UNIVERSIDAD DE CHILE

Web User Behavior Analysis

by

Pablo Enrique Roman Asenjo

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy in Engineering System

in the

Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Industrial

March 2011

"Scientia vincere tenebras"

(VUB's Motto)

Resumen

Desde los orígenes de la Web en el CERN, ha existido una pregunta recurrente entre los investigadores y desarrolladores: ¿Cual es la estructura y contenido correcto para que un sitio Web atraiga y/o retenga a sus visitantes? En parte, la respuesta a esta interrogante, se encuentra fuertemente relacionada con una mayor comprensión de las motivaciones que posee un usuario al visitar un sitio. Tradicionalmente, se han utilizado algoritmos de minería de datos (Machine Learning) para extraer patrones de comportamiento del usuario web, a partir de los cuales se elaboran estrategias para mejorar el sitio Web. El presente trabajo describe un nuevo enfoque, que aplica teorías sobre la neurofisiología de la toma de decisiones para describir el comportamiento de navegación del usuario web. Lo anterior nos lleva a la siguiente hipótesis de investigación: ”Es posible aplicar teorías de la neurofisiología de la toma de decisiones para explicar el comportamiento de navegación de los usuarios Web”. En esta tesis, se propone un modelo estocástico para describir el proceso de navegación del usuario Web, basado la teoría neurofisiológica de la toma de decisiones LCA (Leaky Competing Accumulator), la cual describe la actividad neuronal de diferentes regiones de la corteza cerebral durante el proceso de determinación, hasta que se alcanza un cierto umbral que gatilla la decisión. Esta clase de modelos han sido estudiados y testeados experimentalmente por más de 40 años. De acuerdo al modelo presentado, un usuario web se enfrenta a la decisión de elegir que hipervínculo visitar, conforme a sus propias motivaciones. El proceso se repite en cada visita a las páginas hasta salir del sitio. En el caso del usuario web, la mayor fuente de datos respecto de su comportamiento de navegación y preferencias queda almacenada en archivos de Web Log, los cuales dan cuenta de cada una de las acciones que un usuario ha efectuado cuando visita a un sitio. Dependiendo de la cantidad de visitas del sitio, estos archivos pueden contener millones de registros, constituyendo una de las mayores fuentes de datos sobre comportamiento humano en la Web. Sin embargo, estos archivos también contienen registros que no son necesarios para el análisis del comportamiento del usuario web, por lo que se requiere de una etapa de pre-procesamiento que asegure la calidad de los datos con que se calibrará el modelo. En concreto, se requiere reconstruir las secuencias de páginas visitadas (sesiones) de cada visitante, el contenido de texto y la estructura de hipervínculo del sitio Web. Para estos fines, fueron desarrollados nuevos algoritmos basados en programación entera para la extracción óptima de las sesiones de usuario. Se experimentó con datos provenientes del sitio Web de nuestro departamento (DII) el cual cumple ciertas características acordes con los supuestos del modelo. En cuanto a los algoritmos del pre-procesamiento de sesiones se obtuvo una performance (F-score) del 72% versus un 60% de los algoritmos tradicionales. En relación al modelo de simulación, los parámetros fueron ajustados por medio del método de máxima verosimilitud, usando las sesiones obtenidas. Se concluye que cerca del 70% de la distribución real de sesiones se recupera mediante este método. Este es un importante avance debido a su rendimiento sobresaliente en relación a algoritmos tradicionales de Web mining que alcanzan un 70% de éxito solo en transiciones de un paso, es decir de una página a otra. Las distribuciones de tiempos también alcanzan un gran ajuste a una ley de potencia que también se observa en la realidad. Por lo tanto, se prueba la plausibilidad de la hipótesis.

Abstract

From the very beginning of the origins of the Web at CERN, one big question remains unanswered: What is the optimum structure and content of a web site in order to attract the maximum interest of visitors? The answer to this question is closely related to the understanding of web user purposes and their motivations for visiting a web site. It is clear that by having more accurate knowledge about the interests and preferences of web users, better content and structure can be offered. A web site service can succeed in personalizing the web user's experience. The analysis of human behavior has been conducted within such diverse disciplines as psychology, sociology, economics, linguistics, marketing, and computer science. Hence a broad theoretical framework is available, with a high potential for application into other areas of knowledge, in particular the analysis of web user browsing behavior. These previous disciplines use surveys and experimental sampling for testing and calibrating their theoretical models. For the web user the major source of data comes from web logs, which store every visitor's action on a web site. Such files could contain millions of registers depending on the web site traffic, constituting a major data source about human behavior. The present work describes a novel approach by applying a neurophysiological theory of decision making for describing web user browsing behavior. The research hypothesis is: "It is possible to apply neurophysiology's decision making theories to explain web user navigational behavior using web data." Such an analysis of web user behavior requires a non-trivial data pre-processing stage, in order to obtain the sequence of web pages (session) for individual visitors, the text content and the hyperlink structure of the web site. Traditionally, the next step consists of applying data mining techniques for identifying and extracting patterns of web user browsing behavior. An important contribution of this work corresponds to the data pre-processing stage. Furthermore, data quality needs to be ensured, since the calibration of web user models is sensitive to the data set. Novel algorithms are developed, based on integer programming, and are used for the optimal extraction of web users' sessions. A dynamic stochastic model of the decision making process is proposed. Such a model is based on the Leaky Competing Accumulator (LCA) model of neurophysiology. It describes the neural activity of different brain regions during the subject resolution of the decision, by means of a stochastic process that evolves until an activity reaches a given threshold that fires the decision. Such a class of stochastic processes applied to decision making has been experimentally studied for nearly forty years. In this context, a web user is confronted with deciding which link to follow according to his/her own purposes, and the process is repeated again for each visited page until leaving the web site. The parameters of the model are adjusted by means of the maximum likelihood method, using the observed sequences of pages and the time spent on each of them. It is concluded, that nearly a 70% of the real distribution is recovered by this method. This is a important advancement since it prove to overcome traditional web mining algorithm. Then the hypothesis is proved to be plausible.

Publication related with this thesis

Exploration related with web usage mining where first presented during the XIII Latino-American Congress (CLAIO 2006) [RV06] and the XIV Latin Ibero-American Congress on Operations Research (CLAIO 2008) [RV08b].

It was presented at CLAIO 2008 a first approach using integer programming for data pre-processing [DRV08a], the 2008 WI-IAT conference [DRV08b] published by IEEE Computer Society, the 2009 IFIP TC7 Congress [DRV09a], the Revista de Ingenieria de Sistemas published by the Institute of Complex System [DRV09c], and in the Procs. of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES 2009 [DRV09b] published as Lecture Notes LNAI 5711 by Springer.

An invited lecture in the 6th Atlantic Web Intelligence Conference (AWIC 2009) published as part of the Advances in Intelligent and Soft Computing Series by Springer on the Advances in Intelligent Web Mastering-2 book [RV09b], presented a stochastic model of web user. The novel Physiological based model was presented [RV09a] on the AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA 2009) in Arlington-USA and published as technical notes of the AAAI. The calibration model was presented at the [RV10b] WI-IAT conference Toronto-Canada Published by IEEE Press and in local congress [RV10c, RV10a].

Two book chapters have been published in the book "Advanced Techniques in Web Intelligence-1" by Springer [RLV10, RDV10] related to Web Usage Mining and referred to pre-processing techniques. A journal paper [RDVL10] was submitted to the international journal of Intelligent Data Analysis consisting in a large scale study of the sessionization based on integer programming.

Further exploration with other models were presented in [ARRV10, LRV10b] as a engineering thesis co-guidance.

Acknowledgements

First, I want to thank my wife Yumi for her kind comprehension and support of my PhD work. Performing doctoral studies impose severe restriction to the family group, thus Yumi faced with decision the issues and help me to accomplish my goals. My son Felipe was born during my doctoral work and i devote to him the accomplishment of this thesis.

My advisor's Dr. Juan Velasquez has been the main supporter of my doctoral activities. Many unexpected obstacles appeared during the development of this thesis work, and Juan faced them all with notable enthusiasm. Without his strong support, this thesis would never comes to light. He also guided me in the art of scientific publishing and working, and I must acknowledge him for his great contributions to my formation.

I thank very much Dr. Robert Dell, who kindly guided me in the field of applied operation research when pre-processing techniques were developed. We performed several papers together, enabled by the synergy originated in the application of operation research techniques to the web mining field.

I greatly acknowledge Dr. Víctor Parada and Dr. Sebastián Ríos for their support and thesis revision.

I would like to thanks the faculty for my first year grant support. This first financial aid helped make stronger the decision to engage in the doctoral adventure. I thank to Conicyt for award me with the national doctoral grant that allows me full dedication for research activities. I greatly acknowledge the Complex Engineering System Institute and the associated doctorate program that yield financial aids for conference assistance and economic support during time extension of the thesis work. I acknowledge FONDEF by mean of project DOCODE D08I1015 for financial aids at the final stage of the thesis.

Contents

List of Figures

List of Tables

Abbreviations

BCM	**B**ipartite **C**ardinality **M**atching
CEL	**C**hoice **E**vidence **L**evel
DES	**D**iscrete **E**vent **S**ystem
DFT	**D**ecisio **F**ield **T**heory
LCA	**L**eaky **C**ompeting **A**ccumulator
LIP	**L**ateral **I**ntra **P**arietal cortex
ML	**M**achine **L**earning
MT	**M**iddle **T**emporal
NAL	**N**eural **A**ctivity **L**evel
OU	**O**rnstein **U**hlenbeck
QDT	**Q**uantum **D**ecision **T**heory
RUM	**R**andom **U**tility **M**odel
SIP	**S**essionization **I**nteger **P**rogram
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
WUM	**W**eb **U**sage **M**ining

Symbols

I	Choice Evidence Level (**CEL**) Vector	[Spikes/s]
W	iid White Noise Vector	[Spikes]
X	Neural Activity Level (**NAL**) Vector	[Spikes/s]
σ^2	White Noise Variance for Stochastic Process	$[s^{-1}]$
κ	Lateral Inhibition	$[s^{-1}]$
λ	Activation Decay	$[s^{-1}]$
β	CEL's Normalization Factor	$[s^{-1}]$
o	Label for the order in a session	
r	Label for a register in a log file	
i, j	Label for a page (choice) in a web site	
X_{ros}	Binary integer variable for sessionization.	
$N(a, b)$	Random variable with normal distribution of media a and variance b^2	
V_i	Utility for choice i	
μ	Text preference vector	
F	Force vector in the LCA model	
$\phi(X, t)$	Probability density of having the neural system on the state X on time t	
Ψ	Absorbing surface	
Δ	Reflecting surface	
Ω	The domain of the neural state variable X	
$p(i, t)$	Probability of choosing the option i on time t.	

Dedicated to my wife Yumi and my son Felipe

Chapter 1

Introduction

This study presents a dynamic model of human decision making behavior. The current approach focuses on a narrow class of individuals: web users. From mathematical psychology studies several models have been presented on this subject, but few have so far been applied to the engineering field. The primary purpose of this research is to predict changes in the navigational behavior of the web user based on historical data. Many other fields have this challenge as their purpose: for example economics attempts to predict consumer/producer demand/offer; sociology attempts to predict mass behavior. The restricted Web User/Site system is in the end a human-machine interaction system. The importance of this study lies in revealing the plausibility of its further extension into other areas. The introductory chapter furthers this novel research area by reviewing other approaches, machine learning techniques and applications.

1.1 Relevance to Human Behavior

Since the epoch of early civilization, humanity has faced the challenge of understanding itself. Traders anticipate peoples' needs, politics calculates the move with the best political outcome and generals decide the position of the army. Human beings live together in societies constituting complex systems of interdependence. One stepping stone for the construction of a better society consists in having sufficient knowledge of human behavior.

Economics relates to understanding the way that resources and agents participate socially in the production, exchange, distribution, and consumption of goods and services. Today, government economic departments are one of the most critical administrative areas of a country. The current "sub-prime mortgage crisis" relates directly to past economic decisions, having as consequences losses on the order of several trillion dollars in the financial markets [Ely09]. This discipline relies on describing the system of an agent's decision for consumption or production of goods

and the system's resultant equilibrium. More recently, new research directions on incorporating sociological and psychological models into economics have produced the field called behavioral economics. The 2002 Nobel Prize winner Daniel Kahneman received the distinction for "having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty" [Kah03]. Furthermore, a more accurate understanding of the economic agent's behavior will help in controlling the stability of markets.

Social science has recently used many modern tools like dynamic and stochastic systems for describing the structures of social change [CFL09]. Every social system has naturally been described as a highly complex network of interaction that has been considered impossible to represent mathematically. Nevertheless, with the help of abstraction many models have been designed for explaining particular social phenomena. For example, applications to politics are revealed by the modeling of the mass's opinion dynamics. A model for voter dynamics [SMS08] could help to understand the mass's voting intentions.

In a business environment, marketing is the set of processes that helps to determine customer preferences and demand for products and services, recommending strategies for sales and communications. Google is a company that bases its main business on selected publicity on a search result page. Google's marketing strategy is based on web users' preference rankings for web pages. A simple stochastic model for web user browsing (Page Rank algorithm [BP98]) generates stationary probabilities for calculating this ranking. The exact formulation of the current algorithm used in Google Search remains secret, since it generates 6,520 millions US$ of net profit per year (2009 [Goo09]). Therefore, the Internet has become a new frontier in the marketing field that promises commercial success, but at the cost of the need to have accurate knowledge about the web user.

Some successful examples in e-commerce like Amazon could be mentioned. Amazon is a USA company that is mainly designed as an online book store, but has also moved into trade in electronic devices, furniture and other goods. Amazon dot com is considered to be the first promoter of online shopping with prosperous business results. Its annual net incomes were about US$ 902 million in 2009. Amazon relies on online recommendations to customers according to their detected pattern profiling. Such technologies are based on predicting the preferences of a web user based on his/her observed navigational behavior.

Netflix is another dot-com company with US$ 115.8 million in 2009 net income, which focuses on DVD movie rental. This company in 2006 offered a one million-dollar contest for improving by ten percent the effectiveness of its movie-recommending algorithm. However, the prize was only claimed in September 2009 after roughly four years of worldwide competitors' attempts [Loh09]. The winner used a very specific data mining algorithm [Wol92] similar to many others. The main conclusion is an example of the difficulty of modeling human behavior. Despite the

current advances in new algorithms, the problem remained unsolved for 4 years. This thesis presents a novel point of view based on well-proven physiological mathematical models about human decision making. This approach intends to apply the results of other fields to the realm of engineering.

1.2 The Web User/Site System

In order to study web user navigational behavior it will be important to clarify the system first (Figure 1.1). Web users are considered human entities that, by means of a web browser, access information resources in a hypermedia space called the World Wide Web (WWW). Common web users' objectives are information foraging (looking for information about something), social networking activities (e.g. Facebook), e-commerce transactions (e.g. Amazon Shopping), bank operations, etc. On the other hand, the hypermedia space is organized into web pages that can be described as perceived compact subunits called "web objects." The design of web pages is created by "web masters" that are in charge of a group of pages called a "web site." Therefore, the WWW consists of a vast repository of interconnected web sites for different purposes.

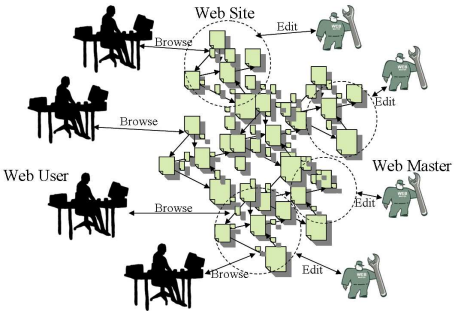

FIGURE 1.1: First level description of the Web User/Site System.

On a second level of abstraction, web pages are distributed by "web servers," and "web browsers" are used to trigger the mechanism. A web page is an HTML-encoded document, which contains hyperlinks to other pages. The content of a web page corresponds to the visual text and multimedia that a web user perceives when the browser interprets the HTML code. The web structure corresponds to a graph of pages interconnected by hyperlinks. Actually, both content and structure have been highly dynamic since the web 2.0 application began personalizing web sites to the current web user like Facebook or Amazon. In this context, the system constituted by web users that navigate on web sites has a highly complex description.

The framework [BCR06] was the first application of a model of human behavior on the web like the Page Rank algorithm [BP98]. This model has had a very positive reputation since it was first applied to the Google search engine. It anticipates the user visiting any link on the page with equal probability. This stochastic model also includes the probability of restarting the

process by recommencing the navigation in another uniformly distributed page that results in an ergodic Markov chain. This simple stochastic process adheres to stationary probability rules. In this sense the most important or interesting pages carry the highest probability of a random user visiting it. This process creates a ranking for pages used in web search engines.

Further refinements of this idea are used in [Tom98], where web users are considered as flows over the network of hyperlinks. Other approaches relate to models for evaluating the navigability of web sites. In particular [ZLW07a], incorporates common browsing actions that a web user performs such as: terminating a session, proceeding to, going back, staying and jumping to. Each one of these actions relates to a probability value that is incorporated into a Markov chain model.

1.3 Data Mining

Data mining is therefore the automated extraction of predictive information from generally large databases. This definition has an implicit assumption that a statistical methodology is used. Data mining is by nature proactive in the sense that the user must tune the algorithm, because in the process some unexpected pattern is discovered that the miner has to interpret. Nevertheless much of the prospective effort can be automated. Data mining is automated pattern discovery on data, and for such purposes machine learning algorithms are used. Data mining in fact uses generic tools and does not rely on the internal physical processes from which data is generated. However, once patterns are discovered, machine learning algorithms can be used as a predictive modality. Successful applications on business and science have been performed such as credit scoring, weather prediction, revenue management, customer segmentation, and many others.

Nevertheless, data mining is not a silver bullet for many problems including human behavior. The hierarchy cascade data mining process has shown that many human-controlled iterations are needed for finally adjusting a model to data and to applying it in a predictive mode. Nowadays new computer improvements allow having a more automatic process for adjusting machine learning models which make up the new generation of intelligent applications. An editorial in Wired magazine entitled “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” claims that current intelligent applications are quite powerful enough to handle any given real-world complexity, making many theories obsolete [And08]. Of course the explosive increment of data had an impact on the precision of many modeling techniques. Yet the proposal seems too optimistic to be true, and there are some fundamental reasons for discarding this extremely naive proposal.

First, the automatic intelligent method improves the accuracy of predictions by training models on the available data. The problem is that future scenarios need to belong to the same kind of

data. For instance if we try to model the trajectory of the Apollo mission by a machine learning approach, perhaps the model will perform well. But if a new factor like cosmic dust enters into the scenario, the machine learning model will fail. On the other hand Newton's theory does not change in this case and the new factor can be included as a viscosity term on the theoretical model. Second, the large quantities of data presented here do not compare with the large quantities of data that could be extracted from nature. A simple fluid motion in a box contains an Avogadro number of particles ($N = 10^{23}$) and a much larger number of possible configurations ($N!$), therefore the problem is that if a computer were capable of handling such an amount of information, it would have to be constructed of a number of memory pieces much larger than the estimated number of atoms in the universe. Of course, fluid theory can model the average motion of particles in a fluid, and this is the best-known way to handle this problem. Third, intractable computational complexity is easy to find in a simple natural system like the traveling salesman problem, since it belongs to the class of NP-Hard problems. Nevertheless, there are several heuristics based on machine learning methods, but they show average performance in some cases and worse performance in others.

The Internet has become a regular channel for communication, most of all for business transactions. Commerce over the Internet has grown to higher levels in recent years. For instance, e-shopping sales have been dramatically increasing recently, achieving a growth of 17% in 2007, generating revenue of $240 billion/Year in the US alone [GD08]. This highlights the extent of the importance of acquiring knowledge on how the Internet monitors customer's interactions within a particular web site.

One can analyze this new technological environment using traditional marketing approaches, but the Internet invites new methods of determining consumer's genuine needs and tastes. Traditional market surveys serve no purpose in reflecting the actual requirements of customers who have not been precisely defined in the web context. It is well known that web users are ubiquitous. In that sense, a marketing survey compiled in a specific location in the world does not carry clear statistical significance in another. However, online queries should improve this issue by requesting that each visitor answers as many focused questions as they can [Kau09], but apart from predicting future customer preferences, online surveys can improve the effectiveness of web site content strategy.

According to the WIC (Web Intelligence Consortium) the term "Web Intelligence" corresponds to "Research and development to explore the fundamental roles and impact of Artificial Intelligence and advanced Information Technology on the next generation of web-empowered systems, services, and activities." In this context, Web usage mining (WUM) can be defined as the application of machine learning techniques on web data for the automatic extraction of behavioral patterns from web users. In this sense, web usage patterns can be used for analyzing web user preferences. Traditional data mining methods need to be pre-processed and adapted

before being employed on web data. Several efforts have been made to improve the quality of the resulting data. Once a repository of web user behavior (Web Warehouse) is available [VP08], specific machine learning algorithms can be applied in order to extract patterns regarding the usage of the web site. As a result of this process, several applications can be implemented on adaptive web sites, such as recommender systems and revenue management marketing, among others.

1.4 The Hypothesis

This thesis focused on the following hypothesis:

It is possible to apply neurophysiology's decision making theories to explain web user navigational behavior using web data.

This declaration includes the proposals:

- **To use web data to study of human behavior on the web**: Traditional studies in psychology obtain data from wired-brain subject and surveys. In this case, web data is a enormous repository of human activities on the web.
- **To use integer programming techniques for better solving the combinatorial problem of session reconstruction**: Sessionization has been traditionally heuristic. Formal optimization methods improve accuracy of the retrieved sessions.
- **To adapt a psychology's model of decision to explain web user activities on the web**: Neurophysiology's model are mostly used for research in Psychology and few application have been today formulated.
- **To enable a simulation model that reproduce web user trail in a web site**: Such application must fit after a large number of simulation the observed session distribution.
- **To propose a mechanism for parameter's fitting of the model**: The problem correspond to a high dimensional differential problem that suffers from the "curse of dimensionality." Approximations that solve this issue are proposed.

This thesis pretend to expose the feasibility of the application psychology's models in complex engineering systems as the Web. Next generation of web-based system will be strongly founded on web user behavior characterization, then this thesis gives a framework for such purposes. Further applications could be envisaged since human behavior is a part of other complex system fields.

1.5 The Proposed Model

While current approaches for studying the web user's browsing behavior are based on generic machine learning approaches, a rather different point of view is developed in this thesis. A model based on the neurophysiology theory of decision making is applied to the link selection process. This model has two stages, the training stage and the simulation stage. In the first, the model's parameters are adjusted to the user's data. In the second, the configured agents are simulated within a web structure for recovering the expected behavior. The main difference with the machine learning approach consists in the model being independent of the structure and content of the web site. Furthermore, agents can be confronted with any page and decide which link to follow (or leave the web site). This important characteristic makes this model appropriate for heavily dynamic web sites. Another important difference is that the model has a strong theoretical basis built upon physical phenomenon. Traditional approaches are generic, but this proposal is based on a state-of-the-art theory of brain decision making.

The proposal is based on the model named LCA (Leaky Competing Accumulator) [UM01]. This model associates the neural activity levels (NAL) of certain brain regions with a discrete set of possible choices. Those NALs (X_i) evolve according to a stochastic equation (1.1) during the agent's decision making process until one of the NAL's values reaches a given threshold (Figure 1.2). The stochastic equation depends on the choice evidence levels (CEL's). A CEL (I_i) is a neural activity level of a brain region that is associated with a unique choice, and whose value anticipates the likelihood for the choice before the decision is made.

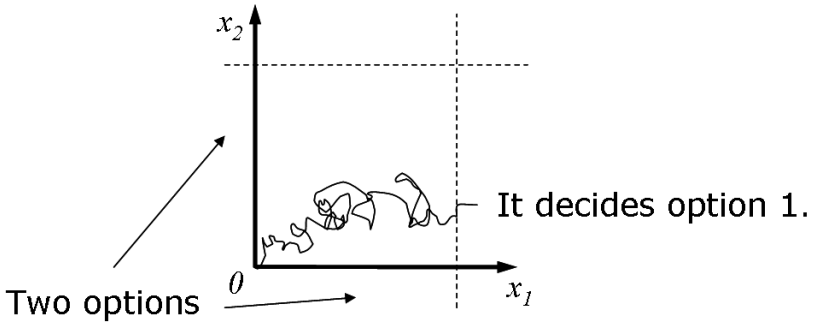

FIGURE 1.2: A diffusion-based decision making model. The first coordinate to reach the threshold corresponds to the final decision.

Models like the LCA stochastic process have a long history of research and experimental validations, most of which have been carried out in the last 40 years [Lam68, Sch01, Sto60a, Rat78]. However no engineering application has been proposed until now. This work assumes that those proven theories regarding human behavior can be applied and adapted to describe web user behavior, producing a more effectively structured and specific machine learning model. The approach consists in applying the LCA model to predicting the web user's selection of pages (session). This proposition was based on experimental validation. A web user faces a set of

discrete decisions that corresponds to the selection of a hyperlink (or leaving the site) as in figure 1.3.

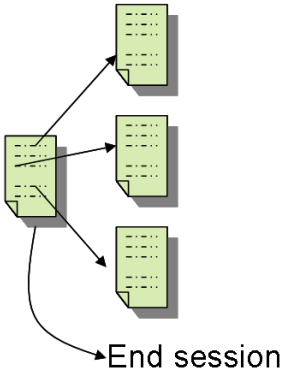

FIGURE 1.3: Navigational options on a web page.

Here, the LCA model simulates the artificial web user's session by estimating the user's page sequences and furthermore by determining the time taken in selecting an action, such as leaving the site or proceeding to another web page. Experiments performed using artificial agents that behave in this way highlight the similarities between artificial results and a real web user mode of behavior. Furthermore, the performance of the artificial agents is reported to have similar statistical behavior to humans. If the web site semantic does not change, the set of visitors remains the same. This principle enables the predicting of changes in the access pattern to web pages related to small changes in the web site that preserve the semantic. The web user's behavior could be predicted by simulation and then services could be optimized. Other studies on ant colony models [AR03] relate directly to general purpose clustering techniques.

The neurophysiology of decision making [UM01, BBM+06] and the Random Utility Model of discrete choices [McF73] are the bases of the model. In the field of mathematical psychology, the Leaky Competing Accumulator (LCA) model describes the neurophysiology of decision making in the brain [UM01]. It corresponds to the time description of the subject neural activity of specific zones i in the subject's brain.

$$dX_i = (-\kappa X_i - \lambda \sum_{j \neq i} f(X_j) + I_i)dt + \sigma dW_i \tag{1.1}$$

The dynamics of the web user are intrinsically stochastic. For each decision i a region in the brain is associated, which has a neuronal activity level (NAL) $X_i \in [0, 1]$. If a region i_0 reaches an NAL value equal to one then the subject makes the decision i_0. The NAL's X_i is time dependent, which dynamic is stochastic as shown in the equation 1.1. The coefficients are interpreted as: κ which is a dissipative coefficient, λ is related to competitive inhibition between choices, I_i is the supporting evidence of choice i and σ is the variance of the white noise term dW_i. The function $f(.)$ corresponds to the inhibition signal response from other neurons, usually modeled

as a sigmoid (near to linear) or linear in this case ($f(x) = x$). The parameter I_i in the LCA theory is interpreted as likelihood values regarding the importance of choice i for the subject. Other diffusion models have been proposed but all have been proven to be equivalent to LCA [BBM+06].

Web users are considered stochastic agents [RV09a, DRV09c]. Those agents follow the LCA stochastic model dynamics (Equation 1.1), and maintain an internal state X_i (NAL's values) with some white noise dW_i. The available choices lie in the links on a web page, including the probability of leaving the web site. Agents make decisions according to their internal preferences using a utilitarian scheme.

The CEL's values are the forces that drive those equations (1.1). Furthermore, those values are proportional to the probability $P(i)$ of the discrete choices ($I_i = \beta P(i)$), which are usually modeled using the Random Utility Model (RUM). Discrete choice preferences have been studied in economics to describe the amount of demand for discrete goods where consumers are considered rational as utility maximizers.

The utility maximization problem regarding discrete random variables results in a class of extreme probability distributions, in which the widely-used model is the logit model (Equation 1.2) and where probabilities are adjusted using the known logistics regression [NM94]. The logit probability distribution $P(i)$ anticipates every possible choice on the page j and has a consumer utility V_j.

$$P(i) = \frac{e^{V_i}}{\sum_{j \in C} e^{V_j}} \tag{1.2}$$

The logit model has been successfully applied to modeling the user's quest for information on hypertext systems [Pir09] resulting in an advancement for adaptive systems. The utility function should depend on the text present in links that the agent interprets and through which it makes the decision. Hence the assumption is that each agent's link preferences are defined by its TF-IDF text vector μ [MS99]. The TF-IDF weight μ_k component represents the importance for the web user for the word k. Furthermore, an agent prefers to follow similar links to its vector μ. The utility values (equation 1.3) are given by the dot product between the normalized TF-IDF vector μ and L_i that represents the TD-IDF weight text vector associated with the link i.

$$V_i(\mu) = \frac{\mu \bullet L_i}{|\mu||L_i|} \tag{1.3}$$

The resulting stochastic model (equation 1.1) is dependent on the parameters $\{\kappa, \lambda, \sigma, \beta, \mu\}$ and the set of vectors $\{L_i\}$. The first four parameters must be considered as universal constants of neurophysiology, yet the μ vector is an intrinsic characteristic of each web user. In this sense,

the real web user's mode of behavior as observed within a web site corresponds to a distribution of users.

Collected data from a real web server contains the behavior of a variety of different users. A parameter inference should be performed on the distribution of the μ vectors to discover the web user's preferences. A web user is considered memoryless, making decisions without considering the previous pages visited, but having a purpose driven by μ. A special link corresponding to the decision of leaving the web site is presented on every page, with a fixed probability transition analogous to the random surfer teleportation operation [BCR06]. Each artificial user ends up following a trail $((p_1,t_1),...,(p_L,t_L))$ of pages $\{p_o\}$ with the visitor's time durations on the site $\{t_o\}$, until the moment the user decides to leave the L step.

1.6 Real-World Applications

Current web 2.0 applications [UBL+08] are highly popular on the Web and have enormous potential for commercial use. Nevertheless, the analysis of web usage is obfuscated by the dynamic characteristic of the content and structure of web sites. Today's applications of web mining have been fully explored in terms of both fixed content and structure with the help of natural language processing techniques and specialized machine learning algorithms. A recommender system is a typical web intelligence application. Recommendation is based on a web user profile recognition procedure, which filters information for content on a web page. Such a system is based on detected common patterns of web user browser behavior.

In this thesis a mechanism is proposed which is based on simulations for updating visitor patterns and predicting future behavior. The proposed system is based on the ability of the decision model 1.1 to reproduce the sequence of visited pages. Once the model presented in the previous section is calibrated, it can be possible to obtain by Monte Carlo techniques the distribution of navigation trails. As a sub-product of the calibration mechanism it results in the dispersion of web users' keyword interests. By using both tools it is possible to build an automatic mechanism for giving the best recommendation to web users in order to enhance the web site experience.

1.7 Aim and Scope

This thesis focuses on the study of a model of the web user's browsing behavior based on the state of the art of the neurophysiology of decision making. The general objective is to develop a stochastic model of the web user's behavior, with the goal of analyzing the net variation of navigational preferences on changes in web site content and structure. For these purposes,

new techniques and algorithms for data pre-processing will be developed in order to ensure data quality. The simulation topic is important both in the context of the stochastic model and experimentation. Mathematical analysis of the system is performed upon perturbation of the system. Finally, new perspectives are identified, since models of human behavior on the web could be applied to other human activities.

1.8 Organization of this Document

This thesis presents a novel model for the web user's behavior incorporating the results of research from other areas (e.g. psychology). In order to organize a self-contained document on data mining web usage research, the perceptual decision making and Random Utility Models are reviewed. An extensive description of data pre-processing is included since it is fundamental for calibration operations. Finally, the model is presented and experimental results are examined. The descriptions of the individual chapters are:

- **Chapter 2, Related Work** : In this chapter traditional data mining approaches are described and applications are explained.
- **Chapter 3, Data Pre-Processing** : This chapter presents current techniques for data pre-processing and cleaning related to web user navigational behavior. A novel approach based on integer programming is discussed.
- **Chapter 4, Psychology of Decision Making** : A review of psychological research on decision making is presented, with emphasis on the LCA model [UM01]. It also includes the Random Utility Model approach, which was originally covered in the field of psychology by Luce [LS65] and later by McFadden in the field of economics [McF73].
- **Chapter 5, A Stochastic Model of the Web User** : The core mathematical description of the proposed model is presented.
- **Chapter 6, A Framework for Web User Simulation and Model Calibration** : The software design of a simulator for the model is described. Any performed simulation needs to have a web site structure and content, but the parameters of the model also need to be explicitly given. The methodology for fitting the parameters of the stochastic equation is presented.
- **Chapter 7, Experimental Results** : The whole methodology presented in this thesis is tested on the web site of the Department of Industrial Engineering of the University of Chile. The web site, because of its simpler design is ideal for testing the theory and models presented here in an elementary fashion.

- **Chapter 8, Conclusion and New Perspectives** : Experimental results are analyzed and concluding remarks are presented. Many of the assumptions can be avoided with some adjustment for making the procedure more general. New perspectives are open, since the human behavior model can be applied in a variety of other fields.

Chapter 2

Related work

This thesis relates directly with the web usage mining (WUM) community, psychology of perceptual choice community and econometric methodology. Nevertheless, the physical environment where the theory is developed corresponds to the Web realm. This multidisciplinary conjunction produces a novel point of view for the analysis of the web user. In the following chapter a review of the related work of the converging disciplines is presented.

2.1 Characterizing the web user browsing behavior

As described in [JZM04, MPTM08, SCDT00, THLC09, VP08], Web usage data can be extracted from different sources, with web logs considered one of the main resources for web mining applications. The variety of different sources carries a number of complexities in terms of data pre-processing and furthermore these are associated with the incompleteness of each source. As a solution to this, several pre-processing algorithms have been developed [VP08]. Further sources like the hyperlink structure and the web content complement the extracted Log information, providing a semantic dimension of each user's action. Furthermore, in terms of the web log entries, several problems must be confronted. Overall, a web log in itself does not necessarily reflect a sequence of an individual user's documented access. Instead it registers every retrieval action but without a unique identification for each user.

2.1.1 Representative variables

The web user browsing behavior can be described by three kinds of data: the web structure, the web content and the web user session. The first is directly related with the environment. The third describes the click stream that each Web User performs during his visit to the web site.

- **The web structure:** A Web Site can be represented as a directed graph $G(N,V,T)$, consisting of a collection of n nodes $N = \{1,\ldots,n\}$ and vertexes $V = \{(i,j)/$ *a web link point from i to* $j\}$ with text content $T = \{T_i\}$. A node i from G correspond to a web page with text content T_i , The representation of the content will be described later. Two special nodes need to be individualized, as they have no correspondence with any real page. This is because they represent the exit/entrance to the web site and each node consists of a link to the "exit or entrance" node. This representation has the advantage of explicitly including all transition between nodes, which is useful for stochastic process descriptions.

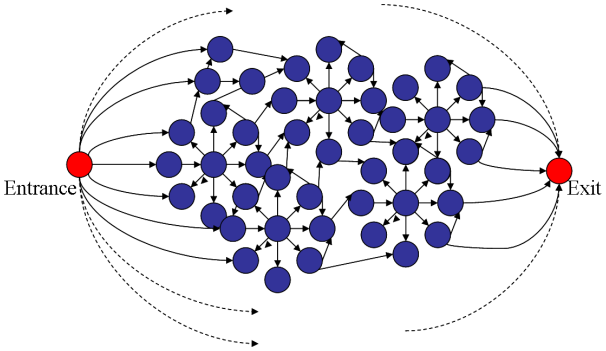

FIGURE 2.1: Static graph web site structure representation.

Nevertheless, this description of a Web Site can be considered as a first approximation of the real hyperlink structure. The notion of a web page consisting of static content and unique URL can not fit the dynamical case. Web sites are being continuously updated, persistently modifying links and content, including those that depend on web user profile adaptive Web Site changes . On static pages with frames this concept is also challenged since the page is a composite. As stated in the Internet Report of China [GD09], the rate of web pages that moves from static to dynamic content is close to one.

Considering web 2.0 sites , the latter model seems obsolete. However, this approximation is valid under some general circumstances. Informative sites that are updated on a regular periodic basis like those of newspapers can be represented under this graph representation within a definite period of time. Blogs can also be represented as such considering that the replies are increasingly added due to the fact that their nodes consist of post and reply.

More generally a time dependent graph structure of a variety of web objects and its associations can be defined for general purpose sites. For application, the analysis should be adjusted to the simpler model depending on the particular web site. Other representations of the structure of the web site are discussed in the data pre-processing Chapter.

- **The web content:** It corresponds to the web user perception semantic information of each visited page from a Web Site. On the earlier Internet this data corresponded mainly to the text content. Nowadays Web 2.0 sites represent a much more complex picture. The web content is more dynamic and constituted by a rich variety of media (text, images,

video, embedded application, etc.). Web pages are composites of web objects that have semantic values. Several valuations of the semantic have been proposed and are revised in the chapter which describes semantic issues.

Natural language processing for semantic extraction has been a large subject of study from when information retrieval systems began and it is still a large unsolved problem for reliable and automatic operational system [MS99]. Despite its limitation, some approximations to the problem have been proposed based on special representation of text and similarity measures at times have reasonable results. Instead of extracting the exact semantic, the notion of similar semantics between texts has demonstrated a more intelligent approach. The representations proposed are assigned to each term p in a page j, a weight ω_{pj} representing its semantic relevance. In this sense, a column vector d_j of this matrix ω represents approximately the text semantic of the page j and row vector $t_p^\dagger$ represents the term p semantic related with the set of documents. This approach is also termed "Bag of Word" because it does not take into account the semantic relations and syntax of phrases. In this way every equal word has the same importance independently of the context. Similar approaches can be extended to non-textual content like web object, where meta-data play a fundamental role in representing the semantic.

Furthermore, dynamic content implies time and user dependence of the semantic. As web applications become more complex than the standard representation of the content the semantics become more inaccurate. Specific semantic context that group a variety of content must be tailored for each specific application.

- **The web user session:** Web User visits a Web Site represented by the browsing trajectories that is categorized as a session [SMBN03]. A session s is a sequence $s = [(i_1,\tau_1),\ldots,(i_L,\tau_L)] \in S$ of pages $i_k \in N$ and time $\tau_k \in \mathbb{R}^+$ spent by a Web User. The size $L = \|s\|$ of a session corresponds to the number of nodes without considering the sink and source. In this representation the time associated with both the source and sink nodes and the duration of a session $\mathcal{T} = \sum_k \tau_k$ is the sum of all visitor's times spent on the site.

  Nevertheless, if sessions are not explicitly given they must be reconstructed from other sources like web logs. When they are specially retrieved, some privacy concerns [MS08, JJJ06] arise that complicate its implementation. On the other hand, session retrieval from web logs have less privacy issues since the data is stored anonymously. The process of extracting sessions has been reviewed in the chapter entitled data pre-processing.

Nevertheless, web data has some concerns. However, a further problem is associated with this data: the high dimensionality. Data mining algorithms suffer from the so called "curse of dimensionality" phenomenon. Over the years the processing of such data has been specialized as the Web Mining discipline, in particular when the purpose of such analysis is related to the behavior of the Web User. In such cases, it is called Web Usage Mining.

Also there are some evident problems connected with the web usage data. For example, the high diversity of some web pages; search engines that allow users to directly access some part of the web site; a single IP address with single server sessions; single IP address with multiple server sessions; multiple IP address with single server sessions; multiple IP address associated with a single visitor; and multiple agents associated with a single user session [VP08]. Additionally, a user's activation of the forward and reverse browser button is often not recorded in the web log because, in most cases, the browser retrieves the page from its own cache. A proxy server, acting as an Internet web page cache serves to reduce network traffic, and can also capture web requests that are not recorded in a web log [Gla94].

Browsing data has been recently considered in WUM, where the scroll-bar, select and save-as user interactions with the web site [THLC09, THS08]. Furthermore, semantic considerations have been proposed by different authors, where Plumbaum et al. in [PSK09] uses the open standard of Microformats in order to add semantic information on the web page. This is a similar method as proposed in [SSST07], for which JavaScript events (gathered with AJAX) are associated with key concepts in the portal, providing a context of such events and linking valuable usage information with the semantic of the web site.

As stated in [HMS09], WUM presents different challenging problems in terms of the pre-processing of the usage data. Web-logs are larger in size, for both data volume and dimensionality, where sparse data set representations of web users are needed to be transformed into a more accurate user behavior representation. One of the problems associated with this lies in the high dimensionality, which increase the computational complexity of the mining process. Secondly, the data scarcity results in the mining algorithms have the ability to extract meaningful and interesting patterns in the user browsing behavior.

Each WUM technique requires a model of user behavior per web site in order to define a feature vector to extract behavior patterns. Usually, the model contains the sequence of pages visited during the user session and some usage statistics, like the time spent per session, and pages viewed, amongst other information gathered. Now difficulties can be encountered when a page is loaded into a given web browser, the request for the web site objects can be logged separately, for which a series of page can be viewed associated with the same session.

All of the latter leads to the fact that web usage data needs different pre-processing techniques before analyzing user behavior [SCDT00]. However, one of the main tasks present in WUM is the determination of the web user sessions based on the web usage data collated from a given web site (sessionization). It is well known that strategies for sessionization can be classified as reactive and proactive [HNJ08].

Proactive sessionization strategies capture a rich collection of a web user's activity during the visit to a given web site. However, this practice is considered invasive, and even forbidden in

some countries [SMBN03], or regulated by law to protect the user's privacy [Lan00]. Examples of these methods include cookie oriented session retrieval [BMSW01], URL rewriting [FL03], and web tracking software, close to spyware, installed on the user's computer (or browser) to capture the entire session [MS08].

Reactive sessionization strategies have less privacy concerns because they are design to use only the web log entries information, which excludes explicit user information [SMBN03]. However, a web log only provides an approximate way of retrieving a user's session for previously stated reasons. This reinforces the need to reconstruct a user's session from the information available (sessionization). Prior work on sessionization has relied on heuristics [BHS99, CMS02, SMBN03] which have been applied with a high degree of success on a variety of studies that include web user navigational behavior, recommender systems, pattern extraction, and web site keyword analysis [VP08].

2.1.2 Empirical statistics studies about web usage

Human behavior shows some predictive regularity on the averages, contrary to the free will hypothesis. Some of those regularities are observed on the distributions of session in different kinds of web sites [HPPL98]. With the help of such regular statistics of the human behavior, Web Usage Mining can be tailored to fit those conditions. Several stochastic models have been theorized in order to mathematically explain this result [HPPL98, VOD+06, Vaz05], but nothing is related intrinsically to the physical phenomena. Some other models based on the neurophysiology of the decision making process have also been proposed [RV09b].

Data mining processing on web data should show results that are in agreement with the observed universal probability distributions. As was commented, web user's actions on a web site follow regular patterns. This is additional information can reduce the size of the feature space for machine learning algorithms. When such a reduction is available, algorithms have a narrow region for working resulting in better performance and accuracy. Nevertheless, procedures must be adapted for fitting such statistical constraints [LZY04, OC03]. Understanding those statistics results in a better standard and quality for users [Whi07].

Some important statistical empirical studies are summarized below.

- **Session length distribution**: Empirical studies over different web sites shows that the distribution of session size follows a common shaped function having an asymptotic heavy tail. Following [HPPL98] an Inverse Gaussian distribution ties in well with reality, and it was called the universal law of surfing. In some work a Zipf distribution (power law) has been observed [LBL01] reflecting a real session, but this distribution is used to approximate the Inverse Gaussian since its tails decay much slower than a Gaussian.

Application of this kind of regularities enables the tuning up of systems like web crawlers [ByC04] and session retrieval from log files [DRV08b]. This web usage regularity has also been exploited for mining [LZY04].

- **Information seeking behavior**: While studies focus on algorithms for pattern extraction, few studies relate to web tasks that users perform. Furthermore, the manner in which people seek information through the web can be classified through cognitive styles [WD07]. Studies differentiated two main kind of web user: Navigators (17% of total of users) and Explorers (3%). The other 80% correspond to a large variety of purposes, each one with smaller portion of the total. The first, maintain consistency on the sequence of pages of visited pages. Navigators seek information sequentially and revisit the same sites frequently. The second have highly variable patterns of navigation. Explorers have a tendency to query web search pages frequently, revisit pages several times and browse a large variety of web sites. These statistical studies highlight the impact of classical navigational pattern extraction. Navigators will have the most influence on data regardless of having only 17 percent of total use. Nonetheless, a whole skew distribution of cognitive styles have been reported [WD07], beginning with Navigators and ending with Explorers. The study of this distribution needs to be taken in account for further specialization of usage mining studies. A large study of cognitive information seeking has been investigated [IJ05] showing that context have influence implication for each particular web user behaviors. Others sources focus [Kel07] on the statistics of the task that web users perform. Such taxonomy becomes associated with a web user who perform transactions (e.g. reading emails 46.7%), Browsing (e.g. news reading 19.9%), Fact Finding (e.g. looking for whether 16.3%), Info Gathering (e.g. job hunting 13.5%) and a 1.7% non-classified. These four categories specify a simpler structure concerning the web usage.

- **Web user habits**: Web user's habits have changed since the Internet became more sophisticated. Nevertheless, with current new web 2.0 applications web logs are becoming a much more intricate data source. Web users browsing behavior are changing, using fewer backtracking support tools such as the back button. However there is an increasing usage of parallel browsing with tabs and new windows.

- **The inter-event time for web site visit**: Similar heavy tailed distribution [OWHM07] has been measured on the time spent on pages throughout wide variety of the pages.

2.1.3 Amateur and expert users

Users can be grouped in two classes: experienced and inexperienced or "amateurs" [VP08]. The latter is unfamiliar with the process of accessing web sites and possibly dealing with web technology. Their behavior is characterized by erratic browsing and sometimes they do not find

what they are looking for. The former are users with web site experience and with some standard knowledge of web technology. Their behavior is characterized by spending little time on pages with low interest and thus concentrating on the pages they are looking for and where they spend a significant amount of time. As amateurs gain experience, they slowly become experienced users who are aware of a particular web site's features. Therefore recommendations for change should be based on those users.

On the one hand, amateur users correspond to those unfamiliar with a particular web site and probably with web technology skills [VP08]. Their browsing behavior is erratic and they often do not find what they are looking for. On the other hand, experienced users are familiar with this or similar sites and have a certain degree of web technology skills. They tend to spend little time visiting low interest pages and concentrate on the pages they are looking for on which they spend a significant amount of time. As amateurs gain skills they slowly become experienced users, and spend more time on pages that interest them.

2.2 Representing the web user browsing behavior and preferences

Regarding data mining purposes there are two kinds of structures that are used: feature vectors that correspond to tuples of real numbers, and graph representation where numeric attributes are associated with nodes and relations. The most used representation corresponds to feature vectors with information about sessions, web page content and web site structure. Feature vectors are employed for traditional web mining in an unsupervised a supervised fashion. Furthermore, graph data structure is used for graph mining techniques or rule extraction.

2.2.1 Vector representations

A session is a variable length data structure that is not directly usable for most data mining algorithms. Web user activities can be extracted in several ways for the summarization of a session which is codified by mean of weighting ω_i the usage per page i. The vector $v = [\omega_i] \in V^n \subset \mathbb{R}^n$ has a dimension n corresponding to the number of different web pages. The cardinality of the set $|V| = m$ is equal to the number of collected sessions in the pre-processing phase. The sequence logic of each session is not reflected in this representation. Despite this simplification, this methodology has been used with success in several web mining applications.

Several methods exist to evaluate the weight ω_i. The simplest weighting schema correspond to assigning a binary value $\omega_i \in \{0, 1\}$ represented if the page is used (1) or not (0) on this session [MDLN01]. More information can be incorporated extending the binary passage of the web page by the visit duration fraction. In this case the weight remains in the interval $\omega \in [0, 1]$

for normalization purposes. The fraction of time remaining in a web page is supposed to be an indicator of the quality and interest of the content [MDLN02]. Other weighting measures attempt at using other prior information about pages for measuring the significance of each page [MDLN02].

2.2.2 Incorporating content valuations

Hypermedia can be measured primarily by its text content. Natural Language Processing techniques consist of measuring text by means of different multidimensional representation. The most common valuation is the vector space model, where a document i is represented by a vector $m_i = [m_{ij}]$. Each component represents a weight for the importance of the word j in document i. This model use the "Bag of Word" abstraction, where any sentence structure is disregarded in favor of simple word frequencies. Furthermore, this semantic approximation has shown accurate results for data mining application. Several weighting schemes have been used with different results. The simplest is the binary weighting scheme where $m_{ij} = 1$ if the term j is present in document i. The most used weighting scheme is the TF-IDF that combines the frequency of term j in the documents i and the frequency of document containing the term i. Recently the weight has been constructed with the help of a machine learning feature selection mechanism [LTSL09].

The text weighting scheme m_{ij} enriches the information provided by the feature vector of web usage. The visitor behavior vector $v = [(m_i, \omega_i)]_{i=1}^{n}$, where each component represents the content and page importance.

2.2.3 Web object valuation

Nowadays, web sites become highly dynamic applications. The visual presentation of a web page on a browser can not be identified correctly with a URL. A variable multiplicity of hypermedia could appear on browser presentation for the same URL. Nevertheless, embedded visual entities on web pages seem to be a more reliable concept. An object displayed within a web page is termed a Web Object. Despite the complex semantic analysis of multimedia, meta data is used to define the Web Object within it (Figure 2.2).

Meta data that describes web objects constitute the information source for building the vector representation of content. The user's point of view is the principal research topic from which Web Object techniques have been developed. In this way the content and appearance of a web page is combined for processing. Different ways have been developed to describe web pages based on how the user perceives a particular page.

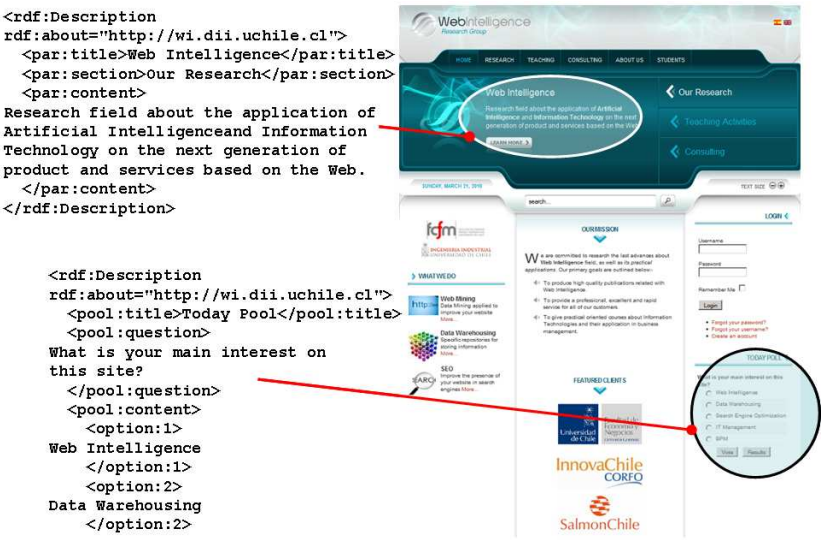

FIGURE 2.2: Some Web Object identified visually on browser and its XML representation.

Web Object research has been carried in the following work: Web site Key objects identification [DV09], Web Page Element Classification [BR09], Named Objects [SK09] and Entity extraction from the Web.

2.2.4 Graph representation

Graph mining uses graph theoretical constructs and algorithms for discovering patterns in data [CF06]. In [GO03] web user trails are converted into a weighted graph using a similarity measure between session. Nodes correspond to sessions and arcs are labeled by the value of the similarity measure between both nodes. This similarity measure is used to overlap degrees between sessions, but any other measure could be adopted. The resulting structure is a direct representation of web user behavior similarity.

2.2.5 The high dimensionality of representation

Usage data correspond to a high dimensional representation. Considering that a medium sized web site contains thousand of pages and there are around ten thousand terms, therefore the feature vector dimensionality correspond to at least 10^4 components. Automatic data mining methods based on similarity measure suffers from the "curse of dimensionality". This problem corresponds to the exponential growth of the size of the search space for a data mining algorithm. Performance issues are one of the principal problems that could render the problem intractable. Recent studies reveal that a distance based algorithm is affected since the numeric difference between distance measure collapse in higher dimensional space. Similarity measure based algorithms (e.g. Clustering) are highly affected by this phenomenon.

Feature selection is a technique for data refinement that has been recently used for alleviating the higher dimensionality problem [ITP07, XdSCP08]. A radical method of reducing the dimension of the feature vector is by way of a supervised approach to the text representation of web pages [RV08a]. In this case an expert categorizes the semantic of the web pages which helps to reduce the dimension of the feature vector.

2.3 Extracting patterns from web user browsing behavior

As defined by Srivastava et al. in [SCDT00], "Web usage mining (WUM) is the application of data mining techniques to discover usage patterns from Web data, in order to understand the needs of Web-based application". Given this, one of the most challenging topics is the understanding and discovery of usage patterns from human users. Moreover, to analyze and predict human behavior is its main characteristic, which is differentiated from Web Structure Mining and Web Content Mining, where techniques used are from a different nature and span between researchers and practitioners. It is relevant to point out that WUM is sometimes referred to as click stream analysis [SCDT00], considered as the analysis and extraction of underlying patterns from an aggregated sequence of page visits from a particular web site user navigation.

The interest in WUM is growing rapidly in the scientific and commercial communities, possibly due to both its direct application to web personalization and the increased complexity of web sites [VP08], and Web 2.0 applications [PSK09]. In general, WUM uses traditional behavioral models, operations research and data mining methods (which will be intensively reviewed in this chapter) to deal with web usage data. However, some modifications are necessary according to their respective application domain, due to the different types of web usage data.

In general terms, two families of techniques have been used to analyze sequential patterns: deterministic and stochastic techniques. Each one of these techniques gathers different optimization methods, from data mining, operations research, and stochastic modeling, where different approaches have been adopted to compensate for the lack of analysis of some of these techniques present.

2.3.1 Clustering analysis

Clustering user sessions can be used, essentially, for grouping users with common browsing behavior and those interested in pages with similar content [SCDT00]. In both cases, the application of clustering techniques is straightforward when extracting patterns to improve the web site structure and content. Normally, these improvements are carried out at site shut-down.

However, there are systems that personalize user navigation with on-line recommendations about which page or content should be visited [PE00, VEY+04]. In [JK00, RB03], the *k*-means clustering method is used to create navigation clusters. In [XZJL01], the usage vector is represented by the set of pages visited during the session, where the similarity measure considers the set of common pages visited during the user session.

Following previously stated reasoning, in [JK00], the pages visited are also considered for similarity with reference to the structure of the web site as well as the URLs involved. In [HWV04] and [RB03], the sequence of visited pages is incorporated into a similarity measure added to the page usage data. In [VYAW04], it is proposed that, despite the page sequence, the time spent per page during the session and the text content in each page is included in the similarity measure.

Also, clustering techniques such as Self-Organizing Features Maps (SOFM) have been used to group user sessions and information that could lead to the correct characterization of the web user behavior [VP08]. In this context, Velásquez et al. in [VWYA04] proposed two feature vectors, composed of the information related to the text content from a web site, and a feature vector related to the usage, including information such as the time spent on the visited page by the web user. With this, authors obtained the relevant words from the web site (or keywords), methodology which was extended by Dujovne et al. in [DV09] to determine the web site key objects.

Graph clustering techniques are applied to graph web data representation; through which [GO03] the time spent on each page was considered for similarity calculation and graph construction. The representation of a session is created by using a similarity measure on the sequence of visited pages. A weighted graph is constructed using a similarity measure, and this in turn is achieved by employing an algorithm of sequence alignment [GO03]. This representation is further processed using a graph clustering algorithm for web user classification. Association rules can also be extracted [DUO07].

Recently, Park et al. in [PSJ08], refers to the relevance of clustering in WUM that aims to find groups whose common interests and behavior are shared. In this work, the question is whether sequence based clustering performed more effectively than frequency based, and thus acquired the best results a sequence based fuzzy clustering methodology proposed by the authors. Likewise, another recent clustering based methodology in WUM was proposed by Rios et al. in [RV08a], where a semantic analysis was developed by considering a concept-based approach for off-line Web site enhancements. Here, the main development was the introduction of concepts into the mining process of WUM, carried out by a hybrid method based on the evaluation of web sites using Nakanishi's fuzzy reasoning model, and a similarity measures for semantic web usage mining.

2.3.2 Decision rules

Decision rule induction is one of the classification approaches widely used in web usage mining [SCDT00]. Its practical results are sets of rules that represent the user's interests. In WUM, the association rules are focused mainly on the discovery of relations between the pages visited by the users [MCS99]. For instance, an association rule for a Master and Business Administration (MBA) program is mba/seminar.html mba/speakers.html, showing that a user who is interested in a seminar tends to visit the speaker information page. Based on the extraction rules, it is possible to personalize web site information for a particular user [MCS00].

2.3.3 Integer programming

For deterministic sessionization, Dell et al. in [DRV08b] present a session reconstruction algorithm based on integer programming. This approach presents different advantages to address linking structure constraints presented by the time sequence of the web log entries, finding the best combinations of paths that fulfill these constraints. Recent advances on integer programming and optimization [Bix02], shows promising for solving some hard combinatorial problem in acceptable timing. This method will be presented with further details in the data pre-processing chapter.

2.3.4 Markov chain models

Several statistical models for web surfing and web user behavior have been developed in [BL00, CZ03, EVK05, JPT03, SF98]. Here Markov models have been proposed for the modeling of the behavior of the web user. In this cases, a web site is considered as an undirected graph of pages, in which the web user passes from one page to another [JPT03, RV08c] with an estimated probability.

On the basis of a large number of users, and taking into consideration that the web site browsing is performed during a large period of time, it is possible to predict transition probabilities between pages. It can be considered that Web users have limited memory about visited pages, for which a transition probabilities can be considered independent of older transition probabilities. If transition probabilities are independent in more than k previous stages then the chain is known as a k-order Markov chain. Let X_o be the page visited at the step o and then there a k-order Markov chain must have property presented in equation 2.1,

$$P(X_o|X_{o-1},\ldots,X_1) = P(X_o|X_{o-1},\ldots,X_{o-k}) \tag{2.1}$$

The latter expression represents a stochastic process with a k-step memory. A k-order Markov chain can be represented as a first-order Markov chain by re-labeling techniques [Res92].

Once the web user behavior is modeled, and its flow probabilities determined, different levels of analysis and information can be extracted. Therefore, the prediction of the session size distribution, sequence of pages more likely to be visited, the mean time that a user can spend on the site or the ranking of the page that is most likely visited, are some of the possible outcomes that could be determined for the decision making process in a given web site.

Usually, a web site has a large amount of web pages and building transition matrices for the Markov model could be costly. For this reason, some authors [AS06] recommend the reduction of the dimensionality using clustering techniques over web pages. Then, the site transitions could be interpreted as web users changing from different clusters. The predictive power of the Markov chain has been studied by Sarukkai in [Sar00], where the next browsing step of a web user is constructed taking the next link as the one with maximum probability. However, some authors proposed higher order Markov models to use for this task [DJ02, ZAN99], considering that lower order Markov models have been found with poor predictive power [CHM+00].

2.3.5 Mixture of Markov models

Mixtures of Markov chains, a discrete set of Markov chains that represents different web user groups with different browsing behavior [CHM+00, SH03], has been proposed as an alternative for the web usage pattern extraction with emphasis on the different types of users a site might have. In this case, the independence relation with the past behavior is represented by an extended mixture coefficient $P(k) = \lambda_k$ into $P(X_o|X_{o-1},\ldots,X_1) = \sum_{k=1}^{K} P(X_o|X_{o-1},k)P(k)$, which represents K different browsing behavior determined by further data mining techniques.

2.3.6 Hidden Markov models

As an extension of the traditional Markov chain modeling, Hidden Markov Models (HMM) have been used to model the stochastic representation of the web site. It considers hidden states of the usage that will represents web user's underlying patterns in their behavior [FHK03, SSQ03, WGH+00]. Overall, this stochastic modeling tool has been used to determine frequent navigation patterns, combining web usage data and WCM techniques to build a predictive model.

2.3.7 Conditional random fields

Other approaches used to predict the web user behavior are based on Conditional Random Fields (CRF) [LMP01]. CRFs are a probabilistic framework generally used for the classification of sequential data. In the WUM context, Guo et al. in [GRP08] aimed to predict all of the probable subsequent web pages for web users, comparing its results to other well known probabilistic frameworks, such as plain Markov chains and HMMs. Later, in [GRP09], an Error Correcting Output Coding (ECOC) of the CRF was proposed for the prediction of subsequent web pages on large-size web sites, extending previous development to a multi-class classification task for the web site prediction problem. It compared its results against single multi-label CRFs which outperformed the proposed method.

2.3.8 Variable length Markov chain (VLMC)

Another probabilistic framework developed for the web user behavior modeling was proposed by Borges et al in [BL07], using as Variable Length Markov Chain (VLMC). There models are based on a non-fixed memory size, for which its usage in web browsing behavior modeling is considerable. In [BL07], the main purpose for the researchers was to incorporate the dynamic extension of some web navigation sessions. They also aimed to extend the plain Markov chain method into a more general predictive tool. In this work, the authors proved that the usage of such techniques increases the prediction accuracy, as well as the summarization ability of the Markov chain, an effective tool for the personalization of web sites, given the web usage data gathered by practitioners.

2.3.9 Ant colony models

Ant colonies have been used to learn the web usage [LRV10a]. The model is based on a simplification of a model based on the neurophysiology of the human decision making [RV09b] and the random surfer [BP98]. The model called the "Ant Surfer" is where agents evolve like in the original random surfer model. The Ant Surfer start in a random page and continue browsing with probability p or return to the nest with probability $(1-p)$. The agent objective is foraging for information that is accumulated and its satiation is modeled with a threshold on the accumulated information utility. When the threshold is reach then it returns to the nest. The agent positioned on the page i select the hyperlink j to follows with probability P_{ij}, that correspond to the Logit model with utility given by the similarity measure between the ant text preference μ and the hyperlink text [RV09b]. This model is applied to extract the web user preference and to predict web usage. Others models relate to other Markovian models for measuring the navigability of a web site [ZLW07b] proposing similar web user models.

2.3.10 Matrix factorization methods

The NetFlix price [Loh09] has been a corner stone evident to the magnitude of the Web Usage problem. A one million dollars price was announced for improving 10% the RMS error of the current NetFlix movie rental recommendation. The algorithm is based on matrix factorization which has been proved superior to similarity measure based techniques [KBV09].The problem lies in mapping web usage vector and product vector (books, movies). In this model a join vector (u_l, p_k) of web user l behavior vector u_l and the product k feature vector p_k and is linearly mapped to a factor space of dimensionality f.This problem stems from the family of singular value decomposition [Pat07] where the linear mapping is partially known. The algorithm solves a minimum error square problem between known factor space rating values and user's and product data [TPNT07].

2.4 Application of web usage mining

Web usage mining enables the analysis of the habits of a web user browsing in a web site. Furthermore knowing the user's interest and browsing behavior can be used to improve a web site or build new web applications. The web usage mining could be used with automatic On-line algorithm, or in an Off-line fashion, supervised techniques. The general framework for the application of the web usage mining is the adaptive web sites [VP08]. The automatic personalization of the web site to the web user's tastes, habits or marketing recommendation is only one of the techniques used on adaptive web sites for modifying or updating the content or structure.

2.4.1 Adaptive web sites

Adaptive Web Site are systems which adapt their content, structure and/or presentation of the Web Objects, to each individual user's characteristics, usage behavior and/or usage environment [KKA+08]. Adaptive sites provide users with personalized services, recommended services and content according to the user's profile acquired by the system [HX09, VP08]. The server load can be optimized since hyperlink demands can be forecast and automatic balances could be performed. The topology of the web site can be modified to the web user interest. Different aspects of managing a web site have benefited from this technology. Marketing purposes have highly improved for adaptive sites. Usability trends are solved by mean of specific user requirement.

2.4.2 Web personalization

Web Personalization improves the web site structure based on interaction with all visitors. Profiling is the principal processing that must be performed for those purposes. Using the profiling information, an automatic classification of a web user should return the profile association with an objective (e.g. product). There are two kind of personalization [MT07, VP08] based on the degree of conflict with the current semantic of the web site.

Tactical adaptation: It does not affect the overall structure of the web site as a semantic consistency is maintained and can be implemented by automatic systems. Such kinds of systems are autonomous and the whole design of the web site contemplates dynamic changes. Web sites like Amazon, NetFlix and others implement this kind of personalization.

Strategic adaptation: It must have the agreement of the owner of the web site, since the suggested changes are in conflict with the original orientation of the web site. Off-line recommendations are in general used for this kind of adaptation where owner's feedback is part of the component of the process [RV08a].

2.4.3 Recommendation

Recommendation is also based on profiling and its objective is to retrieve a product that is most likely to be selected by the current web user. Furthermore, the profiling processing for web user's usage can be described from two different points of view depending on, whether the profile is pre-established or is created on the run. The general process is called "Filtering", which is described depending on the orientation [KBV09].

Content Filtering: Categories are created according to the nature of the Web User. The information about a user is retrieved from customer databases and associations with objectives like promotions or products are performed by a trained classifier.

Collaborative Filtering: The profile is created on the run based on the history of user's interaction with the web site. The technique relates to the associations between products and relationship with users history. Some approaches relate to the product's neighborhood in which similar ratings on other products is provided by similar users. This technique is called the Neighborhood Method, where category clusters are defined by the user's past product rating. For instance a book titled "Calculus" is visited in the neighbor of Authors like "Spivak" or "Apostol" as web users visit on a web book selling site. Latent Factor Model is another approach

based on recognizing factor variables that help to measure how useful a product is for a user. Once the factor is discovered, the importance of the produce is determined for a user. Matrix factorization methods have been used for implementing Latent Factor Models such as the 2009 NetFlix winner [Loh09].

2.5 Web user simulation

Simulation has been a default tools for exploring complex problems. Nevertheless one of the common problem about simulation is to guarantee that simulated observable values will correspond to real observed values. A simulation is always based on a model of a real system and an algorithm. The procedure should **ensure** over a large number of iterations that observables will converge to solution of the proposed model. Stochastic computer simulation (henceforth just "stochastic simulation") refers to the analysis of stochastic processes through the generation of sample paths (realizations) of the processes. Stochastic simulation relates with execution of Monte Carlo related algorithm for obtaining a suited probability distribution. The web user/web site system is a complex system where time and trails over web pages are the observable results. This thesis propose to use stochastic simulation of web user as a way to anticipate change in behavior in relation to changes in the web site.

2.5.1 The random surfer

The first model of web user was applied in obtaining a ranking of importance of web page. It was a naive stochastic process [BCR06] that describe the probabilistic navigation of a web user without considering the content of web pages. The asymptotic probability to be in a page was considered as a ranking number constituting the Page Rank algorithm [BP98]. The process consider a web user following with uniform probability each link of a page or with probability d to restart the process in any other uniform page. The stationary probability can be calculated by mean of an iterative procedure, as well by simulation.

2.5.2 An extended surfer

Recently [ZLW07b] an extended version of the random surfer has been proposed. It consists of extending the navigational action than the surfer can perform: going back, staying in the page and jumping to anther page; which have a respective probability value. This stochastic model was used to build a navigability measure based on the stationary probability of never leaving the web site.

2.5.3 Aggregated model of surfer

Another approach is based on aggregated fluxes of visiting traffic [Tom98] along the network of pages. The normalized traffic by a node will correspond to the probability of a visit. In this case a maximum entropy problem was stated subject to flux conservation entities. This model predict fluxes of visit per node.

2.6 Summary

Web Usage Mining has been studied for more than ten year with successful application. Nevertheless, vast quantity of new research in the area of applied behavioral sciences and other new trends such as matrix factorization methods have revolutionized the field of web mining. NetFlix price has demonstrated that traditional web usage mining techniques have little impact on real world issues. However, recent advances in this field have reopened promising results.

Chapter 3

Data pre-processing

The Web has become the primary communication channel for many financial, trading, and academic organizations. Large corporations such as Amazon and Google would not exist without the Web and they rely on web data as an important source of customer information. For more than 10 years, there have been numerous methods proposed for extracting knowledge from this web data [PTM02, Mob06]], where pre-processing the data is a critical step toward pattern identification. This chapter explore different techniques for data pre-processing and cleaning. It is presented a novel optimal schema for extracting web user's session from web logs.

3.1 Data sources

Soft computing techniques such as Neural Networks, Fuzzy Logic, Bayesian methods, and Genetic Algorithm are commonly applied to web data to help cope with uncertainty and imprecision [Mob06]. However, these soft computing techniques do not always take care of the time dependent and high dimensionality of web data, suggesting one take considerable care to avoid the celebrated term "Garbage in, Garbage out" [Han99] . Web data can be placed into three categories: Web Structure, Web Content and Web Usage [Liu09].

Web Structure Data corresponds to the hyperlink network description of a web site. This oriented graph in today's web 2.0 [UBL+08] becomes more dynamic and depends on a web user's actions [POSPG09]. Web 2.0 applications distinguish themselves from previous software because they take full advantage of dynamic pages and encourage social interaction such as in Facebook and Twitter. This structure was traditionally static [CDK+99] but new web applications suggest the need for a time dependent graph.

Web Content Data corresponds to the text content and relates to the field of information retrieval. There is a long history of research on both text retrieval and representation spanning more than fifty years. Traditional representations, based on word frequencies of text, are the Vector Space Model or Bag of Word Model [MS99]. The focus of recent research is on more accurate representations of the semantic of text [WH06, Hen04, JS07, JM09, She09] and on coping with the more dynamic nature of text. Text now changes depending on both the time and circumstances. The semantic of a web page also changes due to multimedia objects. Parsing web pages, automatic interpretation of objects and pre-processing such information is part of ongoing research [SK09, UFTS09].

Web Usage Data corresponds to the trail of pages, also called a session, that tracks each web user while browsing a web site. Monitoring a web user's session can be a violation of privacy [CCW+07] and is forbidden by law in several countries [Lan00]. For example, installing tracing software in the web user's browser is equivalent to spyware. There are other less intrusive ways of retrieving sessions; the most popular is by using a web log, a file that records each page retrieved from a web site. This data source is anonymous, but simultaneous interactions of web users complicates a session's identification. With the advancement of browser technology, for instance the Opera software, session retrieval has become more complicated as more sophisticated algorithms for pre-loading and buffering pages have been implemented. New studies [DRV08b, DRV09c] relate to increasing the accuracy of sessionization.

Pre-processing web data has some well known issues [RHGJ06]. Parsing problems relate to different version of HTML, DHTML and non-compliant codes. Dynamic content generated via embedded code like JavaScript or server side content generation render text content extraction unfruitful through usual parsing methods. Pages with frames produce different presentations to a user that can be interpreted as a "pageview" [Mob06] that group pages and other objects together.

A pageview abstract is a specific type of user activity such as reading news, or browsing search results. Even when a session is considered as a sequence of pageviews, dynamic pages produce a large number of possible object combinations for each pageview. Log files are influenced by new browser technology (linkprefetching) [KK03], enhanced cache management and also by network caching devices such as proxies and parallel browsing consisting of tabs and pop-ups.

The quality of pre-processing has been quantified using a set of measures for each data type. New measures have also been used in the case where uncertain results are obtained [DRV08b].

An efficient storage and data representation is also important because companies like Google have large data storage that incorporates most of the Web.

3.2 The nature of the web data: general characteristics and quality issues.

Web data is not a random collection of user's interactions, pages and content. There is some regularity that remains constant across the sites and groups of web user sessions (e.g., [HPPL98]). This should impact pre-processing [DRV08b] and the evaluation of web mining.

Knowing prior web data regularities could significantly improve data mining. For example, for pattern recognition, semi-supervised clustering techniques have been popular in recent years. The results found by these algorithm are considerably better [Nad07] because they use domain data descriptions that refine the search space where an unsupervised machine learning algorithm works [GKB09]. The resulting subspace region is less of a factor with the additional domain information. In the field of data cleaning, deviations identify outliers or automatic anomaly detection [May07].

3.2.1 Web content

Web content information retrieval has been studied for many years. For example, the evaluation of a text for word frequencies. Today, Web content is based on multimedia in which rich text features enhance a web user's experience. In this context, extracting the semantics relies on the discovery of compact structures or a web object that represents a component in a web page.

Pure web texts do not differ from a traditional corpus and empirical studies show statistical regularities on the text distributions. The well known Zipf Law from 1932 [IG04, MS99] states that in a corpus the ranked number of words per page is the power $\alpha \simeq 1$ (see equation 3.1) . Today, linguistics agree that this rule assumes speakers simplify communication by using a small pool of words that can be retrieved efficiently from their memory. Furthermore, the listeners simplify communication by selecting words with a single unambiguous meaning. Other models relate to a stochastic diffusion model for generalizing Zipf's Law [KKKM05] and illustrate a new perspective on the subject.

$$P(n) = \frac{n^{-\alpha}}{\sum_{k=1}^{\infty} k^{-\alpha}} \tag{3.1}$$

One of most celebrated distribution is the Heaps' Law [LdW05] used in recent applications to describe the Internet. The heap law describes the number of unique word in a text as a power with exponent β representing its size or the number of word of a page. It has been found that $\beta = 0.52$ [LB05] has greater accuracy on Internet pages. The study then estimates the number of n-tuples of words m having the expected number of hits on search engines, obtaining log-normal results. Also these distributions are used to measure similarity between pairs of words for grouping purposes. This kind of clustering result is useful for grouping terms on text that better capture the semantic for further processing.

Thanks to the advent of Web 2.0 applications , the content of web pages has become more dynamic. Updates on web sites have been a prolific subject of study since web search engines are based on the accuracy of indexed terms and pages. The study [FMNW03] includes 151 millions Web pages browsed once a week for a three month period. The findings signified that 22% of pages were deleted, and 34.8% of pages had content updates (larger pages where more frequently updated). From other studies [OP08] the information longevity is not related to the extent of updates on the web page. It also reveals that dynamic changes can be placed into two main categories: Churn Updating behavior (33%) related to repeatedly overwriting content and Scroll Updating behavior (67%) related to lists of updates (e.g., news).

In a recent study [ATDE09] over 55,000 web pages with a diversity of visitation pattern indicated a higher dynamic content than previous studies. Consequently, further refinement of the web content representation and pre-processing must be taken into consideration in order to manage this dynamic content. Some other studies reveal that in a one hour period 26% of the visited pages during the study were modified or updated in some way and 66% of those pages were visited the following day, of which 44% consisted of query parameters. Earlier studies, (e.g., [CGM00]) report only 23% of the pages consisted of dynamic content for the period of one day.

3.2.2 Web site structure

Early studies of web site structure suggested a static graph [CDK+99] with pages as nodes and hyperlinks as edges. However, the link structure is as dynamic as content and develops exponentially [BYP06]. Despite the changing structure, large-scale statistical analyzes of the hyperlinks [BYCE07, KDNL06] show a power law distribution ($p(x) \sim x^{-\alpha}$) has a good fit for several structural measures (categorized by the exponent α). The study [BYCE07] suggests that the number of pages per web site is a power law with exponent $\alpha \in [1.2, 1.6]$ [DKM+02], the number of pages that have a given number of in-links have $\alpha \in [1.6, 2.1]$, the number of

pages with a given number of out-links has a piecewise power law with $\alpha_1 \in [0.3, 0.7]$ and $\alpha_2 \in [1.9, 3.6]$, and finally the ranking number (Page Rank) has $\alpha \in [1.8, 1.9]$. The distribution of the age of a page corresponding to the *last modified* value register is a Poisson process [CGM03] with decay parameter $\lambda \in [1.6, 2.3]$.

An important observation from current studies is that the notion of a page is losing its importance as a fundamental information source in its description of the overall web structure [BYP06]. In todays dynamic web applications, what is commonly described as a web page depends on the parameters that the application receives. Today, a complete web site could be served by a unique application file in which each page is represented by the URL's query parameters.

As stated by [DS04], the temporal component of the graph evolution has not been conclusively studied by the web mining community. They consider three levels of study: temporal "single nodes", temporal "sub-graphs" and "whole graph" analysis. The single node study examines how frequently a page is accessed during a time period in which there are no changes to a page. In this case, the data mining approach consists of clustering over page content. The subgraph level consists of finding the period of time where a small level of change has occurred within a structure based on graph properties such as order, size, components, "Max Authorities", "Hub", and "Page Rank". The "whole graph" analysis focuses on identifying a set of measures for building a features vector of the graph. Like in the subgraph level, the features could be represented by basic graph properties (e.g., order and size) or derived properties (e.g., "Max Hub" score and "Average Hub" score). A current means of research is to explain the changes of such measures in time.

3.2.3 Web user session

The celebrated Law of Surfing [HPPL98, Whi07, HW07] was noticed ten years ago as a distribution that regularly fits the number of sessions of a given size. This distribution was recognized as an inverse Gaussian (see equation 3.2) whose parameters change depending on the Web site ($E[L] = \mu, Var[L] = \mu^3/\lambda$). This law is also approximated by the Zipf's laws simplifying the calculation processing because in log scale the fitting problem reduce to linear regression.

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} e^{-\lambda(L-\mu)^2/2\mu^2 L} \tag{3.2}$$

Some empirical studies have been performed on browsing customs [TG97, CP95, WOHM06], with recent changes on web user general behavior reported in [WOHM08]. The change in browsing habits can be explained by the highly dynamics environments and new Internet services. The use of a browser's interface widget orchestrates the changes in frequencies. For

example users backtracking on certain sites are declining. Earlier studies [WJH09] report 30% back button activation. However, new studies [OWHM07] report only 14% re-visitation patterns. This behavior is most likely also the result of new applications incorporating backtracking support within a web page. The use of the forward button also reported decrease usage; decreasing from 1.5% to 0.6%. The reload button usage has also decreased from 4.3% to 1.7%.

New windows and submitting activities reported an increase in total user's actions of ($\sim$ 10%), as they are key features of dynamic pages. Despite this, the link click stream seems to maintain its percentage of actions in web browsers at 44%. As reported in [OWHM07], search engines, now considered new browsing assisting tools, carried out 16% of page visits. Another new behavior also appears as navigating simultaneously with new windows and tabs, where an average of two pages are opened throughout a session. It confirms that parallel navigation is not the exception but the rule.

The same study [OWHM07] shows a heavy tail distribution of the average stay time on a page where a considerable fraction of the visits last a very short time. Revisiting pages or scanning behavior corresponds to a fraction of the entire user's action. Despite these observations, a direct correlation is found between time staying and the number of hyperlinks on the page (also with the number of words).

[AC08, Ale09] perform other empirical work on navigation of electronic documents like pdf and MS Word. Despite the many differences between users, they report some similarities. A power law distribution models the number of times a document opens. They discover some interesting patterns by mean of a tracking system on the user interface action. The period of time spent reading a section of the text appears to relate to the interest of the user and to the quality of content. In these cases, the main methods of navigating correspond to scrolling operations, where distances traveled in relation to session size are revealed to have a power law distribution. They report the same behavior on the mean percent of time in text regions with a heavy tail distribution. A general rule for human behavior on information seeking seems to rule the interactions on Internet and document seeking behavior.

Web usage research is mainly driven by click stream data (web user's sessions). A recent study on more precise biometric data in conjunction with Web Usage Data shows that click streams are a bias source of information [JGP+07, GFL08]. Study [GFL08] uses eye monitoring techniques to focus on the click stream in a Google search result page. This reflects how web user's behave in relation to the texts and links. Eye tracking data provides important information about the web user's cognitive processing that have a direct relation to their actions. This study highlights

that click-streams are directly influenced by the order of the presented search result; the first links on a page are likely to be chosen as they tend to be inspected more. The first visual portion of the web page also has a direct influence on the click decision as a web user prefers to process this first portion. In this way, sessions obtained from web logs do not directly reflect web user behavior. The visual disposition of the elements on the web page must also be taken in consideration.

3.2.4 Privacy issues

Legal issues must be considered when collecting web usage data [IH07]. The extraction of information from web data can lead to the identification of personal data that can be illegal to collect. Any general solution to issues of what to collect must balance four forces: laws, social norms, the market, and technical mechanism.

Privacy relates to "The control of information concerning an individual," "the right to prevent commercial publicity of one's own name and image," "the right to make certain personal and intimate decisions free from government interference," and "to be free from physical invasion of one's home or person" [MB09]. Certainly, a whole spectrum of solutions exists but all of them need to be tested under current laws and social norms. While one method may protect individuals, it could compromise social information (e.g., racial discrimination) and can carry legal prosecutions.

Legislation on privacy control is still unclear [MB09]. Despite the W3C organization producing a policy called the P3P Platform for Privacy Preferences [Lin05]. This policy stipulates that information about web users must be protected by privacy laws and the users themselves informed about their rights. It suggests [MB09] a continuous process of negotiating, with relevant third parties, an optimum or acceptable level of disclosure of personal information in an on-line environment." Hence, an optimal level of privacy strives to return the privacy control to the web user, and declares that browsing preferences must determine the level of privacy.

A user's privacy can be resolved by incorporating a "slider" on a web browser that controls behavior monitoring. The P3P protocol should be the mechanism that determines and illustrates the level of privacy a user is entitled to in each web site. The use of this protocol is increasing among web sites [RBDM07]. For instance, with a common session tracking mechanism (section 3.5.2.1) the P3P protocol defines the following cookie persistent or non-persistent, first party or third party, non-personal or personal. "Persistent" cookies correspond to a permanent cookie repository that can store all interaction within the web. "First party" cookies send information

only to the web server of the page while "Third party" allows to distribute data to any other server. "Personal" cookie stores any personal web user data. According to this protocol, the most privacy compliant cookie is a non-persistent, first party and non-personal and the worst privacy compliant cookie is a persistent, third party and personal. This information will be present on the P3P protocol and future non-compliant privacy tracking session methods will be banned by most users [MB09].

3.2.5 Quality Measures

Ensuring web data quality is important for data mining. We present different quality measures for different types of web data. Quality metrics also enable the implementation of pre-processing filters.

Web structure data

The discipline of graph mining [CF06] relates to the analysis of the web structure's intrinsic graph properties. Data quality depends on the application where the data is used. For example, in web structure mining, large web crawling results need to be evaluated for the page rank value as a measure of the importance of a page [SDM+09]. The Page Rank value is an indicator of the visibility that a web page has on a search engine. Another metric corresponds to the hub and authority scores [CF06] that concentrates on out-going and in-going links. [RKJ08] discusses more indirect measures based on the web user action on the search results.

$$c_i = \begin{cases} \frac{n_i}{k_i} & k_i > 1 \\ 0 & k_i \in \{0, 1\} \end{cases} \tag{3.3}$$

There are graph theoretical measures related to community structure [CF06] that are useful for controlling data quality. The clustering coefficient (Equation 3.3 where k_i is number of neighbors of node i and n_i the number of edges between them) reflects the degree of transitivity of a graph. Other measures relate to identifying the number of connected sub-graph components. Thus a disconnected web graph is an indicator of some kind of problem in the data collection or a recent change. Another important indicator is the resilience value [CF06]. This value is obtained by finding the minimum cut-set cardinality of the graph. A cut-set is a set of links that split the graph into two components. A data set with a high resiliency value should be the most representative.

Web content data

Web pages are special cases in the context of text pre-processing, since HTML tags contain both embedded and miscellaneous information (e.g., advertisement and logos). It is estimated that an average of 13.2% of the embedded information on web pages is noise related [ZCL04]. A

similar value found between vector model text content and a predefined set of topics provides an indication of the quality of the content data [ZCL04]. Filtering embedded information knowing this measure helps with pre-processing text data.

Web usage data

The set of web users' sessions correspond to web usage data. The quality of such large data could be tested using the session size distribution, which follows a site independent distribution law [HPPL98]. For simplicity, this distribution is approximated by a power law [LBL01]. The goodness of fit value to this distribution is a quality measure of the session set [DRV08b, DRV09c, DRV09b].

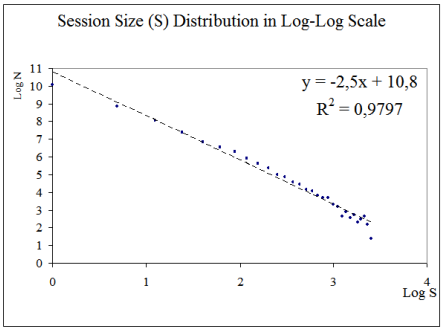

FIGURE 3.1: Session size distribution shows an approximate power law for a cookie sessionization. Data extracted April-June 2009 from http://dii.uchile.cl web site.

If real sessions are available (for example using logged users), then the "degree of overlap" between sessions can be compared [SMBN03]. It is defined as the maximal averaged fraction of real sessions that are recovered using the sessionization. Another wide spread measure [BTCF09] corresponds to the precision and recall scheme. The precision p is a measure of exactness that can be defined as $p = N_p/N_s$, where N_p is the number of maximal patterns positively recovered and N_s is the number of maximal patterns recovered from the sessionization. The recall measure is a measure of completeness that can be defined as $r = N_p/N_r$, where N_r is the number of correct maximal patterns. An accuracy value can be defined as a geometric mean $a = \sqrt{p * r}$. Better accuracy means a better sessionization method. Simulation can also be used to obtain artificial sessions [SMBN03, BTCF09]. Such simulations should be "human compliant" with statistical behavior following known strong regularities [HPPL98]. Some such simulation results are available [DRV10].

General measures

Data mining algorithms are influenced by data quality[YZZ03]. For instance, a neural network is particularly sensitive to this issue as reported in [YWL06]. One way to handle this issue is by using anomaly detection algorithms. Because an anomaly is based on parameter settings

it can be considered as a metric of data quality. Support vector machines (SVM) have been largely used for these purposes [WWZ05]. SVM can learn complex regions and find outliers. The fraction of outliers gives a measure of the quality of data. This general measuring principle can be applied to web feature vectors as defined in the section 3.5.1. Furthermore, finding the outliers implies a mechanism for data cleaning.

3.3 Transforming hyperlinks to a graph representation.

With the advent of search engines, the field of structure retrieval has been widely studied. The web structure as a graph representation has it own specific data mining process [CF06]. Large scale retrieval schema over all the Web has been implemented with success. Web page structure must be retrieved by sampling web pages following the observed hyperlink structure. This process is called Web Crawling [Cas04] and involves the storage of the hyperlink structure and web page content.

3.3.1 Hyperlink retrieval issues

A major difficulty for hyperlink retrieval is large volumes and the frequent rate of change of web data [ATDE09]. The amount of data slows the complete retrieval process. A large crawling of the web must select an update strategy such as selecting the pages most likely to change first. The crawler must have a set of policies regarding: the page selection method, the revisit schedule, a politeness policy, and parallel processing. Several strategies are available based on incomplete information: Breadth search ordering, Page rank based, prediction of changing pages [KDNL06]. Other issues relate to the imperfect mapping between URL and the page that is visually seen by users [RHGJ06]. A common HTML structure like frameset, groups a set of URLs in the same visual presented page. Hence a frameset produces confusion in the process of mapping URLs to web pages. Nevertheless, a solution considers the group of web pages as a "pageview" object [Mob06].

3.3.2 Crawler processing

The high rate of web page changing [ATDE09] implies that the time of page retrieval could be near the time of page content expiration. It is important to prioritize the most important pages for retrieval, and that requires a measure of importance for the selection algorithm. The revisiting policy should seek to maintain the freshness of the pages, and once predicting an updated page, it is inserted on the pile for crawling (Figure 3.2). The strategy for revisiting pages could be using the same frequency for all (Uniform case), proportional to its registered updating

frequency, or other estimation for the time of page obsolescence. A recent method based on the longevity of web pages [OP08] optimizes the cost of page retrieval using a generative model. Web pages are modeled as an independent set of content regions of three types: Static, Churn Content, Scroll Content. Such content types have specific lifetime probability distributions that are fitted with observed data. The page lifetime is then estimated using this distribution. The retrieval mechanism must ensure is does not overload the web server. This is called the web crawler's politeness policy, if they are not compliant, the retrieval program could be blocked on the network by third parties. The visit interval setting for the same web server should have a lower limit (usually 10 seconds) that ensures a minimal impact on the web server. This produces a slow retrieval rate because a web site can have anywhere from a thousand to a million pages. Finally a parallel algorithm reduces the processing time but should be designed to avoid retrieving the same page twice.

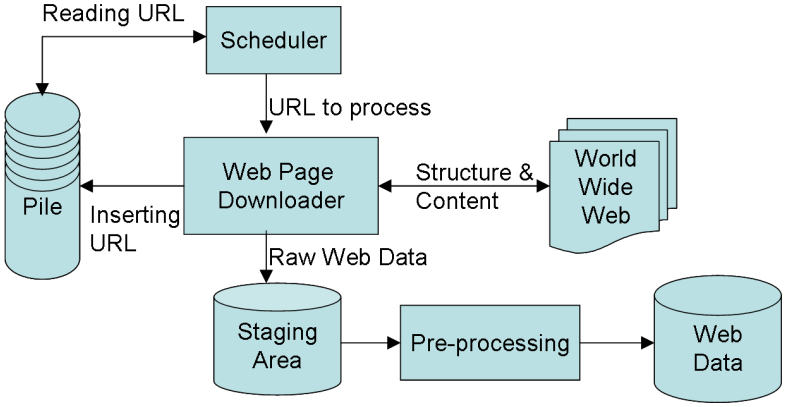

FIGURE 3.2: High level Crawler Structure.

3.3.3 Large sparse distributed storage

A crawler that inserts data directly into a database results in poor performance [Cas04]. Retrieved data is therefore stored on a intermediary format for further processing. Recent advances on storage for distributed systems [CDG+06] give some insight into efficiently connecting a web crawler with structured data storage. The Google's Bigtable storage engine is a distributed storage engine for peta byte size. It is implemented over several projects including web indexing. It consists of a three dimensional map of strings that are indexed by row, column and timestamp. A hyperlink structure fits well in this schema because a row can represent the page of a hyperlink, the column can correspond to the pointed page, and the timestamp represents when the link is retrieved. This storage facility is implemented using the Google File System and operates over a shared pool of servers. It is built on top of the Google SSTable file format that provides a persistent ordered immutable map from keys to values.

3.4 Transforming web content into a feature vector.

The information presented on web pages is complicated by text and multimedia content. This information is used for page classification [QD09], page summarization [Seb02], entity extraction [UFTS09] and semantic processing. This data needs to be filtered because many objects present on web pages are not related with the stated objectives (e.g., advertising). Furthermore text content by itself is noisy, for example sentences contain many word that have a poor influence on its semantic content. Text algorithms are also dependent on the text's language. [ZCL04] present a collection of text preprocessing algorithms for data mining purposes.

3.4.1 Cleaning web content

Web content is full of noisy items such as navigation sidebars, advertisement, copyright, and notices that can mask the principal content of a web page. Using automatic detection methods, [CKP07] claim that noise is more than 50% of the total content. An earlier method, [DMPG05], segments a web page into blocks and separates blocks into informative and non-informative. The segmentation is realized by using the DOM (W3C's Document Object Model), which decomposes the web page in a hierarchical tree structure. The objective of this method is to identify nodes from this structure that are non-informative. A node is non-informative if all its sub nodes are non-informative. The method consists of building a classifier for this property. Once all the nodes from the DOM tree representation of a web page have been labeled as non-informative then the top level non-informative nodes are removed.

At the sentence level, a cleaning consists of removing each Stop Word. A Stop Word corresponds to illatives, pronouns and others words that do not contain much semantic value by itself. This kind of preprocessing is fundamental for the "Bag of Word" model where the semantic is approximated by a set of words without considering the sentence and paragraph structure semantic. An additional step is Porter Stemming that reduces each word to it root [Por06].

In English, the algorithm for stemming is simple because it corresponds to identifying the suffix and removing it. In other languages, it can be more complicated so several language specific stemming algorithms exist [SPM09]. Another approach for stemming sentences consists of clustering words based on a corpus statistic [BP07]. The algorithm is trained on a corpus that finally defines a representative word for a set of words, the stemmed text has been shown to produce better results for automatic classification.

The changing content of a web page [ATDE09] must be managed using time window processing [JI06] where it is assumed that content remains constant during a defined period of time. This implies a content feature vector update using a temporal index.

A promising direction on text pre-processing is the Word Sense Disambiguation technique [TVN10] (WSD). It addresses the problem of selecting the most appropriate sense for a word with respect to its context. The technique consists of selecting the most appropriate meaning for each word using a semantic model of the text. It was reported that WSD boosted the further information retrieval and data mining processing [WDT09].

3.4.2 Vector representation of content

The simplest text representation consist in a vector $V = [\omega_i]$ of real components indexed by word. This representation comes from the Bag of Word abstraction. Despite the approximation, this model gives remarkable good results in information retrieval [MS99]. The value of each component of the vector ω_i is called a weight for word i. There are several weighting schema for word (Term). The most common term weighting schema is the TF-IDF schema. Table 3.1 presents the most common weightings.

The replacement of a document by a "Bag of Word" inevitably involves a loss of information. For instance, "The Matrix" and "Matrix" represent a film and a mathematical term. Ontology gives a more correct description of the semantic of objects on a web page. Once the ontology annotation on web objects is complete it is possible to transform it into a vector representation [CFV07]. But first an automatic semantic annotation should be performed [PKO+04].

3.4.3 Web object extraction

Extracting information automatically or semi-automatically from web data has become more difficult as web sites have adopted multimedia technologies to enhance both their content and their presentation. Some of the most successful sites provide video streaming and picture sharing. Unlike text, the content of multimedia formats within web pages can be understandable only to humans. At the most, some technical information such as the color of a histogram or wave patterns for pictures and sounds can be obtained automatically. Given this, a different approach to data extraction for these formats must be adopted.

The use of meta data to describe the content of any multimedia format allows the creation of automatic or semiautomatic ways of extracting information. This enables the web page to be

$f(.)$	Calculation
$binary$	$binary(\omega_i) = 1$ if the term i is present on the document k or 0 if not.
tf_k	Term frequency on the document k, $tf_k = n_{ik}/N_k$, n_{ik}/N_k is the frequency of the term i on the document k.
$logtf_k$	$logtf_k = log(1 + tf_k)$
ITF	Inverse Term Frequency, $ITF = 1 - 1/(1 + tf)$
idf	Inverse Document Frequency, $idf = log(N/n_i)$, where n_i is the number of document having the term ω_i and N the total number of document.
$tf.idf$	$tf.idf = tf * idf$ [MS99]
$logtf.idf$	$logtf.idf = log(1 + tf)idf$
$tf.idf - prob$	The probabilistic approximation of the value $tf.idf$ using an estimator for idf
$tf.chi$	Use of the χ^2 feature selection measure, $tf.chi = tf * \chi^2$
$tf.rf$	$tf.rf = tf * log(1 + n_{ik}/max\{1, n_{ik}^-\})$, where rf is the relevance frequency [LTLS05, LTSL09]
$tf.ig$	$tf.ig = tf * ig$, where ig is the information gain (Kullback Leibler divergence [EC06]).
$tf.gr$	$tf.gr = tf * gr$, where gr is the information gain ratio [Mor02].
$tf.OR$	$tf.OR = tf * OR$, where OR is the Odds Ratio [GJM01].

TABLE 3.1: Common weighting Schema. Index i is for a unique term, index k for a unique document. A document k is represented by a vector $[\omega_{ik}]_{i=1,...,N_k}$, where $\omega_{ik} = f(term\ i\ in\ document\ k)$. n_{ik} is the times the term i appears in the document k and n_{ik}^- correspond to the negative number of appearances predicted for the word i according to a trained algorithm. N_k is the number of terms in documents k.

described as a series of objects brought together in an ordered manner similar to a structured manner in which text and multimedia formats are displayed within a page.

An object displayed within a web page is termed a Web Object. One definition is found in [DV09], a web object corresponds to text and multimedia content that a user identifies as a compact unity. Using this definition every part of a web page can be describe as an object, a picture a sound or even a paragraph. An advantage here is that the content of a website is described not by the site itself, but in the metadata used to define the Web Objects within it.

Another significant advantage of using Web Objects is that any two objects can be easily compared. This can be achieved by defining a similarity measure that uses the metadata that describes the content of an object. This enables complex comparisons between objects that do not require the same format. For example if a web page contains only one picture and the accompanying text describes the picture in detail, the metadata for both the picture and the text focus on the content rather than on the format. Therefore by using a similarity measure both

objects can be discovered as equivalent.

The development of Web Object techniques is focused mainly through the user's point of view, as they are able to describe a website taking into consideration both the content and appearance of a web page rather than only the data which it contains. Different ways have been developed to describe web pages based on how the user perceives a particular page [BR09].

A large degree of research has recently been carried out in the field of Web Objects; in this section some of these are described focusing in mainly four areas: Web site Key object identification [DV09], Web Page Element Classification [BR09], Named Objects [SK09] and Entity extraction from the Web.

Web site Key objects Identification: In this work [DV09], Web Objects are defined using a specially created metadata model and a similarity measure to compare two objects. Data mining reconstructs the user sessions from the web server log and objects are created from the web-pages of a site. By inspecting the users' content preferences and similar sessions (clustering), website key objects can be found that reflect the objects within a site that captivate a user's attention.

Web Page Element Classification: This work [BR09] creates a method for detecting the interesting areas in a web page. This is accomplished by dividing a web page in visual blocks and detecting the purpose of each block based on their visual features.

Named Objects: The manner in which users' interpret a web page is the focus of this work [SK09] where a user's perception of a web page is obtained through the user's intention. Web Design Patterns are then selected based on a user's intention. These named objects are used as the basis of mining methods which allows Web Content Mining.

Entity Extraction from the Web: A Web knowledge extraction system is created in this work [DFIN08, GZDN09], which uses Concepts, Attributes and Entities as input data. By modeling this using ontology, facts from generic structures and formats are extracted. Subsequently a self supervised learning algorithm automatically estimates the precision of these structures.

3.5 Web session reconstruction.

Sessionization is the process of retrieving web user sessions from web data. Web usage mining highly depends on the correct extraction of Web User sessions [DT09]. Several methods exist and can be classified according to the level of web user personal data (privacy protection).

Proactive methods directly retrieve a rich set of information about the operation of a web user. Examples are cookies based sessionization methods where personal data and activities are stored and retrieved. Other proactive methods consist of installing a tracking application on a user's computer that enables the capture of each interaction with the browser interface. In these cases, privacy issues are raised, and in some countries it is forbidden by law. A further possibility is the use of login information in order to track the web user's actions. In this case, a disclaimer agreement is required between the company and the web user in order to enable the tracking of personal information.

Reactive sessionization corresponds to indirect ways to obtain anonymous sessions. The primary source of data for reactive sessionization is the server Web Logs containing all activities of all web users that excludes personal identifiers. Several heuristics have been used to reconstruct sessions from web logs as individuals can not be uniquely identified. Recently, integer programming has also been used to construct sessions and conduct additional analyzes on sessions [DRV08b, DRV09c].

Click stream analysis is contributing to new browser technology such as link prefetching [KK03, KT05] from the Mozilla Firefox browser [Cor] together with enhanced web page caching from the Opera browser [ASA]. Link prefetching corresponds to loading links in the background that can be visited in the future allowing a faster browsing experience. In this case, the access register in the log corresponds to a machine operation and does not reflect human behavior. This increases the difficulty of constructing sessions from web logs.

Sessionization based on web logs passes several processing phases [DT09]. Firstly, data acquisition is performed during a web server's operation, where for each HTTP request a register is recorded in the web log. During data collection, files are selected and scanned and information stored in temporal repositories for further transformation. Data cleaning consists of selecting only the valid registers: this includes discarding robots, exploits and worm attacks; and finally only html files are selected (discarding multimedia, images, etc). The chunking process [DRV08b] splits the large set of valid log registers into smaller sections based on the same IP, agent combination and a time threshold between consecutive registers. The last partition does

not ensure unique sessions since multiple web users can share the same IP and agent, but it simplifies further processing. A chunk is the main data unit through which a sessionization algorithm retrieves the web user's trails.

3.5.1 Representation of the trails

Sessions have different representations [DGEU07] depending on what further data mining is to be conducted. At least three historical representation of a session have been used.

Weight usage per pages:
Following earlier work on web user profiling [MDLN02], sessions are processed to obtain a weight vector $V = [w_1, ..., w_n]$ of normalized time of visit for each page. The vector's dimension is n corresponding to the number of pages in the web site, and the index relates with the corresponding page. Vector entries are zero when the page is not visited and proportional to the time spent on the page when the user had visited it. This representation is useful for web user's profiling based on clustering. The weight can be tested with many other representative measures. A binary weighting scheme could be used ($w_i \in \{0, 1\}$) that is suitable for association rule analysis [MDLN01]. In this case, the weight takes value one if the page is visited during the session. Another variation makes use of prior knowledge of each page [MDLN02].

Graph of sessions:
[GO03] considers the time spent on each page for similarity calculations. The representation of a session is given by the sequences of visited pages. A weighted graph is constructed using a similarity measure and an algorithm of sequence alignment [GO03]. This representation is further processed using a graph clustering algorithm for web user classification.

Considering text weighting:
Other works [VYA+03, VP08] propose to take into account the semantic of text and the time used. An important page vector is defined containing the page (text content) and the percentage of time spent on each page sorted by time usage and selecting only a fixed number of pages with maximum time usage. This representation is generalized for multimedia content using web objects. Clustering of important page vectors are perform via the SOFM (Self Organized Feature Map) [Koh88] algorithm obtaining categorizations for web user.

3.5.2 Proactive sessionization

Proactive sessionization directly record a web user's trails but there are some implementation subtleties. We detail a variety of proactive session retrieval methods. indexProactive Sessionization

3.5.2.1 Cookie based sessionization method

Cookies are a data repository in the web browser that can be read and updated by an embedded program on the web page. This mechanism, usually used for persistence purposes, can uniquely identify the web user by storing a unique identifier and sending this information to the web server (Figure 3.3).

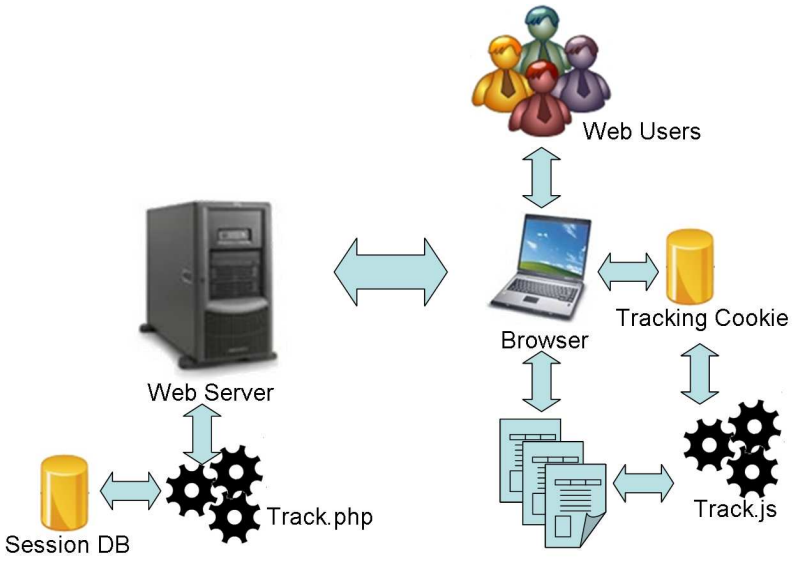

FIGURE 3.3: Cookie Sessionization: An embedded script (Track.js) updates the cookie with a unique identification per each user and sends it along with the current URL to the web server web application (Track.php) for recording the page on the session's database.

The method of session processing with cookies is simple in appearance but has it own issues. Security and privacy issues (already mentioned) cause web users to often disable the cookie facilities of their web browsers. Two DOM methods are available for cookie processing, the "onload" method that can be replaced by direct execution of code on HTML and the "onbeforeunload" that is more restricted for browsers executing at the moment of leaving the page. Entering and leaving a web page events could be recorded on the web server for calculating the sequence of pages and time of visit per page. Different browser execution policies gives different pattern access to the same session. Heuristic methods should be used for obtaining sessions for some undetermined cases (Table 7.2). For example, if it was registered as an enter event on a page and then a leaving event from another page, the time of visit for both page

could be approximated as half of the whole period. A more precise heuristic could estimate the visit time based on similar known sessions. Another refining could be the loading time of the web page. There are 11 combinations of event considering the previous page and the next page; impression occurs in some cases when the page corresponds to the first or last page.

Despite the problem of security and privacy of cookies some advances have been made in order to secure it usage [YXW07]. It consists of automatic validation of cookies for safety of the web user, based on automatic classification algorithms. Others protocols like P3P [Lin05, RBDM07] relate to the web site declaration of a cookie's degree of private data retrieval.

3.5.2.2 Tracking application

A spyware application monitors and stores events on the host machine. They can have links to criminal activity because they can be used to retrieve personal credit card numbers and passwords. These extreme tracking applications are a world wide Internet security problem, but some applications have been designed for scientific purposes, for example, AppMonitor [Ale09] that tracks low level events such as mouse clicks on Windows OS for Word and Adobe Reader.

3.5.2.3 Logged users

Web applications with login authentication can maintain a register of a user's trails on a proprietary database. This can be implemented by storing each page visit during the web user's visit. This is the simplest and most reliable way to track sessions but requires client permission and authentication.

3.5.3 Reactive sessionization

Sessions obtained from log files are anonymous because web user identity does not appears explicitly in the registers. Of course, this complicates identification of a web user's trail (Reactive Sessionization). Web users with a similar profile could be accessing the same part of a site at the same time resulting in registers that appears shuffled on Logs. Additionally, web page caching (e.g., Back Button) from browsers and proxy servers can contribute with missing Log registers. Untangling sessions from Log files requires some assumptions like the maximum time spent by web users on a session, the web site's topology compliance and semantic content of web pages.

3.5.3.1 Traditional heuristic sessionization

Web Logs maintain four important field for each register: The web user's IP address, the date and time of the request, the url requested before of the current (when activated), and the browser agent identification (Agent) [Zaw02]. Using those data a number of processing methods for session retrieval have been historically used. Traditional heuristics start from a partition of the Log register set by grouping according to the same IP number and Agent description. An IP number and agent can not uniquely describe a web user since Internet Service Provider (ISP) multiplex the same IP number over several users using the network address protocol (NAT).

Time oriented heuristic:
One of the most popular methods for session retrieval is based on a maximum time a user stays at a web site. After partitioning by IP and Agent, the register is further sliced considering the accumulated time of the visit. There is a tradition to use 30 minutes for the time window [SMBN03]. This time is based on empirical experience of the web data mining community. Nowadays, this number doesn't have an explanation, but it does seem to provide reasonable results [VP08]. Another less used approach is to limit the time spent per page.

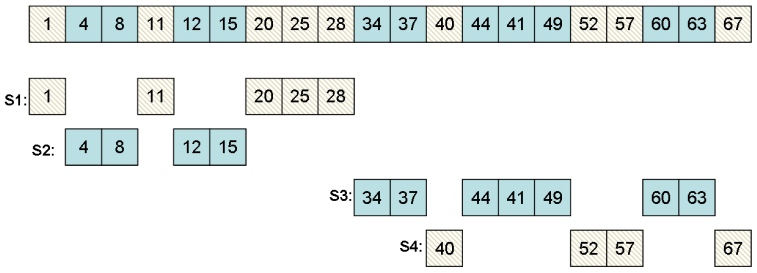

FIGURE 3.4: Time Based Sessionization example: a log register indexed by time (second) is segmented in two groups (IP/Agent) generating four session. A timeout occurs after registers 15, 28 and 63.

Web site topology oriented heuristics
The site topology motivates another heuristic where web users strictly follow the hyperlink structure of a web site. If a register can not be reached from the last register in a web log, it starts a new session [CMS99, SMBN03]. The heuristic scans registers with the same IP and Agent, starting a new session each time a register can not be followed by the previous (Figure 3.5). Of course, this does not uniquely identify the individual path and such a heuristic has difficulty when two or more users follow the same path at more or less at the same time [DRV08b]. This is the case for web site that has frequently accessed content with only a few ways of accessing it, for example, the financial news from a news web site. When a browser or proxy cache is activated, "path competition" can be used for reconstructing the missing registers and conforming a session [SMBN03] by selecting the shortest path for missing registers. If the

referrer field of the log file is activated, the path competition heuristic can be enhanced because the previous pages are given [LFM08].

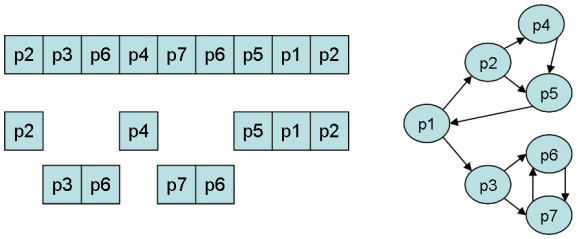

FIGURE 3.5: Topology Based Sessionization example: Given the sequence of pages in the log register (left hand) and the web page's structure (right hand) if a page (p3) does not follows the web site hyperlink structure then it start a new session (p4,p7,p5).

3.5.3.2 Ontology based sessionization

This method consists of enriching web registers with the semantic information of the visited URL. It is based on the assumption that each web user has a defined purpose that is reflected in the sequence of pages that it visits. The ontology description of each URL must be stated before [JJ04, Jun04] and a semantic distance matrix Δ_{ij} is calculated from the URL i to j. Sessions are constructed by including the URL that has the nearest semantic distance. Another approach consists in defining predefined sub-trails [KC06] categorized by ontology.

3.5.4 Sessions in dynamic environments

Static assumptions for web usage mining are a common hypothesis for data pre-processing. However in web application like CRM and recommender systems dynamic content is not the exception but the rule. In this dynamic context, the notion of dynamic URL becomes a valid representation [NSS+08]. Query parameter values from the URL and application database contribute to map the content with the dynamic URL into a hierarchical semantically structure $(i, j)|i : parent,\ j : child$ of semantic label.

The evolution of a web site produces changes in the behavior of the web users. Thus, web user sessions should be comparable for a period of time where changes can be considered minimal. Those periods $\{T_1, ..., T_k\}$ depend on web site managers' updating procedures and they need to be defined. Then, a given period of sessionization has to be performed using the available semantic labels in the page. Periods of sessionization can then be compared to the analysis of the user profile evolution [NSS+08]. The use of semantic labeling provides a comparison between the sessions belonging to different version of the same web site.

3.5.5 Identifying session outliers

An important phase of data pre-processing is data cleaning. A first stage of the sessionization process cleans the Log file erasing register from robot, viruses/worms, hacking attempt and others. But in general not all of those unwanted data can be removed. The sessions set could be refined detecting different modes of unusual behavior [SL08, JJ04]. Some recent studies [SL08] use the 1% tail of the Malanobis distance to locate rare behavior sessions. Others studies relate to semantic characteristics of the sessions [JJ04] which are obtained using an Ontology-Oriented heuristic.

3.6 Integer programming sessionization

We present two optimization models for sessionization. Our optimization models group log registers from the same IP address and agent as well as ensuring the link structure of the site is followed in any constructed session. Unlike the heuristics that construct the sessions one at a time, each of our optimization models constructs all sessions simultaneously.

Each constructed session from a web server log is an ordered list of log registers where each register can only be used once in only one session. In the same session, register $r1$ can be an immediate predecessor of $r2$ if: the two registers share the same IP address and agent; a link exists from the page requested by $r1$ to the page requested by $r2$; and the request time for register $r2$ is within an allowable time window since the request time for register $r1$.

3.6.1 Bipartite cardinality matching

The first optimization model we present for sessionization is bipartite cardinality matching (*BCM*) [e.g., AMO93]. The BCM problem seeks a matching of maximum cardinality in a bipartite undirected network. We construct the bipartite undirected network from our web server log so that the matching of maximum cardinality is equivalent to minimizing the number of sessions. There are several specialized algorithms for solving BCM with $O(\sqrt{n}m)$ where n is the number of nodes and m is the number of arcs [e.g., AMO93]. In our network, each register is represented by two nodes; one on the *from* side (representing an immediate predecessor) and one on the *to* side (representing an immediate successor). Figure 3.6 shows a six register example.

On each side, we order the nodes (registers) in order of increasing time as recorded in the web server log. An arc exists from a node on the *from* side, $r1$, to a node on the *to* side, $r2$, if the register corresponding to $r1$ could be an immediate predecessor of $r2$. For the Figure 3.6

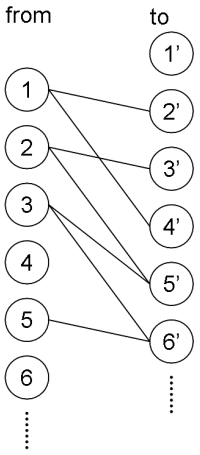

FIGURE 3.6: Bipartite maximum cardinality matching. Each register is represented by two nodes. An arc exists if the register on the from side can be an immediate predecessor of the node on the to side.

example, we assume seven arcs exist.

Given a solution, we construct the sessions from the matching. A node on the *from* side that is not matched is the last node in a session. A node on the *to* side that is not matched is the first node in a session. A session follows the connected segments in the matching where (to aid in visualizing the sessions) we add directed arcs (from the *to* side to the *from* side) connecting nodes representing the same register. Figure 3.7 provides a solution to the example of Figure 1. Nodes 4 and 6 end sessions, nodes $1'$ and $4'$ start sessions, that results in two sessions (1-2-3-5-6 and 4).

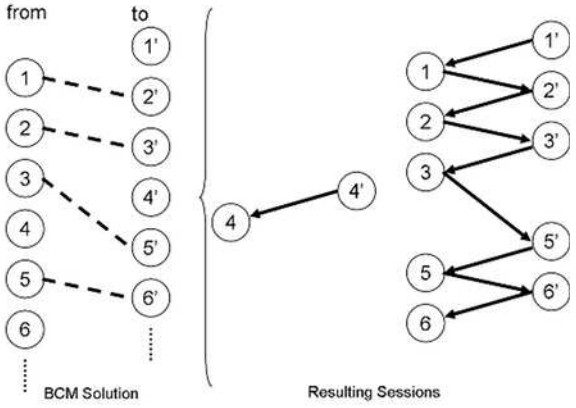

FIGURE 3.7: The maximum cardinality matching has four arcs. We construct two sessions from the matching: 1-2-3-5-6 and a session with only register 4.

3.6.2 An integer program for sessionization.

An optimal solution to our constructed BCM for sessionization provides the lower limit on the number of sessions for a given log file. Such an optimal solution has intuitive appeal because fewer sessions should result in fewer, less interesting, sessions that consist of only one or two registers. The upper limit on the number of sessions for a given log file is the number of registers; each register is its own session. To construct other solutions for sessionization, we formulate an integer program (slightly generalized from [DRV08b]) that allows us to reward characteristics of each session.

Our sessionization integer program (*SIP*) uses a binary variable X_{ros} that has value one if a log register r is assigned as the *oth* position during session s and zero otherwise. Each index r identifies a unique register, each index s identifies a unique user session, and the index o is the ordered position of a register during a session. We present the SIP formulation below in NPS standard format [BD07].

3.6.2.1 Indices

o	Order of a log register visit during a session (*e.g.*, $o = 1,2,\cdots,20$). The cardinality defines the maximum size of a session.
p, p'	Web page.
r, r'	Web server log register.
s	Web user session.

3.6.2.2 Index Sets

$r' \in bpage_r$	The set of registers that can be the register immediately before register r in the same session. Based on: - pages that have a link to the page of register r - IP address matching of register r and register r' - agent matching of register r and register r' - time of register r and register r'. Of course, r can not occur before r' but we assume a user defined minimum and maximum time between two consecutive registers in the same session.
$r \in first$	set of registers that must be first in a session.

3.6.2.3 Data [units]

Used to produce the index sets above:

$time_r$	the time of register r [seconds].
ip_r	the IP address for register r.
$agent_r$	the agent for register r.
$page_r$	the page for register r.
$\underline{mtp}, \overline{mtp}$	the minimum, maximum time between pages in a session [seconds].
$adjacent_{p,p'}$	one if a page p has a link to page p'.

Used in formulation:

C_{ro}	the objective function coefficient for register r assigned to the oth position in a session.

3.6.2.4 Binary Variables

X_{ros}	1 if log register r is assigned as the oth request during session s and zero otherwise.

3.6.2.5 Formulation

Maximize $\sum_{ros} C_{ro} X_{ros}$

Subject to:

$$\sum_{os} X_{ros} = 1 \qquad \forall r \qquad (1)$$

$$\sum_{r} X_{ros} \leq 1 \qquad \forall o, s \qquad (2)$$

$$X_{r,o+1,s} \leq \sum_{r' \in bpage_r} X_{r',o,s} \qquad \forall r, o, s \qquad (3)$$

$$X_{ros} \in \{0,1\} \forall r, o, s,$$

$$X_{ros} = 0, \forall r \in first, o > 1, s$$

The objective function expresses a total reward for sessions where a reward of $\sum_{r,o' \leq o} C_{r,o'}$ is obtained for a session of size o. As an example, setting $C_{ro} = 1 \ \ \forall r, o = 3$ and $C_{ro} = 0 \ \ \forall r, o \neq 3$ provides an objective function for maximizing the number of sessions of size three. Section 3.9 reports on how we varied the values of C_{ro} and the results obtained. Constraint set (1) ensures each register is used once. Constraint set (2) restricts each session to have at most one register assigned for each ordered position. Constraint set (3) ensures the proper ordering of registers

in the same session. $X_{ros} \in \{0,1\} \forall r, o, s$ defines variables as binary. To improve solution time, we can fix (or eliminate) a subset of these binary variables to zero ($X_{ros} = 0, \forall r \in first, o > 1, s$). After forming the set $bpage_r$, the set $first$ is easily found ($r \in first$ if $bpage_r = \emptyset$).

3.7 Variations

We develop variations of our optimization models to further explore the likelihood of specific sessions and characteristics of sessions. Specifically, we find the maximum number of copies of a given session, the maximum number of sessions of a given size, and maximum number of sessions with a given web page requested in a given position for each session.

3.7.1 Finding the maximum number of copies of a given session

To find the maximum number of copies of a given session possible from the registers of the web server log, we have two cases.

When each page in the session is visited only once, this can be modeled as a maximum flow problem [e.g., AMO93]. The maximum flow problem seeks a solution that sends the maximum flow through a network from a source node to a sink node. We construct the network with a node for each register that corresponds to a page of the session, a source node, and a sink node. An arc exists: from the source node to each node that corresponds to the first page in the session; between any two nodes where one can be an immediate predecessor of the other for the session; from each node that corresponds to the last page in the session to the sink; and from the sink to the source. The arc from the sink to the source has unlimited capacity; all other arcs have an upper capacity of one.

We have a network with side constraints when one or more pages in the session repeats. The network is much like the previous case; we have a source node and a sink node, but we must now also keep track of the node's order in the session. For each position in the session, we have a node for each register that corresponds to the page occurring in that order for the session. Therefore, we replicate a node corresponding to the same register the number of times its page repeats in the session. An arc exists: from the source node to each node that corresponds to the first page for the first position in the session; between any two nodes where one can be an immediate predecessor of the other for the session of interest; from each node that corresponds to the last page in the last position for the session to the sink; and from the sink to the source. Again, the arc from the sink to the source has unlimited capacity; all other arcs have an upper

capacity of one. For each page that repeats in the session, there is a constraint to restrict the total flow out of all nodes corresponding to the same register to be less-than-or-equal-to one.

3.7.2 Maximizing the number of sessions of a given size

Using our SIP of section 3.6.2, we can find the maximum number of sessions of a given size l by setting the objective function coefficients to $C_{ro} = 0\ \forall r, o \neq l$, and $C_{ro} = 1\ \forall r, o = l$. With such coefficients, only a session of size $\geq l$ has value one. An optimal solution could include sessions greater than size l but any such session can be split into two sessions; one consisting of the first l registers and the other consisting of all the rest of the registers.

3.7.3 Maximizing the number of sessions with a given web page in a given order

Using our SIP, we can also find the maximum number of sessions possible with a specific page p in a specific order o in a session. The cost coefficients $C_{ro} = 1$ when the register r corresponds to the specific page p and o to the given order, we set $C_{ro} = 0$ otherwise.

3.8 Web server log test data

We consider a university web site (http://www.dii.uchile.cl) that hosts the main page of the Industrial Engineering Department of the University of Chile, sub-sites of research groups, personal homepages, a web mail site, academic programs and related project sub-sites. As a general purpose site, it has a lot of diversity and reasonably high traffic, although much of this traffic comes from web mail which we do not consider.

3.8.1 The shape of web server log data

3,756,006 raw registers were collected over a time window of one month, April 2008. We want to find sessions consisting of just web pages so we filtered out multimedia objects (*e.g.,* pdf, jpg, and avi), faulty requests (HTTP errors), web spider requests, web mail requests, hacking attempts (very quick and continuous request from the same IP on some login pages), and monitoring tasks. The final total for our study is 102,303 clean registers of static html pages as well some dynamic pages (php and jsp), with a total of 172 different pages. Of these, 9,044 registers correspond to visits to the root page of the site.

We find that only a few IP addresses account for the vast majority of all clean registers. Over 98 percent (16,785 out of a total of 16,985 unique IP addresses) have less than 50 register for the

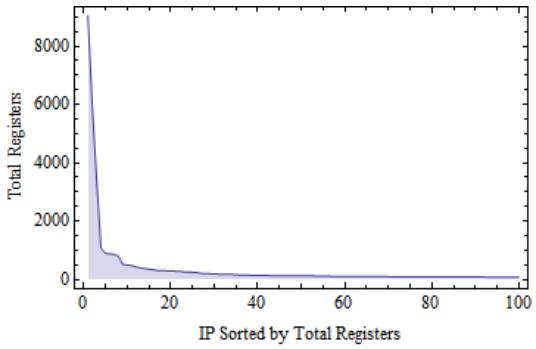

FIGURE 3.8: The number of registers for the 100 IP addresses that account for the most registers.

entire month. Figure 3.8 displays the number of registers for the 100 IP addresses that account for the most registers.

We also found how many unique web pages are visited by each IP address; we find that IP addresses that visit many unique web pages tend to have more diverse sessions. Figure 3.9 shows the number of unique pages requested by the 2,000 IP addresses that account for the greatest number of unique page requests. Of the IP addresses not shown, almost 84 percent (14,265 out of 16,985) visit three or less different pages for the entire month.

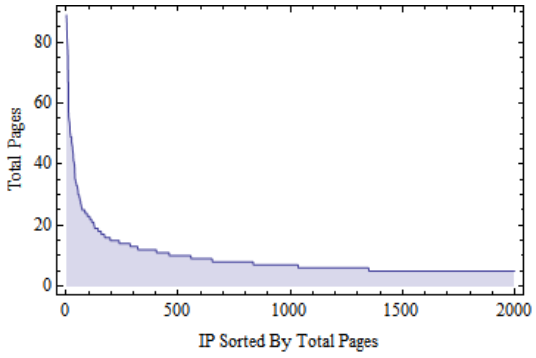

FIGURE 3.9: The 2,000 IP addresses that account for the greatest number of unique page requests.

We recovered the hyperlink structure of the web site using a web crawler [MB98]. We obtain 172 pages with 1,228 links between them, for pages identified in the previous cleaned log registers. We store the information in a relational database (MySQL) that includes tables for unique IP addresses, unique page identifiers, and unique links between pages and the registers. The database maintains relational constraints between tables in order to ensure data consistency.

3.8.2 Data selection

We select registers for our sessionization study with filtering by IP address. For each IP address, we find the number of registers and a measure of the diversity of the pages visited over the registers. The measure of diversity we use is entropy, $S = \sum_p f_p Log_N(1/f_p)$, where f_p is the frequency of page p occurrence over all register entries for the same IP address and N is the number of unique pages visited by the same IP address. S takes values from zero to one. When the value of S is near zero, most register entries are for the same page, if the value is near one all the pages are visited with similar frequency. Figure 3.10 plots, for each IP address, the number of registers versus S. There are many IP addresses with diversity near zero (visiting one page most of the time) and many IP addresses with high diversity but a low number of registers for the entire month. We concentrated on the IP addresses with high diversity and a high number of registers reckoning that these are the most interesting (and most difficult to solve) for sessionization.

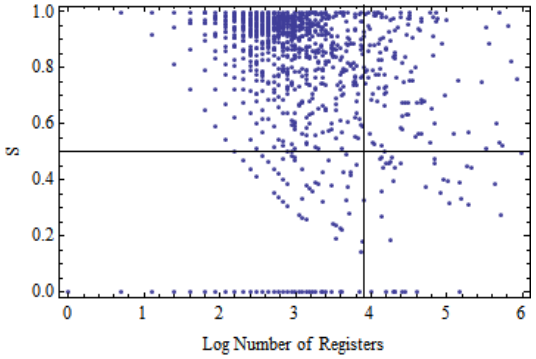

FIGURE 3.10: Log number of registers vs. S for each IP.

Selecting the IP addresses in the upper right rectangle of Figure 3.10 (more than 50 registers and S greater than 0.5) results in 130 IP addresses with 17,709 registers (17.3% of the total number of registers).

3.9 Results

Without knowledge of the actual sessions, that give rise to a specific web server log, it is not possible to know exactly how realistic any proposed session are. One measure that is available is how well the distribution of session sizes matches the empirically observed power law distribution [HPPL98, VOD+06]. We use linear regression on the logarithm of the size and the logarithm of the number of sessions. We report the regression correlation coefficient and standard error as measures of sessionization quality. The closer the correlation coefficient is to one and the standard error is to zero, the better the sessionization result. We also use the

variations of our optimization models (sections 4.1-4.3) to further explore the sessions.

It is easy to construct instances of our SIP that can not be solved. For example, a web server log of 100,000 registers, such as the one we consider in our computational study, allowing a maximum of 5,000 sessions, and a maximum session size of 20 produces 10^{10} binary variables and even more constraints. Fortunately, preprocessing allows us to partition an instance from a web log into *chunks,* where a chunck is a subset of register such that no register in one chunk could ever be part of a session in another. This is easily accomplished by partitioning chunks so that each one corresponds to a unique IP and agent combination. Even after this partitioning, a chunk may contain too many registers to be solved easily. In such cases, a chunk may be further partitioned whenever the time difference between two consecutive registers (where the registers are sorted by time) exceeds $\overline{mtp}$, the maximum time permitted between registers in the same session

Heuristic rules could be employed to continue to reduce the size of a chunk but such rules were not needed in our computational study. In fact, we avoided making a chunk too small because there is a fixed time associated with generating and solving each chunk. For our computational work, we used a minimum chunk size of 50 and $\overline{mtp} = 300$. This results in 403 different chunks.

All computation is done using a 1.6Ghz PC with 2 Gbs of RAM. We generate BCM and SIP instances using GAMS [GAM08] and solve them using CPLEX version 10.1.0 with default settings [Gal00], controlled by a php script and MySQL 5.0.27 [e.g., Aul04] as a data storage engine.

3.9.1 BCM results

We solve the 403 instances of BCM using the CPLEX linear programming solver. The largest instance consists of about 1,500 binary variables and 200 constraints. Total solution time for all 403 instances is less than five minutes. Using a specialized BCM algorithm would reduce this time. The solution results in 12,366 sessions. The longest session has 41 registers (size 41) and the second longest session is only 14 registers. Considering all sessions, we get $R^2 = 0.88$, a standard error of $std = 1.23$ and a exponent of -2.93. The regression fit predicts less than one session for a size of 10 or more registers. Considering just sessions of up to 14 registers (eliminating the session of size 41), we get a correlation coefficient of $R^2 = 0.98$, a standard error of $std = 0.39$ and an exponent of -3.85 (Figure 3.11).

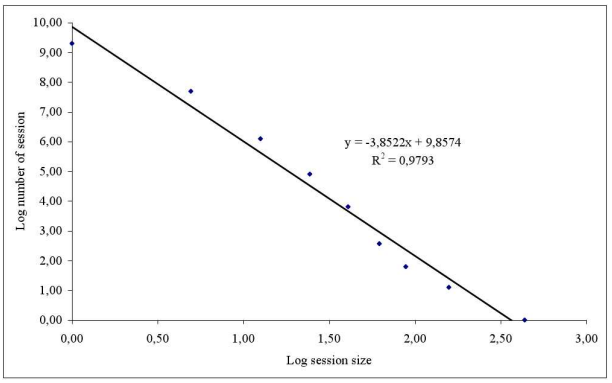

FIGURE 3.11: Number of session vs. size in logarithm scale for BCM (section 3.6.1).

3.9.2 SIP results

The 403 different instances of our SIP (with $\underline{mtp} = 0$, $\overline{mtp} = 300$, and a maximum session size of 20) range in size from about 4,000 to 376,000 binary variables and 7,800 to over 340,000 constraints. For each instance of SIP, we set the maximum time limit to 300 seconds and the relative gap to one percent.

We performed many experiments with the objective function coefficients C_{ro} and found most sets of values that reward sessions of longer size produced good quality sessions. Table 3.2 presents five sets of objective function coefficients along with the resulting correlation coefficient and standard error. For instance, we solve SIP for all 403 chunks using the fifth set of coefficients, $C_{ro} = o^2 \ \forall r, o$, and found the resulting correlation coefficient of 0.92 and standard error of 0.72. Figure 3.12 shows how many of each session size are found by our integer program for sizes two and higher, the power law distribution fit, and the resulting correlation coefficient for the third set of objective function coefficients.

	$C_{ro} =$	R^2	StdErr	Total Sessions
1	$1/\sqrt{o}$	0.88	1.10	12,502
2	$Log(o)$	0.93	0.66	12,403
3	$3/2Log(o) + (o-3)^2/12o$	0.94	0.59	12,403
4	o	0.93	0.63	12,409
5	o^2	0.92	0.72	12,410

TABLE 3.2: Five sets of different objective function coefficient values and the resulting correlation coefficient, standard error and the total number of sessions.

The computation time was similar for all sets of coefficients. We report details for the third set. Over 95% (382 out of 403) of the chunks obtained a solution within one percent of optimal. The average relative gap for these 382 chunks was 0.1%. The average generation and solution time for these chunks was only 4 seconds. The 300 second limit was reached in only 21 out of the 403 chunks. Figure 3.13 shows solution time by the number of binary variables. Above 200,000

discrete variables we see the instances that reached the 300s time limit.

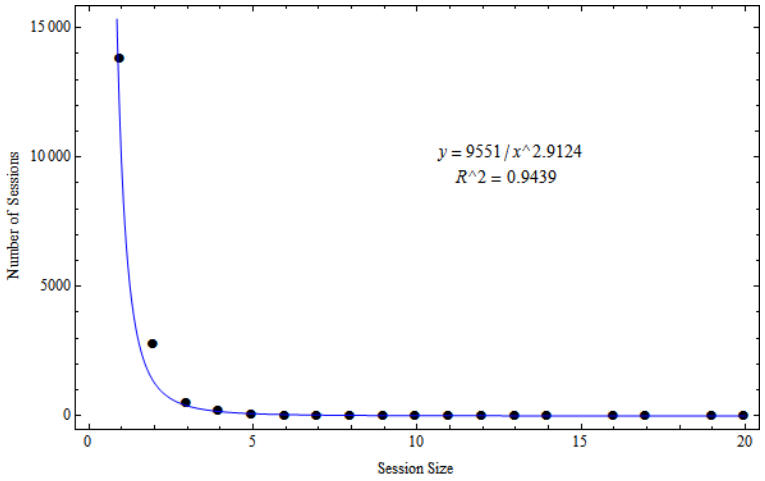

FIGURE 3.12: Session size found and the power law distribution fit.

For the 21 chunks reaching the limit, the average relative gap was 30%. We increased the solution time for several of these 21 chunks and found in most cases the additional time did not change the resulting sessions but did improve the bound on the theoretically best solution and therefore improved the relative gap. A further division of these 21 chunks into smaller chunks is another possible approach we tried on a few of these chunks to improve the gap but this too did not change the sessions.

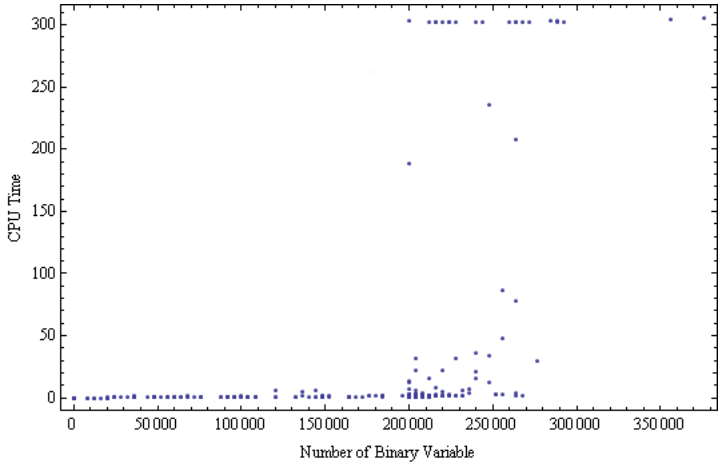

FIGURE 3.13: Solution time in seconds vs. number of binary variables.

For a variety of chunks, we also doubled the value of $\overline{mtp}$ to 600 seconds and the resulting sessions were almost identical to those found with $\overline{mtp}$ = 300. We also increase the maximum permitted session size from 20 to 40 and this results in only 10 sessions longer than 20.

3.9.3 Comparison with a time-oriented heuristic

We compare our results with a traditional sessionization timeout heuristic on all clean registers. The timeout heuristic is substantially faster (only 13 seconds) but results in a distribution of sessions with only a $R^2 = 0.92$ correlation coefficient (not as good as the $R^2 = 0.98$ found by BCM or most of the solutions found by SIP) and a standard error of 0.64 (nearly twice the standard error of 0.38 found by BCM or most of the solutions found by SIP). Figure 3.14 provides a comparison of the standard error for both the timeout heuristic and BCM.

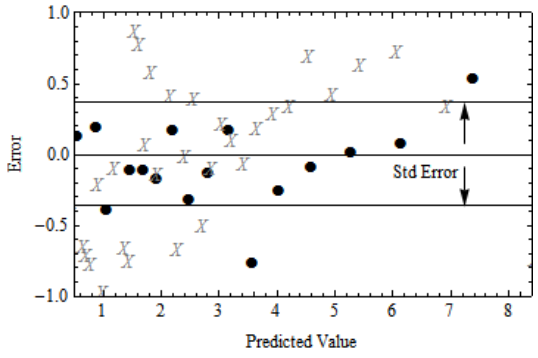

FIGURE 3.14: Predicted value error for the timeout heuristic shown as (X) and BCM shown as ($\bullet$). Also shown is the standard error (0.39) band for the integer program.

3.9.4 The maximum number of copies of a given session

BCM and SIP provide a similar number of total sessions, 12,366 for BCM compared with 12,403 to 12,502 for SIP. When we compare the sessions produced, we find some differences. We use sessions of size four as an example in this section. The first two graphs in Figure 3.15 show the most frequent sessions of size four (sorted from highest to lowest) for BCM and SIP (with $C_{ro} = 3/2Log(o) + (o-3)^2/12o$). Table 3.3 provides BCM and SIP results for 10 sessions of size four. We see both BCM and SIP find session 1 most frequently (41 times), but then differences occur. Session 18 (not shown) is the next most frequent session for SIP (with 6 sessions) but it is only the seventh most frequent session for BCM (with 3 sessions). Session 6 doesn't occur at all in SIP and only one time in BCM.

By maximizing the number of copies of a given session (section 3.7.1), we find the maximum number of times the session could occur (Figure 3.15 and Table 3.3). This provides us some additional guidance on how likely the session is to occur. In Table 3.3, we provide these values in the "max" column. For sessions of size 4, we see there are only a few sessions that can possible occur very frequently. Specifically, only seven sessions could occur 20 or more times for the entire month. Only 17 sessions could occur more than 10 times for the entire month. It takes an average of about one second to solve for a given session and chuck. 403 chucks and 80 sessions takes a total of approximately 9 hours. All instances are solved to optimality.

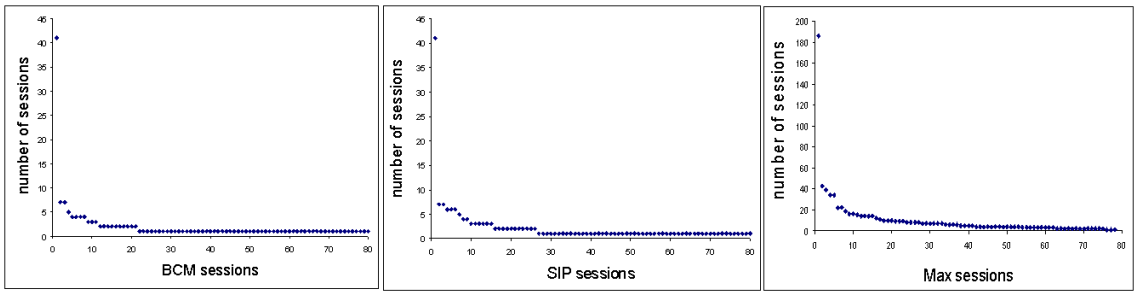

FIGURE 3.15: Number of session of size 4 resulting from BCM, SIP and maximizing the number of a given session.

Session	*BCM*	*SIP*	*max*
1	41	41	186
2	5	3	43
3	4	5	39
4	7	4	34
5	4	4	34
6	1	0	22
7	1	0	22
8	7	0	19
9	1	0	16
10	1	0	16

TABLE 3.3: The top 10 most likely sessions based on finding the maximum number of a specific session possible compared with the number of sessions found by BCM and SIP.

3.9.5 Maximum number of sessions of a given size

We find the maximum number of sessions of a given size for use in other research [RV08c]. Specifically, for a specific *size* session, $C_{ro} = 0 \; \forall r$, $o \neq size$, and $C_{ro} = 1 \; \forall r$, $o = size$. Results for size two to ten are shown in Table 3.4. We see that relatively few sessions of size six (or higher) are possible. Note that the maximum number of sessions of size one is the number of registers.

Size	Num. Sessions
2	3,000
3	1,509
4	755
5	435
6	257
7	181
8	135
9	98
10	82

TABLE 3.4: The maximum number of sessions possible of a given size.

3.9.6 Maximum number of sessions with a page in a fixed position

What are the most likely second or third pages in a session? Web site designers often ask such questions. Table 3.5 and Figure 3.16 shows results obtained by the SIP variation in section 3.7.3 along with BCM and SIP (with $C_{ro} = 3/2Log(o) + (o-3)^2/12o$) results for the third position. There are only 4 pages that could have been the third page visited in 100 or more sessions; only 19 pages that could have been the third page visited in 30 or more session. The BCM and SIP session have only some minor differences when compared to each other and "max" for this session characteristic. Solution time is about 29 hours for the 403 chunks and 58 pages; an average solution time of under 5 seconds per instance. No instance reached the 300 second time limit.

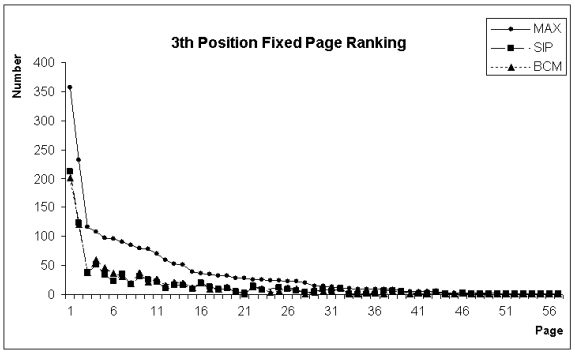

FIGURE 3.16: Maximum number of sessions with a page in the 3rd position, compared with SIP and BCM.

Page	*BCM*	*SIP*	*max*
1	201	212	356
2	121	124	232
7	38	38	116
13	60	51	108
5	46	34	97
8	37	23	96
4	31	35	91
9	19	18	86
6	38	31	80
10	22	26	78

TABLE 3.5: The top 10 pages that could be in the 3rd position comparing with BCM and SIP results.

Considering the effect of cache devices on web logs

As presented in section 3.2 the use of the back button in web browsing produces missing registers in web logs. The maximal reward per session can be modified by introducing a new variable Y_{ros} corresponding to the use of a back button action [DRV09c]. A session in this case could be

identified to include a session with back button usage and constraints can be added on the total number (or percent) of sessions that include the use of the back button.

3.10 Performance measures

The problem of sessionization can be visualized as a machine learning classification problem. Register of page retrieval are classified where labels correspond to the session identifiers. Furthermore, sessionization is an unsupervised machine learning problem. The labeling process falls in the category of multiple-classification. In this case, performance measures like precision and recall have been adapted [TK07] for multiple labels. Precision and recall are classical performance measures from information retrieval for binary classification [BYRN99].

We consider, in the case of the problem of sessionization, exact matching of real session for constructing precision and recall measures. In the set of all real sessions C it is identifying the matching subset $\mathcal{M} \subset C$ of session that are completely identified and $\mathcal{S}$ the set of all session estimated by sessionization. In this case precision and recall are simply defined by equation 3.4 and 3.5.

$$p = \frac{|\mathcal{M}|}{|\mathcal{S}|} \tag{3.4}$$

$$r = \frac{|\mathcal{M}|}{|C|} \tag{3.5}$$

Both measure can be resumed using the harmonic mean $F = \frac{2pr}{p+r}$ resulting in an F-score. Using this measure, different algorithms for sessionization can be compared with a better one having a higher F value.

The same analysis can be performed counting the number of registers. However, as seen in the previous section data analysis, there exists a strong correlation between the number of registers and the number of sessions. Furthermore, counting the number of sessions or number of registers could be strongly related on the limit of large numbers.

3.11 Testing algorithm's performance

SIP processing

The SIP algorithm depends on the definition of the coefficient set $\{C_{ro}\}$. According to previous studies [DRV08a], the choice $C_{ro} = 3/2Log(o) + (o-3)^2/12o$ gives a better matches the distribution of size of the session extracted to the empirical distribution. Therefore, we use this value in our processing.

We solve the 15 month of SIP using the CPLEX linear programming solver. Each month take in average roughly 7 hours to be processed, completing more than 120 hour. The 71,080 different instances (Chunk) of our SIP (with $\underline{mtp} = 0$, $\overline{mtp} = 300$, and a maximum session size of 30) range in size from about 12,000 to 942,000 binary variables and 11,800 to over 783,358 constraints (reach in may 2010). For each instance of SIP, we set the maximum time limit to 300 seconds and the relative gap to one percent. This time limit is reach in average for only 28 chunks per month. However, those chunks correspond to the 30% of the processing time related to the integer linear problem.

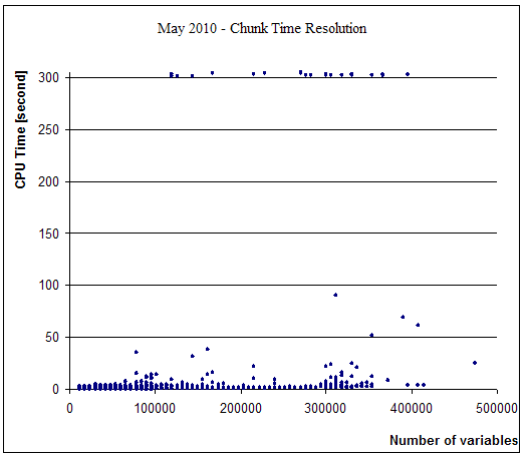

FIGURE 3.17: Solution time in seconds vs. number of binary variables.

The complexity of the SIP problem can be illustrated for May 2010. In this month, 27 chunks reach the limit having an average relative gap of 30% (Figure 3.17). With more than 100,000 integer variables sessionization problem with higher complexity becomes to appears, while the rest of the chunks are processed in less than 100 second.

BCM processing

We solve the 15 month of BCM using the CPLEX linear programming solver. Each month take on average 33 minutes to be solved with 100% of optimality. In this case BCM has a theoretical guarantied polynomial time for its resolution [CLRS01]. Furthermore, all chunks reach optimality. The number of different instances (Chunk) is 71,080 like in SIP since conditions are the same.

Traditional processing

We solve the 15 month of BCM using a traditional timeout heuristic. Each month takes on average at least one minutes to be solved. We compare our results with a traditional sessionization

timeout heuristic on all clean registers. The timeout heuristic is the fastest method.

3.11.1 Comparison of sessionization methods

We use the precision-recall performance measure for evaluating the proposed sessionization method. Both SIP and BCM have better results when compared to the timeout heuristic 3.6. Sessions found by optimization methods have nearly a 25% more sessions that match reality. However, comparing SIP and BCM, SIP performs slightly better than the other.

Method	Precision	Recall	F-Score
SIP	0.7788	0.6696	0.7201
BCM	0.7777	0.6671	0.7182
Timeout	0.5091	0.6996	0.5893

TABLE 3.6: Performance measure of the three sessionization method (15 month).

Comparing the performances by month shows the stability of the evaluation. Figure 3.18 illustrate the performance by month. The optimization methods clearly dominate over the whole period. MIP is still slightly better than BCM on most of the months.

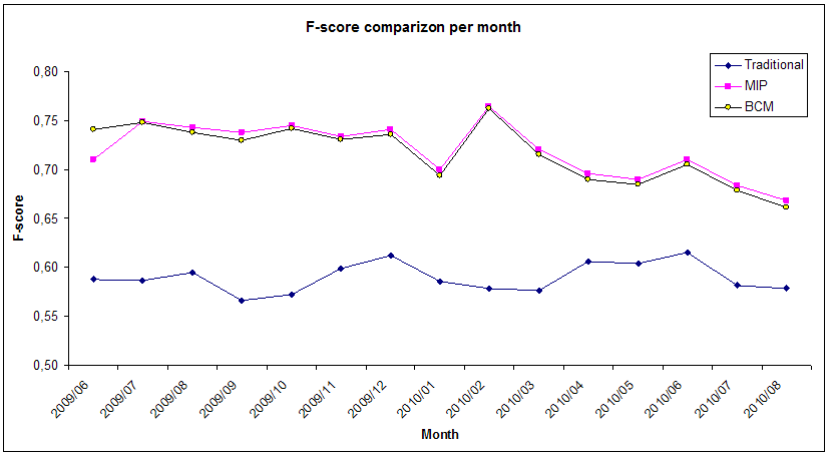

FIGURE 3.18: Monthly evaluation of performance.

Optimization method are limited by the maximum session size parameter, which is set at 30 for this experiment. Much larger sessions are found by the traditional timeout method. The optimization model also dominate when viewed by length(Figure 3.19).

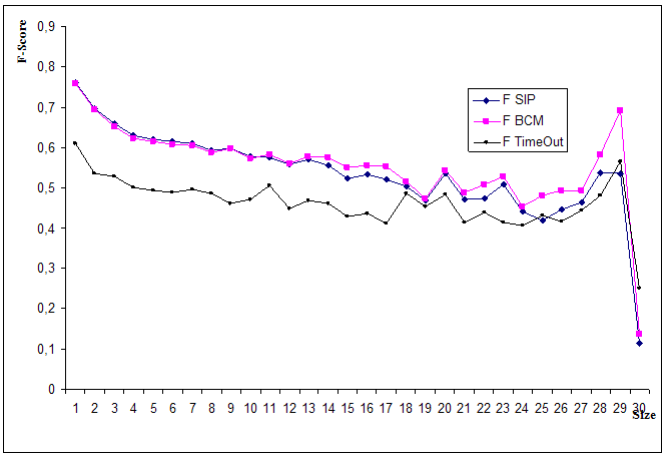

FIGURE 3.19: Performance grouped by session size.

3.12 Discussion

Web data is a complex and noisy data source, yet empirical regularity rules its statistical description. Several pre-processing techniques have been developed in the last ten years to support web mining and address the changing characteristics of the Web. The primary web data extracted are the hyperlink structure, the content, and the usage of a web site. All have some collection issues.

A web crawler collects the hyperlink structure but the time changing characteristic of the Web must be handled with page selection methods, revisit schedules, a politeness policy, and parallel processing. The challenges with web content are determining a weighting scheme as well as dealing with the visual representation and the dynamism of a website. Web usage data can be obtained indirectly from web logs or by direct retrieval. Integer programming has recently shown promise as an indirect method. Data preparation is a significant effort and a necessary cornerstone for web mining.

We present a new approach for sessionization using bipartite cardinality matching and integer programming. When compared to a commonly used heuristic, we find the sessions produced by our approach better match an expected empirical distribution. We also provide variations of our optimization models to further explore the likelihood of specific sessions and characteristics of sessions. Specifically, we find the maximum number of copies of a given session, the maximum number of sessions of a given size, and maximum number of sessions with a given web page requested in a given position for each session. The optimization method is proved to be a fruitful framework for the analysis of session.

At first sight, BCM method is the most practical since it solves more quickly than sip with similar result. However, SIP is a much more customizable method. We can visualize a methodology where the objective function coefficient C_{ro} are fitted to improve empirical distribution

functions. Such possibilities open the way for a much more accurate unsupervised method for extracting sessions from log files.

Chapter 4

Psychology of Decision Making

This thesis relates the application of psychological theories about decision making to web engineering problems. Psychology has been studying the process of decision making from ancient Greek times ("perception" for Protagoras). More recently, economists have applied such concepts to explain consumers' behavior. Both disciplines focus on their own specific goals, yet nowadays application from one field into another has produced fruitful results and resulted in several Nobel prizes. In this chapter a review of current psychological models of the decision making process are presented.

Cognition is the mental faculty of processing information, applying knowledge and changing preferences. In this sense, perception is the cognitive process of attaining awareness or of understanding sensory information.

Psychology has been engaged in several paradigms which focus on describing the mind's choice selection process. Roughly two categories can be identified, perceptual choice-based and motivational valued-based choice selection. The first one relates to a large series of neurophysiology experiments, and takes into account the time evolution of neural tissue electric potential firing rates before reaching a decision. This point of view is related to a mesoscopic dynamical description of natural science phenomena. The second approach has traditionally been the basis of many economic theories. This is the utilitarian scheme that represents a subject's preferences by a function dependent on the goods characteristic called "utility". In this theory, subjects tend to select the alternative with maximal utility. Nevertheless, this point of view does not consider the underlying characteristics of the process of decision making in isolation, rather it can be considered more as a static description.

Furthermore, two currents separate cognitive science, symbolicism and connectionism. The symbolic point of view considers that the mind works as a symbol manipulator and processor

within an algorithmic level framework. Computer chess games are examples of a further application of symbolicism. The connectionism point of view considers that cognition can be modeled as the processing of interconnected neural networks. The first approach corresponds to a higher level of analysis, where artificial intelligence and economics have their foundations, recovering empirical phenomena but not physical ones. The second approach relates to mimicking physical phenomena using simpler abstractions based on first principles, but not without difficulties in explaining higher-level mind results. On the symbolic side, approaches from utilitarian theory to games theory can be found, which are known as "preferential choices."

However, a degree of controversy exists between researchers of both sides in defining a unified point of view for describing the decision making process. Nevertheless recent advances in computational neuroscience suggest a promising new research direction, despite symbolism having already reached a mature level. This thesis relates with this emergent point of view, which has a long history of experimental observation of the microscopic neural mechanism. Two divergent theories for connectionist neurocomputational decision making are explained in the next sections.

4.1 Theories of Preferential Decision Making

Preferential choice [RBM06] in psychology relates to the inherently subjective preferences that arise as a mixture of thoughts, motives, emotions, beliefs and desires. Subjectivity comes from the fact that the preferences of a particular individual are not easily evaluated against objective criteria without knowledge of the individual's goals. In economics, preferential choices are defined in term of rationality's consistency principle (e.g. "utility maximization"). Nevertheless, people are imperfect information processors and limited in knowledge, computational power and time availability. Furthermore, violations of rationality principles have been observed in many experiments [RBM06]. Further extension and replacement via traditional axiomatic economics has been proposed under the name of "NeuroEconomics" [GCP08], which incorporates psychologically-based preferential choices for describing the behavior of economic agents.

4.1.1 Tversky's Elimination by Aspect (EBA)

Tversky's model, stated in 1972, describes an algorithm for defining the decision regarding a subject's discrete choice. The algorithm was originated through behavioral studies on decision making. Each possible choice is categorized by a set of discrete aspects Π. By means of fixing one of them, a strict subset of all possible choices remains. EBA repeatedly fixes one discrete aspect until a unique option is left. The simplified Monte Carlo algorithm follows:

1. **Initiate** each aspect $\alpha \in \Pi$ with a weight utility u_α. Then a probability is defined of selecting this aspect $P_\alpha = u_\alpha / \sum_{\alpha' \in \Pi} u_{\alpha'}$. Consequently, fixing an aspect to α_0 reduces the choice set Γ to $\Gamma_{\alpha_0} \subsetneq \Gamma$. The actual set of the selected option is $\omega = \phi$ void.
2. **Generate a random number** ρ that follows the probability distribution $\{P_\alpha\}$.
3. **Select** the aspect α_ρ, and insert into the set of selected option $\omega := \omega \cup \{\alpha_\rho\}$.
4. **Reduce** the set of possible aspects $\Pi := \Pi \backslash \Pi_{\alpha_\rho}$.
5. If $|\Pi| > 1$ return to step 2.
6. **Return** the selected choice k that corresponds to the unique element contained in $\Pi = \{k\}$.

This algorithm (heuristic) is called a "non-compensatory" strategic rule, since only one aspect is selected each time. In its first formulation it does not consider a previous order of attributes, but can be generalized using conditional probability resulting in a hierarchy of aspects [TS79]. "Loss of aversion" is modeled by the introduction of an asymmetric function in the domain of losses [TS93].

4.1.2 Prospect Theory

This theory appeared as a critique to Expected Utility Theory (EUT) [KT79] as a descriptive model of decision making under risk. The approach considers behavior characteristics like "loss of aversion" from the field of psychology. People tend to underweight outcomes that are merely probable in comparison to outcomes that are obtained with certainty. Another effect considered is that people tend to discard attributes that are shared by all choices (the "isolation effect").

Subjects are confronted by the selection of a kind of contract called a "prospect." Prospects are quite similar to the "lotteries" of EUT, but the formalism is presented as it was conceived. A prospect can be represented by a tuple $(x_1, p_1; \ldots; x_n, p_n)$, where x_i is the outcome to be obtained with probability p_i on n exclusive possibility. Traditional expected utility values are computed as $U = \sum_{i=1}^{n} p_i u(x_i)$ but in this approach this value is adjusted by rescaling function $v(x)$ and $\pi(p)$. Furthermore, $V = \sum_{i=1}^{n} \pi(p_i) v(x_i)$ is interpreted like an expected utility value.

The function π reflects the impact of the probability value on the subject's appreciation of the whole prospect and must accomplish $\pi(0) = 0, \pi(1) = 1$. It is important to indicate that probabilities are never exactly perceived by the subject so other factors can influence such function π. This function collects some behavioral observations: overestimation of rare events, overweighting and sub-certainty for the additional value of probability. Those properties have been shown to imply that $Log(\pi)$ is a monotone convex function of $Log(p)$ and with limits on values of p near 0 and 1 (relaxing condition $\pi(0) = 0, \pi(1) = 1$).

The function $v(x)$ assigns a subjective value to the outcomes, which measure preferences with respect to a reference point reflecting gains and losses. The human perceptual machinery has been observed to respond to changes instead of absolute values. Many sensory and perceptual studies suggest that the psychological response is a concave function in the range of gains, convex for losses, and even steeper for gains than for losses. The function will have an S-shape with $v(0) = 0$.

4.1.3 Expected Utility Theory (EUT)

This approach is related to decisions under uncertainty and risk [MCWG85] that occur for example in insurance strategy. Risky alternatives have a discrete set of outcomes, which are described by the lottery concept. A lottery is a tuple $L = (p_1, \ldots, p_n)$ where $\sum_{i=0}^{n} p_i = 1, p_i \in [0, 1]$ is the multinomial distribution of outcomes occurrence.

The concept of utility in this stochastic framework is recovered as the Von Newman-Morgenstern expected utility function $U(L) = \sum_{i=1}^{n} p_i u_i$ where u_i is identified as the utility value for the case i. Furthermore, one lottery is preferred versus another by a subject if its expected utility value "i" is higher. If the preference relation between lotteries is constrained by the continuity and independence axiom, then it can be represented as an expected utility. The continuity axiom establishes that small changes in probabilities do not affect the preference ordering. The independence axiom establishes that convex combination with a third lottery does not affect the preference ordering. This is called the expected utility theorem.

Nevertheless the flexibility given by the last theorem allows the incorporation of risk aversion in the theory as special utility functions.

4.1.4 The Random Utility Model

A popular model in economics is the random utility model [AL95], where subjects are utility maximizers but have uncertainty about the utility value. The randomness is attributed to imperfections in the rational process. Nevertheless, subjects decide to maximize their random utility resulting in a probability distribution for each choice. The random component error of the utility is supposed to be additive to a classical utility function, and dependent on the assumption of its distribution. In this way there are at least three models assuming linear, normal (Probit) and Gumbel (Logit) distribution.

The model is based on the following assumptions:

- Individuals are faced with selecting an option from a set of finite and uncorrelated alternatives. Alternative j has associated attributes quantified by the variables X_{jkq} for individual q.
- Individuals have common motivations for making decisions.
- Individuals behave rationally (utility maximizer) and possess perfect information.
- Each alternative j is associated with a utility function U_{jq} related to the individual q.
- The utility has a stochastic error component $U_{jq} = V_{jq} + \epsilon_{jq}$. The term $V_{jq}(X)$ corresponds to the traditional notion of utility depending on the feature values $\{X_{jkq}\}$ of the choice j. The term ϵ_{jq} is a stochastic iid term, with average $E(\epsilon_{jq}) = 0$.
- A probability value P_{jq} of choosing the option j is given if the individual q utility is greater than all others values $U_{jq} \geq Uiq$, $\forall i$. Then the probability is given by $P_{jq} = P(\epsilon_{iq} - \epsilon_{jq} \leq V_{jq} - V_{iq},\ \forall i)$.
- If the density of the error term is $g(\epsilon)$ then the probability P_j of choosing the option j is $P_j = \int_{-\infty}^{+\infty} g(\epsilon_j)[\prod_{i \neq j} \int_{-\infty}^{V_j - V_i + \epsilon_j} g(\epsilon_i) d\epsilon_i] d\epsilon_j$

Several efforts have been made for arriving at a simpler expression for P_j in order to have a simpler model. The choice probability will depend on the $g(\epsilon)$ distribution. The most used case is the Multinomial Logit Model (MNL) where $\epsilon \sim \text{Gumbel}(\mu, \theta)$, which results in the Logit distribution (4.1).

$$P_j = \frac{e^{\beta V_j}}{\sum_{i=1}^{n} e^{\beta V_i}} \tag{4.1}$$

The Gumbel distribution is an extremum class that is invariant to the maximum operation, in the sense that the maximum of Gumbel's distributed variable is also Gumbel distributed. The MNL is also equivalent to the problem of maximum entropy having been restricted to an expected utility value. One important property recovered from this model is the independence of irrelevant alternatives, according to which the relative probability of selecting two alternatives does not change if we add a third independent alternative. However, this property is violated in much behavioral experimentation.

Variations of the MNL are Hierarchical Logit (HL), Mixed Logit (ML), Heterocedastic Extreme Value Model (HEVM), and Probit Model (PM). All of them result from relaxation of the assumptions of the MNL.

Hierarchical Logit (HL) considers relaxing the independence of error ϵ_i assumption, so that independence is then established by groups called "nests". A nest is represented by the probability

of choosing the nest, and probabilities within the nest are conditionals. For each nest j a utility nest value is given by $V_j = \phi_j Log(\sum_{i \in \text{nest}_j} e^{\frac{1}{\phi_j} V_i})$. Hence the last formula is consistent with a nest of a single choice, since it collapses to the utility of the corresponding choice. Furthermore, alternatives are grouped by nested levels of independence, where probabilities are given by Bayes conditional probability. The probability of an independent choice j in a nest i is given by top-level nest $P_{ij} = P_i P_{i|j} = (e^{V_i} / \sum_{l \in \text{toplevel}_j} e^{V_l})((e^{V_j/\phi_i} / \sum_{k \in \text{nest}_i} e^{V_k/\phi_i})$.

The mixed Logit (ML) goes further, including in the utility description a random component that depends on the data. For instance the first proposal consists of $U_{jq} = V_{jq} + \epsilon_{jq} + \sum_k \eta_{jkq} X_{jkq}$, where η_{jkq} is a random variable with $E(\eta) = 0$ and a different pattern of correlation and heteroskedasticity.

Another extension of the MNL is the Probit model where "ϵ" is considered distributed by normal distribution with $E(\epsilon) = 0$ and an arbitrary covariance matrix Σ. Since the difference between two normal distributions is also a normal distribution, the probability P_i of choosing the option i is given by $P_i = \int_{-\infty}^{V_i - V_1} \int_{-\infty}^{V_i - V_2} \dots \int_{-\infty}^{V_i - V_j} N(0, \Sigma_{\epsilon_j - \epsilon_i}) d\epsilon$.

Finally, the heterocedastic model (HEVM) can be stated on the basis of different variances for ϵ.

4.2 Connectionist Theories

The brain, particularly the human one, is one of the most complex systems that science has attempted to describe. Having more that 10^{11} neurons and 10^{14} connections, how does this machine achieve consciousness? Neuroscience takes this question into account by attempting an answer based on neuronal mechanisms in the brain. This is a much more fundamental point of view compared with the previous symbolic-based theories. In this case a much higher complexity is exhibited since consciousness is a time-dependent phenomenon, and macroscopic descriptions depend on an enormous number of microscopic events.

Current connectionism is synonymous with "computational neuroscience." This has as a definition "to study the brain as a computer, and using the computer to study the brain." The hypothesis is that neuronal electric potential variations relate with states that codify sensory feelings and abstraction. Over the last half century, several models have been proposed for explaining the reaction time of a decision. Perceptual choice is a research area where the processing and context of alternatives is studied. It considers as basic processing concepts the accumulation (i.e. integration) of perceptual data values and controlling the criteria for such values. A Neuronal Activity Theory is described and related to the decision actions of individuals.

Paraphrasing Busemeyer's paper [BT93]: "The first unavoidable fact about human decision making is that preferences are inconsistent (Classical Economic Rationality). We propose that this inconsistency arises from changes in preference over time and that this process of change must be rigorously specified so that it can be evaluated as a viable scientific explanation. Any psychological theory of decision making must be capable of predicting how choice probability changes as a function of the events and payoffs that define each pair of actions. This first fact rules out the deterministic derivatives of expected utility theory and points to the need for probabilistic account. It resumes the aims of the neurocomputing point of view for decision making."

4.2.1 Neuroscience Foundations

The first model of the neuron dates from 1930 [Hil36], where electrical potential in nerves was described as an accommodation of excitatory inputs from other neurons, and which theory required that a given threshold slope was needed for obtaining a response. Learning principles based on neural activity were presented in 1949 by Hebb [Heb49] establishing that repeated excitation of a neuron by another strengthened over time augmenting the efficiency of the process. This process inspired the model of artificial neural networks in artificial intelligence. The strength of a connection between neurons is represented by a number called weight. This value represents the fraction of activity exchanged from one neuron to another. Then Hebb's learning rule stated that those weights change for learning purposes.

In 1952 Hodgkin and Huxley [HF52] proposed a differential equation description for the conductance of the neuronal cell membrane, for which work they received the Nobel prize in 1963. Such an equation described the electric potential V across the membrane and ion conductance g_x (x=Na,K). The evolution is presented in the following nonlinear multivariate differential equation:

$$C_m \frac{dV}{dt} = -g_L(V - V_L) - g_{Na} m^3 h(V - V_{Na}) - g_K n^4 (V - V_K) \tag{4.2}$$

$$\frac{dm}{dt} = \alpha_m(V)(1 - m) - \beta_m(V)m \tag{4.3}$$

$$\frac{dh}{dt} = \alpha_h(V)(1 - h) - \beta_h(V)h \tag{4.4}$$

$$\frac{dn}{dt} = \alpha_n(V)(1 - n) - \beta_n(V)n \tag{4.5}$$

Computational theories of the brain have been extended by the study of cerebellum functionality. Its best known function is fine control of body movement. The learning aspect of this structure

was first explained in 1969 by Marr [Mar69]. It was shown that the cerebellum's neural connections encode and control fine motor skills. The first proposed learning model for the cerebellum was the "Perceptron" [Alb71].

Neural associative memory [Wil69] was first proposed in 1969. In this model, neural weight interconnections are binary-valued and updated by Hebbian rules. In this way learning is represented by those weight patterns. This is called reinforcement learning, which has been biologically inspired by the neural learning of bees [MDPS95], and experimentally checked.

Decision making reaction time was first covered by Stone in 1960 [Sto60b]. An empirical model was proposed for describing a choice's reaction time together with agreement on the psychological basis and experimental data. The proposed algorithm was based on the famous Wald's Sequential Probability Ratio Test (SPRT) for deciding between two alternatives. The subject perceived noise evidence (x) about two independent choices. The test consisted in comparing two thresholds for deciding the option, and the value used was the accumulated log ratio between probabilities $p_i(x)$ to choice option i. This procedure was proved to be optimal in the sense that the average number of steps needed for making the decision was minimal with respect to other mechanisms based on the statistical test [WW48]. Despite the statistical assumptions of the nature of this model, it was the basis for the understanding of further stochastic theories.

4.2.2 Biological Insight on Decision Making

Neurocomputing models are rather simpler representations of what really occurs in a brain. The aim is to obtain theories that predict physically measurable values of psychological phenomena, and the first class objects in the field are neurons. Neurocomputing establishes mathematical models based on neuronal experimental facts from psychology. Neurons are distinguished from other cells by the experimental fact that they can be excited electrically by other neurons as described by formulas 4.2 to 4.5. Using a new method for measuring neuronal activity, a 2002 experiment on rhesus monkeys revealed how decisions based on visual stimuli correlate with the middle temporal area (MT) of the brain [RS02]. The experiment consisted in presenting a screen with random moving points, where subjects had to decide to move their eyes in one of two directions. Each monkey was implanted with electrodes recording the activity of 54 neurons of the Lateral Intra-Parietal cortex (LIP) and in the MT. Figure 4.1 presents a rough description of the anatomy of a human brain, where the LIP region on the brain's right upper corner and the MT area in the bottom zone are present in most mammals. Eye movements were tracked on a horizontal and vertical plane. The subjects learned to respond to a reward of fruit juice if they indicated by eye movement the correct direction of movement of points on the screen. The screen was presented with a random movement of dots and the reaction

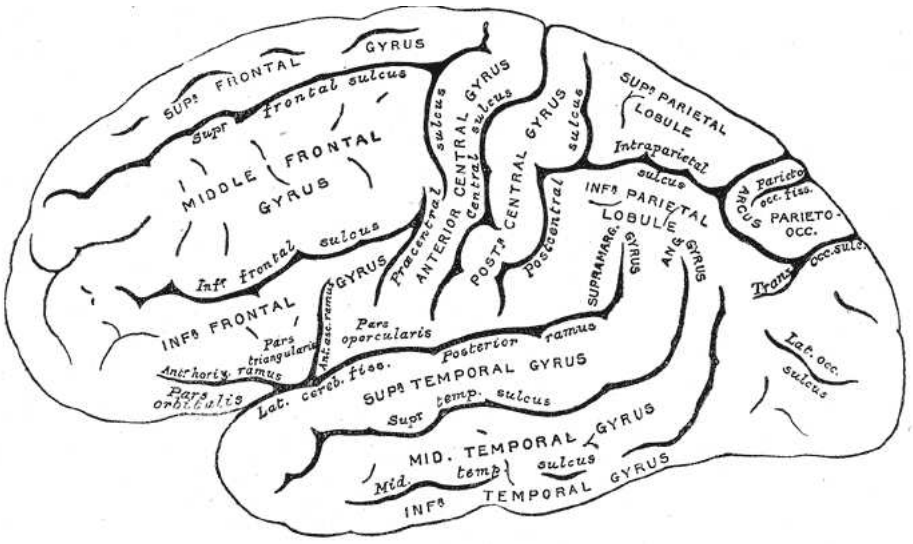

FIGURE 4.1: Human Brain Anatomy (wikimedia commons repository, from public domain ebook H. Gray Anatomy Descriptive and Surgical, 1858, UK).

time of the decision was recorded. The experimental data was interpreted at a physiological level. Before reaching a decision the increase and reduction of the spike rate was observed. When the activity had reached a threshold value and the decision process was complete, the monkey responded 50 msec later, and the LIP activity terminated. The process suggested an accumulation of information toward a threshold. The LIP area is then a temporary storage area of partially-recorded information, and is in charge of processing the decision by means of a kind of sequential analysis processing analogous to Wald's SPRT.

The MT area appears to be a temporary storage and processing device for visual information that is transmitted to the LIP area for further processing. Earlier studies [BSNM06] showed a linear correlation of visual stimuli on the neural activity of the visual cortex MT. The experiment was the same as [RS02], measuring neuronal activities by mean of electrodes.

More recently in [HDS06], the same experiment was modified to induce neuronal stimulation of the monkey brain's LIP area. The purpose of this external stimulation was the observation of an artificially-induced decision on the subject. The experiment was set up first identifying those neurons presenting activity and which correlated with each decision choice. Neurons associated with one alternative were micro-stimulated by electrodes. Half of the measurements were taken with the bias of micro-stimulation of the correct choice and the other half of the wrong one. The result was that micro-stimulation produced an effective bias on the monkey's choice, but if the evidence of the moving dots was strong enough (i.e. a clearer movement) then the bias was smaller. Furthermore, the micro-stimulation effect was small, but significant. Another effect noticed was an increased time taken for the wrong decision and a faster response for stimulation on the correct answer. MT area (Figure 4.1) micro-stimulation was also explored but the results were weaker than for LIP. Choices seemed to have been affected as if the dot motion

perception were distorted to the point of reaching logical inconsistency, generating artificial evidence over visual perception, but having the duration of the artificial stimuli. Finally the experiment concluded that decisions in the experiment were strongly correlated with the neural activity levels of specific neurons in the LIP that receive sensory evidence from the MT areas, using a threshold level mechanism.

Furthermore, the neurophysiology of decision making has been explored with multiple-choice decision making [CKS08]. The experiment was similar to the previous using a screen with random dots moving principally in four or two directions. 90 electrodes per brain on the LIP area cortex were installed in monkeys trained by reward to identify the busy pattern. It was observed that it took a longer time to decide between four alternatives than two. An important effect was the reduction of the maximum activity level in the LIP (threshold) compared with the two-choice option. The interpretation is related to a kind of trade-off adjustment of the threshold for gaining speed in decision versus accuracy. As the number of alternatives increases, the brain needs to accumulate more information in order to make a decision and takes more time. The time usage per decision seems to be limited by an internal mechanism based on the implementation of a deadline. Previous findings about information accumulation in the LIP were thereby confirmed.

Moreover this decision mechanism is purely based on statistical efficiency and is more automatic than if driven by a higher level of reasoning. Psychologists argue that those empirical mechanisms are followed by a reliability assumption about the decision. A recent experiment [KS08] presented the correlation of the LIP areas with the degree of confidence. The certainty of the decision in the experiment was measured by implementing the decision to opt-out with a small reward revealing the confidence degree.

4.2.3 Wiener Diffusion Process

In 1978, Ratcliff [Rat78] postulated a memory-based decision theory using a simple "random walk" process that was tested with experimental data. The model reproduced the distribution of reaction times and error latency. The model considered for each choice was a random walk with drift 4.6, but the variable X does not represent any biological value since it take negatives values.

$$dX_i = I_i dt + \sigma dW_i \tag{4.6}$$

The last stochastic process is called a Wiener process. Each time a comparison over two thresholds ($\pm a$) is made, a decision can be considered to have been made when any comparison terminates in a hit with an upper limit ($+a$), or all processes reach a lower limit ($-a$), stopping without a decision. Once a decision is ready, a response triggers another mechanism for response. The entire process is driven by the $\{I_i\}$ set of drift, the alternative with the bigger value

having the best possibility of being chosen. In this work the drift value is called "relatedness" of the alternative to the problem decision, which integrates partial information about the supporting evidence for the choice. Furthermore, this decision model uses the same criteria as the Wald test after which the stopping rule ensures a minimum decision time for a given accuracy [Sto60b]. Fortunately this model has exact solutions for decision time and distributions [Rat78] that are used for parameter-fitting purposes using the maximum likelihood method.

Another criterion for decision definition is establishing a timeout limit. Under this consideration the process loosens the variability of the time to decide, replacing it by the timeout limit parameter of the model t_c. Figure 4.2 describes the evolution of the variable X. The model parameter

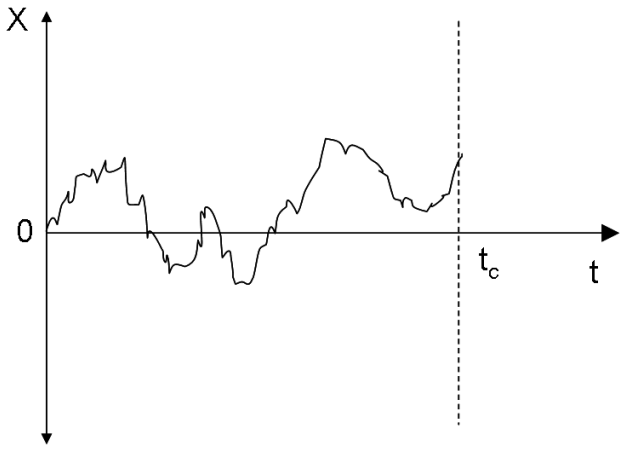

FIGURE 4.2: Weiner Process with timeout t_c.

magnitudes "I_t" reflect the difficulty in reaching a decision, since for smaller values the time to make a decision in the threshold's model can be a larger value. In the absence of noise $\sigma = 0$ the model is interpreted as the subject reaches the decision with the maximal I_i value. In this case the time used to reach a decision is deterministic and tc = $t_c = MaxI_i/a$.

The performance of models is measured by the error rate, which is then considered as a mechanism of decision under noisy evidence. The error rate is a function of the model parameter I_i, σ and a. In the case of a model with a threshold and two choices ($I = I_1 - I_2$) the error rate is $P(a) = 1/(1 + e^{\frac{2Ia}{\sigma^2}})$. In a case with a timeout $P(t_c) = \int_{-\infty}^{-I\sqrt{t_c}/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{I^2}{2}}$ is the error rate.

This model was extended [RVZM] to allow some parameters to be Gaussian random variables. In this case additional parameters like the mean and variance of the Gaussian distribution need to be adjusted. This change improves the degree of fit to the experimental results. Additionally, the starting point for Brownian motion is considered to be an additional parameter of the model and is considered to follow a uniform distribution.

A boundary mechanism was considered for modeling threshold stopping time. In this case an absorbing boundary was used for terminating the process (no time pressure) [MIOK88].

Reflecting boundaries were also used for controlling the integrated information, like a lower limit on accumulated value.

Under the same diffusion framework, the problem of word recognition was studied [MIOK88]. In this work lexical decision tasks were tested on human subjects and decision time recorded.

4.2.4 Decision Field Theory (DFT): The Ornstein-Uhlenbeck Process

Busemeyer et al. in 1993 [BT93] elucidated the Decision Field Theory (DFT), a mathematical description of decision making with a cognitive basis. This framework aimed to provide an explanation of experimental violation of some common economic decision making principles such as stochastic dominance, strong stochastic transitivity, independence over alternatives and serial position effect on preferences. Furthermore, it explained speed-accuracy trade-off effects in decision making, the inverse relation between choice probability and decision time, changes in the direction of preferences under time pressure, a slower decision time for avoidance as compared with approach conflict, and preference reversal between choice and selling price measures of preferences. The first presentation of this theory considered the mentioned characteristics and constructed it on this empirical evidence.

The theory is described by a time-dependent vector $X = [X_i(t)]_{i=1}^{n}$ a n-dimensional preference state vector [BJJT06], which coordinates $X_i \in \mathbb{R}$ relative to the alternative i. At $t = 0$ such a vector represents preferences before any information about actions is considered, such as memory from previous experience ($\sum_i X_i(0) = 0$), but if a decision doesn't have an antecedent then the process starts from zero ($X_i = 0 \forall i$). The decision making process consists of the time evolution formulated in discrete time step h 4.7.

$$X(t) = S \cdot X(t-h) + V(t) \tag{4.7}$$

$$S(h) = Id - h\Gamma \tag{4.8}$$

$$X(t) = \sum_{k=0}^{T} S^k \cdot V(t-kh) + S^T \cdot X(0) \tag{4.9}$$

Where S is the nxn feedback matrix, $V(t)$ is a noise input vector called the valence, and Id is the nxn identity matrix. When $h \to 0$, X is a continuous Ornstein-Uhlenbeck process, and if $\Gamma = 0$ then it is reduced to the Weiner process. The matrix $\Gamma = [\gamma_{ij}]$ is symmetrical ($\gamma_{ij} = \gamma_{ji}$) and diagonal elements are equals ($\gamma_{ii} = \gamma$). If such a matrix represents intensity of connection, then diagonal elements represent self-feedback and off-diagonal elements represent lateral inhibitory connections that can vary over conceptual distance between choices.

The stochastic valence vector is decomposed into $V(t) = C \cdot M \cdot W(t)$. The matrix $C = [c_{ij}]$ is called the contrast matrix, which is designed to compute the (dis)advantage of an action relative

to others, and its value $c_{ij} = \delta_{ij} - 1/(n-1)(1-\delta_{ij})$ (δij is the Kronecker delta). The matrix $M = [m_{ij}]$ is interpreted as the affective evaluation of the consequence of performing the choice j but not the choice i. The stochastic vector $W(t) = [w_i(t)]$ corresponds to the weight of the result of the attention process for each choice. For instance, if the decision is about a characteristic of a visual stimulus, then such a vector corresponds to the result of the processing of the visual cortex. Such a vector is supposed to fluctuate, representing changes of attention over time, and is considered as a stationary process with $E(W) = \bar{W}h$. In this case the product $\mathrm{Cov}(W) = \Psi h$ is an average evaluation of each action at a particular moment. Finally, V will represent advantage or disadvantage over the average of the other action at any given moment.

DFT is based on the Weiner process including a dissipation term with parameter λ (4.10). The new equation introduces the property of asymptotic attraction for the larger t of the variable X. Notice that for $\lambda > 0$ this equation has an attracting fixed point $X_i^0 = I_i/\lambda$, otherwise this point is unstable. A property acquired with the introduction of this new term is the "recency effect." Furthermore, early accumulations of information from I_i are erased from the new term, in this case recent stimuli drive the time evolution of X.

$$dX_i = (-\lambda X_i + I_i)dt + \sigma dW_i \tag{4.10}$$

Thresholds are incorporated in the theory in the same way as for the Weiner process.

4.2.5 Leaky Competing Accumulator Model (LCA)

In 2001 Usher and McClelland [UM01] published a new diffusion model for decision making that incorporated the dissipation term of Busemeyer's DFT and a lateral inhibition effect. This new term is included for taking relative accumulated evidence into account.

$$dX_i = [I_i - \lambda X_i + \alpha f_i(X_i) - \beta \sum_{j \neq i} f_i(X_i)]dt + \sigma_i dW_i \tag{4.11}$$

The equation 4.11 presented in the 2001 paper considers the following term. I_i is an evidence value in favor of the alternative i that is accumulated from other devices such as the visual cortex and serves as an input to the LCA process. Those values are supposed to be constrained as $I_i \geq 0$ under neurophysiology reasoning. Evidence values are accumulated in the variable X_i in favor of the alternative i. The λ parameter takes into account the decay from DFT. The α parameter is a recurrent excitatory source coming from the unit i and modulated by the function $f_i() \geq 0$. Lateral inhibition between accumulator units is controlled by the β parameter and considers equal effect for all units, but is modulated by the function $f_j() \geq 0$. The accumulated values are considered biological values, like neural activity (rate of spikes), which are then restricted to being positive. Function $f_i()$ is near to a linear neural response, and an approximation could

be considered as $f_i(x) = x$. This is a difference from previous models, since the aim is that X be adjusted to the real biological value of neural activity. Hence DFT and the Weiner process use variables that take negative values. The model then considers that a decision begins when the first accumulator X_i reaches $i^* = \text{ArgMax}_i(X_i(t^*))$ where $X_{i^*}(t^*) = X^*$ on the threshold X^*. Otherwise, if the phenomenon is time constrained (i.e. with a timeout) then the actual maximal value is used as the decision. Without loosening of generality, I_i values are supposed to be $\sum_i I_i = 1$ since its value could be considered constant during the decision process. This is valid for short-term decision problems.

In the linear approximation of the f_i function and time-constrained decision case, there are exact solutions for the probability density of the process. In [UM01] the process was analyzed for the case of two dimensions and proved to be an Ornstein-Uhlenbeck process with a solution in terms of $x = X_1 - X_2$ following a Gaussian distribution 4.12 with a time-dependent mean 4.13 and variance 4.14 (see Appendix A).

$$x \sim N[\mu(t), \sigma(t)] \tag{4.12}$$

$$\mu(t) = \frac{I_1 - I_2}{\lambda - \alpha - \beta}(1 - e^{-(\lambda-\alpha-\beta)t}) \tag{4.13}$$

$$\sigma(t) = \frac{\sigma}{\sqrt{\lambda - \alpha - \beta}}\sqrt{1 - e^{-2(\lambda-\alpha-\beta)t}} \tag{4.14}$$

$$\epsilon(t) = \frac{2\mu(t)}{\sigma(t)} \tag{4.15}$$

This model exhibits asymptotic behavior for the average $\mu(t) \to \frac{I_1-I_2}{\lambda-\alpha-\beta}$) and variance ($\sigma(t) \to \frac{\sigma}{\sqrt{\lambda-\alpha-\beta}}$). For measuring the accuracy of the model 4.15, it was defined using a signal theory accuracy ratio between the separation of the mean of success and failure versus the variance [Wic02]. This value is asymptotically $\epsilon(t) \to \frac{2(I_1-I_2)}{\sigma\sqrt{\lambda-\alpha-\beta}}$.

4.2.6 Quantum Probabilities Approach

A novel approach for mathematical modeling of decision making is the quantum dynamic approach [BWT06]. In such a model the mathematical framework of quantum mechanics is applied for explaining decision making. Nevertheless no neurophysiology principles have been used for supporting those assumptions, but the formalism is used as a richer formulation that better explains some known violations of rational theories [PB09]. The new approach is known as Quantum Decision Theory (QDT), and claims that the brain is a quantum computer.

A Markov process has been used with considerable success in explaining aspects of decision making. Quantum dynamics has many similarities with Markov processes. Quantum states are related to complex probability amplitudes $\varphi \in \mathbb{C}$, which inner products between them are transition probabilities between their respective states. The evolution rule is similar to Kolmogorov's

forward equation for probability P_{ij} transition $\frac{dP_{ij}}{dt} = Q \cdot T(t)$. In the case of quantum evolution the state evolves according to 4.16, where H is the Hamiltonian operator.

$$\frac{d\varphi}{dt} = -iH\varphi \tag{4.16}$$

A basic example considers the quantum-based two choice problem [BWT06]. States are described by a level of confidence $l \in \{0, 1, \ldots, m\}$ that indicates a measure of the likelihood of the decision where zero is indifference. The decision is binary $+/-$. Hence, the number of states is $m+1$, since a codification of possible states is the set of integers $\{-m, -(m-1), \ldots, m-1, m\}$. The system is represented by linear combinations of wave functions representing those states. The Hamiltonian matrix $H = [h_{ij}]$ for this problem is defined in 4.17.

$$h_{ij} = -j\mu\delta_{ij} - \frac{\sigma^2}{\Delta^2}(\delta_{i-1,j} + \delta_{i+1,j}) \tag{4.17}$$

If the initial state of the system is given as φ_0, then in a time t the result is that $\varphi(t) = e^{-iHt}\varphi_0$. If the subject has a deadline in which to decide, the process stops and a measurement of the level of confidence is performed. A quantum measurement corresponds to a protection of the wave function of the system in the subspace that represents the measurement. In this case a choice corresponds to a projection to a positive confidence value for choice (+) and negative for choice (−). Probabilities are then inner products of those projections on the wave function.

More recently [BPF10], this approach was applied to the prisoner dilemma and compared to the Markov approach.

4.2.7 Grossberg's Nonlinear Connectionist Theory of Decision Making

Grossberg [GG87] in 1987 stated a model based on neural network dynamics for explaining decision making under risk. This theory is called Affective Balance Theory (QABT). The theory states that the stimulus is not the only factor in decision making, but the cognitive context alters the interpretation and consequent behavior. This model incorporates emotional and physiological factors. The dynamics of the theory become non-linear at each step. Nevertheless, it is presented as an example of a further enhancement of the theory.

4.3 Discussion

From earlier times static heuristic models of decision making like the Utilitarian Scheme, Random Utility Models, and Expected Utility Theory have been used for describing decision making with considerable success. Nevertheless, axiomatic restrictions (e.g. weak stochastic transitivity) that these models impose are violated in experiments [RBM06], thus constituting paradoxes. The main reason is that people are imperfect information processors and limited in knowledge and processing capabilities.

Time-dependent stochastic processes of decision making were explored in psychology using first principles, thus giving rise to the field of NeuroComputing. Furthermore, models like the Weiner Process, DFT and LCA successfully explain many aspects of the decision making of individuals. Such theories were consistently tested in the laboratory by means of experiments from neuronal activity level to final decision making. However this is still an active research topic, offering promising new proposals like QDT.

One author argues that [Gro00]: "Thus, just as in the organization of the physical world with which it interacts, it is proposed that the brain is organized to obey principles of complementarity, uncertainty, and symmetry-breaking. In fact, it can be argued that known complementary properties exist because of the need to process complementary types of information in the environment." Indeed, throughout the history of science, theories based on first principles have been more robust and have had more predictive power. The Apollo mission was successful in planning trajectories using Newton's theory calculations. In the same way, neurophysiology's inspired theory of decision making aims to produce a kind of law of decision making describing choice probabilities and time.

On the other side of the spectrum, machine learning pretends to explain everything with complex algorithms trained on observed data. Wired magazine published in 2008 [And08] an article claiming the end of theories, based on computer power and the information already accumulated. They claim, "Learning to use a "computer" of this scale may be challenging. But the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." Machine learning could be an efficient answer to theoretical approaches. Nevertheless, machine learning is restricted to the data used for learning. Nothing justifies using the trained algorithm to predict behavior with a variation on the condition of the problem. First principle theories predict changes under such conditions; ergo they are still the basis for further scientific advancement.

Chapter 5

A Stochastic Model of the Web User

Data mining techniques have been used by e-businesses for decades [VP08] using generic machine learning algorithms. This research surpasses traditional models, by using a time-dependent dynamic model of web user browsing that captures the statistical behavior of an ensemble of real web users. Such an approach is based on the neurophysiology representation of decision making in the brain of an agent, known as the LCA model [UM01].

Two different ways exist for describing a natural phenomenon, generic methods and first principle-based methods. Generic methods relate to general mathematical models that are fitted to available data in order to predict and describe observations. Data mining techniques are based on generic methods for discovering unexpected regularities in unknown data. Machine learning algorithms are further optimized for handling large amounts of data. Furthermore, without any preconceived understanding of the observed data, data mining techniques constitute an outstanding tool for analysis. If on the other hand, first principles are known to rule the observed data, then a much better approximation of the phenomenon can be obtained if the mathematical model adopts such conditions.

First principle mathematical models do not discard the use of generic machine learning techniques. The first principle description only partially describes the real behavior of the phenomenon, since a theory of everything has yet to be discovered, or perhaps never will be. Effects not covered by the theory are mainly represented by parameters of the models. For instance, the Newtonian law that describes a harmonic oscillator requires the value of a mass and spring constant. Such parameters represent atomic interactions that are abstracted and simplified into one adjustable value. Machine learning techniques help to adjust such parameters using available data. In this sense, both approaches are intrinsically complementary for better modeling of natural phenomena.

As presented in chapter 4, the human decision process is a natural phenomenon which has been studied by the field of psychology for decades. First principle mathematical models have been tested experimentally and related with the physical description at the level of neural electrical activity in the brain. Human decision making behavior is described by a stochastic process that has many parameters. This theory is applied to each navigational decision process of a web user confronted by a web page, who is considering each hyperlink as a possible choice. The proposed model predicts a probability distribution for each choice, as well as the time taken to reach a decision. Such probabilities are dependent on parameters that need to be fitted to specific algorithms that are described in Chapter 6.

In this chapter the proposed model of web user navigation behavior based on first principles is described.

5.1 The Time Course of The Web User

As has been reviewed in Chapter 2, the current models of a web user's navigational behavior have related more to describing the visited combination of pages than the timed sequence of them. The present chapter proposes a dynamic model that describes the forces involved in link selection or otherwise leaving the site, recovering the jump sequence between pages.

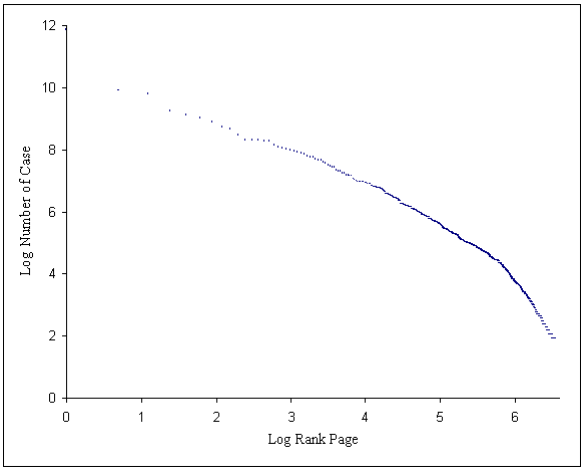

FIGURE 5.1: Log-Distribution of the ranked first page on a session (source: web site http://www.dii.uchile.cl).

A web user's visit starts when he reaches a page from a chosen web site, which is defined as the initial page on the trail of page sequence (also called a "session.") Initial page arrivals have a very typical distribution (Figure 5.1), and the implicit mechanism for reaching such pages could be from a search result (e.g. Google), navigator bookmarks, navigator suggestions and links from other web sites. As an example over a period of 15 month the first page distribution on the http://www.dii.uchile.cl web site has the distribution shown at the figure 5.1. The root page has

nearly the 35% of all first visits, while the rest have scarcely less than 5% (Table 5.1).

Page	Percent of visits to a first page
/	34.6
/ cea/sitedev/cea/www/index.php	4.8
/ webmgpp/	4.4
/ cea/sitedev/magcea/www/index.php	2.5
/ diplomas/	2.2
/magister_doctorados/	2.0
/ mba/paginas/superior2.html	1.5
/ingenieria_civil_industrial/	1.4
/ mgo2007/	1.2
/educacion_continua/	1.0
/ diplomas/pages/inteligencia2010.htm	1.0
/ mbaft/	1.0

TABLE 5.1: Most visited first pages in session (web site: http://www.dii.uchile.cl).

Intuition indicates that the first page will define the semantic of further navigation of a user on a web site. However, other factors appear to influence this decision. In the case of a university computer center, the first browser page is usually set to a given department's page, in which case the default browser's page visits do not influence the semantic of further visits. Many first-time visitors to the web site (beginners) will typically try a first page that probably does not clearly represent their purposes, since they are unfamiliar with the site. We can conclude that most of the web user's intentions are contained in the"jump" decisions between pages, rather than the first page visited. Furthermore, visits consisting of a single page could be discarded from the analysis, and first page distribution is exogenous to this model.

During a visit to a web page, the web user has to decide which link will be selected, or finally leave the web session. Web users are modeled as information foragers [Pir09]. Furthermore, they experience a degree of satisfaction with consuming the information included on web pages. This idea is influenced by the economic theory of utility maximization, where a web user is a consumer of information, selecting the link that most satisfies him. However, a model which only considers this dynamic factor would produce a web user that never stops navigating. For example the Random Surfer [BCR06] Model is a naive description of a web user that has no interest at all in the web page content. Furthermore, a random surfer does not rely on web page content, he/she only uniformly decides on the next link to follow, or leaves the site with probability d. Even more, this probability d seems to be a constant over all sites. If we include the exit choice, a new option corresponding to leaving the web site is incorporated. As a first approximation this option should be tested on real data.

The previous description will not be complete if decision time is not considered in the evolution of the system. The perceptual choice theory described in Chapter 4 describes the time to decide based on a stopping time for the diffusion model. That conforms to the specifications for the navigation of an artificial agent on the web, where at each page it decides which option (either click on link or leave) to choose according to the LCA model. Nevertheless LCA depends on the perception of the text on the web page, which would correspond to a higher cognitive area's probabilistic processing results. Such a kind of probability value has already been explored in economics according to the random utility model, where in this case a utility for content will be explored.

5.2 Assumptions and Approximations

The further simplified dynamics of the agent described in the previous section will approximate the observed navigational behavior of web users if some assumptions can be verified. If not, some approximation should be considered.

- **Web browsing behaviour is characterized only by jumps** : This assumption relegates the first page distribution and sessions of one page as exogenous.
- **Independence of available choices and random utility** : This assumption is related with the Logit model [McF73] . However, links could be repeated on a web page introducing a correlation. Nevertheless such a situation can be detected and considered as the same choice. If sub-groups of links are related by similarity the model is much more complicated. Those cases can also be explored using Conditional Logit models.
- **Utility depends only on text** : In a general case the selection of a link will also depend on its position on the screen [JGP+07, GFL08] or its visual presentation. Furthermore, considering a web site with low complexity and simple content, this condition could be valid regarding the site studied in this thesis. The Industrial Engineering departmental site is devoted to information about projects, courses, academic programs and staff. People that interact with this site are students, professors and secretaries, who have the purpose of searching for very specific information. Otherwise, if the web pages are complex in content disposal, the visual component of the utility should be considered. Identifying and processing the text content associated with a link could be a complex task. A good approximation, if we consider that the web site structure is consistent with the content, would be to associate the text of a link with the content of the pointed-to page.
- **Independence of the visited trail** : We suppose that web users react mechanically while visiting a web site. This is the case when they have a low frequency of visits to the web

site. Furthermore, if they only react to the best-identified link for finding very specific information, then the next step performed would depend on the actual page selected. Otherwise, the model should incorporate a memory of the visited pages in the dynamics.

- **No information satiety** : Web users are considered as information foragers with an infinite capacity, but having a fixed probability of leaving. Limits on consumption can be modeled as an accumulated utility [LRV10a]. A simple way to include information satiety is by making leaving implicit in the probability choice. Further research on this topic could be performed by considering in the utility a negative term that accumulates while the web user visits the site [LRV10a]. In this way once the user reaches a level of satiety, it will be more probable that he/she leaves the site than continues navigating. Another way to include satiety is to consider the leaving probability as dependent on the already visited texts. The more similar the already visited text is to the text preference of the web user, the higher the satiety is and also the likelihood of leaving.

- **Rational web user** : A random utility model considers that a subject attempts to maximize the benefits of visiting a web page within a degree of uncertainty. Utility is defined using natural language processing theory. Further refinement of the preference scheme can be attempted by means of the maximum entropy principle.

- **Correctness of web site information** : Of course all the web site content and structure must be consistent; otherwise even real users will not behave normally.

- **Web pages with little content** : Web users are supposed to behave more automatically. In this case Wikipedia, which is richer in textual content, is a bad example, since visitors are required to perform more complex processing within the content. The presented framework does not include higher levels of conscious processing. The LCA model considers time reaction and choices after the likelihood of each option (I_i) has already been obtained. However, connectionist models for describing such values [GG87] can be included for further research.

- **Web pages with simple content** : Text is assumed to be sufficient to analyze the perceived page semantic. If the semantic includes the visual disposition, and graphics become important, then the web user's information processing becomes more complex.

- **Web user's information processing time is negligible** : In the same way as the previous consideration, a consequence of simple content is the scant time needed for "understanding" the web page semantic.

5.3 A Model From Psychology

The LCA model reviewed in chapter 4 is used for describing the time evolution of probabilities regarding a web user's decision making. This model has the advantage of being easily extended by plugging stochastic force into the evolution equation in order to capture new behavior.

This model associates the Neural Activity Levels (NAL) of certain brain regions with a discrete set of possible choices. Those NALs (Y_i) evolve according to a stochastic equation (5.1) during the agent's decision making process until one of the NAL's values reaches a threshold equal to one (Figure 5.2). In the previous case, the agent makes the decision that corresponds to the choice associated with the maximal NAL. The stochastic equation depends on the Choice Evidence Levels (CELs). A CEL (I_i) is the neural activity level of a brain region that is associated with a unique choice, and whose value anticipates the likelihood for the choice before the decision is made.

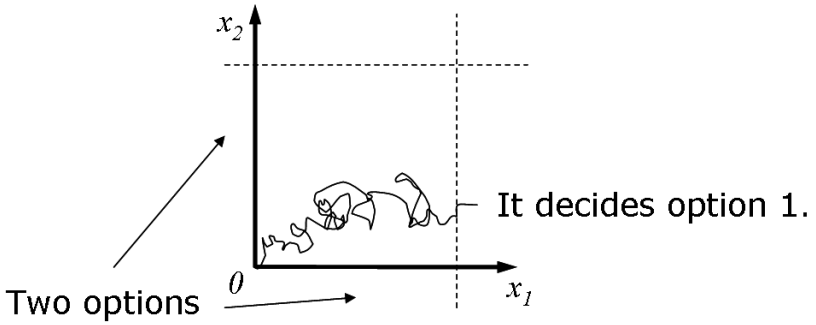

FIGURE 5.2: A diffusion-based decision making model. The first coordinate to reach the threshold corresponds to the final decision.

Models like the LCA stochastic process have a long history of research and experimental validations, most of which have been carried out in the last 40 years [Lam68, Sch01, Sto60a, Rat78]. However, few engineering applications have been proposed up to now. This work assumes that those proven theories on human behavior can be applied and adapted to describe web user behavior, producing a more effectively structured and specific machine learning model. The approach consists in applying the LCA model to predicting the web user's selection of pages (session). This proposition was based on experimental validation. A web user faces a set of discrete decisions that corresponds to the selection of a hyperlink (or leaving the site).

The LCA model is applied to simulate the artificial web user's session by estimating the user's page sequences, and furthermore by determining the time taken in selecting an action, such as leaving the site or proceeding to another web page. Experiments performed using artificial agents that behaved in this way highlighted the similarities between artificial results and a real web user mode of behavior. Furthermore, the performance of the artificial agents was reported to have statistical behavior similar to humans. If the web site semantic does not change, the set

of visitors remains the same. This principle enables the predicting of changes in the pattern of access to the web page, which in turn is related to small changes in the web site that preserve the semantic. Web user behavior could be predicted by simulation, and then services could be optimized. Other studies on ant colony models [AR03] relate directly to general-purpose clustering techniques.

The neurophysiology of decision making [UM01, BBM+06] and the random utility model of discrete choices [McF73] are considered to model the web user's behavior. In the field of mathematical psychology, the Leaky Competing Accumulator (LCA) Model describes the neurophysiology of decision making in the brain [UM01]. It corresponds to the time description of the subject neural activity of specific zones $\{i\}$ in the brain.

$$dX_i = (F_i^D + F_i^C + F_i^E)dt + \sigma dW_i \tag{5.1}$$

$$F_i^D = -\kappa X_i \tag{5.2}$$

$$F_i^C = -\lambda \sum_{j \neq i} f(X_j) \tag{5.3}$$

$$F_i^E = I_i \tag{5.4}$$

For each decision i a region in the brain is associated, which has a neuronal activity level (NAL) activity $X_i \in [0,1]$. If a region i_0 reaches an NAL value equal to one, then the subject makes the decision i_0. The NAL's X_i are time-dependent, which dynamic is stochastic as shown in the equation 5.1. Several forces (F^D, F^C, F^E) drive the system including the stochastic force σdW_i. The forces are interpreted as: F^D is the dissipative force in the case and is responsible

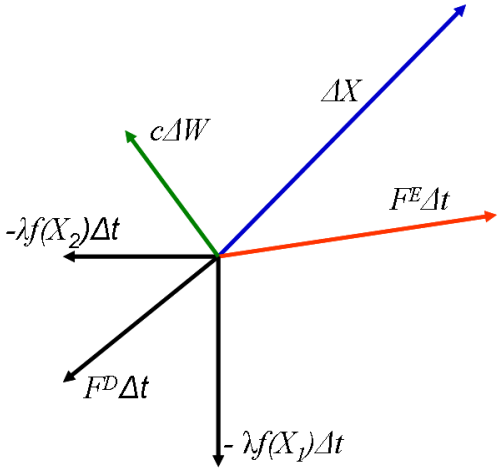

FIGURE 5.3: Forces of the LCA's system in 2-D.

for vanishing memory if no other interactions are present, F^C is the competing term related to inter-inhibition of neural connection observed in real network tissue, and F^E corresponds to the likelihood value that other cognitive processing gives to each choice.

FE is called the evidence term. In the case of decisions involving the visual discrimination of options (Figure 5.4), such values are the results of processing by the visual cortex area. Experimental evidence supports this affirmation as described in Chapter 4. The experiment on visual discrimination with random moving points explores the way in which the visual cortex furnishes the likelihood value for the direction of the movement. Those brain's pre-processing likelihoods directly reinforce the neural activity X. The parameters of the theory are interpreted

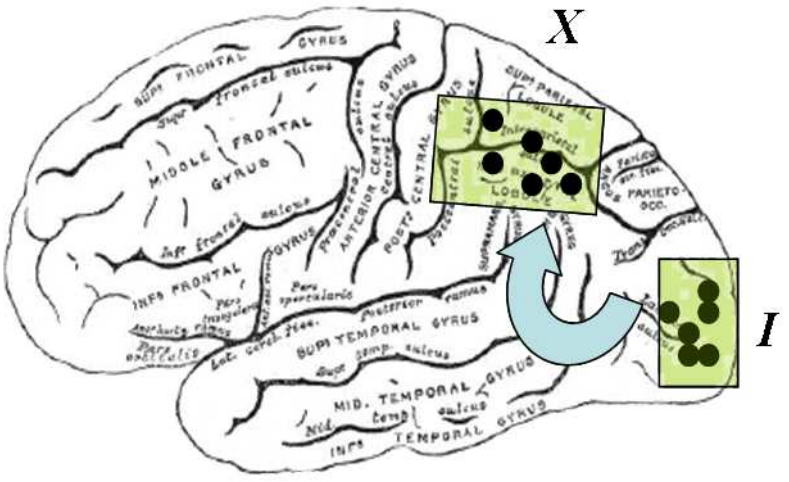

FIGURE 5.4: Evidence Forces from the Visual Cortex.

as: κ which is a dissipative coefficient, λ is related to competitive inhibition between choices, I_i is the supporting evidence of choice i and σ is the variance of the white noise term dW_i. The function $f(.)$ corresponds to the inhibition signal response from other neurons, usually modeled as a sigmoid (near to linear) or in this case linear ($f(x) = x$). The parameter I_i in the LCA theory is interpreted as likelihood values regarding the importance of choice i for the subject. Other diffusion models have been proposed, but all have been proven to be equivalent to LCA [BBM+06].

Web users are considered stochastic agents [RV09b, RV09a, RV10b, RV10c, RV10a]. Those agents follow LCA stochastic model dynamics (Equation 5.1), and maintain an internal state Y_i (NAL's values) with some white noise dW_i. The available choices, including the probability of leaving the web site, lie in the links on a web page. Agents make decisions according to their internal preferences using a utilitarian scheme as shown in the next section 5.4.

An important observation about the equation 5.1 is that it resembles Newton's equations. As a matter of fact it is called a Langevin's Equation, since it incorporates a stochastic force σdW. The variable X is measured as the rate of electric signal spike per second, since it could be considered as a kind of velocity. Furthermore $\frac{dX}{dt}$ can be understood as acceleration. Forces in this model can easily be included as an additive term Fdt opening increased possibilities of expanding the decision making model to its influence on other cognitive phenomena.

5.4 A Random Utility Model for Text Preference

The CEL's values (I) are the main forces that drive the decision system (5.1). Furthermore, we model those values as proportional to the probability $P(i)$ of the discrete choices ($I_i = \beta P(i)$), which are usually modeled using the Random Utility Model. Discrete choice preferences have been studied in economics to describe the amount of demand for discrete goods where consumers are considered rational as utility maximizers.

The utility maximization problem regarding discrete random variables results in a class of extreme probability distributions, in which the widely-used model is the Logit model (Equation 5.5) and where probabilities are adjusted using the known logistics regression [NM94]. The Logit probability distribution of a choice i anticipates every possible choice on the page j and has a consumer utility $V_j + \epsilon_i$, where ϵ_i is the random part of the utility. If ϵ_i behaves like iid Gumbel's distribution probability, then the choice probability is given by 5.5. Such an assumption comes from the fact that Gumbel's behavior is indifferent under the maximization of several Gumbel's random variables.

$$P(i) = \frac{e^{V_i}}{\sum_{j \in C} e^{V_j}} \tag{5.5}$$

The Logit model has successfully been applied to modeling a user's search for information on a hypertext system [Pir09], resulting in improved adaptive systems. The utility function should depend on the text present in links that the user then interprets and by means of which he/she makes the decision.

In order to understand text preference measures for utility construction, it is better to look at the regularities found within texts. A statistical power law fits ($P(x) \sim x^{-(1+\alpha)}$) to the word ranking x frequencies [CS03]. This surprising fact demonstrates that the most important word (best ranked x) should behave directly proportional to the frequency in log-log scale. In consequence, if we consider the document frequency K_i/M (K_i is the number of documents where the word i appears and M the total number of documents), then the value $IDF_i = -Log(K_i/M)$ should behave directly proportional to the log probability that a word i appears on a document. This finding improves the measure of the importance of a word as the TF-IDF$_{ij} = -TF_{ij} \cdot Log(K_i/M)$ [Rob04], where TF_{ij} is the frequency of appearance of the word i in document j.

This numerical vector representation of text document content is not complete without some measure of similarity between objects. Several algorithms for data mining are based on similarity measures for extracting patterns. A similarity measure is some relaxed form of distance, where "similar" objects concentrate around fixed values that take the similarity measure.

Intuition about the functional form of a similarity measure tells us that the similarity between vectors TF/IDF should be invariant under expansion ($sim(x,y) = sim(\lambda x), \lambda y)$) because of the lack of units of the component. The simplest functional form to have this property of homogeneity of degree 0 is the cosine $sim(x,y) = cos(x,y) = (x \bullet y)/(\|x\|\|y\|)$ where "$\bullet$" is the dot vector product.

Several drawbacks have been identified with these models, the most important consisting in the high dimensionality of the space. High dimensionality immediately induces the so-called "curse of dimensionality." It consists in the exponential loosening of the discriminating efficiency of distances and volumes. For instance, one problem is that traditional methods for finding the neighbor of a vector are useless in high-dimensional space [BGRS99]. The worst problem is that phenomena are exponential in the number of dimensions. Another problem is that the bag-of-words model has a typical dimensionality of thousands. This natural limitation of the method adds a large amount of noise to the data mining process, and computing time grows exponentially with the dimensionality of a large set of algorithms. Other problems are associated with synonymy and the semantic context of words creating a great deal of noise, but despite these results valuable information has still been obtained with this method [VP08].

Hence the assumption is that each agent's link preferences are defined by its TF/IDF text vector μ [MS99]. The TF/IDF weight μ_k component is interpreted as the importance for the web user of the word k. Furthermore, an agent prefers to follow similar links to its vector μ. The utility values (equation 5.6) are given by the dot product between the normalized TF/IDF vector μ and L_i that represents the TF/IDF weight text vector associated with the link i.

$$V_i(\mu) = \frac{\mu \bullet L_i}{|\mu||L_i|} \tag{5.6}$$

The resulting stochastic model (equation 5.1) is dependent on the parameters $\{\kappa, \lambda, \sigma, \beta, \mu\}$ and the set of vectors $\{L_i\}$. The first four parameters must be considered as universal constants of neurophysiology, yet the μ vector is an intrinsic characteristic of each web user. In this sense, the real web user's mode of behavior as observed on a web site corresponds to a distribution of users.

A web user is considered a stochastic agent that "surfs" the Web according to the stochastic decision rule 5.1 with preferences defined by 5.6. Parameters like $\{\kappa, \lambda, \sigma, \beta\}$ represent the physiological constants of neural tissue that every human must share. However the vector μ should be interpreted as the web user text preference at the moment of visiting the web site. We are assuming the web user does not change his/her intention during a session and leave the web site according to constant probability. The vector μ drives each web user's behavior. In this model web user profiling is in direct relation with the μ vector distribution.

5.5 Differential Equation for Probability Distributions

Numerical methods for calibrating the proposed model need to be tailored to fit the current mathematical description. Well-known statistical models are based on definite probability distributions (normal, exponential, etc.) which are dependent on the parameters. Maximum-likelihood methods maximize the probability of finding the parameters of such probability distributions in observed data. This technique is called "parametric inference". However the present model is far from having a simple or exact solution, despite its easy simulation. The problem becomes to find the probability distribution function and the parameters that describe the model. This kind of problem falls in the category of "non-parametric inference". For those purposes the mathematical behavior of probability distributions must be analyzed in detail.

The statistical description of the system is analyzed based on The Kolmogorov forward (or Fokker Plank) differential equation on the following set (equations from 5.7 to 5.11).

$$\frac{\partial \phi}{\partial t} = \sum_i \frac{\partial}{\partial X_i}[-\phi F_i + \sigma^2/2 \frac{\partial \phi}{\partial X_i}] \tag{5.7}$$

$$F_i = -\kappa X_i - \lambda \sum_{j \neq i} f(X_j) + I_i \tag{5.8}$$

$$\phi(X,t) = 0,\ \forall t > 0,\ X \in \Psi \tag{5.9}$$

$$\hat{n}(X) \bullet (\sigma^2/2 \nabla \phi + \phi F) = 0,\ \forall X \in \Delta, \hat{n}(Y) \perp \Delta, t > 0 \tag{5.10}$$

$$\phi(X,0) = \delta(X) \tag{5.11}$$

The function $\phi(X,t)$ is interpreted as a joint probability density at time $t < T$, where T is the time of the first hit to the boundary $\Psi = \bigcup_i \Psi_i$, $\Psi_i = \{X | X_i = 1\}$. The restriction 5.9 indicates considering a case when the variable X has never reached the Ψ barrier. Furthermore, $\phi(X,t)$ is the probability density of the subject to have a neural activity level of X without already having reached a decision. This absorbing barrier Ψ is compounded by perpendicular planes Ψ_i that have the coordinate $X_i = 1$, which represents the events of decision reaching i (Figure 5.5). The process initiates at $t = 0$ with a distribution concentrated on $Y = 0$ (equation 5.11) as a Dirac delta, since it is determined that the subject has no neural activity on X.

The equation 5.7 represents the dynamics of the time evolution, where F_i is a force (equation 5.8) deduced from the stochastic equation 5.1. Since the neurophysiological variables X_i are positive, in equation 5.10 a reflective boundary condition must be valid in the set $\Delta = \{X | \exists i,\ X_i = 0\}$. That considers the perpendicular component of the probability flux on this boundary is null.

The process evolves while X remains in the interior of the domain Ω that corresponds to a hypercube of side 1 (Figure 5.5).

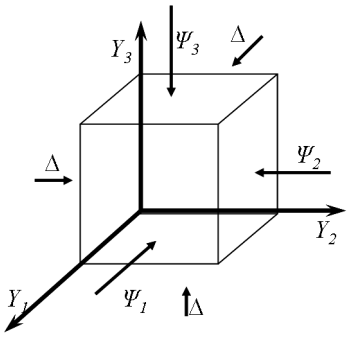

FIGURE 5.5: The Domain Ω and Boundary Topology $\partial\Omega = \Delta \bigcup \Psi$ for the Stochastic Process for three Decision.

$P_0(t)$ is the time-dependent probability of not reaching a decision. It can be expanded by the sum over all possible choices of partial probabilities, in which case $P_0(t) = \int_\Omega \phi(X,t)dX$. We can notice that initially at $t = 0$ using the condition 5.11 this probability is $P_0(t = 0) = 1$, which is interpreted as the system starting without a decision. Furthermore, in $t > 0$ the system diffuses and the probability $P_0(t)$ decreases and $\lim_{t\to\infty} P_0(t) = 0$. This property follows directly from the fact that the process has absorbent states on boundary Ψ [Res92], and the process limit is a stationary state of $\phi(X, t = \infty) = 0$.

The probability $P(t)$ of reaching a decision in time t is the complement of the previous probability $P(t) = 1 - P_0(t)$. The distribution density $p(t)$ is given by the derivative of $P(t)$ as follows in 5.12.

$$p(t) = \frac{\partial P(t)}{\partial t} = -\frac{\partial P_0(t)}{\partial t} = -\int_\Omega \frac{\partial \phi(X,t)}{\partial t}\, dX = \int_\Omega \nabla \cdot J\, dX = \oiint_{\partial\Omega} J \cdot dS \tag{5.12}$$

$$J_i = \phi F_i - \sigma^2/2 \frac{\partial \phi}{\partial X_i} \tag{5.13}$$

In equation 5.12 uniform convergence is used to distribute the time derivative and the Fokker-Planck equation is used in its continuity form $\frac{\partial \phi}{\partial t} + \nabla \cdot J = 0$ according to the flux expression 5.13. Stoke theorem is used in the integration domain Ω of figure 5.5. The Flux $J \cdot dS$ is interpreted as the probability of crossing an infinitesimal area in time t. Furthermore, the boundary of the Ω region can be decomposed on several disjoint sets according to $\partial\Omega = (\bigcup_i \Delta_i) \bigcup (\bigcup_i \Psi_i)$. The surface integral can be separated on each subset.

$$p(t) = \sum_k \oiint_{\Delta_k} J \cdot dS + \sum_k \oiint_{\Psi_k} J \cdot dS \tag{5.14}$$

However, in each plane Δ_i the orthogonal flux vanishes according to the reflective boundary condition 5.10. Only the term corresponding to the Ψ_j set from equation 5.14 remains. The probability density $p(i,t)$ of making the decision i in time t can be identified by restricting the equation 5.14. In the case that the decision is i, all terms with $k \neq i$ vanish, since no flux flows

over Ψ_k, in which case this probability is given by the total flux over the surface Ψ_i.

$$p(i,t) = \oiint_{\Psi_i} J \cdot dS = \int_0^1 \cdots \int_0^1 J_i|_{X_i=1} \prod_{k\neq i} dX_k = -\frac{\sigma^2}{2} \int_0^1 \cdots \int_0^1 \frac{\partial \phi}{\partial X_i}|_{X_i=1} \prod_{k\neq i} dX_k \tag{5.15}$$

The expression 5.15 is derived using the border condition $\phi = 0$ on Ψ. The minus sign is explained by the fact that if $\phi \geq 0$, then on Ψ the derivative are negative $\nabla\phi \cdot dS < 0$ so the term is positive. The probability $p(i,t)$ expression is used to construct the maximum likelihood of optimization problems for model calibration.

5.6 Solution of Fokker-Planck Equation

The partial differential equation system 5.7 for neural activity density probability seems at first sight to be rather complex. Nevertheless exact solutions can be explicitly found for the unconstrained case without considering reflective 5.10 and absorbing conditions 5.9. Those functions have the potential to form the basis for constructing a solution that fulfils both missing restrictions. The system of equation 5.7 is revealed to be similar to an Ornstein-Uhlenbeck process [Cha43]. Two different sets of solutions are presented.

5.6.1 Solving the Generalized Ornstein-Uhlenbeck Problem

The Ornstein-Uhlenbeck problem is a diffusion process with a drift term. It has a known solution [Cha43] that resembles a Gaussian distribution with a constant-velocity moving average and a time-dependent variance factor. The partial differential equation to solve is presented in equation 5.16.

$$\partial_t \phi(X,t) = \nabla \cdot [-(I - \omega X)\phi(X,t) + \frac{\sigma^2}{2} \nabla \phi(X,t)] \tag{5.16}$$

Where ω is a symmetric Toeplitz matrix, where diagonal elements are equal to κ and off-diagonal elements are all equal to λ. This $n \times n$ matrix can be diagonalized using a transformation R. Appendix A section A.2 shows the details about the derivation of the diagonalization problem.

$$\omega = RDR \tag{5.17}$$

$$R^{\dagger} = R = R^{-1} \tag{5.18}$$

$$D_{ij} = \begin{cases} \kappa - \lambda & \text{if } i = j < m \\ \kappa + (n-1)\lambda & \text{if } i = j = m \\ 0 & \text{if not} \end{cases} \tag{5.19}$$

$$R_{ij} = \frac{Cos(2\pi i j/n) + Sin(2\pi i j/n)}{\sqrt{n}} \tag{5.20}$$

Note the transformation R is independent of λ and κ. The transformation given by $Y = R(X - tI)$ produces the separable partial differential equation 5.21 since D is a diagonal matrix 5.19 (see Appendix A section A.3).

$$\partial_t \phi(Y,t) = \nabla \cdot (DY\phi(Y,t) + \nabla\phi(Y,t)) \tag{5.21}$$

A solution could be found by separating $\phi(Y,t) = \prod_{i=1}^{n} \phi_i(Y_i,t)$, where each ϕ_i follows the equation:

$$\partial_t f(y,t) = \gamma\partial_y(yf) + (\sigma^2/2)\partial_y^2 f \tag{5.22}$$

The coefficient $\gamma = D_{ii}$. With the following transformation (5.23 and 5.24) the equation becomes a diffusion equation (5.25), See Appendix A section A.4.

$$h(y,T) = e^{-\beta t} f(ye^{-\beta t}, t(T)) \tag{5.23}$$

$$\frac{dt}{dT} = \frac{e^{-2\beta t}}{\sigma^2/2} \Rightarrow T(t) = \frac{\sigma^2}{4\beta}(e^{2\beta t} - 1) \tag{5.24}$$

$$\Rightarrow \partial_T h = \partial_y^2 h \tag{5.25}$$

A solution for the diffusion equation is known (5.26), yet any derivative of h on y is also a solution (5.27) since this function belongs to class C^∞. This fact generates a whole set of solutions for the diffusion equation. The functions $H_k(z)$ are called the Hermite's polynomials of degree k, and form a complete base for the L^2 space (see Appendix A section A.5).

$$h(y,T) = \frac{e^{-y^2/4T}}{\sqrt{4T}} \tag{5.26}$$

$$h_k(y,T) = \partial_y^k h(y,T) = \frac{H_k(\frac{y}{\sqrt{4T}})e^{-y^2/4T}}{(4T)^{(k+1)/2}} \tag{5.27}$$

Finally, a solution for equation 5.22 is given by equation 5.28 according to reversing the previous transformation.

$$f_k(x,t) = e^{\beta t} h_k(e^{\beta t}x, \frac{\sigma^2}{4\beta}(e^{2\beta t} - 1)) \tag{5.28}$$

Then the multivariate solution to the equation 5.16 is given by 5.29.

$$\phi_{\vec{k}}(X,t) = \phi_{(k_1,\ldots,k_n)}(X,t) = \prod_{i=1}^{n} f_k([R(X - tI)]_i, t) \tag{5.29}$$

Where $[\ldots]_i$ is the i component of a vector, $\phi_{\vec{k}}(X,t)$ will be a polynomial solution of degree $d = \sum_{i=1}^{n} k_i$ on the variable X. The absorbing and reflective boundary conditions have not been considered in the derivation of this solution set, so they are not satisfied.

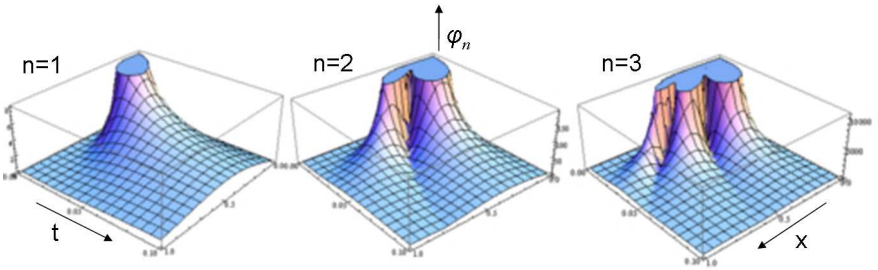

FIGURE 5.6: The solution 5.28 shape for n=1,2, and 3.

5.6.2 A Decay Time Solution

If we consider a class solution that decreases exponentially with time, then a solution of the equation 5.16 results with $\alpha > 0$. In this case we use the matrix R as in equation 5.20.

$$\phi(X,t) = e^{-\alpha t}\varphi(X) \tag{5.30}$$

$$Y = RX \tag{5.31}$$

$$\bar{I} = RI \tag{5.32}$$

$$0 = \sum_{i=1}^{n} \partial_i[(\bar{I}_i - D_{ii}Y_i)\varphi + \frac{\sigma^2}{2}\partial_i\varphi] + \alpha\varphi \tag{5.33}$$

If we assume a set of $\{\alpha_i | \alpha_i > 0\}$ where $\alpha = \sum_{i=1}^{n} \alpha_i$, in this case the equation can be separated. If $\varphi(X) = \prod_{i=1}^{n} g_i(X_i)$ then a set of equations (5.34-5.37) can be stated.

$$\forall i,\ g_i'' - 2(a_i y_i - b_i)g_i' + c_i g_i = 0 \tag{5.34}$$

$$a_i = D_{ii}/\sigma^2 \tag{5.35}$$

$$b_i = \bar{I}_i/\sigma^2 \tag{5.36}$$

$$c_i = 2(\alpha_i - D_{ii})/\sigma^2 \tag{5.37}$$

Changing variables to $z_i = \sqrt{a_i}y_i - b_i/\sqrt{a_i}$ then the equation becomes 5.38.

$$g_i'' - 2z_i g_i' + \frac{c_i}{a_i} = 0 \tag{5.38}$$

Solutions for this equation are confluent hypergeometric functions. In the case when $c_i/2a_i$ is an integer, the solution is a Hermite's polynomial $H_{\frac{c_i}{2a_i}}$. Furthermore, it is dependent on parameters $\alpha_i = D_{ii}(k_i+1)$, since $\frac{c_i}{2a_i} = k_i \in \mathbb{N}$. For an integer choice $k = [k_1, \ldots, k_n] \in \mathbb{N}^n$ the solution becomes a product of:

$$e^{-D_{ii}(k+1)t} H_k\left(\frac{\sqrt{D_{ii}}}{\sigma} Y_i - \frac{\bar{I}_i}{\sigma\sqrt{D_{ii}}}\right) \tag{5.39}$$

In this case polynomial solutions become more suitable for border conditions, since for a given degree products of the last functions could be adjusted for any t. The resulting solution is 5.40 realizing that $\alpha_k = \sum_i D_{ii}(k_i+1) = Tr(\omega) + \sum_i D_{ii}k_i = \kappa n + (\kappa-\lambda)\sum_i^{n-1} k_i + (\kappa+(n-1)\lambda)k_n = \kappa(n+\sum_i^n k_i) - \lambda(\sum_i^{n-1} k_i - (n-1)k_n)$. The degree of the resulting polynomial is $deg(\phi_k) = \sum_{i=1}^n k_i$.

$$\varphi_k(X,t) = e^{-\alpha_k t}\prod_{i=1}^{n} H_{k_i}\left(\frac{\sqrt{D_{ii}}[RX]_i}{\sigma} - \frac{[RI]_i}{\sigma\sqrt{D_{ii}}}\right) \tag{5.40}$$

5.6.3 Approximating Solution for LCA Including Border Condition

The LCA model's principal difficulty corresponds to the mixed absorbing and reflecting border condition, hence the unconstrained cases have explicit analytical solutions. The importance of having a solution for this problem lies in the calibration of the dynamic model 5.1 for adjusting to the observed behavior.

5.6.3.1 Fictitious Forces

Using physics intuition on the Langevin's equation 5.1, border conditions can be mimicked by additional forces. Indeed the reflective boundary in one dimension can be replaced by a large positive force for $X < 0$, and nearly zero for $X > 0$. In such a case a more formal representation of the reflective force is given by $F^R(X) = X^{2N}$, with a large integer N.

In the same way the absorbing boundary implies that if X reaches this state it will never return to the domain Ω. This condition is analogous to the reflective boundary and implemented by including a force that moves the system away once it reaches the border $\Psi \subset \partial\Omega$. In one dimension this absorbing force is given by $F^A = (1-X)^{2N}$ with a large integer N. In Figure 5.7 the boundary force $F^R + F^A = X^{2N} + (1-X)^{2N}$ is presented for several values of N.

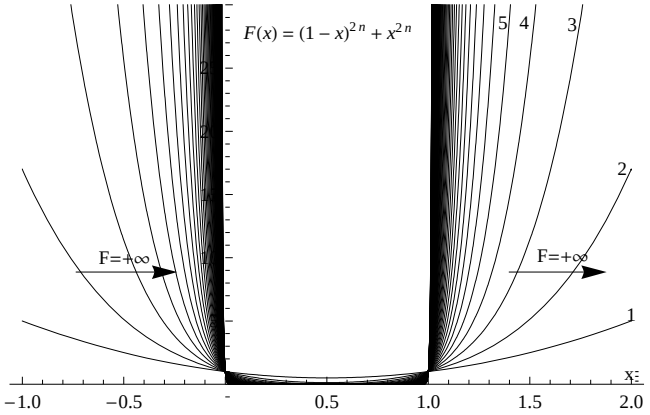

FIGURE 5.7: Boundary Forces in 1-D.

This force has a polynomial form, whose advantage is to help in calculations involving integrals of Hermite polynomials. Such a force 5.41 replaces the boundary condition as an additive term in 5.7, and on the limit when $N \to \infty$.

$$F_i^{\mathrm{B}} = X_i^{2N} + (1 - X_i)^{2N} \tag{5.41}$$

Therefore numerical methods involving solving partial differential equations require smooth initial and border conditions, and then this approximation with $N \to \infty$ introduces an easier way to solve the problem. Initial conditions 5.11 become the only restriction that needs to be imposed.

The initial condition 5.11 is a complicated restriction to be included in a numerical solver program. The Dirac Delta in a partial differential equations is used in the sense of distribution theory as a linear functional that evaluates function on one point (zero). Despite the Schwartz's distribution formalism, numerical manipulation can be done in the sense of limiting functional kernels and working in the dual space. A typical smooth-everywhere kernel is the heat kernel that limit the Dirac Delta 5.42 when $N \to \infty$.

$$\delta(x) \equiv \lim_{N \to \infty} \frac{N e^{-(Nx)^2}}{\sqrt{\pi}} \tag{5.42}$$

Furthermore, the particular LCA condition imposes that the delta be located on 0, which corresponds to the boundary of the domain Ω. Consequently, if boundaries are replaced with the above fictitious forces then this approximation remains smooth even if it is located on the border.

5.6.3.2 The Tent Approximation

Another approach for approximating the solution of the boundary problem is the Tent Approximation [RTR01]. It consists of using a linear combination of exact solutions of the unconstrained problem, and adjusting the boundary conditions as much as possible. On the border of the domain Ω a set of fixed discrete points is chosen in order to fix such a condition within them. The name "tent" comes from this punctual approximation on the border just like adjusting a canvas with pegs on its edge.

The methodology consists of choosing a finite set of exact unconstrained solutions $\{f(t)\phi_k(X)\}$. A solution is built as a weighted sum $\sum_k a_k \phi_k(X)$ of such a solution. On the boundary $\partial\Omega$ the linear restriction 5.43 and 5.44 must be valid on a fixed discrete set of points $b_i \in \Psi \mathrm{and} d_i \in \Delta_i$. The approximation will be tighter if the number of points is higher. The weights are calculated solving the resulting linear system on variables a_k and c_k.

$$\sum_k a_k \phi_k(b_i) = 0 \tag{5.43}$$

$$\sum_k c_k \nabla \phi_k(d_i) \cdot \hat{n}_j = 0 \tag{5.44}$$

The result of this homogeneous system is spanned by a finite set of solutions. The initial condition is adjusted for $t = +0$ approximating the delta function on this set.

5.7 Discussion

The model presented has the disadvantage of the complexity of the mathematical representation of the solutions. In spite of an intricate mathematical description, the model is based on first physiological principles of decision making with experimental validation. These characteristics suffice for developing the presented model on behalf of adjusting the theoretical dynamics to the observed fact.

Assumptions are crucial for the validity of this model of web usage. Visits to text-based web sites with simpler content are expected to be described by this approach. The purposes of the visit are considered fixed and no new objectives appear when visiting the web site. However, extensions are contemplated, but with more complexity included in the model. The probabilistic model does not have exact solutions.

Furthermore, analytical solutions for the free case are developed that can then be used as a functional basis for constructing the solutions of this system. Nevertheless, numerical solution suffers from the discontinuities over boundary conditions. Approximations on this scheme can be constructed on the basis of smooth functions using physical intuition about the phenomena.

Chapter 6

A Framework for Web User Simulation and Model Calibration

Computer simulation is performed by a program that implements a model of a complex phenomenon. Simulations is often used for exploring the behavior of the time evolution of a system, obtaining insight into a complex dynamic that does not have a complete analytical description. On the other hand, mathematical models are simpler abstractions of a much more complex system with the objective of recovering the dynamics of the phenomena. In spite of the desired simplicity, most models have difficult to analyze solutions that are far too complex from having a closed analytical form.

A model of dynamic phenomena can be used in two different ways for predicting future behavior. The first alternative corresponds to a mathematical method for finding an exact or approximate solutions for the evolution of the variables of the system. This provides much more predictive power is needed at the cost of increased mathematical effort. The other approach consists of the numeric integration of the descriptive equation for building the whole development of events in small incremental steps. This method is called computer simulation, and has the advantage of requiring less knowledge about the real solution of the dynamical system. However, the more ignorant one is about the model's solutions, the more computational power and resources are needed to perform the simulation.

The Web User/Web Site system is rather complex since it incorporates human behavior in often dynamically-generated pages. At first sight a complete mathematical description of the individual session evolution would seem to be impossible. Nevertheless, this thesis proposes a stochastic mathematical description of such phenomena as described in chapter 5. Averages, moments, and distributions obtained from the simulation are adequate to represent the observations. In this sense the mechanism is called "Stochastic Simulation" since some of the variables of the system are random.

On the other hand, simulation requires all of its parameters to have a definite value in order to recover predictability. The process of finding such values is called calibration, since the objective is to adjust the model to recover as much of the observed behavior as possible. In this case, observed web user's sessions are used for reaching a maximal likelihood.

6.1 Stochastic Simulation

Stochastic simulation depends on algorithms and methods of approximating statistical properties of stochastic processes by means of random sampling and integrating values. The most celebrated technique for stochastic simulation is the Monte Carlo method, from which many other algorithms derive. There are three such kinds of simulation.

The first depends on a generator of sequences of values that follows a particular distribution, also called a Random Number generator. This kind of simulation of random variables is based on a pseudo-random number algorithm for the uniform case, the accuracy of which is measured using statistical tests. The most popular method is the linear congruent generator ($x_k = (ax_{k-1} + c)(\text{mod } M)$, but it is known to fail in some cases by presenting regularities in its behavior. The second kind of simulation depends on the generation of a more complex distribution based on a Random Number generator as in stochastic process simulation. In this case, the model of the specific phenomenon is derived from the specification of random variables with given distributions. Therefore, a discrete Markov process relates to a jump probability between states, and simulation must be based on a random generator that simulates those probabilities. The Monte Carlo method is the basis for generating sequences of the desired statistical behavior.

This simulation-dependent architecture drives the design of any software for these purposes. The third kind of simulation is more complex, involving both a previous simulation and its previous application to a model of a complex phenomenon. This level relates to the concrete application, and multiple parameters are available to be tuned. Once the simulation is executing, relevant magnitudes are sampled obtaining statistical measures and distributions. This three-level description of simulation architecture helps to identify sources of errors in the final simulation results. Several statistical measures of the accuracy of the simulation are available in each level. Such checking is necessary to support the affirmation that the simulation approximates what is pretended to be simulated.

6.1.1 Monte Carlo Method

The Monte Carlo Method is a class of algorithms that, using repeated sampling of certain outcomes, constructs convergent statistics. The Law of Large Numbers is used in this case

for ensuring convergence of a resulting stadigraph. The general steps of this class of algorithm are:

1. **Basic Initialization**, in particular $k = 1$.
2. **Definition** of the input domain. In this step continuous variables are discretized and limits are fixed.
3. **Random Generation** of input in the defined domain. This step uses a random number generator from the described level-one simulation and performs operations described in the model over the defined domain.
4. **Statistics calculation**. Some of them need to be used as convergence control parameters.
5. **Contrasting a convergence criterion**. If convergence is reached, then end, if not $k := k + 1$ and return to step 3.
6. **Validation** involves error and measure contrasting.

6.1.2 Monte Carlo Quality

According to Sawilowsky [Saw03] the characteristics of a High Quality Monte Carlo simulation are:

- **The pseudo-random number generator must have certain characteristics.** Statistical test deviation from realization sequences of the expected random number is basically a standard statistical goodness-of-fit test. A traditional random number generator has a period when the random number begins to be repeated. Such a period should be as long as possible.
- **The pseudo-random number generator produces values that pass tests for randomness**. Standard tests include the χ^2-test and Kolmogorov-Smirnov statistic. For a uniform distribution test, the interval $[0, 1]$ is split in n equal sub-intervals and the stadigraph is given by $\chi^2 = \sum_{k=1}^{n} (O_k - N/n)^2 / O_k$ where O_k is the number of observations in the interval k with N essay. This test is a standard χ^2_{n-1} with $n-1$ degree of liberty. The Kolmogorov-Smirnov test has a more complicated distribution in which values are tabulated. In this case, the statistic is given by $\max_x |\hat{F}_N(x) - F(x)|$, where $\hat{F}_N$ is the empirical cumulative distribution of the n observation by interval and F is the theoretical distribution. Significance acceptance levels of a test are usually up to 5% probability to be wrong about the assumption.

- **The number of repetitions of the experiment is sufficiently large to ensure accuracy of results**. A simple random walk process has a Large Number Law convergence of $1/\sqrt{N}$. Furthermore, the average stability is used as an estimator for the convergence criteria.

- **The proper sampling technique is used and the algorithm used is valid for what is being modeled**. Sampling is the process of obtaining the right distribution stated in the mathematical statistical model. Several techniques have been developed for each special case.

- **The study simulates the phenomenon in question.** As mentioned before, the statistical model must rely on first principles in order to be adjusted to reality. Physical assumptions involve approximation schemas where other effects are neglected or incorporated as parameters of the model. Observed distributions and moments must fit the real data as measured by standard error measures.

6.2 A naive queuing network theory approach for web usage

A Discrete Event System (DES) is a popular architecture for stochastic simulation. Such software simulates a discrete state whose time transition occurs as a sequence of discrete transition epochs. The implementation is based on the equivalence of such stochastic processes with a queuing system including a variety of interconnected servers and scheduling. The classes of problem that can be simulated with those standardized software packages are generalized semi-Markov processes. The components of a DES include a clock, a heap of events, a Random Number generator, a scheduler, a number of tasks, a memory, and a statistic processor. The system does not differ too much from an operative system, the heap of events indicating the discrete set of conditions that triggers the execution of a particular task that in turn triggers other events. Task execution corresponds to finding the stochastic time for its completion, and triggers are based on random number generation. DES tasks and triggering conditions are specified by a program. The execution cycle roughly consists of recovering the next event from the heap, executing the corresponding task, updating the heap with the new events, executing the statistics processor for this step, and updating the clock for the next event processing.

The Web User/Web Site system can be modeled as a particular queuing network where service time depends directly on a web user's text preference and web page texts. Web pages are servers with infinite capacity ($G/G/\infty$) but with a definite time of information service given by the LCA model. A web user's own characteristics have a direct influence on the service time. However, since one web user does not have to wait for another web user, queues are not formed. The web user navigates through the network of servers satisfying information needs.

The web user is considered memoryless, making decisions without considering the previous pages visited, but with a purpose driven by μ. A special link corresponding to the decision of leaving the web site is presented on every page, with a fixed probability transition analogous to the random surfer teleportation operation [BCR06]. Each artificial user ends up following a trail $((p_1,t_1),...,(p_L,t_L))$ of pages $\{p_o\}$ with the visitor's time durations on the site $\{t_o\}$, until the moment the user decides to leave the L step. Arrivals to the web site are considered exogenous and other considerations stated in this chapter are also directly used in the Jackson Network Model [GH98].

However, DES packages are not the best choice for simulating a web user.

- First, efficiency is a requirement in intensive simulation for web site optimization, and multipurpose scheduling management of those packages results in a lot of overhead on computation cost.
- Second, servers (web pages) have virtually infinite attention capacity for web users. In such a case the whole package scheduling logic is useless.
- Third, the service time (time to decide) is given by the result of the LCA stochastic model. Such random variables depend on the complex stopping-time problem, which does not have an analytic expression. Most of the DES package can call on external procedures for such service time simulation, thus adding a bit of complexity to the use of those systems.
- Fourth, web structure and content have constantly changed over time. This factor introduces a degree of difficulty for implementing the simulation in DES. Despite the dynamics, the hyperlink structure is a complex and sparse graph that requires a special memory structure for efficient data management.
- Finally, specific software for web usage can easily be customized for high-performance parallel computing. As a matter of fact, since web users navigate without interaction between each other, parallel simulations can be implemented by distributing each web user calculation to a different processor.

Simulation of a web user's navigation is better implemented by means of special-purpose software. Furthermore, the sequencing of pages and the time taken for reaching a navigational decision requires special attention and are discussed in the next sections.

6.3 Stochastic Simulation of Web User

Simulation of web usage is implemented using two main levels. The first level is driven by navigational decisions on pages, recovering a simulated visit trail. The second level consists of

finding the choice of the next action and the time taken for it. Nevertheless, the complexity of the decision of which navigational operation the user will take is driven by the model in chapter 5. The overall algorithm for simulating a single web user session consists of the following steps. This process is repeated for obtaining an approximation of statistical values.

Simulation of a Single Web Navigation: The session's generation

1. **Select a random initial page** p according to the observed empirical distribution of the first session pages. Probabilities are considered by the observed rate of hit per initial page.
2. **Simulate the web user's next hyperlink selection**, obtaining the next page p' and the time used for the decision.
3. **Update the states**: Store the current page p into session and $p := p'$.
4. **If** the page p corresponds to the sink node then **the session is finished**.
5. **If** not then **return to the step 2**.

Nowadays Monte Carlo techniques are simpler and easier to implement, yet this method has slow convergence rates. Nevertheless, this procedure should be executed on several threads in order to ensure faster statistical convergence. This single agent procedure should be executed with different text vector preferences for recovering the associated different modes of behavior on the web site. The navigational dynamic of choice is proposed to be described by the LCA 5.1 stochastic process. A schema of processing the hyperlink choice is detailed in the following algorithm.

Simple Simulation of a Navigational Decision

1. **Initialize** the vector $Y = 0$ having the same dimension as links in the page p that includes a component for leaving the site option (sink node). Set the session size variable $s = 0$. Initialize the user text preference vector μ and prepare the vector L_i as the TF-IDF text vector associated with each link i on page p.
2. **Evaluate** the CEL vector I (same dimension as Y) using the similarity measure with μ and the text vector L_i associated with the link i as shown in equation 5.6. In this case we approximate L_i by the $TF-IDF$ values of the pointed-to page.
3. **Perform one-step iteration** for the evolution of the vector Y. The Euler method uses a time step h as shown in equation 6.2 and is the most-used technique for this step.
4. **If** components Y_i remain under the threshold 1 return to the step 3; if one or more components become negative, reset them to 0.

5. If not, when **the threshold is reached**, proceed to select the coordinate j' with maximum value $Y_{j'}$. The simulated choice is j' and the time taken for the decision corresponds to the sum of each increment h.

6.3.1 Simulation of a stochastic equation

The problem is to generate a sample path $\{X(t)\}_{t=0}^{\tau}$ by simulation, where $\tau = \inf t > 0 | X(t) \in \text{int}(\Omega)$ is the stopping time. The error criterion is dependent on the type of application. The motivation consists in having a simulation model that on average behaves like the observed averages. In this sense, if $\hat{X}(t)$ is the simulated variable, then the associated error is described in equation 6.1.

$$\epsilon_0 = \mathbb{E} \int_0^{\tau} |\hat{X}(t) - X(t)| dt \tag{6.1}$$

There exist several methods for the simulation of stochastic equations. All are based on small steps h of time iteration. The simulated path is the sum of small increments $X_h^n = \sum_{k=1}^{n} \Delta X_k^h$. For the case of simple Brownian motion the error of the numeric approximation is proportional to $\epsilon_0 \propto \sqrt{h}$.

- **The Euler method** is a numeric resolution of the stochastic equations [AG07] where time discretization was performed using an increment of h as shown in equation 6.2. The term $\Delta^h W_i \sim N(0,h)$ is a Gaussian noise with variance h. The term $a()$ is identified as the dt term of the LCA equation, considering function $f()$ as linear and the term $b()$ is the variance σ.

$$Y_i^h(t) = Y_i^h(t-h) + a_i(t-h, Y^h)h + b_i(t-h, Y^h)\Delta^h W_i \tag{6.2}$$

  The previous equation 6.2 corresponds to a strong Taylor [KP95] approximation on the first order. The general stochastic differential equation 6.3 can be expressed as an Ito integral equation [Oks02], where vector a and b are functions depending on X and t. Nevertheless, a more general expression can be studied by extending both functions to matrices [KP95]. The recurrence 6.2 derives from an integral stochastic equation 6.4, where integrals are approximated to the first order.

$$dX_i(t) = a_i(t, X(t))dt + b_i(t, X(t))dW_i \tag{6.3}$$

$$X_i(t) = X_i(0) + \int_0^t a_i(s, X(s))ds + \int_0^t b_i(s, X(s))dW_i \tag{6.4}$$

  On the assumption of a finite average of X, bounded L^2 distance $E(|X(0) - Y(0)^{\delta}|^2)^{1/2} \leq K\delta^{1/2}$, the function F(t,x)=(a(t,X),b(t,X)) fulfill the 1-norm Lipschitz condition, $|F| \leq K(1+|X|)$, and $|F(t,X) - F(s,X)| \leq K_2(1+|X|)|t-s|^{1/2}$ then the Euler's approximation is bounded by a constant in the average [KP95]. That correspond to the uniform error of

equation 6.5.

$$E(|X - Y^{\delta}|) \leq C\delta^{1/2} \tag{6.5}$$

A strong definition of convergence corresponds when $E|X - Y| = O(h^{\epsilon})$ and $\epsilon > 0$. This is based on the previous inequality $\epsilon = 1/2$ and considering the strong condition stated for. The boundary condition can be simulated straightforwardly but not without numerical difficulties [AGP95]. A reflective boundary $\Delta \subseteq \partial\Omega$ can be simulated considering $\tilde{Y} = \text{ArgMin}_x\{\|x - Y\|, x \in \Omega\}$ if Y is the iteration variable. In this case the order of convergence is $1/2$ even in a higher order of approximation, the reason seeming to be a misconsideration of the time remaining in the border Δ. The case of adsorbing boundaries (Ψ) can be simulated simply by stopping the process when Y crosses Ψ and can be improved [GM10].

- **The bisection method** Numerical methods for solving a differential equation rely on discretization of the continuous domain. Nevertheless, the resolution of the used grid has a direct influence on the accuracy of the approximation. The time dimension is partitioned on equal intervals of size h, then discrete variables are given by $X_n^h = X(nh)$ and the stopping time approximation is $\tau_n = h \cdot \inf\{n | X_n^h \notin \text{Int}(\Omega)\}$. Nevertheless, τ_n is an overestimation as a result of the infimum construction. A finer grid over t improves the degree of approximation with a smaller h. The problem is to which degree of approximation the grid is going to be refined. In this sense, the method of bisection [AG07] helps to evaluate the order of magnitude of the discretization scheme.

- **The Milstein Scheme** The error of the Euler approximation remains bounded if a and b are nearly constant (6.5). However, in other cases, higher-order approximation is recommended. According to a higher-order derivative using the Ito Lemma, the Euler recurrence 6.2 is corrected by an additional term in the equation 6.6. The method is based on approximating the integral $\int_0 hb(t, X(t))dW \sim b(0, X(0))W(h) + \frac{1}{2}b(0, X(0)) \sum_i \frac{\partial b(0,X(0))}{\partial X_i}(\Delta^h W^2 - h)$.

$$Y_i^h(t) = Y_i^h(t-h) + a_i(t-h, Y^h)h + b_i(t-h, Y^h)\Delta^h W_i + \frac{1}{2}(\sum_k b_k \frac{\partial b_i}{\partial X_k})(\Delta^h W_i^2 - h) \tag{6.6}$$

Under conditions similar to the Euler method, it can be proved [KP95] that this method has uniform convergence with the exponent $\delta = 1$.

- **Refining Taylor Based Iteration**: Some consideration of using a variable discretization of time seems to improve the efficiency and accuracy of the algorithm. Using the natural time scale $\frac{dT}{dx} = e^{\int^y 2a/b^2}$ the one-dimensional stochastic equation becomes a simple Brownian motion [GL97]. Nevertheless, the difficulty comes from the inversion of the function T. On the other hand, higher-order schemes use the Ito-Taylor expansion in $X(0)$.

- **Exact Simulation:** Previous approaches are general and rely on a small-step approximation based on function $a(t, X(t))$ and $b(t, X(t))$. However, if the stochastic equation is

known to have exact or approximate solution paths, then simulation could be performed in a more precise way [Gil96]. Furthermore exact simulation is defined as generating a sampling of random variables using a known probability distribution [BR05, BPR06].

$$dX_i = (I_i - \omega_i X_i)dt + \sigma dW_i \tag{6.7}$$

Fortunately the Ornstein-Uhlenbeck (OU) process stochastic equation 6.7 can be solved exactly by mean of a time transformation 6.8. This process as seen in chapter 5 is intrinsically related with the LCA model.

$$X_i(t) = X_i(0)e^{-\omega_i t} + \frac{I_i}{\omega_i}(1 - e^{-\omega_i t}) + \frac{\sigma}{\sqrt{2\omega}} W_i(e^{2\omega_i t} - 1)e^{-\omega_i t} \tag{6.8}$$

This kind of simulation is called "exact simulation", according to 6.9 where $N(0,1)$ is a normal random variable with variance 1 and mean 0. Of course averages are always an approximation arrived at by running the simulation many times , and the term "exact" refers only to the differential relationship.

$$X_i(t) = X_i(0)e^{-\omega_i t} + \frac{I_i}{\omega_i}(1 - e^{-\omega_i t}) + \sigma \sqrt{\frac{1 - e^{-2\omega_i t}}{2\omega_i}} N(0,1) \tag{6.9}$$

Boundary conditions have special considerations, since it is not known a priori at which time the trajectory will hit the borders. According to [BR05] in the case of simulation of the first boundary hitting time, the following algorithm is used.

1. Initialize a small time interval $[a,b]$, and consider a starting value $X_a \in \Omega$.
2. Perform the exact simulation on N moment $\{t_k | t_0 = a,\ t_N = b\}$ resulting in the point sequence X_{t_k}.
3. If a point X_τ hits the boundary $\partial\Omega$ (approximated hit) then return to $\tau = min\{t_k | X_{t_k}$ 'hit' $\partial\Omega\}$.
4. Otherwise, set $a := b$, $X_a := X_b$, rebuild the time interval, and go to 2.

The step 3 depends on the condition of X reaching the boundary Ω. This condition is verified if the state X enters or crosses a tolerance region defined by $B = \{x \in \Omega / \ \|x - y\| < \epsilon, y \in \partial\Omega\}$. The step 2 is restarted until no boundary is crossed and for the Ornstein-Uhlenbeck process the simulation is given by 6.9.

Reflective boundaries (Figure 6.1) are considered in the simulation by restarting on the step 2, and with departure from the nearest feasible point in Ω outside the reflective set B. In the case of an adsorbing boundary the process ends at the moment of the first entrance to the adsorbing part of the set B.

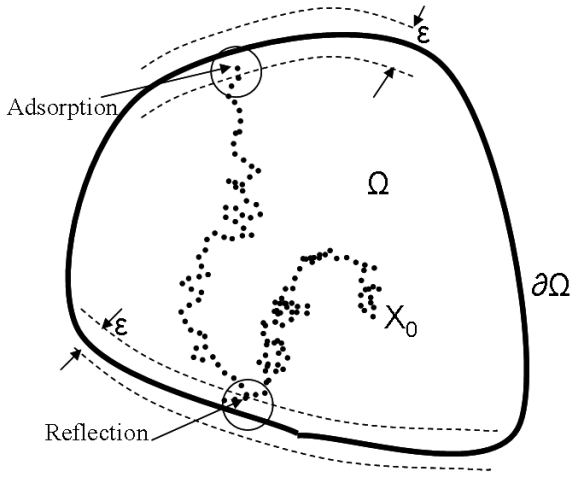

FIGURE 6.1: An exact simulation path with reflective and adsorbing boundaries.

6.3.2 Simulation of the web user decision

According to this thesis, the LCA stochastic model rules the web user's navigation decision. Such a model has the particularity to be related to the Ornstein-Uhlenbeck (OU) process that can be simulated with a high degree of accuracy using exact simulation (Equation 6.8). Using this fact, this thesis proposes to adapt LCA for exact simulation as in OU process. Furthermore, LCA differs from the OU process by a linear transformation R over X (see appendix A) and the boundary condition from section 5.1.

Hopefully the transformation R has the following property $R = R^{\dagger} = R^{-1}$ that implies that the transformed random white noise RdW continues to be a random white noise vector. The R transformation is detailed in the appendix section A.2 and can be applied over identity 6.8 for recovering the LCA dynamic equation. Furthermore, simulation for a stochastic equation 5.1 can be derived from exact simulation of the OU process including border conditions. For a point in the interior of Ω, whose time increment is sufficiently small to not reach the boundaries, then the stochastic evolution of the vector X is given by 6.10.

$$X(t) = e^{-\omega t}X(0) + M(\omega,t)I + \sigma K(w,t)Z \tag{6.10}$$

The random vector Z has each component behaving as normal variables $Z_i \sim N(0,1)$. Such a vector equation includes a matrix ω whose components are $\omega_{ij} = (\kappa-\beta)\delta_{ij} + \beta$. ω which can be diagonalized $\omega = RDR$ (see appendix A.2) using the orthogonal transformation $R_{ij} = \frac{1}{\sqrt{n}}[Cos(\frac{2\pi kj}{n}) + Sin(\frac{2\pi kj}{n})]$, where $[D]_{ij} = (\kappa-\beta)\delta_{ij}$ if $i \neq n$ and $D_{nj} = (\kappa+(n-1)\beta)\delta_{nj}$. Functions over the matrix ω are defined over each diagonal element as $F(\omega) = RF(D)R$, where $[F(D)]_{ij} = F(D_{ii})\delta_{ij}$. Furthermore, the exponential function on 6.10 is a nxn square matrix defined as before

and the rest of the terms are described by equation 6.11 and 6.12.

$$M(\omega,t) = (1-e^{-\omega t})\omega^{-1} \tag{6.11}$$

$$K(\omega,t) = [\frac{1}{2}(1-e^{-2\omega t})\omega^{-1}]^{1/2} \tag{6.12}$$

Such a term reflects consistency on $t = 0$ with equation 6.10, since $M(\omega,0) = 0$ and $K(\omega,0) = 0$. Despite the apparent complexity of matrix function, building such matrices by computer program are straightforward. Each hyperlink on a page is mapped to a coordinate in X, hence X_n is associated with the decision of terminating the session.

The vector I corresponds to the available likelihood of a navigational choice calculated by the expression 5.5 based on the text content of the pointed-to pages. Nevertheless, the component I_n relates to the navigational exit that does not involve text content. This problem is modeled considering for each text vector L_i in 5.6 with an additional component representing a fictitious term, the utility contribution of which represents the probability of leaving the web site. In this case, the text vector L_n is considered as $[L_n]_k = \delta_{mk}$ and any others consider $[L_i]_m = 0$ with m to be equal to the number of terms.

An efficient and accurate algorithm for simulating the web user is described in the following algorithm 6.1.

Step	Calculation
1.	**Initialization:** Vector X is initialized near 0 and t with a small value. Vector I is calculated according to the current page. Set the time step h and $k = 1$.
2.	**Evaluation:** The matrices $e^{-\omega(t+kh)}$, $M(\omega,t+kh)$ and $K(\omega,t+kh)$ are evaluated according to the expression 6.11 and 6.11.
3.	**Exact simulation:** $X(t+kh) = e^{-\omega(t+kh)}X(t) + M(\omega,t+kh)I + \sigma K(w,t+kh)Z$, where Z is a generated vector of normal $N(0,1)$ components.
4.	**Reflective border verification:** If $X(t+kh)$ reach or cross a small neighbor of the reflective region Δ, then set $X(t+kh) = X^*$ to be the point on the straight line on the border of the neighbor of Δ. Furthermore set $k = k+1$ and return to step 2.
5.	**Adsorbing border verification:** If $X(t+kh)$ reach or cross a small neighbor of the adsorbing region Ψ, then the decision time is $\tau = t+kh$ and the decision taken is $i^* = \text{ArgMax}\{X_i\}$. The simulation on this stage is over.
6.	**Goto next point:** Otherwise set $k = k+1$ and return to step 2.

TABLE 6.1: Algorithm for simulation of a navigational decision.

6.3.3 Mass visit simulation to a web site

We already described the algorithm for simulating a single web user session. However, real-world observation of visits to a web site shows a more detailed structure. Mass visits to a web

site turn single web user's parameters into stochastic variables.

- **First page in a session:** Arrivals to the web site are considered exogenous to the dynamic model, although arrivals are considered stochastic, which distribution should be estimated. First pages could be simulated by the Monte Carlo method based on empirical distribution. Frequencies of first arrival are recorded for simulating the first pages of the session.
- **Web user text preference vector:** Mass visits correspond to a variety of preferences. Once having estimated the multivariate distribution for vector μ, then a set of web users characterized by the μ distribution needs to be simulated for recovering observations.

6.4 Calibration of the LCA decision model

The stochastic dynamic of a web user depends on parameters that must be known before running any simulation. Parameters can be adjusted in order that the model's prediction conforms to the observed visits to the web site. Nevertheless, the LCA model for web user behavior depends on a special set of parameters. Some parameters are real values, and others are functions. Moreover, the web user text preference vector μ is considered to be described by a multivariate distribution.

Web users that visit a web site are described by a variety of objectives defined by the μ distribution. State-of-the-art web usage mining algorithm helps to find web user profiling in ways similar to such a distribution. This thesis proposes a novel web user profiling method based on adjusted users' text preference distributions and other parameters for enabling web user simulation.

6.4.1 The parameter's description

The NAL vector X evolves according to the equation 6.13 and the boundary condition. Parameters are $\{\kappa, \lambda, \sigma, \beta, \mu\}$ and correspond to physiological values related to web user neural tissue properties and information-seeking objectives (μ). All values and vector components are real and positive values. The web site structure and content contribute values for the L_i text vector and defining the dimension of the space of choices. While L_i is considered to be normalized, any normalization of μ and logit parameters is considered to be adsorbed by the last vector. However as a first stage toward parameter estimation the β coefficient on the stochastic process should be considered as the real vector parameter I, and if $\beta = 1$ then $\sum_i I_i = 1$. As mentioned before, the component n is associated with the session termination choice.

$$dX_i = (-\kappa X_i - \lambda \sum_{j \neq i} X_j + \beta \frac{e^{-\mu \cdot L_i}}{\sum_j e^{-\mu \cdot L_j}})dt + \sigma dW_i \tag{6.13}$$

If the domain $X \in \Omega$ has been set as the hypercube of side 1, then no parameter has been involved in sizing the region. Moreover the time scale parameter needs to be adjusted in order to recover the observed values. The equation 6.13 could be simplified assuming that time was scaled for making the factor $\beta = 1$ for the CEL value as defined in 5.4.

The values $\{\kappa, \lambda, \sigma, \beta\}$ should be considered the same for every visitor of the web site, since they correspond to common human body properties shared by all. Moreover, the μ or I vectors are intrinsic properties of each web user that defines his/her further explorations on the web site. According to [BUZM07] for $n = 2$ such parameters are estimated as $\sigma = 0.33$, $O(I) \sim 0 - 10$, $\kappa \sim \lambda \sim 1 - 10$, and [SN98] $\sigma^2 \sim 1.5 I_i$. For the sake of simplicity such values could be considered fixed at the moment of calibrating the model. However, they could also be included as variables in the non-linear optimization problem for maximum likelihood.

Nonetheless, observed sessions are not performed by agents with the same purposes. Recorded observations consist in a distribution of web users according to a set of intentions $\{I^u |\ u :$ web user$\}$. Such a set must be described by a multivariate statistical distribution $\rho(I)$, where I is now a random vector on $\mathbb{R}^n_+$.

The distribution ρ is a difficult parameter to adjust. A single value is an unknown point in space, but a distribution has an infinite number of unknown points to be adjusted. Such a difficulty could be simplified considering that the average $\bar{I} = \mathbb{E}(I)$ of the distribution ρ accounts for most of the cases found in reality. This approximation should be valid for instance if ρ is a sharp normal distribution around the average, which could be the case of a single-purpose web site with a small number of pages.

6.4.2 Semi-Parametric Estimation

In the last section, the difficulty of having to estimate a distribution as a parameter was described. Using a kind of average $\bar{I}$ as a partial solution could be a way for finding the calibration of the model. Maximum likelihood is a well-known technique for stochastic model calibration. The probability of observing the available data (likelihood) is maximized using as variables the unknown parameters, subject to the restrictions of the theoretical model. The solution is interpreted to be optimal in the sense of being the most probable according to the observation.

The calculation for obtaining a kind of average $\bar{I}$ vector is based on observed data from a real web site. For a given possible transition from the page i to j, a number n_{ijk} of observed clicks measure the time spent on the website t_{ijk}. In this context the log-likelihood is given by the equation 6.14, where $p(i, j, t|I)$ is the probability transition in equation 5.15 (expanded in 6.15) for a given value of I. According to section 5.5 the value of 6.15 is positive since $\phi \geq 0$ on Ω and $\phi = 0$ on Ψ, then the derivatives are negative. The constraint stated in 6.16 considers $\beta = 1$

and I to be a probability.

$$\max_{I,\kappa,\lambda,\sigma,\phi} S = \sum_{ijk} n_{ijk} Log(p(i,j,t_{ijk}|I)) \tag{6.14}$$

$$p(i,j,t|I) = -\frac{\sigma^2}{2} \int_0^1 \cdots \int_0^1 \frac{\partial \phi}{\partial X_j}|_{X_j=1} \prod_{k \neq j} dX_k \tag{6.15}$$

$$\sum_k I_k = 1,\ I_k > 0,\ \sigma > 0,\ \kappa > 0,\ \lambda > 0 \tag{6.16}$$

$$\phi(X \in \Psi, t) = 0 \tag{6.17}$$

$$\hat{n} \cdot ((I - \omega X)\phi - \sigma^2/2 \nabla \phi)|_{X \in \Delta} = 0 \tag{6.18}$$

$$\phi(X,0) = \delta(X) \tag{6.19}$$

The maximization problem is restricted by the set of the equation from 5.7 to 5.15 using a double index (i,j) instead of a single one, representing a transition from page i to j. Ψ_{ij} which corresponds to the hyperplane $\{X \in \partial[0,1]^n \mid X_j = 1\}$ and ϕ density is the solution to the Fokker-Plank equation A.25 with constraints.

Furthermore, the tent approximation (section 5.6.3.2) could be used considering $\phi = \phi(X,t;I)$ as a linear combination of the exact solution of the Fokker-Planck equation from 5.40 $\phi \sim \sum_{d=0}^{D} \sum_{\sum_i k_i = d} a_k \phi_k$ fulfilling the border condition 6.17, 6.19 6.18 on a finite number of points (equation 6.20). In this case, the weights a_k of such linear combinations are incorporated as variables to the system including equations related to boundary condition. If the weight a_k^* is an optimal solution, then the distribution $\sum_k a_k^* \phi_k$ approximates ϕ in the sense of being the most likely function of degree D.

$$\phi(X,t) \sim \sum_{d=0}^{D} \sum_{[\sum_i k_i = d]} a_k e^{-\alpha_k t} \prod_{j=1}^{n} H_{k_j}\left(\frac{\sqrt{D_{jj}}[RX]_j}{\sigma} - \frac{[RI]_j}{\sigma\sqrt{D_{jj}}}\right) \tag{6.20}$$

The previous expression relies on a combinatorial partition of the integer d represented by the integer vector $k = [k_1, \ldots, k_n]$. Therefore, probability can be calculated by replacing such an expression in 6.15 resulting in 6.21.

$$p(i,j,t|I) \sim -\frac{\sigma^2}{2} \sum_{d=0}^{D} \sum_{[\sum_i k_i = d]} a_k e^{-\alpha_k t} \frac{\partial}{\partial X_j} \left[S_{k_l}(j, X_j, D, I, \sigma) \right]_{X_j=1} \tag{6.21}$$

$$S_{k_l}(j, X_j, D, I, \sigma) = \int_0^1 \cdots \int_0^1 \left(\prod_{l=1}^{n} H_{k_l}\left(\frac{\sqrt{D_{ll}}[RX]_l}{\sigma} - \frac{[RI]_l}{\sigma\sqrt{D_{ll}}}\right) \right) \prod_{k \neq j} dX_k \tag{6.22}$$

Despite the presence of a polynomial on the integral 6.22 it is difficult to obtain an explicit expression for $S_{k_l}(j, X_j, D, I, \sigma)$. It turns out much more unmanageable to calculate this integral by using a partition on the set Ψ. Indeed if each dimension is partitioned in 100 parts and there

are 21 links on the page i then the number of points turns out to be an astronomical 100^{20}. Fortunately, a much simpler algorithm is proposed for calculating such an integral.

Symbolic integration can manage very efficiently the computation of $S_{k_l}(j, X_j, D, I, \sigma)$. A first observation is that the multivariate integrated function is a polynomial on variables $\{X, I, a\}$ and a rational function on $\{D, \sigma\}$. Insomuch, integrals over variable X are straightforwardly calculated and evaluated by symbolic integration.

This observation is important since it drastically reduces the computational complexity of the inference algorithm. Furthermore, derivatives of $S_{k_l}(j, X_j, D, I, \sigma)$ on I, σ, κ, and λ can be directly extracted after symbolic processing. In this way, traditional non-linear optimization methods can be used on this system using the resulting evaluation of the function S.

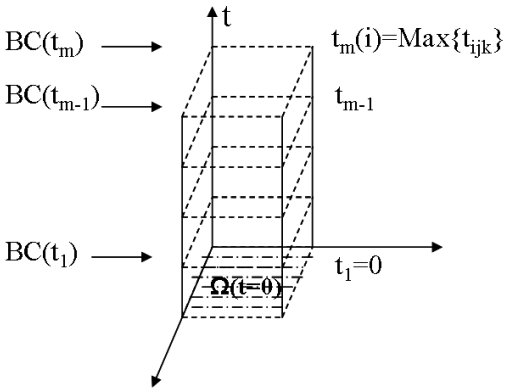

FIGURE 6.2: The border condition over time evolution is a cylinder.

Border condition 6.17 and 6.18 must be accomplished by the function ϕ for all $t \geq 0$. This implies in the tent framework an approximation of a cylinder-like manifold (Figure 6.2) for generating point restriction. The time dimension is sliced in the m point where the last corresponds to the maximum time spent on the current web page (t_m).

For each time-slice, border conditions are checked generating a set of linear restrictions for variables a_k. For an adsorption boundary condition, $q1$ points X_l^p are chosen on Ψ generating a total of $q1 \cdot m$ linear restrictions on a_k considering all time slices. Those restrictions 6.23 replace the restriction 6.17 in this approximation. Reflection restrictions 6.24 are linear on variable a_k and are $n \cdot q0 \cdot m$, where $q0$ is the number of points per face Δ_i.

$$0 = \sum_{d=0}^{D} \sum_{[\sum_i k_i = d]} a_k e^{-\alpha_k t_l} \phi_k(X_l^p),\ X_l^p \in \Psi,\ p = 1, \ldots, q1 \tag{6.23}$$

$$0 = \sum_{d=0}^{D} \sum_{[\sum_i k_i = d]} a_k e^{-\alpha_k t_l} \left((I_i - [\omega X]_i)\phi_k - \sigma^2/2 \frac{\partial \phi_k}{\partial X_i} \right)_{X_l^p},\ X_l^p \in \Delta_i,\ p = 1, \ldots, q0 \tag{6.24}$$

The initial condition 6.19 states that all the probability density is concentrated on $X = 0$ just like a Dirac's delta function. Nevertheless, the implementation of such a condition needs be approximated. As a matter of fact, no Dirac's delta is observed in nature, instead very intense but finite and localized signals are noticed. A positive function on Ω can be envisaged at $t = \epsilon$ that fulfills border conditions, being concentrated near 0.

However, the previous approach seems feasible if the dimension n is sufficiently small for using a number of points in Ω for covering the variance of the initial condition and border condition. Furthermore, giving a particular set of points that covers the variance of an initial comes near to solving the equation on a particular time.

Another approach empowered by symbolic processing uses fictitious forces 5.41 for emulating border conditions. Such forces are polynomial functions on variable X when the propagator operator 6.25 (see A.4) can be approximated to any order in t symbolically. Such an operator performs the evolution of a system that fulfills an equation $\frac{\partial \phi}{\partial t} = L\phi$ according to the equation 6.26.

$$U(t) = e^{tL} \tag{6.25}$$

$$U(t')\ \phi(X,t) = \phi(X,t+t') \tag{6.26}$$

Furthermore, the thin-step function 6.27 satisfies the border condition 6.17, 6.18 and equation A.25 since it is zero anywhere in Ω but outside $[\epsilon/2, 3\epsilon/2]$.

$$\phi(X, t=\epsilon) = \begin{cases} \epsilon^{-n} & X \in [\epsilon/2, 3\epsilon/2]^n \\ 0 & \sim \end{cases} \tag{6.27}$$

Nonetheless, the previous discontinuous function is not suitable for propagation according to this method. First, it must be approximated by a polynomial series as in 6.20 resulting in an accumulated solution near 0 fulfilling the LCA equation. Hermite polynomials are a basis of orthogonal functions as noted in A.90. Function 6.27 is approximated by a linear combination of ϕ_k, thus border conditions are increased as needed as the number of functions are higher.

Approximating boundary conditions by using the propagator operator generates a symbolic solution for a time t from polynomial approximation 6.27. The likelihood problem does not depend on the a_k parameter being simpler to implement, since it does not depend on a discretization scheme over a high-dimensional space. This is a very important observation. The number of variables $|a_k|$ is nearly exponential since it depends on the number of partition of d (degree of the polynomial).

For example, if the number of variable X is 20 and the maximal degree of the polynomial is 8 then the number of variable $|\{a_k\}|$ is $3,108,105$. This fact introduce a limit into the degree of the polynomial versus the available computational capabilities. Nevertheless, once initial

solution will consist in a polynomial objective function of degree equal to d and with a number of variable $|\{I_i\}|$.

This approximation has the cost of not accomplishing exactly the border condition. The part corresponding to the reflective border will contribute as an additional probability mass to the total distribution. However, the part corresponding to the absorbing border will quickly vanished since it is on the border. This suggest to reinforce the approximation at the reflecting boundary.

6.4.3 Non-parametric estimation: The maximum likelihood variational problem

There are two approaches for the estimation of the μ vector parameter. A simpler approach consists in calculating the I that maximizes the likelihood function, providing a kind of average tendency of web users for text preferences. On the other hand, if one considers I not a single value but a distribution, this introduces a radical change regarding the mathematical problem. In this case probability decomposes as $p(i,t) = \sum_I P(i,t|I)P(I)$ where $P(I)$ is the unknown distribution of users with I text preference given by equation 1.3. The first problem is related to the multivariate optimization, the second concerns the infinite dimensional problem.

$$\max_{P(I)} S = \sum_{ijk} n_{ijk} Log(\sum_I p(i,j,t_{ijk}|I)P(I)) \tag{6.28}$$

For the continuous case, an approximation related to series expansion is used. Based on recent advances in non-parametric inference [DW00], a discrete version in Fourier coefficient of the equation set is formulated as a restriction of a maximal likelihood problem for the distribution of the observed number n_{ijk} of transitions k measuring the time t_{ijk} for each choice $i \to j$. This approach takes into consideration a discretization on the time variable. Real data is recorded in an integer number of seconds according to the discretization scheme. Information constraints are included either as a unimodal condition, monotonic conditions, smoothness, heavy tail, available moment of time decision and the Fourier series version of the differential system (see [DW00]).

$$\phi(Y,t) = \sum_r a_r \varphi_r(t,X) \tag{6.29}$$

$$\varphi_k(t,Y) \in Fourier \tag{6.30}$$

$$P(I) = \sum_{i,s} b_{is} \varphi_s(I) \tag{6.31}$$

Hence, the resulting non-linear optimization problem could be solved numerically on variables $\{a_r, b_{is}\}$ for approximating the multivariate function $P(I)$. Nevertheless, a high-dimensional grid involves an unmanageable number of variables.

The following consideration results in considerable improvement of the non-parametric algorithm performance.

- **Symbolic processing:** If the system manipulates expression symbolically, the whole process of optimizing likelihood is benefitted. No discretization of Ω is needed, since the values of functions are calculated by delayed evaluation. The cost is ensuring that expression must fulfill the condition of the problem and depends on expression representation in memory for performance of the manipulation. In the process of finding optimal likelihood values for parameters, derivation and integration operations are performed. Symbolic representation should be simple enough to have explicit derivatives and integrals.

- **Propagator operator:** Having the ϕ distribution depending on X on a fixed time t_0 it is possible to know the distribution on a time $t > t_0$ by using the propagator operator. This propagator must ensure the border condition. But according to its exponential character, only an approximated version can be implemented $U(t) = 1 + tL + t^2L^2 + ...$ The operator L is the infinitesimal generator of the transformation U and for unconstrained diffusion it corresponds to the Ornstein-Uhlenbeck differential operator. If ϕ is implemented by symbolic processing, then the propagator must be adequately applied to symbolic expression.

- **Polynomial orthogonal set of solutions:** Polynomials are objects that can be better implemented on a computer. Integration and differentiation are easily implemented over symbolic polynomial objects.

- **Clustering of session:** Finding the distribution of vector I is a difficult task. First of all, such a distribution does not relate with text choice since it reflects the distribution of different kinds of subjects classified by their text preference. However, two assumptions involve further simplification of the inference process. Similar sessions should group similar web user text preferences and distribute like-multivariate normal distributions within each cluster. Therefore, each cluster ζ has a function $P_\zeta(I)$ represented by equation 6.32. Parameters of the model correspond to the average vector per cluster $\bar{I}_\zeta$ and the variance matrix Σ_ζ.

$$P_\zeta(I) = \frac{1}{\sqrt{2\pi|\Sigma_\zeta|}} e^{-\frac{1}{2}(I-\bar{I}_\zeta)^\dagger \Sigma^{-}1(I-\bar{I}_\zeta)} \tag{6.32}$$

  Probability distribution is given by the sum of partial probabilities 6.33.

$$P(I) = \sum_\zeta P_\zeta \tag{6.33}$$

  Cluster set $\{\zeta\}$ can only be obtained by hierarchical clustering techniques, given the discrete character of the similarity between trails. The reason is that there is no way to define a middle point between two sessions. Similarity relates with the size $|\mathrm{MCS}(s_1, s_2)|$ of the maximal common subsequence [SMBN03] in the sense that sessions with a maximal

degree of common path are considered. Finally the similarity function between s_1 and s_2 is defined by $2|\mathrm{MCS}(s_1, s_2)|/(|s_1| + |s_2|)$. Transitions are then segregated by the identified clusters, and values for the restricted set $\bar{I}_\zeta, \Sigma_\zeta$ are found by the optimization procedure.

However, semi-parametric inference can still be used considering only one cluster. In such cases the procedure is much simpler but the mass of web users seems to be better adjusted by equation 6.33.

6.5 Computer implementation

The more highly elaborated the mathematical description of the numerical algorithm is, the more precise the implementation will be. The system is mainly composed of three stages: model calibration, simulation platform and application. Simulation platforms iterate Monte Carlo generation of web user trails on the basis of a parameter set. Calibration is an optimization-based system for adjusting such parameters. Simulation is further applied in web site optimization as a way to perform prediction on web usage when the site changes.

Computer implementation is the final stage after mathematical analysis and deals with its own set of problems. Limited memory and computation power are some of the reasons to search for a fair approximation, and the solution is called engineering. This thesis proposes to make intensive use of symbolic processing to avoid the common problem of discretization, and insofar as possible, the analytic solution. A simulation process is proposed to use an exact scheme for generating a path. A calibration process is proposed to use symbolic processing of the exact solution of the unconstrained LCA problem.

6.6 Web site optimization algorithm

Web site quality can be measured by several indicator measures based on the content and structure. An optimal web site should score higher on such measures as well as on the direct opinion of web users. A commonly-used indicator corresponds to the time spent by a web user on the web site, as a larger value is interpreted as better. Therefore, by using a simulation-mechanism change, the web site could be tested in order to improve such indicators.

1. Configure a set of stochastic agents $\{A_I\}$ behaving with the discovered navigational dynamics and according to the observed usage distribution of web user text preferences.

2. Select a set $\{R_a\}_{a=1,\ldots,N}$ of limited-variation structures of the web site R that preserves graph connectivity. The selection could be random considering a maximum number of new links and deletions.

3. For each $\{R_a\}$ the set $\{A_I\}$ is simulated, obtaining τ_a
4. The structure a^* with maximal τ_{a^*} is selected as the new web site structure R.
5. If the variation of the average time τ_{a^*} with respect to the previous one is negligible, then stop the algorithm, otherwise return to 2.

The agent simulation platform works on the basis of a given distribution of web user text preference vectors. Once this distribution is obtained from web usage data, the simulation platform then generates the different frequency distribution usages (session, time, etc.) On the other hand, a common measure of the usage of the web site lies in the average time spent per session τ. This measures the number of clients viewing the web site.

Web structure is improved by means of maximizing such usage restricted by graph connectivity (without considering hyperlink direction). Simulation with artificial agents results in artificial usage path data for calculating the usage measure. The iteration consists of the following simple algorithm.

6.7 Discussion

The complexity of simulating the trail of a web user comes from implementing the individual decision process. However, straightforward exact simulation can be implemented by using Ito integration. Such a simulation reproduces the navigational path based not only on text and hyperlink structure, but also on web user text preferences and the subject's neurophysiologic properties. Such values assigned to each particular web user need to be known in order to perform simulation.

Calibration is the mechanism that finds such parameters. Nevertheless, the traditional mechanism of calibration fails in the sense that the problem has an intrinsically high dimensionality. The proposal is to use symbolic processing to overcome this issue. Once the simulation machine is calibrated, experimentation upon changing web site configuration could be performed since the model does not depend on the web site but the web user.

Chapter 7

Experimental results and setup

This chapter relates with the engineering of implementations and experimentation of models. Pre-processing and setting is definitely the largest part of the experimental work. In chapter 3 several techniques have been exposed for pre-processing web data. This thesis is related to building an artificial system that behaves like a real web user. A mathematical framework was built to support theoretically the agent configuration for simulation. However, the simulation system needs to be calibrated using historical data in order to adjust internal parameters. Data required consist in web user trails in a web site, the text seen in that occasion since web sites change with time, and the hyperlink structure present at each visit.

Today web site content and structure are dynamic. This observation represents a challenge since a historic web site record needs to store the whole site over time to capture changes. Specific design and assumptions must be taken into account for extraction and storage. A main concern relates with the volume of data. Considering that a small web site contains of the order of thousands of pages and millions of characters from text, indexing and compressed representation becomes an important point for retrieval.

Web user sessions are not explicit in web data. Trails are not recorded until an explicit mechanism allows precisely recording them or inference is used. For this thesis an accurate integer programming method was developed in (chapter 3) the case where web log are the only source of past web user visits. Nevertheless, having the possibility to use cookie based session retrieval, then a much more precise set of session can be extracted. The number of accumulated clicks in medium sized web site per month are typically 10^6 in order of magnitude. The implementation must ensure that both the recorded structure and the set of session be consistent; in the sense that pages exist in both databases and hyperlink being present at the correct time.

As mentioned in chapter 6, calculation involving density for LCA model should use symbolic expression object in order to avoid problems with higher dimensionality. Nevertheless, there

exists several systems implementing such capabilities. A requirement for choosing a symbolic manipulator is the fact that mathematical expression be first class object of the system in order to facilitate the implementation of the resolution likelihood problem. In this sense, Wolfram's Mathematica™[Mae00] software manages such object with optimization procedures simultaneously.

Calibration and simulation are processes that need to be verified. The design is based on estimation of underlying probabilities distribution. A number of measures are presented for testing model accuracy. Simulation should also serve to generate measures for testing calibration.

This chapter describes the whole process from pre-processing to application of the simulation platform.

7.1 The big picture

Remembering the hypothesis: "**It is possible to apply neurophysiology's decision making theories to explain web user navigational behavior using web data.**" In this context (Figure 7.1), the experimental part is threefold. Stochastic equations description of decision making is proposed to allows the simulation of web user visit to a web site. However, parameters need to be obtained by calibration methods. The last operation depends on data quality. The proposal was experimented on the university's departamental web site "http://www.dii.uchile.cl", whose description is exposed on next section.

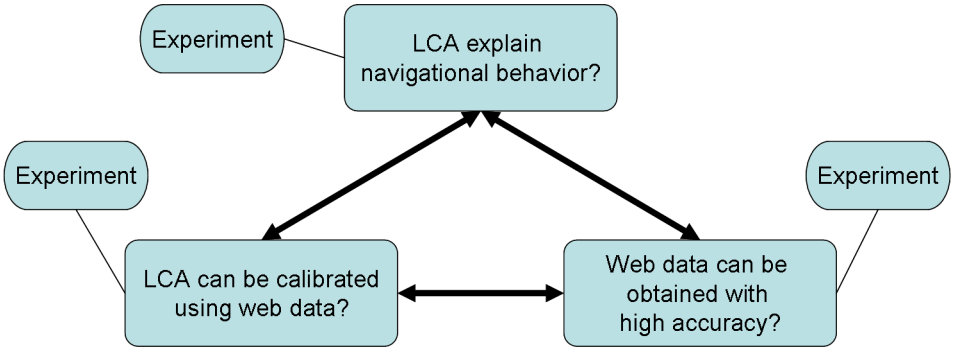

FIGURE 7.1: Experiment related with the hypothesis.

1. **Web data retrieval accuracy**: In the case where only web-log are available novel method involving integer programing has been proposed and tested has been revised on chapter 3. Fortunately, it was granted a faculty permission to include a cookie tracking device (see section B.4) in a subset of the web site. Furthermore, web content retrieval was implemented as part of a crawler system. Data quality is analyzed by looking at the

resulting distribution. This step pretend to demonstrate the feasibility of obtaining quality web data.

2. **Calibration of the stochastic model**: Using the web data a calibration process is performed for finding optimal parameters according to the likelihood.

3. **Simulation of the calibrated model**: Using a discretization of the stochastic equation simulation of the model are executed as testing the model as well the calibration.

Both three step are finally tested at the simulation, where observed session distributions should be recovered. This is presented as experimental validation of the proposed hypothesis.

7.2 Web Site Description

In this research we used five sub-sites belonging to the Industrial Engineering Department. The main departmental site, three sub-sites from master degree program, and a project web site have nearly thousand web pages. Each one has its own characteristics in terms of content and structure and there is no homogeneity in relation to the process of web construction. Only one uses a content management system which allows standardizing the adding of content, but in the others the insertion is manual.

The main topics addressed on these web sites include:

- General information about the Industrial Engineering Department
- Faculty staff
- Description of the undergraduate and post graduate programs
- News and Information about upcoming events and conferences

This web site has the particularity of having a lower degree of complexity and changes on the web site are minimal comparing with others. Those characteristic make sessions on simpler and ideal for the study of web usage mining. Real session are retrieved and stored in the format of web log in order to test sessionization.

7.2.1 The shape of sessions

We find that only a few IP addresses account for the vast majority of all clean registers. Over 98 percent have less than 50 register for the entire month. Figure 7.2 displays the number of

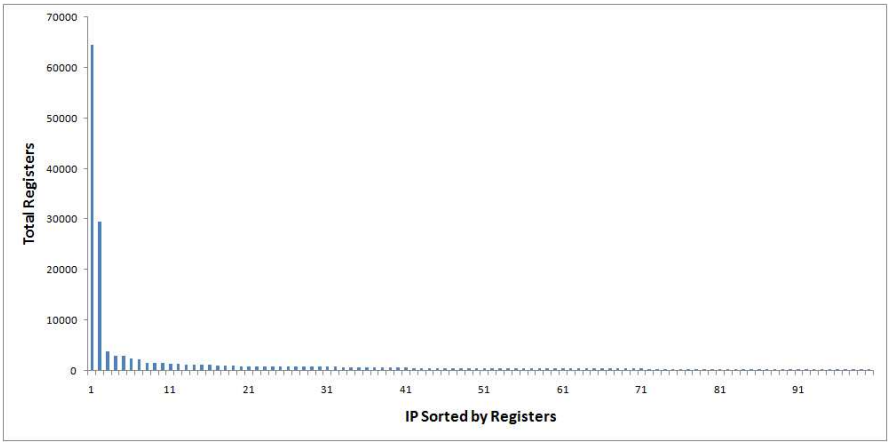

FIGURE 7.2: The number of registers for the 100 IP addresses that account for the most registers.

registers for the 100 IP addresses that account for the most registers. We also found how many unique web pages are visited by each IP address; we find that IP addresses that visit many unique web pages tend to have more diverse sessions. Figure 7.3 shows the number of unique pages requested by the 2,000 IP addresses that account for the greatest number of unique page requests. Of the IP addresses not shown, almost 84 percent visit three or less different pages for the entire month. We store the information in a relational database (MySQL) that includes

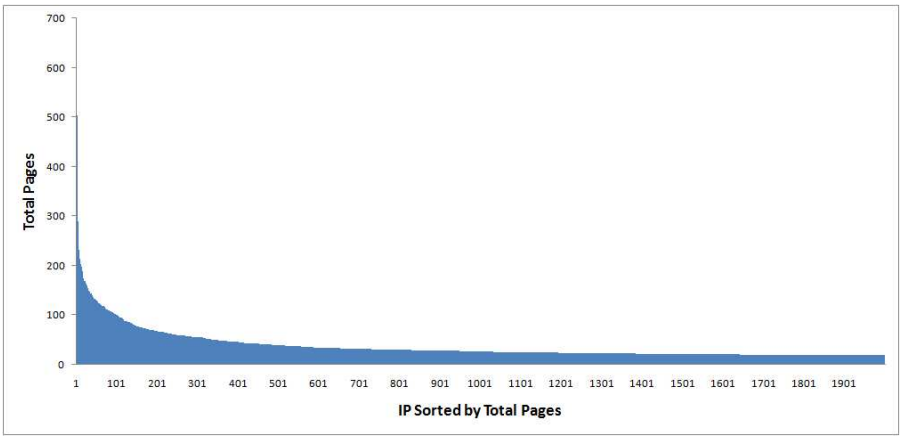

FIGURE 7.3: The 2,000 IP addresses that account for the greatest number of unique page requests.

tables for unique IP addresses, unique page identifiers, and unique links between pages and the registers. The database maintains relational constraints between tables in order to ensure data consistency.

7.2.2 The hyperlink structure

The number of hyperlink change over month. The average number of hyperlink on the web site is 4058 with a average change per month of -109 (2.7%) that means the number of hyperlink

was decreasing. The number of pages also change, with an average per month of 691 different pages with a change about of −15 pages per month (2.2%). Those pages correspond to the sub-site where the agreement allows to investigate web user session. The figure 7.4 shows the

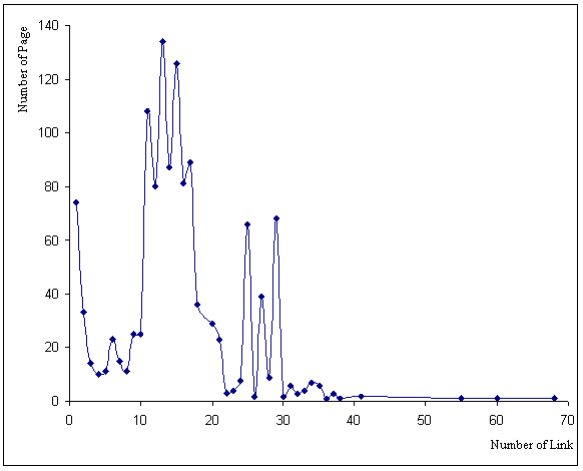

FIGURE 7.4: The distribution of out-link per page.

typical distribution of links in a month per page. A characteristic number of hyperlink per page is 20. This fact is important since is define the dimension of the neural activity of a subject in this page.

7.2.3 The content

The web site contains nearly 17,000 stemmed terms. Furthermore, the content is much more dynamics that the structure. Changes in content reach in average near 347,000 changes in term frequency per month. This correspond to the updates that are performed on daily news over all the site. Each page has in average 193 different term.

7.3 Pre-processing

Data retrieval and preparation deserve a especial treatment.

7.3.1 Web content and structure

It was decided to use the Websphinx java library for implementing a crawler. This system periodically (everyday 6P.M. and 12P.M.) inspect recursively all pages from the web site, extracting hyperlink to other pages and the text content. Text was filtered eliminating stop word and realizing an stemming process as seen in chapter 3. As a result of the processing of one

page the following data is obtained:

- Unix timestamp: unique time value of the retrieval of the next object.
- A URL and page Title: The url serve as unique identifier of the web page.
- A set of link: Each hyperlink will be recursively visited later. But now are added as hyperlink.
- A set of stemmed word: It is a list of each term found on the page with its corresponding number of appearance.

All those data are stored in a relational database that have the following structure in figure 7.5. This database works with the following idea, there exist permanent pieces of information that

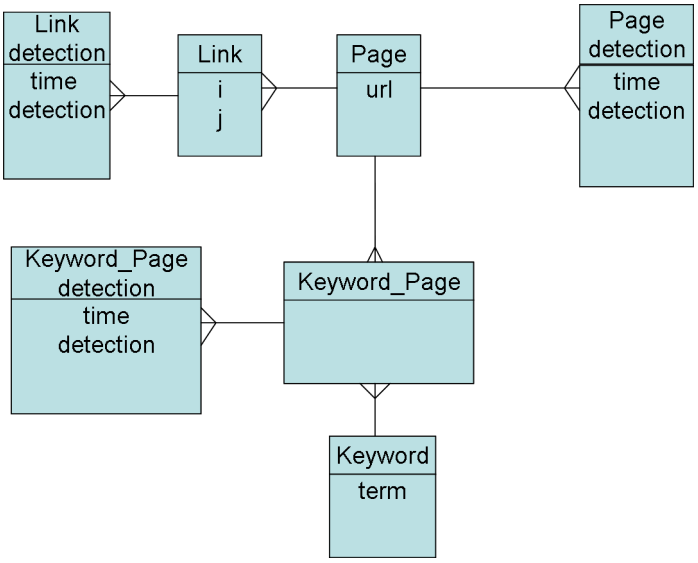

FIGURE 7.5: The database for content and structure storage.

are stored only once. Such data are keyword that at some moment will reach more than 20,000 term but after that the growing of the table will be negligible. The same phenomena occurs with page represented a it URL and hyperlink. On the other hand, other attributes like number of appearances and time validity change periodically. Such dynamic properties are stored on tables with the postfix "detection". In this way it was possible to store the whole history of changes in the web site.

7.3.2 Web usage data: sessions

A cookie sessionization method was used to extract sessions for measuring the performance of the integer programming methods. The web site was adapted to track web user navigational actions in order to recover sessions. The method used was based on cookies, identifying individual

session but without storing personal information. A JavaScript program was included on each web page for tracking navigational actions using a unique anonymous identifier for each user.

The data recollection process started in June of 2009 and finished in August 2010. Web user sessions were obtained through a cookie-based sessionization method, which store the anonymous user identifier. This technique uses both client side and server side elements to track web user behavior and reconstruct automatically the sessions.

First, a script file is embedded in every web page. This insertion is generated through an automatic process which picks every web page from the server and writes within the <head> HTML tags the corresponding path to the JavaScript file.

When a user begins to navigate the web site, this script sets a unique identifier based on a randomly generated number and uses a cookie for storing the navigation details. This cookie is updated along the session path and the corresponding collected data is retrieved by the JavaScript file and sent to the web server for final storage using a through a PHP file.

Collected data was stored using the following structure in table 7.1.

Name	Type	Description
id_log_session	int	primary key (autoincrement)
id_session	varchar	randomly generated value that represent the session identifier
time	int	UNIX timestamp
ip	varchar	web user IP address
ip	varchar	web user IP address
host	varchar	web host of the requested web page
uri	varchar	web uri of the requested web page
event	varchar	identification of entering a page (IN) or leaving it (OUT)
query	varchar	query parameters passed to the url
agent	varchar	web user UserAgent

TABLE 7.1: Data retrieved from JavaScript event and cookies.

Within the period of inspection, $1,224,812$ rows were inserted in the table, containing 382.047 sessions, $121,968$ different IP addresses and $1,227$ different web pages.

A cleaning process was performed to ensure quality and validity of data. This process included the deletion of Uri's referencing non web page documents (images or files) and fames. It also eliminated the records from robot crawling activity.

The cookie-based sessionization method has remarkable advantages regarding the automatic identification of web user sessions and the possibility to avoid the usual processing of web logs, which commonly represents a huge task in terms of both time and resource usage.

The drawbacks of this method are related with its dependency from the "onload" and "onbeforeunload" JavaScript functions used for identifying when a user enters or leaves a page. These

functions are handled and executed in varied manners by different web browsers, which in some cases generates wrong insertions or missing values in the column "event" in Table 7.1.

Also, as the cookie based method registers both entering a leaving of a page, in an ideal case two records are inserted in the table for a particular web page access. This overstocking of data is not useful for the next steps of the research.

To fix these problems, the table was reprocessed using a heuristic based algorithm which analyze the "event" field and taking into account all the possible combinations of the IN/OUT values, allowed to perform the following:

- Assign only one record for a particular web page access
- Estimate the visit time of a web page in a determined session (see Table 2)
- Produce a data transformation from string to integer values to adapt the data for subsequent analysis.
- Ensure a correct visit order for every component of particular web user session.

Page 1		Page 2		Page 3		
(E)IN	(F)OUT	(A)IN	(B)OUT	(C)IN	(D)OUT	TIME Page 2
1	1	1	1	1	1	B-A
1	1	1	0	1	1	C-A
0	1	1	1	0	1	B-A
1	0	1	1	1	0	B-A
1	1	0	1	1	1	B-F
0	0	1	1	0	0	(D-A)/2
1	0	1	0	1	0	C-A
0	1	0	1	0	1	B-F
1	0	0	1	1	0	(B-E)/2
0	0	1	0	0	0	Indeterminate
0	0	0	1	0	0	Indeterminate

TABLE 7.2: Simple heuristic for visit time estimation with cookie (1: the event is registered, 0: if not).

After the process, a clean data set was obtained with a total of 708.007 rows, containing 360.748 sessions, 114.041 different IP addresses and 1.192 web pages.

Although there was a significant reduction in the total record number (42.1%), it did not affect substantially the number of key components of the table 7.2:

- 5.6% of reduction in terms of web user session
- 6.4% of reduction in terms of unique IP addresses

- 2.8% of reduction in terms of web pages

The shape of the resulting session is presented on figure 7.6. The number of visitor (Session) appears to be nearly the half of the number of registers. Variation in the number of visits to pages at the web site are on average 12% of the total per month. Seasonal effect are clearly represented on the figure 7.6; since during the summer in January and February the academic activities are minimal. Nevertheless, visits to the web site during June to November are almost constant with small variations of 5%. This data set including session identification is used for testing the accuracy of the proposed sessionization method. The distribution of session size appears to have

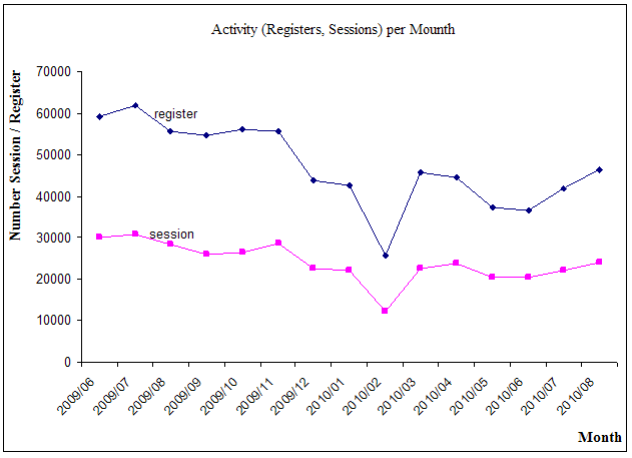

FIGURE 7.6: Seasonal activity of the web site in term of number of visitor (session) and registers (June 2009 - August 2010).

a good linear approximation in Log-Log scale. More precisely the distribution has a better good piecewise linear as shown in figure 7.7. The stability of the log-linearity is illustrated on figure

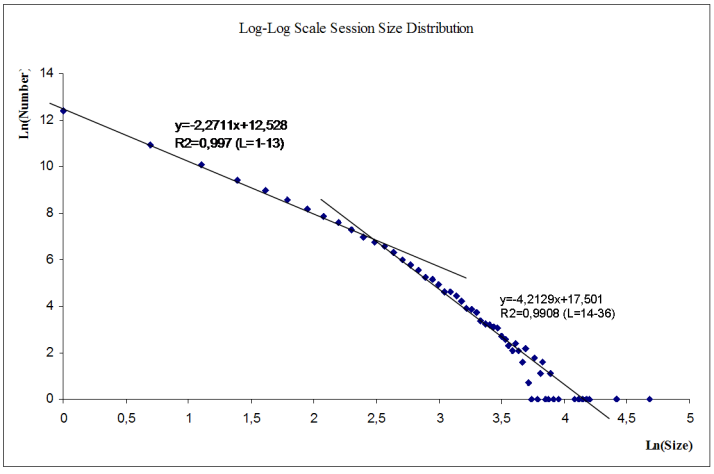

FIGURE 7.7: Piecewise linear distribution approximation for session size in Log-Log scale.

7.8 reveals to be a property from sessions. The figure shows slope and constant variation over the 15 month, showing nearly constant behavior. Slope correspond to the power of the Zipf law,

resulting in an average of -2.8 with a standard variation of 0.1. The constant results in a value of 10.6 with a standard variation of 0.4. This empirically observed power law distribution was

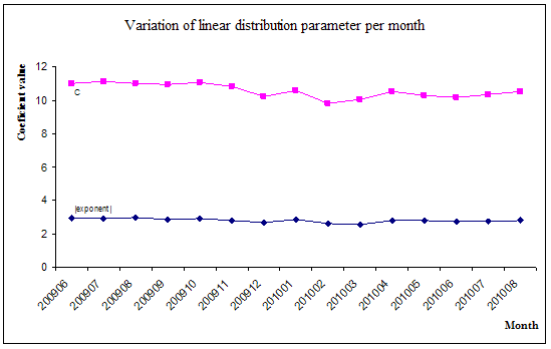

FIGURE 7.8: Log-Linear regression coefficient value variation over month.

known as the "Web Surfer Law" [HPPL98, VOD+06] and with a correlation coefficient of $R^2 = 0.98$. In the case when information about real session is not available, then this property is the only reference to the quality of the sessionization process [DRV08a]. Moreover, the adjustment to power law is clearly approximate, but its stability suggest that fitting quality measure reveals quality of sessions.

7.3.3 Homologation of databases

Two different sources were used for obtaining the web site's data and the usage path. However discrepancy are always presents according to the different nature of the sources. Both databases need to be linked by the unique identification of web pages. Nevertheless, it was observed pages that only belong to one of the sets. The main reasons were: isolated sub-sites that are not reachable from the main page then the crawler will never find such object.

A process of directed crawling and URL standardization produce finally a database with consistent cookie session and web site's data.

7.4 Calibration of parameters

As was mentioned, parameter of the decision model separated on two: The evidence vector (I) and neural tissue constant ($\kappa, \lambda, \beta, \sigma$). An assumption was performed considering simulation for a fixed evidence vector, built on base of a μ preference text vector that resumes the web site most important word. This was achieved cutting the 10% of the higher TF-IDF term and the lower 5% values in the whole site. Then the stochastic model results simpler with only four scalar parameter.

The calibration was resolved using a Monte Carlo simulation iteration where changes on parameter where tested and best matching with time distribution of session where used. The following result were obtained.

1. λ : 0.4
2. κ : 0.2
3. σ : 0.03

The distribution obtained match the asymptote linear parameter (Figure 7.9) with a 10% of error.

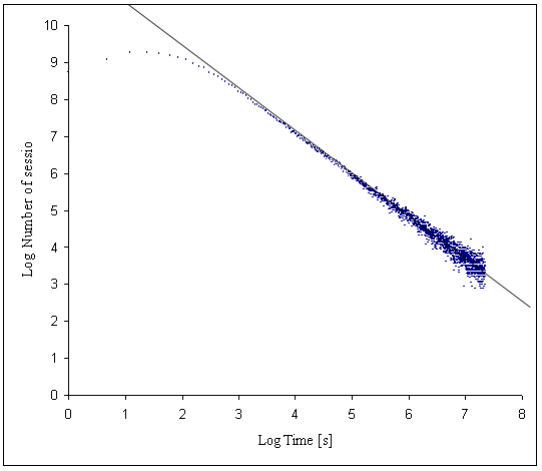

FIGURE 7.9: The distribution of time duration of a session in Log-Log Scale.

Those parameter where then fixed for performing the calibration of the evidence vector. The symbolic based algorithm were stated and a single vector were fit. An interesting fact is that is nearly similar to the vector obtained by the most important word in the web site. A 40% of error is identified in the session distribution simulation. The process take nearly 10 hour.

A distribution of vector were obtained clustering the session by using the longest common subsequence distance for clustering. In this the clustering process were stopped when 10 cluster were detected. The same than before process of calibration were performed using this method on each cluster and subset of visited pages. The results were astonishing, nearly $8,3\%$ of error in the simulated session distribution were obtained.

7.5 Experimenting Simulation

As we have seen session length follows a typical distribution [HPPL98]. The simulated "average web user" follows a distribution of session length similar to the empirical. The relative error is

of 8,1%, less than 1% of error in log scale. Distribution error remains more or less constant for session with a capacity of less than 15 pages consisting of 0,3% of error (Figure 7.10). It is not surprising that the leaving-site probability chosen was equal to the sum that was empirically calculated.

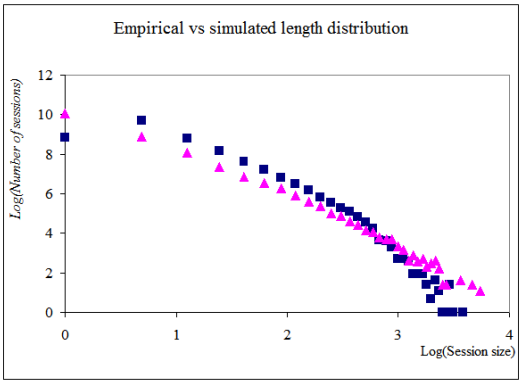

FIGURE 7.10: The distribution of session length Empirical (Squares) vs. Simulated (Triangles) in Log scale.

The common random surfer was simulated and the session length distribution was recorded. Nevertheless it was observed a larger error on the distribution of the session length, because the session stopped at much shorter number of steps (5 on average). This simple model does not reflect the user's time loosing much of the behavioral information from sessions.

However several others parameters variations were carried out that did not succumb to such degrees of adjustment. Therefore this is a relevant result. Figure 7.11 compares the visited pages frequencies of the simulated behavior with the experimental behavior, with nearly 5% of error and a 50% of variance. The time behavior of the model is revealed to have the same power law shape than with real session (Figure 7.12).

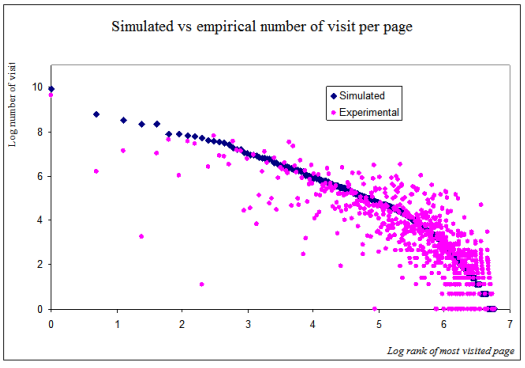

FIGURE 7.11: The distribution of visit per page on simulated vs. experimental session.

The regression results obtained from the distribution of the μ text preference vector show that most probably (maximal $P(u)$) have the following word highly ranked. 3 vectors were selected on the upper 70% of probability. A sample set of the words obtained by this method is presented:

1. **Management, Engineering, Enterprise, Society**. Interpretation: related to industrial engineering field.
2. **Mgpp, Council, Bank, Description. Interpretation**: Word related to a degree master in public politics (Mgpp).
3. **Capital, Market, Sustainability, Characterization**. Interpretation: Word related to economics.

Those results show great accuracy in describing the interest of real visitor to the web site of the Industrial Engineering Department of the University of Chile.

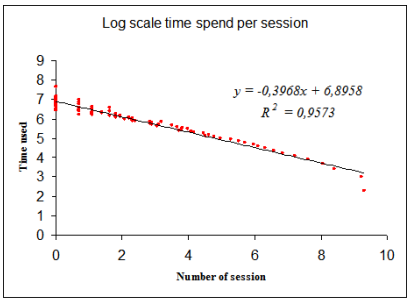

FIGURE 7.12: The ranked time spends on a session in log scale.

It was known that traditional machine learning result in 50% of effectiveness for rebuilding the distribution of session [VP07]. The efectiveness of this method is nearly 80%.

The most important testing of the model is related to future behavior on based on past training. The effectiveness of this method on next 2 month after calibration were about of 73%. The value is specifically high but it is lower that when testing on the calibration set. However, this number still overcome traditional web mining algorithm.

7.6 Discussion

40,000 web users were simulated using the Euler discrete version of the equation 6.2 [AG07]. Both the web site structure, text content on the academic at a departmental web site was used (http://www.dii.uchile.cl). The distribution of session length has a relative error of 8.1%. If the distribution of the session size is studied, a power law of exponent -2.6 adjusts with notable

precision and with a quadratic error of 0.4. This adjustment has been observed in real web user sessions [HPPL98] and demonstrates the quality of these artificial sessions.

Nearly a 70% of the real distribution is recovered by this method. This is a important advancement since it prove to overcome traditional web mining algorithm. Then the hypothesis is proved to be plausible.

To sum up, this method is a plausible way to simulate navigational behavior of web users that are in accordance with statistical behavior of real sessions. Nevertheless, this work describes the theory and methodology for setting a web user simulation platform. Further work should relate to improving the discrete choice utility to quadratic term and expanding this methodology to other fields such as Market Research.

Chapter 8

Conclusion and New Perspectives

Web usage mining has been for ten years the cornerstone of adapting web sites to the necessity of users. In spite of a vast literature on this subject Netflix's prize has demonstrated that traditional web usage mining techniques have limited impact on real world issues. On the other hand, applied behavioral sciences describes first principle decision making models that are suitable for being applied to Web usage.

This thesis proposed to study and apply Psychology based theory of decision making to web usage in order to describe a user's sessions. The mechanism used for such purposes was the neurophysiology LCA (equation 1.1) model of decision making. It is adapted to predict the next page in a session obtaining the sequence of visited pages. This corresponds to a stochastic simulation scheme that is handled by Monte Carlo techniques. Once the model presented on the last section is calibrated, it is possible to obtain the distribution navigation trails using Monte Carlo techniques. As a sub product of the calibration mechanism it results the dispersion of web user's keyword interest. With both tools it is possible to build an automatic mechanism for giving the best recommendation to web users in order to enhance the web site experience.

The general objective of this thesis is developing a stochastic models of web user's behavior, for analyzing variation of navigational preferences after changes in web site content and structure. However, the presented model is not simple and presents a challenge by itself. New techniques for solving such model and data pre-processing algorithm are presented. The simulation topic is relevant for experimentation on web site configuration. Mathematical analysis of the system is performed based on exact solution to the unconstrained problem. Finally new perspectives are identified since human behavior model on the web could be applied to other human activities.

Web data processing problem has a set of stages, which have particular considerations. The first and most time consuming is data pre-processing. The extracted primary web data consist of the hyperlink structure, the content, and the web site usage. Each data kind has it own

issues. Web data pre-processing is needed to calibrate the model, and has its particular issues. Web user past session cannot be obtained directly, this must be inferred. Common techniques for session inferences are demonstrated to have nearly a 60% precision, while the developed integer programming method is about 80%. Since calibration methods require precise data, the proposed integer programming method give feasibility to continue applying the simulation framework, even in the case when only web logs are available.

The integer model for pre-processing is based on bipartite cardinality matching (BCM) and integer programming. The method suggests further application to explore the likelihood of specific sessions and theirs characteristics. As a sub-product of this work, it is possible to find the maximum number of copies of a given session, the maximum number of sessions of a given size, and maximum number of sessions with a given web page requested in a given position for each session. In this sense, optimization method for sessionization is a fruitful framework for the analysis of web log. Another important finding is that the BCM reaches nearly the same performance as the integer program, and BCM problem can be solved in polynomial time.

Another problem covered is the dynamic nature of web sites. Pages and hyperlink are updated periodically, and then what a web user confronts one day is different on another. The proposed model describes navigational decision making according to the exposed content and structure of a web site. A web crawler collects the hyperlink structure and content. Page selection methods, revisit schedules, a politeness policy, and parallel processing are component of such system. In order to approximate dynamics of the web site, changes are recorded on a daily base by stemmed terms, pages, and hyperlinks. Term repository quickly reaches a stable (slow increasing) number of words, pages, and hyperlink repositories also present the same behavior. But relationship repository between them grows proportionally to the changes performed on the web site, since such storage is indexed by retrieval time. Such a data warehouse allows performing any processing on web data and ensures consistency for further processing.

There exist several models in Neuro-Computing that can be applied to web user. Most of them describes stochastic aspects of the decision making process. Time dependent stochastic process of decision making where explored in Psychology using **first principles** constituting the field of Neuro-Computing. Furthermore, models like Weiner Process, DFT, QDT, and LCA successfully explain many aspects of individuals decision making. Such theories were consistently tested on laboratory by mean of experiments based on neuronal activity level. Therefore, LCA and DFT theories dominate the field. Both describe the stochastic process of decision making and are proven to be equivalent in many aspects. However, in this thesis the model chosen corresponds to LCA because of its uses definite physical variables like neural activity, thus other incorporate artificial quantities (DFT). The theory appears to be easily modified to include other effects.

The connection with text content is made on the basis of the bag of word model, the logit discrete choice approach, and the LCA's vector I. Such vector is interpreted as a value computed by

higher conscious level of processing likelihood about decisions, which each component correspond to a preconceived probability of which choice is better. This probability is approximated by the logit model of discrete choice using a utility function per web page. A web user will prefer a web page which text context is more similar to its own text preference. In this sense web user's navigational behavior is categorized by a text preference vector, and the perceived utility is proposed to be the cosine similarity measure with the page's bag of word text vector. This proposal completes the definition of web user in the LCA framework.

A major disadvantage of the model is its mathematical difficulty to have approximate solutions. The problem originates on the naturally high number of dimension for the differential problem (typically 20) and non-standard border conditions. Differential problems are commonly solved by a discrete mesh on the domain. Therefore, with such a number of dimensions, a mesh can be easily compounded of $100^{20} = 10^{40}$ cells which is computationally intractable. In spite of the intricate mathematical description, the model is based on physiology principles of decision making validated experimentally. This characteristic suffices for developing the presented model on behalf of adjusting the theoretical dynamics to the observed fact.

An approach based on symbolic processing and exact polynomial solution of the unconstrained problem was proposed to avoid a dimensional explosion. The propagator operator $U(t') = e^{t'L}$ transforms solutions on time t of the LCA equation with border condition, to another time $t + t'$. This operator was constructed in term of the LCA differential operator L and can be applied to an initial condition in order to recover the dynamics in any t. Using exact polynomial solution of the unconstrained problem to approximate a delta on $t = \epsilon$ and ensuring nearly the border condition, could be propagated obtaining a solution in any t. Since solution functions are based on Hermite's polynomials, all derivation and integration could be performed exactly and symbolically. In this sense, the high dimensionality of the problem has no further influence on the symbolic problem.

A maximum likelihood problem for calibrating model's parameters is envisaged. However, distribution functions are managed symbolically in the optimization process to avoid dimensionality problem. Again, polynomial functions are an advantage for the needed operation at the moment of finding the optimum since derivatives can be evaluated exactly. A I vector could be inferred using this technique, representing an average likelihood for text preference. However, web user should represent a distribution of different preferences. In this case the inference expands to a distribution of such vector $P(I)$. Clustering methods helps in this direction predefining a compound distribution of preference based on multivariate normal distributions. Such approximation, results in discrete and limited number of variables.

Once parameters are fitted, simulation of web users can be performed. Simulations are based on exact solution for the Ornstein-Uhlenbeck stochastic equation instead of using an approximated method based on stochastic Taylor theorem. This generates more precise path and faster

convergence for distributions. The resulting system is expected to work in a slightly modified version of the web site predicting distribution of visits to the web site. This assumption is based on the theory that is independent of the visited web site. The only possible change not cover is changes on the text preference distribution that is supposed to be exogenous. If the changes on the web site are smaller and do not drastically change the overall semantic then the simulation is expected to predict changes in navigation.

Since the simulation successfully predicts navigation changes then changes on a web site can be investigated for optimizing measures of web site usability. Such a possibility drastically changes the concept of an adaptive web site since it is possible to predict the impact of a change based on historical visits. Before this advancement, suggestions were performed and validated only after introducing them on the web site using trial and error. Now, the navigational improvement is based on an optimization method where adjustment can now be predicted. This result has the importance to leverage adaptive web site processing into a more scientific field.

8.1 Further extension

This thesis presents a set of assumptions that are used for building the theory, and on this basis extension are envisaged as arguments of plausibility. Improvement on retrieval techniques could be explored at the light of the proposed model of web browsing. Web user operations and perception of the browser are parts of what can be improved in the model. Further mathematical extensions of the stochastic model can be developed. A more efficient implementation schemes could be explored. The construction of a free platform for web experimentation could be designed and implemented for web research.

It has been shown that log analysis for session retrieval was based on natural properties of sessions, using integer programming. An extension could be performed including probabilities for sessions in the case of conflict disentangling simultaneous sessions present in a log sequence. It is plausible, that the presented stochastic model for sessions, could be adapted to perform more exact session retrieval by using maximum likelihood. In this case, the optimization model becomes non-linear, but iterative simulations over sessions could be performed to extract most probable sessions.

Navigational operations of web user are much more complicated than simple hyperlink clicking, the back and forward browser's button correspond to some of the set of possible navigational events as well. Other navigational operations are: using bookmarks, tab or other window browsing. All those operations modify the presented model choice for web user. However, all browsers are implemented with the same set of events and could be considered as universal. Those events can be abstracted and included into web site structure representation, in the same

way that session termination event was included. Probability for such navigator events could be represented as logit probabilities, modulated by empirical probabilities of using those features instead of link clicking. These probabilities can also be calibrated from historical data, using the presented framework.

Web page semantic has not been considered in the model, since the bag of word model dismisses those relations. Visual disposition and presentation of hyperlink on a web page, could also influence the decision about the next page to visit. This phenomena should be included in the I vector, since in this case, relates with brain's visual processing and semantic resolution. In the proposed framework a utility function resume preferences. Text semantic and/or visual disposition, should be represented by a numeric vector, so it enters on the utility definition.

LCA model could be extended to incorporate more neural structure as found in DFT model. Neural weighting matrices represent inhibition/reinforcement of neural connection with the matrix ω adopting a Toeplitz form. Those matrices classify interaction intensity connection by diagonal band. Furthermore, this form is suitable for automatic diagonalization. A richer neural connections scheme should also be explored since other psychological phenomenas are expected to be described by such models [Gro00].

The algorithm for sessionization, calibration, and simulation is suitable to be implemented efficiently. Sessionization could use a specific network flow algorithm for solving the BCM, instead of using a commercial solver. Calibration is based on symbolic computation of polynomials, which could be implemented by special trees of monomials. Hermite polynomials are implemented by a variety of libraries and polynomials can always be reduced to standard forms. Finally, simulations could be implemented easily using parallel threads in order to have faster convergence. With such an implementation, a complete system of web research could be envisaged controlling the whole process of data extraction, pre-processing, training and experimental simulation.

Appendix A

On the LCA partial differential equation.

The LCA model relates on the specification of the stochastic evolution of neural activation values, the constrains of such variables, and the process for reaching a decision (Stopping time or timeout). The stochastic process is giving in general by equation A.1 for n decision. The vector $X = [X_i]_{i=1}^n$ represents neural activity level on the brain device for decision making (e.g. Lateral Intraparietal Cortex). The vector $I(t) = [I_i]_{i=1}^n$ represents input of evidence level for alternatives from brain areas like Middle Parietal cortex. The whole system is driven by stochastic forces given by the vector $W = [W_i]_{i=1}^n$ which component are independent standard Gaussian white noise.

$$dX_i = [I_i(t) - \kappa_i X_i - \beta_i \sum_{j \neq i} f_i(X(t))]dt + \sigma_i dW_i \tag{A.1}$$

$$\kappa_i \geq 0 \tag{A.2}$$

$$\beta_i \geq 0 \tag{A.3}$$

$$\sigma \geq 0 \tag{A.4}$$

$$f_i(X) \in C^2[0,1] \tag{A.5}$$

$$X_i \geq 0 \tag{A.6}$$

$$X_i(0) = 0 \tag{A.7}$$

Variables are constrained to be positive since are considered biological values (equation A.6). The system starts at the origin (equation A.7) until the subject reaches a decision. Two paradigms of reaching a decision have been established: by reaching a neural activity threshold and by timeout. The first paradigm can be stated mathematically as a stopping time problem. The time τ of first coordinate reaching the threshold 1 (equation A.8) corresponds to the time of reaching

a decision i^* (equation A.9).

$$\tau = \text{Min}_t\{t|\exists i,\ X_i(t) > 1\} \tag{A.8}$$

$$i^* = \text{ArgMin}_i\{t|X_i(t) > 1\} \tag{A.9}$$

In the timeout paradigm, the decision time τ is given and decision correspond to i^* of equation A.10.

$$i^* = \text{ArgMax}_i\{X_i(\tau) > 1\} \tag{A.10}$$

A.1 The forward Kolmogorov (Fokker-Plank) equation derivation

The last formulation of LCA model is important in simulation, but in order to explore it is better to use the join probability $\phi(X,t)$. That probability evolves according to a partial differential equation called Kolmogorov's Forward equation [Res92]. It is presented a demonstration of the Fokker-Plank equation based on [Cha43]. It corresponds to the older physics approach, but with great intuitive significance.

Considering discrete step of time Δt and space $\Delta X = [\Delta X_i]$. The white noise component of the stochastic equation A.1 describes the **density transition probability** $P(X \rightarrow X + \Delta X) = P(\sigma \Delta W) = G(X, \Delta X)$ that occurs in a variation ΔX given X.

$$\Delta X_i - F_i \Delta t = \sigma \Delta W_i \tag{A.11}$$

$$\text{Where,} \quad F_i = I_i - \kappa_i X_i - \beta_i \sum_{j \neq i} f_i(X) \tag{A.12}$$

$$\text{Then,} \quad G(X, \Delta X) = P(\sigma \Delta W) = \frac{1}{(2\pi\sigma^2 \Delta t)^{n/2}} e^{-\frac{\sum_i (\sigma \Delta W_i)^2}{2\sigma^2 \Delta t}} = \frac{1}{(2\pi\sigma^2 \Delta t)^{n/2}} e^{-\frac{|\Delta X - F \Delta t|^2}{2\sigma^2 \Delta t}} \tag{A.13}$$

$$\text{Since,} \quad \Delta W = B\Delta t \tag{A.14}$$

Where B is a vector of independent random variable with normal distribution $N[0,1]$ [Res92]. The Kolmogorov equation relates to use total probability decomposition (A.15) over all transition based on ΔX in order to calculate the probability in a small time step Δt in the future. For

this reason, it is named *Forward*.

$$P(X,t+\Delta t) = \int_{R^n} P(X-\Delta X,t)P(X-\Delta X \to X)d\Delta X = \int_{R^n} P(X-\Delta X,t)G(X-\Delta X,\Delta X) \quad \text{(A.15)}$$

$$P(X-\Delta X,t) = P(X,t) - \sum_i \frac{\partial P(X,t)}{\partial X_i}\Delta X_i + 1/2 \sum_{ij} \frac{\partial^2 P(X,t)}{\partial X_i \partial X_j}\Delta X_i \Delta X_j + O(\Delta X^3) \quad \text{(A.16)}$$

$$P(X,t+\Delta t) = P(X,t) + \frac{\partial P}{\partial t}\Delta t + O(\Delta t^2) \quad \text{(A.17)}$$

$$G(X-\Delta X,\Delta X) = G(X,\Delta X) - \sum_i \frac{\partial G}{\partial X_i}\Delta X_i + 1/2 \sum_{ij} \frac{\partial^2 G}{\partial X_i \partial X_j}\Delta X_i \Delta X_j + O(\Delta X^3) \quad \text{(A.18)}$$

Then, replacing A.16, A.17, A.18, and A.13 into A.15, result in equating coefficient of the same order in Δt. First, integral in ΔX can be calculated exactly as Gaussian integrals and evaluate as moments of such random variable.

$$< \Delta X >= \int_{R^n} \Delta X \; P(X \to X+\Delta X) d\Delta W = F\Delta t \text{ ,by A.13} \quad \text{(A.19)}$$

$$< \Delta X_i \Delta X_j >= \int_{R^n} \Delta X_i \Delta X_j \; P(X \to X+\Delta X) d\Delta W = \sigma^2 \Delta t \delta_{ij} \text{ ,by A.13} \quad \text{(A.20)}$$

Then the expansion follows the steps on A.21 naming $\phi(X,t) = P(X,t)$ as the probability density and conserving linear term in Δt (equivalently quadratic in ΔX). The derivatives on X are factored out from integrals.

$$\begin{aligned}
\phi(X,t) + \frac{\partial \phi}{\partial t}\Delta t = \quad & \text{(A.21)} \\
\int_{[0,\epsilon]^n} (\phi - \sum_i \frac{\partial \phi}{\partial X_i}\Delta X_i + 1/2 \sum_{ij} \frac{\partial^2 \phi}{\partial X_i \partial X_j}\Delta X_i \Delta X_j) \; \times \\
\times (G - \sum_i \frac{\partial G}{\partial X_i}\Delta X_i + 1/2 \sum_{ij} \frac{\partial^2 G}{\partial X_i \partial X_j}\Delta X_i \Delta X_j) \; d\Delta X = \\
\phi - \sum_i \frac{\partial \phi}{\partial X_i} < \Delta X_i > + 1/2 \sum_{ij} \frac{\partial^2 \phi}{\partial X_i \partial X_j} < \Delta X_i \Delta X_j > + \\
- \sum_i \phi \frac{\partial < X_i >}{\partial X_i} + \sum_{ij} \frac{\partial \phi}{\partial X_i} \frac{\partial < \Delta X i \Delta X_j >}{\partial X_i} + 1/2 \sum_{ij} \phi \frac{\partial^2 < X_i X_j >}{\partial X_i \partial X_j}
\end{aligned}$$

Derivative of products are reconstructed and factor of Δt are equated giving the following relation A.22.

$$\frac{\partial \phi}{\partial t}\Delta t = -\sum_i \frac{\partial}{\partial X_i}(< \Delta X_i > \phi) + 1/2 \sum_{ij} \frac{\partial^2}{\partial X_i \partial X_j}(< X_i X_j > \phi) \quad \text{(A.22)}$$

$$\frac{\partial \phi}{\partial t} + \nabla \cdot J = 0 \quad \text{(A.23)}$$

$$J = \phi F - \nabla(\sigma^2 \phi) \quad \text{(A.24)}$$

The last two equations result from A.22, A.19, and A.20 stating the probability evolution as

a flux conservation equation. The last derivation is general, and can be extended for the case where $\sigma = \sigma(X)$ is a function of X and also where correlation in noise is introduced. Finally, the LCA's join probability density time evolution is given by the equation A.25.

$$\frac{\partial \phi}{\partial t} = -\sum_i \frac{\partial}{\partial X_i}(I_i - \kappa_i X_i - \beta_i \sum_{j \neq i} f_i(X))\phi + \frac{1}{2}\sigma^2 \sum_i \frac{\partial^2 \phi}{\partial X_i^2}, \tag{A.25}$$

A.2 On the eigenvalues and eigenvectors of a Toeplitz matrix

The non-linear partial differential equation A.25 can be linearized in the case when $f_i(X) = X_i$. In such case the competing factors are supposed to be equal [UM01] $\beta_i = \beta$. Also dissipation $\kappa_i = \kappa$ is equal.

$$\frac{\partial \phi}{\partial t} = -\nabla \cdot (I\phi - \omega X \phi) + \frac{1}{2}\sigma^2 \nabla^2 \phi \tag{A.26}$$

$$\omega = \begin{pmatrix} \kappa & \beta & \cdots & \beta \\ \beta & \kappa & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta \\ \beta & \cdots & \beta & \kappa \end{pmatrix} \tag{A.27}$$

The matrix ω is a Toeplitz matrix [Mey01] and must be analysed in order to study the solution of the linearised equation A.26. Further generalization of the matrix ω is considering a equal element bands in the matrix, identifying with Decision Field Theory [BT93]. Neuron's connections belonging to the same band are interpreted as having the same distance. Then ω matrix represents interaction between activation levels over distance effects.

Nevertheless, the proposed matrix A.27 corresponds to a circulant matrix, which diagonal form is known [Meh88]. Considering the determinant $\det(\omega - \lambda\, Id)$ is a polynomial, which roots are eigenvalues. Such polynomial factorize easily observing that adding each remaining column $j-1$ multiplied by θ^j where $\theta^n = 1$ then the determinant can be factorised according to A.28 with $\kappa + \beta \sum_{k=1} \theta^k - \lambda$. Note that this transformation is equivalent to a discrete Fourier transform.

$$\det(\omega - \lambda\, Id) = \begin{pmatrix} \theta^0(\kappa + \beta \sum_{k=1}^{n-1} \theta^k - \lambda) & \beta & \cdots & & \beta \\ \theta^1(\kappa + \beta \sum_{k=1}^{n-1} \theta^k - \lambda) & \kappa & \ddots & & \vdots \\ \theta^2(\kappa + \beta \sum_{k=1}^{n-1} \theta^k - \lambda) & \beta & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \beta \\ \theta^{(n-1)}(\kappa + \beta \sum_{k=1}^{n-1} \theta^k - \lambda) & \beta & \cdots & \beta & \kappa \end{pmatrix} \tag{A.28}$$

Hence determinant is linear by column indicate an eigenvalue of $\lambda = \kappa + \beta \sum_{k=1}^{n-1} \theta^k = \kappa + \beta(\theta^n - \theta)/(\theta - 1) = \kappa - \beta$, since $\theta^n = 1$. The same occurs with any other column l and adding any other column multiplied by θ^l, considering $\theta = e^{\frac{2\pi i}{n}}$. In such case, the eigenvalue are $\lambda_l = \kappa + \beta \sum_{k=1}^{n-1} \theta^{kl}$ and can be reduced to the expression with two cases (equation A.29).

$$\lambda_l = \begin{cases} \kappa - \beta & l < n \\ \kappa + (n-1)\beta & l = n \end{cases} \tag{A.29}$$

Eigenvectors can be identified from $S^1 = (\omega - (\kappa - \lambda)Id) = [1]_{ij}$ and $S^2 = (\omega - (\kappa + (n-1)\lambda)Id)$. The first matrix S^1 spawn a degenerate space with $(n-1)$ dimension for the Eigenvalue $\kappa - \lambda$. Vectors $v = [v_i]$ belonging to this space must fulfil $\sum_i v_i = 0$. They can be expanded in terms of the complex root of the unity $\{e^{\frac{2\pi i}{n}kj}\}$. The second matrix S^2 has the null vector consisting of all components equal to one.

Nevertheless, it is necessary to find an orthogonal real vector basis. The proposed complex basis (root of the unity) is orthogonal, because it corresponds to a discrete Fourier transform [Meh88]. Moreover, the null vector of S^2 is orthogonal with them and corresponds to take $j = n$. However considering only the real component $Cos(\frac{2\pi kj}{n})$ does not reach orthogonal relationship. Several orthogonalization procedures can be chosen (e.g. Gram-Smith), but a global rotation and taking the real part could be a sufficient condition for imposing orthogonality. In this case using rotation with angle of $-\pi/4$, a real orthogonal basis can be found (equation A.30).

$$R_{kj} = \frac{1}{\sqrt{n}}[Cos(\frac{2\pi kj}{n}) + Sin(\frac{2\pi kj}{n})] = Re(\frac{1}{\sqrt{n}} e^{\frac{2\pi kj}{n} - i\pi/4}) \tag{A.30}$$

The orthogonality is recovered from discrete Fourier identities. The demonstration follows on next lines. The Discrete Fourier Transformation (DFT) matrix is defined by $A = [\frac{1}{\sqrt{n}} e^{\frac{2\pi i}{n}kj}]_{kj \in \{1 \cdots n\}}$ and properties of this matrix are listed.

$$A = A^T \tag{A.31}$$

$$A^{-1} = A^* \tag{A.32}$$

$$AA^*_{jk} = \sum_l^n \frac{1}{n} e^{\frac{2\pi i}{n}(j-k)l} = \begin{cases} \sum_l^n \frac{1}{n} = 1 & j = k \\ \frac{1}{n} \frac{e^{2\pi i(n+1)(j-k)/n} - e^{2\pi i(j-k)/n}}{e^{2\pi i(j-k)/n} - 1} = \frac{e^{2\pi i(j-k)/n} - e^{2\pi i(j-k)/n}}{e^{2\pi i(j-k)/n} - 1} = 0 & j \neq k \end{cases} \tag{A.33}$$

$$A^2 = \sum_l^n \frac{1}{n} e^{\frac{2\pi i}{n}(j+k)l} = \begin{cases} \sum_l^n \frac{1}{n} = 1 & j + k \in \{n, 2n\} \\ \frac{1}{n} \frac{e^{2\pi i(n+1)(j-k)/n} - e^{2\pi i(j-k)/n}}{e^{2\pi i(j-k)/n} - 1} = \frac{e^{2\pi i(j+k)/n} - e^{2\pi i(j+k)/n}}{e^{2\pi i(j+k)/n} - 1} = 0 & j + k \notin \{n, 2n\} \end{cases} \tag{A.34}$$

$$A^3 = A^* \tag{A.35}$$

$$A^4 = Id \tag{A.36}$$

The demonstration of the inverse DFT A.33 use the geometric series results. In the same way A.34 results in a matrix filled with zeros unless of the band $k + j = n$ and the corner $k = j = n$ filled with one. The cube A.35 is obtained using the square. Using those properties and considering

the real matrix $R = \frac{1}{\sqrt{2}}(e^{i\eta}A + (e^{i\eta}A)^*)$, it is possible to set η in order to recover $R^2 = Id$ obtaining real orthogonal Eigenvectors of the matrix ω.

$$R^2 = \frac{1}{2}(e^{2i\eta}A^2 + 2AA^* + e^{-2i\eta}(A^2)^*) = \frac{1}{2}((e^{2i\eta} + e^{-2i\eta})A^2 + 2Id) \tag{A.37}$$

$$\text{If, } e^{2i\eta} + e^{-2i\eta} = 0 \tag{A.38}$$

$$\text{Then, } \eta = -\pi/4 \tag{A.39}$$

$$R = \frac{1}{\sqrt{2}}(\frac{A}{\sqrt{i}} + \sqrt{i}A^*) = \frac{1}{\sqrt{2n}}[Cos(\frac{2\pi kj}{n}) + Sin(\frac{2\pi kj}{n})]_{ij\in\{1\cdots n\}} \tag{A.40}$$

In this way the matrix ω has the following diagonal representation.

$$\omega = RDR \tag{A.41}$$

$$R_{kj} = \frac{1}{\sqrt{2n}}[Cos(\frac{2\pi kj}{n}) + Sin(\frac{2\pi kj}{n})]_{ij\in\{1\cdots n\}} \tag{A.42}$$

$$D_{kj} = \begin{cases} \kappa - \beta & k = j < n \\ \kappa + (n-1)\beta & k = j = n \\ 0 & \text{otherwise} \end{cases} \tag{A.43}$$

Further generalization of the matrix ω in circulant form, could be attempted to be calculated since eigenvalue equation that involves circulant matrices can be expressed as a convolution. Hence the Fourier transform of a convolution is a simple product and the eigenvalues can be explicitly calculated. Moreover, any Symmetric Toeplitz matrix should have the same Fourier basis of Eigenvectors. An interesting property is that R matrix does not depend on parameters κ and β.

A.3 On the Symmetries of the LCA Partial Differential Equation

The symmetries of an equation are helpful features for resolving it. A symmetry occurs when an invertible transformation of functions and variables, results in an equation of the same class. The LCA partial differential $\frac{\partial\phi}{\partial t} = -\nabla\cdot(F\phi) + \frac{1}{2}\sigma^2\nabla^2\phi$ A.25 equation has several symmetries that are presented. The transformation on (X,t,ϕ) induces transformations on pairs (F,σ).

- **Translation with constant velocity:** New coordinates are introduced in a frame moving with constant velocity V. This transformation is useful for constant elimination from F.

$$\begin{Bmatrix} X & \to & X + Vt \\ t & \to & t \end{Bmatrix} \Longleftrightarrow \begin{Bmatrix} F & \to & F + V \\ \sigma & \to & \sigma \end{Bmatrix} \tag{A.44}$$

The previous relations A.44 come from derivative transformations rules A.45.

$$\left\{\begin{array}{ccc} \frac{\partial}{\partial X_i} & \rightarrow & \frac{\partial}{\partial X_i} \\ \frac{\partial}{\partial t} & \rightarrow & \frac{\partial}{\partial t} + \sum_j V_j \frac{\partial}{\partial X_j} \end{array}\right\} \tag{A.45}$$

- **Coordinate Rotation:** A rotation $R \in SO(n)$ correspond to a ortogonal real linear transformation $R^{-1} = R^T$ (T is the transpose). Matrices R are used for separating mixed variables by diagonalization procedure.

$$\left\{\begin{array}{ccc} X & \rightarrow & RX \\ t & \rightarrow & t \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{ccc} F & \rightarrow & R^{\dagger}F \\ \sigma & \rightarrow & \sigma \end{array}\right\} \tag{A.46}$$

Those transformations are derived using the following derivative operations.

$$\left\{\begin{array}{ccc} \frac{\partial}{\partial X_i} & \rightarrow & \sum_j [R^T]_{ij} \frac{\partial}{\partial X_j} \\ \frac{\partial}{\partial t} & \rightarrow & \frac{\partial}{\partial t} \end{array}\right\} \tag{A.47}$$

- **Gauge Invariance:** The electrodynamic field first named gauge transformation for potentials that are changed by adding functions that produce null effect on electromagnetic equations. Further transformations of this kind in other fields were named also as *Gauge*. In this case the transformation is as follows.

$$\phi \rightarrow \phi + \chi \tag{A.48}$$

$$\text{Where, } \nabla \cdot (\chi F + \frac{1}{2}\sigma^2 \nabla \chi) = 0 \tag{A.49}$$

A.4 On the solution of the Ornstein Uhlenbeck (OU) equation

The Ornstein Uhlenbeck equation corresponds to the Fokker-Plank partial differential equation of a stochastic process where $F = I - \lambda X$. Hopefully, exact solutions for the free evolution of this process are known. Nevertheless, once border conditions are imposed on this system, then a more complex evolution is revealed. Solutions for this equation A.50 are presented based on past studies on this subject.

$$\frac{\partial \phi}{\partial t} = L_{OU}\phi \tag{A.50}$$

$$L_{OU}\phi = -\sum_{i=1}^{n} \frac{\partial (I_i - \lambda_i X_i)\phi}{\partial X_i} + \frac{1}{2}\sigma^2 \sum_{i=1}^{n} \frac{\partial^2 \phi}{\partial X_i^2} \tag{A.51}$$

- **Formal Solution:** The OU linear equation can be solved formally considering the linear operator L_{OU} in A.51, based on derivatives on t applied to the exponential factor a L_{OU}

operator. The operator $e^{tL_{OU}}$ is called the propagator since $e^{\Delta t L_{OU}}\phi(X,t) = \phi(X,t+\Delta t)$.

$$\phi(X,t) = e^{tL_{OU}}\phi(X,0) \tag{A.52}$$

Furthermore, if the function ϕ can be expanded on a complete set of eigenvector $\{\phi_\eta\}_{\eta\in\Lambda}$ A.53 and eigenvalues Λ, then using linearity of operator on A.52, the expression A.54 is reached. The coefficient c_η corresponds to the initial condition expansion $\phi(X,t=0)$ in the OU's eigenvector.

$$L_{OU}\phi_\eta(X) = \eta\phi_\eta(X) \tag{A.53}$$

$$\phi(X,t) = \sum_{\eta\in\Lambda} c_\eta e^{t\eta}\phi_\eta(X) \tag{A.54}$$

- **Solving the 1-d eigenvalue problem:** The one dimensional linear problem $L_{OU}\phi = \eta\phi$ A.55 is a well known problem on the field of special functions. The solutions are given by the family of Hypergeometric functions [AS65] as follows, where constants have been renamed.

$$g'' - (a-bX)g' + (c-\eta)g = 0 \tag{A.55}$$

$$\text{If,} \quad z = \frac{a}{\sqrt{2b}} - \frac{\sqrt{b}X}{\sqrt{2}} \tag{A.56}$$

$$\text{Then,} \quad g'' + 2zg' + 2(\tfrac{c-\hat{\eta}}{b})g = 0 \tag{A.57}$$

The equation A.57 can be arranged to a well known equation using the transformation $g(z) = e^{-z^2}y(z)$.

$$g' = e^{-z^2}[y' - 2zy] \tag{A.58}$$

$$g'' = e^{-z^2}[y'' - 4zy' + 2(-1+z^2)y] \tag{A.59}$$

$$y'' - 2zy' + 2(\frac{c-\hat{\eta}}{b} - 1)y = 0 \tag{A.60}$$

A.60 is named the Hermite equation [AS65]. Imposing that solution are polynomial then the constant $(\frac{c-\hat{\eta}}{b} - 1) = n$ is an integer. In that case solutions are given by a complete set of functions called Hermite Polynomials $\{H_n(z)\}_{n=0,1,\dots}$. Furthermore, the associated eigenvalues are given by A.61.

$$\hat{\eta}_n = c - b(n+1),\ n \in \mathbb{N} \tag{A.61}$$

However, the one dimensional diffusion equation reduces to A.55 using $a = \frac{2I}{\sigma^2}$, $b = \frac{2\omega}{\sigma^2}$, $c = b$, and $\hat{\eta} = \frac{2\eta}{\sigma^2}$ resulting in a negative set of eigenvalues A.62. This is because of the

biological restriction $\omega > 0$.

$$\eta_n = -\omega n, n \in \mathbb{N} \tag{A.62}$$

$$g(X) = e^{-(\frac{I}{\sigma\sqrt{\omega}} - \frac{\sqrt{\omega}X}{\sigma})^2} H_n(\frac{I}{\sigma\sqrt{\omega}} - \frac{\sqrt{\omega}X}{\sigma}) \tag{A.63}$$

For the case when regularity conditions on solutions of equation A.60 are relaxed, solutions become related to hypergeometric functions. In order to transform the last equation to a known form, the following change in variables $u = z^2$, $\hat{y}(u) = y(\sqrt{u})$ is performed.

$$y' = 2\sqrt{u}\hat{y}' \tag{A.64}$$

$$y'' = 2\hat{y}' + 4u\hat{y}'' \tag{A.65}$$

$$u\hat{y}'' + (\frac{1}{2} - u)\hat{y}' - \frac{\eta}{2\omega}\hat{y} = 0 \tag{A.66}$$

The equation A.66 is named the Kummer equation [AS65] and has as solution Confluent Hypergeometric functions [KLS10], represented by the symbol ${}_1F_1(\frac{\eta}{2\omega}; \frac{1}{2}; u)$. Finally, the eigenfunctions of the one dimensional problem with continuous eigenvalue η are described by A.67.

$$g(X) = e^{-(\frac{I}{\sigma\sqrt{\omega}} - \frac{\sqrt{\omega}X}{\sigma})^2} {}_1F_1(\frac{\eta}{2\omega}; \frac{1}{2}; (\frac{I}{\sigma\sqrt{\omega}} - \frac{\sqrt{\omega}X}{\sigma})^2) \tag{A.67}$$

- **N-Dimensional OU Solutions:** Fortunately the LCA eigenvalue differential equation A.53 is separable. The matrix ω can be diagonalised according to A.41 using a given matrix R. Using that matrix as a coordinate rotation transformation A.46, the resulting equation could be expressed as A.70 where λ_i are the eigenvalues of R in A.29.

$$\hat{X} = RX \tag{A.68}$$

$$\hat{I} = RI \tag{A.69}$$

$$\sum_{i=1}^{n} [-(\frac{\partial(\hat{I}_i - \lambda_i\hat{X}_i)\phi}{\partial\hat{X}_i}) + \frac{1}{2}\sigma^2\frac{\partial^2\phi}{\partial\hat{X}_i^2}] = \eta\phi \tag{A.70}$$

Separation of variables is achieved in A.70 looking for solutions that are product of one coordinate function $\phi(X) = \prod_{i=1}^{n} g_i(X_i)$. In such case as A.70 is a sum of operation on ϕ that only depend on the variable X_i, each g_i must fulfil the one dimensional equations A.71 with $\sum_i \eta_i = \eta$.

$$-\eta_i g_i(\hat{X}_i) - (\frac{\partial(\hat{I}_i - \lambda_i\hat{X}_i)g_i(\hat{X}_i)}{\partial\hat{X}_i}) + \frac{1}{2}\sigma^2\frac{\partial^2 g_i(\hat{X}_i)}{\partial\hat{X}_i^2} = 0, \forall\, i \tag{A.71}$$

The last expression A.71 corresponds to the one dimensional LCA problem. Considering restricted polynomial solution a multi-dimensional solution for the eigenvalue problem is

A.72.

$$\phi_{\eta_\theta}(X,t) = \prod_{k=1}^{n}[e^{-m_k\lambda_k t-(\frac{I_k}{\sigma\sqrt{\lambda_k}}-\frac{\sqrt{\lambda_k}[RX]_k}{\sigma})^2} H_{m_k}(\frac{I_k t}{\sigma\sqrt{\lambda_k}}-\frac{\sqrt{\lambda_k}[RX]_k}{\sigma})] \quad (A.72)$$

$$\eta_\theta = -\sum_{k=1}^{n} m_k\lambda_k \quad (A.73)$$

$$\theta = (m_1,\cdots,m_n),\ m_k \in \mathbb{N} \quad (A.74)$$

Considering H_{m_k} are polynomials of degree m_k, then product of them are multi-variate polynomial of degree $o = \sum_{k=1}^{n} m_k$. The eigenvector expansion A.54 must fulfill the restrictions of the problem, which are border and initial conditions. To control the order of the approximating polynomials, the eigenfunction can be chosen in order to have a partial ordered set of functions ϕ_o with order o.

- **Other traditional solutions:** The OU process has been largely studied in statistical physics [Cha43] for diffusion in fluids. Using the transformation $X \to R(X - It)$ according to A.44, with R as in A.41, and A.46, the Fokker-Plank equation is reduced to A.75.

$$\frac{\partial\phi}{\partial t} = \sum_{i=1}^{n}\frac{\partial}{\partial X_i}[\lambda_i X_i\phi + \sigma^2/2\frac{\partial\phi}{\partial X_i}] \quad (A.75)$$

$$\phi(X,t) = \prod_{i=1}^{n}\phi_i(X_i,t) \quad (A.76)$$

$$\frac{\partial\phi_i}{\partial t} = \frac{\partial}{\partial X_i}[\lambda_i X_i\phi_i + \sigma^2/2\frac{\partial\phi_i}{\partial X_i}] \quad (A.77)$$

Considering separating variables as in A.76, produces n uni-dimensional equations are obtained A.77. Changing variables by $Y_i = e^{\lambda_i t}X_i$, the equations changes to A.78. Such expression is a 1-D diffusion equation with a time dependent diffusion coefficient ($\sigma^2/2\lambda_i^2 e^{2\lambda_i t}$).

$$\frac{\partial\phi_i}{\partial t} = \sigma^2/2\lambda_i^2 e^{2\lambda_i t}\frac{\partial^2\phi_i}{\partial Y_i^2} \quad (A.78)$$

In such case time can be changed in order to obtain a classical diffusion equation. Such transformation considers a time transformation, whose derivative reproduces the time varying coefficient A.79.

$$\frac{dt}{dT} = \frac{1}{\sigma^2/2\lambda_i^2 e^{2\lambda_i t}} \quad (A.79)$$

$$T(t) = \frac{\sigma^2}{2\lambda_i}(e^{2\lambda_i t} - 1) \quad (A.80)$$

Using this transformation for time, the resulting equation is a simple diffusion A.81, which has the heat kernel (Figure A.1) as a solution A.82.

$$\frac{\partial \phi}{\partial T} = \frac{\partial^2 \phi_i}{\partial Y_i^2} \tag{A.81}$$

$$\phi_i(Y_i, T) = \frac{e^{-\frac{Y_i^2}{4T}}}{\sqrt{4\pi T}} \tag{A.82}$$

The solution of the original equation is reconstructed by applying backward all the performed transformations.

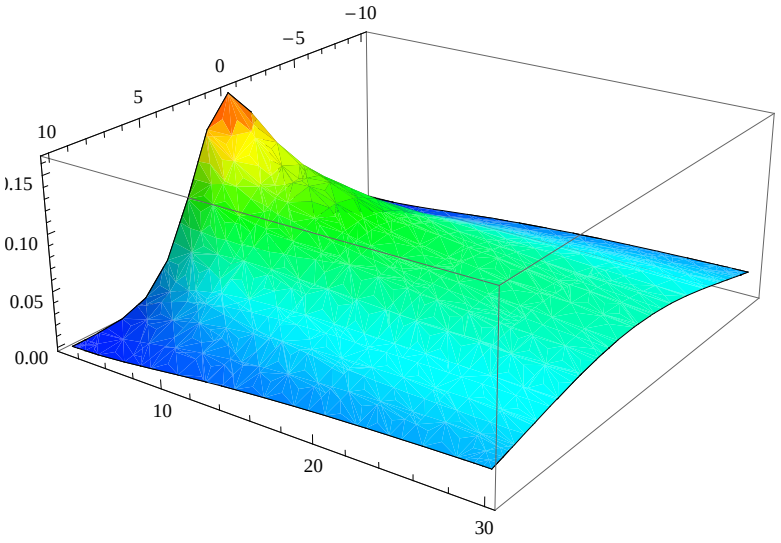

FIGURE A.1: The Heat Kernel Function in (x,t) space.

$$\phi(X,t) = \prod_{i=1}^{n} [(2\pi\sigma^2\lambda_i(e^{2\lambda_1 t} - 1))^{-1/2} e^{\frac{[R(X-It)]_i^2}{2\sigma^2\lambda_i(e^{2\lambda_i t}-1)}}] \tag{A.83}$$

$$\phi(X,t) = (2\pi\sigma^2)^{-n/2}|\Sigma_1|^{-1} e^{(X-It)^T\Sigma_2(X-It)} \tag{A.84}$$

$$|\Sigma_1| = |\omega||e^{2\omega t} - Id| \tag{A.85}$$

$$\Sigma_2 = RKR,\ K_{ij} = \frac{e^{\lambda_i t}}{2\sigma^2\lambda_i(e^{2\lambda_i t} - 1)}\delta_{ij} \tag{A.86}$$

A.5 On the Hermite polynomial basis of function

Hermite polynomials are widely used in problems related to physics and statistics. There exists several ways to introduce them, but in the context of this thesis it is more natural to begin with

the diffusion equation A.87.

$$\frac{\partial \phi}{\partial t} = \frac{\partial^2 \phi}{\partial x^2} \tag{A.87}$$

A well known solution of this equation is the heat kernel A.88 (see figure A.2).

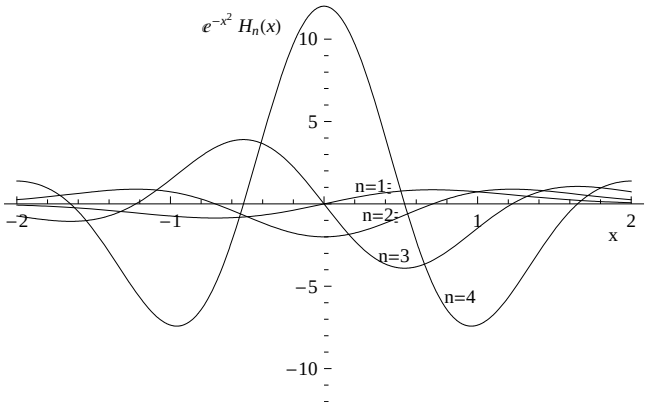

FIGURE A.2: The function $e^{-x^2} H_n(x)$ for $n = 1,2,3,4$ as solution for the diffusion equation.

$$\phi(x,t) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{x^2}{4t}} \tag{A.88}$$

If $f(x,t)$ is a solution of this equation, then its derivative $\frac{\partial f}{\partial x}$ is also a solution. This is considering

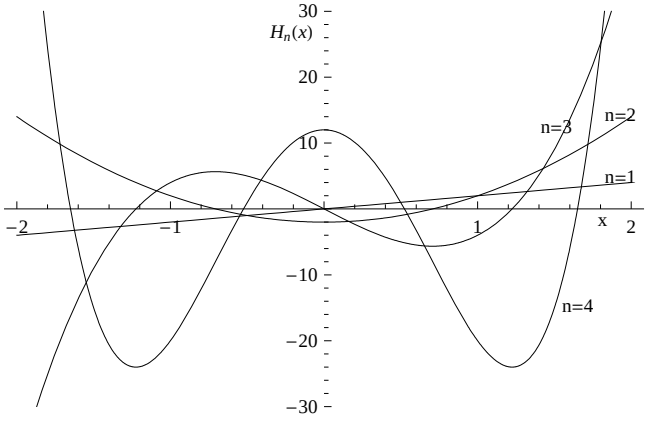

FIGURE A.3: The H_1, H_2, H_3,and H_4 Hermite Polynomials.

$f(x,t)$ having sufficient regularity conditions like being part of $\mathbb{C}^\infty$. Furthermore, $\phi_n = \frac{\partial^n \phi}{\partial x^n}$ is a solution that always is factorized by the exponential term $e^{-\frac{x^2}{4t}}$ and a polynomial. Hence such term corresponds to the Hermite polynomial of degree n. More formal and traditional definition is given by the Rodriguez's Formula A.89 and named $H_n(x)$. A few polynomials are presented on figure A.3.

$$e^{-x^2} H_n(x) = (-1)^n \frac{d^n e^{-x^2}}{dx^n} \tag{A.89}$$

TABLE A.1: First five Hermite Polynomials.

n	$H_n(x)$
0	1
1	$2x$
2	$4x^2-2$
3	$8x^3-12x$
4	$16x^4-48x^2+12$

Remarkable properties can be derived from previous definition. Orthogonality relation A.90 is established using as weight function e^{-x^2}. Noticing that integrating by part several time the product using the Rodriguez's Formula, any polynomial will vanish unless both are of the same order. Normalization constants can also be obtained with the same method. Two recurrence relations A.92A.91 can be obtained from the Rodriguez´s formula over the product and identifying terms as Hermite polynomials. From A.92, the parity A.93 of Hermite polynomials can be obtained and demonstrated by induction.

$$\int_{-\infty}^{+\infty} e^{-x^2} H_n(x) H_m(x) dx = 2^n n! \sqrt{\pi} \delta_{nm} \tag{A.90}$$

$$H_{n+1} = 2xH_n - 2nH_{n-1} \tag{A.91}$$

$$\frac{dH_n(x)}{dx} = 2nH_{n-1}(x) \tag{A.92}$$

$$H_n(-x) = (-1)^n H_n(x) \tag{A.93}$$

$$\tag{A.94}$$

A table (A.1) of the first 5 Hermite Polynomials is presented. Such polynomials could be easily constructed, using the recurrence relation A.91, starting from $H_0 = 1$ and $H_1 = 2x$. Other powerful methods in the study of special function, is the generating function. A generating function $f(x)$ for a sequence c_k corresponds to $f(x) = \sum_k \frac{c_k x^k}{n!}$. In this case the sequence of Hermite polynomials has a clean exponential generating function A.95 representation, directly derived from Rodriguez's Formula and Taylor expansion.

Furthermore, orthogonality A.90 implies linear independence and using the generating function the completeness property of the Hermite set of functions can be demonstrated. It is sufficient to demonstrate that inner product $\int_{-\infty}^{+\infty} f(z) H_n(z) e^{-z^2} dz$ with weight e^{-z^2} is null iff $f = 0$. Considering the function $F(x) = \int_{-\infty}^{+\infty} f(z) e^{-(z-x)^2} dz = \sum_{n=0}^{\infty} \frac{x^n}{n!} \int_{-\infty}^{+\infty} f(z) H_n(z) e^{-z^2} dz = 0$, using the A.95 Hermite generating function expansion, the uniform convergence properties for the sum for interchanging summations and integral, and the null inner product must be 0. However, the Fourier transform of $\mathcal{F}(F(z)) =$
$\mathcal{F}(f(z) * e^{-z^2}) = \mathcal{F}(f(z))\mathcal{F}(e^{-z^2}) = \mathcal{F}(f(z)) \sqrt{\pi} e^{-(\pi y)^2} = 0$ (where $*$ is the convolution operator) iff $\mathcal{F}(f(z)) = 0$ iff $f(z) = 0$. Nowadays, the last argumentation of completeness A.96 can be

formalized using the Dirac delta distribution representation.

$$e^{2xt-t^2} = \sum_{k=0}^{\infty} \frac{H_k(x)t^k}{n!} \tag{A.95}$$

$$\sum_{k=0}^{\infty} \frac{H_n(x)H_n(y)}{2^n n!\sqrt{\pi}} = \delta(x-y) \tag{A.96}$$

A.6 On generating probability densities that satisfy reflecting condition

Earlier known methods for finding a solution of a stochastic process that include a reflecting barrier correspond to the reflection principle [Cha43]. If the probability density $\phi_a(x,t)$ of a 1-d Brownian motion that starts on $t=0$ at $a>0$ is given, then the density of the system constrained by a reflecting barrier on $x=0$, is given by equation A.97.

$$\phi_a^R = \frac{1}{c}[\phi_a(x,t) + \phi_{-a}(-x,t)] \tag{A.97}$$

The border condition for reflection on $x=0$ is $\frac{d\phi_a^R}{dx}|_{x=0} = 0$, this means that no probability flux cross the barrier $x=0$. In this sense, ϕ_a^R satisfy border condition and the diffusion equation. Nevertheless, normalization is satisfied since both term in A.97 are normalized and then needs to be divided by $c = \int_{x>0}[\phi_a(x,t) + \phi_{-a}(-x,t)]dx$. Such term is independent of t since $\frac{dc}{dt}$ by the diffusion equation integrate a derivative that cancels on borders $\{0,\infty\}$.

This results can be generalized to several dimensions. The important point on the previous derivation, was the cancellation of derivatives on the barrier $x=0$. Furthermore, a transformation $R(x) = -x$ was applied over the parameter a and variable x in order to change the sign of the derivative for canceling on the barrier. In several dimensions, the whole group of reflection over reflective planes needs to be considered for enabling the perpendicular gradient cancellation on planes $\Delta_i = \{X \in \mathbb{R}^n \mid X_i = 0\}$.

The group of reflection is constructed by using representation by linear transformation $Q_i(X)$, where $[Q_i(X)]_i = -X_i$ and $[Q_i(X)]_{j\neq i} = X_j$. The closure is then built by composition. The complete group consists of the set $\mathfrak{R} = \{Q_{i_1}\dots Q_{i_g} \mid i_a \neq i_b,\ a \neq b,\ g = 1,\dots,n\}$. Some properties are well known, $\forall\ A, B \in \mathfrak{R}$, then $A^2 = 1$ and $AB = BA$.

If $\phi(X,t)$ satisfies the diffusion equation on $\mathbb{R}^n$, then a solution $\phi^R(X,t)$ that fulfills reflecting conditions on all planes Δ_i (or $\frac{\partial\phi^R(X,t)}{\partial X_i}|_{X_i=0}\ \forall i$) can be constructed as equation A.98.

$$\phi^R(X,t) = \sum_{Q\in\mathfrak{R}} \phi(QX,t) \tag{A.98}$$

The demonstration of such construction is by induction on n, since it holds for $n = 1$. And use the reduction $QX|_{X_i=0} = Q'X|_{X_i=0}$ where Q' is a product of all component of Q but no Q_i. Moreover, when restricted to Δ_i some terms of such sum A.98, reduces in one the dimension, then induction can be applied. For each element Q that contains a Q_i when restricted to Δ_i, is equivalent to Q' so $Q = Q'Q_i$. In such cases, $\frac{\partial\phi(QX,t)}{\partial X_i}|_{X_i=0} = -\frac{\partial\phi(Q'X,t)}{\partial X_i}|_{X_i=0}$ and both terms cancel in A.98 because $Q' \in \mathfrak{R}$. On the other hand, if a term Q' does not contain Q_i then there exists a term $Q = Q'Q_i \in \mathfrak{R}$ that will cancel. Furthermore, the function $\phi^R(X,t)$ satisfies reflection conditions on Δ as $\frac{\partial\phi^R(X,t)}{\partial X_i}|_{X\in\Delta_i} = 0$, and it is a solution of the diffusion equation due to invariance to transformation $X \to QX$, because $Q^2 = 1$.

The Ornstein-Uhlenbeck process has a rather different feature from the simple diffusion. Only in the case when the matrix $\omega = 1$ (case without lateral inhibition), the operation of reflection could be successfully applied. The border conditions relates with the orthogonal component of probability flux, that do not correspond to a simple gradient. This illustrates some of the different particularities of this random process.

A.7 On the stochastic calculus of the LCA process

More accurate simulation techniques [BR05] require a finner analysis of the LCA's stochastic equation. Stochastic calculus [Oks02] is related with consistent definitions of integrals and derivatives of Brownian processes.

Itô calculus formalize differential of functions of stochastic processes. The differential definition is called the Itô Lemma as in equation A.99.

$$df(t,X) = \frac{\partial f}{\partial t}dt + \sum_i \frac{\partial f}{\partial X_i}dX_i + \frac{1}{2}\sum_{ij} dX_i \frac{\partial^2 f}{\partial X_i \partial X_j}dX_j \tag{A.99}$$

Using this property over $X_i e^{-\omega_i t}$ and integrating from 0 to t, we obtains the set of equation A.100.

$$X_i(t) = X_i(0)e^{-\omega_i t} + \frac{I_i}{\omega_i}(1 - e^{-\omega_i t}) + \frac{\sigma}{\sqrt{2\omega}}W_i(e^{2\omega_i t} - 1)e^{-\omega_i t} \tag{A.100}$$

Performing a rotation using ω_i as the eigenvalues of the matrix ω we obtain a solution based on matrices function given by 6.10.

Appendix B

Mathematica Programs

Code are generally not included in thesis document. However in this case Mathematica's codes are shorter and serves as a proof of concept of theoretical results presented here. Mathematica is a functional computer language like LISP. Every program is a function application and evaluation of expressions. Every object in Mathematica is an expression like a mathematical equations. For this reason theoretical mathematical expression has a direct representation on this language.

On the other hand, Mathematica is a symbolic manipulator. Expression are evaluated in any case, even when unevaluated variables are present. That makes it a powerful native virtual machine expression manipulator. All the rest of operation in Mathematica that are common to many other mathematical packages use as input such abstract expression.

This is the main reason for using this software as an exploratory tool for solution of the LCA equation. The requirement are to perform symbolic manipulation while solving an optimization problem. Despite of further more efficient implementation could be designed, first a full exploration of this complex mathematical problem is required and more efficient implementation correspond to a far different problem.

B.1 Toeplitz Matrix Operations

Important part of the resolution correspond to the explicit diagonalization of the ω matrix of any dimension. Such codes are included here.

Omega matrix: This square matrix represent neural interaction intensity. It depends on the number of decision to perform n (the dimension number), parameter κ (k) and λ (l).

```
w[n_,k_,l_]:=Table[If[i==j,k,l],{i,1,n},{j,1,n}]
```

Rotation matrix R: This matrix describes the base of orthonormal eigenvector of the ω matrix. It only depends on the number of decision n.

```
R[n_] :=
 Table[(Cos[2 Pi i j/n] + Sin[2 Pi i j/n])/Sqrt[n], {i, 1, n}, {j, 1,n}]
```

Eigenvalues of Omega: Those values are simple linear combination of parameter κ and λ, that are indexed by i. One eigenvalue has degeneracy $(n-1)$ and comes first on the labeling sequence.

```
d[i_, k_, l_, n_] := If[i < n, k - l, k + (n - 1) l]
```

This definition could be immediately tested for $n = 5$ (B.1).

```
FullSimplify[R[5].w[a, b, 5].R[5]]
```

$$\begin{pmatrix} a-b & 0 & 0 & 0 & 0 \\ 0 & a-b & 0 & 0 & 0 \\ 0 & 0 & a-b & 0 & 0 \\ 0 & 0 & 0 & a-b & 0 \\ 0 & 0 & 0 & 0 & a+4b \end{pmatrix} \tag{B.1}$$

B.2 Explicit Symbolic Basis Set of Function

Polynomial solution chosen are those found in equation 5.40. Using the previous definition the explicit solution for the unconstrained case are given by the following definitions.

One-dimensional solution: In base of those function the N-dimensional case is built. It depends on the number of dimension n and two other parameters that will be described later. The variable is x that corresponds to the rotated neural activity (RX).

```
f[x_, a_, b_, n_] := HermiteH[n, Sqrt[a] x - b/Sqrt[a]]/Sqrt[2^n n!]
```

N-Dimensional Eigenfunction of the LCA operator: Those functions (g) have been shown to be product of the previous defined function. The function g depends on the neural activity vector x, a vector of integer z of the same dimension than x where $\sum_i z_i$ is the dimension of the resulting polynomial, parameter κ (k), λ (l), σ (s), and vector I (ii).

```
a[i_, k_, l_, s_, n_] := d[i, k, l, n]/s^2
b[i_, ii_, s_, n_] := (R[n].ii)[[i]]/s^2
```

```
g[x_, z_, k_, l_, s_, ii_] :=
 FullSimplify[
  Product[f[(R[Length[x]].x)[[i]],
    a[i, k, l, s, Length[x]], b[i, ii, s, Length[x]], z[[i]]],
    {i, 1, Length[x]}]]
```

The resulting function could be inspected graphically using proper features of Mathematica.

```
gg[x1_, x2_] := g[{x1, x2}, {5, 4}, 1.1, 1, 0.06, {0.4, 0.6}, 2]
Plot3D[gg[x1, x2], {x1, 0, 1}, {x2, 0, 1}, PlotRange -> Full]
```

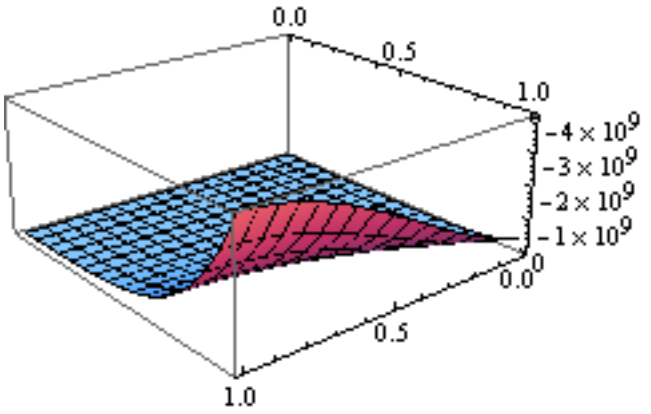

FIGURE B.1: The shape of an Eigenfunction of the LCA operator in 2-D.

The probability flux operator: In order to test that the corresponding set of function it is needed to build the differential problem. The flux vector operates on an abstract function ϕ (Phi), a vector of integer components, the neural activity vector x, the CEL vector I, κ (k), λ (l), σ (s), and the time t.

```
Flux[\[Phi]_, z_, x_, ii_, k_, l_, \[Sigma]_, t_] :=
 FullSimplify[(ii - (w[k, l, Length[x]].x)) \[Phi][x, z, ii, k,
     l, \[Sigma], t] + (\[Sigma]^2/2) Table[
     D[\[Phi][x, z, ii, k, l, \[Sigma], t], x[[i]]], {i, 1,
      Length[x]}]]
```

The LCA operator: It is built in base of the Flux operator.

```
L[ff_, z_, x_, ii_, k_, l_, s_, t_] :=
 Sum[D[Flux[ff, z, x, ii, k, l, s, t][[i]], x[[i]] ], {i, 1,
   Length[x]}]
```

LCA solutions

```
\[Phi][x_, z_, ii_, k_, l_, s_, t_] :=
 Exp[-alfa[z, k, l] t] g[x, z, k, l, s, ii]
```

Solution presented here could be tested immediately by mean of applying such operators, for instance $n = 3$.

```
FullSimplify[
 L[\[Phi], {3, 3}, {x1, x2}, {i1, i2}, a, b, Sqrt[2], t] -
  D[\[Phi][{x1, x2}, {3, 3}, {i1, i2}, a, b, Sqrt[2], t], t]]
0
```

The 0 result reveal Mathematica's symbolic operation proved for this particular parameter case, the proposed multivariate function is a solution of the unconstrained LCA equation.

B.3 Approximating solution to the constrained problem

The approximating scheme is based on the propagator operator, which includes a penalization force for asymptotic recovering of border condition. Penalization forces are defined by:

```
Penalization[x_, m_] := (x^(2 m) + (x - 1)^(2 m) - 2^(-2 m + 1))/
  m^(1/2)
```

The corresponding generating operator becomes:

```
PenalizedFlux[ff_, z_, x_, ii_, k_, l_, \[Sigma]_, t_, m_] :=
 Flux[ff, z, x, ii, k, l, \[Sigma], t] +
  Penalization[x, m] ff[x, z, ii, k, l, \[Sigma], t]
PenalizedL[ff_, z_, x_, ii_, k_, l_, s_, t_] :=
 Sum[D[PenalizedFlux[ff, z, x, ii, k, l, s, t][[i]], x[[i]] ], {i, 1,
   Length[x]}]
```

Finally the penalized propagator becomes:

```
PowerPenalizedL[ff_, z_, x_, ii_, k_, l_, s_, t_, m_] :=
 If[m = 0, ff[x, z, ii, k, l, s, t],
  PenalizedL[PowerPenalizedL[ff, z, x, ii, k, l, s, t, m - 1], z, x,
```

```
   ii, k, l, s, t]]
PenalizedU[ff_, z_, x_, ii_, k_, l_, s_, t_, m_] :=
 Sum[PowerPenalizedL[ff, z, x, ii, k, l, s, t, h], {h, 0, m}]
```

B.4 Collecting Data

This section includes the code used to that stores session from cookies. It is presented in this way for clarifying the operation.

The JavaScript code

```
/**
* @author Pablo Roman (proman@ing.uchile.cl)
* @version 1.0 2009
* Web Log generator based on cookie tracking
* store user anonymous sessions.
*/
function getCookie(NameOfCookie) {
if (document.cookie.length > 0) {
begin = document.cookie.indexOf(NameOfCookie+"=");
if (begin != -1) {
begin += NameOfCookie.length+1;
end = document.cookie.indexOf(";", begin);
if (end == -1) end = document.cookie.length;
return unescape(document.cookie.substring(begin, end));
}
}
return null;
}

function setCookie (name, value, lifesecond, access_path, domain) {
  var cookietext = name + "=" + escape(value);
    if (lifesecond != null) {
      var today=new Date();
      var expiredate = new Date();
      expiredate.setTime(today.getTime() + 1000*lifesecond);
      cookietext += "; expires=" + expiredate.toGMTString();
    }
```

```
    if (access_path != null) {
      cookietext += "; PATH="+access_path ;
    } if (domain != null) {
      cookietext += "; domain=" + domain;
    }
   document.cookie = cookietext;
   return null;
}

function trackCookie(callid) {
   var dummyImage, cookie, query, query2, query3;

// setting cookie if not set
   cookie=getCookie('DIIWUM');
   if(cookie==null|cookie=="") {
cookie='DII'+Math.random();
setCookie('DIIWUM', cookie, 1800);
   }

   dummyImage = new Image();
   url=document.URL;
   //**********
   query="/track.php?x="+url;
   //**********
   query2="&y="+cookie;
   query3="&z="+callid;
   dummyImage.src=query+query2+query3;
   return dummyImage;
}

function trackOUT() {
   trackCookie("OUT");
}

var x=trackCookie("IN");
window.onbeforeunload = trackOUT;
```

The php code

```
<?php
/**
 * Tracking de usuarios por pagina
 * @author Pablo Roman
 * @email 'proman@ing.uchile.cl'
 * @date 2009/04/03
 * @version 1.0
 */
##############################################################
// Genera output de codigo de un gif de 1x1
class gifpix {
        var $start = '474946383961010001008000000';
        var $color = 'ffffff';
        var $black = '000000';
        var $marker = '21f904';
        var $transparent = '01';
        var $end = '000000002c0000000001000100000020244010003b';
        function hex2bin($s) {
                for ($i = 0; $i < strlen($s); $i += 2) {
                        $bin .= chr(hexdec(substr($s,$i,2)));
                }
                return $bin;
        }
        function create($color = -1) {
                if (($color!= -1) && (strlen($color)==6)) {
                        $this->transparent = '00';
                        if ($color == '000000')
                                $this->black = 'ffffff';
                        $this->color = $color;
                }
                $hex = $this->start.$this->color.$this->black.$this->marker.$this->tr

                return $this->hex2bin($hex);
        }
}

// genera regitro de session de usuario con cookieid|tiempo|IP_client|url_host|url_pa

Class LogSess {
```

```
        var $arch="";
        function LogSess($name="") {
                $this->arch=$name;
                if ($name=="") exit();
        }
        function register($value="", $host="", $url="", $callid="", $query=""){
                if ($url=="" || $value=="") {
                        exit();
                } else { // No requiere LOCK segun documentacion PHP4-5
                        $hjhg=@fopen($this->arch,'a');
                        @fwrite($hjhg,$value."|".time()."|".$_SERVER['REMOTE_ADDR']."
.$callid."|".$query."|".$_SERVER['HTTP_USER_AGENT']."\n");
                        @fflush($hjhg);
                        @fclose($hjhg);
                }
        }
}
################################################################
// creando un gif transparente de 1x1 como salida
@header("Content-Type: image/gif");
$gifpix = new gifpix();
print $gifpix->create();

// chequeando url y cookie
$url=$_REQUEST["x"];
$val=$_REQUEST["y"];
$callid=$_REQUEST["z"];
$purl=@parse_url($url);
$pval=split("\.",$val);

if ($purl && strcmp($pval[0],'DII0')==0){ // si es url valido
        // generamos el registro
        $log=new LogSess("/home/areas/web2006/LOG/sesslog.txt");
        $log->register($pval[1], $purl["host"], $purl["path"], $callid, $purl["query"]
} else { // no es url valido
        exit();
}
?>
```

Bibliography

[AC08] J. Alexander and A. Cockburn. An empirical characterisation of electronic document navigation. In *GI '08: Proceedings of graphics interface 2008*, pages 123–130, Toronto, Ont., Canada, Canada, 2008. Canadian Information Processing Society.

[AG07] S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer, 2007.

[AGP95] Soren Asmussen, Peter Glynn, and Jim Pitman. Discretization error in simulation of one dimensional reflecting brownian motion. *The annals of applied probability*, 5(4):875–895, 1995.

[AL95] Moshe Ben Akiva and Steven Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, 1995.

[Alb71] James Albus. A theory of cerebellar function. *Mathematical Biosciences*, 10(2):25–61, 1971.

[Ale09] J. Alexander. *Understanding and Improving Navigation Within Electronic Documents*. PhD thesis, University of Canterbury, Christchurch, New Zealand, 2009.

[AMO93] Ahuja, T. Magnanti, and J. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[And08] Chris Anderson. Wired Magazine, Editorial, June 2008.

[AR03] Ajith Abraham and Vitorino Ramos. Web usage mining using artificial ant colony clustering and genetic programming. In *Procs. Of the 2003 IEEE Congress on Evolutionary Computation (CEC2003)*, pages 1384–1391, 2003.

[ARRV10] E. Andaur, S. Rios, P. E. Román, and J. D. Velásquez. "best web site structure for users based on a genetic algorithm approach". In *"The First Workshop in Business Analytics and Optimization (BAO 2010)"*, Santiago, Chile, January 2010. "E. Andaur Engineering thesis co-guidance by P. E. Roman".

[AS65] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1965.

[AS06] Padmapriya Ayyagari and Yang Sun. Modeling the internet and the web: Probabilistic methods and algorithms. by pierre baldi, paolo frasconi, padhraic smith, john wiley and sons ltd., west sussex, england, 2003. 285 pp isbn 0 470 84906 1. *Inf. Process. Manage.*, 42(1):325–326, 2006.

[ASA] Opera Software ASA. Opera browser. http://www.opera.com.

[ATDE09] E. Adar, J. Teevan, S.T. Dumais, and J.L. Elsas. The web changes everything: understanding the dynamics of web content. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291, New York, NY, USA, 2009. ACM.

[Aul04] Charles Aulds. *High performance MySQL*. O'Reilly Media, 2004.

[BBM+06] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, 4(113):700–765, 2006.

[BCR06] A. Blum, T.-H. H. Chan, and M. R. Rwebangira. A random-surfer web-graph model. In *Proceedings of the eigth Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmics and Combinatorics.*, pages 238–246. Society for Industrial and Applied Mathematics, 2006.

[BD07] G. G. Brown and R. F. Dell. Formulating integer linear programs: A rogues' gallery. *Informs Transactions on Education*, 7(2):1–13, 2007.

[BGRS99] K. Beyer, J. Goldstein, R. Ramakristan, and U. Shaft. When is nearest neighbor meaningful. *DataBase Theory.*, 15:217–235, 1999.

[BHS99] B. Berendt, A. Hotho, and G. Stumme. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems*, 1(1):5–32, 1999.

[Bix02] Robert E. Bixby. Solving real-world linear programs: A decade and more of progress. *Operations Research*, 50(1):3–15, 2002.

[BJJT06] Jerome R. Busemeyer, Ryan K. Jessup, Joseph G. Johnson, and James T. Townsend. Building bridges between neural models and complex decision making behaviour. *Neural Networks*, 19(8):1047 – 1058, 2006. Neurobiology of Decision Making.

[BL00] José Borges and Mark Levene. Data mining of user navigation patterns. In *WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, pages 92–111, London, UK, 2000. Springer-Verlag.

[BL07] Jose Borges and Mark Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Trans. on Knowl. and Data Eng.*, 19(4):441–452, 2007.

[BMSW01] B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. In *Proc.of the Workshop on Web Mining, First SIAM Internat.Conf. on Data Mining*, pages 7–14, 2001.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[BP07] Narayan L. Bhamidipati and Sankar K. Pal. Stemming via distribution-based word segregation for classification and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(2):350–360, 2007.

[BPF10] Jerome R. Busemeyer, Emmanuel M. Pothos, and Riccardo Franco. A quantum theoretical explanation for probability judgment errors. *Psychology Revue Letter*, 2010. SUBMITTED TO.

[BPR06] A. Beskos, O. Papaspiliopoulos, and G. O. Roberts. Retrospective exact paths with applications. *Bernoulli*, 12(6):1077–1098, 2006.

[BR05] A. Beskos and G. O. Roberts. Exact simulation of diffusions. *Annals of Applied Probability*, (15):2422–2444, 2005.

[BR09] R. Burget and I. Rudolfova. Web page element classification based on visual features. *Intelligent Information and Database Systems, Asian Conference on*, 0:67–72, 2009.

[BSNM06] Kenneth H. Britten, Michael N. Shadlen, William T. Newsome, and J. Anthony Movshon. Response of neurons in macaque motion signals. *Visual Neuroscience*, 10:1157–1169, 2006.

[BT93] J. R. Busemeyer and J. T. Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459, July 1993.

[BTCF09] M.A. Bayir, I.H. Toroslu, A. Cosar, and G. Fidan. Smart miner: a new framework for mining large scale web usage data. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 161–170, New York, NY, USA, 2009. ACM.

[BUZM07] R. Bogacz, M. Usher, J. Zhang, and J. McClelland. Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transaction of the Royal Society B*, 362(1485):1655–1670, 2007.

[BWT06] Jerome R. Busemeyer, Zheng Wang, and James T. Townsend. A quantum dynamics of human decision making. *Journal of Mathematical Psychology*, (50):220–241, 2006.

[ByC04] R. Baeza-yates and C. Castillo. Crawling the infinite web: Five levels are enough. In *In Proceedings of the third Workshop on Web Graphs (WAW*, pages 156–167. Springer, 2004.

[BYCE07] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2), May 2007.

[BYP06] R. Baeza-Yates and B. Poblete. Dynamics of the chilean web structure. *Comput. Netw.*, 50(10):1464–1473, 2006.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[Cas04] Carlos Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, Santiago, Chile, 2004.

[CCW+07] S. E. Coull, M. P. Collins, C. V. Wright, F. Monrose, and M. K. Reiter. On web browsing privacy in anonymized netflows. In *SS'07: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, pages 1–14, Berkeley, CA, USA, 2007. USENIX Association.

[CDG+06] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: a distributed storage system for structured data. In *OSDI '06: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation*, pages 15–15, Berkeley, CA, USA, 2006. USENIX Association.

[CDK+99] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure. *Computer*, 32(8):60–67, 1999.

[CF06] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[CFL09] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591+, 2009.

[CFV07] Pablo Castells, Miriam Fernandez, and David Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. on Knowl. and Data Eng.*, 19(2):261–272, 2007.

[CGM00] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[CGM03] Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. *ACM Trans. Internet Technol.*, 3(3):256–290, 2003.

[Cha43] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15(1):1–89, Jan 1943.

[CHM+00] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Visualization of navigation patterns on a web site using model-based clustering. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–284, New York, NY, USA, 2000. ACM.

[CKP07] Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. Page-level template detection via isotonic smoothing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 61–70, New York, NY, USA, 2007. ACM.

[CKS08] Anne K. Churchland, Roozbeh Kiani, and Michael N. Shadlen. Decision making with multiple choice. *Nature Neuroscience*, 11(6):693–702, June 2008.

[CLRS01] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2001.

[CMS99] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1:5–32, 1999.

[CMS02] R. Cooley, B. Mobasher, and J. Srivastava. Towards semantic web mining. In *Proc. in First Int. Semantic Web Conference*, pages 264–278, 2002.

[Cor] Mozilla Corporation. Mozilla firefox browser. http://www.mozilla.org.

[CP95] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. In *Computer Networks and ISDN Systems*, pages 1065–1073, 1995.

[CS03] R. Cancho and R. Sole. Least effort and the origins of scaling in human language. In *National Academy of Science USA*, volume 100, pages 788–791, 2003.

[CZ03] Xin Chen and Xiaodong Zhang. A popularity-based prediction model for web prefetching. *Computer*, 36(3):63–70, 2003.

[DFIN08] Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, and Wolfgang Nejdl. Semantically enhanced entity ranking. In *WISE '08: Proceedings of the 9th international conference on Web Information Systems Engineering*, pages 176–188, Berlin, Heidelberg, 2008. Springer-Verlag.

[DGEU07] G. N. Demir, M. Goksedef, and A. S. Etaner-Uyar. Effects of session representation models on the performance of web recommender systems. In *ICDEW '07: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 931–936, Washington, DC, USA, 2007. IEEE Computer Society.

[DJ02] Xing Dongshan and Shen Junyi. A new markov model for web access prediction. *Computing in Science and Engg.*, 4(6):34–39, 2002.

[DKM+02] S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Internet Technol.*, 2(3):205–223, 2002.

[DMPG05] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles. Automatic identification of informative sections of web pages. *IEEE Trans. on Knowl. and Data Eng.*, 17(9):1233–1246, 2005.

[DRV08a] R. F. Dell, P. E. Román, and J. D. Velásquez. Identifying web user session using an integer programming approach. In *Procs. of The XIII Latin Ibero-American Congress on Operations Research (CLAIO 2008)*, Cartagena, Colombia, 2008.

[DRV08b] R. F. Dell, P. E. Román, and J. D. Velásquez. Web user session reconstruction using integer programming. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 385–388, Washington, DC, USA, 2008. IEEE Computer Society.

[DRV09a] R. F. Dell, P. E. Román, and J. D. Velásquez. Fast combinatorial algorithm for web user session reconstruction. In *Procs. of 24th IFIP TC7 Conference*, Buenos Aires, Argentina, 2009.

[DRV09b] R. F. Dell, P. E. Román, and J. D. Velásquez. "user session reconstruction with back button browsing". In *Lecture Note in Computer Science, LNAI 5711. Procs. Of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES 2009*, pages 326–332, Santiago, Chile, September 2009.

[DRV09c] R. F. Dell, P. E. Román, and J. D. Velásquez. Un método de optimización lineal entera para el análisis de sesiones de usuarios web. *Revista de Ingenieria de Sistemas*, 23:109–124, 2009.

[DRV10] R. F. Dell, P. E. Román, and J. D. Velásquez. "optimization models for construction of web user sessions". Working Paper, 2010.

[DS04] P. Desikan and J. Srivastava. Mining temporally evolving graphs. In B. Mobasher, B. Liu, B. Masand, and O. Nasraoui, editors, *Webmining and Web Usage Analysis (WebKDD'04)*, 2004.

[DT09] R. Das and I. Turkoglu. Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Syst. Appl.*, 36(3):6635–6644, 2009.

[DUO07] Gül Nildem Demir, A. Sima Uyar, and Sule Gündüz Ögüdücü. Graph-based sequence clustering through multiobjective evolutionary algorithms for web recommender systems. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1943–1950, New York, NY, USA, 2007. ACM.

[DV09] L. E. Dujovne and J. D. Velásquez. Design and implementation of a methodology for identifying website keyobjects. In *Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference, KES 2009, Santiago, Chile, September 28-30, 2009, Proceedings, Part I*, volume 5711 of *Lecture Notes in Computer Science*, pages 301–308, 2009.

[DW00] M. X. Dong and R. J-B Wets. Estimating density functions: a constrained maximum likelihood approach. *Journal of Nonparametric Statistics*, 12(4):549–595, 2000.

[EC06] Shinto Eguchi and John Copas. Interpreting kullback-leibler divergence with the neyman-pearson lemma. *J. Multivar. Anal.*, 97(9):2034–2040, 2006.

[Ely09] Bert Ely. Bad rules produce bad outcomes: Underlying bad public-policy causes of the u.s. financial crisis. *Cato Journal*, 29(1):93–99, 2009.

[EVK05] Magdalini Eirinaki, Michalis Vazirgiannis, and Dimitris Kapogiannis. Web path recommendations based on page ranking and markov models. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 2–9, New York, NY, USA, 2005. ACM.

[FHK03] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, and Jon M. Kleinberg. Fast algorithms for large-state-space hmms with applications to web usage analysis.

In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003.

[FL03] F. Facca and P. Lanzi. Recent developments in web usage mining research. In *DaWaK*, pages 140–150, 2003.

[FMNW03] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM.

[Gal00] Thomson Gale. Ilog cplex. *Manufacturing Automation*, 9, 2000.

[GAM08] GAMS Development Corporation. General algebraic modeling system (gams), 2008. Software available from http://www.gams.com.

[GCP08] Paul W. Glimcher, Colin Camerer, and Russel A. Poldrack. *Neuroeconomics: Decision Making and the Brain*. Academic Press, 2008.

[GD08] K. Grannis and E. Davis. Online sales to climb despite struggling economy, 2008. According to Shop.org/Forrester Research Study.

[GD09] K. Grannis and E. Davis. China internet network information center, 14th statistical survey report on the internet development of china 2009, 2009. According to http://www.cnnic.net.cn/uploadfiles/pdf/2009/10/13/94556.pdf.

[GFL08] L. Granka, M. Feusner, and L. Lorigo. Eye monitoring in online search. In R.I. Hammoud and T. Ohno, editors, *Passive Eye Monitoring*, Signals and Communication Technology, pages 347–372. Springer Berlin-Heidelberg,, 2008. Part VI.

[GG87] Stephen Grossberg and William E. Gutowski. Neural dynamics of decision making under risk: Affective balance and cognitive-emotional interactions. *Psychological Review*, 94(3):300–318, 1987.

[GH98] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory*. Willey-Interscience, third edition, 1998.

[Gil96] Daniel T. Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical Review E*, 54(2):2089–2091, 1996.

[GJM01] Rayid Ghani, Rosie Jones, and Dunja Mladenic. Mining the web to create minority language corpora. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286, New York, NY, USA, 2001. ACM.

[GKB09] Nico Görnitz, Marius Kloft, and Ulf Brefeld. Active and semi-supervised data domain description. In *ECML PKDD '09: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–422, Berlin, Heidelberg, 2009. Springer-Verlag.

[GL97] J. G. Gaines and T. J. Lyons. Variable step size control in the numerical solution of stochastic differential equations. *SIAM J. Appl. Math.*, 57:1455–1484, October 1997.

[Gla94] Steven Glassman. A caching relay for the world wide web. In *Selected papers of the first conference on World-Wide Web*, pages 165–173, Amsterdam, The Netherlands, The Netherlands, 1994. Elsevier Science Publishers B. V.

[GM10] Emmanuel Gobet and Stephane Menozzi. Stopped diffusion processes: boundary correction and overshoot. *Stochastic Processes and their Applications*, 120:130–162, 2010.

[GO03] Şule Gündüz and M. Tamer Özsu. A web page prediction model based on click-stream tree representation of user behavior. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–540, New York, NY, USA, 2003. ACM.

[Goo09] Google. Google investor relation: Finacial table. http://investor.google.com/financial/tables.html, 2009.

[Gro00] Stephen Grossberg. The complementary brain: Unifying brain dynamics and modularity. *Trend in Cognitive Science*, (4):233–246, 2000.

[GRP08] Yong Zhen Guo, Kotagiri Ramamohanarao, and Laurence A. F. Park. Web page prediction based on conditional random fields. In *Proceeding of the 2008 conference on ECAI 2008*, pages 251–255, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[GRP09] Yong Zhen Guo, Kotagiri Ramamohanarao, and Laurence A. Park. Grouped ecoc conditional random fields for prediction of web user behavior. In *PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 757–763, Berlin, Heidelberg, 2009. Springer-Verlag.

[GZDN09] Julien Gaugaz, Jakub Zakrzewski, Gianluca Demartini, and Wolfgang Nejdl. How to trace and revise identities. In *ESWC 2009 Heraklion: Proceedings of the 6th European Semantic Web Conference on The Semantic Web*, pages 414–428, Berlin, Heidelberg, 2009. Springer-Verlag.

[Han99] D.J. Hand. Statistics and data mining: intersecting disciplines. *SIGKDD Explor. Newsl.*, 1(1):16–19, 1999.

[HDS06] Timothy D. Hanks, Jochen Ditterich, and Michael N. Shadlen. Microstimulation of macaque area lip affects decision-making in a motion discrimination task. *Nature Neuroscience*, 9(5):682–689, April 2006.

[Heb49] Donald Hebb. *The organization of behaviour : a neuropsychological theory*. L. Erlbaum Associates, 1949.

[Hen04] Svetlana Hensman. Construction of conceptual graph representation of texts. In *HLT-NAACL '04: Proceedings of the Student Research Workshop at HLT-NAACL 2004 on XX*, pages 49–54, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[HF52] A. L. Hodgkin and Huxley A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117(4):500–544, 1952.

[Hil36] A. V. Hill. Excitation and accommodation in nerve. *Proceedings of the Royal Society B*, 119:305–355, 1936.

[HMS09] Tahira Hasan, Sudhir P. Mudur, and Nematollaah Shiri. A session generalization technique for improved web usage mining. In *WIDM '09: Proceeding of the eleventh international workshop on Web information and data management*, pages 23–30, New York, NY, USA, 2009. ACM.

[HNJ08] Paul Huntington, David Nicholas, and Hamid R. Jamali. Website usage metrics: A re-assessment of session data. *Information Processing & Management*, 44(1):358–372, January 2008.

[HPPL98] B. Huberman, P. Pirolli, J. Pitkow, and R. M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.

[HW07] B.A. Huberman and F. Wu. The economics of attention: maximizing user value in information-rich environments. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 16–20, New York, NY, USA, 2007. ACM.

[HWV04] Birgit Hay, Geert Wets, and Koen Vanhoof. Mining navigation patterns using a sequence alignment method. *Knowl. Inf. Syst.*, 6(2):150–163, 2004.

[HX09] Wang Hongwei and Liu Xie. Adaptive site design based on web mining and topology. In *CSIE '09: Proceedings of the 2009 WRI World Congress on Computer*

Science and Information Engineering, pages 184–189, Washington, DC, USA, 2009. IEEE Computer Society.

[IG04] P.G. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 767–778, New York, NY, USA, 2004. ACM.

[IH07] G. Iachello and J. Hong. End-user privacy in human-computer interaction. *Found. Trends Hum.-Comput. Interact.*, 1(1):1–137, 2007.

[IJ05] P. Ingwersen and K. Jirvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, first edition, 2005.

[ITP07] H. Hannah Inbarani, K. Thangavel, and A. Pethalakshmi. Rough set based feature selection for web usage mining. In *ICCIMA '07: Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, pages 33–38, Washington, DC, USA, 2007. IEEE Computer Society.

[JGP+07] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

[JI06] A. Jatowt and M. Ishizuka. Temporal multi-page summarization. *Web Intelli. and Agent Sys.*, 4(2):163–180, 2006.

[JJ04] Jason J. Jung and Geun-Sik Jo. Semantic outlier analysis for sessionizing web logs. In *ECML/PKDD Conference*, pages 13–25, 2004.

[JJJ06] A. Juels, M. Jakobsson, and T.N. Jagatic. Cache cookies for browser authentication (extended abstract). In *SP '06: Proceedings of the 2006 IEEE Symposium on Security and Privacy*, pages 301–305, Washington, DC, USA, 2006. IEEE Computer Society.

[JK00] A. Joshi and R. Krishnapuram. On mining web access logs. In *Proc. of the 2000 ACM SIGMOD Workshop on Research Issue in Data Mining and Knowledge Discovery*, pages 63–69, 2000.

[JM09] Sabine Janzen and Wolfgang Maass. Ontology-based natural language processing for in-store shopping situations. In *ICSC '09: Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pages 361–366, Washington, DC, USA, 2009. IEEE Computer Society.

[JPT03] Søren Jespersen, Torben Bach Pedersen, and Jesper Thorhauge. Evaluating the markov assumption for web usage mining. In *WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management*, pages 82–89, New York, NY, USA, 2003. ACM.

[JS07] Wei Jin and Rohini K. Srihari. Graph-based text representation and knowledge discovery. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 807–811, New York, NY, USA, 2007. ACM.

[Jun04] Jason J. Jung. Ontology-based partitioning of data steam for web mining: A case study of web logs. In *ICCS 2004, 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part I*, pages 247–254, 2004.

[JZM04] Xin Jin, Yanzan Zhou, and Bamshad Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 197–205, New York, NY, USA, 2004. ACM.

[Kah03] Daniel Kahneman. *The Nobel Prizes 2002*, chapter Maps of Bounded Rationality:A PERSPECTIVE ON INTUITIVE JUDGMENT AND CHOICE. Nobel Foundation, 2003.

[Kau09] A. Kausshik. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*. Sybex, 2009.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[KC06] N. Khasawneh and C. Chan. Active user-based and ontology-based web log data preprocessing for web usage mining. In *2006 IEEE / WIC / ACM International Conference on Web Intelligence(WI 2006), Hong Kong, China*, pages 325–328. IEEE Computer Society, 2006.

[KDNL06] Y. Ke, L. Deng, W. Ng, and D.L. Lee. Web dynamics and their ramifications for the development of web search engines. *Comput. Netw.*, 50(10):1430–1447, 2006.

[Kel07] M. Kellar. *An Examination of User Behaviour during Web Information Tasks*. PhD thesis, Dalhousie University, Halifax, Nova Scotia, Canada, 2007.

[KK03] Y. Kim and J. Kim. Web prefetching using display-based prediction. In *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, page 486, Washington, DC, USA, 2003. IEEE Computer Society.

[KKA+08] Constantinos Kolias, Vassilis Kolias, Ioannis Anagnostopoulos, Georgios Kambourakis, and Eleftherios Kayafas. Enhancing user privacy in adaptive web sites with client-side user profiles. In *SMAP '08: Proceedings of the 2008 Third International Workshop on Semantic Media Adaptation and Personalization*, pages 170–176, Washington, DC, USA, 2008. IEEE Computer Society.

[KKKM05] V.V. Kryssanov, K. Kakusho, E.L. Kuleshov, and M. Minoh. Modeling hypermedia-based communication. *Information Sciences*, 174(1-2):37–53, 2005.

[KLS10] Roelof Koekoek, Peter A. Lesky, and René F. Swarttouw. *Hypergeometric Orthogonal Polynomials and Their q-Analogues (Springer Monographs in Mathematics)*. Springer, 2010.

[Koh88] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. pages 509–521, 1988.

[KP95] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1 edition, 1995. Second Corrected Printing.

[KS08] Roozbeh Kiani and Michael N. Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(8):759–764, May 2008.

[KT79] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

[KT05] Javed I. Khan and Qingping Tao. Exploiting webspace organization for accelerating web prefetching. *Web Intelli. and Agent Sys.*, 3(2):117–129, 2005.

[Lam68] D. R. J. Laming. *Information theory of choice reaction time*. Wiley, 1968.

[Lan00] D. Langford. *Internet ethics*. MacMillan Press Ltd, 2000.

[LB05] J. C. Lansey and B. Bukiet. Internet search result probabilities, heaps' law and word associativity. *Journal of Quantitative Linguistics*, 16(1):40–66, 2005.

[LBL01] M. Levene, J. Borges, and G. Loizou. Zipf's law for web surfers. *Knowl. Inf. Syst.*, 3(1):120–129, 2001.

[LdW05] D.C. Van Leijenhorst and Th.P. Van der Weide. A formal derivation of heaps' law. *Inf. Sci. Inf. Comput. Sci.*, 170(2-4):263–272, 2005.

[LFM08] Y. Li, B. Feng, and Q. Mao. Research on path completion technique in web usage mining. *Computer Science and Computational Technology, International Symposium on*, 1:554–559, 2008.

[Lin05] J. Linn. Technology and web user data privacy: A survey of risks and countermeasures. *IEEE Security and Privacy*, 3(1):52–58, 2005.

[Liu09] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[Loh09] Steve Lohr. A 1 million dollars research bargain for netflix, and maybe a model for others. *New York Times*, 2009.

[LRV10a] P. Loyola, P. E. Román, and J. D. Velásquez. Colony surfer: Discovering the distribution of text preferences from web usage. In *Procs. Of the First Workshop in Business Analytics and Optimization (BAO)*, 2010.

[LRV10b] P. Loyola, P. E. Román, and J. D. Velásquez. "ant colony surfer: Discovering the distribution of text preferences from web usage". In *"The First Workshop in Business Analytics and Optimization (BAO 2010)"*, Santiago, Chile, January 2010. "P. Loyola Engineering thesis co-guidance by P. E. Román".

[LS65] R. Luce and P. Suppes. Preference, utility and subjective probability, in luce, bush and galanter (eds.), handbook of mathematical psychology iii. *Handbook of Mathematical Psychology III, Wiley*, pages 249–410, 1965.

[LTLS05] Man Lan, Chew-Lim Tan, Hwee-Boon Low, and Sam-Yuan Sung. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1032–1033, New York, NY, USA, 2005. ACM.

[LTSL09] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):721–735, 2009.

[LZY04] Jiming Liu, Shiwu Zhang, and Jie Yang. Characterizing web usage regularities with information foraging agents. *IEEE Trans. on Knowl. and Data Eng.*, 16(5):566–584, 2004.

[Mae00] Roman Maeder. *Computer Science with Mathematica: Theory and Practice for Science, Mathematics, and Engineering*. 2000.

[Mar69] D. Marr. A theory of cerebellar cortex. *Journal of Physiology*, 202(2):437–470, 1969.

[May07] D. Maynor. *Metasploit Toolkit for Penetration Testing, Exploit Development, and Vulnerability Research*. Syngress, first edition, 2007.

[MB98] Robert C. Miller and Krishna Bharat. Sphinx: A framework for creating personal, site-specific web crawlers. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, pages 119–130, 1998.

[MB09] M. Moloney and F. Bannister. A privacy control theory for online environments. In *HICSS '09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Washington, DC, USA, 2009. IEEE Computer Society.

[McF73] D. McFadden. Is conditional logit analysis of qualitative choice behavior. *Zarembka (ed.), Frontiers in Econometrics, Academic Press*, 1973.

[MCS99] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *KDEX '99: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, page 19, Washington, DC, USA, 1999. IEEE Computer Society.

[MCS00] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.

[MCWG85] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1985.

[MDLN01] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In *WIDM '01: Proceedings of the 3rd international workshop on Web information and data management*, pages 9–15, New York, NY, USA, 2001. ACM.

[MDLN02] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Min. Knowl. Discov.*, 6(1):61–82, 2002.

[MDPS95] P. Read Montague, Peter Dayan, Christophe Person, and Terrence J. Sejnowski. Bee foraging in certain environment using predictive hebbian learning. *Nature*, 377(26):725–728, 1995.

[Meh88] Madan Lal Mehta. *Matrix Theory: Selected Topics and Useful Results*. Les edition de physique, 1988.

[Mey01] Carl Meyer. *Matrix Analysis*. SIAM: Society for Industrial and Applied Mathematics, 2001.

[MIOK88] D.E. Meyer, D.E. Irwin, A.M. Osman, and J. Kounios. The dynamics of cognition and action: mental processes inferred from speed-accuracy decomposition. *Psychol Rev*, 95(2):183–237, 1988.

[Mob06] B. Mobasher. Web usage mining. In B. Liu, editor, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*, chapter 12. Springer Berlin-Heidelberg, 2006.

[Mor02] Tatsunori Mori. Information gain ratio as term weight: the case of summarization of ir results. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[MPTM08] F. Masseglia, P. Poncelet, M. Teisseire, and A. Marascu. Web usage mining: extracting unexpected periods from web logs. *Data Min. Knowl. Discov.*, 16(1):39–65, 2008.

[MS99] C. D. Manning and H. Schutze. *Fundation of Statistical Natural Language Processing*. The MIT Press, 1999.

[MS08] Viktor Mayer-Schonberger. Nutzliches vergessen. In *Goodbye privacy grundrechte in der digitalen welt (Ars Electronica)*, pages 253–265, 2008.

[MT07] Alexander Mikroyannidis and Babis Theodoulidis. Heraclitus: A framework for semantic web adaptation. *IEEE Internet Computing*, 11(3):45–52, 2007.

[Nad07] D. Nadeax. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD thesis, University of Ottawa, Ottawa, Canada, 2007.

[NM94] W. Newey and D. McFadden. Large sample estimation and hypothesis testing. *HANDBOOK OF ECONOMETRICS, North Holland: Amsterdam*, 4:2111–2245, 1994.

[NSS+08] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain. A web usage mining framework for mining evolving user profiles in dynamic web sites. *IEEE Trans. on Knowl. and Data Eng.*, 20(2):202–215, 2008.

[OC03] C. Olston and E.H. Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Trans. Comput.-Hum. Interact.*, 10(3):177–197, 2003.

[Oks02] Bernt K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 5 edition, 2002.

[OP08] C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 437–446, New York, NY, USA, 2008. ACM.

[OWHM07] H. Obendorf, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 597–606, 2007.

[Pat07] A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, 2007.

[PB09] Emmanuel M. Pothos and Jerome R. Busemeyer. A quantum probability explanation for violation of rational decision theory. *Proceedings of The Royal Society B*, 276(1165):2171–2178, 2009.

[PE00] Mike Perkowitz and Oren Etzioni. Towards adaptive web sites: conceptual framework and case study. *Artif. Intell.*, 118(1-2):245–275, 2000.

[Pir09] P. Pirolli. Power of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42:33–40, 2009.

[PKO+04] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.

[Por06] M. F. Porter. An algorithm for suffix stripping. *Electronic Library and Electronic Systems*, 40:211–218, 2006.

[POSPG09] R. Peña-Ortiz, J. Sahuquillo, A. Pont, and J.A. Gil. Dweb model: Representing web 2.0 dynamism. *Comput. Commun.*, 32(6):1118–1128, 2009.

[PSJ08] Sungjune Park, Nallan C. Suresh, and Bong-Keun Jeong. Sequence-based clustering for web usage mining: A new experimental framework and ann-enhanced k-means algorithm. *Data Knowl. Eng.*, 65(3):512–543, 2008.

[PSK09] Till Plumbaum, Tino Stelter, and Alexander Korth. Semantic web usage mining: Using semantics to understand user intentions. In *UMAP '09: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization*, pages 391–396, Berlin, Heidelberg, 2009. Springer-Verlag.

[PTM02] S. K. Pal, V. Talwar, and P. Mitra. Web mining in soft computing framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, 13:1163–1177, 2002.

[QD09] X. Qi and B.D. Davison. Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2):1–31, 2009.

[Rat78] R. Ratcliff. A theory of memory retrieval. *Psychological Review*, (83):59–108, 1978.

[RB03] T. A. Runkler and J. C. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, February 2003.

[RBDM07] I. K. Reay, P. Beatty, S. Dick, and J. Miller. A survey and analysis of the p3p protocol's agents, adoption, maintenance, and future. *IEEE Transactions on Dependable and Secure Computing*, 4:151–164, 2007.

[RBM06] Jörg Rieskamp, Jerome R. Busemeyer, and Barbara A. Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.

[RDV10] "P.E. Román, R.F. Dell, and J.D. Velásquez". *"Advanced Techniques in Web Intelligence"*, chapter "Advanced Techniques in Web Data Pre-Processing and Cleaning". "Springer", "2010".

[RDVL10] P.E. Román, R.F. Dell, J.D. Velásquez, and P. Loyola. Optimization models for sessionization. *Intelligent Data Analysis*, 2010. Submitted to.

[Res92] Sidney I. Resnick. *Adventures in stochastic processes*. Birkhauser Verlag, Basel, Switzerland, Switzerland, 1992.

[RHGJ06] S. Rugaber, N. Harel, S. Govindharaj, and D. Jerding. Problems modeling web sites and user behavior. In *WSE '06: Proceedings of the Eighth IEEE International Symposium on Web Site Evolution*, pages 83–94, Washington, DC, USA, 2006. IEEE Computer Society.

[RKJ08] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 43–52, New York, NY, USA, 2008. ACM.

[RLV10] "P. E. Román, G. L'Huillier, and J. D. Velásquez". *"Advanced Techniques in Web Intelligence"*, chapter "Web Usage Mining". "Springer", "2010".

[Rob04] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004, 2004.

[RS02] J. D. Roitman and M. N. Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, 22:9475–9489, 2002.

[RTR01] RTR. The tent approximation for partial differential equations with border condition. Private Communication Department of Physics, Faculty of Engineering, University of Chile., 2001.

[RV06] P. E. Román and J. D. Velásquez. Improving a web site using keywords. In *Procs. of The XIII Latin Ibero-American Congress on Operations Research (CLAIO 2008)*, Montevideo, Uruguay, 2006.

[RV08a] Sebastián A. Ríos and Juan D. Velásquez. Semantic web usage mining by a concept-based approach for off-line web site enhancements. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 234–241, Washington, DC, USA, 2008. IEEE Computer Society.

[RV08b] P. E. Román and J. D. Velásquez. Markov chain for modeling the web user behavior. In *Procs. of The XIII Latin Ibero-American Congress on Operations Research (CLAIO 2008)*, Cartagena, Colombia, 2008.

[RV08c] Pablo E. Román and Juan D. Velásquez. Markov chain for modeling web user browsing behavior: statistical inference. In *XIV Latin Ibero-American Congress on Operations Research (CLAIO)*, 2008.

[RV09a] P. E. Román and J. D. Velásquez. Analysis of the web user behavior with a psychologically-based diffusion model. In *Of the AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures, Arlington, USA.*, Arlington, Washington DC, USA, 2009. Technical Paper of the AAAI.

[RV09b] P. E. Román and J. D. Velásquez. A dynamic stochastic model applied to the analysis of the web user behavior. In Snasel et al., editor, *The 2009 AWIC 6th Atlantic Web Intelligence Conference*, pages 31–40, Prague, Czech Republic, 2009. Invited Lecture, in Intelligent and Soft Computing Series, Advances in Intelligent Web Mastering-2.

[RV10a] P. E. Román and J. D. Velásquez. "artificial web user simulation and web usage mining". In *"The First Workshop in Business Analytics and Optimization (BAO 2010)"*, Santiago, Chile, January 2010.

[RV10b] P. E. Román and J. D. Velásquez. "stochastic simulation of web users". In *Procs. of the 2010 IEEE/WIC/ACM International Conference*, Toronto, Canada, September 2010. IEEE Press.

[RV10c] P. E. Román and J. D. Velásquez. "the time course of the web user". In *Second Workshop on Time Use Observatory (TUO2)*, San Felipe, Chile, March 2010.

[RVZM] R Ratcliff, T Van Zandt, and G McKoon.

[Sar00] Ramesh R. Sarukkai. Link prediction and path analysis using markov chains. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 377–386, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

[Saw03] Shlomo S. Sawilowsky. You think you've got trivials? *Journal of Modern Applied Statistical Methods*, 2(1):218–225, 2003.

[SCDT00] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 2(1):12–23, 2000.

[Sch01] J. D. Schall. Neural basis of deciding, choosing and acting. *National Review of Neuroscience*, 2(1):33–42, 2001.

[SDM+09] Marc Spaniol, Dimitar Denev, Arturas Mazeika, Gerhard Weikum, and Pierre Senellart. Data quality in web archiving. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*, pages 19–26, New York, NY, USA, 2009. ACM.

[Seb02] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[SF98] Myra Spiliopoulou and Lukas Faulstich. Wum: A web utilization miner. In Paolo Atzeni, Alberto O. Mendelzon, and Giansalvatore Mecca, editors, *WebDB*, volume 1590 of *Lecture Notes in Computer Science*, pages 184–103. Springer, 1998.

[SH03] R. Sen and M. Hansen. Predicting web user's next access based on log data. *J. Comput. Graph. Stat.*, 12(1):143–155, 2003.

[She09] Shady Shehata. A wordnet-based semantic model for enhancing text clustering. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 477–482, Washington, DC, USA, 2009. IEEE Computer Society.

[SK09] V. Snásel and M. Kudelka. Web content mining focused on named objects. In *(IHCI) First International Conference on Intelligent Human Computer Interaction*, pages 37–58. Springer India, 2009.

[SL08] N. Sadagopan and J. Li. Characterizing typical and atypical user sessions in clickstreams. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 885–894, New York, NY, USA, 2008. ACM.

[SMBN03] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs Journal on Computing*, 15(2):171–190, 2003.

[SMS08] Casey M Schneider-Mizell and Leonard M Sander. A generalized voter model on complex networks. Technical Report arXiv:0804.1269, Department of Physics, University of Michigan, Apr 2008. 15 pages, 3 figures.

[SN98] M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of NeuroScience*, 18(20):3870–3896, 1998.

[SPM09] M. V. B. Soares, R. C. Prati, and M. C. Monard. Improvement on the porter's stemming algorithm for portuguese. *IEEE Latin America Transaction*, 7(4):472–477, 2009.

[SSQ03] Jing Shi, Fang Shi, and HangPing Qiu. User's interests navigation model based on hidden markov model. In Guoyin Wang, Qing Liu, Yiyu Yao, and Andrzej Skowron, editors, *RSFDGrC*, volume 2639 of *Lecture Notes in Computer Science*, pages 644–647. Springer, 2003.

[SSST07] Kay-Uwe Schmidt, Ljiljana Stojanovic, Nenad Stojanovic, and Susan Thomas. On enriching ajax with semantics: The web personalization use case. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, *ESWC*, volume 4519 of *Lecture Notes in Computer Science*, pages 686–700. Springer, 2007.

[Sto60a] M. Stone. Models for choice reaction time. *Psychometrika*, (25):251–260, 1960.

[Sto60b] Mervyn Stone. Models for choice-reaction time. *Psychometrika*, 25(3):251–260, 1960.

[TG97] L. Tauscher and S. Greenberg. Revisitation patterns in world wide web navigation. In *Procs. of the Conference on Human Factors in Computing Systems*, pages 22–27, March 1997.

[THLC09] Yu-Hui Tao, Tzung-Pei Hong, Wen-Yang Lin, and Wen-Yuan Chiu. A practical extension of web usage mining with intentional browsing data toward usage. *Expert Syst. Appl.*, 36(2):3937–3945, 2009.

[THS08] Yu-Hui Tao, Tzung-Pei Hong, and Yu-Ming Su. Web usage mining with intentional browsing data. *Expert Syst. Appl.*, 34(3):1893–1904, 2008.

[TK07] G. Tsoumakas and I. Katakis. Multi label classification: An overview. *International Journal of Data Warehouse and Mining*, 3(3):1–13, 2007.

[Tom98] J. A. Tomlin. , "a new paradigm for ranking pages on the world wide web", www2003, may 20-24, 2003, budapest, hungary. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.

[TPNT07] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Major components of the gravity recommendation system. *SIGKDD Explor. Newsl.*, 9(2):80–83, 2007.

[TS79] Amos Tversky and Shmuel Sattath. Preference trees. *Psychological Review*, 86(6):542–573, 1979.

[TS93] Amos Tversky and I. Simonson. Context-dependent preferences. *Management Science*, 39(10):1179–1189, 1993.

[TVN10] George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. An experimental study on unsupervised graph-based word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010.*, pages 184–198, 2010.

[UBL+08] C. Ullrich, K. Borau, H. Luo, X. Tan, L. Shen, and R. Shen. Why web 2.0 is good for learning and for research: principles and prototypes. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 705–714, New York, NY, USA, 2008. ACM.

[UFTS09] D. Urbansky, M. Feldmann, J. A. Thom, and A. Schill. Entity extraction from the web with webknox. In *6th Atlantic Web Intelligence Conference (AWIC)*, Prague, Czech Republic, 2009.

[UM01] M. Usher and J. McClelland. The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 2(1):550–592, 2001.

[Vaz05] Alexei Vazquez. Exact results for the barabasi model of human dynamics. *Physical Review Letters*, 95(24):248701, 2005.

[VEY+04] Juan D. Velásquez, Pablo A. Estévez, Hiroshi Yasuda, Terumasa Aoki, and Eduardo S. Vera. Intelligent web site: Understanding the visitor behavior. In Mircea Gh. Negoita, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES*, volume 3213 of *Lecture Notes in Computer Science*, pages 140–147. Springer, 2004.

[VOD+06] A. Vazquez, J. Gama Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and Albert-Laszlo Barabasi. Modeling bursts and heavy tails in human dynamics. *PHYSICAL REVIEW E*, 73(3):036127, 2006.

[VP07] J.D. Velasquez and V. Palade. A knowledge base for the maintenance of knowledge. *Journal of Knowledge Based Systems*, 1(20):238–248, 2007.

[VP08] J.D. Velásquez and V. Palade. *Adaptive web sites: A knowledge extraction from web data approach*. IOS Press, Amsterdam, NL, 2008.

[VWYA04] Juan D. Velásquez, Richard Weber, Hiroshi Yasuda, and Terumasa Aoki. A methodology to find web site keywords. In *EEE '04: Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, pages 285–292, Washington, DC, USA, 2004. IEEE Computer Society.

[VYA+03] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber, and E. Vera. Using self organizing feature maps to acquire knowledge about visitor behavior in a web site. *Lecture Notes in Artificial Intelligence*, 2773(1):951–958, September 2003.

[VYAW04] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.

[WD07] Ryen W. White and Steven M. Drucker. Investigating behavioral variability in web search. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007.

[WDT09] Peter Wittek, Sándor Darányi, and Chew Lim Tan. Improving text classification by a sense spectrum approach to term expansion. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 183–191, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[WGH+00] Shi Wang, Wen Gao, Tiejun Huang, Jiyong Ma, Jintao Li, and Hui Xie. Adaptive online retail web site based on hidden markov model. In *WAIM '00: Proceedings of the First International Conference on Web-Age Information Management*, pages 177–188, London, UK, 2000. Springer-Verlag.

[WH06] Yong Wang and Julia Hodges. Document clustering with semantic analysis. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, page 54.3, Washington, DC, USA, 2006. IEEE Computer Society.

[Whi07] R. W. White. Investigating behavioral variability in web search. In *In Proc. WWW*, pages 21–30, 2007.

[Wic02] Thomas D. Wickens. *Elementary Signal Detection Theory*. Oxford University Press, 2002.

[Wil69] Daniel Willshaw. Non-holographic associative memory. *Nature*, 222:960–962, 1969.

[WJH09] S.S. Won, J. Jin, and J.I. Hong. Contextual web history: using visual and contextual cues to improve web browser history. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1457–1466, New York, NY, USA, 2009. ACM.

[WOHM06] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer. Off the beaten tracks: exploring three aspects of web navigation. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 133–142, New York, NY, USA, 2006. ACM.

[WOHM08] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer. Not quite the average: An empirical study of web use. *ACM Trans. Web*, 2(1):1–31, 2008.

[Wol92] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[WW48] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948.

[WWZ05] J. Wang, X. Wu, and C. Zhang. Support vector machines based on kmeans clustering for real time business intelligence systems. *Int. J. Bus. Intell. Data Min.*, 1(1):54–64, 2005.

[XdSCP08] Geraldo Xexeo, Jano de Souza, Patricia F. Castro, and Wallace A. Pinheiro. Using wavelets to classify documents. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:272–278, 2008.

[XZJL01] Jitian Xiao, Yanchun Zhang, Xiaohua Jia, and Tianzhu Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian database conference*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.

[YWL06] Lean Yu, Shouyang Wang, and K.K. Lai. An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, 18:217–230, 2006.

[YXW07] C. Yue, M. Xie, and H. Wang. Automatic cookie usage setting with cookiepicker. In *DSN '07: Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 460–470, Washington, DC, USA, 2007. IEEE Computer Society.

[YZZ03] X. Yan, C. Zhang, and S. Zhang. Toward databases mining: Pre-processing collected data. *Applied Artificial Intelligence*, 17(5-6):545–561, 2003.

[ZAN99] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. Predicting users' requests on the www. In *UM '99: Proceedings of the seventh international conference on User modeling*, pages 275–284, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc.

[Zaw02] Jeremy D. Zawodny. *Linux apache web server administration*. Sybex; 2 edition, 2002.

[ZCL04] Z. Zhang, J. Chen, and X. Li. A preprocessing framework and approach for web applications. *J. Web Eng.*, 2(3):176–192, 2004.

[ZLW07a] Y. Zhou, H. Leung, and P. Winoto. Mnav: A markov model-based web site navigability measure. *IEEE Trans. Softw. Eng.*, 33(12):869–890, 2007.

[ZLW07b] Yuming Zhou, Hareton Leung, and Pinata Winoto. Mnav: A markov model-based web site navigability measure. *IEEE Trans. Softw. Eng.*, 33(12):869–890, 2007.

Index

Z