New Research Directions in Search Results Clustering

Claudio Carpineto, Andrea Bernardini,
Massimiliano D'Amico, Gianni Romano
Fondazione Ugo Bordoni
Rome, Italy
{carpinet, abernardini, romano}@fub.it
mas.damico@gmail.com

ABSTRACT

We discuss which are the main research themes in the field of search results clustering and report some recent results achieved by the Information Mining group at Fondazione Ugo Bordoni.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

1. SEARCH RESULTS CLUSTERING

Search results clustering organizes search results by topic, thus providing a complementary view to the flat list returned by document ranking systems. This approach is especially useful when document ranking fails. Besides allowing direct subtopic access, search results clustering reduces information overlook, helps filtering out irrelevant items, and favors exploration of unknown or dynamic domains.

Search results clustering is related to, but distinct from, conventional document clustering. When clustering takes place as a post-processing step on the set of results retrieved by an information retrieval system on a query, it may be both more efficient, because the input consists of few hundred of snippets, and more effective, because query-specific text features are used. On the other hand, search results clustering must fulfill a number of more stringent requirements raised by the nature of the application in which it is embedded; e.g., meaningful cluster labels, low response times, short input data description, unknown number of clusters, overlapping clusters.

A comprehensive survey of search results clustering, including issues, techniques, and systems is given in [4]. In the remainder of this paper we point out interesting research directions.

1.1 Description-centric clustering algorithms

Given that search results clustering systems are primarily intended for browsing retrieval, a critical part is the quality of cluster labels, as opposed to optimizing only the clustering structure. In fact, the algorithms for performing search results clustering cover a spectrum ranging from data-centric to description-centric techniques, depending on whether the priority is given to cluster formation or cluster labeling.

One of the most recent examples of the latter category is KeySRC (Keyphrase-based Search Results Clustering), described in [1]. This system generates clusters labeled by keyphrases. The keyphrases are extracted from the generalized suffix tree built from the search results and merged through an improved hierarchical agglomerative clustering procedure, representing each phrase as a weighted document vector and making use of a variable dendrogram cut-off value. KeySRC is available at http://keysrc.fub.it.

1.2 Performance evaluation measures

Internal validity measures and comparison with ground truth results are two common ways of evaluating clustering partitions, but they have the disadvantage that the performance of the system in which the document partition is encompassed is not explicitly taken into account. As the intended use of search results clustering is to find documents relevant to the single query's subtopic, it may be more convenient to evaluate the performance on a retrieval oriented task. However, the classical measures related to subtopic retrieval, such as subtopic recall, subtopic precision, and subtopic MRR, assume that the system output consists of a ranked list and thus they are not directly or easily applicable to clustered results, Furthermore, they strictly focus on subtopic coverage; i.e., retrieving at least one relevant document per subtopic.

To address these limitations, we presented a new evaluation measure inspired by Cooper's expected search length: *Subtopic Search Length under k document sufficiency* (kSSL). The idea is to consider the number of elements (cluster labels or search results) that the user must examine to retrieve a specified number (k) of documents relevant to the single subtopics of a query. The shorter the search length, the better the system performance. It is assumed that both cluster labels and search results are read sequentially from top to bottom, and that only cluster with labels relevant to the subtopic at hand are opened. The main advantages of kSSL are that it is suitable for both ranked lists and clustered results and that it allows evaluation of full subtopic retrieval (i.e., retrieval of multiple documents relevant to a query's subtopic). A full description of kSSL is given in [1].

1.3 Test collections

There is almost a complete lack of test collections with subtopic relevance judgments. Two exceptions are the collections developed at the TREC Interactive track, which is

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

small and primarily focuses on the instances of a given concept (e.g., 'what tropical storms – hurricanes and typhoons – have caused property damage and/or loss of life'), and at Image CLEF, which is mainly about geographical diversity of photos associated with a given topic (e.g., 'images of beaches in Brazil').

We created two new test collections for evaluating subtopic retrieval, namely AMBIENT and ODP-239. AMBIENT (AMBIguous ENTries) consists of 44 topics extracted from the *ambiguous* Wikipedia entries, each with a set of subtopics and a list of 100 ranked search results manually annotated with subtopic relevance judgments. AMBIENT is fully described in [3] and is available at http://credo.fub.it/ambient.

ODP-239 consists of 239 topics, each with about 10 subtopics and 100 documents associated with the subtopics. The topics, subtopics, and their associated documents were selected from the Open Directory Project (www.dmoz.org). The distribution of documents across subtopics reflects the relative importance of subtopics. ODP-239 can be downloaded from http://credo.fub.it/odp239.

1.4 Applications in mobile search

The features of search results clustering appear very suitable for mobile information retrieval, where a minimization of user actions (such as scrolling and typing), device resources, and amount of data to be downloaded are primary concerns. Furthermore, such features seem to nicely comply with the most recently observed usage patterns of mobile searchers.

We implemented two mobile clustering engines (for PDAs and cellphones) and evaluated their retrieval performance [3]. We found that mobile clustering engines can be faster and more accurate than the corresponding mobile search engines, especially for subtopic retrieval tasks. We also found that although mobile retrieval becomes, in general, less effective as the search device gets smaller, the adoption of clustering may help expand the usage patterns beyond mere informational search while mobile.

1.5 Meta search results clustering

Just as the results of several search engines can be combined into a meta search engine, the outputs produced by distinct clustering engines can be merged into a meta clustering engine. Currently, there are many different web clustering engines but no attempts has still been made to combine them, to the best of our knowledge.

We studied the problem of meta search results clustering, that has unique features with respect to the relatively well understood field of general meta clustering. After showing that the combination of multiple search results clustering algorithms is empirically justified, we developed a novel meta clustering algorithm that maximizes the agreement between the outputs produced by the input clustering algorithms [5]. The novel meta clustering algorithm applied to web search results is both efficient and effective.

1.6 Clustering versus diversification of search results

Re-ranking search results to promote diversity of top elements is another approach to subtopic retrieval that has received much attention lately. Clustering and diversification of search results are thus different techniques with a similar goal, i.e., addressing the limitations of the probabilistic ranking principle when a topic has multiple aspects of potential interest and the relevance criterion alone is not sufficient.

These two techniques have not been compared so far. We performed a systematic evaluation of several clustering and diversification algorithms using multiple test collections and evaluation measures [2]. It turns out that diversification works well when one wants to get a quick overview of documents relevant to distinct subtopics, whereas clustering is more useful when one is interested in retrieving multiple documents relevant to each subtopic.

1.7 Other research directions

There are further directions that have started to be explored recently by other research groups. They mainly aim to improve the quality and effectiveness of the search results clustering process. A non-exhaustive list is given below.

– Personalized search results clustering

– Integrating external knowledge (e.g., thesauri, metadata, folksonomies, past queries) with search results clustering

– Semi-supervised search results clustering

– Temporal search results clustering

– Visualization of clustered search results

– Search results clustering and faceted hierarchies

2. REFERENCES

[1] A. Bernardini, C. Carpineto, and M. D'Amico. Full-Subtopic Retrieval with Keyphrase-Based Search Results Clustering. In *Proceedings of 2009 IEEE/WIC/ACM International Conference on Web Intelligence, Milan, Italy*, pages 206–213. IEEE Computer Society, 2009.

[2] C. Carpineto, M. D'Amico, and G. Romano. Evaluating subtopic retrieval system performance: clustering versus diversification. Submitted.

[3] C. Carpineto, S. Mizzaro, G. Romano, and M. Snidero. Mobile Information Retrieval with Search Results Clustering: Prototypes and Evaluations. *Journal of American Society for Information Science and Technology (JASIST)*, 60(5):877–895, 2009.

[4] C. Carpineto, S. Osiński, G. Romano, and D. Weiss. A survey of Web clustering engines. *ACM Computing Survey*, 41(3), 2009.

[5] C. Carpineto and G. Romano. Optimal Meta Search Results Clustering. Submitted.