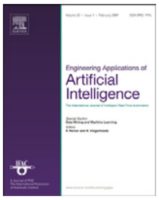
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Extracting significant Website Key Objects: A Semantic Web mining approach

Juan D. Velásquez *, Luis E. Dujovne, Gaston L'Huillier

Department of Industrial Engineering, Universidad de Chile, Republica 701 - P.O. Box 8370439, Santiago, Chile

A R T I C L E I N F O

Available online 15 March 2011

Keywords:
Website Key Objects
Web user preferences
Web usage mining
Semantic Web mining
Self Organizing Feature Maps
Ontology Engineering
Web Intelligence
Clickstream analysis
Web personalization

A B S T R A C T

Web mining has been traditionally used in different application domains in order to enhance the content that Web users are accessing. Likewise, Website administrators are interested in finding new approaches to improve their Website content according to their users' preferences. Furthermore, the *Semantic Web* has been considered as an alternative to represent Web content in a way which can be used by intelligent techniques to provide the organization, meaning, and definition of Web content. In this work, we define the Website Key Object Extraction problem, whose solution is based on a Semantic Web mining approach to extract from a given Website core ontology, new relations between objects according to their Web user interests. This methodology was applied to a real Website, whose results showed that the automatic extraction of Key Objects is highly competitive against traditional surveys applied to Web users.

1. Introduction

The rapid growth of the World Wide Web, the assembly of large-scale volumes of Web data, and ever exponentially increasing applications have led to the development of ever smarter approaches to extract patterns and build knowledge with the aid of artificial intelligence techniques. These techniques have been used, together with information technology, in a wide range of applications. This is where semantics, social network analysis, Web structure, content, usage, and other aspects have already been and will increasingly be included in many application domains.

One of such domains is related to how Web users browse the Webpages and Websites looking for information. Often, they require and there is a greater possibility of them staying or returning to the Website if they find the content they are searching for in a Website. For this, Website administrators intend to reach the highest user base they can, therefore it is within their interest to provide accurate and correct content (Velásquez and Palade, 2008).

However, different difficulties are present in this application domain. On the one hand, Web users' interests often change and it is often unclear to assume at first sight what the users' interests are. On the other hand, whether the content has been correctly presented is a relevant question for any Website administrator. Furthermore, the content may be presented in several formats, which could stem from free text to images or videos. In this sense, not only it is unclear what is the content that users are looking for, but also their preferences in terms of the format that should be considered.

A typical Website is composed primarily of a free text being formatted within the limitations imposed by the HTML standard. However, they also consist of other data formats such as images, videos, etc. An example of this is the most successful sites of the so-called Web 2.0 such as You Tube where the main focus of interest lies in Web videos. A drawback to this trend is that the formats shown above do not provide information regarding their content which can easily be retrieved by a computer and, therefore, a small degree of content analysis can be carried out in relation to them.

In Web mining, several techniques have been created to discover the problem stated above, focusing mainly in text-based Websites (Velásquez et al., 2005; Wang et al., 2005), leaving aside what other presentation formats, images, or flash animations present on the most successful Websites nowadays. Taking this into consideration the main idea is to define WebObjects which could represent any content in a Website independently of the format in which it is presented and to discover which of this WebObjects attracts user interest. These objects are named Website Key Objects.

The main contribution of this work is a methodology that enables the extraction of significant Website Key Objects, following a Semantic Web mining approach. In this case, Semantic Web mining will be considered as using data mining algorithms in order to extract relevant information from the Semantic Web representation of a given Website. Specifically, the idea is to extract a new relation between structured components from a Website (WebObjects),

* Corresponding author.
E-mail addresses: jvelasqu@dii.uchile.cl (J.D. Velásquez), dujovne@dcc.uchile.cl (L.E. Dujovne), glhuilli@dii.uchile.cl (G. L'Huillier).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.02.001

represented by a simple core ontology. This relation is extracted from the Web user's perspective, represented by their collected sessions, from which patterns are extracted.

This paper is organized as follows. Section 2 provides an overview on related work for significant Website objects extraction. In Section 3 the proposed Semantic Web mining approach to find Website Key Objects is presented. Then, in Section 4 an application of our work for a Chilean Geographical Information Systems company[1] is described. Finally, conclusions and future work are presented in Section 5.

2. Related work

Different approaches have been previously used to find relevant information for end-users in Web mining applications. In this context, traditional techniques, such as Web content, usage and structural mining, have been taken into account by most researchers. Furthermore, Web semantic mining techniques have been proposed for extracting relevant information from a given Website. In this section, the main contributions on traditional Web mining techniques, as well as semantic approach, are reviewed.

2.1. Significant information extraction using Web mining

Significant information extraction from Web content has been a major focus for many researchers, where different degrees of information, such as words, text passages, or WebObjects, have been taken into account. Furthermore, many methodologies have been proposed, and some of the most relevant approaches will be discussed in the following.

The methodology created by Velásquez (to appear) and Velásquez et al. (2005) for finding Website keywords forms the basis of this work, in which information retrieval and Web usage mining techniques were used within the Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996) framework, to find the keywords that define the search process for a group of users. The process described is based on five fundamental steps. The first one is associated with the Vector Space Model definition from a given Website (Salton et al., 1975) and the processing of Weblogs, in order to include the end-user information. Afterwards, this methodology focusses on finding the relationship between the page interest and time spent, as well as selecting the most important pages from the extracted user sessions. Finally, by using different clustering techniques (particularly *k*-Means and Kohonen Self-Organizing Feature Maps), the process to discover Keywords in clusters takes place. According to Velásquez (to appear), this methodology can be used in order to improve a given Website information to enhance the general content of the Website.

The attempts for attracting users to Websites have been made since the Web became a massive source of information, and the study of usability in Websites has been one of the most widespread research domains. One of the first approaches to create Website usability patterns is the Common User Access (CUA) proposed by Berry (1998). Another approach described by Nielsen (2006) focuses in how to present the content in terms of typography, design, presentation of elements and other end-user properties associated with their interaction with visualization components. These patterns have shown to be effective, but they lack information about the direct feedback for the users of the site.

Several researches have been done lately in the field of WebObjects. In this section some of these methodologies will be described focusing mainly on three of them: WebPage Element Classification (Burget and Rudolfová, 2009), Named Objects (Tiwary et al., 2009), and Entity extraction from the Web (Urbansky et al., 2008).

In terms of WebPage Element Classification, the work proposed by Burget and Rudolfová (2009) focuses on the fact that in normal pages found across the World Wide Web, most additional information such as copyright notices or advertisement influences in a negative manner the results of the Web. To avoid this, a method for detecting the interesting areas in a Webpage from a human reader approach is created. This is accomplished by dividing a Webpage into visual blocks and detecting the purpose of each block based on their visual features.

The relation of the users description of Webpages is the focus of the approach presented by Tiwary et al. (2009), where the perception of a Webpage is obtained through the intention of users. This intention delivers information for both the user and the Webmaster and is the basis of Named Objects which allows the mining of patterns within Websites by using a Web Design Pattern approach. This named objects are used as the basis of mining methods which allows Web Content Mining.

A Web knowledge extraction system is proposed by Urbansky et al. (2008), which uses Concepts, Attributes and Entities as input data. By modelling this using an ontology, facts from generic structures and formats are extracted. Afterwards a self-supervised learning algorithm automatically estimates the precision of these structures.

Significant information extraction in Web mining has been developed from different perspectives. One of the leading approaches was proposed by Gao et al. (2005), whose research is based on determining which is the information on a given Web that is most interesting to end-users. In this approach, information retrieval techniques (Jr and Ziviani, 2004) are used along with Web usage mining to infer the user preferences found in objects. Also, microformats (Khare and Çelik, 2006) have been previously used as a mechanism to add semantics into a given Website, and improving the information extraction from the usage data of different Websites (Plumbaum et al., 2009).

2.2. Significant information extraction using Semantic Web mining

In terms of Semantic Web mining or approaches that extract relevant patterns from the Semantic Web, different techniques have been proposed. One of the first approaches (Li and Zhong, 2003) presents an ontology representation of user profiles in order to design efficient Web mining models. This approach is one of the first to be considered oriented towards the correct characterization of user profiles according to a semantic representation.

In Li and Zhong (2003), researchers extended previous ontology towards a more flexible and capable ontology for further applications related to the usage of Web user profiles for different Web mining applications. Furthermore, in Li and Zhong (2007), introduced pattern taxonomy and Ontology mining into a general semantic representation.

Despite different WebObject learning approaches being developed, one of the main difficulties for their correct analysis is the comparison between their different formats. In Sahami (2006), an approach to set a common representation between WebObjects using natural language processing and semantics to set their differences has been proposed.

In terms of identifying WebObjects using Semantic Web mining, Chambers et al. (2006) present WebLearn, whose objective is to identify WebObjects by using the semantic content of the written language surrounding WebObjects.

According to Stumme et al. (2006), the baseline representation of the semantic information for WebObjects is one of the main concerns that Semantic Web mining researchers has to take into

[1] http://www.dmapas.cl (online: accessed 13 December 2010)

consideration. Likewise, in Berendt et al. (2002) ontology learning (Maedche and Staab, 2001; Poon and Domingos, to appear; Tsoi et al., 2009) approaches must be considered in terms of applying Semantic Web mining approaches. These approaches are based on basic pattern recognition techniques to extract, prune, refine, and reuse of Web information to set the basics for ontology learning from an architecture structure.

Most of the ontology learning approaches are based on textual analysis of documents (Zavitsanos et al., 2010), where using probabilistic reduction techniques the semantic representation of a given corpus is identified. Different approaches have been used in the past to extract ontologies from document-based environments, where one of the most promising automatic techniques are based on latent semantic analysis tools (Paaß et al., 2004), where using topic modelling (Blei et al., 2003) and other information theoretic approaches (Papadimitriou et al., 1998; Baeza-Yates and Ribeiro-Neto, 1999) have been widely used.

3. A methodology to extract Website Key Objects

In this section, the proposed approach to extract the Website Key Objects is presented. Firstly, the problem definition and the general notation of terms are introduced. Secondly, Web content and Web usage mining terms and methodologies used to achieve the Website Key Object definition are presented. Finally, the core methodology and main contribution of this work are detailed.

3.1. *Problem definition and general notation*

In the following, WebObjects and Website Key Objects terms are presented, as well as the proposed ontological representation and mathematical notation. Likewise, the Website Key Object (WSKO) extraction problem is introduced.

3.1.1. *WebObjects*

WebObjects may represent structured text or any other multimedia format present in a Website. In order to process their content using computer software it is necessary to include metadata that describes them. The definition given in a later section allows a human to comprehend the nature of a WebObject. However, a computer is unable to understand this so the following definition of a WebObject is introduced.

Definition 1 (*WebObject*). "It is a structured group of words or a multimedia file present within a Webpage that has metadata for describing its content".

The implementation of WebObjects can be made in several ways because it relies heavily on the ontology used to describe them. In this work a simple ontology was introduced based on the work by the MPEG to include metadata in videos. In this sense an XML document will be associated with each WebObject present in a Webpage. Despite the complex semantic analysis of multimedia, metadata is used to define the WebObject within it. In this sense, a set of N objects is defined by $\mathbf{x}=\{x_1,\ldots,x_N\}$, and each object is defined by a set of M concepts, $\mathbf{c}=\{c_1,\ldots,c_M\}$.

In our approach, the usage of metadata to describe WebObjects will be considered as the basis to constitute the information source, in order to build a vectorial representation of its content. However, the end-user's point of view will be considered as the principal research topic of this approach. Therefore, the contents of the Website and Weblogs are combined for processing.

3.1.2. *Website Key Objects*

Having described what WebObjects are, we introduce the term Website Key Object (WSKO) as follows,

Definition 2 (*Website Key Objects*). "WebObjects or groups of WebObjects that attract the Web users attention".

Key Objects can be considered as elements on a given Website that provide knowledge of both content and formats that appear interesting to end-users. Enhancements can be made in presentation as well as in content when Key Objects are identified, and used to improve the structure of a Website.

In order to accomplish this, and to propose which objects are the ones that must be taken into account when a given Website is re-engineered, the extraction of Key Objects problem must be solved.

Definition 3 (*Website Key Object Extraction*). The Website Key Object (WSKO) Extraction problem is defined as setting an order relation $\preceq_{KO}$ between a list of WebObjects from a given Website, taking into account the relevance of WebObjects for the Website users.

In order to achieve an accurate WSKO extraction, the Website can be represented as a core-ontology (Stumme et al., 2006), from which concepts and its relations will be used as input to the WSKO extraction process.

In general terms, the Key Objects can be represented as an order relation from the end-user perspective, where each object's relevance is inferred from the usage of the given Website.

Definition 4 (*Key Objects core ontology*). According to Stumme et al. (2006), a Website Key Objects' core ontology is represented by the tuple

$$\mathcal{O} := (\mathcal{C}, \leq_{\mathcal{C}}, \sigma, \leq_{\mathcal{R}}, \mathcal{A})$$

with the following properties:

- A set of concepts defined by the set $\mathcal{C}$ as, $\mathcal{C}=\{\text{Website}(s),\text{Webpages}(\mathbf{w}),\text{WebObjects}(\mathbf{x}),\text{WebConcepts}(\mathbf{c})\}$
- The concept hierarchy (or organizational part-of hierarchy) $\leq_{\mathcal{C}}=\{s,w_i,x_j,c_k\}$, $\forall w_i \in \mathbf{w}, x_j \in \mathbf{x}, c_k \in \mathbf{c}$. This hierarchy can be interpreted that a given Website (s) is represented by Webpages ($\mathbf{w}$), each Webpage ($w_i \in \mathbf{w}$) can be represented as the composition of WebObjects ($\mathbf{x}$), and each WebObject (x_j) is characterized by a composition of WebConcepts ($\mathbf{c}$).
- Set of part-of relations $\mathcal{R}=\{r_1,r_2,r_3\}$ where r_1=site-has-page, r_2=page-has-object, and r_3=object-has-WebConcepts{⁎⁎}.
- The signature σ is defined for relations $r_i \in \mathcal{R}$ according to $\sigma(r_1)=(s,w_i)$, $\sigma(r_2)=(w_i,x_j)$, $\sigma(r_1)=(x_j,c_k)$.
- Relations' hierarchy ($\leq_{\mathcal{R}}$) in this case will be represented by the identity, given the flat relation between concepts.
- Logical axioms are represented by the empty set $\mathcal{A}=\emptyset$.

In this case, Logical axioms are not necessary given that the ontology concepts interaction will be considered as static, whose representation is sufficiently formalized by using an Resource Description Framework (RDF) (Klyne and Carroll, 2004) schema. Further developments on using the extracted Key Objects for decision making, such as the introduced by Chambers et al. (2006) could be considered as future work.

Previous ontology can be described as a graphical representation, where the Website, Webpages, Webobjects, and their WebConcepts are related according to the relation set $\mathcal{R}$ and their hierarchy $\leq_{\mathcal{C}}$, as shown in Fig. 1.

The proposed ontology sets a common ground to characterize each object, independently from the original format from which it was created. By using this ontology, it is possible to make a pairwise

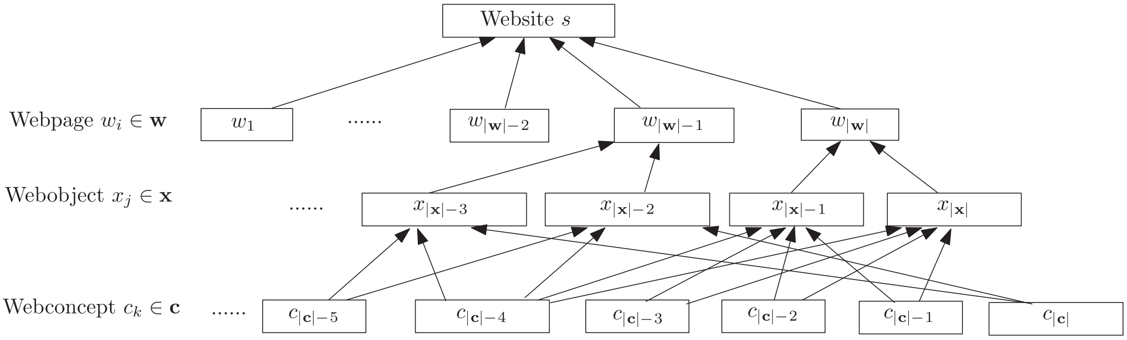


Fig. 1. Graphical representation of the proposed ontology.

conceptual comparison between objects, disregarding their original format. Overall, WebConcepts can be represented as keywords used by the Website administrator to define a given object. Likewise, WebMetaConcepts will be considered as groups of keywords or categories, associated with a more general concept than the one's used in the set of WebConcepts. However, WebMetaConcepts will not be considered in this work as part of the ontology and will be used as categories.

The structure depicted in Fig. 1 can be represented as a Directed Acyclic Graph which leads to a simple model of a given Website. This model allows to reflect the main interactions between the ontology's components, such as the hierarchy between Websites, Webpages, WebObjects, and WebConcepts.

3.2. Comparing WebObjects

In order to compare two WebObjects, we consider that each of them is represented by a group of concepts that defines a given object's content. Considering this shared representation among all objects, a distance for comparing objects $do: |\mathbf{c}| \times |\mathbf{c}| \rightarrow [0,1]$ is proposed by using a given distance measure (e.g. an edit distance).

Given two objects x_i and x_j such that $|x_i| = N$ and $|x_j| = M$, where $N,M \geq 0 \wedge N \leq M$, and given $x_i \rightarrow c_k^1, k \in \{1,\ldots,M\}$ as the kth concept of the object x_i. We need to find a representation in which all WebObjects can be compared. A compare function that could be used, but not limited to, is based on the edit distance between the pairwise alignment between WebConcepts of objects x_i and $x_j, \forall x_i, x_j \in \mathbf{x}, i \neq j$ represented by Algorithm 1.

Algorithm 1. Pairwise Concept Alignment Between Two WebObjects.

```
Require: $x_i, x_j \in \mathbf{x}, \tau$
Ensure: Aligned objects $\{x_i, x_j\}$
seq($x_i$,$x_j$) ← 0
for $c_k \in \{x_i \rightarrow c\}$ do
  for $c_l \in \{x_j \rightarrow c\}$ do
    if $c_k$.Equals($c_l$) then
      seq($x_i$,$x_j$) ← seq($x_i$,$x_j$) + 1
    else if $c_k$. Synonym($c_l$) then
      seq($x_i$,$x_j$) ← seq($x_i$,$x_j$) + 0.5
    end if
    if seq($x_i$,$x_j$) > $\tau$ then
      align($x_i$,$x_j$,$k$,$l$) ← Pair concept $c_k$ with $c_l$ for both $x_i$
      and $x_j$ vectors
    end if
  end for
end for
```

Once all the concepts are paired, the object concepts are ordered in such a way that every concept is in the same relative position in relation to each object. Then a string that consists of a symbol representing each concept is created to represent each object. This string has the structure shown in the expression (1),

$$x = WebConcept_1, \ldots, WebConcept_N \Rightarrow x = c_1, \ldots, c_N \tag{1}$$

where $c_k \in \{x_i \rightarrow c\}$ represents a set of all concepts in object x_i. *Equals* and *Synonym* functions are defined according to a word comparison whose outputs are `True` if compared words are equals or synonyms, or its output is otherwise `False`.

As objects are characterized by a sequence of different symbols each representing a certain category, two objects may be compared by using an edit distance. The differences noted by using this distance will introduce a notion of conceptual similarity between each object as their symbols represent their fundamental concepts. For this, the idea is to pair the most similar concepts between each object and then compare the objects based on a comparison of this paring.

3.3. Sessionization and approximated time spent in WebObject

The methodology gathers the information related to the user behavior from Weblog analysis through the application of a sessionization process to the Weblog. The process applied is taken from Velásquez (to appear), where he combines the approach proposed by Berendt and Spiliopoulou (2001), with a stemming process (Porter, 1980) of the Website.

A normal Weblog considers, among other information, the page a host requested and a timestamp for the request. By reconstructing the user sessions, it is possible to determine how much time each user spends on a given Webpage. However, it is not possible to determine how much time that user spends in a certain object within that page.

The analysis should be made under the assumption that every user spends an equal amount of time in each object that defines a page. If this assumption is made, the analysis that could define Website Key Objects would be merely the analysis of pages browsed by users and a definition drawn on the basis of the most popular objects. To avoid this, an approximation of the time spent by each user is obtained by making a survey over a controlled group of users. The purpose of this survey is to analyze which objects where more appealing to users in each page, so a grade was awarded to every object in a given Webpage.

This survey delivers an approximation of the objects most interesting within a certain page for each user. Using this and weighing it with the time spent by every user on that particular page gives an approximation of the time spent by users in every object within a Webpage.

3.4. Important Object Vector

Following previous works (Velásquez, to appear), an *Object Visitor Vector* (OVV) whose components store the objects visited and time spent by the user during his/her session, can be defined as follows,

Definition 5 (*Object Visitor Vector (OVV)*).

$\theta = [(\vec{\mathbf{x}}_1, \vec{T}_1), \ldots, (\vec{\mathbf{x}}_n, \vec{T}_n)]$

where $\vec{\mathbf{x}}_i = (x_k^i, \ldots, x_l^i)$ is the list of objects belonging to the ith page visited and $\vec{T}_i = (t_k^i, \ldots, t_l^i)$ the respective percentage of time spent by the user during the session seeing each object.

By selecting the ι objects from OOV, the Important Object Vector (IOV) is created as shown in Definition 6.

Definition 6 (*Important Object Vector (IOV)*).

$\Psi_\iota(\theta) = [(\lambda_1, \mu_1), \ldots, (\lambda_\iota, \mu_\iota)]$

where $(\lambda_\iota, \mu_\iota)$ is the component that represents the ιth most important objects and the percentage of time spent on it by session.

Let δ, γ be two IOV, a similarity measure between them is calculated by using the following expression:

$$so(\delta,\gamma) = \frac{1}{\iota}\left(\sum_{k=1}^{\iota} \min\left\{\frac{\mu_k^\delta}{\mu_k^\gamma}, \frac{\mu_k^\gamma}{\mu_k^\delta}\right\} * do(\lambda_k^\delta, \lambda_k^\gamma)\right) \quad (2)$$

where $do : |\mathbf{c}| \times |\mathbf{c}| \rightarrow [0,1]$ is the similarity between two WebObjects (e.g. an edit distance).

3.5. User sessions clustering

Different clustering techniques can be applied to create user session clustering. In this work, two algorithms will be used and will be cross-checked to prove that the created clusters are similar. These algorithms are Kohonen's Self-Organizing Feature Maps (SOFM) (Kohonen et al., 2001) and k-Means (MacQueen, 1967; Hartigan and Wong, 1979).

The SOFM machine learning algorithm is a special type of neural network where a typically two-dimensional grid of neurons is ordered so it reflects changes made in the n-dimensional vector that neurons represent. In this particular case, these vectors can be considered as IOVs. SOFM works with the concept of neighborhoods among neurons, where within the grid, some neurons are considered as neighbors and furthermore changes in one neuron will affect their neighbors.

SOFM requires rules for updating the weights of neurons, this is achieved generically by rule (3),

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad (3)$$

where $m(*)$ is the weight, $h(*)$ is a monotone decreasing function depending on the radius of the neighborhood, and $x(t)$ is the example presented to the network. In our case, special consideration should be taken for IOVs composed by times and objects. The latter cannot be weighted straightforward, for which a vector of differences between objects in a SOFM's neuron is given (Eq. (4)),

$$D = \{do(x_i, x_j)\}_{i,j=1}^{N,M}, \quad i \neq j \quad (4)$$

3.6. Website Key Object Extraction

Finally, according to previously described terms, the Website Key Object Extraction process is introduced in Algorithm 2.

Algorithm 2. Website WebObject Extraction.

```
Require: $\mathbf{x}, \mathcal{T}$
Ensure: $\preceq_{KO}$
$align(\mathbf{x})$ according to Algorithm 1
Compute $D$ according to Eq. (4)
Using $\mathcal{T} = \{T_1, \ldots, T_n\}$, determine OVVs and IOVs according to Eqs. (5) and (6) respectively
Clusterize objects according to $D$ and similarity measure in Eq. (2)
for each cluster $k \in K$ do
  for each $x_i \in k$ do
    $count_i \leftarrow count_i + 1$
  end for
end for
$\preceq_{KO} \leftarrow$ Ordered list according to $count_i, \forall i \in \{1, \ldots, |\mathbf{x}|\}$
```

As described in Algorithm 2, for a given Website, an initial ontology learning step must be realized. In this work, the core ontology proposed is based on a simple representation of the Website according to their Webpages, WebObjects, and WebConcepts for each object. This can be developed by using both automatic or manual processing, whose structure can be easily represented by an XML schema or an RDF-like ontology representation. Afterwards, all objects are compared with their respective concepts by using a pairwise alignment procedure, and then an edit distance can be computed (e.g. Levenshtein, 1966) and the Levenshtein distance is computed.

Once the distance matrix between all objects is computed, by using the Webpages' Weblogs, all objects are grouped together according to their relevance for the end-users. Finally, after concluding the clustering of different objects, their respective frequency was determined by clustering algorithms. This methodology's performance can be tested against a survey where different end-users vote for the most relevant Objects within each Webpage.

4. Practical application

The site chosen to test the proposed WSKO Extraction approach belongs to a Chilean geographical information systems service provider, known as DMapas.[2] The site is written completely in Spanish and is composed by 27 static Webpages. Its content is represented by free-text, images, and flash animations. Weblogs correspond to the month of June 2007, composed of 31.756 requests. In particular, this site has the following characteristics:

- All pages address different information, and if two pages share similar information it is presented with a different focus.
- The users are interested in a certain set of pages and not interested in the remainder.
- The Website is maintained by a Webmaster who can choose if a page stays in the site based primarily on its success to attract users attention.

Given this, the Website ontological representation of objects, the similarity between them, the sessionization process, the approximate time spent in each object, the clustering process, and the extracted Website Key Objects are presented with their respective results and discussion as follows.

As described in Section 3, for the given Website (in this case DMapas), an initial ontology learning step must be realized. It can be developed by using a manual process performed by Webmasters and experts on this Website. This core ontology was translated into an RDF-like representation. Afterwards, all objects were compared with their respective concepts, which were defined by end-users, by using a pairwise alignment procedure,

[2] DMapas: http://www.dmapas.cl (online: accessed 13 December 2010).

and then the Levenshtein distance is computed. Once the distance matrix between all objects are computed, by using the Webpages' Weblogs, all objects are grouped together according to their relevance according to end-users. Finally, their frequency on different clusters determined by clustering algorithms, and Website Key Objects can were retrieved.

The complete evaluation of previously described steps, as well as results obtained, and discussion of relevant points, are extensively presented in the following section.

4.1. Site objects and ontological representation

The site has 40 objects, out of which 26 are composed of free text within tables, 11 as images and three as flash animations. Three hundred and forty-four WebConcepts were associated with these objects and the WebConcepts categorized into one of the 12 categories created (WebMetaConcepts). Depending on the context in which an object is positioned, two of their defining concepts can belong to different categories even if they are identical.

The Web ontology is represented by an XMLschema, whose structure is based on the ontology for the DMapas Website (Fig. 2). An example of this representation is presented as follows.

```
 < ?xml version = }1.0} encoding = }UTF − 8} ? >
 < xsd : schema
targetNamespace =  }http : //www.dmapas.com/core}
  xmlns : wko = }http : //www.dmapas.com/core}
  xmlns : xsd = }http : //www.w3.org/2001/XMLSchema}
  elementFormDefault = }qualified} attributeFormDefault = }unqualified} >

   < xsd : complexType name = }object} >
   < xs : attribute name = }Cartography1} type = }xs : String}
    use = }required}/ >
    < xs : attribute name = }objectType} type = }xs : String}
    use = }required}/ >
    < xsd : sequence >
      < xsd : element name = }WebConcept} minOcurrs = }1}maxOcurrs = }unbounded} >
        < xsd : complexType >
          < xsd : simpleContent >
            < xsd : extension base = }xsd : String} >
              < xsd : attribute name = }Information}
                type = }xsd : string}/ >
            < /xsd : extension >
          < /xsd : simpleContent >
        < /xsd : complexType >
      < /xsd : element >
      < xsd : element name = }WebConcept} minOcurrs = }1}maxOcurrs = }unbounded} >
        < xsd : complexType >
          < xsd : simpleContent >
            < xsd : extension base = }xsd : String} >
              < xsd : attribute name = }GIS}
                type = }xsd : string}/ >
            < /xsd : extension >
          < /xsd : simpleContent >
        < /xsd : complexType >
        < /xsd : element >
        < /xsd : sequence >
      < /xsd : complexType >
 < /xsd : schema >
```

This representation was extended over the 40 WebObjects in all 27 Webpages of the selected Website. This XML representation supported the core ontology $\mathcal{O}$ described in Section 3.1.2, and was built completely manually. For further applications in large-scale Websites, automatic ontology learning and extraction (Tsoi et al., 2009), among other ontology engineering techniques (Poon and Domingos, to appear), could be considered as an extension of this work.

A proof of concept was performed before the clustering process began proving that the similarity measure created was suitable enough to compare two objects at a conceptual level. This proof took into consideration a dataset of four objects, two of which where flash animations depicting demos of Geographic Information Systems (GIS) solutions that the Company provides, the remaining two where free text within tables. The first describes GIS from a technical point of view, defining what they are and how they operate. The other shows information about the company, their owners, and employees.

Before the tests were performed, the experts created Table 1, which shows how similar these objects are from the users' point of view. The tests that were performed given as results in Table 1, consisted of similarities in the range [0,1] where $do(x_i,x_j)=1$ means that objects x_i and x_j are identical.

By comparing the results from Table 1, it is easy to see that the calculations are correct, and provides results according to the conceptual similarities given by experts.

4.2. Sessionization process

The sessionization process was implemented in the PHP language using a reactive strategy (Velásquez and Palade, 2008)

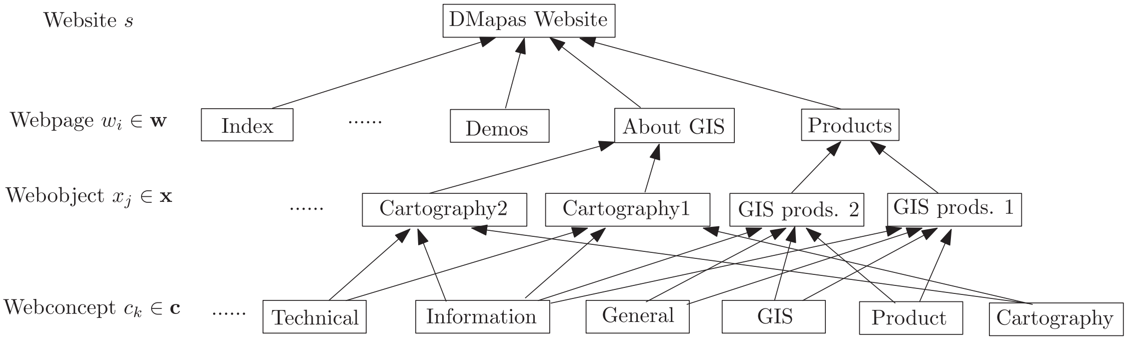


Fig. 2. Graphical representation of the proposed ontology for the DMapas Website.

Table 1
Comparison between selected objects from the users point of view.

Object x_1	Object x_2	Relation	$do(x_1,x_2)$
Demo 1	Demo 1	Identical	1
Demo 1	Demo 2	Very similar	0.929
Demo 1	About GIS	Similar	0.6
Demo 1	About the company	Not similar	0.286
Demo 2	Demo 2	Identical	1
Demo 2	About GIS	Similar	0.6
Demo 2	About the company	Not similar	0.286
About GIS	About GIS	Identical	1
About GIS	About the company	Totally different	0
About the company	About the company	Identical	1

Table 2
Estimating the WebObject importance from the user point of view.

Page ID	x_1	x_2	x_3	x_4	Favorite	*Score* (x_1)	*Score* (x_2)	*Score* (x_3)	*Score* (x_4)
4	4	5	–	–	4	5	5	–	–
17	15	16	–	–	15	8	2	–	–
20	19	22	21	20	19	5	3	1	1
21	23	24	–	–	23	7	3	–	–
22	25	26	–	–	25	8	2	–	–
23	28	27	–	–	28	6	4	–	–
24	29	30	31	–	29	6	2	2	–
26	33	35	34	–	33	7	2	1	–
29	38	39	–	–	38	5	5	–	–

which was limited to 30 min per session. The process considered replacing the pages with objects leading to an expansion of the Weblog. This was considered due to the fact that requests for one page can represent one or more objects with their corresponding times spent by users, provided by the results of the survey. The result of the sessionization process provided 12.608 sessions. After this, all the objects with navigation time zero were eliminated, resulting in 5.866 sessions over 19.282 requests. This provided an average of 3.29 objects per session.

To create the IOV, the number of requests considered in each session was calculated by taking the mean number of objects per session and adding the standard deviation which lead to all sessions having six or more requests. Only 815 out of the 12.608 sessions applied to this constraint, which was softened to all sessions consisting of five or more requests. Therefore, the final number of used sessions was 1.463.

4.3. Approximated time spent in each WebObject

A survey was taken over a group of 10 users, two of them were the experts who had an extensive knowledge of the Website, four users were DMapas customers who had visited the Website but had a partial knowledge of it, the rest were new users who had seen the site for the first time. This ensured that a diverse users were surveyed. Before the survey was taken, every user was introduced to every object in the Web page, and all their characteristics were clearly defined and explained.

After the objects were introduced, three questions were asked. The first question for each user was “Which object was the most appealing to you within the whole Website?”, the second question was to make a top 20 list with all the objects of the site being the first most appealing for them. Finally for each page which had two or more objects, 10 points must be awarded between all of them having the most points the object which was the most appealing to the user within a certain page. Table 2 is an example about the answers given by one user. It was applied on Webpages with two or more objects, because in the case one of object per page, the assumption is the object concentrates its whole attention of the user during his/her visit to the page, i.e., the time spent in the page is the same that in the object.

Information in Table 2 is interpreted as follows: one page can contain two or more objects (until four), the column *Favorite* shows the favorite object. Regarding the last four columns, each score was given by users. For instance, the page ID 17 has the objects IDs 15 and 16. From these two objects, the favorite one for the user is the ID 15 because 8 points were given, where only 2 points were given to object ID 16.

By using the survey's information, the average of points assigned per WebObject were calculated. This information was used to distribute the time spent per page by user, obtained from Weblogs, over all objects during the user session according to their weights. After applying the sessionization process and by using the objects' weights, both Object Visited Vectors (OVVs) and Important Object Vectors (IOVs) can be created and used as input for clustering algorithms.

4.4. Clustering process

The clustering algorithms developed in this work were SOFM and *k*-Means. Both algorithms were implemented over an `Intel T2300 Core Duo` running at `1.63 GHz` with `1GB RAM` on `Windows XP` operational system. The SOFM network was implemented in `Python` using {12 × 12,14 × 14,18 × 18,24 × 24} neurons in the chart with a toroidal topology, The *k*-Means algorithm was implemented in `Java 1.42` using the number of clusters extracted from the best SOFM representation. In terms of computational time, by using this architecture and programming languages, the SOFM algorithm ran approximately 2 h, while *k*-Means took approximately 15 min.

4.4.1. Clustering results

The main algorithm for obtaining the clustering over WebObjects was SOFM. However, these results were cross-checked with *k*-Means. By using SOFM, in the best clustering approach for the 24 × 24 neurons architecture, nine clusters were clearly identified as illustrated in Fig. 3.

WebObjects that belonged to each cluster for the 24 × 24 neurons SOFM can be seen in Table 3.

It is important to notice that each cluster was formed by either one or two neurons in the chart so they may have 5 or 10 objects that represent them. In some cases an object can appear twice in a cluster. Then each of the objects was labelled with its main concept, and then, a label for each cluster was created, which described the main content for each cluster shown in Table 3.

Furthermore, it can be inferred from Table 3 that all objects in cluster 9 are related to technical information about GIS and cartography that DMapas company provides. Analogously, the remained clusters where labelled according to the main content of the objects they contain. This result was checked using the *k*-Means algorithm, which discovered five well-defined clusters, presented in Table 4.

Five clusters ($k=5$) were chosen according to the evaluation of $k \in [2,12]$. The evaluation began $k = 12$, which is the number of clusters generated by the SOFM algorithm, plus a slack of +3. Then, the k parameter evaluated was decreased until all the clusters found were acceptable in terms of their interpretation.

4.5. Website Key Objects

In order to discover the Website Key Objects, all objects present in every cluster were counted. These results are presented in Table 5, where objects which appeared the most were considered to be the Website Key Objects.

It can be seen that from top 10 objects, seven are presented in text format, two as flash animations, and only one corresponds to an image. They focus on a small part of the company's site omitting a large quantity of information that administrators assumed to be very useful to users.

Table 3
Similarity measures between objects.

Cluster	Objects	Labels
1	{3,4,5,11,32}	Cartography
2	{11,12,19,33,35}	Geobusiness
3	{7,8,9,11,38}	Demos
4	{4 × (2),11,33,34,35,38 × (2),39}	Geobusiness and GIS
5	{1,3,4,12,32}	The Company and Cartography
6	{3,4,5,7 × (2),8 × (2),11,39}	Demos and Cartography
7	{2,7,8,11,38}	Demos and GIS
8	{1 × (2),2 × (2),12 × (2),13,38 × (2),39}	The Company and GIS
9	{2 × (2),3,4,11 × (2),32,38 × (2),39}	Cartography and GIS

Table 4
Labeled clusters discovered with *k*-Means.

Cluster	Objects
1	Cartography
2	Geobusiness
3	Cartography and GIS
4	Geobusiness and GIS
5	Demos and Cartography

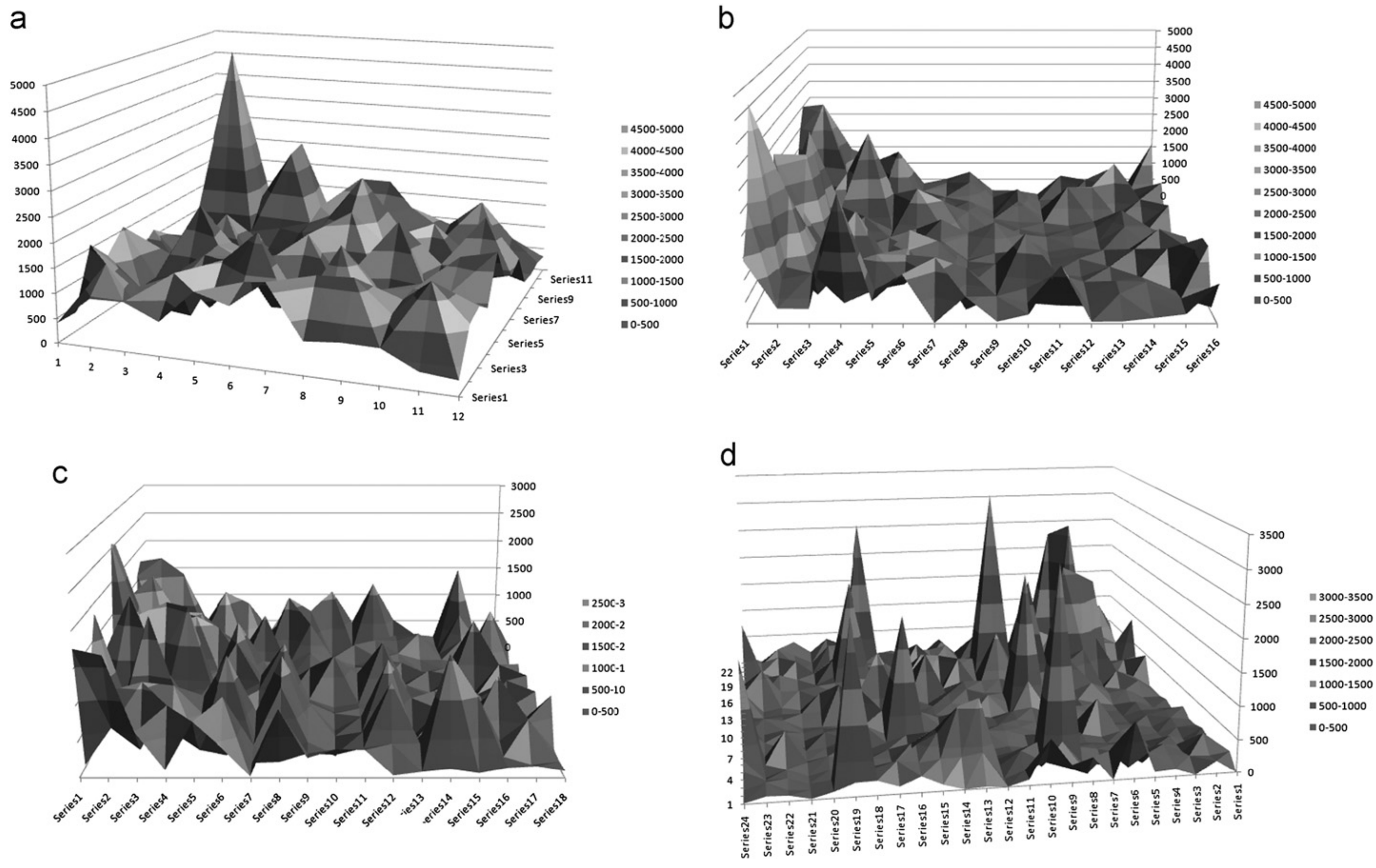


Fig. 3. Self-Organizing Feature Maps clustering results for grids of 12 × 12 neurons (a), 14 × 14 neurons (b), 18 × 18 neurons (c), and 24 × 24 neurons (d).

Table 5
Website Key Objects.

Object	Type	Concept	Count
Index	Flash	General information about DMaps products	28
Cartography 1	Text	Technical information about Cartography	23
GIS products 1	Text	General information about GIS products	17
GIS products 2	Text	General information about GIS products	14
Cartography 2	Text	Technical information about Cartography	14
Cartography products	Text	Information about Cartography provided by DMaps	13
About GIS	Text	General information about GIS systems	10
Geobusiness 3	Image	Information about Geobusiness applications	9
Cartography 3	Text	Technical information about Cartography	9
Demo 2	Flash	Demonstration of a GIS application	10

Table 6
Effectiveness of the extracted Website Key Objects tested.

#	Including the Website Key Object?	Acceptability opinion				
		Irrelevant	Moderately irrelevant	Some information	Moderately relevant	Relevant
1	Yes	–	–	1	7	2
2	Yes	–	–	1	6	3
3	No	5	3	2	–	–
4	No	–	7	3	–	–
5	No	2	8	–	–	–

Table 7
Most appealing objects for Web users determined by the number of appearances in the overall survey evaluation.

Object	Number of appearances
Index	10
Geobusiness	9
Demo 2	8
About the company	7
About GIS	7
SIG products 2	7
Cartography 2	7
Cartography 1	6
Demo 1	6
Cartography products	5

4.6. *Verification of Website Key Objects*

The accuracy of the extracted Key Objects by the algorithm was proved by comparing the results with a survey taken by a controlled group of users. They looked at five pages, two of them with Website Key Objects, and the others were extracted randomly from the site. Next, users were asked to answer which of the five Webpages shown was the most appealing for them. Table 6 shows the results of the Website Key Object effectiveness.

Web users showed a positive receptivity towards pages containing Key Objects, considering them explicitly interesting and with relevant information. This means that for these users, WebObjects showed significant information, demonstrating that Website Key Objects can be used for attracting the Web user's attention.

Finally, as a benchmark, Web users were asked about which objects shown in Webpages can be considered as more relevant. These results are shown in Table 7.

When comparing results of this survey with the algorithm's extracted Key Objects, a difference of only two objects is detected, which leads to a match of 80% between the algorithm's detected objects and those preferred by the controlled group. If the analysis is extended to the top 15 objects, the accuracy rises to 87%, which shows a positive relation between the algorithm and the survey results.

5. Conclusions

In this work, a methodology for identifying Website Key Objects is introduced. Website Key Objects are the most appealing objects for users within a Website. This methodology is a generalization of a prior developed by Velásquez et al. (2005) for identifying Website Keywords. Our approach is based on the fact that there is a correlation between the time spent by a user in a certain page during a session and the interest the user has in its content.

In order to develop this methodology a definition of a WebObject was created, and particularly a definition for Website Key Objects, which are those objects in a Website that drives the attention of users. The definition of these objects enables the characterization of the conceptual content represented by a simple core ontology. The Website Key Object Extraction Problem (WSKOP) is aimed towards the definition of a new relation between the core ontology's WebObjects. This relation is an ordered list of Website Key Objects, according to the Web user preferences.

This characterization delivers a common ground over which any object can be defined with no restrictions regarding the format in which it is presented to the user. This allows a clear conceptual definition for each object. Additionally, by using this ontology a similarity measure was introduced which enables a quantifiable conceptual comparison between two WebObjects, even though these might not share the same format.

The proposed methodology was applied in a real Website, for which its Key Objects were extracted, whose results were compared with Web user surveys. Results showed that our approach is similar by at least 80% of those Website Key Objects described by Web users, which leads us to the conclusion that an automatic approach which is scalable and easy to implement could enhance the Webmaster's labor on what information and what format should be considered to update a given Website.

The knowledge acquired from the application of the methodology to a real site allows a Webmaster to know the preferences of the user base. It also enables the possibility of enhancing the Website by empowering the information that users are looking for and also presenting it in an appealing format.

5.1. *Future work*

The methodology to discovering Website Key Objects relies heavily on two factors, an ontology used to define WebObjects and an approximation used to determine how much time a user spent looking for a certain object within a Webpage. By using privacy-preserving data mining, an end-user profiling algorithm could be included to improve the modelling and inclusion of different types of users to extract a more representative list of Key Objects. Also, the information gathered by eye-tracking devises could be included as Web usage data to extend the possibility to analyze more complex Websites.

Also, in this work, all metadata was created manually mainly taking into consideration that the test Website was composed of static pages and that its cardinality was small. The model used in this work is a relatively basic and easy to implement ontology.

The strong development of metadata applied to the Web allows for the creation of more advanced metadata models which enables the creation of a more complex and expressive ontology (e.g. considering the hierarchy between WebConcepts). This would result in a more precise definition of an object which would lead to a more precise comparison between two objects. Furthermore, in this work metadata was incorporated manually to the test Website. This could be made in reasonable time because the test Website of a limited size and of static nature. If a larger or dynamical site were used this task would have been time consuming.

Acknowledgments

This work was supported partially by the FONDEF project DO8I-1015 and the Millennium Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16). Authors would like to thank anonymous reviewers for their interesting and useful comments on the preliminary version of this article.

References

Baeza-Yates, R.A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Berendt, B., Hotho, A., Stumme, G., 2002. Towards semantic web mining. In: ISWC '02: Proceedings of the First International Semantic Web Conference on the Semantic Web. Springer-Verlag, London, UK, pp. 264–278.

Berendt, B., Spiliopoulou, M., 2001. Analysis of navigation behavior in web sites integrating multiple information systems. The VLDB Journal 9, 56–75.

Berry, R., 1998. Common user access—a consistent and usable human–computer interface for the saa environments. IBM Systems Journal 3 (27), 281–300.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Burget, R., Rudolfová, I., 2009. Web page element classification based on visual features. In: Proceedings of the 1st Asian Conference on Intelligent Information and Database Systems ACIIDS 2009, Dong Hoi, VN. IEEE CS.

Chambers, N., Allen, J., Galescu, L., Jung, H., Taysom, W., 2006. Using semantics to identify web objects. In: AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence. AAAI Press, pp. 1259–1264.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. Ai Magazine 17, 37–54.

Gao, X., Murugesan, S., Lo, B., 2005. Extraction of keyterms by simple text mining for business information retrieval. In: Proceedings of the International Conference on e-Business Engineering (ICEBE05). IEEE Computer Society, pp. 332–339.

Hartigan, J., Wong, M., 1979. A k-means clustering algorithm. Journal of the Applied Statistics 28, 100–108.

Jr, A.P., Ziviani, N., 2004. Retrieving similar documents from the web. Journal of Web Engineering 2 (4), 247–261.

Khare, R., Çelik, T., 2006. Microformats: a pragmatic path to the semantic web. In: Proceedings of the 15th International Conference on World Wide Web, WWW '06. ACM, New York, NY, USA, pp. 865–866.

Klyne, G., Carroll, J.J. (Eds.), 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. World Wide Web Consortium.

Kohonen, T., Schroeder, M.R., Huang, T.S. (Eds.), 2001. Self-Organizing Maps. Springer-Verlag New York, Inc, Secaucus, NJ, USA.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 705–710.

Li, Y., Zhong, N., 2003. Ontology-based web mining model: representations of user profiles. In: WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence. IEEE Computer Society, Washington, DC, USA, 96 pp.

Li, Y., Zhong, N., 2006. Mining ontology for automatically acquiring web user information needs. IEEE Transactions on Knowledge and Data Engineering 18 (4), 554–568.

Li, Y., Zhong, N., 2007. Ontology based web mining for information gathering In: WImBI'06: Proceedings of the 1st WICI International Conference on Web Intelligence Meets Brain Informatics. Springer-Verlag, Berlin, Heidelberg, pp. 406–427.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics Probability, vol. 1. University of California, pp. 281–297.

Maedche, A., Staab, S., 2001. Ontology learning for the semantic web. IEEE Intelligent Systems 16 (2), 72–79.

Nielsen, J., 2006. Prioritizing Web Usability. New Riders, Berkley.

Paaß, G., Kindermann, J., Leopold, E., 2004. Learning prototype ontologies by hierarchical latent semantic analysis. In: Abecker, A., Bickel, S., Brefeld, U., Drost, I., Henze, N., Herden, O., Minor, M., Scheffer, T., Stojanovic, L., Weibelzahl, S. (Eds.), LWA. Humbold-Universität Berlin, pp. 193–205.

Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S., 1998. Latent semantic indexing: a probabilistic analysis. In: PODS '98: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. ACM, New York, NY, USA, pp. 159–168.

Plumbaum, T., Stelter, T., Korth, A., 2009. Semantic web usage mining: using semantics to understand user intentions. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (Eds.), User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science, vol. 5535. Springer Berlin/Heidelberg, pp. 391–396.

Poon, H., Domingos, P. Unsupervised ontology induction from text. In: ACL'10: Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics, to appear, http://portal.acm.org/citation.cfm?id=1858681.1858712.

Porter, M.F., 1980. An algorithm for suffix stripping. Program. Automated Library and Information Systems 3 (14), 130–137.

Sahami, M., 2006. Mining the web to determine similarity between words, objects, and communities. In: Sutcliffe, G., Goebel, R. (Eds.), FLAIRS Conference. AAAI Press, pp. 14–19.

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM archive 18 (11), 613–620.

Stumme, G., Hotho, A., Berendt, B., 2006. Semantic web mining—state of the art and future directions. Journal of Web Semantics 4 (2), 124–143.

Tiwary, U., Siddiqui, T., Radhakrishna, M., Tiwari, M., 2009. Web content mining focused on named objects. In: Proceedings of the First International Conference on Intelligent Human Computer Interaction (IHCI 2009).

Tsoi, L.C., Patel, R., Zhao, W., Zheng, W.J., 2009. Text-mining approach to evaluate terms for ontology development. Journal of Biomedical Informatics 42 (5), 824–830.

Urbansky, D., Feldmann, M., Thom, J., 2008. Entity extraction from the web with webknox. In: Proceedings of the 13th Australasian Document Computing Symposium, December 8, 2008, Hobart, Australia.

Velásquez, J.D. Web site keywords: a methodology for improving gradually the web site text content. Intelligent Data Analysis 15 (1), to appear.

Velásquez, J.D., Palade, V., 2008. Adaptive Web Sites: a Knowledge Extraction from Web Data Approach. IOS Press.

Velásquez, J.D., Ríos, S., Bassi, A., Yasuda, H., Aoki, T., 2005. Towards the identification of keywords in the web site text context: a methodological approach. Journal of Web Information Systems 1 (2), 11–15.

Wang, J., Peng, H., HU, J., 2005. Automatic keyphrases extraction from document using backpropagation. In: Proceedings of the 4th International Conference on Machine Learning and Cybernetics, pp. 3770–3774.

Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S., 2010. Learning subsumption hierarchies of ontology concepts from texts. Web Intelligence and Agent Systems 8 (1), 37–51.