Wright State University

CORE Scholar

Kno.e.sis Publications

The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis)

7-2005

# A Modular Approach to Document Indexing and Semantic Search

Dhanya Ravishankar

Krishnaprasad Thirunarayan
*Wright State University - Main Campus*, t.k.prasad@wright.edu

Trivikram Immaneni

Follow this and additional works at: https://corescholar.libraries.wright.edu/knoesis

 Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

## Repository Citation

Ravishankar, D., Thirunarayan, K., & Immaneni, T. (2005). A Modular Approach to Document Indexing and Semantic Search. .
https://corescholar.libraries.wright.edu/knoesis/882

This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.


# A Modular Approach to Document Indexing and Semantic Search

Dhanya Ravishankar, Trivikram Immaneni
Krishnaprasad Thirunarayan
Department of Computer Science & Engineering
Wright State University
Dayton, OH-45435, USA

# Talk Outline

- Goal (*What?*)
- Background and Motivation (*Why?*)
- Implementation Details (*How?*)
- Evaluation and Applications (*Why?*)
- Conclusions

WRIGHT STATE UNIVERSITY

# Goal

- Develop a modular approach to improving effectiveness of searching documents for information
- Reuse and integrate mature software components

# Background and Motivation

- Improve recall using information implicit in the English language
- Improve precision and recall using domain-specific information implicit in the document collection
- Assist manual content extraction by mapping document phrases to controlled vocabulary terms (domain library)
  - NSF-SBIR Phases I and II with Cohesia Corp.

- Enable extensions
  - Spell check input query
  - Organize search results through grouping
    - Improve precision thro sense-disambiguation
- Enable experimentation
  - Investigate empirical relationship between significant eigenvalues in the Singular Value Decomposition (SVD) and the number of document clusters using benchmarks.

# Implementation Details (*How?*)

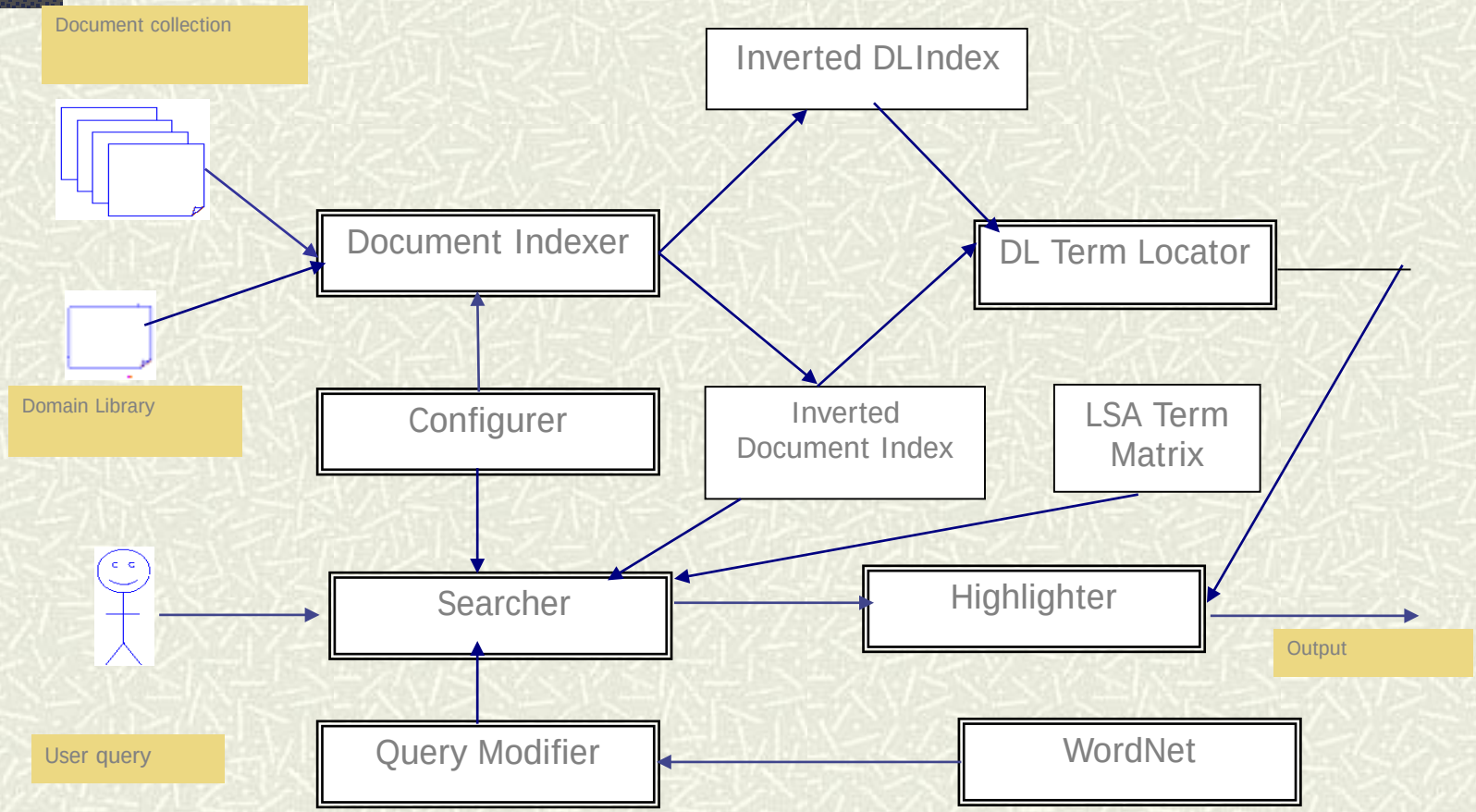
# Tools Used

- Apache's Lucene APIs
  - A high-performance, Java text search engine library with smart indexing strategies.
- WordNet and Java WordNet Library
- NIST and MathWork's Java Matrix package (JAMA) for LSI
- Domain-specific controlled vocabulary for Materials and Process Specs

- Jazzy, a Java Open Source Spell-Checker

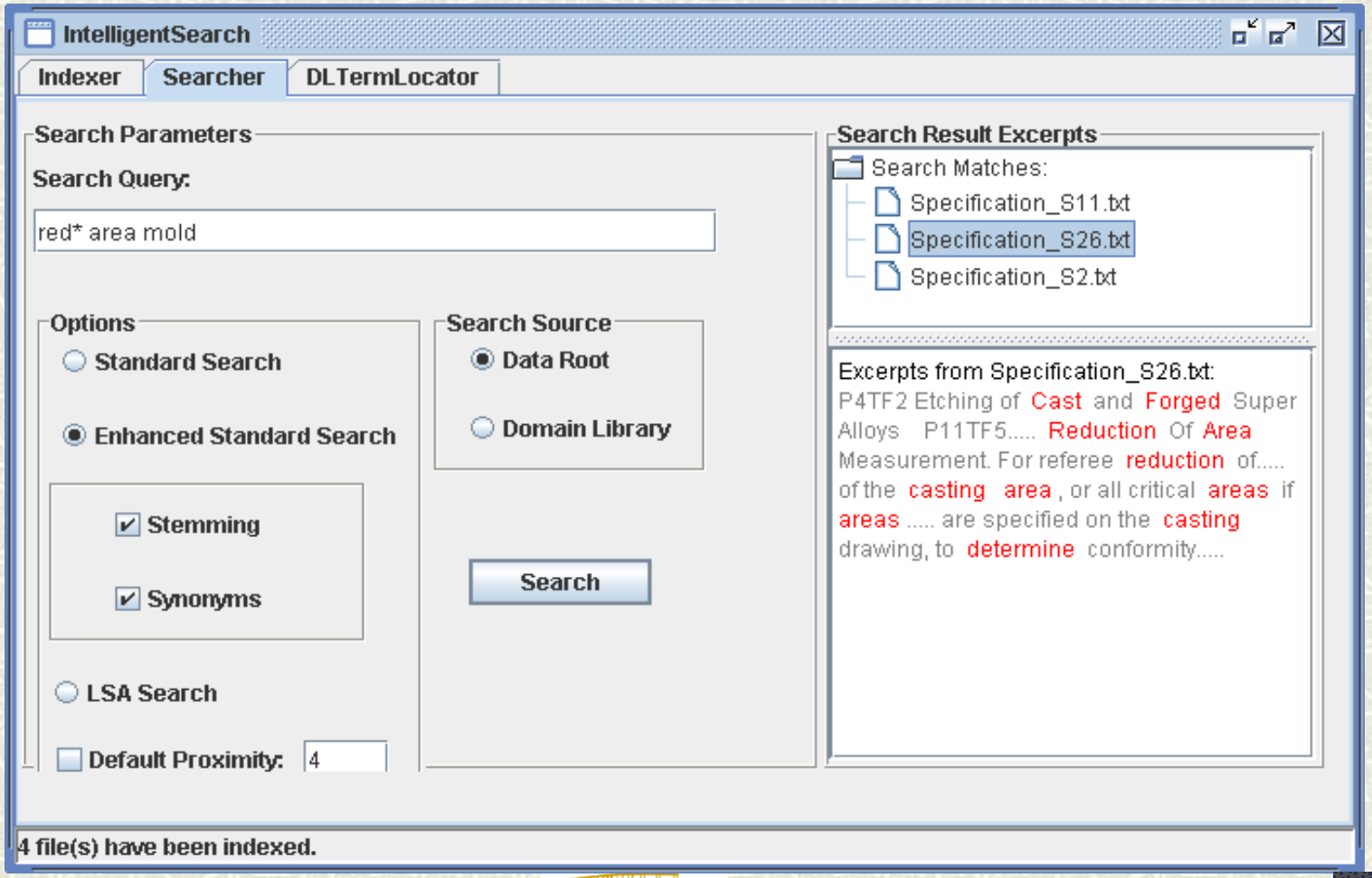
- MEDLINE dataset
- 20-Newsgroups dataset
- Reuters-215781 newswire stories datasets

# Architecture of Content-based Indexing and Semantic Search Engine



# Evaluation and Application (*Why?*)

# Enhanced search illustrating wildcard pattern and synonym expansion



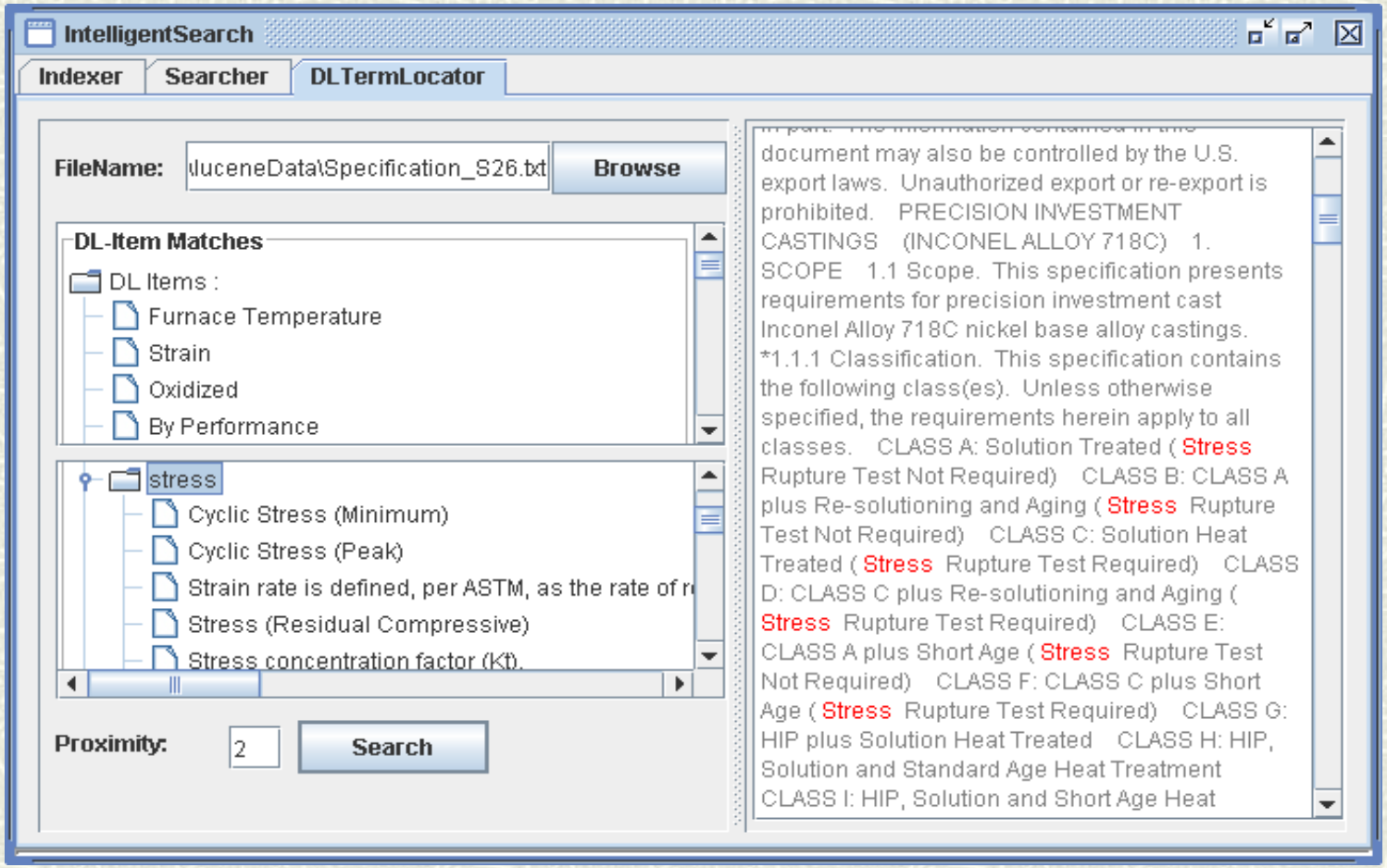
# More examples

- Syntactic variations
  - test certificate ≈ certificate of test ≈ test certification
- Semantic invariance
  - tensile strength ≈ ductile force
  - part number ≈ part and lot number
  - insufficient immunity ≈ immune deficiency
  - causes cancer ≈ induces cancer ≈ reasons for cancer

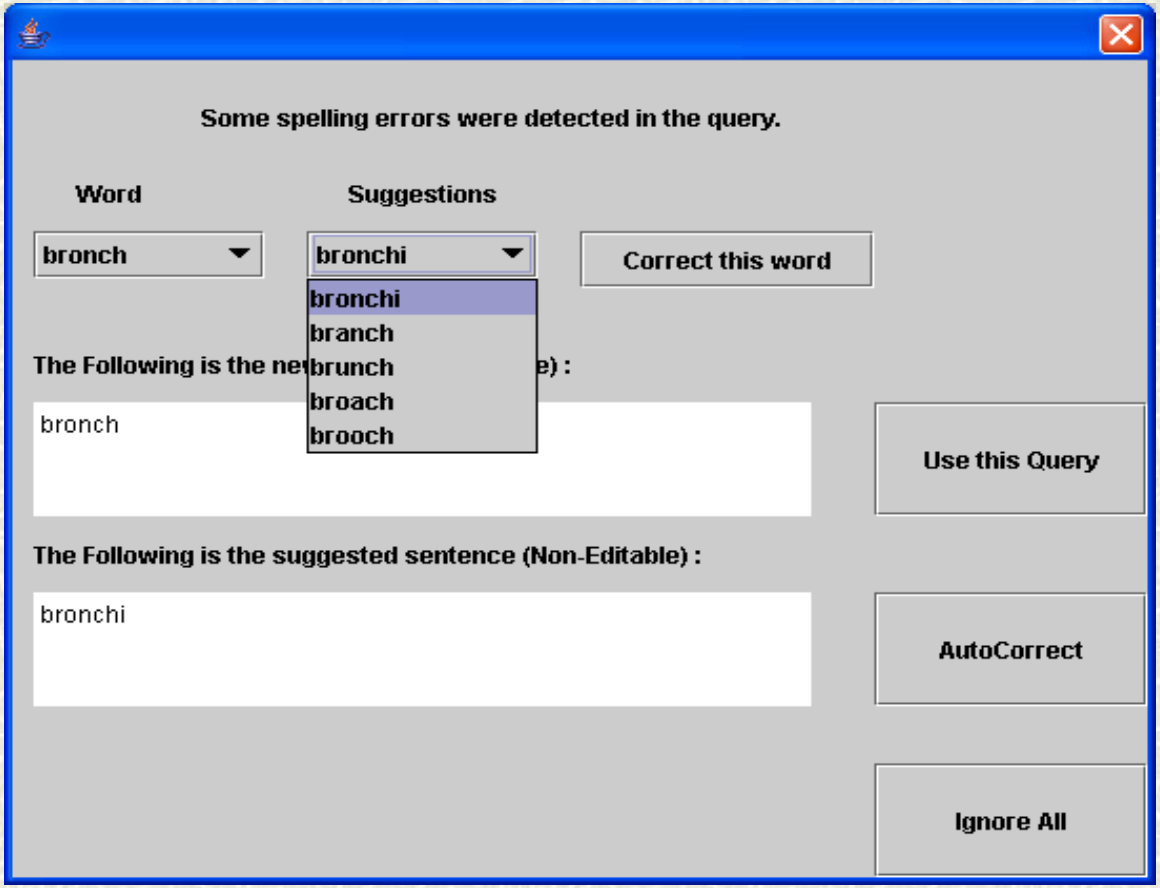
# Recall and Precision on MEDLINE collection with Different Search Strategies

| Query | Enhanced Search | | LSA Search | |
|---|---|---|---|---|
| | **Recall** | **Precision** | **Recall** | **Precision** |
| "electron microscopy of lung or bronchi" | 0.86 | 0.2 | 0.91 | 0.5 |
| "the crossing of fatty acids through the placental barrier. normal fatty acid levels in placenta and fetus" | 0.96 | 0.08 | 0.85 | 0.63 |
| "the use of induced hypothermia in heart surgery, neurosurgery, head injuries and infectious diseases." | 0.96 | 0.07 | 0.82 | 0.3 |
| "bacillus subtilis phages and genetics, with particular reference to transduction." | 1.0 | 0.12 | **0.95** | **0.83** |

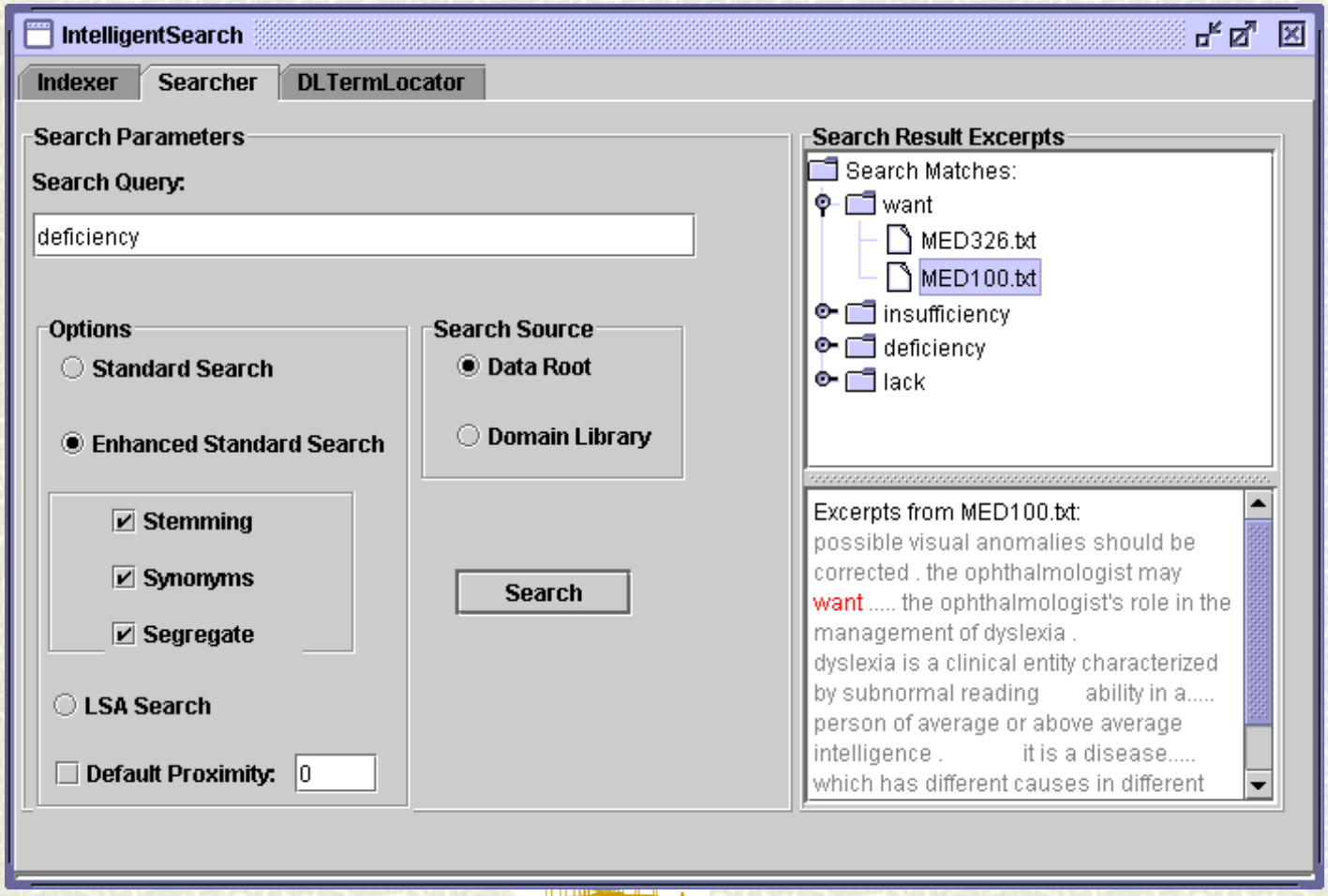
# Matching DL Items; DL Term and its location in the document



# Spell-checking input dialog



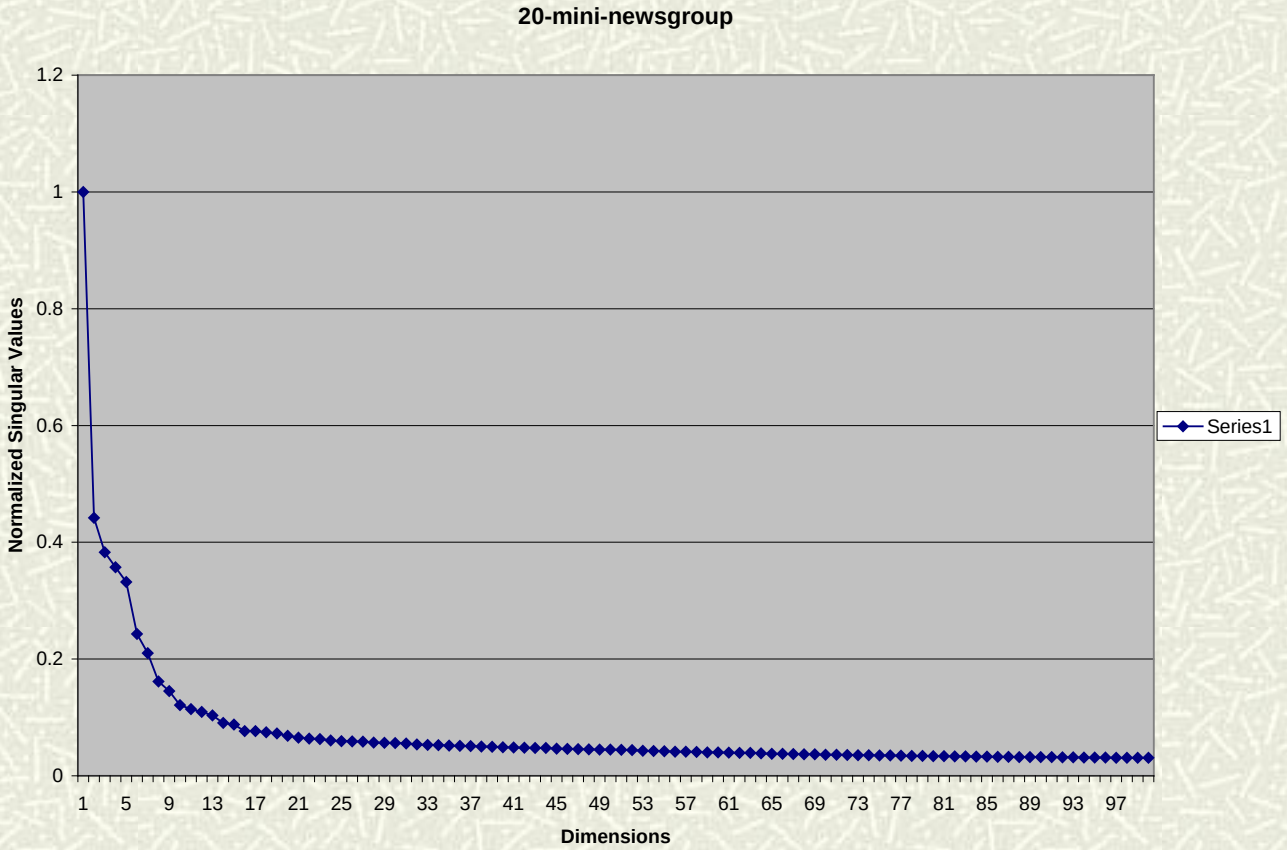
# Grouping retrieved results



# LSI and Clustering

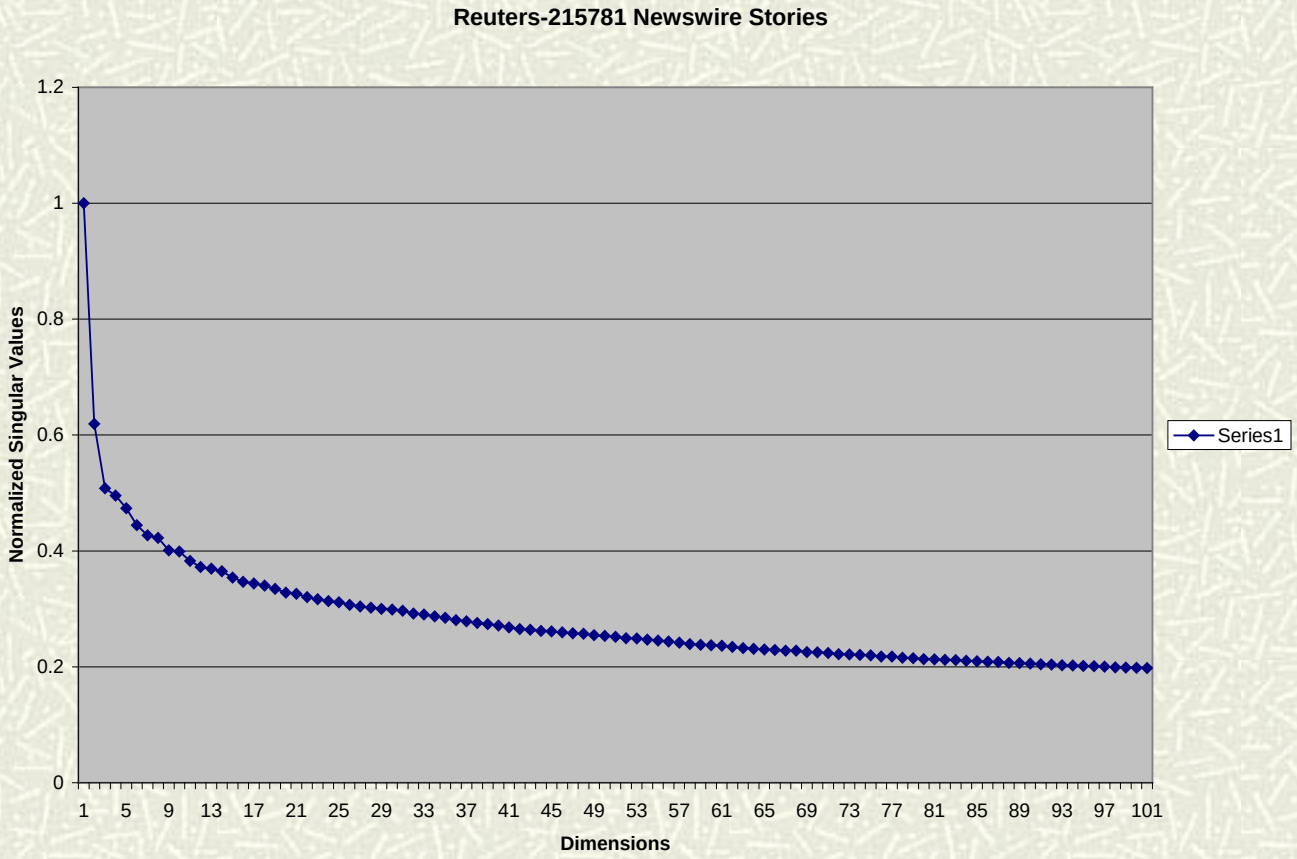
- Exploring relationship between the number of significant eigenvalues and the number of document clusters
  - 20-Mini-Newsgroup dataset
    - 2000 postings, 20 groups
  - Reuters-215781 Newswire Stories dataset
    - Used 2000 stories at a time, 70 topics

# 20-Mini-Newsgroup dataset results (eigen value reduction = 1/7)




WRIGHT STATE UNIVERSITY

# Reuters-21578 newswire dataset results (eigenvalue reduction = 1/5)



# Conclusions

- Search flexible and effective
  - In future, incorporate domain-specific context for word-sense disambiguation
- LSI is memory and CPU intensive, and could not run with full datasets (only 2K docs used) on a 2.53 GHz, 1GB m/c
  - In future, run on more powerful server machine

- Useful assistance for manual content extraction from materials and process specs, given the controlled vocabulary
- In future, this framework / infrastructure usable for experiments with expressive, context-aware, and scalable search.