International Journal of Information Processing, 1(2), 49 - 57, 2007
ISSN: 0973-8215
I. K. International Publishing House Pvt. Ltd., New Delhi, India

Web Document Clustering Using Document Index Graph

B. F. Momin[a*] , P. J. Kulkarni[a†] and A. A. Chaudhari[a‡]

[a]Dept. of Computer Science and Engineering, Walchand College of Engineering, Sangli, India.

Document Clustering is an important tool for many Information Retrieval (IR) tasks. The huge increase in amount of information present on web poses new challenges in clustering regarding to underlying data model and nature of clustering algorithm. Document clustering techniques mostly rely on single term analysis of document data set. To achieve more accurate document clustering, more informative feature such as phrases are important in this scenario. Hence first part of the paper presents phrase-based model, Document Index Graph (DIG), which allows incremental phrase-based encoding of documents and efficient phrase matching. It emphasizes on effectiveness of phrase-based similarity measure over traditional single term based similarities. In the second part, a Document Index Graph based Clustering (DIGBC) algorithm is proposed to enhance the DIG model for incremental and soft clustering. This algorithm incrementally clusters documents based on proposed cluster-document similarity measure. It allows assignment of a document to more than one cluster. The DIGBC algorithm is more efficient as compared to existing clustering algorithms such as single pass, K-NN and Hierarchical Agglomerative Clustering (HAC) algorithm.

1. INTRODUCTION

The World Wide Web is rapidly emerging as an important medium for the dissemination of information related to wide range of topics. This increases need of techniques to unveil inherent structure in the underlined data. Clustering is one of these. Clustering enables one to discover hidden similarity and key concepts. Any clustering technique relies on concepts such as a data representation model, a similarity measure, a cluster model, a clustering algorithm.

Most of the document clustering techniques are based on single term based models such as vector space model [1]. These methods use similarity measures such as cosine measure or jaccard measure. These methods make use of single term analysis only and do not use word-proximity feature or phrase-based analysis.

Though researchers tried to take advantage of phrase based model using different techniques such as Inductive Logic Programming (ILP)[2], by applying different NLP techniques, the results were not encouraging. This is because extraction of such phrases is computationally intensive task. Hence researchers focused on statistical phrase extraction [3]. Statistical phrase is represented by any sequence of words that appear continuously in text. N-grams (sliding window algorithm) [4] and suffix tree model [5] are used to extract statistical phrases. N-gram method suffers from drawbacks that it only considers fixed length phrases and as document size grows, dimensionality increases tremendously. Suffix tree model finds out any length common phrases but it suffers from high redundancies stored in the form of suffixes. To overcome these disadvantages K. Hammouda and S. Kamel proposed DIG model to find out matching phrases [6].

In this paper, we first used DIG model to demonstrate effectiveness of phrase based similarity over term based similarity, and then we proposed DIGBC algorithm to cluster documents efficiently. The paper is organized as follows: Section 2 reveals the structure assigned by HTML and weights assigned to different text parts ac-

*san_bfmomin@sancharnet.in,bfmomin@yahoo.com
†pjk_walchand@rediffmail.com
‡amolac_02@yahoo.com
The work is part of the project titled "Data Mining for Very Large Databases (VLDB)" funded under Research Promotion Scheme [RPS], All India Council for Technical Education [AICTE], New Delhi, INDIA. The author B. F. Momin is Principal Investigator of this project.

this, it showed better performance. Hence this algorithm is scalable for moderate to large dataset.

Table 2: Result For Comparison of DIGBC

Algorithm	F-Measure	Entropy
	DS1	
DIGBC	0.87	0.16
Single Pass	0.67	0.02
K-NN	0.59	0.03
HAC	0.86	0.1
	DS2	
DIGBC	0.530	0.20
Single Pass	0.544	0.269
K-NN	0.531	0.170

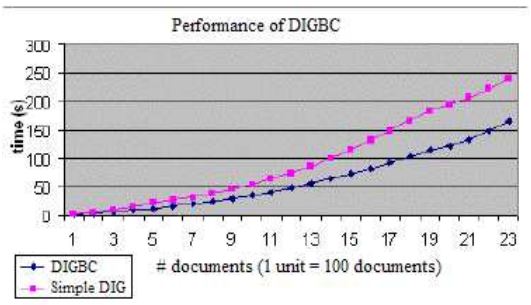

Figure 4. Performance of DIGBC

6. CONCLUSIONS

The system presented is extension of Document Index Graph Model. First part of system uses the DIG model to measure weighted phrase-based similarity between web documents. It performs phrase matching and similarity calculation between documents in a very robust, efficient, and accurate way. The quality of clustering achieved using this model significantly surpasses the traditional vector space model based approaches. The second part is extension made to DIG model. It allows us to embed clustering algorithm in DIG construction and phrase matching algorithm. To find out cluster-document similarity, a similarity measure is devised which appropriately weight the factors affecting similarity value. It shows better performance in exchange of small extra storage space. Potential applications of this framework include automatic grouping of search engine results, phrase-based information retrieval, detection of plagiarism and many others. There are a number of future research directions to extend and improve in this work. One direction that this work might continue on is to improve on the accuracy of document-cluster similarity calculation and the threshold value determination to achieve better quality.

REFERENCES

1. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. *Proc. of KDD-2000 Workshop Text Mining*, pages 3-7, August 2000.
2. Mary Elaine Calif and Raymond J. Mooney. Applying ILP-based Techniques to Natural Language Information Extraction: An Experiment in Relational Learning. *Working Notes of the IJCAI-97 Workshop on Frontier of Inductive Logic Programming*, pages 7-11, Japan, 1997.
3. M.F. Caropreso, S. Matwin, and F. Sebastiani. Statistical Phrases in Automated Text Categorization. *Technical Report IEI-B4-07-2000*, pages 1-15, Italy, 2000.
4. Lee, Min-Jae Lee. n-Gram/2L: A Space and Time Efficient Two-Level n-Gram Inverted Index Structure. *Proceedings of the 31st VLDB Conference*, pages 325-328, 2005.
5. Sven Meyer zu Eissen, Benno Stein, Martin Potthast. The Suffix Tree Document Model Revisited. *Proceedings of the International Conference on Knowledge Management*, pages 596-603, 2005.
6. Khaled M. Hammouda, Mohamed S. Kamel. Efficient Phrase-Based Document Indexing For Web Document Clustering. *IEEE Transactions on Knowledge and Data Engineering*, October 2004.
7. M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130-137, July 1980.
8. D. Lin. An Information-Theoretic Definition of Similarity. *Proc. 15th Int'l Conf. Machine Learning*, pages 296-304, 1998.

Bashirahamad F. Momin received the B.E. and M.E. degree in Computer Science and Engineering from Shivaji University, Kolhapur, India in 1990 and 2001 respectively. He is doing Ph.D. in Computer Science and Engineering at Jadavpur University, Kolkata, India. He had worked in industry. Since 1996, he is working as faculty in Department of Computer Science and Engineering at Walchand College of Engineering, Sangli, India. His research interest includes pattern recognition and its applications, data mining and soft computing techniques. He is a Principal Investigator of research project funded by AICTE, New Delhi, India. He is a student member of IEEE, IEEE Computer Society; IUPRAI USA; Advanced Computing and Communication Society, Bangalore, India.

P. J. Kulkarni completed his graduation in 1979 from Government College of Engineering, Pune, India in Electronics and Telecommunication branch. He carried out his post graduation and Ph.D. research work in Electronics Engineering from Shivaji University, Kolhapur, India in 1986 and 1993 respectively. He has to his credit 9 national papers and 12 international papers. His interest areas are Artificial Neural Network and Genetic Algorithm, Digital Image Processing, Artificial Intelligence, Information Retrieval, Data Mining, Advanced Computer Architecture. He is currently working as Professor and Head of Computer Science and Engineering Department in Walchand College of Engineering, Sangli, India.

Amol A. Chaudhari received the B. E. degree in Computer Engineering from the Pune University, India, in 2004. He received the M. E. degree in Computer Science and Engineering from Shivaji University, India, in 2006. His research interest are in Data Mining, especially Web Mining and knowledge discovery from text data using machine intelligence techniques based on efficient structures and algorithms.