A Review on Clustering of Web Search Result

Mansaf Alam and Kishwar Sadaf

Department of Computer Science, Jamia Millia Islamia
New Delhi, India
{mansaf_alam2002,kishwarsadaf}@gmail.com

Abstract. The over abundance of information on the web, makes information retrieval a difficult process. Today's search engines give too many results out of which only few are relevant. A user has to browse through the result pages to get the desired result. Web search result clustering is the clustering of results returned by the search engines into meaningful groups. This paper throws light and categorizes various clustering techniques that have been applied on the web search result.

Keywords: Information Retrieval, document-clustering, web search result.

1 Introduction

The information available on the web is unstructured, disorganized, dynamic and heterogeneous in nature and enormously large. Moreover the process of retrieval is highly affected by the vague query put up by the average user. Today's search engines return too many results which are not necessarily relevant to the user's need. Usually user has to traverse several search result pages to get to the desired result. A way of assisting users in finding what they are looking for quickly is to group the search results by topic. The user does not have to reformulate the query, but can merely click on the topic most accurately describing his or her specific information need. This grouping of result is called Clustering. More specifically, it is a process of grouping similar documents into clusters so that documents of one cluster are different from the documents of other clusters. There are many web clustering engines available on the web (Carrot2, Vivisimo, SnakeT, Grouper etc) which give the search results in forms of clusters. A web clustering engine takes the result, returned by the search engine as input and performs clustering and labelling on that result. This process is usually seen as complementary rather than alternative and different to the search engine [1]. The main use for web search result clustering is not to improve the actual ranking, but to give the user a quick overview of the results. Having divided the result set into clusters, the user can quickly narrow down his search further by selecting a cluster. This resembles query refinement, but avoids the need to query the search engine for each step. Web search result clustering has been the focus of IR community since the emergence of web search engine. Therefore numerous works has been done in this area. The Scatter/Gather system by [2] is held as the predecessor and conceptual father of all web search result clustering. Web Search engine is the most commonly used

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 153–159.
springerlink.com

tool for information retrieval on the web; however, its current status is far from satisfaction for several possible reasons [3]:

- Information on the Web multiply continuously;
- Different users have different requirements and expectations for search results;
- Users want whole picture of their search result on the first page of the search engine.
- Sometimes search request cannot be expressed clearly just in several keywords;
- Synonymous and polysemous words make searching more complicated;
- Users may be just interested in "most qualified" information or small part of information returned while thousands of pages are returned from search engine;
- Many returned pages are useless or irrelevant;
- Many useful information/pages are not returned for some reasons.

2 Traditional Clustering Techniques

Clustering in IR context can be classified as pre-retrieval and post-retrieval. In preretrieval clustering approach, all the documents that contain the query terms are retrieved and a clustering is done using some similarity function. The result is then presented to the user. While in postretrieval clustering approach, clustering is applied on the documents that are returned by the search engine. Clustering whether preretrieval or post retrieval can be classified into main two categories: Hierarchical clustering and Flat clustering. Although there are numerous clustering techniques but these clustering methods form the basis for other clustering techniques. Hierarchical clustering methods group the documents into a hierarchical tree structure by *Agglomerative (*bottom-up) approach or *Divisive* (top-down) approach. [4] [5]. Hierarchical methods are widely adopted, but its time complexity of O(n^2) struggle to meet the speed requirements of the web. The K-Means algorithm is the most common flat clustering and comes in many flavors [Steinbach]. Although it has O(n) time complexity, it produces a fixed number (k) of flat clusters and a "bad choice" in the random selection of initial clusters can severely degrade performance.

Above mentioned clustering techniques use the vector based representation of the document where documents are grouped only when they share exact common individual words separately. Frequent itemset clustering technique is characterized by focusing on grouping documents that share sets of frequently occurring phrases**.** In [6] Fung et al propose using the data mining notion of frequent itemsets to cluster documents. Frequent itemsets originate from association rule mining. The idea is that documents that share a set of words i.e. itemsets that appear frequently are related, and this is used to cluster documents.

The traditional clustering techniques can be applied on web search result. In case of hierarchical approach, there is tradeoff between quick result and good quality result. Since web search result clustering is an online process, time can't be traded. Usually operating on document vectors with a time complexity of O(n^2) or more, clustering more than a few hundred snippets is often unfeasible. Another problem is that if two clusters are incorrectly merged in an early state there is no way of fixing this later in the process. Finding the best halting criterion that works well with all queries can also be very difficult. In flat clustering approach, the number of clusters should be known prior to clustering. The search engine returns thousands of documents for a simple query. It is difficult to know in advance that how many clusters will be formed from the numerous documents. Several problems exist with this approach: It can only produce a fixed number of clusters (k). It performs optimally when the clusters are spherical but we have no reason to assume that documents clusters are spherical. Finally, a "bad choice" in the random selection of initial clusters can severely degrade performance.

3 Search Result Clustering

Clustering of web search results has been studied in the area of Information Retrieval (IR). The goal of clustering search result is to give user an idea of what the result contains. This idea is in the form of clusters. Clustering in context of web search result means organizing query result pages into groups based on their similarity between each other. Vivisimo, Carrot2, Kartoo etc are some of common commercial clustering engines available. Search result clustering techniques specific to the search engine result can be broadly classified as content-based and topology-based clustering. Document snippet clustering can be classified as the content-based clustering. Graph based clustering can be categorized as topology-based clustering.

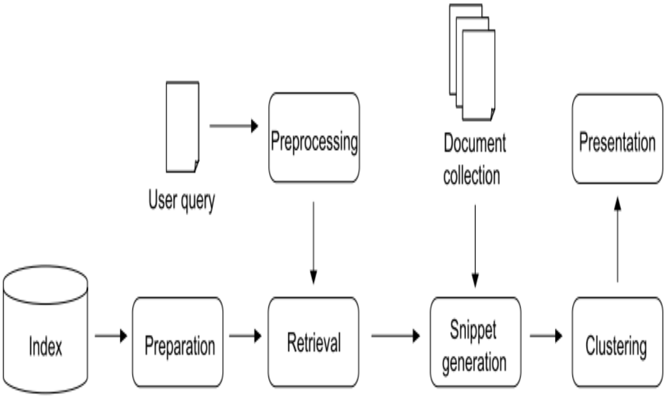

Fig. 1. A generic web search result clustering system using snippets

3.1 Document Snippet Clustering

A common technique used by clustering engines is to cluster so-called document snippets rather than entire documents. Snippets are the small paragraphs often displayed along with web search results to give the user a suggestion of the document contents. Snippets are considerably smaller than the documents (typically only

100-200 characters), thereby drastically reducing the computational cost of the clustering. This is very important since scalability and performance are major challenges for most clustering engines. When building clusters based only on short extracts from the documents, the quality of the snippets returned by the search engine naturally becomes very important. Snippet generation approaches vary from naive (e.g. first words in the document) to more sophisticated (e.g. display the passage containing the most words from the query or multiple passages containing all or most of the query keywords).

Clustering algorithms differ in their sensitivity to document length, but generally the effect of using snippets as opposed to entire documents is surprisingly small as demonstrated by [7]. Only about 15% average loss of precision for the clusters was found when using snippets rather than entire documents. The article suggests that this is caused by the search engines efforts to extract meaningful snippets concerning the user query, which reduces the noise present in the original document so much that the results do not deteriorate significantly. This further emphasizes the importance of high quality snippet extraction for snippet clustering approaches. In [8], Yao et al put forward a token-based web-snippet clustering. Direct probability graph is used to represent the snippet features. The documents which share the same features are grouped into one cluster.

An important snippet-based clustering, Suffix Tree Clustering (STC), is based on the Suffix Tree Document (STD) model which was proposed by Zamir et al [7]. The STC algorithm was used in their meta-searching engine to cluster the document snippets returned from other search engine in realtime. The similarity between documents is based on matching phrases rather than on single words only. A phrase in this context is an ordered sequence of one or more words. The STC algorithm focuses on clustering document snippets returned by the search engine, faster than standard data mining approaches. Its time complexity is linear to the number of snippets, making it attractive when clustering a large number of documents. There are numerous works available, which are derived from STC algorithm [9] [10]. In [11], authors propose an online clustering method using the STC algorithm. This algorithm groups web search results through a hierarchical, semantic and online clustering approach and named as SHOC. It consists of three steps-data collection and cleaning, feature extraction and identifying and organizing clusters. The problem with STC is the use of continuous phrases as the only features measuring similarity between documents. It can cause certain problems in languages where the positional order of parts of speech in a sentence may change. In [12], Osinski proposes a method where first, labels for clusters are defined using the input document snippets and then documents are assigned to these clusters according to their similarity with the labels.

In [13], Mecca et al use Singular Value Decomposition (SVD) on documents returned by the search engine as a whole instead of document snippets. Their algorithm has been integrated with Noodles search engine.

3.2 Graph-Based Search Result Clustering

The documents returned by the search engine in answer of a query can be looked as a subgraph of the whole web graph. The documents to be clustered can be viewed as a set of nodes and the edges between the nodes represent the relationship between them.

The edges bare a weight, which denotes the strength of that relationship. Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge-cut, i.e. the sum of the weights of the edges that are cut, is minimized. Since each edge in the graph represents the similarity between the documents, by cutting the edges with the minimum sum of weights the algorithm minimizes the similarity between documents in different clusters. The basic idea is that the weights of the edges in the same cluster will be greater than the weights of the edges across clusters. Hence, the resulting cluster will contain highly related documents. Sha et al [14] propose a web search result clustering based on lexical graph. Authors show that lexical graph structure is suitable in finding the word relationship and synonyms. The web search result is structured as a graph. They assert that their method performs better than STC and k-means. Navigelli et al [15] use graph-based clustering approach to cluster web search results. They first use graph clustering for word sense disambiguation and then cluster the results based on their semantic similarity.

Search engine like Google uses the hyperlink structure of the web to retrieve query results. This hyperlink structure is basically a directed graph, where a node represents a page and a link is characterize d by a directed edge. The pioneer works in the field of link-based web search are [16] and [17]. They have inspired many other works. Applying clustering on the hyperlink structure of web documents is an evolving area in IR research. Wang et al in [18], propose a web search result clustering which makes use of the hyperlinks between the pages and employs the HITS [16] algorithm and k-means clustering. Authorities are pages that are recognized as providing significant, trustworthy, and useful information on a topic. Hubs are index pages that provide lots of useful links to relevant content pages. PageRank uses an alternative link-analysis method. It ranks pages just by authority. Its applied to the entire web rather than a local neighborhood of pages surrounding the results of a query. In [19], Bradic uses the graph structure of the document that is preserved in the search result. Then this subgraph is partitioned to form topic related clusters.

3.3 Rank-Based and Hybrid Search Result Clustering

Clustering can be applied on the ranked result returned by the search engine or ranking can be done within each clusters formed. Leuski et al [20] propose a method where ranking and clustering are combined. The approach first traverses through the ranked list returned by the search engine until a relevant document is found. This document is then used as a cluster seed and clustering is performed on unexamined documents. Duhan et al [21] combine the power of ranking and clustering. First they cluster the documents in accordance with the query and then apply ranking within each cluster. Combining the topology and contents of the documents on the web, search result clustering can perform proficiently. Wang et al [18] propose a web search result clustering which makes use of the hyperlinks between the pages and employs the HITS algorithm and k-means clustering. Bekkerman et al [22] propose a multiagent, and bidirectional based heuristic search in the web graph to form clusters. They apply beam search in the search result graph in parallel to traditional topical clustering method on the clusters so formed. In [23], authors propose an approach based on the topology i.e. hyperlink and contents of the documents returned by the

Table 1. Search Result Clustering

Clustering Type	Input Data	General Clustering Methods
Snippet-based	Document Snippets returned by the Search Engine	STC, SHOC, SVD and other Hierarchical and flat clustering methods
Graph-based	Underlying Web graph of the search result	Graph Clustering Methods
Hybrid	Underlying web graph and the content of the documents of the search result	Combination of graph and semantic or lexical based clustering methods
Rank-based	Documents returned in the ranked search result	Various Hierarchical and Flat clustering methods

search engine. They first apply heuristic search on the web search result graph to form cluster and then perform Latent Semantic Indexing process in each cluster to derive semantic similarity between documents.

4 Discussion and Future Work

Users want a complete depiction of their search result at once. Clustering is the best possible solution for this problem. Clustering in a data mining setting has been researched for decades. Lately, document clustering used to cluster web search engine results has received much attention. Although commercial clustering engines exist, clustering is yet to be deployed on major search engines like Google. As the primary aim of a search results clustering is to decrease the effort required to find relevant information, user experience of clustering-based search result is of crucial importance. Part of this experience is the speed at which the results are delivered to the user. Ideally, clustering should not introduce a noticeable delay to normal query processing. This is presumably because of the computational overhead caused by data mining methods. It should have a low response time. Another issue related to search result clustering is labelling the clusters. The labels should be such that they must define the clusters. Unfortunately, regardless of how good the document grouping is, users are not likely to click on clusters if the labels are ill-defined. Defining accurate labels for cluster is an interesting and important area of research in the field of IR. Clustering performance is also a major issue, because web users expect fast response times.

References

[1] Carpenito, C., Osinski, S., Romano, G., Weiss, D.: A Survey of Web Clustering Engines II. ACM Computing Surveys 41(3), Article 17 (2009)

[2] Cutting, D.R., Kager, D.R., Pedersen, J.O.: Tukey JW Scatter/gather: a cluster-based approach to browsing large document collections. In: The 15th Annual International ACM Sigir Conference on Research and Development in Information Retrieval (1992)

[3] Wang, Y., Kitsuregawa, M.: Link Based Clustering of Web Search Results. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, pp. 225–236. Springer, Heidelberg (2001)

[4] Han, J., Kamber, M.: Data Mining -Concepts and Techniques. Academic Press (2001)

[5] Steinbach, M., Karypis, G., Kumar, M.: A Comparison of Document Clustering Techniques II. In: KDD Workshop on Text Mining (2000)

[6] Fung, B.C.M., Wang, K., Ester, M.: Hierarchical Document Clustering (2003)

[7] Zamir, O., Etzioni, O.: Web Document Clustering: A Feasibility Demonstration. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46–54 (1998)

[8] Yao, T., Li, J.: A Token-based Online Web-Snippet Clustering Approach based on Directed Probability Graph. Journal of Computational Information Systems 5(3), 1235–1244 (2009)

[9] Branson, S., Greenberg, A.: Clustering Web Search Results Using Suffix Tree Methods. Stanford University (2009)

[10] Janruang, J., Guha, S.: Semantic Suffix Tree Clustering. In: First IRAST International Conference on Data Engineering and Internet Technology, DEIT (2011)

[11] Zhang, D., Dong, Y.: Semantic, Hierarchical, Online Clustering of Web Search Results. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 69–78. Springer, Heidelberg (2004)

[12] Osinski, S.: A Concept-Driven Algorithm for Clustering Search Results. IEEE Intelligent Systems 20(3), 48–54 (2005)

[13] Mecca, G., Raunich, S., Pappalardo, A.: A New Algorithm for Clustering Search Result. Journal of Data & Knowledge Engineering 62(3) (2007)

[14] Sha, Y., Zhang, G.: Web Search Result Clustering Algorithm based on Lexical Graph. Journal of Computational Information Systems 5(1) (2009)

[15] Navigli, R., Crisafulli, G.: Inducing Word Senses to Improve Web Search Result Clustering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2010)

[16] Kleinberg, J.: Authoritative Sources In A Hyperlinked Environment. In: Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, SODA (1998)

[17] Page, L., Brin, S.: Web document clustering: A feasibility demonstration. In: Proceedings of SIGIR 1998, Melbourne, Australia (1998)

[18] Bradic, A.: Search Result Clustering via Randomized Partitioning of Query-Induced Subgraphs. Telfor Journal 1(1) (2009)

[19] Leuski, A., Allan, J.: Improving Interactive Retrieval by Combining Ranked Lists and Clustering. In: Proceeding of RIAO (2000)

[20] Duhan, N., Sharma, A.K.: A Novel Approach for Organizing Web Search Results using Ranking and Clustering. International Journal of Computer Applications 5(10) (2010)

[21] Wang, Y., Kitsuregawa, M.: Link Based Clustering of Web Search Results. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, pp. 225–236. Springer, Heidelberg (2001)

[22] Bekkerman, R., Zilbersteinn, S., Allan, J.: Web Page Clustering using Heuristic Search in the Web Graph. In: Proceedings of IJCAI 2007, the 20th International Joint Conference on Artificial Intelligence (2007)

[23] Alam, M., Sadaf, K.: Web Search Result Clustering using Heuristic Search and Latent Semantic Indexing. International Journal of Computer Applications 44(15) (2012)