Defense Acquisition in Transition

6TH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

Semantic Search

Craig Martell

Assoc. Professor, Computer Science, Naval Postgraduate School

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Semantic Search	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Department of Computer Science,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
NPS?s 6th Annual Acquisition Research Symposium, Monterey CA, 13-14 May 2009

14. ABSTRACT

15. SUBJECT TERMS

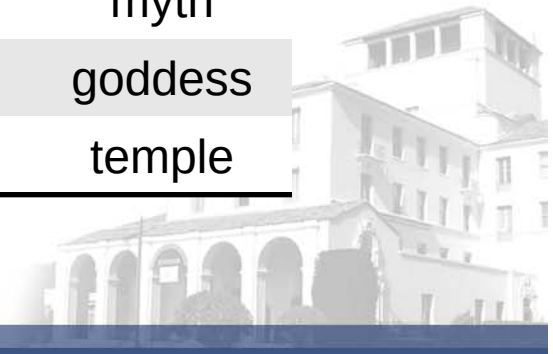
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

Semantic Search

- Don't just search on keywords, but on a model of what the words *mean*.
- *Meaning* is still very difficult to ascertain, but we can estimate it the co-occurrence of other words
- Correlations can be discovered by probabilistic topic models like Latent Dirichlet Allocation or Hierarchical Dirichlet Processes.
- This allows us to find documents that are relevant to the query, even if they do not share keywords with the query.

Example Wikipedia "Topics," Automatically Detected

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
ball	treatment	software	god	greek
play	medical	computer	christian	zeus
team	acupuncture	hardware	church	mythology
player	disease	video	jesus	gods
football	pain	disk	christianity	god
line	studies	computers	believe	son
offensive	evidence	memory	book	aeneas
defensive	effects	bit	christ	myth
pass	found	operating	holy	goddess
field	patients	screen	faith	temple



Our Experiments

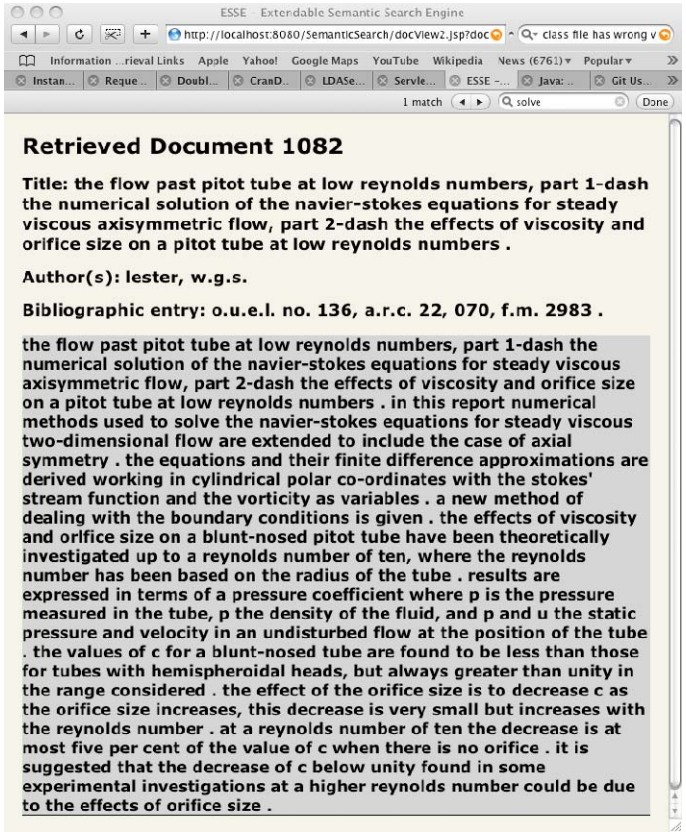
- We tested on a standard benchmark data set from the TREC challenge.
- TREC (Text REtrieval Conference) is a yearly competition sponsored by NIST. They provide standardized datasets for information-retrieval research
- We used the Cranfield dataset for these experiments, since it deals with abstracts of technical documents
- Our target is requirements documents/abstracts).

Simple Example

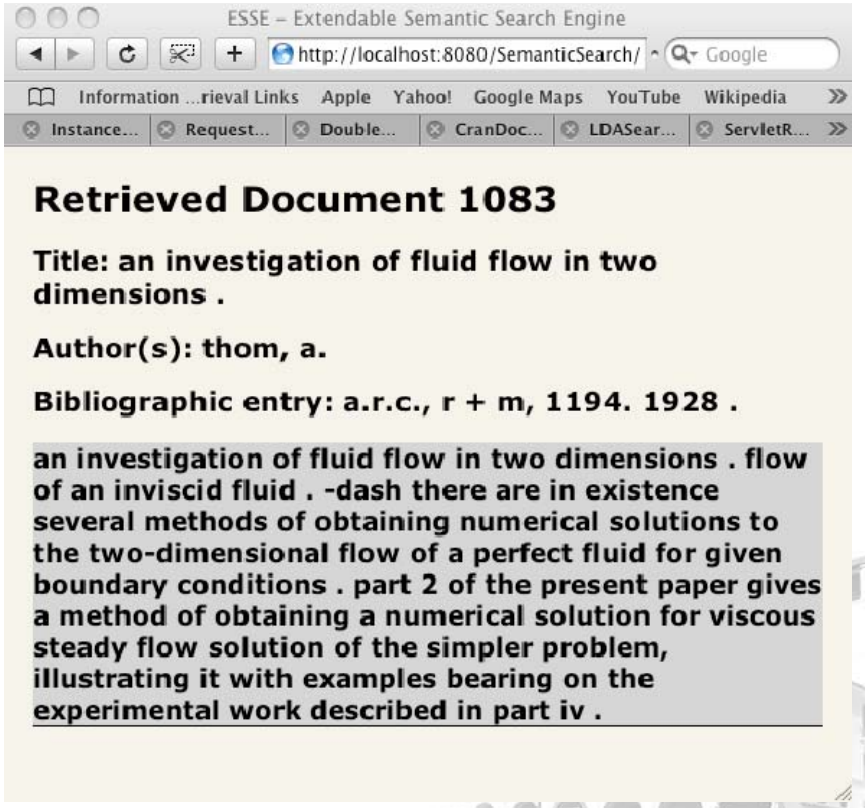
- Query 226 from CRANFIELD:
  - "how should the navier-stokes difference equations be solved"
- Relevant documents are 1063, 1078, 1080-1085, a total of 8 documents.
- A pure key-word based model has an average precision of 37% on this model.
- A pure LDA topic model using 10 Markov chains for sampling gives us 41%
- A combination of both gives us 46% precision.
  - This may seem low, but what it says is that 46% of the results are relevant to the user.
- All of the relevant documents are retrieved by the 22nd document.

Two of the Documents found

A typical keyword result



A typical LDA result



A Closer Look

Document 1082 (found by keyword search)

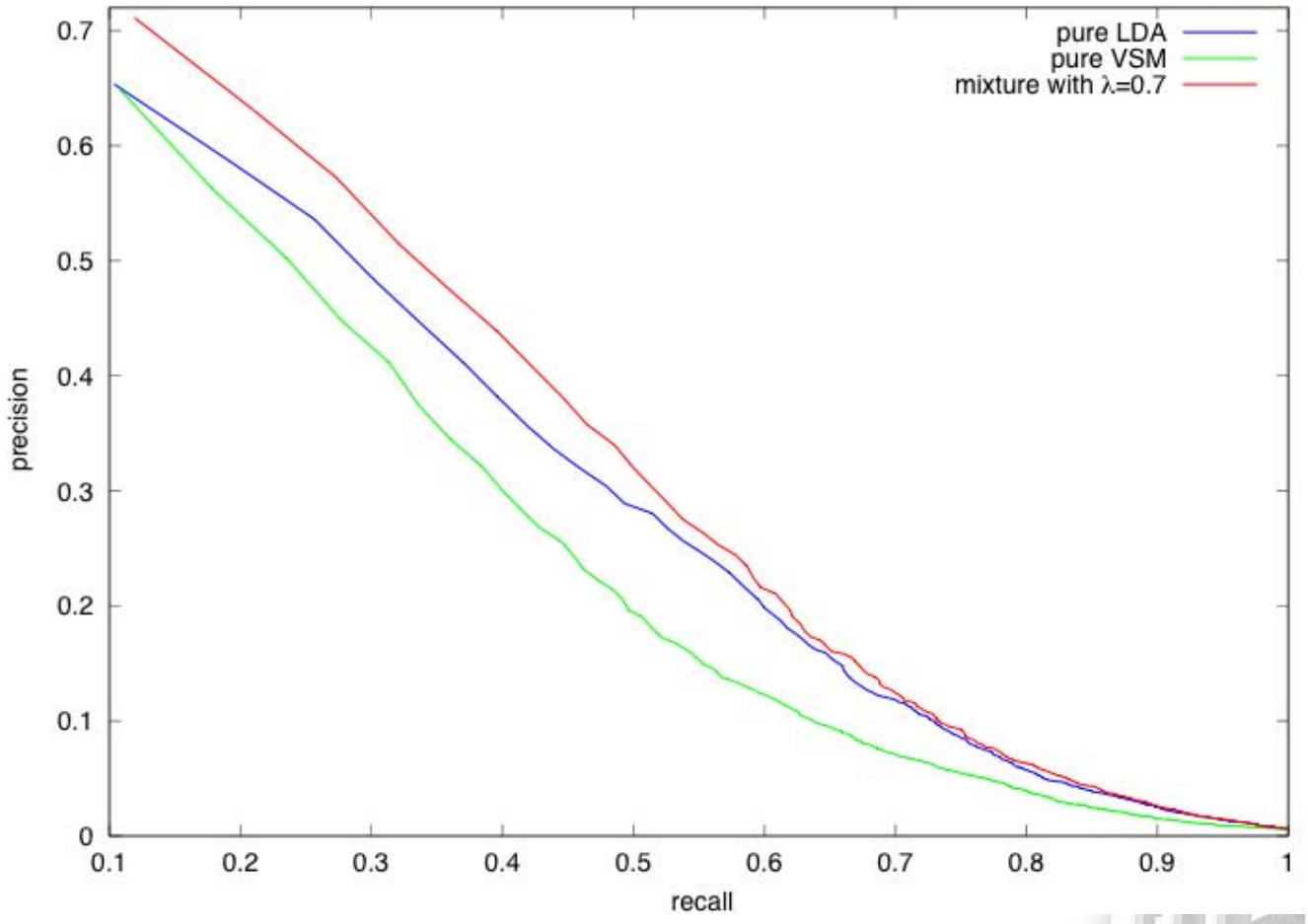
- "… **report numerical methods used to solve the navier-stokes equations for steady viscous two-dimensional flow are extended …**"
- This document is ranked high in a keyword search model.

Document 1083 (not found by keyword search, but by LDA)

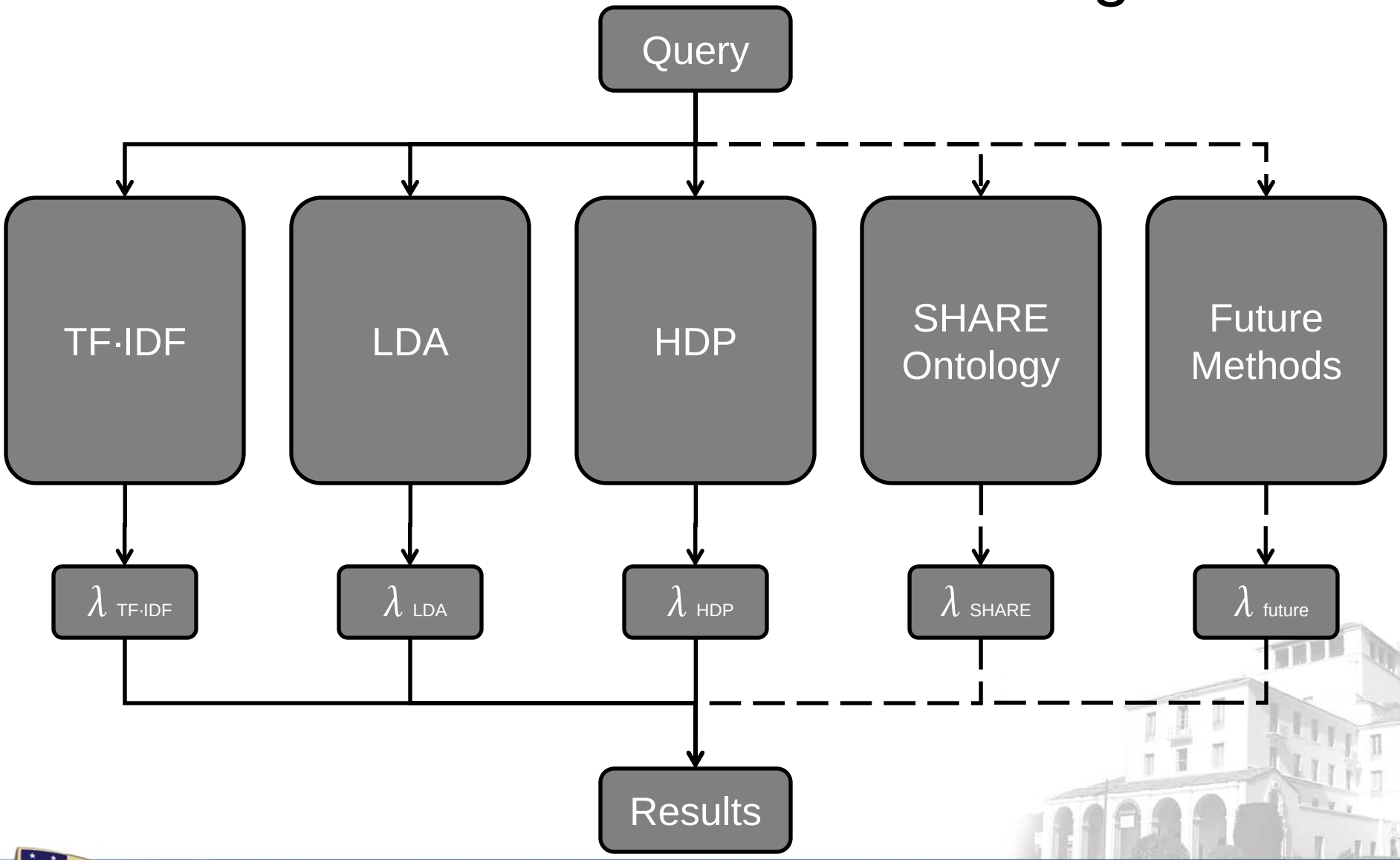
- "… **numerical solutions to the two-dimensional flow of a perfect fluid for given boundary conditions** …"
- This document has very low rank in a keyword search, but is very relevant to our query.
- These words are in the same "topic" as the query, even if they don't *match.*

Remember the query: "how should the navier-stokes difference equations be solved"

The Power of Combination



Modular Semantic Search Engine



Thank You
Questions?

Craig Martell
cmartell@nps.edu