International Journal of Computational Intelligence Research.
ISSN 0973-1873 Vol.2, No.2 (2006), pp. 159–167

Using a Self Organizing Feature Map for Extracting Representative Web Pages from a Web Site

Sebastián A. Ríos[1], **Juan D. Velásquez**[2], **Hiroshi Yasuda**[1] **and Terumasa Aoki**[1]

[1]Research Center for Advanced Science and Technology, University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, Japan
{*srios,yasuda,aoki*}*@mpeg.rcast.u-tokyo.ac.jp*

[2]Department of Industrial Engineering, University of Chile
República 701, Santiago, Chile
jvelasqu@dii.uchile.cl

Abstract: **We introduce a method for improving the web site content through the identification of their most representative web pages. The process begin with the transformation of the web page text content in feature vectors by using the vector space model for documents. Next a Self Organizing Feature Map (SOFM) receive these vectors as input, generating a set of clusters, whose centroids contain the most representative text content for a topic in the site.**
In the web page's vectorial representation, the text content is transformed in a set of numeric values. Then by operation of the SOFM, the cluster's content are vectors whose relation with the web site pages is not clear. By applying a Reverse Cluster Analysis (RCA), it is possible to identify which pages are represented in each cluster. The RCA consists in the comparison among the vectors in each clusters with the page's vector representation. Next the pages whose vectorial representation is near to the cluster's centroid, are extracted.
This approach was tested in a real web site in order to shows its effectiveness. The results indicate that it is possible to identify representative web page in a web site and for this way, improve the site's text content.
Keywords: Neural Networks, Web Content Mining, Self Organizing Maps

I. Introduction

Designers and web masters have made great efforts to achieve continuous improvements of the web site structure and content (only the free text on the page or text plus images, videos, etc). However, this is a non-trivial task because the site must dynamically change in order to permanently satisfy the visitors' requirements.
Usually managers or web masters try to enhance the web site in order to achieve a better visitors browsing experience. These is very important because this way they are able to maintain the existing visitors and to attract new ones [Nie99]. However to define the correct text content is a complex task, due to the fact that visitors requirements and preferences are continuously changing [BGMP01, Nie99]. On the other hand, many researchers have proposed several mathematical tools, processes or methodologies to help improve the web site content, the web site structure and the web site usability, such as Web Content Mining (WCM) [BGMP01, MSB97, RJZ89, RVV$^+$05a, Tur03] , Web Structure Mining (WSM) [PTM02, RVYA05, VRB$^+$05] and Web Usage Mining (WUM) [BS01, MCS00, Per01, SMBN03, VYAW04] respectively.
Our aim in this work is to help manager, web masters or any organization to improve a web site content based on Web Text Mining (WTM) techniques [BS01]. We discover that it is hard to precisely find where the changes have to be applied or to produce a guide on how to focus the efforts and resources to change the Web Site. We help to automatically identify relevant or Representative web pages. This is made using a technique that we have called *Reverse Cluster Analysis* (RCA) [RVV$^+$05c, RVV$^+$05b, RVV$^+$05a]. A small set of pages is the result of this process and these are the pages that should be the main point of attention and resources focus in the begining stage of analysis and enhancements of the site. We successfully test this idea in a real web site of the University of Chile.
The work is organized as follow: In Section 2, we introduce the past works on the subject and also we explain detailed some techniques used and also how we use a SOFM for classification of web pages. Section 3, we introduce the RCA technique and also how we do the automatic Representative pages identification. Next, in Section 4, we show experimental results that support our proposal. Finally, in Section 5, we discuss our work and show a glimpse on our future work.

II. Related Work

Searching the information that best suite the needs and requirements of any web visitor is a very demanding task. The existence of huge amounts of heterogeneous, unlabeled, distributed, time-variant, semi-structured and high-dimensional data [PTM02] together with the changing needs of the visitors requirements makes it very difficult to decide which is the best way to define the correct content of a site [VRB$^+$04]. There exist several approaches that help to do such task. Some researchers make use of Web Mining (WM) techniques. The type of mining technique used depends on the needs of the owner of the site. If enhancements on the site text content are needed then the use of Web Text Mining (WTM) or Web Content Mining (WCM) techniques is a common practice. However, one may need some information about how the web site is being used by the visitors. We use Web Usage Mining (WUM) in this case, which works on the mining of web logs contained in the web server to understand how the visitors browse the Web site. Similarly, we may need to analyze the internal link structure of a web site, this is made using Web Structure Mining (WSM).
Text mining look for the identification of Key Words [VWYA04, VRB$^+$04, VRB$^+$05], Key Sentences or Key Paragraphs [Tur03, MSB97]. Knowledge Discovery in Text (KDT) concerns to the application of Knowledge Discovery in Databases (KDD) techniques over free text. Loh et al. use KDT process for developing a Concept-Based Knowledge Discovery process for web texts [LWdO00]. In that work the KDT techniques are applied over concepts rather than on attribute values, terms or keywords labeling texts. Then statistical analysis are performed to obtain interesting patterns. One of the objectives of Loh was to allow the user to search ideas, ideologies, trends and intentions presents on text.
Some other new approaches aim to make a combination between WTM, WCM, WUM and WSM. For example in [VRB$^+$04, VRB$^+$05], Velasquez et al. propose a methodology to extract Key Words. His approach lay on a process that combined the Web site text content and the visitors' browsing behavior. Using a SOFM they obtain several clusters and then from these obtain the Key Words that are more representative of the visitors' interests. The results are used to improve the Web site structure for a better visitor experience. In order to introduce more semantics to the text mining process Rau et al. in [RJZ89] developed SCISOR (System for Conceptual Information Summarization, Organization and Retrieval) which tries to allow the conceptual access to documents. Similarly, Eirinaki et al. have developed the Semantic Web Personalization System (SEWeP) [EVV03, ELPV04]. The basis of this system relays on an enhanced version of the web logs which are called C-Logs (concept logs). These C-Logs consist of web sites' semantic information that is added to the traditional usage logs in the way of keywords. Afterwards, these C-Logs are used in the mining process to obtain better and broader recommendations.
However, independent of the mining technique chosen, the main goal is to help managers or web masters to improve the Web site. To do so, usually is not enought to just apply one of the above approches, many times the help of the expert of the site (usually the manager, web master or a qualified team) is needed. They should validate the results and perform a usability test to the web site before it goes to production [Nie99, ZGA05].

A. Data Selection and Preprocessing

In order to obtain the best possible results, a web site with a relatively high amount of text is needed. Pages which show the information to the visitor in images, flash text, videos, audio, etc. are not so useful. This is because we are mining text. In other words, we can't extract semantic meaning from an image or video or audio using our SOFM.
After selecting a web site fulfilling the above requirements, we first filter the non-useful words like articles (a, an, the) or pronouns ("he," "she," and "it" for singular, "they" for plural), etc., in order to just apply the clustering algorithm to the most relevant words (nouns, verbs and adjectives).
An interesting problem appears with the plural of words. Should two different words, for instance "car" and "cars", which represent the same concept, be considered as only one term? Another similar problem is verbal conjugations, for instance "drive", "drives", "drove", "driving", etc. To solve this problem we use Porter's algorithm [Por80]. This is a stemming algorithm, which allows us to find the root of the words.
After applying this techniques to the selected web site we reduced the universe of different words of the site by about 64%. This allows the next steps to be faster and more precise. It is very important to mention, that we apply all of these techniques to a web site written in Spanish. That's why we use a modification of the Porter's algorithm to process Spanish language.

B. Web Page Feature Vector

Using the vector space model [SWY75], the web page text content is transformed into a feature vector.
Let W be the number of different words in the entire collection of documents and Q the number of documents. In our case a document would be a web page and the collection of documents the respective web site. A vectorial representation of the web site would then be a matrix M of dimension $W \times Q$ with:

$$M = (m_{ij}) \, \forall i = 1, \ldots, W \ \wedge \ j = 1, \ldots, Q \qquad (1)$$

where m_{ij} is the weight of word i in document j.
This weight must capture the fact that a given word can be more important than another one. For instance, if the word i appears in n_i documents, the expression n_i/Q gives a sense

of its importance in the complete set. The "inverse document frequency" $IDF = \lg(Q/n_i)$ can be used like a weight. However, in this work we use an expression that is known as $TF * IDF$ (Term Frequency Times Inverse Document Frequency), as shown in Eq. 2.

$$m_{ij} = f_{ij} * \lg(Q/n_i) \tag{2}$$

where f_{ij} is the number of times that the i^{th} word in the j^{th} page and n_i is the number of documents containing the i^{th} word. A page p_j is represented by the column j in M, i.e., $p_j \rightarrow (m_{1j}, \ldots, m_{Wj})$ and the distance or similarity measure used is:

$$pd(p_i, p_j) = \frac{\sum_{k=1}^{W} m_{ki} m_{kj}}{\sum_{k=1}^{W} (m_{ki})^2 \sum_{k=1}^{W} (m_{kj})^2} \tag{3}$$

The Eq.2 is known as the dot product between two vectors in Cartesian coordinates. Therefore, to find out how similar or dissimilar are two feature vectors we compute this expression.

C. Using Self Organizing Feature Map (SOFM)

We use SOFM [Koh01] to extract significant patterns from the web page text content. A toroidal topology is used to maintain the continuity of the space [VYAW02, VRB+04, VRB+05]. Then a Gaussian function that depends on the distance from the centroid is used to propagate the learning to the neighbor neurons as shown in Eq.(4). This function allows the centroid neuron to learn the pattern shown. Afterwards the effect of the learning is passed to the neighborhood in smaller degree, inversely proportional to the centroid distance.

$$h_{ci}(t) = \alpha(t) \cdot \exp^{-\frac{\|r_c - r_i\|^2}{2 \cdot \sigma^2(t)}} \tag{4}$$

Finally the neurons learning on each epoch are described using the expression 5, where $x(t)$ is the Example shown to the Network on the epoch t.

$$m_i(t+1) = m_i(t) + h_{ci}(t) \cdot (x(t) - m_i(t)) \tag{5}$$

The SOFM has a restriction called the *Domain Restriction*, this means that the results obtained depends on the examples chosen to train the Network. This is the reason why, if we need to represent the whole web site text content, it is not possible to let some page out from our teaching examples set. Therefore we use 182 examples (one per each real web site page as we mention before).

III. Reverse Clustering Analysis

After finding the clusters, we have the most commonly used words that attract the visitors attention in the whole Web Site, but we know nothing about which are the most relevant web pages. Moreover, if we study the artificial neural network, we only have a vector of frequencies for all the words that compound the web site. One big challenge that we find in this technique is that such vector is far from a web page, because the network at the beginning is randomly initialized. Therefore, the vectors that the resulting clusters may contain usually do not correspond to any real web site's page. In other words, we have found some content patterns using a SOFM, but we don't know, at this point, which are the web pages that these clusters represent? or which are the pages that best matches the content patterns found?.

To answer the above questions a new method is needed. However, as we just mentioned before, it is very hard to find a perfect correspondence between the clusters and real web pages.

Therefore, we apply again the similarity measure between pages Eq.(3), in order to find the documents which are most similar to our clusters. This way we obtain the most relevant pages in the whole web site Fig.1.

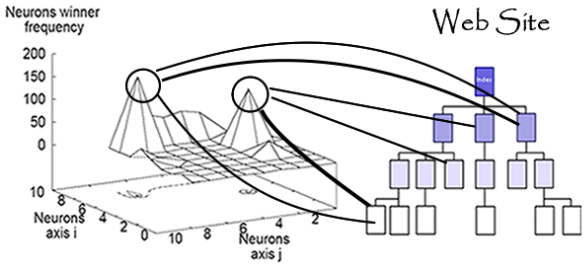

Figure. 1: Cluster analysis to find out the relevant pages

First we need to extract a cluster centroid and its neighbor neurons. Then, for each one of these neurons we compute the similarity between this neuron and each document in the web site using Eq.(3). At the end, we obtain a web page which is the most similar to the neuron used. We call this web page a winner page, and we say that the neuron is *referencing* this web page. We perform these for all the neurons in the cluster and for all the clusters. Therefore, at the end of this process, we found a set of web pages, that are referenced by the cluster neurons and we called it the *relevant pages set*. This is the set of all pages which have more than zero references.

The process mentioned above is called by us as the *reverse clustering analysis*.

A. Extracting the Clusters from the SOFM

At this point we need to know which are the cluster's centroids, and associated neurons to perform the reverse clustering analysis. This task is absolutely critic and must be done carefully in order to obtain reliable information. However, there are many ways to do this, so we focus only in two ways.

$$r \leq \sqrt{(x_c - x)^2 + (y_c - y)^2} \quad (6)$$

First we use a very simple circular neighbor function. This consists in taking all the neurons inside the radius r and looking if there is a local maximum in this vicinity (see Eq. 6). This is the traditional circular function with the origin in (x_c, y_c) this is the position of the possible centroid neuron in our toroidal space. If there is a local maximum, then this is the centroid. Later we take the rest of neurons inside of this cluster centroid and we mark them as part of the centroid Fig.2.

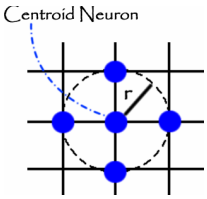

Figure. 2: Circular vicinity for cluster extraction

However, if we use this function the problem is that we can consider more clusters than they really exist, because we do not compare the possible centroid to the vertexes of the grid. That is why we take a square vicinity. For instance, if we take $r = 1$ in Fig.3 then we only compare the centroid to four neurons, the vertices of the square are outside of the circular vicinity.

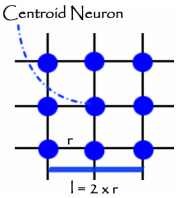

Figure. 3: Square vicinity for cluster extraction

The experimental results show that in fact, with a circular vicinity we find more clusters (34 clusters). As we mentioned before, some clusters are not local maximums and must not be considered as clusters; we should use the square vicinity instead to find local cluster centroids. Using the square vicinity we only obtain 13 cluster's centroids using the side of the square center in (x_c, y_c) parameter in $a = 2$.

The consequence of finding more clusters, which are fake or not, is that we introduce noise and obtain wrong results, marking a web page as important while it is not. Doing so it is possible to observe how the final results behave.

The results could even be more accurate if we take a bigger side of the square. If our SOFM in matrix representation has $(U \times U)$ elements in the limit when $a = card(U)$ only centroid that will be obtained is the maximum of the whole network.

B. Marking and Summarizing the Relevant Web Pages

The next step, after finding our clusters, is to compare the feature vectors of the centroids's neurons with the real pages. Depending on which kind of vicinity used is the number of neurons used. e.g. 4 neighbor neurons plus the centroid in the case of circular vicinity (5 neurons in total) when we use a radius $r = 1$.

Then we apply the similarity measure shown in Eq.(3) to obtain the minimum value between a neuron in the cluster and all web pages in the site. We compare all documents with all neurons, these means all neurons on each cluster. To do this, we define the Page Reference function $PR(n_i, p_j)$ Eq.7, where ζ is the set of clusters centroids plus the associated neurons.

$$\begin{aligned} &PR(n_i, p_j) = Min\{pd(n_i, p_j)\} \\ &\forall j = 1, \ldots, Q \qquad \wedge \qquad \forall i \epsilon \zeta \end{aligned} \quad (7)$$

We can observe in Table 1 part of the results that we obtained when applied Eq.(7). In this table we observe the URL of the web pages that have been referenced at least once by any neuron in any cluster found. Under each URL we show the cluster centroid ID that is referencing that page. For instance, the page "/novedades.htm", is referenced by the cluster which Centroid Id is 0, five times. The page "/mapa.htm" is referenced for the clusters $\{0, 1, 4, 5, 7, 8\}$.

The references from a cluster to the real web site depend on the type of vicinity used. The circular vicinity has five neurons representing each cluster (see Fig.2) but the square vicinity has nine (see Fig.3). For instance, in Table 1 it is possible to observe that the page "novedades.html" and the page "mapa.htm" are both being referenced by the cluster $ID = 0$ however, in the first case there are five references from cluster $ID = 0$ and in the second case there are two references. This means that five neurons of the cluster $ID = 0$ (of the whole set of five neurons that represent a cluster in the circular vicinity or nine neurons in the case of square vicinity) is the most similar to the page "novedades.htm" and another two neurons from the same set of cluster $ID = 0$ are the most similar to the page "mapa.htm". We do not distinguish between the centroid and the neighbors. This means that all the neurons in the cluster have the same weight. On the other hand, both pages are being referenced once by any neuron in the cluster $ID = \{1, 4, 5, 8\}$. This is exactly the same situation explained before.

We need to summarize the references obtained at the end of the process. To do so we simply add the number of references that a real page has. For the page "novedades.htm" we identified 30 references (see Table 1); in the case of the page "/mapa.htm" we found eleven references in total; finally, for the page "servicios.htm" has two references.

We could have found no convergence, these means that each neuron has a different most similar document. That means that each representative real page has only one reference. However, we discover a rather small set of web pages that are the most relevant based on the text content analysis.

URL: *http://escuela.ing.uchile.cl/novedades.htm*
+ Centroid Id: 0 References: 5
+ Centroid Id: 1 References: 5
+ Centroid Id: 2 References: 9
+ Centroid Id: 3 References: 1
+ Centroid Id: 4 References: 1
+ Centroid Id: 5 References: 2
+ Centroid Id: 6 References: 4
+ Centroid Id: 8 References: 3
URL: *http://escuela.ing.uchile.cl/mapa.htm*
+ Centroid Id: 0 References: 2
+ Centroid Id: 1 References: 3
+ Centroid Id: 4 References: 1
+ Centroid Id: 5 References: 1
+ Centroid Id: 7 References: 1
+ Centroid Id: 8 References: 3
URL: *http://escuela.ing.uchile.cl/servicios.htm*
+ Centroid Id: 0 References: 1
+ Centroid Id: 7 References: 1

Table 1: Partial results of RCA, list of clusters and their references to a real page.

IV. A Real Case Application

We applied the process mentioned above to the site of the School of Engineering and Sciences of the University of Chile.[1] This Web Site has 182 web pages. Most of the meaningfull content of the site is contained as free text. They use very few images for the main sections titles. After the filtering and stemming process the number of different only about 4,000 words remain. (from more than 11.000)
The process ran over a DELL server dual Xeon, with 2 Gigs RAM and Linux RedHat 9. It took about 34 hours. The Clustering and Site Classification software was developed using Object Oriented PHP. This way its structure could be easily translated to JAVA or C++.
We applied the process explained in Sections II and III in two different ways to observe how the circular and the square vicinities alter the final result. Also, we alter the size of the neural network selected because we think it is the other variable that could affect the results severely
The artificial neural network used in the first experiment was set in 144 neurons, i.e 12×12 (about 79% of the size of the original space of documents) and it was applied to the examples using 70 epochs. After the application, we found four main clusters (see Fig.5).
The artificial neural network used in the second experiment was set in 100 neurons (about 55% of the size of the original space of documents) and applied the examples using 70 epochs also. After the application, we found five main clusters, and 14 clusters in total Fig.6.
There are several reasons to set the SOFM in 144 and 100 neurons. First, we tried to test if the size of the SOFM can alter the final results of the RCA; second, the process take 34 hours using 144 neurons so we set 100 neurons to make the second experiment faster; third, when using SOFM we try to map our data to a very small space, the first map 144 neurons is about 79% of the document space, but with 100 neurons we have a SOFM of 55% of the size of the original space. The problem using a space similar to the space of the data is that getting a fast convergence is harder than with a small one; even more, obtaining convergence at all is harder.

[1] http://escuela.ing.uchile.cl

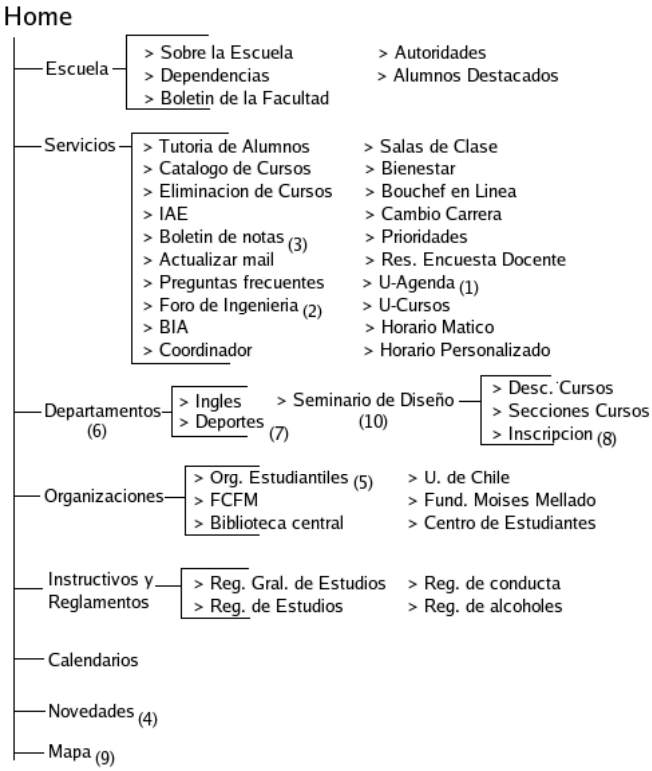

Figure. 4: Map of links of the site used in the experiments (only two levels are shown). To help in the results' visualization we introduce numbers in brackets that correspond to the ones in column (#) of the Tables 2 and 3

.

A. First Experiment Results

The circular vicinity was applied first. The radius was set to $r = 1$, obtaining 36 clusters in the feature map. Afterwards, we applied the similarity measure to find out the most relevant documents. The final result is that 9 pages were found (see Table 2).
Later we applied the square vicinity with the side set as $l = 2 * r$ units to obtain only 16 clusters (see Table 4), this is 55.5% less clusters. Then using the page reference Eq.(7) we gather the guideline set of the most important web pages with only 7 real pages. However, if we observe Table 2, the important web pages are the same. Furthermore, even the order of importance of the pages is the same that results from the circular vicinity. The only difference is that with the square vicinity we do not reference the pages "/sd20a/alumn-

#	Web Page	Circ. Vic.	Sqr. Vic.
1	/agenda/index.html	73	56
2	/foroING/index.html	29	19
3	/Boletin_Notas/index.html	19	9
4	/novedades/novedad_alumnos.php	10	6
5	/organizaciones/estudiantes.htm	8	5
6	/departamentos/index.htm	4	5
7	/departamentos/deportes.htm	4	4
8	/sd20a/alumn-sc.php	2	-
9	/mapa.htm	1	-

Table 2: First experiment results, representative real pages from a 144 SOFM

sc." and "/mapa.htm". The reason for this being that with the square vicinity we only find out real clusters, not fake ones. Therefore, we only identify really important web pages.

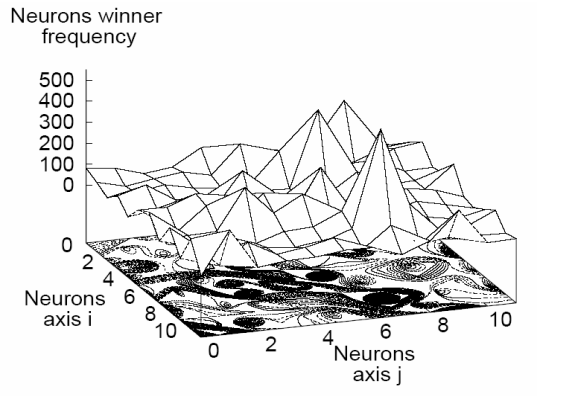

Figure. 5: Clusters found in the first experiment (144 Neurons).

Several pages in the results have "php" extention. Which means that that page is dynamically generated. However, in this case, the "php" program is used only to generate the graphic interface, because we use templates to separate the presentation layer from the content layer. In other words, we have only one content but we can change how this content is shown to the users. This is the reason why we can process "php" pages. To visualize the results shown in Table 2 and 3 into the web site link structure, we provide a two level site map diagram (Fig.4). We added numbers to the links obtained from the RCA. For example, the page "/mapa.htm" is the # 9 page in Table 2 so the reader can see the page with the number 9 in brackets in the Fig.4.

B. Second Experiment Results

The same process was used to extract the clusters from the second SOFM. With the circular vicinity we obtained 27 cluster centroids but we found out only 14 clusters centroids with the square vicinity (see Table 4). It is important to say that not all the 27 or 14 clusters are the most important. In fact, for this case there are only 5 important clusters. In our experiments we used all the clusters found (27 or 14) to make

#	Web Page	Circ. Vic.	Sqr. Vic.
1	/agenda/index.html	37	31
3	/Boletin_Notas/index.html	28	25
2	/foroING/index.html	20	19
6	/departamentos/index.htm	14	10
4	/novedades/novedad_alumnos.php	7	6
8	/sd20a/alumn-sc.php	6	5
10	/sd20a/index.html	4	1
5	/organizaciones/estudiantes.htm	2	1

Table 3: Second experiment results, representative pages from a SOFM of 100 neurons.

the RCA not only those 5 important ones. The reason to do so is that we need to probe the effectiveness of the technique in the worst case, that is considering all the clusters found. Other reason for this is that the experts judgment can tell us if a cluster is or it is not important but an algorithm can not do that beforehand.

After applying the page reference function Eq.(7) the results were only 8 pages, using both methods, the square and the circular vicinities (see Table 3).

This eight pages are the most similar to our clusters, and the real web pages that we could consider as relevant web pages in the whole web site that is composed by 182 pages. Consequently, the University of Chile can now focus its efforts in this reduced set.

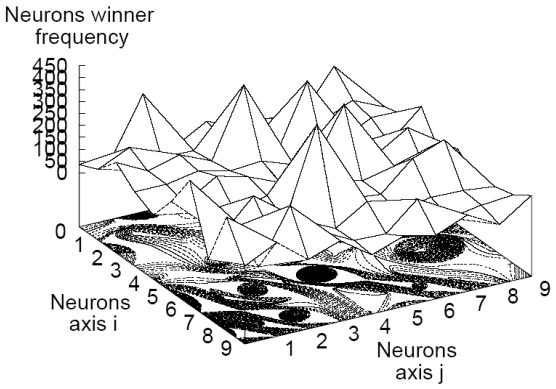

Figure. 6: Clusters found in the second experiment (100 Neurons).

The first page found (see Table 3) was the *School's Agenda* with 31 cluster references; the second page was *Students Grades* with 25 references and the third page was the *Engineering Forum* with 19 references.

It is very interesting that in both experiments the final set found is very small, less than 5% of the whole web site web pages. Moreover, in both experiments (with the circular and square vicinities) the sets of important web pages are almost identical. The only difference that we could observe is in Table 2. In this case, the pages "/sd20a/alumn-sc.php" and "/mapa.htm" are not referenced in the square vicinity. However, the order and importance by reference number is very

	Circular Vic.	Square Vic.
1^{st} Exp.	36	16
2^{nd} Exp.	27	14
Difference	9	2

Table 4: Difference on the number of clusters extracted using circular and square vicinity in both experiments

low.
Furthermore, the both experiments results are very similar, even the three most important pages are the same in the both guidelines sets.

V. Discussion and Future Work

Our results prove that the square vicinity extraction method is better to perform the reverse cluster analysis than the circular vicinity method. As explained before, the reason for this being that we found many more clusters with the circular vicinity than with the square vicinity method because the circular vicinity missed four comparisons. As a result we found out more clusters with the circular approach than with the square one.
If we compare the clusters found with the circular method in both experiments, the variation is huge. First, we found 36 clusters and then we found only 27. The cluster sets found in the second experiment is 25% smaller than the set found in the first experiment. On the other hand, the cluster sets found with the square vicinity in both experiments differ in only 12.5%, corresponding to only two clusters (See Table 4).
These experimental results support the conclusion that the square vicinity detects more real clusters than the circular vicinity. However, if we see the representative real pages in Table 2 and in Table 3, the final representative pages set is almost not affected.
The software developed take about 34 hours, probably because it is made in PHP that is an interpreted language not like JAVA where we have the Bytecode that is intermediate code or C++ that is compiled. Also, the velocity of our software is very different if it is executed from command line or trough the web server (we use apache and mod_php) with any browser; this last manner takes several days to finish the execution. In order to avoid lost of information the software makes periodic back up of all its memory variables (the whole SOFM included). These produce a big delay from a straightforward run.
We need to perform more experiments in other versions of the same site as well as in other organizations web sites, to show more clearly the effects of the vicinity selection in the representative pages set. Although we have shown that we succeeded in finding a set of real web pages.
Another issue to think about is that we considered all the clusters to perform the RCA. It could be interesting to test the technique if we set up a threshold to chose the most important clusters. We do not know if the results converge to a small set of pages or if each neuron reference a different document. Besides, we also considered all the neurons in the cluster of equal weight but it is possible to set up different weights to distinguish between the centroid neurons and the neighbor neurons.
The process mention above we have computed the Page Reference function in Eq. (7) and we always mark one page per iteration. However, we think that it is possible to set up also a threshold in order to discover just the most important references from a neuron to a cluster.
We also mention before that usability tests are very important. We need to perform a survey to the visitors of the web site in order to validate the results obtained with the RCA. However, to do so it could be necessary to change the similarity measure used to incorporate the visitors preferences in the study as we explain in [RVYA05].

Acknowledgments

We would like to thank Pablo Roman, Chief of Development of the School of engineering of the University of Chile, for his enthusiasm and valuable help.
This work has been funded partially by the Millenniumm Scientific Nucleus on Complexes Engineering Systems, Chile.

References

[BGMP01] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Procs. 10th Int. Conf. on World Wide Web*, pages 652–662, Hong Kong, 2001.

[BS01] B. Berendt and M. Spiliopoulou. Analysis of navigation behavior in web sites integrating multiple information systems. *The VLDB journal*, 9:27–75, 2001.

[ELPV04] Magdalini Eirinaki, Charalampos Lampos, Stratos Paulakis, and Michalis Vazirgiannis. Web personalization integrating content semantics and navigational patterns. In *WIDM '04: Proceedings of the 6th annual ACM international workshop on Web information and data management*, pages 72–79, New York, NY, USA, 2004. ACM Press.

[EVV03] M. Eirinaki, M. Vazirgiannis, and I. Varlamis. Sewep: using site semantics and a taxonomy to enhance the web personalization process. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, New York, NY, USA, 2003. ACM Press.

[Koh01] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. 3rd ed. edition, 2001.

[LWdO00] Stanley Loh, Leandro Krug Wives, and José Palazzo M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explor. Newsl.*, 2(1):29–39, 2000.

[MCS00] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.

[MSB97] M. Mitra, A. Singhal, and C. Buckley. Automatic text summarization by paragraph extraction. *In Procs. in ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 31–36, 1997.

[Nie99] J. Nielsen. User Interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.

[Per01] M. Perkowitz. *Adaptative Web Sites: Cluster Mining and Conceptual Clustering for Index Page Synthesis*. PhD thesis, Univerity of Washingtong, 2001.

[Por80] M. F. Porter. An algorithm for suffix stripping. *Program; automated library and information systems*, 14(3):130–137, 1980.

[PTM02] S. K. Pal, V. Talwar, and P. Mitra. Web Mining in Soft Computing Framework: Relevance, state of the art and future directions. *IEEE Transactions on Neural Networks*, 13(5):1163–1177, September 2002.

[RJZ89] Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419–428, 1989.

[RVV+05a] S. A. Ríos, J. D. Velásquez, E. S. Vera, Hiroshi Yasuda, and Terumasa Aoki. Establishing guidelines on how to improve the web site content based on the identification of representative pages. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 284–288, Compiegne, France, September 2005. IEEE Computer Sociecty.

[RVV+05b] S. A. Ríos, J. D. Velásquez, E. S. Vera, Hiroshi Yasuda, and Terumasa Aoki. Improving the web text content by extracting significant pages into a Web Site. In Halina Kwasnicka and Marcin Paprzycki, editors, *5th International Conference on Intelligent Systems Design and Applications,*, pages 32–36, Wroclaw, Poland, September 2005. IEEE Computer Sociecty.

[RVV+05c] S. A. Ríos, J. D. Velásquez, E. S. Vera, Hiroshi Yasuda, and Terumasa Aoki. Using SOFM to Improve Web Site Text Content. In Lipo Wang, Ke Chen, and Yew S. Ong, editors, *Advances in Natural Computation: First International Conference, ICNC 2005*, volume 3611, pages 622–626, Changsha, China, August 27-29 2005. Springer-Verlag GmbH.

[RVYA05] S. A. Ríos, J. D. Velásquez, Hiroshi Yasuda, and Terumasa Aoki. Web Site Improvements Based on Representative Pages Identification. In Shichao Zhang and Ray Jarvis, editors, *AI 2005: Advances in Artificial Intelligence: 18th Australian Joint Conference on Artificial Intelligence*, volume 3809, pages 1162–1166, Sydney, Australia, November 2005. Lecture Notes in Computer Science.

[SMBN03] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS J. on Computing*, 15(2):171–190, 2003.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.

[Tur03] P. D. Turney. Coherent keyphrase extraction via web mining. *In Procs. of the 18^{th} Int. Conference on Artificial Intelligence (IJCAI-03)*, pages 434–439, 2003.

[VRB+04] J. D. Velásquez, S. A. Ríos, A. Bassi, H. Yasuda, and T. Aoki. Identifying keywords to improve a web site text content. 6^{th} *International Conference on Information Integration and Web-based Applications & Services*, pages 39–48, September 2004.

[VRB+05] J. D. Velásquez, S. A. Ríos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.

[VWYA04] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. A methodology to find web site

keywords. *In Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285–292, March 2004.

[VYAW02] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. Voice Codification using Self Organizing Maps as Data Mining Tool. In *Procs. of Second Int. Conf. on Hybrid Intelligent Systems*, pages 480–489, Santiago, Chile, December 2002.

[VYAW04] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *Transactions in IEICE, Special Issues in Information Processing Technology for web utilization*, page to appear, April 2004.

[ZGA05] M. Zviran, Ch. Glezer, and I. Avni. User satisfaction from commercial web sites: The effect of the design and use. *Information and management*, 2211:22, 2005.

Authors' Biographies

Ríos, Sebastián A. is a doctoral student at the RCAST of the University of Tokyo, Japan. He received the B.E on Industrial Engineering on 2001, the B.E on Computer Science and the P.E. on Industrial Engineering on 2003 from the University of Chile, Chile. He has been lecturer since 2002 on the Department of Industrial Engineering of the University of Chile. His research interests consist in Data Mining, Web Mining and Web Semantics.

Velásquez, Juan D. is Assistant Professor with the Department of Industrial Engineering, School of Engineering and Science, University of Chile . He received his B.E. in electrical engineering and B.E. in computer science in 1995, P.E. in electrical engineering and P.E. in computer engineering in 1996, Masters in computer science and Masters in industrial engineering in 2001 and 2002, respectively, from the University of Chile, Chile and his Ph.D. from the University of Tokyo, Japan in 2005. He has been postdoctoral fellow with the Computing Laboratory, University of Oxford, UK in 2005 and Visiting Professor with the Center for Collaborative Research, University of Tokyo, Japan in 2006.

Yasuda, Hiroshi received his B.E., M.E. and Dr.E. from the University of Tokyo, Japan in 1967, 1969, and 1972, respectively. Since joining the Electrical Communication Laboratories of NTT, in 1972, he has been involved in work on video coding, image processing, tele-presence, B-ISDN network and services, Internet and computer communication applications. After serving twenty-five years (1972-1997), his final position being Vice President, Director of NTT Information and Communication Systems Laboratories at Yokosuka, he left NTT and joined the University of Tokyo. He is now Director of The Center for Collaborative Research (CCR). He had served as the Chairman of ISO/IEC JTC1/SC29 (JPEG/MPEG Standardization) from 1991 to 1999, as well as the President of DAVIC (Digital Audio Video Council) from September 1996 to September 1998. He received the 1987 Takayanagi Award, the 1995 Achievement Award of EICEJ, the 1995-1996 EMMY award from The National Academy of Television Arts and Science, and the 2000 Charles Proteus Steinmetz Award from IEEE. He is a Fellow of IEEE, EICEJ, and IPSJ, and a member of Television Institute.

Aoki, Terumasa is a lecturer with the Research Center for Advanced Science and Technology, the University of Tokyo. He received his B.S., M.E. and Ph.D. in information and communication from the University of Tokyo, Japan in 1993, 1995, and 1998, respectively. His current research interests are in the fields of terabit IP router, access control of gigabit LAN/WAN, next-generation video conferencing system, high-efficiency image coding, and management of digital content copyrights. He has received various academic excellent awards such as the 2001 IPSJ Yamashita award, the FEEICP Inose award for 1994, and the 4 other awards.