

An Information Retrieval System for Improving Efficiency in Scientific Literature Searches

Danny Dunlavy

Scientific Research

- First step: literature search
 - Journal articles, tech. reports, preprints
- Information retrieval engines
 - Google: WWW search engine
 - 2.5 billion HTML documents
 - arXiv: preprint server for physics
 - 135,000 preprints
- Need good retrieval methods

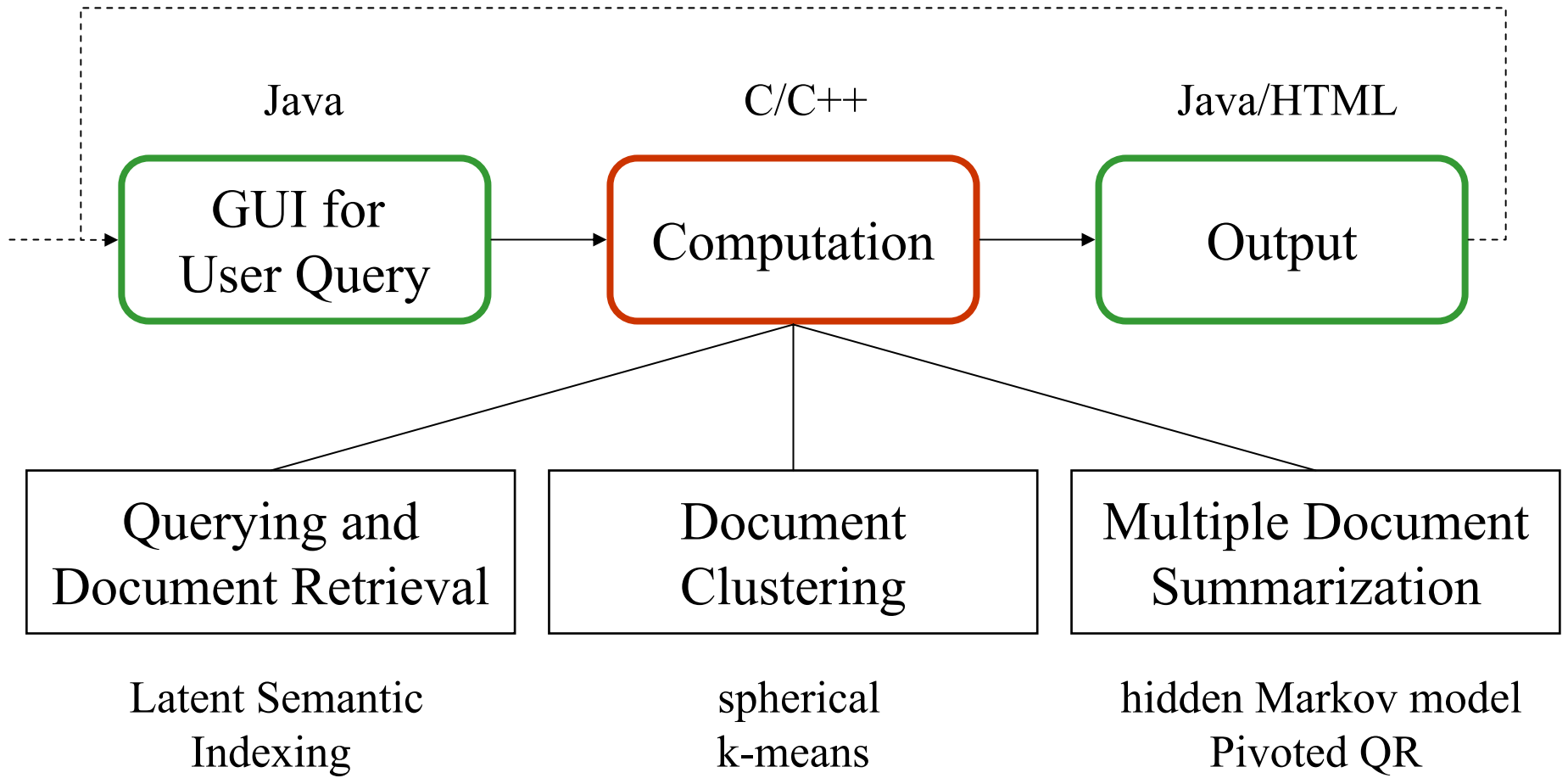
Example

- Query
 - “methods plasma physics”
- Retrieval
 - Google: 27,000 documents
 - arXiv: 350 documents
- Solutions
 - Better query
 - Better retrieval system

Project Proposal

- New information retrieval system
 - retrieve documents relevant to a query
 - separate documents into topic clusters
 - create summary for each topic cluster
- QCS (query, cluster, summarize)

Implementation



Implementation (cont.)

- *Software*

Q: Parallel General Text Parser (*PGTP*)

C: Generalized Spherical K-means (*gmeans*)

S: GNU Hidden Markov Model Library (*ghmm*)

Linear Algebra Package (*LAPACK*)

MPICH, CVS

- *Hardware*

UMIACS PC Linux Cluster (16 nodes)

Validation

- Document Understanding Conferences
 - document set summaries
- Testing
 - Build query from documents sets
 - Run QCS
 - Match output to DUC summaries

Links

- **Project Web Page**

<http://www.math.umd.edu/~ddunlavy/amsc663/index.html>

- **Project Proposal**

<http://www.math.umd.edu/~ddunlavy/amsc663/proposal.pdf>

<http://www.math.umd.edu/~ddunlavy/amsc663/proposal.ps>

- **Project Proposal Slides**

<http://www.math.umd.edu/~ddunlavy/amsc663/proposal-slides.ppt>

<http://www.math.umd.edu/~ddunlavy/amsc663/proposal-slides.pdf>

<http://www.math.umd.edu/~ddunlavy/amsc663/proposal-slides.ps>

References

1. N. Aluthgedara. Recognizing Sentence Boundaries and Boilerplate. In preparation.
2. 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.
3. 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.
4. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391 {407, 1990.
5. I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143 {175, Jan 2001.
6. 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. Volume 2: Draft DUC Papers.