7

Combining SOMs and Ontologies for Effective Web Site Mining

Dimitris Petrilis and Constantin Halatsis
National and Kapodistrian University of Athens
Greece

1. Introduction

The Internet since the late 90s, when it became mainstream, has dramatically changed the way people work, communicate, get educated, socialize and stay informed about current affairs. According to WorldWideWebSize.com (WorldWideWebSize.com, 2010) the Indexed Web contains at least 14.56 billion pages as of August 2010. The estimated minimal size of indexed World Wide Web is based on the estimations of the numbers of pages indexed by Google, Bing, Yahoo Search and Ask. In addition according to Royal Pingdom (Royal Pingdom, 2009) on December 2009 there were 234 million web sites and on September 2009 1.73 billion Internet users.

In the past few years we have also witnessed a new explosion in the usage of the World Wide Web (WWW) with what is commonly referred to as Web 2.0. The term Web 2.0 refers to the set of web sites whose contents are modified by visitor contributions and not only by the webmaster. This includes social networks such as Facebook (Facebook), LinkedIn (LinkedIn), Twitter (Twitter), MySpace (MySpace) as well as blogs and web sites where visitors can share pictures, such as Picasa (Picasa), or video files, such as YouTube (YouTube). The WWW has become an extremely interactive form of communication. The following statistics from Royal Pingdom (Royal Pingdom, 2009) illustrate the rapid penetration of Social Media Networks to everyday life:

- 126 million – the number of blogs on the Internet (as tracked by BlogPulse)
- 27.3 million – the number of tweets on Twitter per day (November, 2009)
- 350 million – people on Facebook
- 50% - percentage of Facebook users that log in every day
- 500,000 – the number of Facebook applications

Bearing in mind these huge numbers it is easy to image the massive amount of information available on the World Wide Web.

On the other hand significant advances in networking technology (such as very fast Internet connections) and search engines have created an impatient Internet culture. Internet users expect to be able to find the information they seek within seconds. Accessing the Internet is no longer an extracurricular activity that people perform at their spare time at home or a business tool used only by large corporations. It has become a common part of our everyday life. Many people have always an Internet connection available to them in their office and home equipment as well as through mobile devices.

Internet visitors are expecting to find information quickly and easily. They can be very harsh in the sense that they will not give a web site a second chance if they cannot find something interesting within the first few seconds of browsing. At the same time web sites are packed with information and hence presenting to every visitor the right information has become very complex. This has created two main challenges when maintaining a web site:

- Attracting visitors, i.e. getting people to visit the web site.
- Keeping visitors on the web site long enough so that the objective of the site can be achieved, e.g. if we are talking about an Internet store to make a sale.

This chapter deals with the second challenge, how to help web site visitors find information quickly and effectively by using clustering techniques. There is a plethora of methods for clustering web pages. These tools fall under a wider category of data mining called Web mining. According to Cooley (Cooley et al., 1997) Web mining is the application of data mining techniques to the World Wide Web. Their limitation is that they typically deal either with the content or the context of the web site. Cooley (Cooley et al., 1997) recognises that the term web mining is used in two different ways:

- Web content mining – information discovery from sources across the World Wide Web.
- Web usage mining – mining for user browsing and access patterns. In this paper we also refer to web usage mining as context mining.

The content of a web site can be analysed by examining the underlying source code of its web pages. This includes the text, images, sounds and videos that are included in the source code. In other words the content of a web site consists of whatever is presented to the visitor. In the scope of this chapter we examine the text that is presented to the visitor and not the multimedia content. Content mining techniques can be utilised in order to propose to the visitors of a web site similar web page(s) to the one that they are currently accessing. Metrics such as the most frequently occurring words can be used to determine the content of the web site (Petrilis & Halatsis, 2008). In this chapter we introduce an ontology-based approach for determining the content of the web site. However, it must be noted that the focus of this chapter is on the usage of SOMs and not on the usage of ontologies. Additional research is required for establishing the additional value of using ontologies for the purpose of context mining.

The page currently being viewed may be a good indicator of what the visitor is looking for, however it ignores the navigation patterns of previous visitors. The aim of context mining techniques is to identify hidden relationships between web pages by analysing the sequence of past visits. It is based on the assumption that pages that were viewed in some sequence by a past visitor are somehow related. Typically context miming is applied on the access-logs of web sites. The web server that is hosting a web site typically records important information about each visitor access. This information is stored in files called access logs. The most common data that can be found in access-logs is the following:

- the IP address of the visitor
- the time and date of access
- the time zone of the visitor in relation to the time zone of the web server hosting the web page
- the size of the web page
- the location (URL) of the web page that the visitor attempted to access
- an indication on whether the attempt to access the web page was successful
- the protocol and access method used

- the referrer (i.e. the web page that referred the visitor to the current page) and
- the cookie identifier

Clustering algorithms can be used to identify web pages that visitors typically visit on the same session (a series of web page accesses by the same visitor). The output of the clustering algorithms can be used to dynamically propose pages to current visitors of the web site.

The problem with most web mining clustering techniques is that they focus on either content, such as WEBSOM (Lagus et al, 2004), or context mining (Merelo et al, 2004). This way important data regarding the web site is ignored during processing. The combination of both content and context mining using SOMs can yield better results (Petrilis & Halatsis, 2008). However, when this analysis takes place in two discreet steps then it becomes difficult to interpret the results and to combine them so that effective recommendations can be made. In this chapter we are going to demonstrate how we can achieve better results by producing a single SOM that is the result of both content and context mining into a single step. In addition we are going to examine how the usage of ontologies can improve the results further.

To illustrate our approach and findings we have used the web pages and access-logs of the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens.

2. Kohonen's self-organising maps

It is not in the scope of this chapter to provide a detailed definition of Kohonen's Self-Organising maps since it is assumed that the reader already has some knowledge regarding this unsupervised neural network technique. According to Kohonen (Kohonen, 2001), the SOM in its basic form produces a similarity graph of input data. It converts the nonlinear statistical relationships among high-dimensional data into simple geometric relationships of their image points on a low-dimensional display, usually a regular two-dimensional grid of nodes. As the SOM thereby compresses information while preserving the most important topological and/or metric relationships of the primary data elements on the display, it may also be thought to produce some kind of abstractions. There are many variations of SOMs (Kohonen, 2001) and in the context of this research we are using the basic form that was proposed by Kohonen.

There is a plethora of different software packages that implement different variations of the SOM. In order to perform our research we use SOM_PAK (SOM_PAK and LVQ_PAK). This package includes command-line programs for training and labelling SOMs, and several tools for visualizing it: sammon, for performing a Sammon (Sammon, 1969) projection of data, and umat, for applying the cluster discovery UMatrix (Ultsch, 1993) algorithm. SOM_PAK was developed by Kohonen's research team.

3. Web Mining

The term Web Mining is often subject to confusion as it has been traditionally used to refer to two different areas of data mining:

- Web Usage Mining - the extraction of information by analysing the behaviour of past web site visitors
- Web Content Mining – the extraction of information from the content of the web pages that constitute a web site.

3.1 Web usage mining

Web usage mining, also known as Web Log Mining, refers to the extraction of information from the raw data that is stored in text files located on the web server(s) hosting the web pages of a web site. These files are called access-logs. Typically each entry in the access log is one line in the text file and it represents an attempt to access a file of the web site. Examples of such files include: static html pages, dynamically generated pages, images, videos and sounds amongst others. A typical access log entry can be seen below:

134.150.123.52 - - [19/Aug/2010:15:09:30 +0200] "GET /~petrilis/index.html HTTP/1.0" 200 4518 "http://www2.di.uoa.gr/gr/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)" 62.74.9.240.20893111230291463

The data of this example is explained in the table that follows:

Data Item	Description
134.150.123.52	The IP address of the computer that accessed the page
-	The identification code (in this case none)
-	The user authentication code (in this case none)
[19/Aug/2010:15:09:30 +0200]	The date, time and time zone (in this case 2 hrs ahead of the timezone of the web server hosting the web site) of the access
"GET /~petrilis/index.html HTTP/1.0"	The request type (GET), the web page accessed and the protocol version
200	The server response code (in this case the page was accessed correctly)
4518	The number of bytes transferred
"http://www2.di.uoa.gr/gr/"	The referrer page
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"	The user agent information, i.e. browser information
62.74.9.240.20893111230291463	Cookie string

Table 1. Data contained in an access-log

There is a large number of software solutions that can perform analysis of the access-logs. Most of these perform simple statistical analysis and provide information, such as the most commonly accessed page, the time of the day that the site has more access, etc. For example WebLog Expert (WebLog Expert) provides the following analysis:

- General statistics
- Activity statistics: daily, by hours of the day, by days of the week and by months
- Access statistics: statistics for pages, files, images, directories, queries, entry pages, exit pages, paths through the site, file types and virtual domains
- Information about visitors: hosts, top-level domains, countries, states, cities, organizations, authenticated users

- Referrers: referring sites, URLs, search engines (including information about search phrases and keywords)
- Browsers, operating systems and spiders statistics
- Information about errors: error types, detailed 404 error information
- Tracked files statistics (activity and referrers)
- Support for custom reports

Such information can provide some valuable information but it does not provide true insight on the navigational patterns of the visitors. Using clustering algorithms more in depth analysis can be performed and we can deduce more valuable information. For example we can identify clusters of visitors with similar access patterns. We can subsequently use this information to dynamically identify the most suitable cluster for a visitor based on the first few clicks and recommend to that visitor pages that other visitors from the same cluster also accessed in the past. There are different methods for performing such clustering ranging from simple statistical algorithms, such as the k-means, to neural network techniques, such as the SOM.

3.2 Web content mining

Web content mining is the application of data mining techniques to the content of web pages. It often viewed as a subset of text mining, however this is not completely accurate as web pages often contain multimedia files that also contribute to its content. A simple example of this is YouTube (YouTube) that mainly consists of video files. This is exactly the most important complexity of web content mining, determining the source of the content. The source code of the web pages, stripped of any tags, such as HTML tags, can be used as input (Petrilis & Halatsis, 2008). However, it is easy to see the limitation of such an approach bearing in mind that as we mentioned other types of files are also embedded in web pages. In addition quite often pages are dynamically generated and therefore we do not know their content in advance. Another additional constraint is the sheer volume of data that is often contained within web pages. In this chapter we attempt to address this issue by proposing an ontology based approach for determining the content of the web pages and for creating suitable input for SOM processing. It is not in the scope of this chapter to elaborate on ontology based techniques and this will be the subject of subsequent research by the authors. However, Paragraph 4 provides further details on our approach.

There are different methods that can be used for web content mining. Simple statistical analysis can provide some level of information such as the most popular words in each page or the most frequent words in the set of all the pages comprising the web site. However, this information is of limited use and does not unveil hidden relationships between web pages. Clustering algorithms can be used to unveil more complex relationships among the web pages by identifying clusters of web pages with similar content. This analysis can be used to dynamically propose web pages to visitors. WEBSOM (Lagus et al., 2004) utilises the SOM algorithm to generate a map that displays to the visitor pages of similar content with the page that is currently being viewed. The recommended pages are topographically placed in the map. The closer a recommended page is to the current location of the visitor within the map, the more relevant the recommendation is. A sample of output of WEBSOM can be seen in Figure 1.

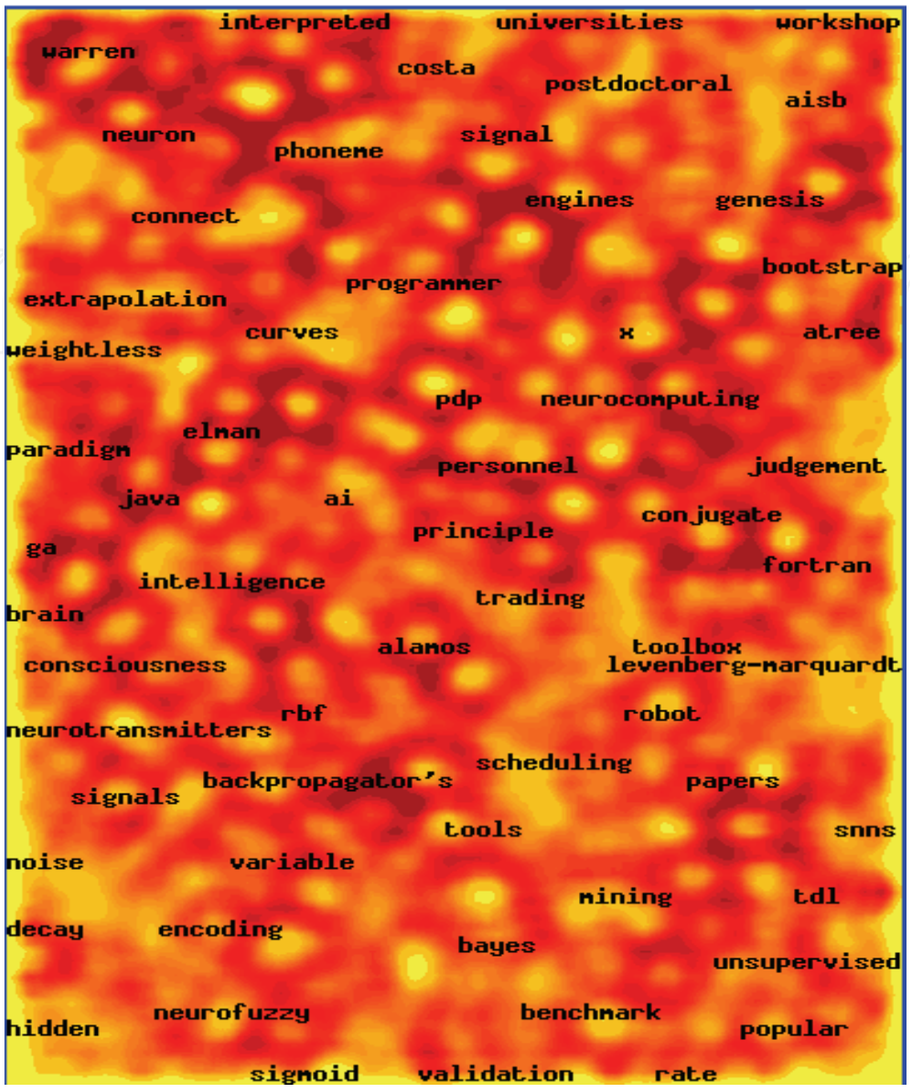


Fig. 1. Example output of WEBSOM

4. Ontology

It is not in the scope of this chapter to provide an in-depth analysis of ontologies and their usage on web mining. However, since a simple ontology has been used to achieve better results in our processing it is useful to provide an overview of ontologies.

Ontology as a term was originally used in philosophy to study the conceptions of reality and the nature of being. Looking at the etymology of the word "ontology", it originates from the Greek word "On" which means "Being". Hence, ontology is the study of "being". Ontology as an explicit discipline was created by the great ancient philosopher Aristotle. According to Gruber (Gennari, 2003) an ontology is an explicit specification of a conceptualization. A "conceptualization" is an abstract, simplified view of the world that we wish to represent for some purpose. According to Katifori (Katifori et al., 2007) it contains the objects, concepts and other entities that are presumed to exist in some area of interest

and the relations that hold them. An ontology is a formal explicit description of concepts in a logical discourse. In ontology concepts are known as classes, the properties of each concept describing various features and attributes of the classes are referred to as slots or properties and the restrictions on the slots as facets. A specific ontology with a set of class instances constitutes a knowledge base.

Ontologies are a very popular tool for adding semantics to web pages in order to facilitate better searching. Luke (Luke et al., 1996) proposes an ontology extension to HTML for exactly that purpose. Berners-Lee (Berners-Lee et al., 2001) suggests the usage of ontologies for enhancing the functioning of the Web with the creation of the Semantic Web of tomorrow. The WWW Consortium (W3C) has created the Resource Description Framework, RDF, a language for encoding knowledge on web pages to make it understandable to electronic agents searching for information. Ontologies are not only used for research purposes but also have many commercial applications. As an example many key players in the WWW, such as Yahoo and Amazon, use ontologies as a means of categorising their web pages.

In the context of the WWW typically the primary use of ontologies is not the description of the domain. It is the definition of the data and its inherent structure so that it can be used more effectively for further processing and analysis. A typical example is the Semantic Web. The goal of the Semantic Web is to make it possible for human beings and software agents to find suitable web content quickly and effectively. The definition of the underlying data itself is not the primary objective.

The focus of our research in the chapter is to achieve better results in clustering web pages by producing a single SOM that is the result of both content and context mining. By introducing the use of a very simple ontology in the content mining part we demonstrate improved results. The tool that was used for creating this simple ontology is Protégé. Protégé is an environment for knowledge-based systems that has been evolving for over a decade (Gruber, 1993). It implements a rich set of knowledge-modelling structures and actions that support the creation, visualization, and manipulation of ontologies in various representation formats. Protégé has been selected because it is one of the most complete packages for the creation of ontologies and at the same time it is very simple to use. In addition a large number of extensions are available (Gruber, 1993). A comprehensive comparison of ontology development environments has been performed by Duineveld (Duineveld et al., 2000).

It is well known and documented that web mining as any other data mining technique can only produce useful results if a suitable data set is used. Hence, it is important to examine the data preparation steps in more detail.

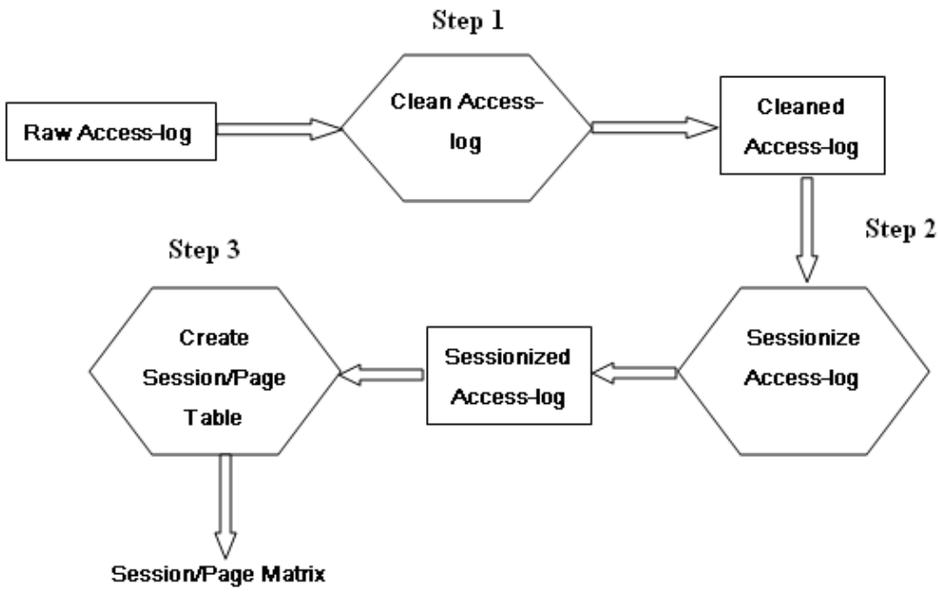
5. Data preparation

As it was previously mentioned the results of any data mining analysis can only be as good as the underlying data. Hence it is important to present the pre-processing steps that are required prior to applying the SOM.

5.1 Data preparation for context mining

As it was previously mentioned web site context mining deals with the analysis of the access-logs that are stored in web servers. Typically the access-logs contain a large amount

of noise. This is data that not only does not add any value to processing but on the contrary skews the results. Each time a visitor accesses a web page, a number of files are being accessed. These may include the main web page (typically HTML), images, videos and audio files. Some of these files, for example a logo that may be present in every web page of the site, generate noise to the access logs. In addition search engines use software agents called web robots that automatically traverse the hyperlink structure of the World Wide Web in an effort to index web pages (Noy & McGuniness, 2001). These software agents perform random accesses to web pages and hence generate access logs entries of no value. Identifying these robot accesses is a difficult task. Another important consideration when processing access-logs is that quite often an IP address does not uniquely identify a visitor. Therefore, we need to introduce the concept of a visitor session. A visitor session for the purposes of our research is a visitor access from a specific IP address within a specific time frame.



	page1	page2	.....	pagen
Session1	1	1	.....	0
Session2	0	0	.....	1
.....	.....	.....	.....	.....
Sessionm	4	0	.....	2

Fig. 2. Data preparation steps for context mining

In order to prepare context related data for input to the SOM the following pre-processing steps were followed that are also depicted in Figure 2:

- Noise Removal - removal of image, video, audio and web robot accesses from the access-logs. It must be noted that in order to simplify the processing all image, video and audio accesses were removed regardless of their content. WumPrep (WumPrep) is used for this purpose. WumPrep is a collection of Perl scripts designed for removing noise from access-logs and preparing them for subsequent processing.

- Session Identification –WumPrep was used to identify visitor sessions and assign to each access a suitable session identifier. Access-log entries with the same session identifier are part of the same session. It must be noted that WumPrep offers the option of inserting dummy entries at the beginning of each session for the referring site, if this is available. We have selected this option as we believe the origin of the access is valuable data.
- Session Aggregation – aggregation of sessions and creation of a session/page matrix that identifies how many times each session visited each of the web pages of the web site.

As a result of the data preparation for content mining we produce a matrix with the rows representing individual sessions and the columns the available web pages. Each row presents which pages and how many times each session visited. A value of zero denotes that the page was not visited by that session; a non-zero value of x indicates that the web page was visited x times during that session.

5.2 Data preparation content mining

In order to depict the contents of the web pages more accurately an ontology of the web site is created. The ontology, despite the fact that it is very simple, provides better results than other techniques such as counting the number of occurrences of words within the web pages (Petrilis & Halatsis, 2008). In the future the authors plan to use a more comprehensive ontology in order to further improve the results.

The ontology describes the set of the web pages that constitute the web site. The main classes, slots and role descriptions are identified. Protégé is used as the visualization tool for the ontology (Protégé). The classes and the value slots have been used to determine the content of each of the web pages. There are six main classes in the ontology that has been created:

- Person –the type of author of the web page
- Web Page – indicates whether it is an internal or an external page
- File – information about the web page file (e.g. name, type, etc)
- Company –company name and type that is associated to the specific web page
- Structure –the place of the web page in the structure of the web site
- URL – information about the URL (static or dynamic and the actual address)

The ontology that was created for the purposes of our processing is depicted in Figure 3. These classes have subclasses, which in turn may have subclasses of their own. In addition classes have slots. As an example the class “URL” has two slots “Static or Dynamic” and “URL”. The first denotes whether the specific web page is statically or dynamically generated and the latter the actual URL of the web page. We have placed great emphasis in encapsulating the structure of the web site. The reason is that in order to get a better understanding of the contents of a web page we need to understand how it relates to other pages within the site.

Using the ontology as a basis we create a matrix with the rows representing individual web pages and the columns the available classes and possible slot values. Each row presents what classes and slot values are relevant to the specific web page. A value of zero denotes that the specific class or slot value is not relevant; a non-zero value indicates that the specific class or slot value is of relevance to the specific web page. The values have been weighted in order to depict the significance of the specific class or slot value to the web page. We apply

greater weights to the classes and slot values that relate to the structure of the web site, since they provide very important information regarding the contents of the web page.

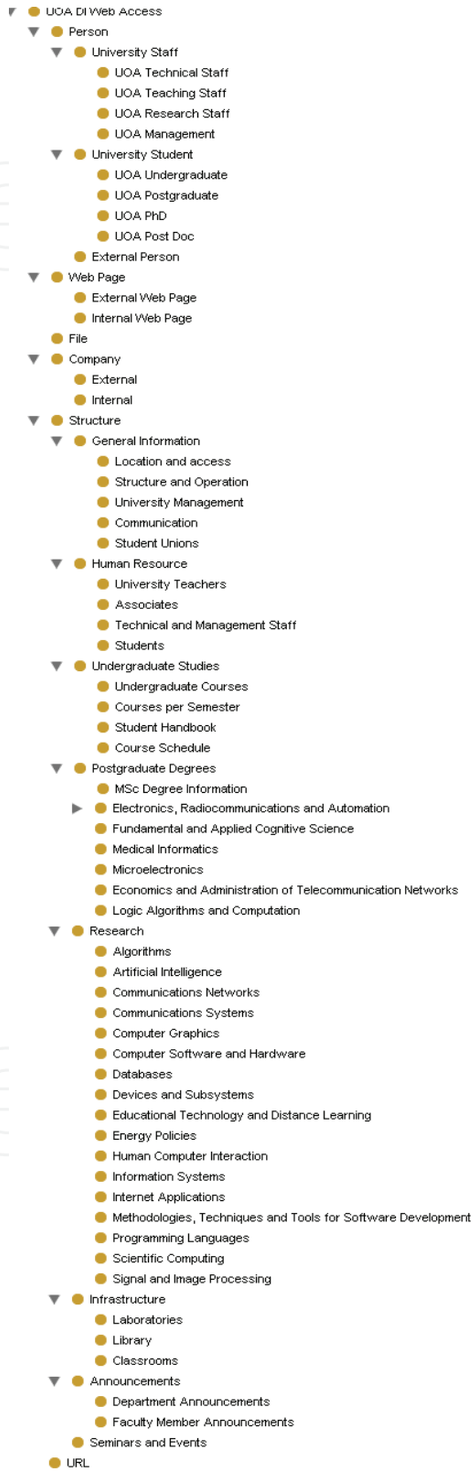


Fig. 3. The Department of Informatics and Telecommunications Ontology

Data Item	Ont. class/slot 1	Ont class/slot 2	…….	Ont class/slot n
Page 1	75	100		0
Page 2	0	100		0
……..	…	…	...	…
Page m	100	75		100

Table 2. Output of the data preparation for content mining

5.3 Combining content and context data

The input data for SOM processing is a combination of the output of the pre-processing steps described in paragraphs 5.1 and 5.2. A matrix is created with the rows representing individual sessions and the columns the available classes and possible slot values of the ontology of the web site. Table 3 shows a sample of the final output of the pre-processing.

Data Item	Ont. class/slot 1	Ont class/slot 2	…….	Ont class/slot n
Session 1	0	100		75
Session 2	75	0		0
……..	…	…	...	…
Session m	100	0		100

Table 3. Final output of the pre-processing

A value of zero indicates that the specific class or slot value is not relevant for the session, whereas a non-zero value denotes that the specific class or slot value is of relevance for the specific session, i.e. to the web pages this session accessed. Additionally a weight is applied to the non-zero values that signifies the relevant of the specific class or slot value to the session. A greater weight is applied to classes or slot values that relate to the structure of the web site, since this is more important in determining the content of the web page.

6. Clustering the data using the SOM

The output that is produced as part of the pre-processing steps described in Paragraph 5 is used as the basis for input to the SOM. The SOM_PAK application has specific formatting requirements and hence the matrix that can be seen in Table 3 is converted to the following format that can be seen in Table 4.

```
<dimensionality>
<class/slot value1 >  <class/slot value2> … < class/slot valuen> <session
id>
```

Table 4. Format of input file to the SOM

A sample of the input file can be seen in Table 5 below:

```
60
100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.121.0.0
0 75 100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.122.0.0
0 0 100 0 75 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.311.0.0
```

Table 5. Sample of the input file to the SOM

The dimensionality indicates the number of columns in the input file and it is a prerequisite for SOM_PAK. Each additional value in the input file denotes the relevance of each ontology class and slot value to the specific session. The session id is used as a label that appears in map produced by SOM and helps us identify individual sessions in the map.

SOM_PAK requires from the user to assign values to some parameters before initiating the processing. These parameter values were selected after evaluating the results with different combination of parameters. To assist with the evaluation of suitable parameters for SOM processing the Sammon projection (Sammon, 1969) of the data is used. The Sammon program of SOM_PAK is used for this purpose. This program provides a quick and easy visual way to evaluate the quality of produced maps for specific parameter values. The selected parameter values can be seen in the table that follows:

Data Item	Value
Neighborhood Type	Hexa
Neighborhood Function	Bubble
Map x size	20
Map y size	8
First Training Period Length	2000
First Training Neighborhood Radius	20
First Training Constant	0.5
Second Training Period Length	8000
Second Training Neighborhood Radius	5
Second Training Constant	0.05

Table 6. SOM Parameter Values

The output of processing is a map in the form of a text file with coordinates. This map is difficult to read and interpret and hence a means of visualising it is required. To visualise the results an UMatrix analysis is applied to the output of the SOM processing. UMatrix analysis provides a visual representation of the map making it easy to identify clusters of sessions. The UMatrix representation uses grey scale values to indicate the distance between nodes. The lighter the colour the closer two nodes are. The darker the colour the greater the distance. Clusters can be easily identified as areas of light hexagons separated by dark hexagons. The output of the UMatrix analysis can be seen in Figure 4.

In the map of Fig. 4 hexagons with a black dot in the centre represent nodes. The labels, such as 1.368.0.0, are session identifiers. The labels have been added to the node that best describes the specific input vector. Some of the clusters of this figure have been highlighted with colours to make it easier to demonstrate the results. Looking at the map of Fig. 4 and examining the underlying data we can quickly draw some conclusions:

- Red oval sessions have all accessed pages written by University staff that are relevant to research on algorithms.
- Sessions in the yellow cluster have accessed pages where the author is a member of the University Staff and/or University Teaching Staff and relate to research and research projects.
- The sessions in the blue cluster have only accessed the University's home page.
- Sessions in the violet cluster have accessed pages that relate to Research, Research Projects or Internet Applications. The authors of these pages are members of the University staff, University teaching staff or University students.
- Sessions in the grey cluster have accessed pages relating to Logic Algorithms and Computation or Internet Applications and were written by University Staff or University teaching staff.
- Sessions in the green cluster have accessed pages relevant to research, research areas, research projects and/or algorithms. These pages were written by University staff, University teaching staff or University students.

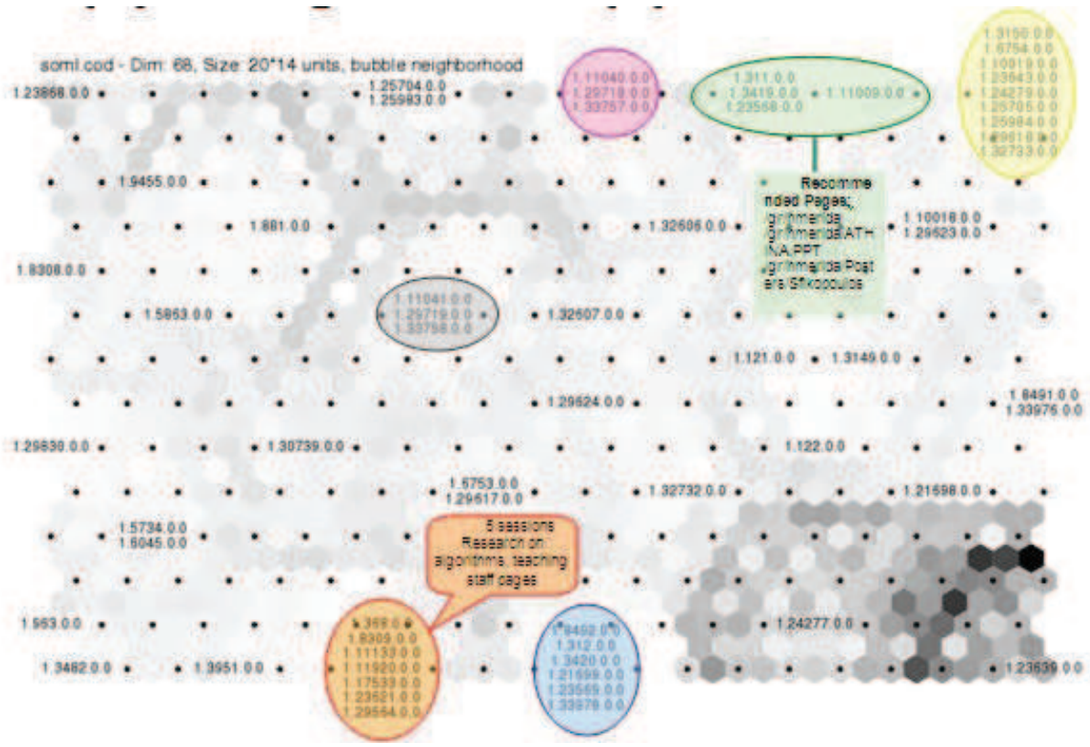


Fig. 4. UMatrix representation of the SOM output

As we have demonstrated by observing the map produced by SOM processing and by examining the underlying data we can quickly and easily extract useful information

regarding the web site. This output could be used to dynamically recommend pages to the visitors of the web site based on the contents of the page that they are currently viewing and on the behaviour of past visitors that have accessed the same or similar pages. As an example if a session accesses page /gr/hmerida then the rest of the pages relevant to the green cluster can be proposed (/gr/hmerida/ATHINA.PPT and /gr/hmerida/Posters/Sfikopoulos)

7. Conclusions

In recent years we have witnessed the rapid expansion of the Internet. There are billions of web pages that are registered by search engines. Web sites tend to increase in size accumulating an ever increasing amount of information. This is especially true for web sites that have been around for a number of years or are updated very often. Web 2.0 and the ever increasing popularity of Social Media Networks have created an Internet culture where visitors are no longer passive but they contribute to the contents of their favourite web sites on a regular basis. This has resulted in web sites that are very complex in their structure. In addition a large number of Internet users have always Internet access available to them through mobile devices. The demand to be able to find information quickly and easily is therefore apparent. Despite the continuous effort to improve the search engines, it is still often a challenge for web site visitors to achieve this.

There is a plethora of commercial applications as well as academic research on predicting web pages that will be useful to a visitor with the final goal of making recommendations to web site visitors. Clustering techniques have demonstrated a relatively good level of success compared to other methods, such as simple statistical applications. However, the current clustering techniques are typically incomplete in the sense they that focus either on the content or the context of the web site. This way important information is ignored when making recommendations because identifying the best web page to recommend depends on both the content of the pages that have been viewed already by the visitor but also on the behaviour of past visitors with similar interests.

In this chapter we present a method that combines both content and context mining. We demonstrate how we can achieve better results by producing a single Self Organising Map that combines data for both the content and context of a web site. Furthermore we demonstrate how a simplistic ontology of the web site can help in determining the content of the web pages. Our approach improves the results of previous research (Petrilis & Halatsis, 2008) and it correctly identifies hidden relationships within the data. In addition the results of the proposed method are easily visualized and interpreted and can be used to dynamically recommend web pages to web site visitors based on both the content of the page they currently viewing but also on the content of similar pages and on past visitor behaviour.

We intend to test our approach on a bigger and more complex web site. In addition it would be interesting to use a more diverse data set. The web sites and web pages of the Department of Informatics of University of Athens are limited in terms of the information they contain. Ideal candidates for our method would be the web site for an online store or an online newspaper and in general web sites with more diverse topics. In addition the ontology that was constructed to depict the contents of the web site pages is very simplistic. We strongly believe that a more comprehensive ontology will yield better results.

Furthermore in the future we plan to integrate the produced SOM with a recommender system that dynamically recommends pages to web site visitors.

8. References

Andrade MA.; Chacón P., Merelo-Guervós J. (1993). Evaluation of secondary structure of proteins from UV circular dichroism spectra, *Protein Eng,* Vol. 6, No. 4, pp. 383–390

Berners-Lee T.; Hendler J. and Lassila O. (2001). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American.com*

Chekuri C. et al. (1996). Web search using automatic classification, *Proceedings of Sixth World Wide Web conference*, San Jose, CA

Cooley R.; Mobasher B., Srivastava J. (1997). Web mining: Information and pattern discovery on the Worldwide Web, *International Conference on Tools with Artificial Intelligence,* pp. 558-567, Newport Beach, CA

Duineveld R. et al. (2000). Wondertools?: a comparative study of ontological engineering tools, *Int. J. Hum.-Comput. Stud.*, Vol. 52, No. 6. (June 2000), pp. 1111-1133

Facebook, http://www.facebook.com

Gruber T. (1993). A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, Vol. 5, No. 2, 199-220, California

Gennari J.H. et al. (2003). The evolution of Protégé: an environment for knowledge-based systems development, *International Journal of Human-Computer Studies*, Vol. 58. No. 1, pp. 89-123

Kaski S. (1997). Computationally efficient approximation of a probabilistic model for document representation in the websom full-text analysis method, *Neural Process Lett*, Vol. 5, No. 2, pp. 69–811

Katifori V. et al. (2007). Ontology Vizualization methods-a survey, *ACM Computer Surveys*, Vol. 39, No. 4

Kohonen T. (2001). *Self-organizing maps*, 3rd edn. Springer-Verlag, Berlin

Lagus K.; Kaski S., Kohonen T. (2004). Mining massive document collections by the WEBSOM method, *Information Sci*, Vol. 163, No. 1–3, pp. 135–156

LinkedIn, http://www.linkedin.com

Luke S.; Spector L. and Rager D. (1996). Ontology-Based Knowledge Discovery on the World-Wide Web, *Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence*, pp. 96-102, USA

Merelo JJ. et al. (2004). Clustering web-based communities using self-organizing maps, In: *Proceedings of IADIS conference on web based communities,* Lisbon, Portugal

Mobasher B.; Cooley R., Srivastava J. (1999). Creating AdaptiveWeb Sites through Usage-based Clustering of URLs. *Proceedings of 1999 workshop on knowledge and data engineering exchange*, pp. 19-26, USA

MySpace, http://www.myspace.com

Netcraft (2008). *Weber Server Survey*, http://news.netcraft.com/archives/web_server_survey.html

Noy NF.; McGuniness D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880

Pang-Ning T.; Vipin K. (2002). Discovery of Web Robot Sessions Based on their Navigational Patterns, *Data Mining and Knowledge Discovery*, Vol. 6, No. 1, pp. 9-35

Petrilis D.; Halatsis C. (2008). Two-level clustering of web sites using self-organizing maps, *Neural Process Lett,* Vol. 27, No. 1, pp. 85-95

Picasa, http://picasa.google.com

Protégé, http://protege.stanford.edu/

Quesada J.; Merelo-Guervós J.J., Oliveras M.J. (2002). Application of artificial aging techniques to samples of rum and comparison with traditionally aged rums by analysis with artificial neural nets, *J Agric Food chem.,* Vol. 50, No. 6, pp. 1470–1477

Romero G. et al. (2003). Visualization of neural network evolution, *Lecture notes in computer science*, Nos. 2686–2687, pp. 534–541, LNCS , Springer-Verlag

Royal Pingdom. (2009). http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/, *Internet 2009 in numbers*

Sammon J.W. Jr. (1969). A nonlinear mapping for data structure analysis, *IEEE TransComput* Vol. 18, pp. 401–409

SOM_PAK and LVQ_PAK, http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml

Twitter, http://www.twitter.com

Ultsch A. (1993). Self-organizing neural networks for visualization and classification , In: Opitz O, Lausen B, Klar R (eds) *Information and classification*, pp. 307–313, Springer, London, UK

Vesanto J. et al. (1999) Self-Organizing map in Matlab: the SOM Toolbox, *Proceedings of the Matlab DSP conference,* pp. 35–40, Espoo, Finland

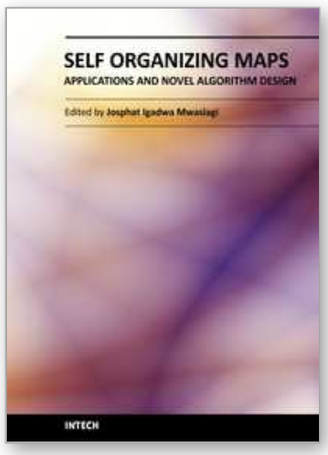
WebLog Expert, http://www.weblogexpert.com/

WorldWideWebSize.com. (2010). http://www.worldwidewebsize.com/, *Daily Estimated Size of the World Wide Web*

WumPrep, http://www.hypknowsys.de

YouTube, http://www.youtube.com

Zhang J.; Caragea D., Honavar V. (2005). Learning ontology-aware classifiers, *Proceedings of the eight international conference on discovery science (DS 2005)*, Springer-Verlag, Berling



Self Organizing Maps - Applications and Novel Algorithm Design
Edited by Dr Josphat Igadwa Mwasiagi

ISBN 978-953-307-546-4
Hard cover, 702 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

Kohonen Self Organizing Maps (SOM) has found application in practical all fields, especially those which tend to handle high dimensional data. SOM can be used for the clustering of genes in the medical field, the study of multi-media and web based contents and in the transportation industry, just to name a few. Apart from the aforementioned areas this book also covers the study of complex data found in meteorological and remotely sensed images acquired using satellite sensing. Data management and envelopment analysis has also been covered. The application of SOM in mechanical and manufacturing engineering forms another important area of this book. The final section of this book, addresses the design and application of novel variants of SOM algorithms.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen