Exploring Web Search Results Clustering

Xiaoxia Wang and Max Bramer
School of Computing, University of Portsmouth, UK
xiaoxia.wang@port.ac.uk, max.bramer@port.ac.uk

Abstract

As the number of documents on the web has proliferated, the low precision of conventional web search engines and the flat ranked search results presentation make it difficult for users to locate specific information of interest. Grouping web search results into a hierarchy of topics provides an alternative to the flat ranked list and facilitates searching and browsing. In this paper, we present a brief survey of previous work on web search results clustering and existing commercial search engines using this technique, discuss two key issues of web search results clustering: cluster summarisation and evaluation and propose some directions for future research.

1. Introduction

Traditional web search engines often return a long list of ranked links in response to user queries. Web users have to go through the long list to identify desired information. This problem gets worse as the web continues to grow. As pre-clustering of the entire corpus (e.g. Yahoo!) "would not be flexible enough to capture the themes of web search results" [5], there are many attempts using post-retrieval document clustering to bring the returned search results into order.

Clustering techniques can be used on search engines to organize retrieved results into a hierarchy of topics based on their similarities. This can help users both in locating desired information more easily and in getting an overview of the retrieved set [22]. The dynamic nature of search results introduces new challenges to document clustering technology. Zamir identified several key requirements of web search results clustering in [23]: Coherent clusters; Ease-of-browsing; Speed.

This paper is a brief survey of research trying to achieve the above key requirements. Section 2 gives an overview of work on web-based clustering techniques. Section 3 and 4 discuss two key issues of ephemeral clustering that have not been well addressed: search results clustering summarisation and evaluation. Section 5 points to future directions. An expanded version of this paper is available as [18].

2. Related work

Scatter/Gather [3] is the first to use clustering technique as a browsing tool in information retrieval. [22, 23] followed this paradigm and proposed the notion of search results clustering. They attempted to cluster "snippets" instead of full web documents. In their Grouper system, STC (Suffix Tree Clustering) treats a document as a string instead of a set of words. The two distinguishing features of STC are: linear time complexity; clustering documents according to shared phrases instead of word frequency. These make it “a substantial momentum” [20] of ephemeral clustering. The Carrot system [20] extended STC’s application into Polish Language by using different stemming techniques. SHOC [25] is based on latent semantic indexing and designed to work in Chinese. Complete phrases and continuous cluster definition were introduced to overcome STC’s limitations. LINGO [15] is a slightly modified version of SHOC. It identifies cluster labels first, and then assigns search results to different groups.

Other methods include combining links and content in a k-means framework [19]; using an N-gram based robust fuzzy relational algorithm in Retriever [9]. Microsoft [24] proposed a system to extract and rank salient phrases based on a regression model, which is trained by human labelled data, but the additional training phase is hard to adapt to the Web [6]. SnakeT [6] took advantage of two offline knowledge bases and attempted to extract sentences involving non-contiguous terms.

In addition to the above academic tools, there also has been a surge of commercial interest in implementing clustering techniques in (meta-) search engines: Vivisimo, Grokker, Clusty and Iboogie provide cluster hierarchies in addition to ranked list; Kartoo and Mooter use a network visualisation interface; Copernic and Dog-pile concentrate on supporting users on query formulation. Among the various clustering search engines, Vivisimo generates very well described thematic groups and can be considered a benchmark in current research [15], but this software is not publicly accessible. Much academic research attempts to address the search results clustering problem, but only SnakeT claims to achieve efficiency and efficacy performance close to Vivisimo.

3. Clusters summarisation

Within the field of IR, document clustering is also known as Automatic Taxonomy Generation (ATG). A key issue of ATG is how to generate appropriate labels for the hierarchical structure. ATG algorithms can be categorized into different types depending on if the taxonomies are generated by clustering words or documents, thus the process of generating clusters summarisation is also different.

3.1 Clusters-come-first approach

The traditional clustering methods such as K-means and AHC (agglomerative hierarchical clustering) fall into this category. The basic idea is representing documents as N-dimensional vectors of word frequencies, where N is the total number of distinct non-stop words in the whole document collection. Once the documents are converted into vectors, appropriate similarity measures and clustering algorithms can be chosen for clustering. Further details can be found in [4]. The set of top ranking words with high occurrence frequency within the cluster can be used as cluster summarisation. We briefly described STC algorithm in section 2. This algorithm is also based on clustering documents but captures shared phrases (contiguous terms) as labels of clusters instead of words with high frequency. The recently developed SnakeT system attempted using knowledge base to enrich the collection of words extracted from snippets to attain cluster labels. Approximate terms (involving non-contiguous terms) were extracted as hierarchy labels and attained a good performance.

From the above examples, we can see the approaches based on clustering documents form concept hierarchy by clustering documents first, then extracting terms from documents within the cluster as cluster summarisation. So it is a clusters-come-first approach and is also called polythetic clustering since the clusters are labelled by multiple concepts (terms).

3.2 Summarisation-come-first approach

The ATG algorithms based on clustering words focus on organizing words according to thesaural relationship [10]. Some are based on analysing the phrase in which a term occurs to infer the term relationships [7]. Some use phrasal analysis in addition to knowledge base to organize terms into a concept hierarchy [21]. A brief description of how they utilise different phrase analyse methods can be found in [16]. Some other methods using term co-occurrence to produce structure of related terms are surveyed in [10].

The approaches in this category first form a concept hierarchy by analysing the relationship between words, then assign documents to appropriate nodes (topics and subtopics). They are also called monothetic clustering as the cluster assignment is based on a single feature. Monothetic clustering is claimed to be well suited for generating hierarchies for search results because user can easily understand clusters described by a single feature [10]. But we believe there have been no formal experimental comparisons between monothetic and polythetic search results clustering, possibly due to the lack of standard evaluation measures in this application area.

There is another approach called co-clustering, which clusters words and documents simultaneously. The details of several examples (FCoDoK, FSKWIC, RPSA) are covered in [10].

4. Clustering evaluation

An important aspect of cluster analysis is the evaluation of clustering results. In this section we introduce the commonly used document clustering evaluation measures and briefly review ephemeral clustering evaluation in the literature.

Clustering results can be evaluated externally by comparing with pre-defined classes in several ways: purity, entropy and mutual information, which are defined in [2]. If a cluster is viewed as the result of a query for a particular category, the F-measure [17] can also be used to evaluate the document clustering results. Because the search results are generated dynamically, there are no predefined categories to compare with. One solution is to manually classify and assign labels to documents [14], or manually assign relevance judgement to each document [22] so that the effectiveness for information retrieval can be evaluated.

The second approach is based on internal criteria when ground truth is not available. Because the goal of clustering is to group a set of points into clusters so that points in the same cluster are more similar than points in different clusters [8], the clustering results can be evaluated by calculating the ratio of the average inter-cluster to intra-cluster distance. [1] proposed two criteria: compactness and separation, which reflect the inter-cluster and intra-cluster similarity. Additionally, how well the labels predict the cluster contents can be measured by Expected Mutual Information Measure (EMIM) [13].

There are also other methods involved in evaluating clustering hierarchy quality: [16] performed a user study to judge the quality of relationship between child and its parent nodes in the hierarchy. [12] also use parent-child pair in evaluation but they are only interested in how many common pairs are shared by two hierarchies. These are classified as relative evaluation measures because only similarity between two hierarchies is of interest in this scenario.

The research from IBM [11] combines the above measures to evaluate clustering hierarchy in terms of six desirable properties. They adopt *compactness* criteria from [1] and interpret separation as *sibling node distinctiveness*, which is more suitable for hierarchy evaluation. The idea of using c*overage* and *reach time* metrics was originally from [12, 13], but the first metric was called reachability in their work. The *reach time* metric measures how quickly a user can reach all relevant documents. *Node label predictiveness* and *general to specific* are difficult to quantify thus user study is used to rate the hierarchy.

5. Summary and future research

In this paper, we present an overview of web search results clustering approaches and discuss two important aspects: clusters summarisation and clustering evaluation. The dynamically generated search results introduce many challenges to clustering techniques. We propose some directions for future research. First, the goal of search results clustering is to provide an efficient searching and browsing tool for online use, thus necessitating accurate cluster summarization. SnakeT's performance makes us believe that using off-line information to aid clusters label extraction/generation is a promising research direction. Second, standard evaluation methods need to be developed so that the monothetic clustering and polythetic clustering algorithms can be compared under a general framework. How to choose suitable objective and subjective measures to make an effective combination in this application area remains an open question.

References

[1] Berry, A. and Linoff, G. (1997) Data Mining Techniques: For Marketing, Sales, and Customer Support, 1 ed., New York, USA, Wiley.

[2] Cover, T. M. and Thomas, J. A. (1991) Elements of Information Theory. Wiley-Interscience.

[3] Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W. (1992) Scatter/Gather: a cluster-based approach to browsing large document collections. Proceedings of the 15th International ACM SIGIR Conference on research and development in information retrieval.

[4] Everitt, B. (1974) Cluster Analysis. London: Heinemann Educational (for) the Social Science Research Council.

[5] Ferragina, P. and Gulli, A. (2004) The Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book Snippets. Technical report, RR04-04 Informatica, Pisa, Italy.

[6] Ferragina, P. and Gulli, A. (2005) A personalized Search Engine Based On Web-Snippet Hierarchical Clustering. Proceedings of the 14th International World Wide Web Conference.

[7] Grefenstette, G. (1994) Explorations in Automatic Thesaurus Discovery. Kluwer.

[8] Jain, A. K. and Dubes, R. C. (1988) Algorithms for Clustering Data. Prentice Hall, New Jersey.

[9] Jiang, Z. H., Joshi, A., Krishnapuram, R. and Yi, L. Y. (2002) Retriever: improving web search engine results using clustering. In Managing Business and Electronic Commerce.

[10] Krishnapuram, R. and Kummamuru, K. (2003) Automatic taxonomy generation: issues and possibilities. Proceedings of fuzzy sets and systems (IFSA), Volume 2715, pages 52-63. Springer-Verlag.

[11] Kummamuru, K., Lotlikar, R. and Roy, S. (2004) A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. SIGIR’04.

[12] Lawrie, D. J. and Croft, W. B. (2000) Discovering and comparing topic hierarchies. Proceedings of RIAO conference, pages 314-330.

[13] Lawrie, D. J. and Croft, W. B. (2003) Generating Hierarchical Summaries for Web Searches. SIGIR’03, Toronto, Canada.

[14] Leouski, A. V. and Croft, W. B. (1996) An Evaluation of Techniques for Clustering Search Results. Technical report IR-76, Department of computer science, University of Massachusetts, Amherst.

[15] Osinski, S. (2003) An algorithm for clustering of web search results. Master Thesis. Poznan University of Technology, Poland.

[16] Sanderson, M. and Croft, W. B. (1999) Deriving concept hierarchies from text, Proceedings of SIGIR, pages 206-213.

[17] Van Rijsbergen, C. J. (1979) Information Retrieval. Butterworth-Heinemann Ltd

[18] Wang, X and Bramer, M. (2006), A review of web search results clustering. University of Portsmouth School of Computing Technical Report.

[19] Wang, Y. and Kitsuregawa, M. (2001) Link based clustering of web search results. Proceedings of the 2nd International Conference on Web-Age Information Management, Xi’An, P.R.China.

[20] Weiss, D. (2001). A clustering interface for web search results in Polish and English. Master Thesis, Poznan University of Technology.

[21] Woods, W. A. (1997) Conceptual Indexing: A better way to organize knowledge, a Sun Labs technical report: TR-97-61. Editor, Technical Reports, 901 San Antonio Road, Palo Alto, California 94303, USA.

[22] Zamir O. and Etzioni, O., (1998) Web document clustering: a feasibility demonstration. SIGIR 98, Melbourne, Australia.

[23] Zamir O. and Etzioni, O., (1999) Grouper: A dynamic clustering interface to web search results, Proceedings of the eighth international world wide web conference, Toronto, Canada.

[24] Zeng, H. J., He, Q. C., Chen, Z., Ma, W. Y. and Ma, J. W. (2004) Learning to Cluster Web Search Results. SIGIR’04, Sheffield, South Yorkshire, UK.

[25] Zhang, D. and Dong, Y. S. (2001). Semantic, Hierarchical, Online Clustering of Web Search Results. In ACM 3rd Workshop on Web Information and Data.