Proceedings of HLT-NAACL 2003
Demonstrations , pp. 11-12
Edmonton, May-June 2003

QCS: A Tool for Querying, Clustering, and Summarizing Documents

Daniel M. Dunlavy
University of Maryland
ddunlavy@cs.umd.edu

John Conroy
IDA/CCS
conroy@super.org

Dianne P. O'Leary
University of Maryland
oleary@cs.umd.edu

Abstract

The QCS information retrieval (IR) system is presented as a tool for querying, clustering, and summarizing document sets. QCS has been developed as a modular development framework, and thus facilitates the inclusion of new technologies targeting these three IR tasks. Details of the system architecture, the QCS interface, and preliminary results are presented.

1 Introduction

QCS is a software tool and development framework for efficient, organized, and streamlined IR from generic document sets. The system is designed to match a query to relevant documents, cluster the resulting subset of documents by topic, and produce a single summary for each topic. Using QCS for IR, the amount of redundant information presented to a user is reduced and the results are categorized by content.

A survey of previous work using a combination of clustering and summarization to improve IR can be found in Radev et al. (2001b). Of existing IR systems employing this combination, QCS most resembles the NewsInEssence system (Radev et al., 2001a) in that both systems can produce multi-document summaries from document sets clustered by topic. However, NewsInEssence is designed for IR from HTML-linked document sets and QCS has been designed for IR from generic document sets. Furthermore, one of the most important aspects of QCS is its modularity, with the ability to plug in alternative implementations of query-based retrieval, document clustering, and summarization algorithms.

2 Querying, Clustering, Summarizing

QCS employs a vector space model (Salton et al., 1975) to represent a set of documents. Choices for the term weighting currently include the following:

- *Local*: term frequency, log, binary
- *Global*: none, normal, idf, idf2, entropy
- *Normalization*: none, normalized

Detailed descriptions of each of these weighting factors as well as strategies for using each of these are presented by Dumais (1991) and Kolda and O'Leary (1998).

The current computational methods used for retrieving a set of documents that best match a query, clustering a set of documents by topic, and creating a summary of multiple documents are as follows:

- *Querying*: Latent Semantic Indexing (LSI)
- *Clustering*: spherical k-means
- *Summarization*: a hidden Markov model (HMM) and pivoted QR

Detailed descriptions of these methods presented in Deerwester et al. (1990), Dhillon and Modha (2001), and Schlesinger et al. (2002), respectively.

The interface to QCS (see Figure 1) consists of a collection of Java^{TM} [1] servlets which format input to and output from QCS via dynamic HTML documents. This approach allows all of the computation and formatting to take place on a Java^{TM} server, with the only requirement on the users' systems being that of an HTML-enabled browser application (e.g., Netscape® 7.0).

3 Results

QCS was tested using data from the 2002 Document Understanding Conference (http://duc.nist.gov/), a conference focusing on summarization and the evaluation of summarization systems. The data consisted of 567 news articles categorized into four types, with one type consisting of articles covering a single natural disaster event reported within a seven day window.

Results of one test producing 100-word extract summaries can be seen in Figure 1, where the query consisted of the words, "hurricane" and "earthquake". The top three scoring clusters contained a total 55 articles (32, 11, and 12, respectively), producing the summaries shown in the figure. The topics of these three summaries were a hurricane near Jamaica, catastrophe insurance, and an earthquake in California, respectively. Despite the limitations of automatic summarization, this example illus-

[1] Java^{TM} is a trademark of Sun Microsystems, Inc.

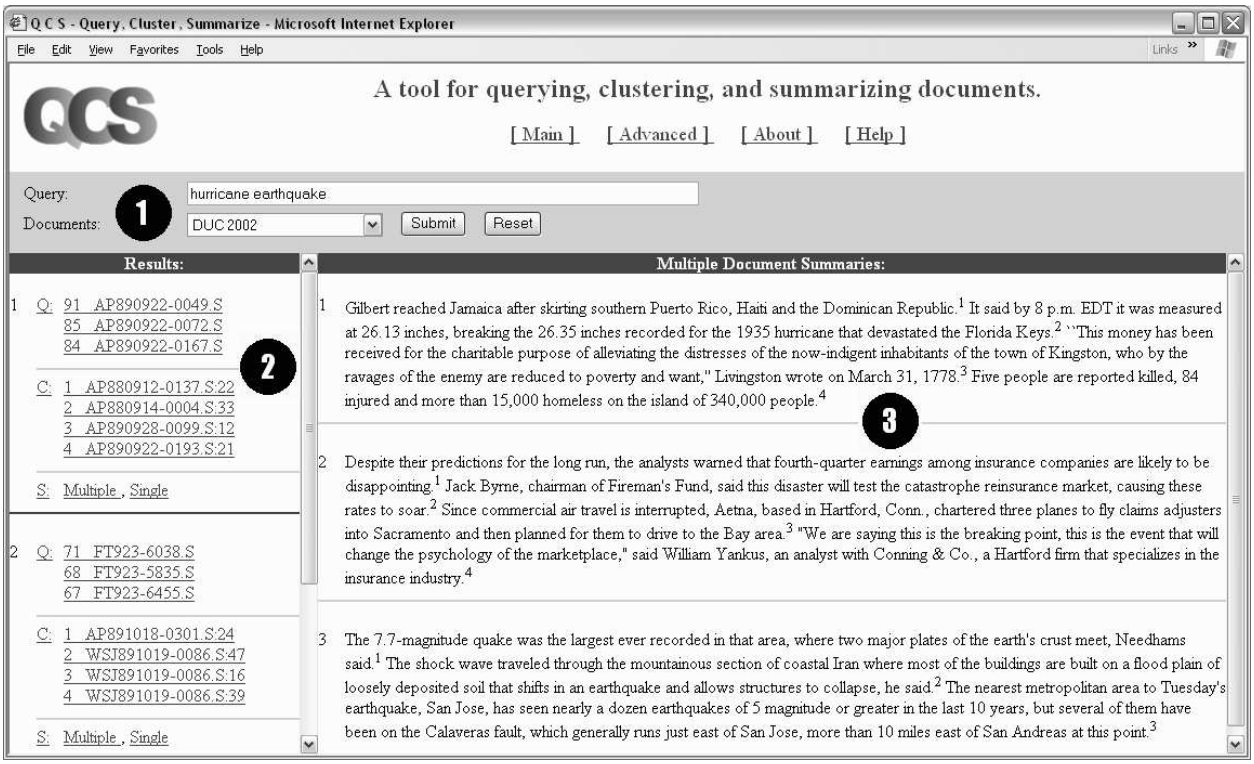

Figure 1: The interface to the QCS system includes 1) an *input* section for the query and choice of document set, 2) a *navigation* section with links to clustered documents (Q: top documents retrieved for the query and their scores, C: documents from which summary sentences were drawn and the sentence indices, S: links to multiple or single document summaries), and 3) an *output viewing* section, which here contains the default output of multiple document summaries for the topic clusters.

trates the utility of summarizing by cluster rather than producing a single summary of the retrieved documents.

Further results are planned for the demonstration, including results of using QCS against the data from the 2003 Document Understanding Conference.

Acknowledgements

The authors would like to thank C. David Levermore and William D. Dorland of the University of Maryland for their helpful remarks concerning the QCS system.

References

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Inderjit S. Dhillon and Dharmendra S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175.

Susan T. Dumais. 1991. Improving the retrieval of information from external sources. *Behav. Res. Meth. Instr.*, 23(6):229–326.

Tamara G. Kolda and Dianne P. O'Leary. 1998. A semidiscrete matrix decomposition for latent semantic indexing in information retrieval. *ACM Trans. Inf. Sys.*, 16(4):322–346.

Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001a. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference*, San Diego, CA.

Dragomir R. Radev, Weiguo Fan, and Zhu Zhang. 2001b. Webinessence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*, Pittsburgh, PA.

Gerard Salton, A. Wong, and C.S. Yang. 1975. A vector space model for information retrieval. *Communications of the ACM*, 18(11):613–620.

Judith D. Schlesinger, Mary Ellen Okurowski, John M. Conroy, Dianne P. O'Leary, Anthony Taylor, Jean Hobbs, and Wilson Harold T. Wilson. 2002. Understanding machine performance in the context of human performance for multi-document summarization. In *Proc. of the Workshop on Automatic Summarization*.