Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 928-931

Clustering the Information Space Using Top-Ranking Sentences: A Study of User Interaction

Anastasios Tombros[1], Joemon M. Jose[1], Ian Ruthven[2] & Ryen W. White[1]

[1]Department of Computing Science, University of Glasgow, Scotland

[2]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland

Abstract: By considering sentences selected by a query-biased sentence extraction model from the top-retrieved documents, we create a personalised information space which is characterised by the presence of search terms. We cluster this information space, and enable searchers to interact with the resulting clusters. In order to examine whether users can recognise, and benefit from, the clustered organisation, we compare user interaction and performance between an actual clustering and a pseudo-clustering of the information space for completing information seeking tasks. The results provide evidence for the utility and meaningfulness of the clustered organisation.

Keywords: information retrieval, WWW, top-ranking sentences, clustering, user interaction

1 Introduction and Motivation

One of the main challenging issues in information retrieval (IR) is the facilitation of efficient and effective access to the large amount of available information. This becomes especially important when users have a vague, poorly defined information need. In such cases searchers may find it particularly difficult to understand the information content of the retrieved document set (referred to as the *information space* in this paper).

In order to facilitate more effective information access, we have previously proposed the clustering of sentences which form part of single document summaries of the top documents retrieved in response to a query (*top-ranking sentences*, *TRS*) (Tombros et al, 2003). In this paper, we use the resulting clusters to present search results to users. Retrieved documents can be accessed through selection from clusters of their corresponding TRS. This approach combines clustering (Willett, 1988) and query-biased text summarisation (Tombros & Sanderson, 1998) in order to present a personalised structured information space to searchers.

The resulting sentence clusters offer a view of the information space which is highly characterised by the presence of query terms. TRS contain a high proportion of query terms, and therefore each sentence can be seen as providing a local context in which these terms occur. Consequently, the information space which corresponds to TRS clusters will be restricted to these local contexts, offering a personalised view to users. We believe that users can benefit from interaction with personalised information spaces, since they may gain a better understanding of the different topics under which their search terms are discussed (this of course assumes that the selected TRS are representative of the way query terms are used in documents).

As an initial investigation of TRS clustering, we compared its effectiveness to that of document clustering in providing access to information which users find useful for completing information seeking tasks (Tombros et al, 2003). Our results demonstrated that TRS clustering provides more effective access to useful information than document clustering.

In this paper we examine a different aspect of TRS clustering by focusing on the way users interact with clusters. Issues relating to the use of document clustering in interactive information retrieval have been examined by other researchers (Hearst & Pedersen, 1996; Wu et al, 2001). The general consensus of such studies is that searchers are able to make use of the clustered organisation of documents in order to effectively access information.

User interaction with TRS clusters, however, is an issue that has not been examined. Although users still interact with groups of related objects (i.e. sentences), the information space which they are presented with is small (only top-retrieved documents and a few TRS extracted from these documents are considered) and

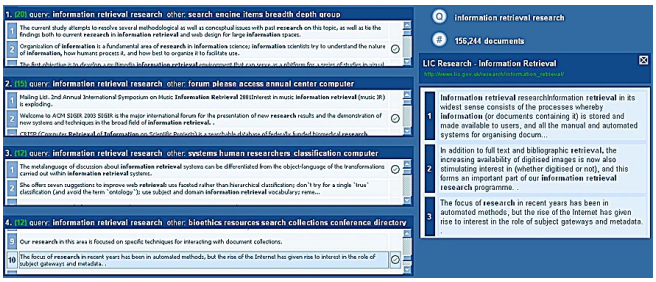

Figure 1: The TRS clustering interface.

highly characterised by the presence of search terms (most of the TRS will contain some of the query terms).

In such a space, it has not been established whether a clustered organisation and presentation of TRS is beneficial to searchers. It may be the case that because of the strong presence of query terms in this space, users are able to effectively access useful information without the need of a clustered organisation. Before further development of our research efforts in the direction of TRS clustering, we need to establish whether this form of clustering is meaningful and useful to searchers. To this end, we monitor user interaction and performance in solving information seeking tasks with a system that uses TRS clustering, and we compare them to when the same set of searchers used a "pseudo-clustered" organisation of TRS.

2 System Details

Both systems used in the study share the same interface functionality. Figure 1 shows the interface used by both systems. In response to a query, the system presents clusters of sentences which are taken from the top thirty documents in the retrieved document set. Sentences within each document are scored based on a number of factors (e.g. their position within a document, the words they contain, and the proportion of query terms they contain). A maximum of four top scoring sentences from each document are selected to form the set of top-ranking sentences. The query-biased sentence extraction model is presented in detail in (White et al, 2002).

The set of TRS is subsequently clustered. The clustering method we employ is the group average method, which has shown to be effective in IR research (Willett, 1988). A fixed number of clusters is generated (seven), both for presentation issues and for consistency between the clustered and pseudo-clustered systems.

Searchers are not shown a list of retrieved document titles and URLs in response to a query. Instead, they are shown a clustered organisation of the set of TRS (left part of Figure 1). Initially, there is no direct association between a sentence in a cluster and its source document. To view the association users must click with the mouse over a sentence. When this occurs, the sentence is highlighted and a window pops up next to the sentence. The window contains the document title, URL and the top-ranking sentences (up to four) for this specific document (right part of Figure 1). To visit the full text of the document searchers can click on the document title. The full text appears in a new window.

Clusters in the interface are ordered and presented based on the average query-biased scores of their comprising sentences. Sentences within each cluster are ordered by their query-biased scores. Each cluster in the interface is described by the number of sentences in the cluster, the query terms which occur in the cluster, and the most frequent terms which occur in the cluster (up to six such terms, excluding very frequent terms such as articles, prepositions, etc.).

The system used to generate a "pseudo-clustered" organisation of TRS is similar to the clustering system with one exception: the query-biased scores of sentences, rather than an actual clustering method, are used to assign them to clusters. More specifically, sentences are ranked based on their scores and split into seven groups (containing equal number of sentences) based on these ranks. Because of the way clusters are ordered in the interface, sentences in the top clusters will have a higher presence of query terms than those in latter clusters. To distinguish between the two systems, we will refer to them as *realCluster* and *pseudoCluster*.

3 Experimental Setup

In this section we describe the experimental setting used to compare searchers' interaction and performance with the two systems described in section 2.

3.1 Research Hypotheses

We assume that the way searchers interact with TRS clusters will reveal whether the presented structure is meaningful and beneficial. More specifically, we expect that when using the realCluster system users will interact with less clusters than when using pseudoCluster. We expect that when using realCluster, sentences which provide access to useful information will be placed in a small number of clusters. On the other hand, when using the pseudo-clustered

organisation, we expect useful sentences to be spread across a larger number of clusters, and we accordingly expect searchers to interact with a larger number of clusters in order to complete tasks. We also expect the clustered organisation offered by realCluster to allow searchers to complete information seeking tasks quicker, have a higher perception of task completion and a higher perception of utility of TRS clusters than when using pseudoCluster.

3.2 Tasks and Participants

We employed 16 searchers in the study. Most were University of Glasgow graduate students and were regular and fluent computer users.

Two task categories were used in the study, each containing three tasks of comparable difficulty. Searchers were asked to select one task from each category depending on their interest. The two categories were general background search (e.g. find information about dust related allergies) and many-items search (e.g. find five hotels in Paris with an online booking service). The tasks were presented in the context of a simulated work task situation (Borlund, 2000). Searchers were given 15 minutes to complete each task, so as to ensure some consistency among subjects.

3.3 Methodology

The experiment followed a within-subjects repeated measures design (i.e. each user performed one task from one category with one system, and one task from the other category with the other system). The order of the presentation of tasks and systems to users was counterbalanced so as to evenly distribute fatigue and learning effects.

Upon arrival searchers were briefed about the experimental process and were given a short tutorial of the interface features incorporated by the system. Searchers were not informed of the purpose of the study until they had completed both tasks. Subsequently, an informal discussion would take place. The same experimenter conducted all user sessions to minimise any biasing effects.

Data was collected by a combination of questionnaire answering and system logging. Questionnaires were issued to users before the start of the study (measuring computer and Internet usage experience, etc.), after the completion of a task (measuring task completion perception, utility of the clustered organisation, etc.), and after the completion of both tasks (asking searchers to rank the two systems). Semantic differentials, Likert scales and open-ended questions were used (Preece, 1994). Through system logging we record a number of events during each task. Such events include the time taken to complete a task, the number and order of TRS and clusters accessed by users, the query terms used and the number of query iterations made.

4 Results

In this section we present results from the participants' questionnaire answers and feedback regarding the interface, and from the analysis of the log data regarding user interaction.

4.1 Interface

All results from user questionnaires were measured on a 5-point scale, where a rating closer to 1 corresponds to a stronger agreement. Testing for statistical significance was done using the Wilcoxon signed-ranks test.

Searchers using realCluster perceived TRS clusters to be more intuitive (2.1 vs. 2.4), clear (1.9 vs. 2.3) and useful (2.2 vs. 2.5) than when using pseudoCluster. Searchers perceived TRS clusters as helping them to complete their search tasks (2.2 vs. 2.7) and helping them to see how their search terms were used in the retrieved documents (2.4 vs. 2.6) more than when using pseudoCluster. Participants perceived the search process as less stressful (2.4 vs. 2.9), more interesting (2 vs. 2.5) and less difficult (2 vs. 2.4) than when using pseudoCluster. No differences were significant.

When asked to state their preference in the two systems, most users rated them equally. However, realCluster was rated higher than pseudoCluster in four out of the five cases where different ranks were given. A number of users also commented on the way cluster contents were represented, by noting that sometimes it was difficult and cognitively demanding to recognise cluster contents from a list of keywords. The significance of cluster representations in interactive environments has been put forward by many researchers (e.g. Kural et al, 2001; Wu et al, 2001).

These results suggest that users generally had a greater perception of satisfaction when using realCluster. However, differences were mainly small and not significant. This can be explained on the basis that the interfaces and the functionality of the two systems were similar. It was difficult for users to perceive the two systems differently. However, the consistency of the results in favour of realCluster, and the comments of the users who ranked the two systems as different (mainly in favour of realCluster) provide evidence that users recognised some difference in the two systems.

Some significant evidence that the two systems were used differently comes from the perception of task completion, where users had a significantly higher perception when using realCluster (1.7 vs. 2.4, p=0.03). This result suggests that the clustered organisation of realCluster was beneficial to searchers for the

completion of tasks. In the next section we examine the data from user logs in order to establish differences in the way searchers interacted with the two systems.

4.2 Interaction

Participants using pseudoCluster made a higher number of query iterations per task (2 vs. 2.4). One possible reason for this might be that searchers did not find the pseudo-clustered organisation helpful in certain cases. In many iterations using pseudoCluster (10 in total), searchers would select a single sentence from a cluster and then formulate a new query almost instantly. Similar cases using realCluster were less (4).

The average number of clusters accessed per user was lower when using realCluster (2.6 vs. 2.7). Searchers using realCluster also accessed more sentences per cluster (4.5 vs. 4.1). An analysis of the log data showed no significant differences in the time taken to complete tasks when using realCluster and pseudoCluster (average time of 10 min. 25 sec. vs. 10 min. 19 sec. per task respectively).

The log data also show that, on average, searchers using realCluster accessed a higher percentage of sentences from a single cluster (64.4% vs. 58.8%, differences not significant). The dispersion of sentence accesses across clusters was higher for users using realCluster (0.22 vs. 0.14 as measured by the standard deviation of sentence accesses per cluster, differences not significant). The larger dispersion of accesses for realCluster suggests that searchers using this system tend to access sentences more unevenly across clusters, focusing on fewer clusters. Searchers using pseudoCluster, on the other hand, tend to access sentences more uniformly across clusters.

The results presented in this section suggest that there are differences in the way searchers interact with the two systems. Users examined fewer clusters when using realCluster, and they also selected a higher percentage of sentences from a single cluster. Searchers also spread their sentence accesses more unevenly across clusters, focusing on fewer clusters than when using pseudoCluster. These results, although not significant, provide a consistent trend which suggests that searchers are able to recognise the clustered structure imposed by realCluster, and to focus on the few clusters which contain information useful for the completion of the task at hand. This consistent trend becomes more important if we view it in conjunction with the significantly higher perception of task completion that searchers had when using realCluster.

5 Conclusions

This work is a preliminary investigation into the utility of top-ranking sentence clustering as a method for the presentation of search results to users. The results of this study provide evidence that searchers are able to recognise, and benefit from, the clustered organisation of the query-biased information space which corresponds to the set of top-ranking sentences.

Based on this evidence, we aim to take our investigation of TRS clustering further. As the next step of our research, we plan to use TRS clustering to structure the personalised information space, and to use the interaction in this space for mining the information need of users in a way similar to (White et al, 2002).

Acknowledgements

The authors wish to thank the searchers who participated in the study. This research is funded by the EPSRC (U.K.) research grant GR/R74642/01.

References

Borlund, P. (2000), Experimental components for the evaluation of interactive information retrieval systems, *Journal of Documentation* 56(1), 71-90.

Hearst, M.A. & Pedersen, J.O. (1996), Re-examining the cluster hypothesis: Scatter/Gather on retrieval results, in *Proceedings of the 19th ACM SIGIR Conference*, pp. 76-84.

Kural, Y., Robertson S.E. & Jones, S. (2001), Deciphering cluster representations, *Information Processing & Management* 37(4), 593-601.

Preece, J. (1994), *Human Computer Interaction*, Addison Wesley.

Tombros, A. & Sanderson, M. (1998), The advantages of query-biased summaries in IR, in *Proceedings of the 21st ACM SIGIR Conference*, pp. 2-10.

Tombros, A., Jose, J.M. & Ruthven, I. (2003), Clustering Top-Ranking Sentences for Information Access, in *Proceedings of the 7th ECDL Conference*, to appear.

White, R.W., Ruthven, I., Jose, J.M. (2002), Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes, in *Proceedings of the 24th ACM SIGIR Conference*, pp. 57-64.

Willett, P. (1988), Recent trends in hierarchic document clustering: A critical review, *Information Processing & Management* 24(5), 577-597.

Wu, M., Fuller, M. & Wilkinson, R. (2001), Using clustering and classification approaches in interactive retrieval, *Information Processing & Management* 37(3), 459-484.