Beyond Precision@10: Clustering the Long Tail of Web Search Results

Benno Stein Tim Gollub Dennis Hoppe

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

The paper addresses the missing user acceptance of web search result clustering. We report on selected analyses and propose new concepts to improve existing result clustering approaches. Our findings in a nutshell are: (1) *Don't compete with a search engine's top hits.* In response to a query we presume search engines to return an optimal result list in the sense of the probabilistic ranking principle: documents that are expected by the majority of users are placed on top and form the result list head. We argue that, with respect to the top results, it is not beneficial to replace this established form of result presentation. (2) *Improve document access in the result list tail.* Documents that address the information need of "minorities" appear at some position in the result list tail. Especially for ambiguous and multi-faceted queries we expect this tail to be long, with many users appreciating different documents. In this situation web search result clustering can improve user satisfaction by reorganizing the long tail into topic-specific clusters. (3) *Avoid shadowing when constructing cluster labels.* We show that most of the cluster labels that are generated by current clustering technology occur within the snippets of the result list head—an effect which we call *shadowing*. The value of such labels for topic organization and navigating within a clustering of the entire result list is limited. We propose and analyze a filtering approach to significantly alleviate the label shadowing effect.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Clustering* H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Search process*
General Terms: Algorithms, Experimentation, Measurement
Keywords: search result clustering, cluster labeling

1. INTRODUCTION

Web search is the task of finding a document in the World Wide Web in order to satisfy a user's information need that is specified as a query. Deriving the "true" information need from a query is a challenge, and search engines often retrieve millions of documents from which only a fraction is relevant for the user. To reduce the negative impact of irrelevant documents, search engines apply a result presentation strategy, which can be characterized as either *relevance-based* or *diversity-based*.

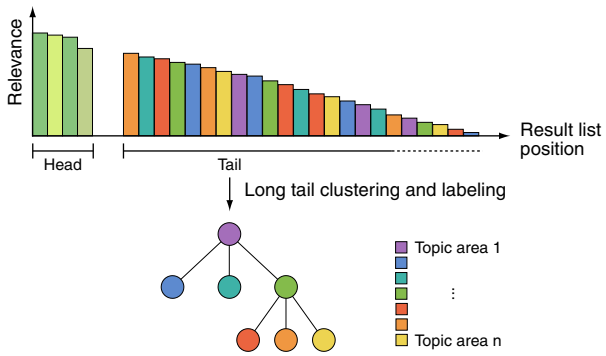

Figure 1: Result presentation by competence partitioning: combine the ranked result list head with a clustering of the result list tail.

The objective of the relevance-based strategy, which is the predominant strategy at this time, is to serve those information needs that are most likely associated with the query. The results are organized as a ranked list. The relevance-based strategy is extremely effective if a user can spot a desired document among the top ranks. If not, the user has to resort to a sequential search in the result list tail, which is ineffective since no topic-specific structuring is provided. Two recent user studies complement these observations impressively: A study from 2010 reviews the top five results of 1 000 queries sampled from the query log of a major search engine and reports that more than 90% of these queries are served excellent by all major search engines [28]. A study from 2008 reveals that only 8% of the users are willing to skim through more than three result pages, which corresponds to less than 30 search results [13].

On the other hand, the objective of the diversity-based strategy is to serve multiple information needs "in parallel". A well-known representative of this strategy is web search result clustering [4]. Search engines that implement this strategy group similar documents into clusters and try to construct descriptive cluster labels to guide users during their search. If the clustering is effective and the labeling is expressive, diversity-based search engines can serve a multitude of information needs associated with a query equally well. Users with uncommon information needs or multi-faceted informational queries especially benefit from the structured information access provided by the clustering.

In this paper we propose to exploit both result presentation strategies by combining the head of a ranked result list with a clustering of those documents found in the tail and not covered by the head. Figure 1 illustrates the paradigm. Notice that this *competence partitioning* strategy is fundamentally different from the current practice of web search result clustering engines: by resorting to the result

This work is supported in part by the German Science Foundation under grants STE1019/2-1 and FU205/22-1.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

list head, we exploit the "wisdom" of the search engine developers. By clustering a filtered result list tail, we prevent a user from sifting through many result pages.

The remainder of this paper is organized as follows. Section 2 gives a focused overview of the state of the art, while Section 3 contributes analytical insights and methodological elements of our approach. (1) A critical analysis of the role that result clustering can play for web search, (2) a quantification of the label shadowing effect, and (3) a filtering approach to exploit the idea of long tail clustering. Compared to existing result clustering approaches we report on less shadowing in the discovered clusters.

2. RELATED WORK

Web search result clustering is a heterogeneous research field, overlapping with and influenced by cluster analysis, cluster labeling, web snippet clustering, subtopic retrieval, search result classification, query type identification, query diversification, and faceted search. This mixture of related fields is not without reason: until this day web search result clustering is not properly solved. A fact which is rooted in the difficulty of the task itself but also in a misunderstanding of the very problem and, last but not least, in the ever changing user expectations and requirements:

- Most web search result clustering systems exploit only the text of the search result snippets, and clustering short texts is difficult [19].
- Is clustering the problem or labeling? Meanwhile, most researchers lean towards the latter [21].
- A user with a known item finding problem has different expectations than a user who wants to overview a new field. A user who experienced the monothetic characteristic of a faceted search application expects the same topical orthogonality from search result clustering [12].

The existing systems for web search result clustering can be distinguished along three dimensions [4]: data-centric, description-centric, and description-aware. *Data-centric systems* give top priority to clustering, while the formation of cluster labels has no effect on the partitioning of the snippets. The labels are usually derived from a cluster's mathematical representation, e.g., a centroid or a medoid under a bag-of-words model, and hence the generated labels form a sequence of probably unrelated words, often lacking understandability. Examples for such systems are WebCat [11] and AIsearch [21]. *Description-centric systems*, by contrast, employ cluster analysis solely for the purpose of discovering topics within a collection. Snippet partitioning is achieved by monothetic clustering, where each feature is used in an isolated manner to partition a collection into (overlapping) clusters. This approach leads to an improved understandability, which is bought with a possibly acceptable decrease in the clustering quality. Examples for such systems are Lingo [18], DisCover [15] and KeySRC [3]. *Description-aware systems*, finally, interweave the processes of clustering and labeling. In the existing systems of this type, such as Grouper [27] or SnakeT [9], labeling affects a polythetic cluster analysis. Although monothetic clustering is preferred, Carpineto et al. [4] note that none of the three system types is per definition superior to the others. The acceptance and usability of a system for web search result clustering depends on the quality of its components. We provide entry points and recent results from the respective fields for those who are interested in background technology:

Cluster Labeling. The salient properties of good cluster labels are comprehensibility, descriptiveness, and discriminative power [21]. Phrases are the basic building blocks of labels, whereas noun phrases [20], named entities [23], and title phrases [10] are recently discussed alternatives to improve comprehensibility. Descriptiveness means that a label should speak for each document in a cluster [25], while discriminative power means that the semantic overlap of a label between two clusters should be minimal.

Subtopic Retrieval. Subtopic retrieval is related to web search result clustering, since one is interested in grouping different aspects of an undirected, informational query. In this regard Bernardini et al. [3] employ keyphrase filtering, while Carpineto and Romano [6] combine the results from various result clustering systems.

Result Classification. At first sight, classifying snippets into a predefined taxonomy (e.g. from the Open Directory Project) sounds perfect in order to group search results into human-understandable categories [16]. However, web directories cannot deal with the dynamics of the web, and the limited number of available categories cannot suit arbitrary queries [17].

Query Diversification. Diversification is a rather new approach to deal with ambiguous queries by re-ranking relevance-based result lists [1, 22, 26]. The objective is to make the head of a result list both maximally relevant and diverse. Query diversification helps users to find relevant information if they are searching only for a specific source; if users are interested in retrieving a set of relevant snippets for a given query, result clustering is superior. In addition, query diversification is limited, because the result list head is short.

Faceted Search. Faceted search can be considered as a "controlled, monothetic cluster analysis". Without doubt, facets excel in satisfying the three label properties mentioned above. The automatic and ad-hoc generation of facets is an active research field, but currently this problem must be considered as unsolved [24]. Table 1 overviews and completes the outlined pros and cons of the result presentation strategies discussed.

Table 1: Comparison of ranked result lists (RRL), query diversification (QD), web search result clustering (WSRC), and faceted search (FS).

	RRL	QD	WSRC	FS
Complexity	low	high	high	very high
Navigational queries	+	+	–	–
Informational queries	–	+	++	++
Common information need	+	+	–	–
Uncommon information need	–	–	+	+
Exploratory search	–	–	+	++
Disambiguation	–	+	++	++
Optimization criterion	relevance	relevance, diversity	diversity	diversity

3. LONG TAIL CLUSTERING

In the following we discuss the competence partitioning strategy as illustrated in Figure 1 and introduce a tailored clustering approach, which we refer to as *LongTailClustering*.

3.1 Towards Better Search Result Clustering

Relevance-based search engines answer navigational as well as the popular non-navigational queries almost perfectly. For such kinds of information needs, web search result clustering and related technology cannot compete. Aside from the navigational overhead (selecting, focusing, and browsing a cluster) the current clustering approaches struggle with the labeling problem as well as with runtime issues [4]. One might argue that providing a comprehensive cluster containing all relevant results is always appreciated. But experience shows that often a fraction of the documents is sufficient to satisfy an information need. Once a relevant result is at the user's disposal, alternative and more robust techniques exist to retrieve additional relevant documents: (1) search engines allow for searching

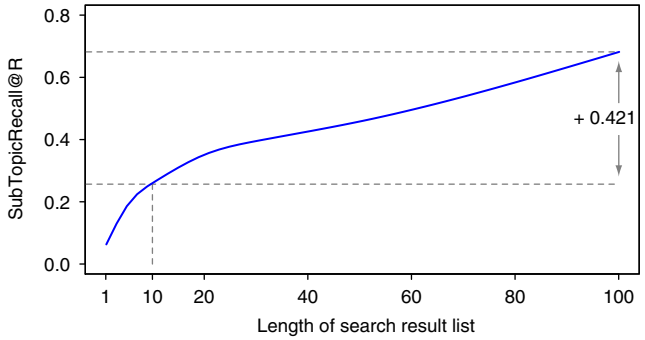

Figure 2: Average subtopic recall for the topics of the AMBIENT dataset. For each topic, we retrieve $N = 100$ search results from eTools (http://www.etools.ch).

similar documents given a particular result, (2) rephrasing the original query inspired by a relevant result became an accepted search strategy [14], and (3) the web itself as a hyperlinked document collection is a great resource for a guided search.

Given the impressive performance of relevance-based search engines, the outlined picture might raise the question whether web search result clustering is of any use. In Section 1, we identified the accessibility of the result list tail as the major weakness of ranked result lists. If we show that the result list tail is valuable and covers further query aspects, a perspective towards better search result clustering emerges.

In order to quantify the potential of the result list tail, we study two datasets: TREC Web Track 2009/2010 [7, 8] and AMBIENT [5]. Each dataset consists of a set T of search topics, with 100 and 44 elements, respectively. For each topic $t \in T$ a number of subtopics is provided. TREC features on average 4.6 subtopics, whereas AMBIENT provides 18 subtopics on average. Every subtopic focuses on a different aspect of t. We form the set S_t as the union of t's subtopics, where we only consider the noun phrases of each subtopic description. Topic terms as well as stopwords present in the noun phrases are discarded. E.g., the topic t="fahrenheit" of the AMBIENT dataset comprises subtopics such as "scale temperature", "michael moore movie", and "band". A standard search engine is queried with each topic $t \in T$ in order to obtain a ranked list $D_t = d_1, \ldots, d_N$ of the top N results, where each $d \in D_t$ denotes a result snippet. A subtopic of t is said to be *covered* by a snippet d, if at least two of its terms occur in d. In the special case of a single term subtopic, the single term has to appear in d. The set of all subtopics of t that are covered by a snippet d is denoted as coverage(d, S_t). We are interested in the subtopic recall for a result list D_t at rank R, which is defined as the fraction of subtopics covered by the first R results [29]:

$$\text{SubTopicRecall@R} = \frac{|\cup_{i=1}^{R} \text{coverage}(d_i, S_t)|}{|S_t|}, \qquad (1)$$

with $d_i \in D_t$. Figure 2 illustrates the average subtopic recall for all topics in T of the AMBIENT dataset. The TREC Web Track dataset reveals equivalent characteristics. The diagram shows that on average the first N=100 results for a topic cover 66.6% of all subtopics. If we define the result list head to comprise the first ten results, it covers 24.5% of all subtopics; additional 42.1% of the subtopics are covered by the result list tail, rendering the tail a valuable information source. In summary, if web search result clustering can be tailored to organize just the remaining subtopics, it will perfectly complement the result list head and enable a more efficient and effective result analysis. The key to such a complementing behavior lies in the alleviation of the shadowing effect.

3.2 The Shadowing Effect

To generate a clustering that complements the head of a result list, the topics of documents that are to be clustered must differ from the topics already covered by the result list head. We refer to the effect of undesired topic repetition as *shadowing*. To quantify the shadowing effect, we perform state-of-the-art result clustering with Lingo on three different kinds of result lists: (a) the unmodified result list D_t to resemble the standard scenario of existing clustering search engines such as Yippy and Carrot Search[1], (b) the tail of D_t, as a naive application of our idea, and (c) a subset of the tail of D_t, which is filtered with respect to the topics of the head. In the following we set the number of generated clusters per clustering (= per topic t) to ten; the respective cluster labels for a clustering are denoted as L_t. Cluster labels $l \in L_t$ are phrases such as "Fahrenheit 9/11" or "Wellington based Web Design Collective". Then, the shadowing at rank R can be expressed by replacing in Equation (1) the set of subtopics S_t with the set of cluster labels L_t:

$$\text{Shadowing@R} = \frac{|\cup_{i=1}^{R} \text{coverage}(d_i, L_t)|}{|L_t|},$$

with $d_i \in D_t$. Figure 3 illustrates the result of this analysis. The shadowing effect when clustering result list (a) is obvious: the Shadowing@10 is 0.417, i.e., about four cluster labels occur in the snippets of the head of the result lists. One might expect that clustering just the tail gives a reasonable smaller shadowing, which is not the case as can be seen in the results for result list (b). The reason for this behavior is that the topics from the head reappear throughout the entire result list. For our two datasets, up to 18% of the search result snippets in the tail cover the subtopics from the head. Finally, the filtered result list (c) reduces the shadowing effect by 31.6%, entailing a Shadowing@10 of only 0.1. Thus, nine out of ten cluster refer to subtopics present only in the tail.

3.3 Long Tail Filtering of Search Results

We now present a basic procedure to filter the result list tail. The aim of the filtering is to exclusively eliminate all results which cover topics from the result list head. We expect to satisfy the following two demands: (1) a reduction of the shadowing effect, and (2) a stable or increased subtopic recall after the filtering. Our procedure is detailed below.

LongTailClustering

Input: $query$
Parameters: $headSize, n$
Output: $tailClustering$

1. ANALYZEHEAD($query$)
$resultList$ = SEARCH($query$)
$headTopics$ = ANALYZE($resultList, headSize$)

2. CLUSTERTAIL($query, headTopics$)
$augmentedQuery = query \wedge \neg headTopics$
$filteredTail$ = SEARCH($augmentedQuery$)
$tailClustering$ = CLUSTER($filteredTail, n$)

return $tailClustering$

LongTailClustering, as presented, is a generic procedure, and the choices of the search engine `SEARCH`, the topic analyzer `ANALYZE`, and the clustering algorithm `CLUSTER` are user-specific. We experimented with a variety of configurations and see further research opportunities for this task. At present, the best overall results are achieved with the following setup: First, given a query, we receive from the meta-search engine eTools a result list of size $N = 200$.

[1] `http://www.yippy.com, http://www.carrotsearch.com`

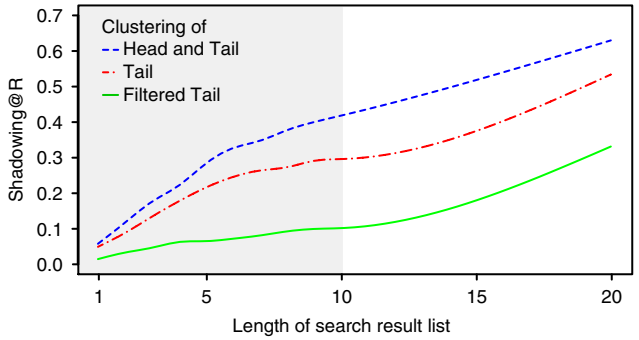

Figure 3: Average shadowing for the topics of the AMBIENT dataset. Three different result sets are considered for the cluster analysis.

In the context of our evaluation, this search engine turned out to provide a higher snippet quality than did Bing or Yahoo. The quality of the snippets is of high importance, since we are interested in finding topics in short texts. For this topic analysis, we extract all noun phrases [2] of maximum length four from the first ten results of the list. Each extracted noun phrase is treated as a topic that has to be filtered from the result list tail in the second step. For this purpose an augmented query is formed, which combines the original query with all noun phrases prefixed by the NOT operator. For instance, the augmented query for "fahrenheit" starts with "fahrenheit -"scale temperature" -"taiwanese boy band" -"video game"". An augmented query can either be used to filter the original result list directly or to send a new request to the search engine. In the latter case, the wisdom of the search engine developers is exploited again. For our experiments with the TREC Web Track and AMBIENT topics, better results are achieved with the former method. For the final clustering, we apply Lingo to the first 90 results of the filtered result list. As already noted and illustrated in Figure 3, the applied *LongTailClustering* procedure reduces the shadowing effect significantly.

In a final experiment, we investigate whether our filtering procedure retains the subtopics that do not appear in the head. For that, we substitute the original result list tail by its filtered version and measure the subtopic recall at rank 100. We find, that although 53.8% of the search results are eliminated in the course of *LongTailClustering*, the filtered result list still reaches a subtopic recall at 100 of 0.64, i.e., contains 96.4% of the subtopics from the original result list (cf. Figure 2). Hence, our procedure successfully implements our idea of moving the focus of web search result clustering to the subtopics that appear exclusively in the result list tail.

4. CONCLUSIONS

A main achievement of our research is that we release web search result clustering from being considered direct competitor to ranked result lists. Instead, we see cluster analysis as a complementary presentation tool to improve the accessibility of relevant documents in the long tail of result lists. This way, web search result clustering goes in line with other advanced search tools such as related search, query suggestions, or the Google Wonderwheel. We have provided the measures SubTopicRecall@R and Shadowing@R to quantify our considerations, and we give empirical evidence for the claimed effects. With *LongTailClustering* we present a procedure that is tailored to the characteristics of our view: *LongTailClustering* is able to cope with the shadowing effect and to increase the density of subtopics that are exclusive for the result list tail. Our current research further analyzes the potential of long tail result clustering and the adaptation of existing clustering technology for this purpose.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proceedings of WSDM 2009*, pages 5–14.

[2] K. Barker and N. Cornacchia. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of AI 2000*, pages 40–52.

[3] A. Bernardini, C. Carpineto, and M. D'Amico. Full-Subtopic Retrieval with Keyphrase-Based Search Results Clustering. In *Proceedings of WI-IAT 2009*, pages 206–213.

[4] C. Carpineto, S. Osiński, G. Romano, D. Weiss. A Survey of Web Clustering Engines. *ACM Comp. Surveys*, 41(3):Article 17, 2009.

[5] C. Carpineto and G. Romano. AMBIENT dataset. `http://credo.fub.it/ambient`, 2008.

[6] C. Carpineto and G. Romano. Optimal Meta Search Results Clustering. In *Proceedings of SIGIR 2010*, pages 170–177.

[7] C.L.A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. `http://plg.uwaterloo.ca/~trecweb/2009.html`, 2009.

[8] C.L.A. Clarke, N. Craswell, I. Soboroff, and G.V. Cormack. Overview of the TREC 2010 Web Track. `http://plg.uwaterloo.ca/~trecweb/2010.html`, 2010.

[9] P. Ferragina and A. Gullì. A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. In *Proceedings of WWW 2005*, pages 801–810.

[10] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani. Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution. In *Proceedings of SPIRE 2006*, pages 25–36.

[11] F. Giannotti, M. Nanni, D. Predreschi, and F. Samaritani. WebCat: Automatic Categorization of Web Search Results. In *Proceedings of SEBD 2003*, pages 507–518.

[12] M.A. Hearst. Clustering versus Faceted Categories for Information Exploration. *Commun. ACM*, 49(4): pages 59–61, 2006.

[13] iProspect.com, Inc. iProspect Blended Search Results Study. `http://www.iprospect.com`, 2008.

[14] R. Jones and K.L. Klinkner. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of CIKM 2008*, pages 699–708.

[15] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, R. Krishnapuram. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proceedings of WWW 2004*, pages 658–665.

[16] Z.-Y. Ming, K. Wang, and T.-S. Chua. Prototype Hierarchy Based Clustering for the Categorization and Navigation of Web Collections. In *Proceedings of SIGIR 2010*, pages 2–9.

[17] R. Navigli and G. Crisafulli. Inducing Word Senses to improve Web Search Result Clustering. In *Proc. of EMNLP 2010*, pages 116–126.

[18] S. Osiński, J. Stefanowski, and D. Weiss. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Proceedings of IIPWM 2004*, pages 359–368.

[19] D. Pinto, J.-M. Benedí, and P. Rosso. Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. In *Proceedings of CICling 2007*, pages 611–622.

[20] J. Stefanowski and D. Weiss. Comprehensible and Accurate Cluster Labels in Text Clustering. In *Proceedings of RIAO 2007*.

[21] B. Stein and S. Meyer zu Eißen. Topic Identification: Framework and Application. In *Proceedings of i-KNOW 2004*, pages 353–360.

[22] A. Swaminathan, C.V. Mathew, and D. Kirovski. Essential Pages. In *Proceedings of WI-IAT 2009*, pages 173–182.

[23] H. Toda and R. Kataoka. A Clustering Method for News Articles Retrieval System. In *Proceedings of WWW 2005*, pages 988–989.

[24] D. Tunkelang. Faceted Search. Morgan & Claypool Publishers, 2009.

[25] D. Weiss. Descriptive Clustering as a Method for Exploring Text Collections. Ph.D. diss., Poznań Univ. of Technology, Poland, 2006.

[26] M.J. Welch, J. Cho, and C. Olston. Search Result Diversity for Informational Queries. In *Proceedings of WWW 2011*, pages 237–246.

[27] O. Zamir and O. Etzioni. Grouper: A dynamic Clustering Interface to Web Search Results. In *Proceedings of WWW 1999*, pages 1361–1374.

[28] H. Zaragoza, B. B. Cambazoglu, and R. Baeza-Yates. Web Search Solved? All Result Rankings the Same?. In *Proceedings of CIKM 2010*, pages 529–538.

[29] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proceedings of SIGIR 2003*, pages 10–17.