COSMOS: Journal of Engineering & Technology
A Refereed Research Journal
Vol 4 / No 1 / Jan - Jun 2014 ISSN: 2231-4210

REVIEW OF WEB PAGE CLUSTERING

K.S. Mishra[1]
Payal Dixit[2]

Abstract

Web page clustering techniques partition a set of web pages into groups of similar pages which forms a single cluster. Web page clustering analysis is preprocessing steps in web mining. Under this context items to be studied are web pages. Various web page clustering techniques are used for grouping web pages in some kind of similarity. Pages in same cluster are treated as single item in web mining analysis. In this paper we have reviewed various techniques of web page clustering i.e. Semantic Clustering, Graph Partitioning for Web Page Clustering, QDC etc.

Introduction

Web page clustering methods group search results in to meaningful clusters that assist users with search refinement; but finding clusters that are semantically meaningful to users is difficult. Web search is not simple because it is hard for users to construct queries that are both sufficiently descriptive and sufficiently discriminating to find just the web pages that are relevant to the user's search goal. Queries are often ambiguous: words and phrases are frequently multi semantic and user search goals are often narrower in scope than the queries used to express them. Web page clustering algorithms identify semantically meaningful groups of web pages and present these to the user as clusters. The clusters provide an overview of the contents of the result set and when a cluster is selected the result set is refined to just the relevant pages in that cluster. We have reviewed various techniques of web page clustering i.e. Semantic Clustering, Graph Partitioning for Web Page Clustering, QDC etc. Web page graph is collection of nodes and arcs, nodes are the different web pages and arcs are the links among these pages. Cooley suggests the usage of some semantically hints to get profit in the mining process of web data, proposing that several analyses cannot be achieved without additional Meta information on structure.

Web Page Clustering Techniques

Web page clustering deals is a set of web pages in which each page similar interrelated information. Web page cluster is create for searching the website pages with respect to the words or keywords. These clusters see under mining process instead of original pages . Semantic, structure, and usage based are three clustering criteria. Clustering approaches are dealing with specific aspects of Web usage mining for the purpose of automatically discovering user profiles. i.e. erkowitz and Etzioni author described the idea of optimizing the structure of Web sites base co-occurrence patterns of pages within usage data for the site. The other author Schechter have developed techniques for using path profiles of users to predict future HTTP requests, which can be used for network and proxy caching. Similarly Spiliopoulou et al, Cooley et al, and Buchner and Mulvenna have applied data mining techniques to extract usage patterns from Web logs, for the purpose of deriving marketing intelligence. Shahabi , Yan and Nasraoui have proposed clustering of user sessions to predict future user behavior.

QDC

Daniel Crabtree described a query directed web page clustering (QDC) algorithm. It gives better clustering performance as compare to other clustering algorithm. QDC have five key aspects i.e. (i) a new query directed cluster quality guide that uses the relationship between clusters and the query, (ii) an improved cluster merging method that generates semantically coherent clusters by using cluster description similarity in additional to cluster overlap, (iii) a new cluster splitting method that fixes the cluster chaining (drifting) problem, (iv) an improved heuristic for cluster selection that uses the query directed cluster quality guide, and (v) a new method of improving clusters by ranking the pages by relevance to the cluster. Author described the algorithm and evaluation of QDC by comparing its performance against other clustering algorithms.

Semantic Clustering

Cooley in proposed the usage of some semantically hints to get profit in the mining process of web data, proposing that several analyses cannot be achieved without additional meta information on structure. Clustering of web pages are based on hierarchies and order of web pages. At lowest level of hierarchy web pages are connected with their similar characteristics and form a cluster or group of pages, these groups are connected at higher level nodes based on

[1]*Professor, SIET, Greater Noida*
[2]*Assistant Professor, SIET, Greater Noida*

semantically affinities. Webpages of similar type information and characteristics are clustered in several product families and later grouped in a cluster for all products, beside other clusters of corporative or support information can also be defined. Semantic hierarchies can be defined many different criteria, depending on the objectives and strategies of this analysis, many different collections of clusters can be provided. In such type of web page clustering techniques domain information are required, it be retrieve from domain expert or from any semantic repository. In this later case, there is a range of possible paths, from META-like information provided on the page contents, to Semantic Web principles, including also CMS-based web sites.

Graph Partitioning for Web Page Clustering

Structure and usage page clustering approaches are used to build web page graph and both are very similar. In the web page graph web pages are represent with nodes and web links shows by arcs. These links can be defined by the actual web links, in the case only web structure is considered or may be weighted by the usage of these transitions. Structure and usage page clustering are both very similar. These two approaches build a web page graph, in which nodes are the different web pages and arcs are the links among these pages. These links can be defined by the actual web links, in the case only web structure is considered or may be weighted by the usage of these transitions. The frequency of transitions analyses by the scanned web page files. In all cases translation of web clustering problem is called graph partitioning. The graph partitioning problem is NP-hard, and it remains NP-hard even when the number of subsets is 2 or when some unbalancing is allowed. There are a lot of graph partitioning algorithms.

Description of Simple partitioning methods, Spectral partitioning methods are as below.

Simple Partitioning Methods:- Purpose of techniques are to generate initial partition sets to be refined by using a local optimization strategy. These initial partitions set's quality are dependent on types of problem but they are often surprisingly good because data locality is often implicit in the vertex numbering. In this category three methods linear scheme, random scheme and scattered methods are found.

(i) linear scheme assigns vertex to sets in order (if we have n vertices and p sets, first n/p vertices will be assigned to set 1, and so on), In the random scheme vertices are randomly assigned to sets preserving balance and In scattered method vertices are processed in order with next vertex being assigned to the smallest set.

Spectral Partitioning:- It use eigenvector of a matrix constructed from the graph to decide how to partition the graph. The connection between eigenvectors and partitions may seem so surprising, but it has been proved that these techniques are quite good at finding the right general area of the graph where cuts should be done. However, they often do not behave properly obtaining fine details. It is therefore advisable to use a local refinement algorithm to improve its results, like the generalized Kernighan-Lin one that will be described next. Kernighan-Lin technique is very old technique and now many extensions and improvements have been done since 1970's. It is most popular graph partitioning techniques. The linear implementation by Fiduccia & Mattheyses is well-known of these improvements and it is often credited with the original algorithm. Kernighan-Lin usually does not find good partitions unless it is given a good initial one. This is why it is generally used as a local optimization technique, where it performs quite better.

Multilevel Kernighan Lin's method the most suitable option to deal with very large problems where high quality partitions are needed. It works by creating a sequence of increasingly smaller graphs approximating the original one, partitioning the smallest graph, and projecting this partition back through the intermediate levels.

Web Page Clustering Algorithm

A web document clustering algorithm partitions a set of web documents into groups of similar documents. These groups of similar documents are called clusters. Clustering algorithm input describe as a target number of clusters N, and a set of documents numbered 1, . . . ,M; in which each document consists of a bag of words from a word vocabulary V and a bag of tags from a tag vocabulary G. and the algorithm output as an assignment of documents to clusters. The assignment is represented as a mapping from each document to a particular cluster $c \in 1, . . ,$ N. This setup is similar to a standard document clustering task, except each document has tags as well as words. Two notable decisions are implicit in our clustering algorithm definition. First, many clustering algorithms make soft rather than hard assignments. With hard assignments, every document is a member of one and only one cluster. Soft assignments allow for degrees of membership and membership in multiple clusters. For algorithms that

output soft assignments, author map the soft assignments to hard assignments by selecting the single most likely cluster for that document. Secondly, our output is a flat set of clusters. In this paper, It focused on flat (nonhierarchical) clustering algorithms rather than hierarchical clustering algorithms. The former tend to be O (kn) while the latter tend to be O (n2) or O(n3) . Since our goal is to scale to huge document collections, we focus on flat clustering. Author seen at two broad families of clustering algorithms. The first family is based on the vector space model (VSM), and specifically the K-means algorithm. K-means has the advantage of being simple to understand, efficient, and standard. The second family is based on a probabilistic model, and specifically derived from LDA. LDA-derived models have the potential to better model the data, though they may be more complicated to implement and slower (though not asymptotically).

Conclusion

In this paper we have reviewed methods and techniques of web page clustering. In this particular case, we have used techniques of graph partitioning to speed up the process and to obtain better association rules by using only relevant information. Then we have presented some of the most important graph partitioning algorithms and the experimental scenario we have worked with. Results of this experiment demonstrate that the improvement introduced by this preprocessing step depends dramatically on the quality of input data. In this case, our study led us to the conclusion that our input data was not good enough. However, we cannot take universia responsible for this problem, as we have said before. Search engines and users' navigational habits have so much to do with these results. For future research, it would be interesting to develop the two lines that we mentioned before to see if we can state that this is the normal way web sites behave and nothing can be done at this respect or if, otherwise, there are some situations where this preprocessing step would be useful.

References

1. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pages 487–499, 1994.
2. R. Banos, C. Gil, J. Ortega, and F.G. Montoya. Multilevel heuristic algorithm for graph partitioning. In Proceedings of the 3rd European Workshop on Evolutionary Computation in Combinatorial Optimization. LNCS 2611, pages 143–153, 2003.
3. R. Banos, C. Gil, J. Ortega, and F.G. Montoya. Partition de grafos mediante optimization evolutiva paralela. In Proceedings de las XIV Jornadas de Paralelismo, pages 245–250, 2003.
4. B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. In Workshop on Web Mining at the First SIAM International
5. Conference on Data Mining, pages 7–14, 2001.
6. A. Buchner and M. D. Mulvenna. Discovering internet marketing intelligence through online analytical Web usage mining. Sigmod Record, 4(27), 1999.
7. T.N. Bui and C. Jones. Finding good approximate vertex and edge partitions is np-hard. Information Processing Letters, 42:153–159, 1992.
8. T.N. Bui and B. Moon. Genetic algorithms and graph partitioning. IEEE Transactions on Computers, 45(7):841–855, 1996.
9. J. Adibi C. Shahabi, A. M. Zarkesh and V. Shah. Knowledge discovery from users Web-page navigation. In Workshop on Research Issues in Data Engineering, Birmingham, England, 1997.
10. Robert Cooley. The use of web structure and content to identify subjectively interesting web. usage patterns. ACM Transactions on Internet Technology, 3(2), 2003.
11. B. Hendrickson and R. Leland. The Chaco User's Guide Version 2.0. Pages 1–44, 1995.
12. A. Joshi O. Nasraoui, H. Frigui and R. Krishnapuram. Mining Web access logs using relational competitive fuzzy clustering. In Eight International Fuzzy Systems Association World Congress, August 1999.
13. M. Perkowitz and O. Etzioni. Adaptive Web sites: automatically synthesizing Web pages. In Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.
14. B. Mobasher R. Cooley and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1(1), 1999.
15. Oren Zamir and Oren Etzioni. Web document clustering: a feasibility demonstration. In Sigir '98.
16. Takuya Maekawa Yutaka Yanagisawa: Web Searching for Daily Living, SIGIR'09, July 19–23, 2009.