Clustering Top-Ranking Sentences for Information Access

Anastasios Tombros[1], Joemon M. Jose[1], and Ian Ruthven[2]

[1] Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, U.K.
{tombrosa, jj}@dcs.gla.ac.uk
[2] Department of Computer and Information Sciences, University of Strathclyde,
Glasgow G1 1XH, U.K.
Ian.Ruthven@cis.strath.ac.uk

Abstract. In this paper we propose the clustering of top-ranking sentences (TRS) for effective information access. Top-ranking sentences are selected by a query-biased sentence extraction model. By clustering such sentences, we aim to generate and present to users a personalised information space. We outline our approach in detail and we describe how we plan to utilise user interaction with this space for effective information access. We present an initial evaluation of TRS clustering by comparing its effectiveness at providing access to useful information to that of document clustering.

1 Introduction

One of the challenging research issues in Digital Libraries is the facilitation of efficient and effective access to large amounts of available information. Document clustering [1] and automatic text summarisation [2] are two methods which have been used in the context of information access in digital libraries.

Document clustering generates groupings of potentially related documents by taking into account interdocument relationships. By taking into account interdocument relationships, users have the possibility to discover documents that might have otherwise been left unseen [3]. Document clusters, effectively, reveal the structure of the document space. This space however, may not help users understand how their search terms relate to the retrieved documents, which can be long and contain many topics. Therefore, the information space offered by document clusters to users is essentially not representative of their queries.

Text summarisation, in the context of information access, offers short previews of the contents of documents, so that users can make a more informed assessment of the usefulness of the information without having to refer to the full text of documents [2, 4]. A particular class of summarisation approaches, query-oriented or query-biased approaches, have proven effective in providing users with relevance clues [4]. Query-biased summaries present to users textual parts of documents (usually sentences) which highly match the user's search terms. The effectiveness of such summaries in the context of interactive retrieval on the World Wide Web has been verified by [4].

The aim of this work is to reveal a personalised information space to users by restructuring the initial document space. To this end, we combine clustering and summarisation in a novel way. We cluster sentences which have been selected by a query-biased sentence extraction model (*top-ranking sentences, TRS*) [4]. The sentences form part of single document summaries which represent top-ranked documents retrieved in response to a query. The resulting sentence clusters offer a view of the initial information space which is highly characterised by the presence of query terms. The overall objective of this approach is to facilitate a more effective user interaction with the personalised information space, and to utilise this interaction for improving the quality of the information presented to users.

In this paper, we mainly focus on two issues. First, we present our approach and its aims in detail in section 2. Then, we present an initial evaluation of TRS clustering by comparing its effectiveness at providing access to useful information to that of document clustering in section 3. We conclude in section 4, where we also outline how we propose to take this work further.

2 Clustering Top-Ranking Sentences

The essence of our approach consists of generating a list of top-ranking sentences for each document retrieved in the top-ranks in response to a user query, and of clustering these sentences. The set of top-ranking sentences constitutes a summary for each of the documents. These sentences are selected through a query-biased sentence extraction model, presented in detail in [4]. Sentences are scored according to factors such as their position within a document, the words they contain, and the proportion of query terms they contain. A number of the highest scoring sentences can then be selected as the summary. Clusters of TRS can be generated by any clustering method, such as hierarchic methods which are commonly used in information retrieval systems [3], or methods which are specifically designed to cluster short textual units (e.g. [5]).

The main function of TRS clusters is to provide effective access to retrieved documents by acting as an abstraction of the information space. Essentially, TRS clusters form a second level of abstraction, where the first level corresponds to summaries (i.e. sets of TRS) of each of the retrieved documents. Instead of interacting with the retrieved document set, users can access documents by browsing through clusters of TRS. Individual TRS are linked to the original documents (or to representations of the original documents, such as titles, summaries, etc.) in which they occur so that users can access the original information.

Sentences within a single TRS cluster will discuss query terms in the context of the same (or similar) topics. This can assist users in better understanding the structure and the contents of the information space which corresponds to the top-retrieved documents. This may be especially useful in cases where users have a vague, not well-defined information need.

It should be noted that the information space which corresponds to clusters of TRS is different to the one which corresponds to the top-retrieved documents. TRS contain a high proportion of query terms, and therefore each sentence can be seen as providing a local context in which these query 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 through interaction with personalised information spaces, since they may gain a better understanding of the different topics under which the query terms are discussed (this of course assumes that the selected TRS are representative of the way query terms are used in documents).

The overall objective of our approach is to utilise information resulting from the interaction of users with this personalized information space in the form of implicit feedback [6]. As mentioned previously, users can access documents, or other shorter representations of documents such as titles and query-biased summaries, by selecting individual sentences in TRS clusters. User interaction with TRS clusters, individual documents and other document representations can be monitored, and the information collected can be used to recommend new documents to users, and to select candidate terms to be added to the query from the documents and clusters viewed. This type of implicit feedback has been used by [6] in order to utilise information from the interaction of users with query-biased document summaries, and has shown to be effective in enabling users to access useful information. The system which combines TRS clustering and implicit feedback is currently under development.

From the previous discussion, some similarities between document and TRS clustering become apparent. Both approaches present an abstracted version of the information space in a structured view which facilitates browsing and interaction. Moreover, both approaches aim to provide users with effective access to useful retrieved information through interaction with the grouped documents or sentences. There are, however, some significant differences between the two approaches. TRS clustering uses finer textual units (sentences instead of full documents), and more importantly, it alters the information space by using textual units which are highly characterised by the presence of query terms. The structuring of the information space by document clustering is not tailored to the query since it offers a grouping of documents which may be long and contain many topics. By using query-biased sentences as the items to be clustered, we offer users a view of the information space which is focused on their query terms.

In the next section, we perform an initial evaluation of the effectiveness of TRS clustering at providing access to useful information. To establish whether pursuing TRS clustering is worthwhile, we compare its effectiveness to that of document clustering. In this way, we can get an indication of whether TRS clustering has the potential to act as a medium for effective information access by improving the quality of the provided information when compared to document clustering. It should be noted that in this initial evaluation we do not use information from the users' interaction with clusters. We plan to evaluate aspects of interaction when the system which combines TRS clustering and implicit feedback is completed.

3 Comparing TRS and Document Clustering

For this study we used 16 queries which represented actual information needs. The queries were generated by 4 users. The average length of the queries was 3.7 terms. Each of the queries was input to a web-based IR system [4] which retrieved and presented to the searcher the top-30 retrieved documents. The full text of each of the web pages was downloaded. Each searcher was asked to examine each of the documents retrieved, and to assign a numerical value to it representing his assessment of how useful he found the document in relation to his query. The assessments were on a scale from 1 (not at all useful) to 10 (very useful). We did not require for the documents to be visited in any particular order, and we allowed users to adjust their assessments as they wished. There was no time limit imposed on users.

For each query Q_i, the top-30 retrieved documents were clustered, generating a document clustering DC_i. The top-ranking sentences for each of these documents (maximum four sentences per document, depending on document length) were also extracted, in a procedure reported in [4]. This generated a respective sentence clustering SC_i. Both document and TRS clustering was performed using the group average link method [3]. It is worth noting that for some queries it was not possible to download all 30 top-ranked documents (for example some documents may not be available). On average, 23 documents were downloaded per query, and 3.2 sentences per document were extracted.

The user assessments were used to assign scores to each of the two clusterings. More specifically, for a document cluster DC_i, the score assigned is a sum of the assessment scores of its comprising documents, normalised by the number of documents in the cluster. For a sentence cluster SC_i, the score assigned is a sum of the assessment scores of the documents in which each of the cluster's TRS belongs to, normalised by the number of sentences in the cluster. The type of clustering which produces the highest score is the one which has the potential to provide users with the more useful information.

3.1 Results

In Table 1 we present a summary of the results for document clustering (DC) and TRS clustering (SC). In columns 2 and 3 we present the average score for the best cluster for each query and for all clusters, respectively. In both cases TRS clustering produces a significantly higher score than document clustering (Wilcoxon signed-ranks test, $p<0.05$). Only in 2 out of the 16 queries document clusters produced a higher score. The average size of the best DC and SC was comparable (6.5 and 7.75 items per cluster respectively). TRS clusters also display a lower standard deviation across all scores (column 4).

In columns 5 and 6 we give the average highest precision and highest recall for SC and DC across all queries for the best clusters. In order to calculate these values, we considered, for each query, the set of documents for which users assigned a score in the range of 7-10 as the set of relevant documents. On average, there were 4.2 such documents per query. We view such documents as being the most useful for users. The values in Table 1 demonstrate that SC show a significantly higher average precision ($p<0.05$), and a higher average recall. The relatively low precision for both types of clustering can be explained on the basis of the relatively few "relevant" documents per query.

Table 1. Summary of results

	Avg. best score	Avg. overall score	Std. deviation of overall scores	Avg. P	Avg. R
DC	4.78	3.18	1.38	0.38	0.73
SC	5.82	3.73	1.12	0.49	0.77

An analysis into the composition of TRS clusters shows that the average size is 5.3 sentences per cluster (compared to 5 documents per cluster for DC). It should be noted that all results presented here for SC have been calculated by considering only one occurrence of TRS from the same document in each cluster. On average, across all queries, 36% of sentences in TRS clusters corresponded to multiple occurrences of TRS from the same original document, a result which is a consequence of the high similarity of TRS from the same documents.

In general, our results suggest that TRS clusters provide access to more useful information than document clusters, and that they also manage to structure the document space in a more effective way than document clusters. Moreover, TRS clusters provided more effective access to the highly useful documents (as these were indicated by the users themselves) than documents clusters.

4 Conclusions

The results we presented in the previous section demonstrate that there is scope for the application of TRS clustering. Although the study was of a small scale, its results are positive and they suggest that TRS clusters have the potential to lead users to parts of the information space which contain useful information.

To the best of our knowledge, combining document clustering and text summarisation to create a personalised information space with the aim of utilising the users' implicit feedback is a novel approach. Document clustering and summarisation are typically combined for the purposes of multiple-document summarisation (e.g. [7, 8]), where sets of related documents, or of their summaries, are clustered in order to select sentences to be included in a summary.

We plan to examine in more detail the characteristics of generated TRS clusters, and to consider the effect of different clustering methods. The effect of the query-biased model, which generates the TRS, on the generated clusters also needs to be considered. In section 3 we presented results for the best SC and DC. Whether users in an interactive environment will be able to recognise the best cluster depends on how cluster contents are summarised and displayed on the interface level. This is a challenging research issue [9] which does not fall within the aims of this paper.

The overall objective of our approach is to integrate TRS clustering in an interactive environment, and to utilise information from the users' interaction with TRS clusters. We believe that users will benefit from interaction with the personalised information space which is generated by TRS clusters. Although devoid of the interaction aspect, the results we reported in this section paper that TRS clusters have the potential to lead users to useful information. We view these results as suggesting that TRS clustering can provide effective access to information, and we plan to build on this research in order to incorporate aspects of interaction in the TRS clustering system.

Acknowledgements

The authors wish to thank Ryen W. White for his help with the top-ranking sentences system. This research is funded by the EPSRC (U.K.) research grant GR/R74642/01.

References

1. Leuski, A., Allan J.: Evaluating a visual navigation system for a digital library. In: Proceedings of the 2nd ECDL Conference, Heraklion, Greece (1998) 535–554
2. Lopez, M.J.M., Rodriguez, M.B., Hidalgo, J.M.G.: Using and evaluating user directed summaries to improve information access. In: Proceedings of the 3rd ECDL Conference, Paris, France (1999) 198–214
3. Willett, P.: Recent trends in hierarchic document clustering: a critical review. Information Procsessing & Management **24** (1988) 577–597
4. White, R.W., Ruthven, I., Jose, J.M.: A task-oriented study on the influencing effects of query-biased summarisation in web searching. Information Procsessing & Management in press (2003)
5. Zamir, O., Etzioni, O.: Web document clustering: A feasibility demonstration. In: Proceedings of the 21st Annual ACM SIGIR Conference, Melbourne, Australia (1998) 46–54
6. White, R.W., Ruthven, I., Jose, J.M.: Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In: Proceedings of the 24th Annual ACM SIGIR Conference, Tampere, Finland (2002) 57–64
7. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the ANLP/NAACL Workshop on Summarization, Seattle, U.S.A. (2000)
8. Zha, H.: Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: Proceedings of the 25th Annual ACM SIGIR Conference, Tampere, Finland (2002) 113–120
9. Kural, Y., Robertson, S.E., Jones, S.: Deciphering cluster representations. Information Procsessing & Management **37** (2001) 593–601