



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

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE MAY 2009		2. REPORT TYPE		3. DATES COVERED 00-00-2009 to 00-00-2009	
4. TITLE AND SUBTITLE Semantic Search				5a. CONTRACT NUMBER	
				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 Same as Report (SAR)	18. NUMBER OF PAGES 10	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

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).



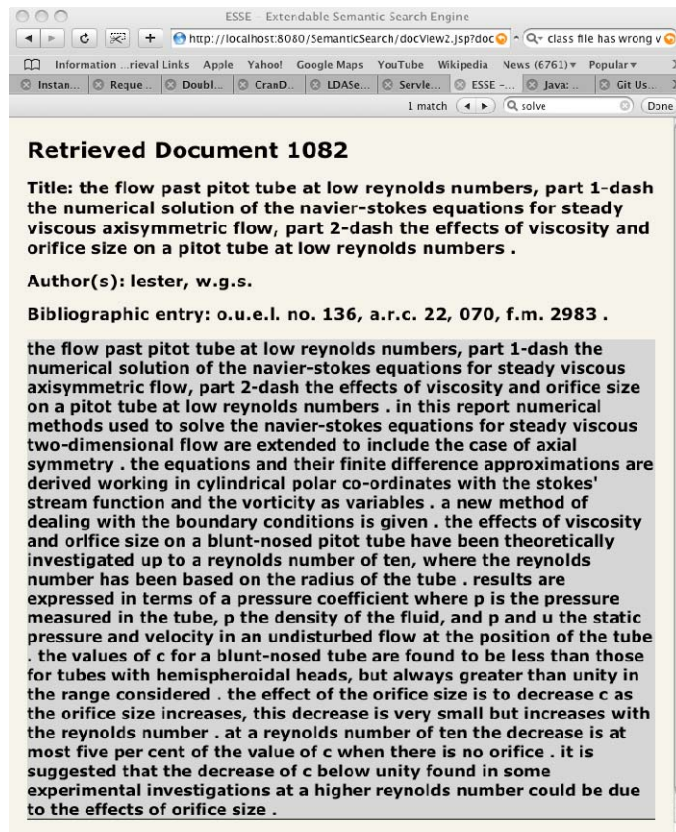
Simple Example

- 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



ESSE - Extendable Semantic Search Engine

http://localhost:8080/SemanticSearch/docView2.jsp?doc=... class file has wrong v

Information ...rieval Links Apple Yahoo! Google Maps YouTube Wikipedia News (6761) Popular >>

Instance... Reque... Doubl... CranD... LDASe... Servle... ESSE ... Java: ... Git Us... >>

1 match <<> solve (Done)

Retrieved Document 1082

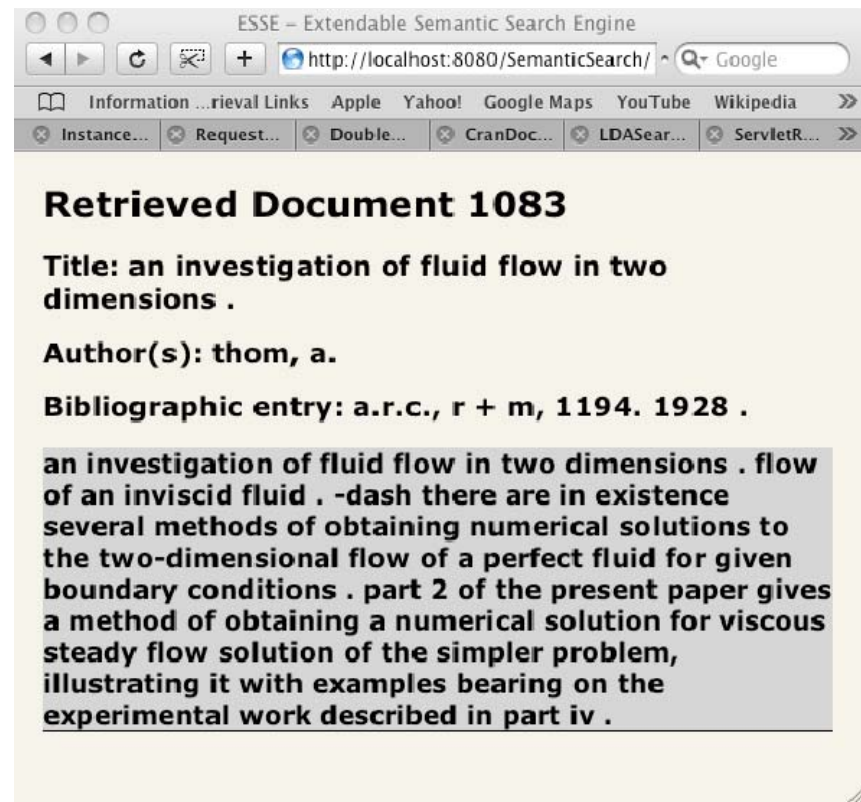
Title: the flow past pitot tube at low reynolds numbers, part 1-dash the numerical solution of the navier-stokes equations for steady viscous axisymmetric flow, part 2-dash the effects of viscosity and orifice size on a pitot tube at low reynolds numbers .

Author(s): lester, w.g.s.

Bibliographic entry: o.u.e.l. no. 136, a.r.c. 22, 070, f.m. 2983 .

the flow past pitot tube at low reynolds numbers, part 1-dash the numerical solution of the navier-stokes equations for steady viscous axisymmetric flow, part 2-dash the effects of viscosity and orifice size on a pitot tube at low reynolds numbers . in this report numerical methods used to solve the navier-stokes equations for steady viscous two-dimensional flow are extended to include the case of axial symmetry . the equations and their finite difference approximations are derived working in cylindrical polar co-ordinates with the stokes' stream function and the vorticity as variables . a new method of dealing with the boundary conditions is given . the effects of viscosity and orifice size on a blunt-nosed pitot tube have been theoretically investigated up to a reynolds number of ten, where the reynolds number has been based on the radius of the tube . results are expressed in terms of a pressure coefficient where p is the pressure measured in the tube, ρ the density of the fluid, and p and u the static pressure and velocity in an undisturbed flow at the position of the tube . the values of c for a blunt-nosed tube are found to be less than those for tubes with hemispheroidal heads, but always greater than unity in the range considered . the effect of the orifice size is to decrease c as the orifice size increases, this decrease is very small but increases with the reynolds number . at a reynolds number of ten the decrease is at most five per cent of the value of c when there is no orifice . it is suggested that the decrease of c below unity found in some experimental investigations at a higher reynolds number could be due to the effects of orifice size .

A typical LDA result



ESSE - Extendable Semantic Search Engine

http://localhost:8080/SemanticSearch/ Google

Information ...rieval Links Apple Yahoo! Google Maps YouTube Wikipedia >>

Instance... Request... Double... CranDoc... LDASe... ServletR... >>

Retrieved Document 1083

Title: an investigation of fluid flow in two dimensions .

Author(s): thom, a.

Bibliographic entry: a.r.c., r + m, 1194. 1928 .

an investigation of fluid flow in two dimensions . flow of an inviscid fluid . -dash there are in existence several methods of obtaining numerical solutions to the two-dimensional flow of a perfect fluid for given boundary conditions . part 2 of the present paper gives a method of obtaining a numerical solution for viscous steady flow solution of the simpler problem, illustrating it with examples bearing on the experimental work described in part iv .



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.

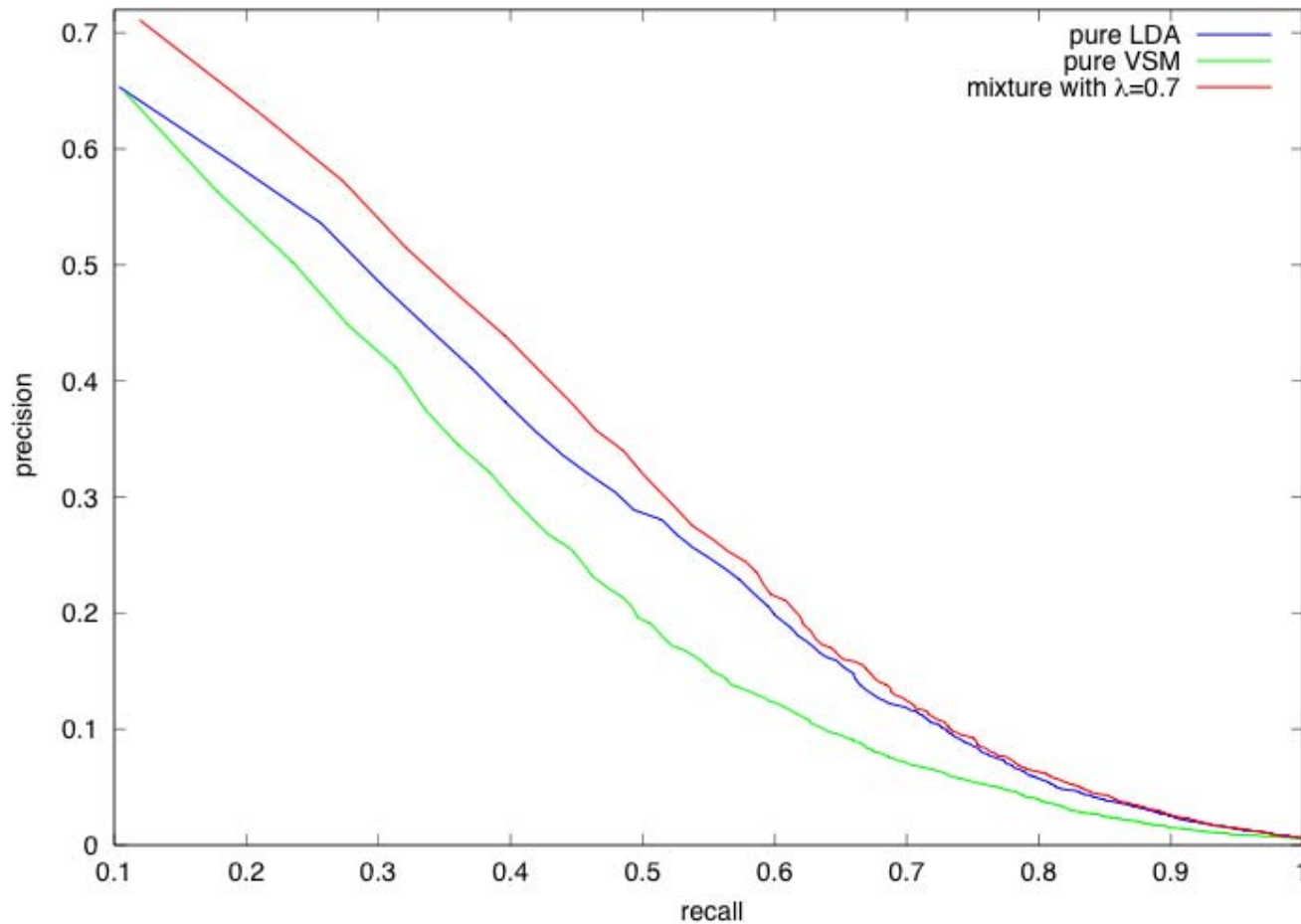
Document 1083 (not found by keyword search, but by LDA)

- “... 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*.

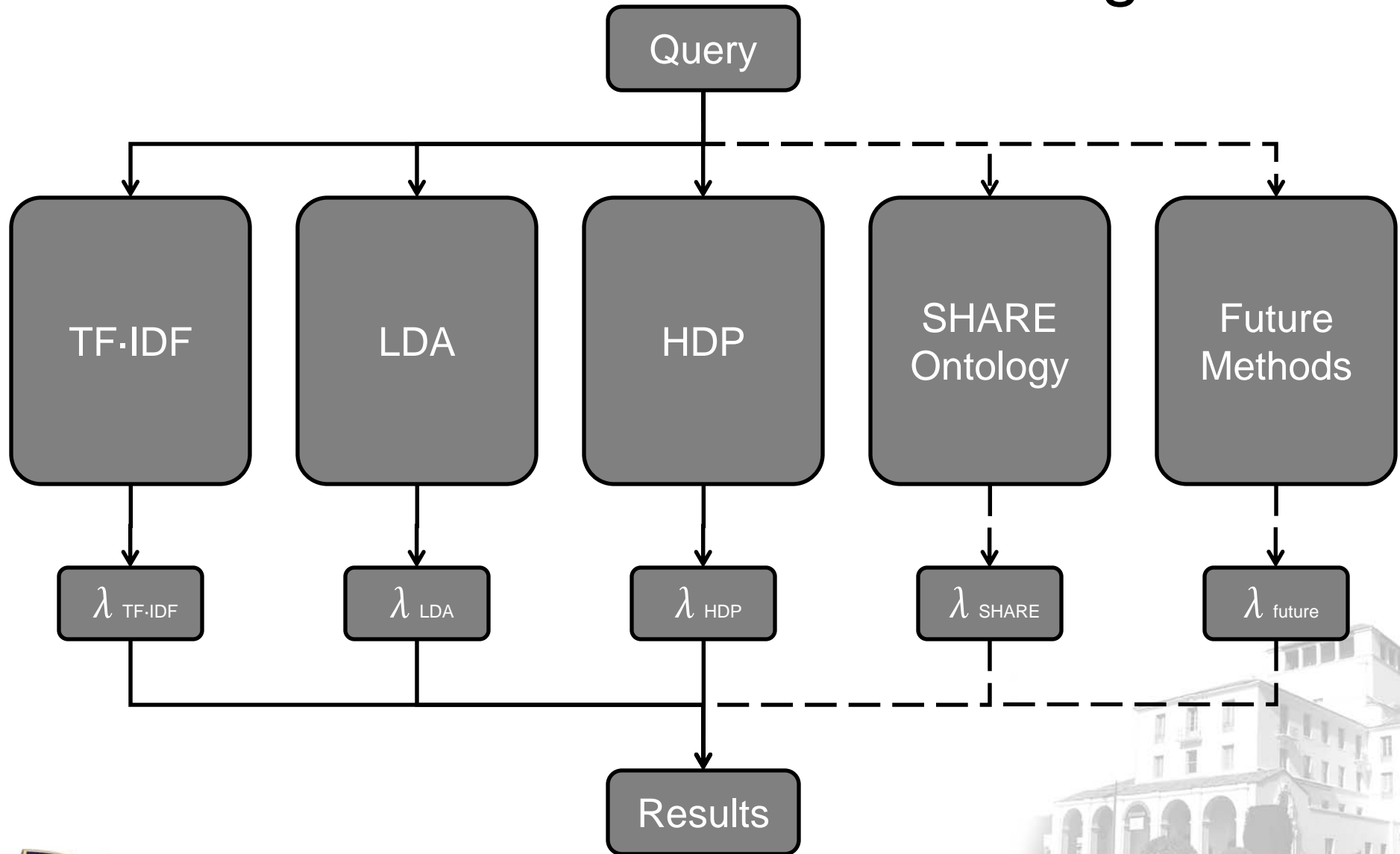
Remember the query: “how should the navier-stokes difference equations be solved”



The Power of Combination



Modular Semantic Search Engine



Thank You Questions?

Craig Martell
cmartell@nps.edu



Defense Acquisition in Transition
6TH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

May 12-14, 2009
Monterey, CA