Incorporating Hyperlink Analysis in Web Page Clustering

Michael Chau
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
+852 2859-1014
mchau@business.hku.hk

Patrick Y. K. Chau
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
+852 2859-1025
pchau@business.hku.hk

Paul J. Hu
School of Acct. & Info. Sys.
The University of Utah
1645 East Campus Center Dr.
Salt Lake City, UT 84112, USA
actph@business.utah.edu

Abstract

The size of the World Wide Web is growing rapidly and it has become a very important source of information that can be useful to various academic and commercial applications. However, because of the large number of documents online, it is becoming increasingly difficult to search for useful information on the Web. General-purpose Web search engines, such as Google and AltaVista, present search results as ranked lists. Such ranked lists can only show users the first few documents of the search results and fail to give them a quick overview of retrieved document set. To address this problem, clustering techniques are often used to group documents into different topics. While traditional clustering algorithms have been applied to Web page clustering, such clustering techniques do not make use of the unique characteristics of the Web, such as its hyperlink structures. In this study, we propose to incorporate hyperlink analysis into the traditional vector space model used in document clustering. Specifically, we will introduce a new metric HFIDF based on link analysis to be used with the traditional TFIDF (term frequency multiplied by inverse document frequency) in similarity measure in clustering algorithms. The proposed study will investigate whether the use of Web structure analysis techniques improve the performance of document clustering in presenting Web search results.

Keywords: Web page clustering, Web mining, information retrieval, search engines

1. Introduction

With billions of pages contributed by millions of individuals and organizations, the World Wide Web is a rich, enormous knowledge base that can be useful to many applications. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. However, it is often not easy for users to search for information in this massive collection. Web search engines such as Google present search results as one-dimensional lists ordered by estimated relevance to the query and page quality. A major drawback of this presentation is that it fails to give users a quick "feel" for the retrieval. Users know little about the returned documents' content until they click on and read each of them. It is highly desirable for a search engine to provide such a "feel" in the form of an overview of the retrieved document set so the user can gain an overall picture of the retrieved set and explore any specific topic of interest. A high-level overview of the search results can also help users determine the relevance of retrieved document sets or reformulate queries based on feedback from the previous search.

As traditional ranked-list presentation lacks the immediate responsiveness desired for high quality information retrieval, document clustering techniques have gained increasing popularity

in recent years. These techniques automatically assign documents to different categories which are decided dynamically according to the collection (unsupervised learning). We believe that existing document clustering techniques can be improved. The reason is that most document clustering algorithms were originally developed for static text collections and rely only on the term information of each document. The unique link structure of the Web, which has been shown to be useful in other Web applications, is not used in the clustering algorithm.

The goal of this research is to study whether the use of Web structure analysis techniques improves the performance of document clustering. The rest of the paper is structure as follows. In Section 2, we review related work in Web document clustering. We discuss our research objective in Section 3, and present our proposed model in Section 4. In Section 5 we discuss our plan for evaluation study and future work.

2. Related Work

Many techniques have been used to categorize large document sets into categories to help users get a quick overview of the document sets. The vector space model, based on TFIDF scores (term frequency multiplied by inverse document frequency), is often used in these techniques to represent the documents. The similarity between a pair of documents is often calculated based on this vector space model using popular similarity metrics such as cosine distance, Euclidean distance, the Dice measure, or the Jaccard's measure (Salton & McGill, 1983).

There are in general two approaches in assigning documents into groups: *text classification* and *text clustering*. Both areas have been studied extensively in traditional information retrieval research. Text classification is the classification of textual documents into predefined categories (supervised learning), while text clustering groups documents into categories dynamically defined based on their similarities (unsupervised learning).

Many studies on text classification have been reported at SIGIR conferences and evaluated on standard testbeds (Yang & Liu, 1999). For example, the Naive Bayesian method and the k-nearest neighbor method have been widely used (e.g., McCallum et al., 1999). Neural network programs have also been applied to text classification, usually employing the feedforward/backpropagation neural network model (Ng et al., 1997; Lam & Lee, 1999). Frequencies or TFIDF scores (term frequency multiplied by inverse document frequency) of the terms are used to form a vector (Salton, 1989) which can be used as the input to the network. Based on learning examples, the network will be trained to predict the category of a document. Text classification has been applied in Web search engines such as NorthernLight (which no longer exists). The format of HTML documents and the structure of the Web also provide additional information for analysis. Examples of such information include the predicted category of neighbors (Chakrabarti et al., 1998), the anchor text pointing to a document (Chau, 2004; Chau & Chen, forthcoming), or the outgoing links to all other documents (Joachims et al., 2001). It has been shown that using such additional information improves classification results.

On the other hand, text clustering tries to assign documents into different categories that are not predefined; all categories are dynamically defined by the algorithm. There are two types of clustering algorithms, namely hierarchical clustering and non-hierarchical clustering. The *k*-nearest neighbor method and Ward's algorithm (Ward, 1963) are the most widely used

hierarchical clustering methods. Willet (1988) provided an excellent review of hierarchical agglomerative clustering algorithms for document retrieval. For non-hierarchical clustering, one of the most common approaches is the K-means algorithm. The centroid position for each cluster is recalculated every time a document is added. The algorithm stops when all documents have been grouped into the final required number of clusters (Rocchio, 1966). Another approach often used in recent years is the neural network approach. For example, Kohonen's self-organizing map (SOM), a type of neural network that produces a 2-dimensional grid representation for *n*-dimensional features, has been widely applied (Lin et al., 1991; Kohonen, 1995; Orwig et al., 1997). One example of applying text clustering to Web applications is popular search engine Clusty (www.clusty.com) which performs dynamic clustering on search results. Similarly, the Grouper system applied the Suffix-Tree Clustering algorithm to Web search results (Zamir & Etzioni, 1999). The self-organizing map (SOM) technique also has been applied to Web applications. A combination of noun phrasing and SOM has been used to cluster the search results of search agents that collect Web pages by meta-searching popular search engines or performing breadth-first search on particular Web sites (Chau et al., 2001; Chen et al., 2002). Readers are referred to Chen & Chau (2004) for a more extensive review on Web page classification and clustering.

3. Research Objective

Hyperlink structure analysis has not been widely used in Web page clustering. Preliminary results have shown that such analysis can improve the performance of Web document clustering (He et al., 2002). Because of the unsupervised nature of clustering, it is a more challenging issue to incorporate link analysis into clustering. In this study, we propose a model for clustering Web search results by incorporating Web structure analysis into document clustering algorithm. We believe that users' Web search performance, in terms of precision, recall, and efficiency, will be improved under our framework that integrates these post-retrieval analyses.

4. Proposed Model

In document clustering, a document is often represented as a vector based on the vector space model (Salton, 1989). More formally, each document D_i is represented by an *n*-dimensional vector:

$$D_i = (w_{i1}, w_{i2}, \ldots\ldots, w_{in}) \quad \text{where } w_{ij} \text{ represents the weight of term } j \text{ in document } D_i. \tag{1}$$

The traditional TFIDF (term frequency multiplied by inverse document frequency) score is used to calculate a term's weight in a document (Salton, 1989):

$$w_{ij} = tf_{ij} \times \log\left(\frac{N}{tdf_j}\right) \quad \text{where} \quad tf_{ij} = \text{the number of occurrences of term } j \text{ in } D_i, \text{ and } tdf_j = \text{the number of documents containing term } j \tag{2}$$

The similarity between documents, which is the key element in most clustering algorithms, is measured based on similarity score such as the cosine product or the Jaccard's measure between the TFIDF scores. This model, which is developed for offline document collection, does not consider Web documents' similarity in link structure. Hyperlink-related features have been shown to be useful in text classification applications (Chakrabarti et al., 1999; Chau, 2004), but have not been tested in document clustering. It has been demonstrated that if two Web pages point to the same document, or are co-cited by the same document, it is more likely that the two

documents are related to the same topic (Davison, 2000). To incorporate Web structure analysis into document clustering, we propose to add link information to the vector space model. In particular, we incorporate the set of q hyperlinks that appear in the document set as our features. If a Web page has a hyperlink to a page, the corresponding feature will be set to 1. Otherwise, it will be 0. These features will be combined with the traditional vector space model (a set of p terms) in the text clustering algorithm. In the revised vector space model, each document D_i is represented by the following vector:

$$D_i = (t_{i1}, t_{i2}, \ldots\ldots, t_{ip}, h_{i1}, h_{i2}, \ldots\ldots, h_{iq}) \quad (3)$$

where t_{ij} represents the weight of term j in document D_i,
h_{ij} represents the weight of hyperlink j in document D_i, and
$n = p + q$

When two Web pages contain the same hyperlink, we assume that the two pages are more likely to be similar to each other than pages that do not share any links. In addition, the similarity also depends on the popularity of the hyperlink. For example, many Web pages have a link to the Yahoo Web site, but pages that share this link are not more likely to be similar to each other than a pair of random pages. On the other hand, if two pages share a link that is more specific (e.g., a link to http://www.jaguar.com), then it is more likely that the two pages are about the same topic (e.g., car and vehicle). It is analogous to the TFIDF score, where terms that appear in a large number of documents (e.g., common words such as "a", "the", "is", etc.) will receive a lower score. Based on this observation, we propose a score called HFIDF (hyperlink frequency multiplied by inverse document frequency). The score is denoted as h_{ij} and is defined as follows:

$$h_{ij} = hf_{ij} \times \log\left(\frac{N}{hdf_j}\right) \quad (4)$$

where hf_{ij} = the number of occurrences of hyperlink j in D_i, and
hdf_i = the number of documents containing hyperlink j

It should be noted that there is a special condition in the calculation of hf_{ij}, where D_i is *located* at the URL specified by hyperlink j (i.e., j points to D_i). In this case, we should assign a high score to hf_{ij} such that another vector D_k containing the hyperlink j (i.e., a link to D_i) will receive a high h_{ik} score with D_i. Therefore, we revise hf_{ij} as follows:

$$hf_{ij} = \begin{cases} \max(hf_{kj}) \text{ for all } k, k \neq i & \text{if } D_i \text{ is located at hyperlink } j, \\ \text{the number of occurrences of hyperlink } j \text{ in } D_i & \text{otherwise.} \end{cases} \quad (5)$$

This HFIDF and TFIDF can be used to represent each Web page in the revised vector space model (in Equation (3)) and the similarity metric (using the cosine or Jaccard's score) used in clustering can be calculated using this model. This can apply to any Web page clustering algorithms that rely on inter-document similarity. These algorithms have been shown effective in facilitating users in their Web searching and browsing activities or traditional data clustering applications (Chen et al., 1996; Steinbach et al., 2000; Zamir & Etzioni, 1999). Traditionally, only term-features (usually TFIDF) are used in these algorithms (e.g., Orwig et al., 1996; Kohonen et al., 2000).

It should be noted that the HFIDF model should only apply to data sets where pages are supposed to share some hyperlinks (e.g., pages that are collected from closely related topics or domains). If the overlaps between hyperlinks are low, the method can be modified by using the

similarity between hyperlinks. For example, hf_{ij} can be represented by the frequencies of terms that appear in the hyperlinks instead of the hyperlinks themselves.

5. Proposed Evaluation and Future Work

We are currently implementing the HFIDF model in different clustering algorithms such as K-means clustering, self-organizing maps (SOM), Suffix-Tree clustering, and agglomerative hierarchical clustering. After successfully implementation, the first step of our evaluation plan is to evaluate the proposed clustering algorithm in a controlled experiment. Each chosen clustering algorithm enhanced by hyperlink information (HFIDF + TFIDF) will be compared with the corresponding clustering algorithm using term information (TFIDF) only. To test the performance of the methods across different domains, the clustering will be performed separately in three to four different search topic areas, such as computer, entertainment, and medical information (Chen et al., 1996). The two methods will also be tested using document collections of different sizes, ranging from hundreds of documents (the typical downloadable pages provided by a search engine) to hundreds of thousands of documents, a size tested in most other clustering applications (Chau et al., 2001; Chen et al., 1996).

The test data set will be created by fetching documents from the Web. The evaluation will be conducted for both English and non-English Web pages (e.g., Chinese or Japanese). The performance of the two methods (with and without using link information) will be compared in terms of cluster precision and cluster recall, based on the standard clusters created manually by domain experts (Roussinov & Chen, 1999). In addition, we will also compare the clusters created by our systems with those provided from open sources on the Web, such as the taxonomy provided in the Yahoo and ODP directories.

Acknowledgment

This project is supported in part by a HKU Incentive Award. We thank Jackey Ng for his help in developing the clustering system presented in this paper.

References

1. Chakrabarti, S., Dom, B., and Indyk, P. (1998). “Enhanced Hypertext Categorization Using Hyperlink,” in *Proceedings of ACM SIGMOD International Conference on Management of Data*, Seattle, Washington, USA, Jun 1998.
2. Chau, M., Zeng, D., and Chen, H. (2001). “"Personalized Spiders for Web Search and Analysis,” in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, USA, June 24-28, 2001.
3. Chau, M. (2004). “Applying Web Analysis in Web Page Filtering,” in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, Tucson, Arizona, USA, June 7-11, 2004.
4. Chau, M. and Chen, H. (forthcoming). "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," *Decision Support Systems*, accepted for publication, forthcoming.
5. Chen, H. and Chau, M. (2004) "Web Mining: Machine Learning for Web Applications," *Annual Review of Information Science and Technology*, 38, 289-329, 2004.
6. Chen, H., Chau, M. and Zeng, D. (2002). “CI Spider: A Tool for Competitive Intelligence on the Web,” *Decision Support Systems*, 34(1), 1-17, 2002.
7. Chen, H., Schuffels, C. and Orwig, R. (1996) “Internet Categorization and Search: A Machine Learning Approach,” *Journal of Visual Communication and Image Representation*, Special Issue on Digital Libraries, 7(1), 88-102.

8. Davison, B. (2000). “Topical Locality in the Web,” in *Proceedings of the 23rd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, Athens, Greece, July 2000.
9. He, X., Zha, H., Ding, C., and Simon, H. (2002). “Web Document Clustering Using Hyperlink Structures,” *Computational Statistics and Data Analysis*, 41, 19-45.
10. Joachims, T., Chistianini, N., Shawe-Taylor, J. (2001). “Composite Kernels for Hypertext Categorization,” in *Proceedings of the 18th International Conference on Machine Learning (ICML’01)*, 2001.
11. Kohonen, T. (1995). *Self-organizing Maps*, Springer-Verlag, Berlin.
12. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). “Self Organization of a Massive Document Collection,” *IEEE Transactions on Neural Networks*, 11(3), 574-585. May 2000.
13. Lam, S.L.Y, and Lee, D.L. (1999). “Feature Reduction for Neural Network Based Text Categorization,” in *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA '99)*, Hsinchu, Taiwan, Apr 1999.
14. Lin, X., Soergel, D., and Marchionini, G. (1991). “A Self-Organizing Semantic Map for Information Retrieval,” in *Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIRIR’91)*, 262-269, 1991.
15. McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (1999). “A Machine Learning Approach to Building Domain-specific Search Engines,” in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99)*, 1999, pp. 662-667.
16. Ng, H. T., Goh, W. B., and Low, K. L. (1997). “Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization,” in *Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR’97)*, 1997, pp. 67-73.
17. Orwig, R., Chen, H. and Nunamaker, J.F. (1997). “A Graphical Self-organizing Approach to Classifying Electronic Meeting Output,” *Journal of the American Society for Information Science*, 48(2), 157-170, 1997
18. Rocchio, J. J. (1966). *Document Retrieval Systems - Optimization and Evaluation*. Ph.D. Thesis, Harvard University.
19. Roussinov, D. and Chen, H. (1999). “Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques,” *Decision Support Systems*, 27(1-2), pp. 67-80.
20. Salton, G. (1989). *Automatic Text Processing*, Reading, MA: Addison-Wesley.
21. Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*, New York, USA: McGraw-Hill.
22. Steinbach, M., Karypis, G., and Kumar, V. (2000). “A Comparison of Document Clustering Techniques,” *Proceedings of the KDD Workshop on Text Mining*, 2000.
23. Ward, J. (1963). “Hierarchical Grouping to Optimize an Objection Function,” *Journal of the American Statistical Association*, 58, 236–244.
24. Willet, P. (1988). “Recent Trends in Hierarchical Document Clustering. *Information Processing and Management*,” 24(5), 577-597.
25. Yang, Y. and Liu, X. (1999). “A Re-examination of Text Categorization Methods, in *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR’99)*, 1999, pp. 42-49.
26. Zamir, O. and Etzioni, O. (1999). “Grouper: A Dynamic Clustering Interface to Web Search Results,” in *Proceedings of the 8thWorld Wide Web Conference*, Toronto, May 1999.