Using SOFM to Improve Web Site Text Content

Sebastían A. Ríos[1], Juan D. Velásquez[2], Eduardo S. Vera[3,4], Hiroshi Yasuda[1], and Terumasa Aoki[1]

[1] Research Center for Advanced Science and Technology, University of Tokyo
{srios, yasuda, aoki}@mpeg.rcast.u-tokyo.ac.jp
[2] Department of Industrial Engineering, University of Chile
jvelasqu@dii.uchile.cl
[3] Center for Collaborative Research, University of Tokyo
[4] On leave from Department of Computer Science, University of Chile
esvera@vp.ccr.u-tokyo.ac.jp, esvera@dcc.uchile.cl

Abstract. We introduce a new method to improve web site text content by identifying the most relevant free text in the web pages. In order to understand the variations in web page text, we collect pages during a period. The page text content is then transformed into a feature vector and is used as input of a clustering algorithm (SOFM), which groups the vectors by common text content. In each cluster, a centroid and its neighbor vectors are extracted. Then using a reverse clustering analysis, the pages represented by each vector are reviewed in order to find the similar. Furthermore, the proposed method was tested in a real web site, proving the effectiveness of this approach.

1 Introduction

From the early stages of the web development, designers and web masters have made great efforts to achieve continuous improvements of web site structure and content. This is a non-trivial task, because the site must dynamically change in order to permanently satisfy the visitors' requirements.

To define the correct text content is a complex task, due to the fact that visitors requirements and preferences are continuously changing. Therefore, it is a hard task to figure out which is the best content for a web site [2]. On the other hand, many researchers have proposed several mathematical tools to help improve the web site content. However, it is hard to discover where the changes have to be applied [11,10], or to produce a guide on how to focus the efforts and resources to change the Web Site.

In this paper, we make use of some Web Content Mining (WCM) techniques in order to find the most relevant pages in the whole web site. These are the pages that should be the main focus of attention for the organizations.

2 Related Work

To produce adequate guidelines on how to make web site improvements, we need to find the most relevant text in a particular site. With this purposes in mind,

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3611, pp. 622–626, 2005.

we apply web content mining (WCM) techniques that consist in several sub processes, which are mainly information selection, pre-processing, generalization (automatically discover general patterns), analysis (validation or interpretation of mined patterns) [4].

2.1 Data Selection and Preprocessing

In order to obtain the best possible results, a web site with a relatively high amount of text is needed, and with few images, flash text, videos, audio, etc.

After selecting a web site fulfilling the above requirements, we first filter the non-useful words in order to just apply the clustering algorithm to the most relevant words (for instance, the prepositions, conjunctions and articles are omitted).

We applied later a Porter's stemming algorithm, which allows us to find the root of the words. After applying this two 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.

2.2 Using Self Organizing Feature Map

We use SOFM of the Kohonen type to extract significant patterns from the web page text content. But first, we used the Vector Space Model, in order to apply this clustering algorithm. The TFIDF was used to obtain the Web Pages feature vectors.

A toroidal topology is used to maintain the continuity of the space [9]. Then a Gaussian function that depends on the distance from the centroid is used to propagate the learning to the neighbor neurons. This function makes that the centroid neuron learn the pattern shown, and then the effect of the learning is passed to the neighborhood in smaller degree, inversely proportional to the centroid distance.

A very important expression is the Eq.(1), which is use as our similarity measure between the Web Pages and the neurons.

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

3 Reverse Clustering Analysis

After finding the clusters, we have the most commonly used words 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 clusters may contain usually do not correspond to a real web site's pages.

That is why a method to find which are the real web pages that the clusters finds relevant 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.1, 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.

3.1 Extracting the Clusters and Marking the Real Pages

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.

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.

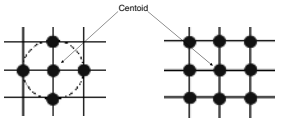

Fig. 1. Square and Circular vicinity for clusters extraction

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 vertices of the grid. That is why we take a square vicinity. For instance, if we take $r = 1$ Fig.1 then we only compare the centroid to four neurons, the vertices of the square are outside of the circular vicinity.

The experiments results show that in fact, with an 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 for 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 $a = 2$.

When we have all the clusters and its associated neurons (9 neurons with the square vicinity and $r = 1$), we compare the feature vectors of the centroids's neurons with the real pages. To do so we use the similarity measure, shown in (1), we calculate the minimum for all the pages. In other words, we obtain the most similar page on the web site for each neuron in each cluster. Formally we define the Page Reference Function $PR(n_i, p_j)$ eq.2, where ζ is the set of clusters centroids plus the associated neurons.

$$PR(n_i, p_j) = Min\{pd(n_i, p_j)\} \quad \forall j = 1, \ldots, Q \ \wedge \ i \ \epsilon \ \zeta \tag{2}$$

Using (2) we obtain the referenced pages's set.

4 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. The number of different words in the whole web site is more than 11,000 but after the preparation process (filtering and stemming) only about 4,000 words remain.

The artificial neural network used in the process was set in 100 neurons and applied the examples using 50 epochs. After the application, we found five main clusters, and 13 clusters in total Fig.2. We selected the 13 clusters found for the next process. We did so, because we intended to find few important web pages in the worst case (using all the clusters found).

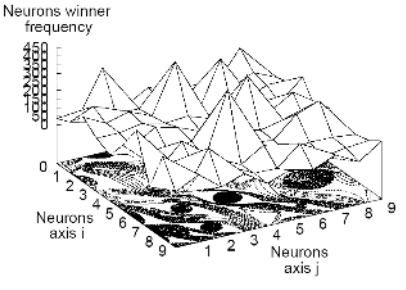

Fig. 2. SOFM for our experiments

After applying the page reference function eq.2 the results were only 8 pages see table 1. This pages are most similar to our clusters based on content, and the real web pages that we could consider as relevant web pages in the whole web site that is composed by 182 pages.

The first page found was the *School's agenda* with 21 cluster references; the second page was *students grades* with 18 references and the third page was the *Engineering Forum* with 9 references.

Table 1. Most representative real pages

Web Page	Cluster References
escuela.ing.uchile.cl/agenda/index.html	21
escuela.ing.uchile.cl/Boletin_Notas/index.html	18
escuela.ing.uchile.cl/foroING/index.html	9
escuela.ing.uchile.cl/departamentos/index.htm	6
escuela.ing.uchile.cl/novedades/novedad_alumnos.php	2
escuela.ing.uchile.cl/sd20a/alumn-sc.php	2
escuela.ing.uchile.cl/sd20a/index.html	1
escuela.ing.uchile.cl/organizaciones/estudiantes.htm	1

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

5 Conclusions

In this paper we prove that is not sufficient to find the most important words in a Web Site, because this information only helps to know which are the possible key words of the Site. However, in those cases that the management of an organization needs an effective guideline on how to focus its efforts and resources to improve the web site, additional analysis is required.

For the above propose we propose a reverse clustering analysis. In order to do so we extract the cluster's centroids using a circular and a square vicinity. We found out that the square vicinity is much better than the circular vicinity because it is more effective in identifying clusters that really exist.

An important contribution of this work is defining the page reference function as a function which compares the real web page documents with the SOFM neuron's feature vector to obtain relevant web site pages.

Furthermore, these concepts were successfully tested in a real web site where we found a set of only 5% of the web pages of the site.

References

1. Berendt, B., Spiliopoulou, M.: Analysis of navigation behavior in web sites integrating multiple information systems. The VLDB journal. **9**(2001)27–75.
2. Buyukkokten, O., Garcia-Molina, H. & Paepcke, A.: Seeing the whole in parts: text summarization for web browsing on handheld devices. Procs. 10th Int. Conf. on World Wide Web, Hong Kong. (2001) 652–662.
3. Chakrabarti, S.: Data mining for hypertext: A tutorial survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining. (2000).
4. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. SIGKDD Explorations. **2(1)**(2000)1–15.
5. Loh, S., Wives, L., de Oliveira, J. P. M.: Concept-based Knowledge Discovery in Texts Extracted from the Web. SIGKDD Explorations. **2(1)**(2000) 2(1):29–39.
6. Nielsen, J.: User Interface directions for the web. Communications of ACM **42(1)** (1999) 65–72.
7. Pal, S. K., Talwar, V., Mitra, P.: Web Mining in Soft Computing Framework: Relevance, state of the art and future directions. IEEE Transactions on Neural Networks. **13(5)** (2002)1163–1177.
8. Salton, G., Wong, A., Yang, C. S.: A vector space model for automatic indexing. Communications of the ACM archive. **18(11)** (1975)613–620.
9. Velásquez, J. D., Yasuda, H., Aoki, T., Weber, R., Vera, E.: Using self-organizing feature maps to acquire knowledge about visitor behavior in a web site. Lecture Notes in Artificial Intelligence. **2773(1)**(2003)951–958.
10. Velásquez, J. D., Weber, R., Yasuda, H., Aoki, T.: A Methodology to Find Web Site Keywords. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service Taipei, Taiwan. (2004)285–292.
11. Velásquez, J. D., Ríos, S., Bassi, A., Yasuda, H., Aoki, T.: Towards the identification of keywords in the web site text content: A methodological approach. International Journal of Web Information Systems. **1(1)** (2005)11–15.