Evaluating subtopic retrieval methods: clustering versus diversification of search results

Claudio Carpineto, Massimiliano D'Amico, Giovanni Romano

Fondazione Ugo Bordoni, Rome

Abstract

To address the inability of current ranking systems to support subtopic retrieval, two main post-processing techniques of search results have been investigated: clustering and diversification. In this paper we present a comparative study of their performance, using a set of complementary evaluation measures that can be applied to both partitions and ranked lists, and two specialized test collections focusing on broad and ambiguous queries, respectively. The main finding of our experiments is that diversification of top hits is more useful for quick coverage of distinct subtopics whereas clustering is better for full retrieval of single subtopics, with a better balance in performance achieved through generating multiple subsets of diverse search results. We also found that there is little scope for improvement over the search engine baseline unless we are interested in strict full-subtopic retrieval, and that search results clustering methods do not perform well on queries with low divergence subtopics, mainly due to the difficulty of generating discriminative cluster labels.

Keywords:
subtopic retrieval, clustering, search results re-ranking, diversification

1. Introduction

In the past few years increasing research has been done on subtopic retrieval, i.e. the task of retrieving documents relevant to specific aspects (or facets, or meanings) of a given query. This task is motivated by the difficulties encountered by users in discriminating the information contained in a conventional list of search results, due to their redundancy and lack of structure. The number of real user queries affected is potentially large, partly because informational queries have been estimated to account for 80% of web queries (Jansen et al., 2008), and partly because today virtually any web query expressed by very few words may have multiple interpretations, depending on the user intent or on the context in which it is issued. Furthermore, there is evidence that the retrieval of more information on specific subtopics of interest is often the primary goal of the efforts of searchers (Xu and Yin, 2008).

Most search systems are not able to deal with subtopic retrieval because they try to present results in descending order of probability of relevance to the user query. The origin of this approach is generally credited to Robertson (1977)'s *probability ranking principle*, although Robertson in his paper recognized the limitations of using a ranking scheme that works document-by-document: 'the major problem appears to lie in the way the principle considers each document

independently of the rest'. It has been shown that assuming that the relevance judgment of any document is independent from the relevance of other documents is suboptimal for the subtopic retrieval task (Gordon and Lenk, 1999). We need instead to take into account the semantic similarity between the documents presented to the user.

One well known strategy to capture the thematic structure of search results is to use clustering. Search results clustering is related to, but distinct from, pre-retrieval applications of document clustering – those aimed at improving the efficiency of query processing (Altingövde et al., 2008) or finding a more effective document ranking (Liu and Croft, 2004). Search results clustering groups the post-retrieval results (usually one hundred) and gives the user the ability to choose the groups of interest in an interactive manner, thus providing direct subtopic access. If the items that relate to the same subtopic have been correctly placed within the same cluster and if the user is able to choose the right cluster from the cluster labels, such items can be accessed in logarithmic rather than linear time. Following Vivísimo's[1] 'description comes first' motto, in the last few years many new search results clustering algorithms optimized for usability have been proposed and several research and commercial web clustering engines have been deployed, surveyed by Carpineto et al. (2009b).

To illustrate, in Figure 1 and Figure 2 we show the output produced by KeySRC[2], a recent web clustering engine (Bernardini et al., 2009), in response to two example queries (as of March 2010). In each figure, the clusters created by the system analyzing the first about 100 retrieved documents are shown on the left, with the original list of search results being displayed on the right. Each cluster is described by a label and by the number of documents contained in that cluster, which can be seen by clicking on the label.

The two queries, used as exemplifications throughout the paper, are 'excalibur' and 'information retrieval'. The query 'excalibur' is ambiguous, with several meanings such as the the name of the sword belonging to King Arthur, the movie, the car, and many others. The first 10 documents returned by the search engine (see Figure 1) cover quite a few distinct meanings, probably because major search engines have started to address the diversity issue in their first results page, although the first two documents are nearly duplicate and the meaning of documents is sometimes not clear from their title (e.g., those related to Arthurian legend). In comparison, the clusters created by KeySRC allow the user to discriminate between a similar or somewhat smaller number of distinct meanings with more ease; e.g., by reading the cluster labels rather than the document titles or snippets.

The query 'information retrieval' is a broad query. The top hits returned by the search engine (see Figure 2) mostly cover introductory (general) documents about information retrieval (e.g., wikipedia, answers.com, glossary) that presumably present a high redundancy and do not look into single aspects. In addition, there are two duplicates. By contrast, the themes identified by the clusters seem quite useful and give a good overview of some of the most interesting subtopics of information retrieval. Perhaps more importantly, for both queries the user can see all the documents that refer to any generated cluster at once, even though they were scattered through the original search list.

The other main approach to subtopic retrieval is diversification of search results. The inner mechanisms of clustering and diversification are similar, but the search interfaces are quite different. Clustering groups similar documents and labels the cluster; diversification considers

[1] Subsequently renamed as Clusty and, more recently, as Yippy (http://yippy.com).

[2] http://keysrc.fub.it/

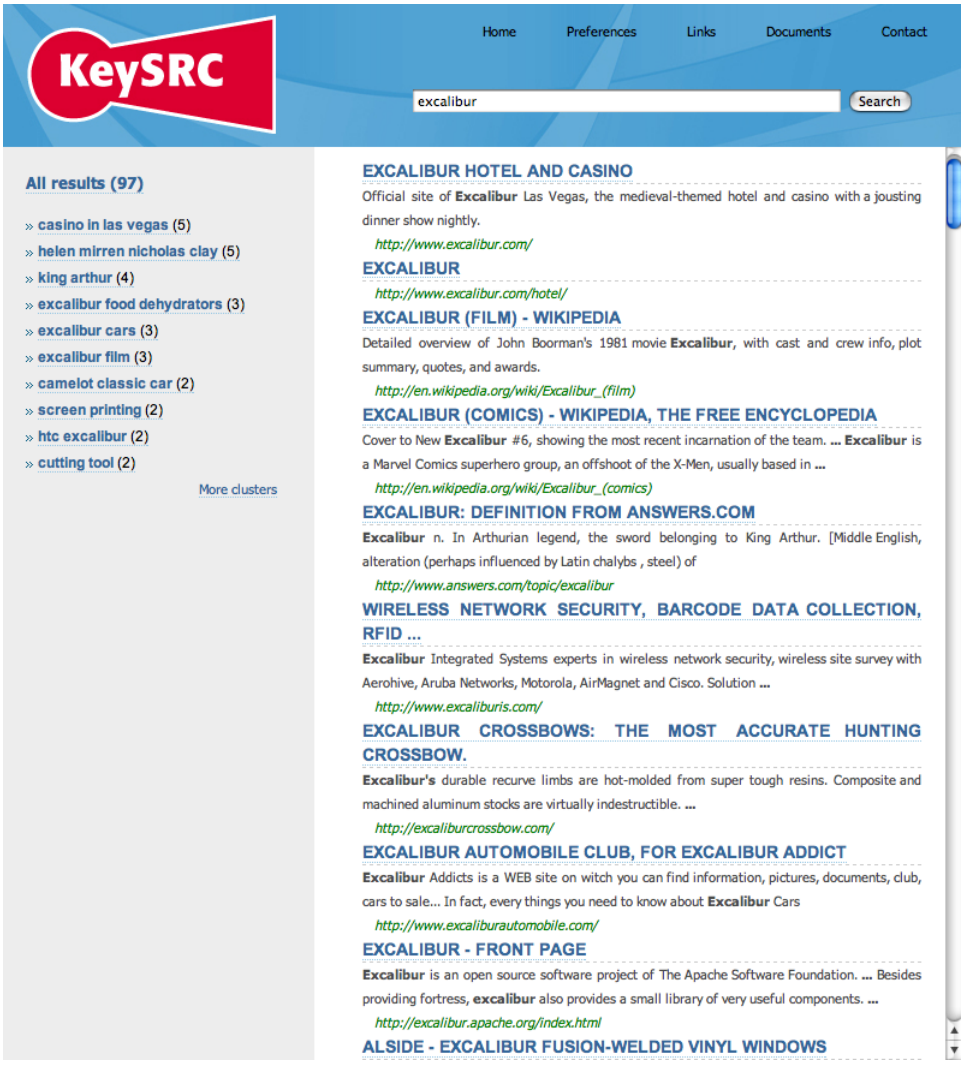

Figure 1: Clustered results for the query ‘excalibur’.

Figure 2: Clustered results for the query ‘information retrieval’.

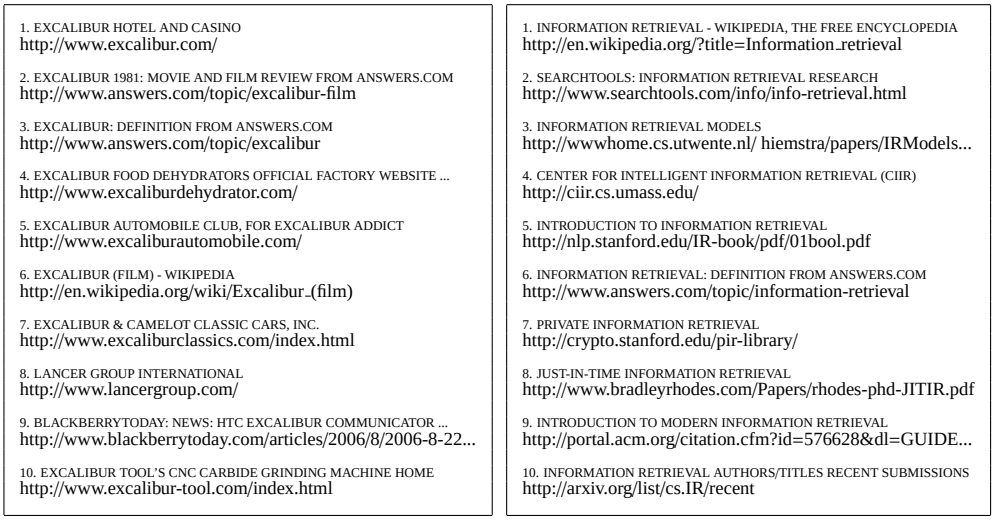

Figure 3: Re-ordered results for the query 'excalibur' (left) and 'information retrieval' (right).

inter-document similarity when ranking, but just presents a result page like any normal search engine.

Diversification tries to optimize the relevance of results at the document set level instead of at the single document level; e.g., (Carbonell and Goldstein, 1998), (Zhai et al., 2003), (Swaminathan et al., 2008), (Agrawal et al., 2009). This is done by re-ranking the list of search results with the goal of delivering a first results page with reduced redundancy. The diversification of results is usually achieved by way of re-ranking functions that trade relevance for novelty: the documents to be selected at each step are compared according to both their estimated relevance to the query and their dissimilarity in content to the subtopics covered by the documents already chosen.

In Figure 3 we show the results of a diversification method on the same queries considered above, namely 'excalibur' and 'information retrieval'. In this particular case, the diversification of search results was performed using again the output produced by KeySRC through a method described below in the paper. Figure 3 suggests that the re-ordered lists were more useful for subtopic retrieval than the original lists. There was an increase in the number of subtopics covered by the titles both for 'excalibur' and 'information retrieval', and the information retrieval subtopics were more focused.

Although there is a rich body of literature on both clustering and diversification, there have been so far no evaluation studies of their relative performance. Earlier work has compared either clustering methods e.g., (Ferragina and Gulli, 2005), (Carpineto et al., 2009a), or diversification methods, e.g., Zuccon and Azzopardi (2010), Santos et al. (2010). To the best of our knowledge, this is the first attempt to compare clustering and diversification of search results using a unified evaluation framework.

In order to allow cross-comparison between methods producing different types of outputs (i.e., lists versus partitions), we adapt subtopic retrieval evaluation measures developed for ranked lists to clustered results. The considered evaluation measures focus on complementary aspects of the subtopic retrieval performance, namely partial subtopic coverage and full subtopic retrieval.

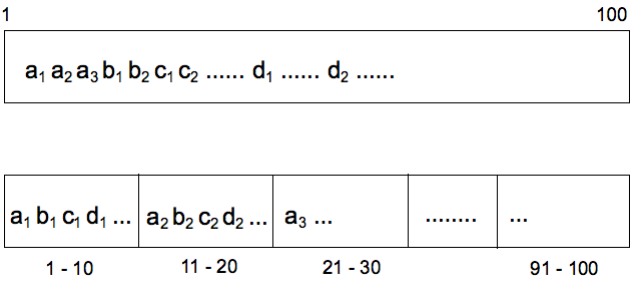

Figure 4: Illustration of minimal document sets retrieval.

In addition, we use two test collections specifically devised for evaluating the subtopic retrieval effectiveness of systems that post-process search results. The two collections contain ambiguous and broad queries, respectively, and are equipped with document-level relevance judgments per subtopic. They have unique features, compared to the other few available data sets for subtopic retrieval evaluation. We compare three strategies of subtopic retrieval methods, namely clustering, diversification, and minimal subsets of results.

The third strategy, i.e., generation of minimal subsets of results (Dai and Srihari, 2005), is an extension of diversification. Rather than creating one small subset of diverse documents at the top of the search results list, the goal is to re-order the whole list in such a way that it becomes a concatenation of multiple subsets of diverse documents. This approach is illustrated in Figure 4. The input consists of 100 documents belonging to subtopics a, b, c, d (among others), and the output is formed by 10 subsets of diversified results, with 10 elements each.[3]

The results of our experiments suggest that different strategies cater for complementary aspects of subtopic retrieval performance. We identify the suitability of each strategy for the various evaluation measures being analyzed and study which are the most critical factors affecting their retrieval effectiveness. Clustering has the greatest potential for the task where lists of results, whether diversified or not, usually fail (i.e., full subtopic retrieval), but the performance of clustering is more sensitive to the similarity of subtopics. In contrast, the diversification methods are more suitable for quick coverage of subtopics, but in practice it may be difficult for them to improve over lists produced by modern search engines.

The remainder of the paper has the following organization. We first address the two main issues involved in the evaluation of subtopic retrieval performance, namely the definition of suitable measures and test collections that accommodate user needs at a finer grain than provided by conventional ranking systems. Then we discuss the main approaches to subtopic retrieval and detail the specific methods used in the experiments. The next sections are dedicated to the experimental results achieved on ambiguous and multi-topic queries, including an estimation of subtopic dissimilarity across test collections and a query-by-query analysis. We finally provide some conclusions.

[3]This is a simplifying assumption for the sake of illustration. In general, the size of each subset is not fixed; e.g., it is determined by a similarity threshold.

2. Evaluation measures

In recent years, several measures have been proposed to evaluate subtopic information retrieval, the best known of which are probably *subtopic recall at rank n* and *subtopic precision at recall r* (Zhai et al., 2003). These two measures adapt classical information retrieval measures based on recall and precision by focusing on partial subtopic coverage, i.e., the ability to find at least one relevant document for as many different subtopics as possible. It is also possible to assess full subtopic retrieval, i.e. the ability to find multiple relevant documents for any subtopic. The latter complementary aspect can be evaluated using *subtopic search length under k document sufficiency* ($kSSL$), measuring the average number of items that must be examined before finding a sufficient number (k) of documents relevant to a query's subtopic (Bernardini et al., 2009). In our experiments we used these three measures, precisely described below. Other evaluation metrics are discussed in Section 2.4.

For each measure, we first give the definition assuming that the input is a flat list of results, and then show how it can be extended to a list of labeled clusters of results. Note that the evaluation of the latter type of search interface is more difficult, because we need to define a user's behavior model that takes into account its more complex features. The typical approach is to revert to a ranked list evaluation, by specifying a sequence of clusters examined by the user. However, most earlier studies assumed that the user is always able to choose the clusters with most relevant documents regardless of the cluster labels. In contrast, we assume that user searches are driven by the content of cluster labels. This seems more realistic and in closer agreement with the findings of some studies of search results clustering usage. Users seem to interact well with cluster labels, and a search session usually consists of displaying the result set of very few selected clusters; e.g., (Hearst and Pedersen, 1996), (Chen and Dumais, 2000), (Koshman et al., 2006).

In addition, we anticipate that the same output clusters may give rise to different sublists of examined items during the interaction between the user and the system, depending on the search task at hand. This approach is similar to the usage-specific notion of *stream* discussed by Azzopardi (2009), where the interactive information retrieval application is in our case a clustering engine. The exact order in which the information items displayed by the clustering system will be examined on the part of the user, for each evaluation measure, is detailed below.

2.1. Subtopic recall at rank n (S–Rec@n)

Consider a query Q with h subtopics $S_1, ..., S_h$ and a ranking $d_1, ..., dm$ of m documents. Let $subtopics(d_i)$ be the set of subtopics to which d_i is relevant. The subtopic recall at rank n, denoted $S–Rec@n$, is defined as the fraction of subtopics covered by the first n documents:

$$S–Rec@n = \frac{\cup_{i=1}^{n} subtopics(d_i)}{h} \tag{1}$$

2.1.1. S–Rec@n for clustered results

The above definition of $S–Rec@n$ applies to ranked lists. For clustered results, we need to define a user's behavior model suitable for this search task. As the goal is to maximize subtopic coverage, the user will take advantage of cluster information to avoid retrieving multiple documents relevant to the same subtopic. Our exact modelization of browsing through the clusters is the following. We assume that the user examines sequentially the cluster list top-down, opens only the clusters with a label relevant to subtopics that have not been already retrieved, and quits

an opened cluster as soon as a document relevant to the cluster's subtopic has been found. When the last cluster has been seen, the search is continued on the full list of documents. At the end of this process, all the documents will have been (re-)ranked accordingly, and we can compute the corresponding $S\text{–}Rec@n$ value using equation 1 (counting examined cluster labels as document snippets). For example, considering the query 'excalibur' in Figure 1, the user will visit all clusters except for 'excalibur film', subsumed by 'helen mirren nicholas clay', and 'camelot classic car', subsumed by 'excalibur cars'.

2.2. Subtopic precision at recall r (S–Prec@r)

Let n_r be the minimum rank at which $S\text{–}Rec@n = r$. The subtopic precision at recall r, denoted $S\text{–}Prec@r$, is defined as the ratio of the number of subtopics covered by the first n_r documents to the number of retrieved documents n_r:

$$S\text{–}Prec@r = \frac{\cup_{i=1}^{n_r} subtopics(d_i)}{n_r} \tag{2}$$

2.2.1. S–Prec@r for clustered results

For clustered results, we assume the same search behavior model as that applied to *S–Rec@n* in Section 2.1.1.

2.3. Subtopic search length under k document sufficiency (kSSL)

Let $p_{S_i,k}$ be the position of the k-th document relevant to subtopic S_i. The subtopic search length under k document sufficiency (*kSSL*) is defined as the mean of the positions $p_{S_i,k}$ of all h subtopics:

$$kSSL = \frac{\sum_{i=1}^{h} p_{S_i,k}}{h} \tag{3}$$

2.3.1. kSSL for clustered results

Turning to clustered results, as in this case the goal is to retrieve multiple documents relevant to the same subtopic, the user will want to inspect each relevant cluster fully. Thus, we assume that the user examines sequentially the cluster list, opening the first cluster with a relevant label to the subtopic, reading sequentially its elements, and then moving to the next cluster with a relevant label in the list, until k relevant search results have been retrieved. The *kSSL* value is the mean, averaged over the set of subtopics, of the number of items (cluster labels or document snippets) that must be examined for each subtopic according to this user behavior model. For example, considering again the query 'excalibur' in Figure 1, for the casino subtopic the user will visit only the cluster 'casino in las vegas', for the film subtopic the two visited clusters will be 'helen mirren nicholas clay' and 'excalibur film', etc.

2.4. Summary and related measures

In Table 1 we summarize the main features of the evaluation measures used in the experiments. Note that these measures do not easily allow comparison across queries having a different number of subtopics. Dividing by the number of subtopics is a simple normalization strategy. A more principled manner consists of normalizing over the best theoretical case (Zhai et al., 2003), but it is computationally very expensive.

Measure	*Browsing model for list of results*	*Browsing model for clustered results*	*Formula*
Subtopic recall at rank n	User reads result list	User reads results in clusters about distinct subtopics	$S\text{–}Rec@n = \frac{\cup_{i=1}^{n} subtopics(d_i)}{h}$
Subtopic precision at recall r	User reads result list	User reads results in clusters about distinct subtopics	$S\text{–}Prec@r = \frac{\cup_{i=1}^{n_r} subtopics(d_i)}{n_r}$
Subtopic search length under k document sufficiency	User reads result list	User reads results in clusters about the same subtopic	$kSSL = \frac{\sum_{i=1}^{h} ps_{i,k}}{h}$

Table 1: Main features of the evaluation measures used in the experiments.

We would also like to emphasize that these measures are not exhaustive, because with ambiguous or broad queries it is more difficult to recognize the underlying search intent and it may be necessary to accommodate multiple user needs. Other metrics have been recently proposed. The α−nDCG measure assumes graded relevance values are available and balances relevance and diversity (Clarke et al., 2008). When the parameter $\alpha = 0$, this measure is equivalent to the nDCG (normalized discounted cumulative gain) measure (Jarvelin and Kekalainen, 2002). *Intent aware measures* (Agrawal et al., 2009) generalize classical information retrieval metrics such as average precision or reciprocal rank by applying them to the single subtopics of a given query and then aggregating the results weighted by the subtopic probabilities. Intent (or subtopic) aware measures model a further evaluation dimension, i.e., when the user wants to retrieve single documents that cover multiple aspects in as much detail as possible.

These measures have become rapidly popular. Most notably, α−nDCG (with $\alpha- = 0.5$) and intent-aware precision at retrieval depth d (assuming equal probabilities) have been used in the diversity task of the TREC 2009 Web track (Clarke et al., 2009). However, they cannot be easily adapted to the clustered results scenario. In particular, it is not clear how to choose a suitable user's browsing model for these search tasks.

3. Test collections

For several years, the TREC 6-8 Interactive Track data (assembled in 2001-2003) has been the only test collection with document-level relevance judgments per subtopic. It contains 20 queries, with a focus on the *instances* of a given concept; e.g., 'what tropical storms – hurricanes and typhoons – have caused property damage and/or loss of life'.

Very recently, the gap between the growth in the availability of Web subtopic retrieval systems and the lack of appropriate tools for their evaluation has started to be addressed. Two early examples are the question-answering collection adapted to subtopic information retrieval by Clarke et al. (2008), and the ImageCLEFPhoto 2008 collection developed by Arni et al. (2008) In both of these collections, however, the single query subtopics have been retrofitted to queries that did not intend to emphasize diversity. Subsequently, two larger test collections have been made available that address the subtopic retrieval task in a more principled manner, namely the ClueWeb09 data set used in the diversity task of the TREC Web Track (Clarke et al., 2009), and the ImageCLEFPhoto 2009 collection (Paramita et al., 2009). They contain fifty queries each, categorized as ambiguous or faceted, with the queries and their subtopics being based on infor-

mation extracted from the usage logs of a commercial search engine. In particular, the subtopics were approximated as query reformulations estimated by means of clustering techniques.

In all these collections, however, the problem of re-ordering (or clustering) the search results is not separated from their retrieval, which makes it difficult to compare systems that only do post-processing of Web search results. In other words, the overall performance depends on the effectiveness of the diversification (or clustering) step as well as on the quality of the initial document ranking, whereas we may want to evaluate the specific behavior of post-processing techniques. We argue that to ensure a more controlled comparison, the same set of search results should be provided as an input to all systems being evaluated.

To tackle these issues Carpineto et al. (2009b) and Carpineto and Romano (2010) introduced two new test collections, termed AMBIENT and ODP-239. The idea is to use several distinct very small collections of mostly relevant documents, one for each query, as opposed to having one large collection containing a small set of relevant documents to each query, with most documents being nonrelevant to any query. Thus, our approach closely mimics the scenario of post-retrieval processing of Web search results.

AMBIENT and ODP-239 contain queries with multiple interpretations and multiple subtopics, respectively, derived from Web knowledge sources (i.e., Wikipedia and the ODP directory). This is another difference to existing collections. The queries of AMBIENT and ODP-239 are artificial, because they are not based on real user queries. In contrast, ClueWeb09 and ImageCLEF-Photo 2009 have a much more realistic set of queries. On the other hand, the subtopics defined in the latter collections may reflect the biases of a certain population of users and/or those of the automatic procedure used to analyze the log and guide the subtopic development, whereas the subtopics in Ambient and ODP-239 may have a more complete and objective definition, as stated in world-class knowledge repositories. These two approaches are complementary.

We now describe each collection in turn as they will be used in our experiments.

3.1. AMBIENT

AMBIENT (which stands for AMBIguous ENTries) consists of 44 queries extracted from *ambiguous* Wikipedia entries, i.e., those with 'disambiguation' in the title.[4] Each query has a set of subtopics (meanings) and a list of 100 ranked search results collected from Yahoo! as of mid-2007 and manually annotated with document-level relevance judgments per subtopic.

The average number of subtopics for each AMBIENT query is 17.95, according to Wikipedia definitions. Roughly half of them were present in the search results (i.e., 7.93 subtopics per query, on average), with an average number of relevant results per *retrieved* subtopic equal to 6.467. It is worthy of note that the search results usually contained several other subtopics that were not contained in the Wikipedia list. In fact AMBIENT measures the ability to retrieve *some* subtopics contained in the search results (i.e., those retrieved by Yahoo! *and* listed in the Wikipedia entry corresponding to the query), not all possible subtopics of a query.

For example, query 10 of AMBIENT is 'Excalibur', which has 26 distinct meanings according to Wikipedia (the complete list is provided in http://credo.fub.it/ambient/). Of these, the following eight were retrieved in the first 100 documents returned by Yahoo!.

[4] See http://en.wikipedia.org/wiki/Wikipedia:Links_to_%28disambiguation%29_pages.

10.1 Excalibur is the mythical sword of King Arthur.
10.2 Excalibur (film), a 1981 film about the legend of King Arthur.
10.4 Excalibur (automobile), a type of"contemporary classic" retro-styled car.
10.5 Excalibur (comics), both a Marvel Comics series set in the United Kingdom and a series featuring Professor X and Magneto in Genosha.
10.10 Excalibur (newspaper), the student newspaper of York University.
10.11 Excalibur (novel), a fantasy novel by Sanders Anne Laubenthal.
10.12 Excalibur (nightclub), a large and well-known nightclub in Chicago, founded in 1989.
10.20 XM982 Excalibur, 155mm extended range artillery projectile being developed by Raytheon and Bofors.
10.21 Excalibur Hotel and Casino, a medieval-themed hotel-casino in Las Vegas, Nevada

The other retrieved documents were often about products and services not covered by the traditional Wikipedia meanings, such as those in Figure 1 that refer to cars, mobile devices, software, and printers, but they may nonetheless be relevant to many users. More details on AMBIENT are provided in (Carpineto et al., 2009a). The collection can be downloaded from http://credo.fub.it/ambient/.[5]

3.2. ODP-239

ODP-239 combines the features of search results data with those of classification benchmarks. It consists of 239 queries, each with about 10 subtopics and 100 documents. The queries, subtopics, and their associated documents were selected from the Open Directory Project (http://www.dmoz.org) as of June 2009, in such a way that the distribution of documents across subtopics reflects the relative importance of subtopics. By ODP document we mean the items contained in the leaf nodes of the directory. Each documents consists of a title (which is the anchor text to the final Web page pointed to by ODP) and a very short description.

The exact procedure to select queries, subtopics, and documents from ODP was the following. For each of the 14 top categories in ODP (we disregarded the 'Regional' and 'World' categories because they do not have thematic subcategories), we chose all their subcategories with at least one hundred documents and six sub-subcategories. Such subcategories formed the set of ODP-239 *queries*. For each query, we chose up to 10 sub-subcategories as ODP-239 *subtopics*, considering the most numerous ones. Then we selected the documents to be assigned to each subtopic. We decided to take into account the relative size of ODP sub-subcategories, just as more popular subtopics get more top results in real web searches. For each subcategory, we randomly selected a set of 100 documents (among those labeled with that subcategory) imposing that the distribution of documents per sub-subcategory in such a set was the same as the distribution in the whole ODP collection. We also set the minimum number of documents per sub-subcategory equal to 4. Due to approximations, the final number of documents per query in ODP-239 was in some cases slightly larger or smaller than 100.

To illustrate, one top entry of Wikipedia is 'Sports', which has about one hundred children categories: 'Adventure_Racing', 'Airsoft', 'Animal Sports', 'Archery', etc. Of these, 33 were selected as queries in ODP-239: 'Adventure_Racing', 'Baseball', 'Basketball', 'Bowling', etc, each with a number of subtopics. For instance, query 221 is 'Sports > Cycling', which has the following 10 subtopics (with the number of associated documents shown in parentheses).

[5] AMBIENT mostly contains single-word queries. A larger version of AMBIENT with multi-word ambiguous queries has been recently made available by Navigli and Crisafulli (2010).

221.1 Sports > Cycling > Organizations (22)
221.2 Sports > Cycling > Travel (21)
221.3 Sports > Cycling > Bike_Shops (18)
221.4 Sports > Cycling > Regional (10)
221.5 Sports > Cycling > Mountain_Biking (10)
221.6 Sports > Cycling > Racing (7)
221.7 Sports > Cycling > BMX (5)
221.8 Sports > Cycling > Human_Powered_Vehicles (4)
221.9 Sports > Cycling > College_and_University (4)
221.10 Sports > Cycling > Personal_Pages (4)

Unlike AMBIENT, all documents are relevant to at least one subtopic and the document-subtopic assignment comes for free. ODP-239 and AMBIENT have complementary aspects: the former collection deals with ambiguous queries and is suitable for information retrieval, the latter is about truly multi-topic queries and is aimed at classification. ODP-239 was first introduced and very briefly described in (Carpineto and Romano, 2010), where it was used to conduct a ground-truth validation of search results clustering systems. The collection is available for download at http://credo.fub.it/odp239.

4. Main approaches to subtopic retrieval and tested methods

We consider three main types of approach to subtopic retrieval that can be applied to post-process search results: clustering of full list of results, diversification of top results, and minimal document set retrieval. For each approach we use three methods. The overall nine methods, although not exhaustive, are representative of a large spectrum of subtopic retrieval techniques. In the following we describe each approach and its relative methods in turn.

4.1. Clustering of search results

In the shift from data-centric to description-centric algorithms, a variety of clustering paradigms have been used, including concept lattices (Carpineto and Romano, 2004), spectral clustering (Cheng et al., 2005), and graph theory (Di Giacomo et al., 2007). The best known approach is probably to use a generalized suffix tree (GST) (Ukkonen, 1995; Andersson et al., 1999), which can be used to extract all the phrases contained in the search results in time linear with the size of the input.

Following Suffix Tree Clustering (STC) (Zamir and Etzioni, 1998), which was the first system to employ GST but was limited by its chaining mechanism for aggregating GST nodes, several other methods have been proposed that select the most informative GST nodes and then build the clusters around them. In this paper we consider three representatives of this clustering paradigm, namely, *Lingo*, *Lingo*3*G*, and *KeySRC*. They are characterized by highly descriptive phrases as cluster labels, and are known to perform well on browsing retrieval tasks (Carpineto et al., 2009b). We did not need to re-implement these algorithms because the results produced by the original versions on our test collections were made available to us by their authors. The three systems are briefly described below.

4.1.1. Lingo

Lingo (Osiński and Weiss, 2005) is a well known successor of STC. Frequent phrases are extracted using suffix arrays (Manber and Myers, 1993) instead of suffix trees, then the frequent phrases that best match certain *latent topics* present in the search results, determined via singular

value decomposition (SVD), are selected (Deerwester et al., 1990), and finally documents are allocated to such frequent phrases. *Lingo* can be tested at http://search.carrot2.org/stable/search.

4.1.2. Lingo3G

Despite similar names, *Lingo* and *Lingo*3*G* are two very different clustering algorithms. While *Lingo* uses SVD as the primary mechanism for cluster label induction, *Lingo*3*G* employs a custom-built metaheuristic algorithm that aims to select well-formed and diverse cluster labels. *Lingo*3*G* is a commercial system developed by Carrot Search (see http://company.carrot-search.com/lingo-applications.html for more information).

4.1.3. KeySRC

One of the most recent examples of description-centric clustering algorithm is $KeySRC$[6] (Keyphrase based Search Results Clustering), described by Bernardini et al. (2009). 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. Hereafter, this clustering method will be referred to as $KeySRC_C$.

4.2. Diversification of top search results

The general problem of minimizing the redundancy of a ranking of documents (or, dually, of maximizing their coverage with respect to different aspects of the query) is NP-hard (Agrawal et al., 2009). Most proposed methods are based on a greedy approximation, termed the *maximal marginal relevance* (Carbonell and Goldstein, 1998), in which documents are re-ranked one at a time choosing at each iteration the document with the best combined score of relevance and diversity.

A general formulation is the following. For each document d in a set of documents D, define a combined relevance-diversity rank for d with respect to a set of documents D^* as:

$$RelDiv(d \mid D, D^*) = \frac{(\beta^2 + 1)\ Rel(d|D)\ Div(d|D^*)}{\beta^2 Rel(d|D) + Div(d|D^*)} \tag{4}$$

where $Rel(d|D)$ and $Div(d|D^*)$ are the relevance rank and the diversity rank for d, respectively, and the parameter β is a weighting factor for the importance of the relevance (or diversity). This function is a generalization of the harmonic mean (i.e., $\frac{2}{\frac{1}{Rel(d|D)} + \frac{1}{Div(d|D^*)}} = \frac{2\ Rel(d|D)\ Div(d|D^*)}{Rel(d|D)+Div(d|D^*)}$, attained with $\beta = 1$), commonly used for the average of rates.[7]

In this paper we use as relevance ranks those provided by the search engine, because they take into account several information sources that are not normally available to research ranking algorithms. Consider also that classical ranking algorithms such as the $tf - idf$ scheme are not directly applicable to post-ranking of search results because the query words are usually present in any retrieved document. The diversity component in Equation 4 can be computed in various ways, e.g., using the similarity of documents (Carbonell and Goldstein, 1998), or their KL divergence (Zhai et al., 2003), or their correlation (Wang and Zhu, 2009).

[6] http://keysrc.fub.it

[7] The same function is often used in information retrieval to combine precision and recall into a single performance measure, termed F_β (van Rijsbergen, 1979).

A second main approach to search results diversification is to explicitly model the query aspects – rather than assuming that similar documents will cover similar aspects – and then avoid choosing documents that can be assigned to the same query aspect. The proposed techniques for modeling query aspects leverage various sources including diverse query reformulations obtained from a query log (Radlinski and Dumais, 2006), the information content of terms relevant to the query (Swaminathan et al., 2008), a predefined query taxonomy (Agrawal et al., 2009), sub-topical clusters (Deselaers et al., 2009), and sub-queries derived from the original query (Santos et al., 2010). In our experiments, we considered three methods spanning both of these approaches. The methods, detailed below, do not require external knowledge for training such as query logs or a classification taxonomy. In addition, they can be easily implemented and are computationally efficient.

4.2.1. Novelty based method

This is one of the earliest approaches to search results re-ranking (Carbonell and Goldstein, 1998). Our implementation of the novelty based method consisted of the following steps.

1. Choose the top ranked document in the original list as the first element of the re-ranked list (*RRL*), and remove it from the original list.

2. For each document d in the current original list *COL*, find $RelDiv(d|COL, RRL)$ (see Equation 4), where the diversity rank $Div(d|RRL)$ is computed by finding the similarity of each document with respect to *RRL* through the following expression:

$$Sim(d|RRL) = argmax_{d_i \in RRL}\ cosine(d, d_i) \tag{5}$$

and then by ranking the documents in *COL* as an inverse function of their similarity to *RRL*. In Equation 5, the *cosine* function is computed representing documents as $tf-idf$ weighted term vectors (up to text normalization).

3. Choose the document with the highest $RelDiv(d)$ as the next element in *RRL* and remove it from the current original list.

4. Repeat steps 2-3 for n times (n is a user-provided parameter).

4.2.2. Coverage based method

This method attempts to maximize the coverage of the total knowledge that exists on the Web about a given query (Swaminathan et al., 2008). Our implementation included the same steps as the novelty based method, except that the value of $Div(d)$ (step 2) is determined as follows.

The diversity rank $Div(d|RRL)$ is computed by finding the *joint coverage* score of the combined set of d and *RRL*, i.e., $C(d \cup \{RRL\})$, and then by ranking the documents in *COL* according to this score.

The joint coverage score is given by:

$$C(d \cup \{RRL\}) = \sum_{t \in T(d \cup \{RRL\})} r(t)\ log_2 \frac{1}{r(t)} \tag{6}$$

where $T(d \cup \{RRL\})$ is the set of terms contained in $d \cup \{RRL\}$, $r(t)$ is the *term-relevance* score, defined as the ratio of the number of documents containing the query *and* the term t to the number of documents containing the query, and the argument of the sum function is the entropy of the term relevance score. Notice that only words appearing together with query terms have non-zero relevance and contribute to the coverage score of the documents that contain them. Although this method has been devised for working with full documents, it is suggested that to

reduce noise only the paragraphs containing the query words should be considered. Thus, using search results snippets can be seen as an approximation of this strategy. Indeed, the authors of this method, termed Essential Pages (EP), recommend that it is used as a post-processing tool of the results returned by a search engine.

4.2.3. Cluster based method

As we hypothesize that documents grouped in a same cluster are relevant to the same subtopic and nonrelevant to the other subtopics, a possible approach to increasing diversity in early ranking is to promote documents belonging to different clusters; e.g., (Dai and Srihari, 2005), (Deselaers et al., 2009). We selected a representative document from each cluster generated by $KeySRC$ and added it to RRL. The list of clusters was visited top-down and the representative was the most relevant document in the cluster according to the ranking function used by the search engine. This diversification strategy is termed $KeySRC_L$.

An alternative means of selecting the best representative in each cluster would be to use the medoid, namely the document closest to the centroid of the cluster. The distance between two documents might be computed using a (dis)similarity function based on the set of terms describing the documents. In our case, however, this is probably not very appropriate because the documents have been clustered using transformed features (i.e., keyphrases) that do not always occur in the same form in the original representation of documents.

4.3. Retrieval of minimal document sets

Minimal document sets retrieval (Dai and Srihari, 2005) is an extension of diversification of top results , i.e., finding the minimal document sets with maximum subtopic coverage. The idea is to create several subsets of diverse documents in such a way that each subset covers all subtopics and does not have redundant information within it. Each set can be seen as a big, composite document, or multimedia object (Lee and Park, 2009), which is distinguishable by its aggregated subtopics.

In our experiments we used three methods to generate and rank relevant document sets. These methods are a simple refinement of those used for diversification of top results. For the novelty based and coverage based methods, we recursively applied the corresponding basic algorithms used for diversification to the documents in the original ranking not chosen by earlier invocations of the algorithms. The diverse document subsets were ranked in the same order as they were generated. The cluster based diversification method was iterated, choosing documents from clusters in a round robin fashion until all clusters became empty, at which point we appended the unclustered documents to the final re-ranked list in the same order as the original ranking.

4.4. Summary

In Table 2 we summarize the main features of the subtopic retrieval methods used in the experiments. The methods are grouped by retrieval strategy, namely clustering, diversification, and minimal set generation. Note that $KeySRC_L$, EP^*, and $KeySRC_L^*$ have not been used elsewhere.

5. Experimental setting

We set the number of clusters to ten, because in our experiments each query contained approximately ten subtopics. Likewise, we set the parameter n used in the diversification methods to ten. We did not try to optimize the method parameters. For the clustering methods we used

Strategy	*Method*	*Description*	*Reference*
Clustering	*Lingo*	Extraction of latent subtopics by SVD	Osiński and Weiss (2005)
Clustering	*Lingo3G*	Metaheuristic search	Proprietary system
Clustering	$KeySRC_C$	Extraction of keyphrases from GST	(Bernardini et al., 2009)
Diversification	*Novelty*	Similarity to seen documents	(Carbonell and Goldstein, 1998)
Diversification	*EP*	Coverage of query knowledge	(Swaminathan et al., 2008)
Diversification	$KeySRC_L$	Representatives of $KeySRC_C$ clusters	
Minimal sets	*Novelty**	Recursive *Novelty*	(Dai and Srihari, 2005)
Minimal sets	*EP**	Recursive *EP*	
Minimal sets	$KeySRC_L^*$	Round robin pick-up from $KeySRC_C$ clusters	

Table 2: Main features of the subtopic retrieval methods used in the experiments.

the default values of the on-line system versions, while for the diversification methods we set $\beta = 0.5$.

The evaluation of clustering performance using the measures described above raised a practical problem. In our modelization of browsing through clustered results, we assumed that user decisions are based on the semantic meaning of the cluster labels. In practice, it is required that each cluster label generated by each system for a given query be tagged as relevant or nonrelevant to any subtopic of that query. For the ODP-239 collection, for example, it takes about $239 \cdot 10 \cdot 10 \cdot 3 = 71,700$ (possibly redundant) relevance judgments, where the four factors are, respectively, the number of queries, subtopics per query, clusters per query, and clustering systems. To reduce the manual labor, rather than using the full set of ODP-239 topics, we randomly selected 44 of them.[8]

The three clustering systems tested in our experiments (i.e., *Lingo*, *Lingo*3*G*, and $KeySRC_C$) were run on the search results associated with each AMBIENT and ODP-239 query. Then, the relevance of each cluster label (produced by any system for any query) to any of the subtopics listed in AMBIENT and ODP-239 for each query was manually assessed by three external subjects using a web based evaluation tool developed for that purpose. The six methods for diversification, and minimal document sets retrieval were run on the initial list of results provided by AMBIENT and ODP-239, without any additional manual intervention. Note that since in ODP-239 all results are equally relevant to a query, the relevance rank $Rel(d|D)$ in Equation 4 was always set to 1.

6. Results for ambiguous queries

In the upper part of Table 3 we report the results of the ten methods on AMBIENT, averaged over the query set. In addition to the nine subtopic retrieval methods, we included as a reference of comparison the results obtained by the original lists of results returned by Yahoo!. The methods are split into four groups: baseline, clustering, diversification and minimal sets (top down). The methods used to generate minimal sets are denoted by an ending asterisk, and the best results are displayed in bold. We now compare the results achieved by each subtopic retrieval strategy.

[8] We tried to automatically assign labels to subtopics by expanding a subtopic with its WordNet synonyms and then requiring that a label have nonzero match to the the expanded subtopic, in order for it to be assigned to that subtopic. In this way, for instance, the cluster label 'film' would match the subtopic 'movie'. However, we found that many subtopics were not covered by any label, and that several assignments made by the automatic procedure were wrong due to disambiguation problems.

	Systems	*S-Rec @5*	*S-Rec @10*	*S-Rec @20*	*S-Prec @.25*	*S-Prec @.50*	*S-Prec @.75*	*S-Prec @1.00*	*kSSL* k=1	*kSSL* k=2	*kSSL* k=3	*kSSL* k=4
AMBIENT	Baseline											
	Yahoo!	0.35	0.49	0.64	0.54	0.34	0.22	0.13	14.1	31.5	40.7	48.1
	Clustering											
	Lingo	0.19	0.38	0.52	0.27	0.17	0.11	0.08	15.0	**24.1**	**31.1**	**36.4**
	Lingo3G	0.18	0.30	0.49	0.26	0.13	0.09	0.07	15.8	27.0	36.0	40.6
	KeySRC$_C$	0.20	0.40	0.55	0.30	0.19	0.10	0.07	14.4	24.3	31.7	36.8
	Diversification											
	EP	0.29	0.43	0.64	0.44	0.29	0.20	0.13	16.0	31.2	40.2	47.3
	Novelty	0.33	0.46	0.65	0.54	0.32	0.20	0.13	14.9	39.12	51.7	59.1
	KeySRC$_L$	**0.43**	**0.56**	**0.68**	**0.69**	**0.47**	**0.26**	**0.15**	**11.4**	31.0	40.8	48.0
	Minimal sets											
	*EP**	0.29	0.44	0.61	0.44	0.29	0.19	0.13	13.8	25.8	33.2	38.0
	*Novelty**	0.33	0.47	0.62	0.53	0.33	0.20	0.13	14.4	31.8	42.5	49.9
	KeySRC$^*_L$	0.43	0.56	0.65	0.68	0.44	0.23	0.15	13.1	26.5	35.7	46.8
ODP-239	Baseline											
	Random list	0.35	0.53	0.73	0.67	0.51	0.33	0.25	15.0	29.4	44.0	58.9
	Clustering											
	Lingo	0.14	0.31	0.52	0.31	0.21	0.13	0.10	22.0	35.0	48.3	63.8
	Lingo3G	0.13	0.28	0.56	0.27	0.20	0.13	0.11	21.5	34.4	48.2	63.3
	KeySRC$_C$	0.19	0.39	0.68	0.46	0.30	0.18	0.14	22.8	40.1	57.3	75.0
	Diversification											
	EP	0.36	0.55	0.75	**0.75**	**0.59**	**0.42**	**0.30**	14.5	30.0	43.6	58.7
	Novelty	0.40	0.62	0.78	0.75	0.57	0.39	0.26	13.3	28.5	43.7	58.8
	KeySRC$_L$	0.36	0.53	0.74	0.63	0.47	0.33	0.25	14.9	29.5	44.1	59.0
	Minimal sets											
	*EP**	0.36	0.55	0.73	0.67	0.51	0.33	0.25	16.2	29.3	43.0	57.0
	*Novelty**	**0.40**	**0.62**	**0.82**	0.75	0.57	0.39	0.26	**12.2**	**25.4**	**38.3**	**52.0**
	KeySRC$^*_L$	0.36	0.53	0.74	0.64	0.48	0.34	0.25	14.9	29.4	43.7	59.0

Table 3: Effectiveness of subtopic retrieval methods on AMBIENT and ODP-239 (restricted to 44 queries).

Clustering methods were in general very good for *kSSL*, with $k > 1$. They consistently improved over the baseline by a clear margin on all data points. The best method was Lingo, with a gain over the baseline of 23.5% for $k = 2$, 23.6% for $k = 3$, and 24.3% for $k = 4$. Note that diversification and minimal sets retrieval were, in general, not able to improve over the *kSSL* baseline. Using a two-tailed paired t test with a confidence level in excess of 95%, we found that the improvements of *Lingo* for *kSSL* ($k > 1$) were always statistically significant over all other methods, except for KeySRC. The variation in *p*-values was large, ranging from 2.6e-7 (over *Novelty* for $kSSL, k = 2$) to 0.028 (over EP^* for $kSSL, k = 3$).

On the other hand, clustering did not perform well on the other evaluation measures, including *kSSL* with $k = 1$. For these measures, where it is sufficient to retrieve only one relevant document per subtopic, the additional cost of browsing through the cluster list incurred by the user is likely to affect the overall performance. Even from a theoretical point of view, the optimal performance of clustering is lower than that of ranked lists. For instance, for $k = 1$, assuming that the query has h subtopics, $min\,kSSL_{cluster} = \frac{2+3+...+(h+1)}{h} = \frac{h+3}{2}$, (i.e., when each of the first n clusters refers to a distinct subtopic and contains a relevant document in the first position), whereas $min\,kSSL_{list} = \frac{1+2+...+h}{h} = \frac{h+1}{2}$ (i.e., when each of the first n search results satisfies a distinct subtopic).

Turning to diversification methods, they achieved dissimilar results. The retrieval effectiveness of $KeySRC_L$ was very good. It obtained the best values for all seven *S–Rec@n* and *S–Prec@r* measures, and also for *kSSL*, with $k = 1$ (i.e., for full subtopic coverage). The last finding is especially remarkable because $KeySRC_L$ was the only method able to improve over the baseline for *kSSL*, $k = 1$. The differences between $KeySRC_L$ and all other methods but $KeySRC_L^*$ were statistically significant across all seven measures, except for $S–Rec@20$. The *p*-values ranged from 9.6e-11 (over Lingo for $S–Prec@.50$) to 0.008 (over EP for $S–Prec@1.00$).

By contrast, EP and *Novelty* were far behind $KeySRC_L$ and they were even almost always worse than baseline. The disappointing performance of EP and *Novelty* can be partially explained by the fact that we did not try to optimize the value of β in Equation 4. On the other hand, improving over the original list of results provided by commercial search engines may not be an easy matter, because it is likely that the diversity issue is being addressed in the first results page.[9] All diversification methods tested in our experiments performed poorly on *kSSL*, $k > 1$. This was to be expected because these methods do not support full-subtopic retrieval.

The strategy based on minimal sets generation seems to combine the advantages of diversification and clustering. Although none of these methods obtained the best absolute value for any evaluation measure,[10] in general they achieved a better balance between the improvements on partial subtopic coverage and on full subtopic retrieval. The performance of $KeySRC_L^*$ was especially noteworthy, because this was the only method to achieve better results than baseline on all 11 evaluation measures. On the other hand, the gain for *kSSL*, $k > 2$, was clearly inferior to that of the best clustering method, and the gain for the other evaluation measures was slightly inferior to the corresponding $KeySRC_L$ diversification method. Interestingly, the best method of both the diversification and minimal sets generation strategies was based on clusters. This a very simple yet relatively novel approach, and there is probably much scope for improving the basic method used in our experiments. Our results suggest that further investigation of cluster based

[9] An interesting discussion of re-ranking and diversification policies of major commercial search engines including Yahoo!, with pointers to relevant patents, is contained in http://www.seobythesea.com/?p=3909.

[10] The results of diversification and minimal set retrieval for $S–Rec@10$ reported in Table 3 did not always coincide due to the presence of ties.

diversification is worthwhile.

7. Results for multi-topic queries

In the lower part of Table 3 we report the results of the ten methods on the ODP-239 test collection. Compared to the analogous results on ambiguous queries, the values show better performance of diversification and minimal sets, and worse performance of clustering.

The diversification and minimal sets methods were almost always better than baseline, but often by a small margin. Note that for this set of experiments the baseline results lists were not generated by a search engine. Rather, they were obtained by iteratively performing a random selection of the documents associated with each query in the ODP-239 collection. The most numerous subtopics were thus more likely to appear in the first results, which increased the scope for diversification policies. The fact that diversification did not clearly outperform random baseline was thus somewhat surprising. This was probably due, in part, to the intrinsic difficulty of the ODP-239 collection, because subtopics are not clearly distinct; see e.g., the subtopics of 'Sports > Cycling' listed in Section 3.2. The problem of discriminating among ODP-239 subtopics will be analyzed on a quantitative basis in the next section. Other reasons for the disappointing performance of diversification techniques are the lack of parameter optimization, already mentioned in Section 6, and the use of strict single-word indexing (while many ODP-239 subtopics are best discriminated by multi-word concepts)

Table 3 shows that the single diversification and minimal sets methods had comparable retrieval effectiveness, although the novelty based and the coverage based methods were in general better than the cluster based method. The best absolute results were obtained by *Novelty** for the *S−Rec@n* and *kSSL* measures, and by *EP* for the *S−Prec@r* measures. Statistical significance was again evaluated using a two-tailed paired t test with 95% confidence limits. The differences obtained by *Novelty** were always statistically significant, with *p*-values ranging from 3.8e-18 (over *Lingo* for *S−Rec*@5) to 0.013 (over *Novelty* for $kSSL, k = 1$), except over *Novelty* for *S−Rec@n* measures. The differences between *EP* and the other methods were always statistically significant, with *p*-values ranging from 5.9e-19 (over *Lingo*3*G* for *S−Prec*@.75) to 0.025 (over *EP** for *S−Prec*@.50), except over *Novelty* and *Novelty** for *S−Prec@r* measures.

The novelty based and coverage based methods probably benefited from the lack of nonrelevant results in the original document set. Another main observation concerning the results shown in the lower part of Table 3 is the low performance of clustering methods. This is discussed in the next section.

As already remarked, the results described in this section were obtained on a restricted version of ODP-239 test collection, formed by cutting down the number of queries from 239 to 44. In order to explore how the query reduction may have affected average performance results, we ran a further experiment. We evaluated the seven methods that did not require manual labeling (i.e., baseline, diversification, and minimal sets) on the full ODP-239 collection, and then compared the obtained performance values to those of Table 3. The results were surprisingly similar, across all measures and systems. For instance, the *kSSL* values for *EP* were 14.5 ($k = 1$), 29.8 ($k = 2$), 43.6 ($k = 3$), 58.9 ($k = 4$); i.e., nearly the same as those reported for the same method in Table 3. This can be seen as an indication that the results obtained for the clustering methods on the 44 topics would scale to the whole collection.

8. Estimation of subtopic dissimilarity across test collections

The $kSSL$ results obtained by all clustering methods on ODP-239 were much worse than the corresponding values on AMBIENT, both in absolute terms and relative to the baseline values. One explanation is that the classes of ODP-239 are more difficult to recover for a clustering method than those of AMBIENT. On the other hand, the comparatively better results achieved by the same clustering methods for *S–Rec@n* and *S–Prec@r*, together with the very good performance of $KeySRC_L$ for all evaluation measures, suggest that documents grouped into different clusters did belong to different subtopics (at least for the *KeySRC* clustering method).

It emerges that the disappointing *kSSL* results were mainly due to poor cluster labeling, because we observed that 54% of the subtopics were not covered by any cluster label. We hypothesized that ODP-239 subtopics were more difficult to label because their language models were more similar. To find experimental evidence for our hypothesis, we performed an estimation of subtopic dissimilarity across the two test collections using the symmetric Kullback-Leibler divergence, also known as Kullback-Leibler Distance (KLD). This measure, already used in information retrieval by Carpineto et al. (2001) and Bigi (2003), among others, was adapted to our problem in the following manner.

Let D_1, D_2, be the composite documents obtained by taking the union of the documents relevant to subtopics S_1, S_2, respectively. The KLD between S_1 and S_2 is given by:

$$KLD(S_1, S_2) = \sum_t \left\{ [(P(t|S_1) - P(t|S_2)] \; log \frac{P(t|S_1)}{P(t|S_2)} \right\} \tag{7}$$

The probabilities are estimated using a back-off smoothing, i.e.,

$$P(t|S_i) = \begin{cases} \psi \frac{c(t,D_i)}{|D_i|} & \text{if } t \in D_i \\ \xi & \text{if } t \notin D_i \end{cases} \tag{8}$$

where $c(t, D_i)$ is the number of times the term t occurs in D_i, $|D_i|$ the number of terms in D_i, ξ is a threshold probability set to the value of the smallest probability of a term in a subtopic, and $\psi = 1 - \sum_{t \notin D_i} \xi$.

The subtopic dissimilarity *S–Diss* for a given query with h subtopics is:

$$S\text{–}Diss = \frac{\sum_{i,j,i \neq j} KLD(S_i, S_j)}{\binom{h}{2}} \tag{9}$$

We computed the *S–Diss* value for each query, and then averaged the results over the set of queries in each collection. We found that the mean subtopic dissimilarity was equal to 9.27 for AMBIENT, whereas it was 6.15 for ODP-239. These results prove that the subtopic language models of ODP-239 were more similar, and thus it may be more difficult to find characteristic and discriminative cluster labels for the clustering algorithms. This is also a confirmation that the problem of cluster labeling may be more difficult than cluster optimization when the subtopic vocabularies overlap.

9. A query-by-query analysis

Table 3 shows that the mean performance of each method is more similar to that of the other methods in the same category. However, as both the clustering methods and the diversification

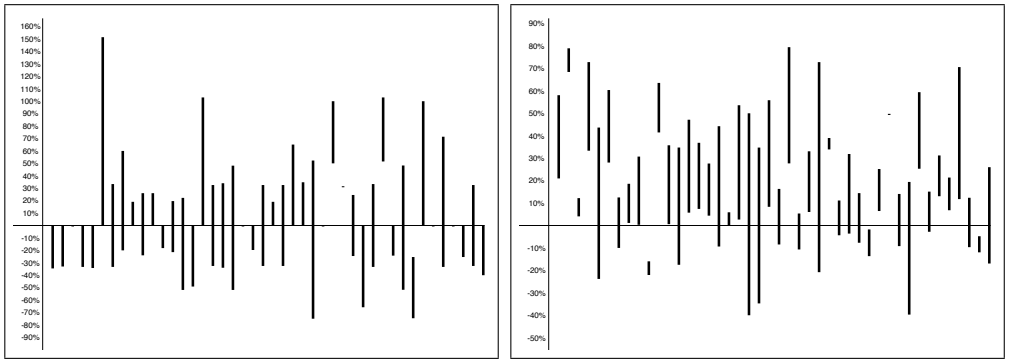

Figure 5: Performance variation of subtopic retrieval methods over the baseline on the 44 individual queries of AMBIENT. The left picture shows the S-Rec@10 variation for the diversification methods, the right picture shows the kSSL ($k = 3$) variation for the clustering methods. In both cases, the Yahoo! list is the baseline.

methods use different mathematical functions, it is conceivable that they would present considerable performance variation on individual queries. We tested this hypothesis through a query by query analysis. For each query and for each tested method, we computed the differences between the retrieval performance of the method and that of the baseline, and then computed the minimum and maximum of such differences.

In the left picture in Figure 5, we show the results for the three diversification methods on AMBIENT, using the list returned by Yahoo! as a baseline and S-Rec@10 as an evaluation measure. The length of each bar depicts the range of performance variations over the baseline (in percentage) attainable by the three methods on each query. In the right picture we plot the analogous results for the three clustering methods, using *kSSL* with $k = 3$. In both cases, the differences were ample, with a lot of scope for performance improvement.

These results suggest trying combination strategies. A first attempt to combine the results of multiple search results clustering algorithms was made by Carpineto and Romano (2010) with very good results. We are not aware of similar approaches for diversification methods, although this seems very promising. Another interesting research direction is the combination of clustering and diversification methods, with the goal of improving performance for those evaluation measures where either strategy, taken in isolation, usually performs poorly.

10. Conclusions

Recent research has focused on the definition of more effective subtopic retrieval strategies under two separate main headings, namely clustering and diversification of search results, but little work has been done to reconcile these different paradigms and identify their relative strengths and weaknesses. In this paper we have taken a step forward to help fill this gap, by providing a unifying evaluation perspective. We compared the effectiveness of three main strategies to subtopic retrieval from search results, using a range of complementary cross-paradigm measures and two suitable test collections with document-level relevance judgments per subtopic. The following major conclusions emerge from the experimental evaluation.

- Clustering is generally good for full-subtopic retrieval where other strategies usually fail, while it is less effective for partial subtopic coverage because of the additional cognitive

effort required on the part of the user to browse through the cluster labels before examining any document. Furthermore, clustering may not work well when the subtopic vocabularies blur, mainly due to the difficulty of producing discriminative cluster labels.

- Diversification is not useful for full subtopic retrieval, at least according to our modelization of this search task, but it can improve partial subtopic coverage. However, even in the latter case, the scope for improvement seems very limited when the original list of result is produced by a modern search engine.

- Retrieval of minimal subsets combines the advantages of diversification for partial subtopic coverage with a better performance on full-subtopic retrieval, although still inferior to that of clustering.

- Combination of multiple subtopic retrieval methods is promising because they present considerable variations on single queries and across different tasks.

In subtopic retrieval there is probably no one-size-fits-all method, because multiple and potentially conflicting user needs may arise; e.g., retrieving as many query interpretations as possible or as much information as possible about a specific interpretation. It would be interesting to see how well the proposed measures and our experimental findings correlate to the user experience; e.g., by means of user evaluation studies focusing on subtopic retrieval. More research is then needed to understand whether the most suitable method for a given query and user can be automatically selected (e.g., by classification of the underlying intent behind the search query). Also, the design of better search interfaces where several methods will coexist and effectively address complementary user needs may be a key to expanding the scope and usability of subtopic retrieval facilities in standard search systems.

11. Acknowledgments

We would like to thank Stanislaw Osiński and Dawid Weiss for running *Lingo* and *Lingo*3*G* on the AMBIENT and ODP-239 test collections, and providing us with the results. We would also like to thank two anonymous reviewers for their valuable comments and suggestions.

References

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S., 2009. Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009), Barcelona, Spain. ACM Press, pp. 5–14.

Altingövde, I. S., Demir, E., Can, F., Ulusoy, Ö., 2008. Incremental cluster-based retrieval using compressed cluster-skipping inverted files. ACM Transactions on Information Systems 26 (3).

Andersson, A., Larsson, N. J., Swanson, K., 1999. Suffix trees on words. Algorithmica 23, 102–115.

Arni, T., Tang, J., Sanderson, M., Clough, P., 2008. Creating a test collection to evaluate diversity in image retrieval. In: Proceedings of the 29th ACM SIGIR Workshop: Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments, Singapore. ACM Press.

Azzopardi, L., 2009. Usage based effectiveness measures: monitoring application performance in information retrieval. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China. ACM Press, pp. 631–640.

Bernardini, A., Carpineto, C., D'Amico, M., 2009. Full-Subtopic Retrieval with Keyphrase-Based Search Results Clustering. In: Proceedings of Web Intelligence 2009, Milan, Italy. IEEE Computer Society, pp. 206–213.

Bigi, B., 2003. Using kullback-leibler distance for text categorization. In: ECIR'03: Proceedings of the 25th European conference on IR research. Springer-Verlag, pp. 305–319.

Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia. ACM Press, pp. 335–336.

Carpineto, C., De Mori, R., Romano, G., Bigi, B., 2001. An information theoretic approach to automatic query expansion. ACM Transactions on Information Systems 19 (1), 1–27.

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

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

Carpineto, C., Romano, G., 2004. Concept Data Analysis — Theory and Applications. Wiley.

Carpineto, C., Romano, G., 2010. Optimal Meta Search Results Clustering. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland. ACM Press, pp. 170–177.

Chen, H., Dumais, S., 2000. Bringing order to the Web: Automatically Categorizing Search Results. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Hague, The Netherlands. ACM Press, pp. 145–152.

Cheng, D., Vempala, S., Kannan, R., Wang, G., 2005. A divide-and-merge methodology for clustering. In: Li, C. (Ed.), Proceedings of the 24th ACM Symposium on Principles of Database Systems, Baltimore, Maryland, USA. ACM Press, pp. 196–205.

Clarke, C. L. A., Craswell, N., Soboroff, I., 2009. Overview of the TREC 2009 Web Track. In: Proceedings of the 18th Text REtrieval Conference (TREC 2009), Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., 2008. Novelty and diversity in information retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore. ACM Press, pp. 659–666.

Dai, W., Srihari, R., 2005. Minimal document set retrieval. In: Proceedings of the of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005), Bremen, Germany. ACM Press, pp. 752–759.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, T. K., 1990. Indexing by latent semanic analysis. Journal of the American Society for Information Science 41 (6), 391–407.

Deselaers, T., Gass, T., Dreuw, P., Ney, H., 2009. Jointly optimising relevance and diversity in image retrieval. In: ACM International Conference on Image and Video Retrieval 2009 (CIVR 2009), Santorini, Greece. ACM Press, New York, NY, USA.

Di Giacomo, E., Didimo, W., Grilli, L., Liotta, G., 2007. Graph Visualization Techniques for Web Clustering Engines. IEEE Transactions on Visualization and Computer Graphics 13 (2), 294–304.

Ferragina, P., Gulli, A., 2005. A personalized search engine based on web-snippet hierarchical clustering. In: Proceedings of the 14th International Conference on World Wide Web, WWW'05, Chiba, Japan. ACM Press, pp. 801–810.

Gordon, M. D., Lenk, P., 1999. When is the probability ranking principle suboptimal. JASIS 43 (1), 1–14.

Hearst, M. A., Pedersen, J. O., 1996. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In: Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval, Zürich, Switzerland. ACM Press, pp. 76–84.

Jansen, B. J., Booth, D. L., Spink, A., 2008. Determining the informational, navigational, and transactional intent of Web queries. Information Processing and Management 44 (3), 1251–1266.

Jarvelin, K., Kekalainen, K., 2002. Cumulative gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20 (4), 422–446.

Koshman, S., Spink, A., Jansen, B. J., 2006. Web Searching on the Vivisimo Search Engine. Journal of the American Society for Information Science and Technology 57 (14), 1875–1887.

Lee, S., Park, J., 2009. A Scoring Function for Retrieving Photo Sets with Broad Topic Coverage. In: Proceedings of the 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, Korea. IEEE Computer Society, pp. 1577–1580.

Liu, X., Croft, B. W., 2004. Cluster-Based Retrieval Using Language Models. In: Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK. ACM Press, pp. 186–193.

Manber, U., Myers, G., 1993. Suffix Arrays: A New Method for On-line String Searches. SIAM Journal on Computing 22 (5), 935–948.

Navigli, R., Crisafulli, G., 2010. Inducing Word Senses to Improve Web Search Result Clustering. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), Cambridge, Massachusetts, USA. pub-acl, pp. 116–126.

Osiński, S., Weiss, D., 2005. A Concept-Driven Algorithm for Clustering Search Results. IEEE Intelligent Systems 20 (3), 48–54.

Paramita, M., Sanderson, M., Clough, P., 2009. Developing a Test Collection to Support Diversity Analysis. In: Proceedings of the 30th ACM SIGIR Workshop: Redundancy, Diversity, and Interdependent Document Relevance, Boston, Massachusetts, USA. ACM Press.

Radlinski, F., Dumais, S., 2006. Improving personalized web search using result diversification. In: Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil. ACM Press, pp. 691–692.

Robertson, S. E., 1977. The probability ranking principle in IR. Journal of Documentation 33, 294–304.

Santos, R. L. T., Peng, J., Macdonald, C., Ounis, I., 2010. Explicit Search Result Diversification through Sub-queries. In: Proceedings of the 32nd European Conference on IR Research (ECIR 2010), Milton Keynes, UK. Springer, pp. 87–99.

Swaminathan, A., Mathew, C., Kirovski, D., 2008. Essential Pages. Tech. Rep. MSR-TR-2008-15, Microsoft Research.

Ukkonen, E., 1995. On-Line Construction of Suffix Trees. Algorithmica 14 (3), 249–260.

van Rijsbergen, K., 1979. Information Retrieval. Butterworth-Heinemann.

Wang, J., Zhu, J., 2009. Portfolio theory of information retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA. ACM Press, pp. 115–122.

Xu, Y., Yin, H., 2008. Novelty and topicality in interactive information retrieval. Journal of the American Society for Information Science and Technology 59 (2), 201–215.

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

Zhai, C., Cohen, W. W., Lafferty, J., 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada. ACM Press, pp. 10–17.

Zuccon, G., Azzopardi, L., 2010. Using the quantum probability ranking principle to rank interdependent documents. In: Proceedings of the 32nd European Conference on IR Research (ECIR 2010), Milton Keynes, UK. pp. 357–369.