

# An Information Retrieval System for Improving Efficiency in Scientific Literature Searches

Danny Dunlavy

# Acknowledgements

- Dianne O'Leary  
University of Maryland
- John Conroy  
Center for Computing Sciences

# Introduction

- Scientific Research
  - Literature search
    - Journal articles, tech. reports, preprints
    - Electronic resources
      - WWW resources: Google
      - Preprint servers: arXiv
      - Online journals
- Need good retrieval methods
  - Problems:
    - Too much information
    - Not enough time

# Motivation

- Query
  - “methods plasma physics”
- Retrieval
  - Google: 27,000/3.2×10<sup>9</sup> documents
  - arXiv: 350/232,000 documents
- Problems
  - Google: too much, redundant (!!)
  - arXiv: limited scope
  - Results: link, title, abstract, etc.

# Project Goals

- The QCS System
  - Query: retrieve relevant documents
  - Cluster: group documents by topic
  - Summarize: one summary per cluster
- Implementation
  - Existing algorithms
  - Parallel computation
  - Platform independence for users

# Examples

- Document Understanding Conference (DUC)
  - Evaluation of automatic summarization systems
- DUC 2002 Test Documents
  - 567 documents, 7767 unique terms (words)
  - AP, LA Times, WSJ, SJMN, FBIS, Fin. Times
  - Four major topic areas (~10 docs/topic)
    - Single natural disaster event within seven day coverage
    - Any single event within seven day coverage
    - Multiple distinct events of a single type
    - Biographical information about a single person

# Representing Documents

- **Vector Space Model**

- Terms:  $\{t_1, \dots, t_m\} \in \mathbf{R}^m$

- Documents:  $\{d_1, \dots, d_n\} \in \mathbf{R}^n$

- Term  $\times$  Document Matrix:  $A$

- $a_{ij}$ : measure of importance of term  $i$  in document  $j$

$$\Rightarrow$$

	$d_1$	$\dots$	$d_n$
$t_1$	$a_{11}$	$\dots$	$a_{1n}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$t_m$	$a_{m1}$	$\dots$	$a_{mn}$

	$d_1$	$d_2$	$d_3$	$d_4$
hurricane	2	1	0	0
earthquake	0	0	1	2
catastrophe	1	1	0	1

# Querying – Goals

- Retrieve documents matching a query
  - Use more than exact matching
    - Pseudonyms: Mark Twain vs. Samuel Clemens
    - Synonyms: method, technique, algorithm
    - Stemming: method, methods, methodology
- Rank documents
  - Function of “matching” terms

# Querying – Details

- Latent Semantic Indexing (LSI)
  - Low rank approximation of term-document matrix via truncated singular value decomposition (SVD)
  - Blurs distinction between terms / between documents
- Scoring documents against query
  - Cosine similarity scores:  $s = q^T A$

	$q$	$A$				$A_2$					
		$d_1$	$d_2$	$d_3$	$d_4$	$d_1$	$d_2$	$d_3$	$d_4$		
hurricane	1	.89	.71	0	0	.78	.78	-.11	.11		
earthquake	0	0	0	1	.89	-.03	.02	.96	.92		
catastrophe	0	.45	.71	0	.45	.59	.60	.15	.30		
$q^T A$		.89	.71	0	0	$q^T A_2$		.78	.78	–	.11

# Querying – Example

- Query: “hurricane earthquake”

- Results:

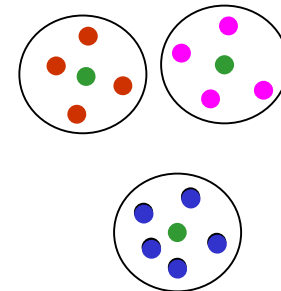
90 Hurricane Latest in String of Disasters to Hit Historic City  
85 Hurricane Forecasters Carry On Amid Chaos  
85 Forecasting Aided By Supercomputers, But Still An Uncertain Science  
84 Killer Storm Hits South Carolina Coast  
83 Scientists: Warming Trends Could Mean Fiercer Hurricanes  
82 City Sends Money To Charleston In Repayment Of 211-year-old Debt  
82 150,000 Take Off As Hugo Takes Aim At Ga., Carolina  
82 Loss Of Life Low Because People Were Prepared  
81 Hurricane Gilbert Heading for Jamaica With 100 MPH Winds  
80 Gilbert: Third Force 5 Hurricane This Century  
... ..

# Clustering – Goals

- Cluster retrieved documents into topics
  - Documents containing related information
  - Allow for variable number of clusters
- Use query results to initialize clustering
  - Clustering from random start is expensive
  - Initial clusters will contain documents that “match” query to the same extent

# Clustering – Details

- Generalized Spherical K-Means
- The Players
  - Documents:  $\{d_j\}$  ( $j = 1, \dots, n$ )
  - Partition/Disjoint Sets:  $\{\pi_l\}$  ( $l = 1, \dots, k$ )
  - Concept vectors (centroids):  $\{c_l\}$  ( $l = 1, \dots, k$ )
- The Game
  - Maximize  $\sum_{l=1}^k \sum_{d_j \in \pi_l} d_j^T c_l$
- The Rules
  - Adaptive  $k$
  - Similarity Estimation



# Clustering – Example

- Query: “hurricane earthquake”
- Results:
  - 83 Scientists: Warming Trends Could Mean Fiercer Hurricanes
  - 81 Hurricane Gilbert Heading for Jamaica With 100 MPH Winds
  - 80 Gilbert: Third Force 5 Hurricane This Century
  - ...
  - 90 Hurricane Latest in String of Disasters to Hit Historic City
  - 85 Hurricane Forecasters Carry On Amid Chaos
  - 82 City Sends Money To Charleston In Repayment Of 211-year-old Debt
  - ...
  - 48 The Bay Area Quake
  - 47 World Series; Earthquakes -- San Francisco
  - 47 Area Where Earthquake Hit Seen As Highly Probable in 1988 Report
  - ...

# Summarizing – Goals

- Create a single summary for each cluster
  - Create summary of each document
  - Create multi-document summary
- Remove redundant information

# Summarizing – Details

- Single Document Summarization

- Mark summary sentences in training documents

- Build probabilistic model

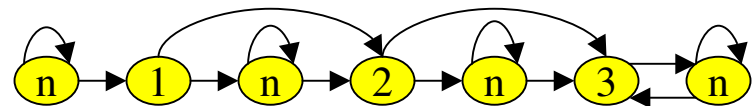
- Hidden Markov Model (HMM)

- Hidden states: {summary, non-summary}

- Observations

- $\log(\#\text{topic terms} + 1)$

- $\log(\#\text{subject terms} + 1)$



- Model state output as multivariate Gaussian

- Score sentences in each document

- Probabilities of sentence being a summary sentence

# Summarize – Details (cont.)

- Multi-document Summarization

- Goal: generate  $w$ -word summaries
- Use HMM scores to select candidate sentences ( $\sim 2w$ )
- Terms as sentence features

- Terms:  $\{t_1, \dots, t_m\} \in \mathbf{R}^m$
- Sentences:  $\{s_1, \dots, s_n\} \in \mathbf{R}^n$
- Scaling:  $\| \mathbf{a} \| = \text{HMM score}$

$$\Rightarrow \begin{array}{c|ccc} & s_1 & \cdots & s_n \\ \hline t_1 & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ t_m & a_{m1} & \cdots & a_{mn} \end{array}$$

- Pivoted QR

- Choose column with maximum norm ( $\mathbf{a}_j$ )
- Subtract components along  $\mathbf{a}_j$  from remaining columns
- Stop: chosen sentences (columns)  $\rightarrow \sim w$  words

- Removes semantic redundancy

# Summarizing – Example

- Query: “hurricane earthquake”
- Results:
  - Hurricane Gilbert (IR mean: 74)
  - Catastrophe Insurance (72)
  - California Earthquake (44)
- Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic. The hurricane, traveling westward across the Caribbean Sea, was upgraded Tuesday to a Category 5, the strongest and deadliest type of hurricane. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. Jamaican Prime Minister Edward Seaga said late Tuesday that at least six people were killed, and an estimated 60,000 were left homeless in “the worst natural disaster in the modern history of Jamaica”.

# Summarizing – Example

- Such increased demand for reinsurance, along with the losses the reinsurers will bear from these two disasters, are likely to spur increases in reinsurance prices that will later be translated into an overall price rise. For example, insurers may seek to limit their future exposure to catastrophes by increasing the amount of reinsurance they buy. Nationwide Insurance Co., a mutual company based in Columbus, Ohio, said Hugo "is the single largest claims disaster" it has seen in its 63-year history. Hugo could have a marginal impact on third-quarter income at Travelers Corp., Aetna Life & Casualty Insurance Co. and Chubb Corp., according to industry analysts.
- The 7.7-magnitude quake was the largest ever recorded in that area, where two major plates of the earth's crust meet, Needhams said. The shock wave traveled through the mountainous section of coastal Iran where most of the buildings are built on a flood plain of loosely deposited soil that shifts in an earthquake and allows structures to collapse, he said. The nearest metropolitan area to Tuesday's earthquake, San Jose, has seen nearly a dozen earthquakes of 5 magnitude or greater in the last 10 years, but several of them have been on the Calaveras fault, which generally runs just east of San Jose, more than 10 miles east of San Andreas at this point .

# Implementation

- Server (SunOS, Linux)
  - Q: GTP (U. Tennessee)
  - C: gmeans (U. Texas, Austin)
  - S: HMM+QR (U. Md., CCS)
- Server Interface (SunOS, Linux)
  - HTTP (Apache)
  - JavaServer (Tomcat)
  - Java-C++ (JNI)
- Client (Any)
  - Java Servlet – Dynamic HTML

# Demo

The logo consists of the letters 'QCS' in a bold, green, sans-serif font. The 'Q' is stylized with a small tail that curves downwards and to the right. The 'C' and 'S' are also bold and rounded.

# Future Directions

- HLT-NAACL 2003 (Edmonton, Canada)
  - Accepted demonstration
- Parameter estimation
- Optimize code
  - Persistent data/variables
  - Document parsing
- Interface development
  - Easier access to parameters

# Conclusions

- Benefits of QCS

- Q: Better than exact matching
- C: Organizes retrieval results → topic clusters
- S: One summary per cluster, reduces redundancy

- Scientific Literature Searches

- QCS → 😊

<http://stiefel.cs.umd.edu:8080/qcs>

# References

- Q M.W. Berry, S. T. Dumais, and G. W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Rev.*, 37(4):573-595, 1995.
- Q M.W. Berry and D.I. Martin. Parallel SVD for Scalable Information Retrieval. *Elsevier Preprint*. 2000
- Q T.G. Kolda and D.P. O'Leary. A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. *ACM Trans. Info. Sys.*, 16(4):322-346, 1998.
- C I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143-175, Jan 2001.
- S J. M. Conroy and D. P. O'Leary. Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition. Technical report, University of Maryland, College Park, Maryland, March, 2001.
- S J. D. Schlesinger, M. E. Okurowski, J. M. Conroy, D. P. O'Leary, A. Taylor, J. Hobbs, and W. H. T. Wilson. Understanding Machine Performance in the Context of Human Performance for Multi-document Summarization. In *Proceedings of the Workshop on Automatic Summarization*, 2002.