Web Site Improvements Based on Representative Pages Identification

Sebastían 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
{srios, yasuda, aoki}@mpeg.rcast.u-tokyo.ac.jp
[2] Department of Industrial Engineering, University of Chile
jvelasqu@dii.uchile.cl

Abstract. Many researchers have successfully shown that web content mining technics and web usage mining techniques can help to find out important patterns on the content and browsing behavior in a site. However, still it is an open problem how to reach a good interpretation of the cluster results after the mining process. We propose a technique called *Reverse Clustering Analysis* (RCA) applied to a Self Organizing Feature Map in order to identify the most representative Web Pages of the Site. Then use this information to perform enhancements in the site. Our mining process is based on the combination of WCM and WUM to find out the content that is most interesting for the visitors. We successfully test our proposal in a real web site.

1 Introduction

To perform improvements in the web site content, first we need to identify which is the most relevant content of the whole web site. Several approaches to do this task have been developed [1]. Techniques like Web Content Mining (WCM), or the use of soft computing to find the representative contents have shown its effectiveness to accomplish such task [1,3].

However, even if we could distinguish the relevant content from the irrelevant content, we have no simple answers to questions like: where should we start making changes to the web site?, which are the most relevant pages of the web site?, should we modify all the web pages?.

The answer to this question is not simple. We propose the *Reverse Clustering Analysis* (RCA) to gather the most relevant web pages from the web site based on a mixed approach of WCM with WUM techniques. These allow us to find out pages which content is not only the most representative from the site but also, the content which is most relevant for the visitors. Then the site owners can use this information for enhance the content of its site.

2 Related Work

Depending on the complexity of the web site, the text content can be written by a professional writer, like a reporter or a linguistic team. Afterwards using a usability test

S. Zhang and R. Jarvis (Eds.): AI 2005, LNAI 3809, pp. 1162–1166, 2005.

[2], the final web site text content is checked before the web site goes to production. This process is human dependant, i.e., it is not possible to semi-automatize the text generation process. Moreover, the expert only have an approximate idea of what the correct content is. Therefore it is highly recommendable to establish a guideline on how to create text content based on small sub set of the whole web pages of the site.

2.1 Web Content Mining and Web Usage Mining Process

In order to extract meaning full patterns from the content of the web sites, the web content mining is a widely used technique. On the other hand, the web usage mining process is widely used to discover visitors browsing behavior. The literature mention four steps to accomplish the web content mining and usage mining processes these are: first, data selection; second, data pre-processing; third, web generalization process (automatically discover general patterns) and fourth, analysis of the patterns (validation or interpretation of mined patterns)[1].

We perform a sessionization process, in this process we take the clean logs and re-generate the sessions [3,6,7].

3 Similarity Measure and Reverse Clustering Analysis

If we assume that the degree of importance in some page content is correlated with the time spent on it by the visitors, we can state that those pages where a visitor spends more time are those more interesting to him. This way, we define a similarity measure that allows to combine the content and the usage Eq.(1) [7].

$$IVS(S^i, S^j) = \sum_{k=1}^{\iota} min(\frac{S^i_\tau(k)}{S^j_\tau(k)}, \frac{S^j_\tau(k)}{S^i_\tau(k)}) * PD(S^i_\rho(k), S^i_\rho(k)) \quad (1)$$

The expression shown in Eq.(1) compares the ι-most important pages into the sessions of two different visitors S^i and S^j. The function $PD()$ is the dot product between two vectors.

In the present work we used a SOFM to find patterns of content that is most interesting for the visitors of the site. Then the RCA process [4,5]. allow us to identify the pages which content is the most interesting from the visitors point of view.

We can write formally in a expression that we call the *Page Reference Function*. This is shown in Eq.(2) [4,5], where n_i is the i^{th} neuron in the cluster neurons set and p_j is the j^{th} real page in the whole site.

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

The result of the RCA process is a small set of real web pages whose references are greater than 0. These set of pages is called the *representative real pages set*. We can consider this set, as the set of the real pages that are the most interesting in content to the visitors. This interpretation of the RCA is based on the method used for clustering the web pages. For example, if we only use text for the clustering algorithm then the result of the RCA is also a *representative real pages set* however, in this case, the pages

in this set are the pages which contain the content that is most representative from the site (based only on its own text)[4,5].

4 Application in a Real Web Site

The whole process explained before was applied to the web site of the School of Engineering and Sciences of the University of Chile.[1]. This Web Site has 182 web pages. We use the March 2005 version of the web site to work.

In the cleaning stage we reduced the number of different words in the data from more than 11,000 to only about 4,000 words, this is after applying filtering and stemming.

On the other hand, we chose only four weeks of logs to perform the sessionization process. The length of the *ι-most important pages vector* was set in three pages. Therefore, we needed the sessions which contain at least three pages visited to create those vectors in order to apply the *3-most important pages vector*. To do so, we sorted the sessions by time spent on each page and then we only kept the three pages where the visitor spent more time.

We perform two different experiments: first, we use a SOFM of about 100 (10x10) neurons and second, we use a SOFM of 64 (8x8) neurons. The epoch parameter was set in $t = 50$ in both experiments. Then for each network, we use circular and square vicinity with $r = 1$. We use both in order to see the effects of the cluster extraction method in the final results.

4.1 First Experiment: SOFM 10x10 Neurons

The network used in this experiment is a 100 neurons network, about 55% of the size of the original space of documents and near 40% from the sessions space. We applied the RCA to discover the most representative pages using the circular and square vicinity. After applying the *Page Reference Function* in Eq.2 the results were only 18 pages using the square vicinity and 20 pages using the circular vicinity. In both cases, these pages represent almost 11% from the whole web site pages (see Table 1).

Using the square vicinity we discover 8 clusters. On the other hand, using the circular vicinity we discover 49 clusters. As mention before, not all the clusters found using circular vicinity are really clusters. If we consider the clusters found using square vicinity as the real clusters then about 80% of the clusters are not really clusters. The interesting result is that even this high difference in the number of clusters found and also, the high noise of the clusters found with circular vicinity the difference in the final result is only two pages (emovil.htm and escuela.htm) that appeared when using the circular vicinity. Moreover, these pages only has 1 reference that is why the importance of those pages is not so high.

On the other side, although the *representative web pages sets* in the experiments are almost the same, the order of those sets is severely altered depending on the method used.

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

4.2 Second Experiment: SOFM 8x8 Neurons

We used a network of 64 neurons, this is about 35% of the size of the original space of documents and near 25% from the sessions space. The number of clusters found using circular vicinity was 32 clusters, however, using the square vicinity we only discover 5 clusters. Again, although this huge difference on the number of clusters detected, we can see that the resulting set of web pages are almost the same in both approaches (see Table 2). With square vicinity we detected 16 representative web pages and with circular vicinity we found 19 pages. Once again the order of the representative web pages is severely altered by the vicinity extraction method.

4.3 Comparison Between Experiments and Discussion

If we compare the results of the first and second experiments using square vicinity, we can see that the difference in the number of clusters found is only 3 clusters. However, the final representative web pages sets are almost the same.

The same thing happen when we compare the both experiments using the circular vicinity. Moreover, the page servicios.htm appear in the first experiment but not

Table 1. First experiment results: Most representative real pages from visitors point of view using circular vicinity and square vicinity (fragment of the whole results)

Web Page (Real URL)	Square Vicinity	Web Page (Real URL)	Circular Vicinity
index_home.php	73	novedades/novedad_alumnos.php	243
novedades/novedad_alumnos.php	60	index_home.php	241
novedades.htm	25	novedades.htm	71
mapa.htm	15	mapa.htm	59
escuela/sobrelaescuela.htm	5	departamentos/index.htm	15
departamentos/index.htm	4	escuela/sobrelaescuela.htm	11
servicios/bienestar.htm	4	escuela/a_destacados.htm	10
escuela/a_destacados.htm	3	escuela/LISTA_2003.html	10

Table 2. Second experiment results: Most representative real pages from visitors point of view using circular vicinity and square vicinity (fragment of the whole results)

Web Page (Real URL)	Square Vicinity	Web Page (Real URL)	Circular Vicinity
novedades/novedad_alumnos.php	45	novedades/novedad_alumnos.php	141
index_home.php	37	index_home.php	124
mapa.htm	14	novedades.htm	55
novedades.htm	11	mapa.htm	49
departamentos/index.htm	5	escuela/a_destacados.htm	13
escuela/a_destacados.htm	4	acad_anual.htm	9
escuela/sobrelaescuela.htm	3	reglam.htm	8
baseorganizaciones.htm	2	baseorganizaciones.htm	8

in the second. Besides, the order of the representative web pages is altered in both experiments.

The most impressive result is that the four representative pages sets obtained are very similar. The representative pages are few and are independent form the size of the network and method used for extraction if we analyze the results.

However, we can not say yet which is the best size of neural network to perform the RCA. Even if the first SOFM (100 neurons) took 5 days and the second (64 neurons) took only 3 and the resulting pages are almost the same. In other web site, the results could be severly affected by the size of the SOFM selected.

5 Conclusions

The discovery of meaningful patterns and its good interpretation is a very difficult and challenging task. We propose to combine the web text mining with the usage mining to find the preferred web site content patterns from the visitors point of view.

Once we found these patterns, we realize that it is not straightforward to state which are the pages that are most representative from the clusters found in a SOFM. This information is very important to begin the improvements of the site content or to plan a strategy to focus the resources for the site enhancements.

We propose a technique called by us *Reverse Clustering Analysis* that allow us to discover which are pages that the SOFM patterns are representing.

We perform two experiments using two different SOFMs and then we applied two different clusters extraction techniques to each one (circular and square vicinity). We successfully found small *representative pages sets* which are about 11% of the whole web site documents.

References

1. Kosala, R. and Blockeel, H.:Web mining research: A survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, 2(1):1–15 (2000).
2. Nielsen, J.: User Interface directions for the web. Communications of ACM, 42(1):65–72 (1999).
3. Pal, S. K., Talwar, V. and Mitra, P.: Web Mining in Soft Computing Framework: Relevance, state of the art and future directions. IEEE Transactions on Neural Networks, 13(5):1163–1177, Sept. (2002).
4. Ríos, S., Velásquez, J., Vera E., Yasuda, H. and Aoki, T: Using SOFM to Improve Web Site Text Content, Lecture Notes in Computer Science, Volume 3611, Pages 622 - 626, Jul (2005)
5. Ríos, S., Velásquez, J., Vera E., Yasuda, H. and Aoki, T.:Establishing guidelines on how to improve the web site content based on the identification of representative pages. To appear IEEE/WI Int. Conf. on Web Intelligence, France, Sept. (2005).
6. Velásquez, J. D., Yasuda, H., Aoki, T., Weber, R. and Vera, E.: Using self-organizing feature maps to acquire knowledge about visitor behavior in a web site. Lecture Notes in Artificial Intelligence, 2773(1):951–958, Sept. (2003).
7. Velásquez, J. D., Ríos, S., Bassi, A., Yasuda, H. and Aoki, T.: Towards the identification of keywords in the web site text content: A methodological approach. International Journal of Web Information Systems, 1:11–15, March (2005).