Advanced Scientific Computing Research
Applied Mathematics
FY 2007 Accomplishment

"Improved Information Retrieval: Organizing Retrieved Documents by Topic, and Automatically Generating Summaries for Each Topic"

Daniel M. Dunlavy*, John von Neumann Fellow, Sandia National Laboratories**
Dianne P. O'Leary, University of Maryland, College Park
John M. Conroy and Judith D. Schlesinger, IDA/CCS

Summary

The modern scientific process requires researchers to search large, global databases of accumulated scientific knowledge on which to base their research. Likewise, new and emerging cyber security issues threaten the open science research enterprise and tools are required to quickly identify malicious behaviors in large textual databases (network logs, email, etc). With my collaborators and as part of my John von Neumann Fellowship, we are investigating novel Information Retrieval (IR) systems that provide access to a vast amounts of reference material while addressing the challenge of effectively presenting only relevant information in response to a query. When using existing IR engines to search through electronic resources, simple queries often return too many documents, including many that are not relevant to the intended search. We have developed a novel integrated information retrieval system—the Query, Cluster, Summarize (QCS) system—which is portable, modular, and permits experimentation with different instantiations of each of the constituent text analysis components. Most importantly, the combination of the three types of methods in QCS improves retrievals by providing more focused information organized by topic that will speed and enhance the scientific process and may serve as a foundational tool for Cyber Security R&D in open science.

We have developed a novel approach for information retrieval in which the following tasks are performed in response to a query:

- relevant documents are retrieved,
- retrieved documents are separated into clusters by topic, and
- a summary is created for each cluster.

The QCS system partitions the code into portable modules, making it easy to experiment with different methods for handling the three main tasks listed above.

The method used for query-based document retrieval in QCS is *latent semantic indexing (LSI)*, which attempts to reveal latent relationships caused by term ambiguity, while preserving the most characteristic features of each document. This allows for conceptual matching of queries to documents, which has been shown to produce better retrieval results when simple or ambiguous queries are used.

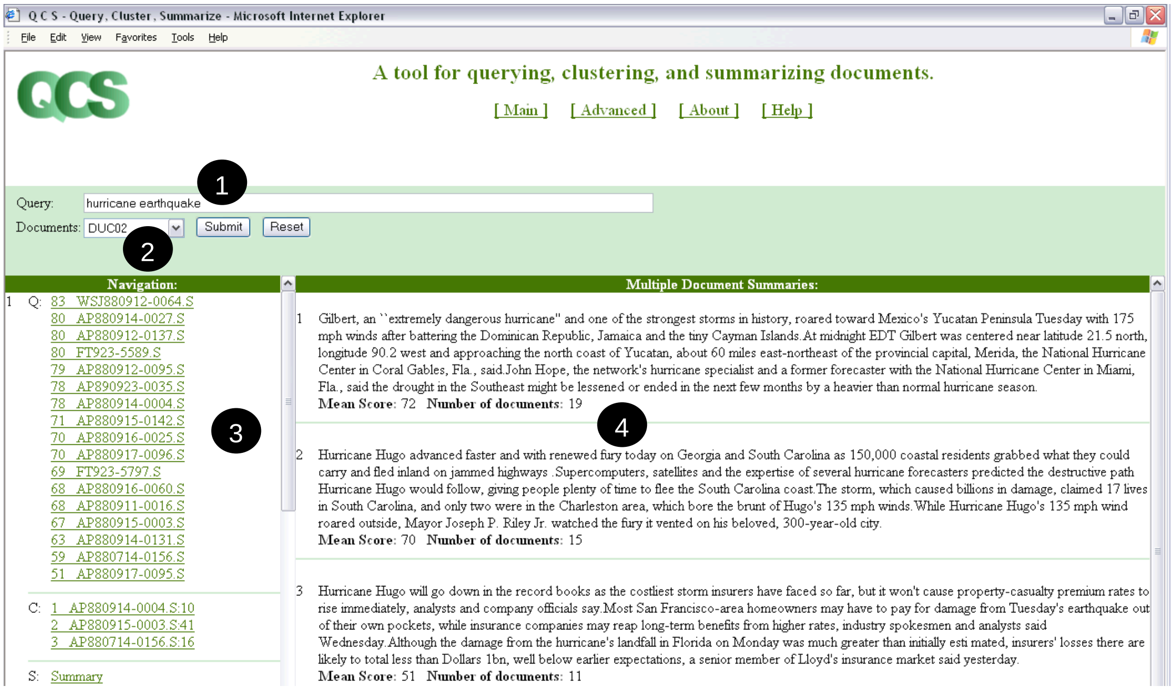
We use the information derived from the query processing phase to cluster documents into a *variable* number of clusters, each representing a single topic. The algorithmic complexity of general clustering methods is not an issue in QCS, as clustering is

*505-284-6092, dmdunla@sandia.gov; **Sandia is a multiprogram laboratory operated by Sandia Corporation, A Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. SAND2007-4460P.

performed only on the retrieved documents (a relatively small set of data), and the rank values produced by LSI are used to estimate the clusters. This results in computation reduced from several hours to several seconds for most tests performed on QCS to date. *Generalized spherical k-means* is currently used for clustering in QCS.

We use the summarization method we developed for the 2002-2004 Document Understanding Conference (DUC) evaluations administered by the National Institute for Standards and Technology (NIST). Single document summaries are produced using a hidden Markov model (HMM) to compute the probability that each sentence is a good summary sentence. A multi-document summary is created for each cluster by choosing a subset of these sentences. In order to remove redundant sentences, a matrix algorithm is applied to vector representations of the terms in the candidate summary sentences. The result is a set of linearly independent vectors (i.e., sentences) ordered by probability of being good summary sentences. Since the vectors model the terms in the sentences, the result is a set of sentences with little semantic redundancy.

We have used QCS for retrieval in two information domains: biomedical abstracts from the National Library of Medicine's MEDLINE database, and newswire documents from the DUC evaluations. Results of using QCS on these sets illustrate the usefulness of this system; in particular, we provide evidence of the value of clustering as a tool for increasing the quality of the summaries. The benefit of using QCS over existing methods is that it is a *fully automatic* system for document retrieval, organization, and summarization.


Q C S - Query, Cluster, Summarize - Microsoft Internet Explorer
File Edit View Favorites Tools Help
QCS
A tool for querying, clustering, and summarizing documents.
[Main] [Advanced] [About] [Help]
1
2
3
4
Submit
Reset
Navigation:
Multiple Document Summaries:

Query: hurricane earthquake

Documents: DUC02

1 Q: 83 WSJ880912-0064.S
80 AP880914-0027.S
80 AP880912-0137.S
80 FT923-5589.S
79 AP880912-0095.S
78 AP890923-0035.S
78 AP880914-0004.S
71 AP880915-0142.S
70 AP880916-0025.S
70 AP880917-0096.S
69 FT923-5797.S
68 AP880916-0060.S
68 AP880911-0016.S
67 AP880915-0003.S
63 AP880914-0131.S
59 AP880714-0156.S
51 AP880917-0095.S

C: 1 AP880914-0004.S:10
2 AP880915-0003.S:41
3 AP880714-0156.S:16

S: Summary

1 Gilbert, an ``extremely dangerous hurricane'' and one of the strongest storms in history, roared toward Mexico's Yucatan Peninsula Tuesday with 175 mph winds after battering the Dominican Republic, Jamaica and the tiny Cayman Islands.At midnight EDT Gilbert was centered near latitude 21.5 north, longitude 90.2 west and approaching the north coast of Yucatan, about 60 miles east-northeast of the provincial capital, Merida, the National Hurricane Center in Coral Gables, Fla., said.John Hope, the network's hurricane specialist and a former forecaster with the National Hurricane Center in Miami, Fla., said the drought in the Southeast might be lessened or ended in the next few months by a heavier than normal hurricane season.
Mean Score: 72 **Number of documents:** 19

2 Hurricane Hugo advanced faster and with renewed fury today on Georgia and South Carolina as 150,000 coastal residents grabbed what they could carry and fled inland on jammed highways .Supercomputers, satellites and the expertise of several hurricane forecasters predicted the destructive path Hurricane Hugo would follow, giving people plenty of time to flee the South Carolina coast.The storm, which caused billions in damage, claimed 17 lives in South Carolina, and only two were in the Charleston area, which bore the brunt of Hugo's 135 mph winds.While Hurricane Hugo's 135 mph wind roared outside, Mayor Joseph P. Riley Jr. watched the fury it vented on his beloved, 300-year-old city.
Mean Score: 70 **Number of documents:** 15

3 Hurricane Hugo will go down in the record books as the costliest storm insurers have faced so far, but it won't cause property-casualty premium rates to rise immediately, analysts and company officials say.Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets, while insurance companies may reap long-term benefits from higher rates, industry spokesmen and analysts said Wednesday.Although the damage from the hurricane's landfall in Florida on Monday was much greater than initially esti mated, insurers' losses there are likely to total less than Dollars 1bn, well below earlier expectations, a senior member of Lloyd's insurance market said yesterday.
Mean Score: 51 **Number of documents:** 11

Figure 0. Screen shot of the Query, Cluster, Summarize (QCS) software tool. A user enters a query in (1), chooses a documents collection to query in (2), navigates the retrieval and topic cluster results in (3), and views multi-document summaries (shown here) and individual documents in (4).

Reference

Dunlavy, O'Leary, Conroy, and Schlesinger. QCS: A System for Querying, Clustering and Summarizing Documents, *Inform. Process. & Manag.*, to appear (online: http://dx.doi.org/10.1016/j.ipm.2007.01.003).

For further information contact:

Dr. Anil Deane
Applied Mathematics Research Program
Office of Advanced Scientific Computing
Phone: 301-903-1465
deane@ascr.doe.gov