Dissimilarities for Web Usage Mining

Fabrice Rossi[1], Francisco De Carvalho[2], Yves Lechevallier[1], and Alzennyr Da Silva[1][2]

[1] Projet AxIS, INRIA Rocquencourt, Domaine de Voluceau,
Rocquencourt, B.P. 105, 78153 Le Chesnay cedex – France

[2] Centro de Informatica - CIn/UFPE
Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brasil

Abstract. The obtention of a set of homogeneous classes of pages according to the browsing patterns identified in web server log files can be very useful for the analysis of organization of the site and of its adequacy to user needs. Such a set of homogeneous classes is often obtained from a dissimilarity measure between the visited pages defined via the visits extracted from the logs. There are however many possibilities for defined such a measure. This paper presents an analysis of different dissimilarity measures based on the comparison between the semantic structure of the site identified by experts and the clustering constructed with standard algorithms applied to the dissimilarity matrices generated by the chosen measures.

1 Introduction

Maintaining a voluminous Web site is a difficult task, especially when it results from the collaboration of several authors. One of the best way to continuously improve the site consists in monitoring user activity via the analysis of the log file of the server. To go beyond simple access statistics provided by standard web log monitoring software, it is important to understand browsing behaviors. The (dis)agreement between the prior structure of the site (in terms of hyperlinks) and the actual trajectories of the users is of particular interest. In many situations, users have to follow some complex paths in the site in order to reach the pages they are looking for, mainly because they are interested in topics that appeared unrelated to the creators of the site and thus remained unlinked. On the contrary, some hyperlinks are not used frequently, for instance because they link documents that are accessed by different user groups.

One way to analyze browsing patterns is to cluster the content of the Web site (i.e., web pages) based on user visits extracted from the log files. The obtained clusters consist in pages that tend to be visited together and thus share some semantic relationship for the users. However, visits are complex objects: one visit can be, for example, the time series of requests sent by an user to the web server. The simplest way to cluster web pages on the server based on the visits is to define a dissimilarity between pages that take into account the way pages appear in the visits. The main problem with this approach is to choose a meaningful dissimilarity among many possibilities.

In this article, we propose a benchmark site to test dissimilarities. This small site (91 pages) has a very well define semantic content and a very dense hyperlink structure. By comparing prior clusters designed by experts according to the semantic content of the site to clusters produced by standard algorithms, we can assess the adequacy of different dissimilarities to the web usage mining (WUM) task described above.

2 Web Usage Data

2.1 From log files to visits

Web usage data are extracted from web server log files. A log file consists in a sequence of request logs. For each request received by the server, the log contains the name of the requested document, the time of the request, the IP address from where the request originates, etc. Log files are corrupted by many sources of noise: web proxies, browser caches, shared IP, etc. Different preprocessing methods, such as the ones described in Tanasa and Trousse (2004a,b), allow to extract reliably visits from log files: a *visit* is a sequence of requests to a web server coming from an unique user, with at most 30 minutes between each request.

While the time elapsed between two requests of a visit is an important information, it is also quite noisy, mainly because the user might be disturbed while browsing or might doing several tasks at a time. In this paper, we don't take into account the exact date of a request. A visit consists therefore in a list of pages of the site in which the order of the pages is important.

2.2 Usage guided content analysis

As explained in the introduction, our goal is to cluster pages of a site by using the usage data. We have therefore to describe the pages via the visits. Let us consider the simple case of a web site with 4 pages, A, B, C and D. Let us define two visits, $v_1 = (A, B, A, C, D)$ and $v_2 = (A, B, C, B)$. The visits can be considered as variables that can be used to describe the pages (which are the individuals). A possible representation of the example data set is given by the following way:

	v_1	v_2
A	$\{1, 3\}$	$\{1\}$
B	$\{2\}$	$\{2, 4\}$
C	$\{4\}$	$\{3\}$
D	$\{5\}$	$\emptyset$

In this representation, the cell at row p and column v contains the set of the position of page p in navigation v. While this representation does not loose any information, compared to the raw data, it is quite difficult to use, as the variables don't have numerical values but variable size set values. Moreover, for voluminous web sites, the table is in general very sparse as most of the visits are short, regardless of the size of the web site.

2.3 Dissimilarities

Our solution consists in combining some data transformation methods with some dissimilarity measure in order to build a dissimilarity matrix for the pages. One of the simplest solutions is to map each set to a binary value, 0 for an empty set and 1 in the other case (then cell (p, v) contains 1 if and only if visit v contains at least one occurrence of page p). Many (dis)similarity measures have been defined for binary data (see e.g. Gower and Legendre (1986)). For WUM, the Jaccard dissimilarity is quite popular (see e.g. Foss et al. (2001)). It is given by

$$d_J(p_i, p_j) = \frac{|\{k | n_{ik} \neq n_{jk}\}|}{|\{k | n_{ik} \neq 0 \text{ ou } n_{jk} \neq 0\}|}, \tag{1}$$

where $n_{ik} = 1$ if and only if visit v_k contains page p_i and where $|A|$ denotes the size of set A.

For the Jaccard dissimilarity, two pages may be close even if one page appears many times in any visit whereas the other one only appears once in the considered visits. This is a direct consequence of the simple binary transformation. Using a integer mapping allows to keep more information: rather than using n_{ik}, we rely on m_{ik} defined as the number of occurrences of page p_i in visit v_k. Among the numerous dissimilarities available for integer valued data table, we retained the cosine and the tf×idf ones. Cosine is defined by

$$d_{\cos}(p_i, p_j) = 1 - \frac{\sum_{k=1}^{N} m_{ik} m_{jk}}{\sqrt{\left(\sum_{k=1}^{N} m_{ik}^2\right)\left(\sum_{k=1}^{N} m_{jk}^2\right)}}, \tag{2}$$

where N is the number of visits. The other dissimilarity is inspired by text mining: tf×idf takes into account both the relative importance of one visit for the page but also the length of the visit. A long visit goes through many pages and the information provided on each page is less specific than for a short visit. The dissimilarity is given by

$$d_{\mathrm{tf}\times\mathrm{idf}}(p_i, p_j) = 1 - \sum_{k=1}^{N} w_{ik} w_{jk}, \tag{3}$$

with

$$w_{ik} = \frac{m_{ik} \log \frac{P}{P_k}}{\sqrt{\sum_{l=1}^{N} m_{il}^2 \log\left(\frac{P}{P_l}\right)^2}}, \tag{4}$$

where P is the number of pages and P_k the number of distinct pages in visit v_k (see for instance Chen (1998)).

3 Comparison of dissimilarities

3.1 Benchmark Web site

Comparison between dissimilarities is conducted via a benchmark web site. We have chosen the site of the CIn, the laboratory of two of the authors. The site consists in dynamic web pages, implemented by servlets. The URLs of the pages are very long (more than one hundred characters) and not very easy to remember, as they corresponds to programmatic call to the servlets. Because of this complexity, we assume that most of the users will start browsing by the first main page of the site and then navigate thanks to the hyperlinks.

The site is quite small (91 pages) and very well organized in a tree with depth 5. Most of the content lies in the leafs of the tree (75 pages) and internal nodes have mostly a navigation and organization role. The hyperlink structure is very dense. A navigation menu appears on each page: it contains a link to the main page and to the 10 first level pages, as well as a link to the parents of the current page and to its siblings in the tree. There is sometimes up to 20 links in the menu which seems too complex in this situation.

The web log ranges from June 26th 2002 to June 26th 2003. This corresponds to 2Go of raw data from which 113 784 visits are extracted.

3.2 Reference semantic

We have classified the content of the site into 13 classes of pages, based on their content. Dissimilarity are compared by building clusters with a clustering algorithm and by comparing the obtained classes to the reference classes. As some of the classes are quite small (see Table 1), we also consider a prior clustering into 11 classes, where classes 9, 10 and 11 of the 13 classes partition are merged (they contain documents for graduate students).

1	2	3	4	5	6	7
Publications	Research	Partners	Undergraduate	Objectives	Presentation	Directory
8	9	10	11	12	13	
Team	Options	Archives	Graduate	News	Others	

Table 1. Size of the prior classes

4 Results

4.1 Partition quality assessment

To compare the dissimilarities presented in section 2.3, we produce homogeneous classes of pages, then we compare these classes with those resulting from the expert analysis on the site reference. For classification, we use a k-means like algorithm adapted to dissimilarity data (see Kaufman and Rousseeuw (1987); Celeux et al. (1989)) and a standard hierarchical clustering based on average linkage.

To analyze the results, we use two criteria. The first algorithm works with a user specified number of classes. We compare the obtained partition with the prior partition thanks to the corrected Rand index (see Hubert and Arabie (1985)). It takes values in $[-1, 1]$ where 1 corresponds to a perfect agreement between partitions, whereas a value equal or below 0 corresponds to completely different partition.

For the hierarchical clustering, we monitor the evolution of the F measure (see van Rijsbergen (1979)) associated to each prior class with the level of the cut in the dendrogram: the F measure is the harmonic mean of the precision and the recall, i.e. respectively of the percentage of elements in the obtained class that belong to the prior class and the percentage of the prior class retrieved in the obtained class. This method allows seeing if some prior classes can be obtained thanks to the clustering algorithm without specifying an arbitrary number of classes. In a sense this analysis can reveal specific weaknesses or skills of dissimilarities by showing whether they can discover a specific class.

4.2 Dynamic clustering

The dynamic clustering algorithm requires a prior number of classes. To limit the effects of this choice, we study the partitions produced for a number of classes from 2 to 20. The results are summarized in Table 2.

Dissimilarity	Rand index	Found classes	min F mesure
Jaccard	0.5698 (9 classes)	6	0.4444
Tf×idf	0.5789 (16 classes)	7	0.5
Cosinus	0.3422 (16 classes)	4	0.3

Table 2. Dynamic clustering results

For a global analysis (corrected Rand index), we indicate the size of the partition which maximizes the criterion. It is clear that tf×idf and Jaccard give rather close results (slightly better for the first one), whereas cosine obtains very unsatisfactory results. For a detailed analysis, we search to each prior class a corresponding class (by means of the F measure) in the set of classes produced while varying the size of the partition from 2 to 20. We indicate the number of perfectly found classes and the worst F measure for the not found classes. The tf×idf measure seems to be the best one. The classes perfectly found by other dissimilarities are also obtained by tf×idf (which finds classes 3, 4, 5, 7, 8, 9 and 12). However, we can notice that the perfectly found classes are in different partitions, which explains the relatively bad Rand indices, compared to the results class by class.

4.3 Hierarchical clustering

We carry out the same analysis for the case of hierarchical classification. We vary here the number of classes by studying all the levels of possible cut in the dendrogramme. We obtain the results summarized in Table 3.

Dissimilarity	Rand index	Found classes	min F mesure
Jaccard	0.6757 (11 classes)	3	0.5
Tf×idf	0.4441 (15 classes)	3	0.4
Cosinus	0.2659 (11 classes)	5	0.4

Table 3. Hierarchical clustering results

In general, we can notice a clear domination of Jaccard and an improvement of the results for this one. The criterion of the average link used here, as well as the hierarchical structure, seems to allow a better exploitation of the Jaccard dissimilarity, whereas the results are clearly degraded for other measures. The results class by class are more difficult to analyze and seem not to depend on measure. However, the satisfactory performances of tf×idf and cosine correspond to a good approximation of the classes for very different cutting levels in the dendrogramme: it is thus not possible to obtain with these measures a good recovery of the set of classes, whereas Jaccard is overall better.

5 Discussion

Overall, the Jaccard dissimilarity appears to be the best one for recovering the prior clusters from the usage data. The tf×idf dissimilarity gives also satisfactory results, while the cosine measure fails to recover most of the prior structure. It is important however to balance the obtained results according to the way prior clusters were designed.

The organization of the CIn web site is a bit peculiar because of the generalization of navigation pages. Class 6, for instance, contains 8 pages that describe the CIn; class 5 contains 6 pages that describe the objectives of the CIn. One of the page of class 5 acts as an introductory page to the detailed presentations of the objectives, sorted into the 5 other pages. This introduces two problems: 1) this page acts as a bridge between the general description of the CIn and the detailed description of its objectives 2) there is no simple way to avoid this page and yet to access to the description of CIn's objectives. The decision to put this page in the prior class 5 has some effect on the Jaccard dissimilarity. Indeed, as there is no simple way to view other pages of class 5 without viewing the bridge page, the Jaccard dissimilarity will tend to consider that the bridge page is close to the other pages. Moreover, as there is no way to reach the bridge page without viewing the main page of the description of the CIn, the Jaccard dissimilarity will have difficulties to separate class 5 from class 6. More generally, the tree structure of the CIn's web site and the navigation (or bridge) pages are quite difficult to handle for the Jaccard dissimilarity. It appears for instance that if we cut the dendrogram constructed via this dissimilarity in order to obtain 13 classes, we face problems that seem to be directly related to the organization of the site. For instance, class 5 and class 6 are merged into one cluster, except for two pages of class 6: the first of those pages is the general presentation of the CIn (a bridge page from the main page to presentation pages) and the localization page that gives instruction to reach the CIn (this page can be accessed directly from the main page).

Tf×idf is less sensitive to this type of problem. The 13 classes obtained from the hierarchical clustering contain one class for all pages of class 5 together with one page from class 6 (a page that describes the mission of the CIn) and another class with all the remaining pages from class 6. However, tf×idf suffers from the reduction of the relevance of long visits induced by its definition. Some pages with a low number of visits tend to appear in longer visits, from people that try to get a general view of the CIn. Clustering tends therefore to produce small classes of pages unrelated to other pages, and then to merge those classes in a quite meaning less.

The case of the cosine dissimilarity is far less clear. That bad results seem to be linked the early creation (in the hierarchical clustering) of a big cluster that mix pages from the research part of CIn's site to the pages for graduate students. The dissimilarity appears to be dominated by some long visits that tend to go through all the pages of the site. The exact source of the limitations of the cosine dissimilarity are still under investigation, however.

It is clear that it would be interesting to investigate how to modify the weighting in the tf×idf to get results closer to the one of Jaccard, will keeping the correct behavior in some circumstances. It seems also important to find a way to take into account both the structure of the site and the visits, because the global organization of the visits seem to be dominated by the structure: it would therefore be interesting to emphasize "surprising" co-occurrence of pages in a visit rather than considering all co-occurrences equality.

6 Conclusion

The results presented here give interesting insight on the adequacy of three dissimilarity measures to a clustering problem related to Web Usage Mining. While they tend to support earlier results Foss et al. (2001) that consider the Jaccard dissimilarity to be well adapted to this type of problem, they also show that the design of the benchmark and the structure of the reference web site can have strong impact on the outcome of the comparison. Further works include the comparison the chosen dissimilarities on other prior clustering of the reference web site as well as an analysis of the effect of the dissimilarities on the results of other clustering algorithm, such as an adapted version of Kohonen's Self Organized Map, as used in Rossi et al. (2005).

Bibliography

G. CELEUX, E. DIDAY, G. GOVAERT, Y. LECHEVALLIER, and H. RALAMBONDRAINY. *Classification Automatique des Données*. Bordas, Paris, 1989.

C. CHEN. Generalized similarity analysis and pathfinder network scaling. *Interacting with Computers*, 10:107–128, 1998.

A. FOSS, W. WANG, and O. R. ZAÏANE. A non-parametric approach to web log analysis. In *Proc. of Workshop on Web Mining in First International SIAM Conference on Data Mining (SDM2001)*, pages 41–50, Chicago, IL, April 2001.

J. GOWER and P. LEGENDRE. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 1986.

L. HUBERT and P. ARABIE. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

L. KAUFMAN and P. J. ROUSSEEUW. Clustering by means of medoids. In Y. Dodge, editor, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416. North-Holland, 1987.

F. ROSSI, A. EL GOLLI, and Y. LECHEVALLIER. Usage guided clustering of web pages with the median self organizing map. In *Proceedings of XIIIth European Symposium on Artificial Neural Networks (ESANN 2005)*, pages 351–356, Bruges (Belgium), April 2005.

D. TANASA and B. TROUSSE. Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59–65, March-April 2004a. ISSN 1094-7167.

D. TANASA and B. TROUSSE. Data preprocessing for wum. *IEEE Potentials*, 23(3):22–25, August-September 2004b.

C.J. VAN RIJSBERGEN. *Information Retrieval* (second ed.). London: Butterworths, 1979.