

Bellcore

 Bell Communications Research

Using Latent Semantic Indexing (LSI) for Information Retrieval, Information Filtering and Other Things

***Susan T. Dumais, Bellcore
Cognitive Technology Conference
April 4, 1997***

Copyright © 1995, Bellcore
All Rights Reserved

Topic Outline

- **The Problem: Retrieving Information from External Sources**
- **A Solution: Latent Semantic Indexing (LSI)**
- **Several Applications**
 - **Information Retrieval**
 - **Information Filtering**
 - **Automatic Assignment of Reviewers**
 - **Bellcore Advisor**
- **Conclusions**

Retrieval of Information from External Sources

- **The Promise: External databases can greatly augment human memory and problem solving**
- **The Reality: It is surprisingly difficult to find information in external databases**
 - **Keyword-based retrieval systems**
 - **“Vocabulary mismatch”**
 - **Implications for retrieval**
 - **Retrieve irrelevant information (50% or more)**
 - **Miss relevant information (routinely 80%)**
 - **Need to capture and exploit structure**

Latent Semantic Indexing (LSI)

Overview

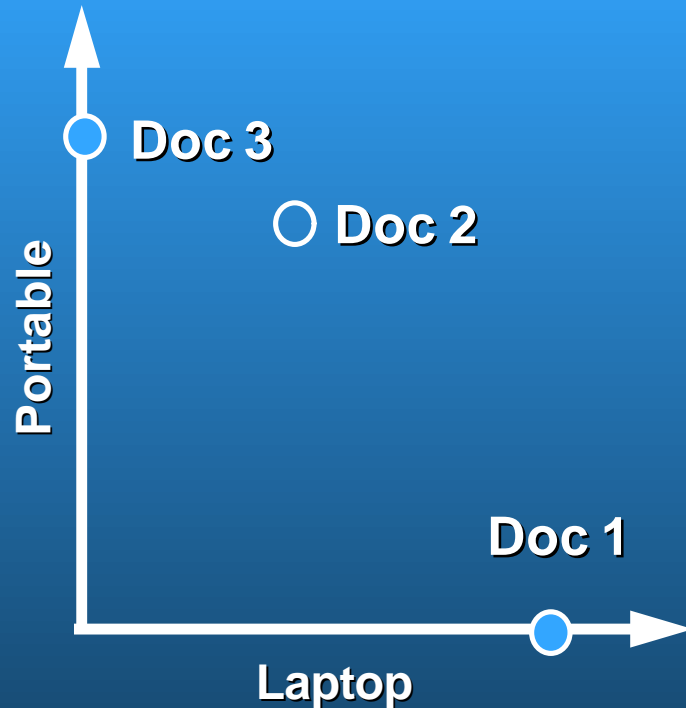
- **Begin with standard term-document matrix**
- **Assume underlying or latent structure in matrix**
- **Use truncated SVD to model latent semantic structure**
- **Use resulting semantic space for retrieval (k~300)**
 - **can retrieve documents that share no words with query**
- **Fully automatic analysis**

- **Geometric representation**

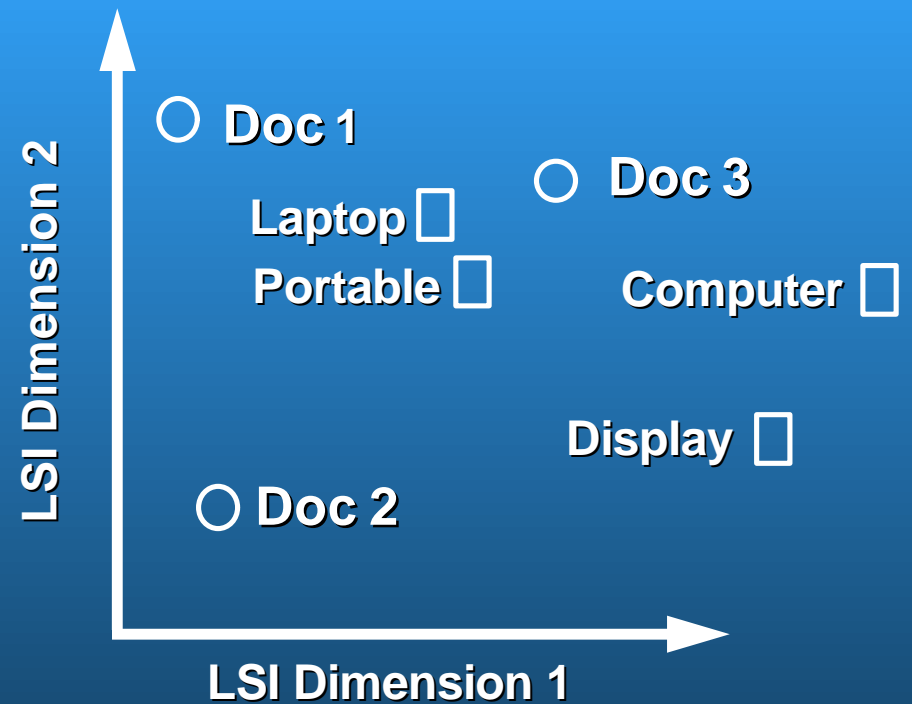
Latent Semantic Indexing (LSI)

Keyword vs. LSI Retrieval

Keyword Retrieval:
Words Unrelated



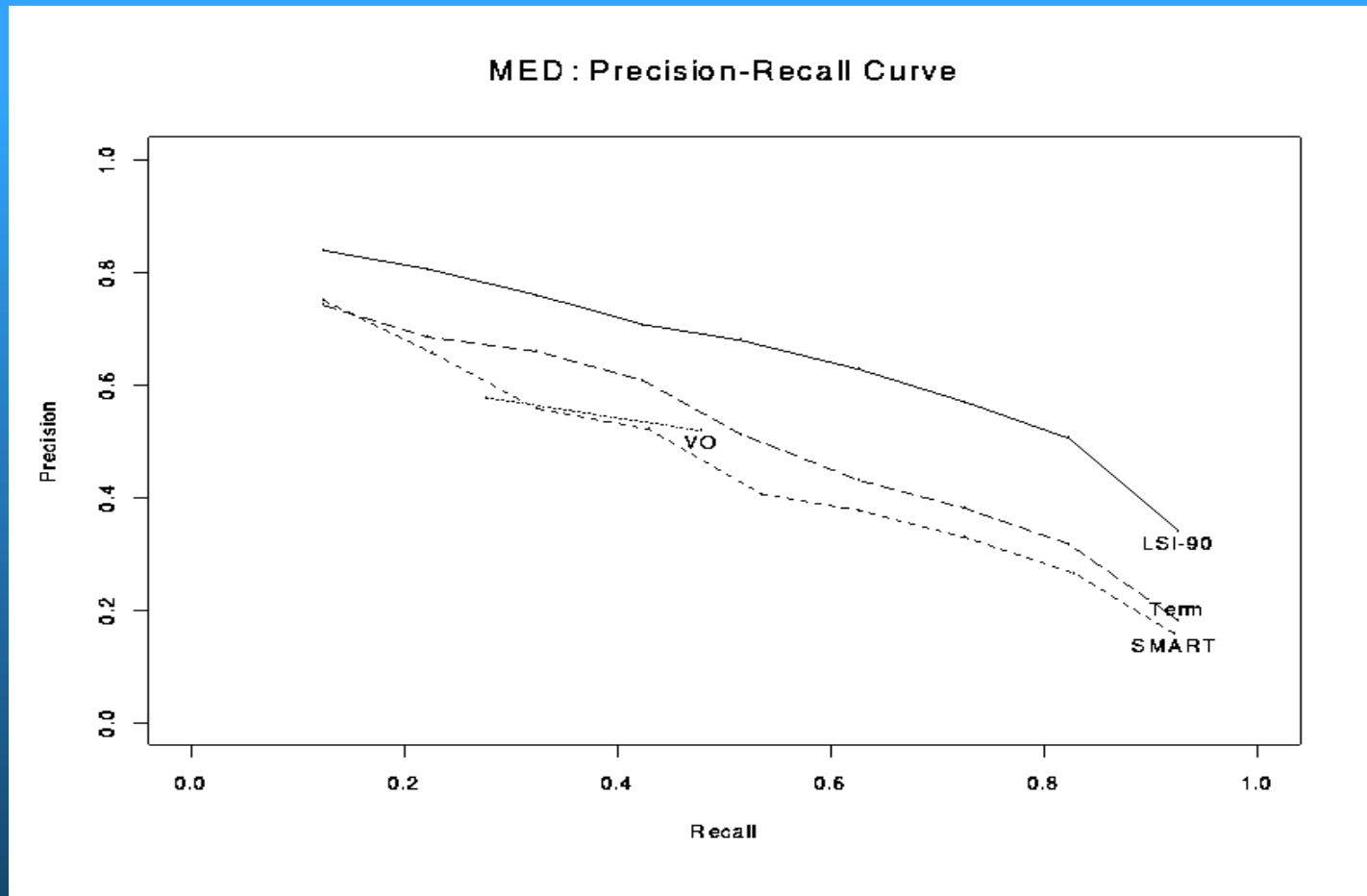
LSI Retrieval:
Similar words associated



Evaluations of LSI for Information Retrieval

- Information Science Test Collections
 - Text objects (“documents”)
 - Queries and relevance judgments
 - Evaluation
 - Precision: $\# \text{ relevant retrieved} / \# \text{ retrieved}$
 - Recall: $\# \text{ relevant retrieved} / \text{total } \# \text{ relevant}$
 - E. G., “Med”
 - 1033 medical abstracts; 5831 terms
 - SVD takes < 2 minutes on Sparc 10
 - For each query, rank abstracts
 - Plot precision/recall curve

LSI Evaluations Med Results

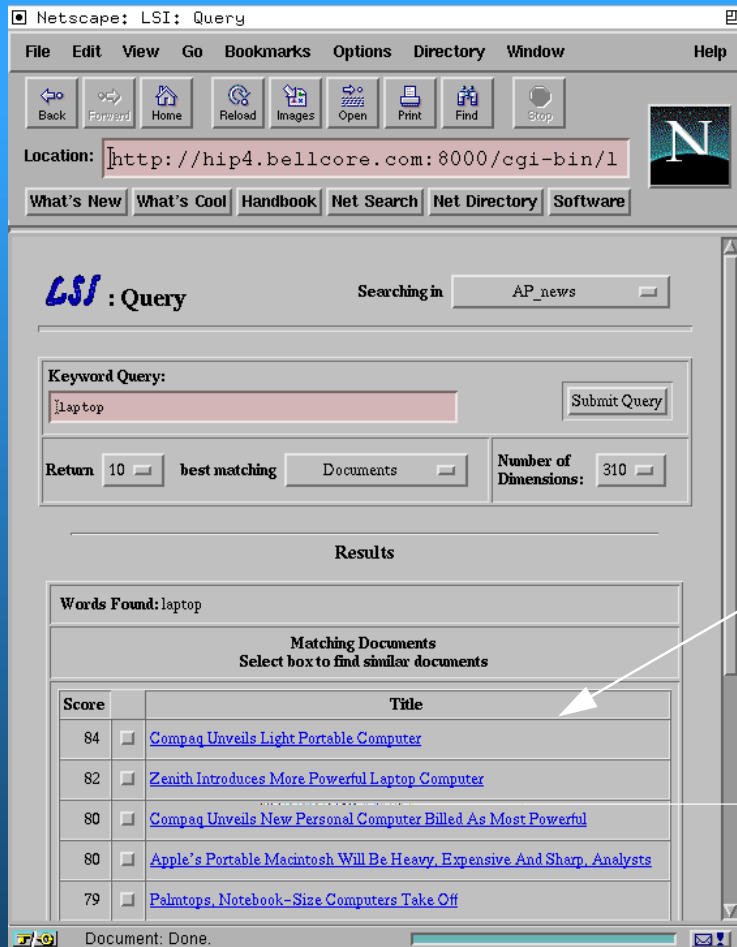


LSI Evaluations Summary Results

Test Collection	LSI	keyword	measure
Med-e	.66	.51	average precision
Med	.52	.46	average precision
Cran	.39	.29	average precision
ADI	.29	.26	average precision
Cisi	.11	.11	average precision
News	.61	.55	average precision
TM	.40	.35	average precision
TREC	.30	.26	average precision
TREC	8676	8043	number relevant
Toefl	53.5	29.5	number correct

Latent Semantic Indexing (LSI)

Best matching document does not contain "laptop"



Netscape: LSI: Query

File Edit View Go Bookmarks Options Directory Window Help

Location:

What's New What's Cool Handbook Net Search Net Directory Software

LSI : Query Searching in

Keyword Query:

Return Number of Dimensions:

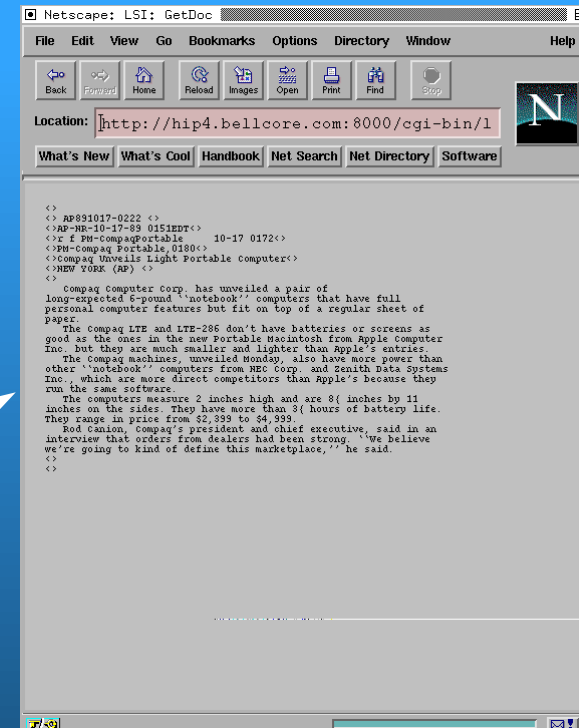
Results

Words Found: laptop

Matching Documents
Select box to find similar documents

Score	Title
84	<input type="checkbox"/> Compaq Unveils Light Portable Computer
82	<input type="checkbox"/> Zenith Introduces More Powerful Laptop Computer
80	<input type="checkbox"/> Compaq Unveils New Personal Computer Billed As Most Powerful
80	<input type="checkbox"/> Apple's Portable Macintosh Will Be Heavy, Expensive And Sharp, Analysts
79	<input type="checkbox"/> Palmtops, Notebook-Size Computers Take Off

Document: Done.



Netscape: LSI: GetDoc

File Edit View Go Bookmarks Options Directory Window Help

Location:

What's New What's Cool Handbook Net Search Net Directory Software

```
<> AP881017-0222 <>
<> CMP-MS-10-17-88 0151EDR<>
<> r f PM-Compaqportable 10-17 0172<>
<> PM-Compaq Portable, 0180<>
<> Compaq Unveils Light Portable Computer<>
<> NEW YORK (AP) <>
<>
Compaq computer corp. has unveiled a pair of
long-expected 6-pound "notebook" computers that have full
personal computer features but fit on top of a regular sheet of
paper.
The Compaq LTE and LTE-286 don't have batteries or screens as
good as the ones in the new Portable Macintosh from Apple Computer
Inc. but they are much smaller and lighter than Apple's entries.
The Compaq machines, unveiled Monday, also have more power than
other "notebook" computers from NEC Corp. and Zenith Data Systems
Inc., which are more direct competitors than Apple's because they
run the same software.
The computers measure 2 inches high and are 8 1/2 inches by 11
inches on the sides. They have more than 3 1/2 hours of battery life.
They range in price from $2,399 to $4,599.
Rod Canion, Compaq's president and chief executive, said in an
interview that orders from dealers had been strong. "We believe
we're going to kind of define this marketplace," he said.
<>
<>
```

Using LSI for Information Retrieval Summary

- **Consistent 20 - 30% retrieval advantages over keyword retrieval**
- **Fully automatic and widely applicable**
 - Different languages
 - Cross-language
- **Flexible input and output options**
 - Query: any combination of words or documents
 - Response: documents or words
- **Web interface and user experiments**

Cross-Language Information Retrieval

- **Query in one language matches relevant documents in same or other languages**
- **State-of-the-art: Machine translation of queries**
 - Requires pairwise lexical resources
 - Expensive to develop; Lack coverage; Fail to adequately handle lexical ambiguity
- **Cross-language LSI (CL-LSI)**
 - Fully automatic corpus analysis
 - No translation

Using LSI for Cross-Language Retrieval After Landauer & Littman (1990)

- 1. *Train* using small set of combined multilingual documents
-> derive inter-relationships among terms**
- 2. *Fold in* monolingual documents**
- 3. *Queries* in either language retrieve the most similar documents regardless of language ... no translation of queries**

Using LSI for Cross-Language Retrieval

1. Train Combined

- “Combined” document from Hansard corpus

Hon. Benoit Bouchard (Secretary of State of Canada): Mr. Speaker, I would like to bring to the attention of the House that today, as Hon. Members are no doubt aware, we are celebrating the anniversary of the proclamation of the Canadian Charter of Rights and Freedoms which took place on April 17, 1982, and also of the coming into effect a year ago of the provisions guaranteeing equality for all members of our society. --- *L'hon. Benoit Bouchard (secrétaire d'Etat du Canada): Monsieur le President, je voudrais porter a l'attention de la Chambre que nous celebrons aujourd'hui, comme le savent les honorables deputés, l'anniversaire de la proclamation de la Charte Canadienne des droits et libertés qui a eu lieu le 17 Avril 1982, ainsi que son parachevement, il y a un an, avec l'entree en viguer des dispositions garantissant l'egalite a tous les membres de notre societe.*

Cross-Language LSI (CL-LSI)

Example Results - Mate Retrieval

- Hansard collection
- Train: 982 combined EF documents (2 mins)
- Foldin: 1500 E documents; 1500 F documents
- Test: Mate Retrieval Test (n=1500 queries)

		Eng->Fr	Fr->Eng	Average
Overlap	CL-LSI	98.3%	98.5%	98.4%
	no-LSI	47.4%	49.5%	48.6%
No Overlap	CL-LSI	98.7%	99.1%	98.9%
	no-LSI	.1%	.1%	.1%

Cross-Language LSI (CL-LSI) Mate Retrieval Test

- **How well does a test document retrieve its cross-language mate?**

Query:

Hon. Erik Nielsen (Deputy Prime Minister and Minister of national Defense):

Mr. Speaker, we are in constant touch with our consular officials in Libya. We are advised the situation there is stabilizing now. There is no immediate threat to Canadians. Therefore my response yesterday, which no doubt the Hon. Member has seen, have not altered.

Cross-Language Mate:

L'hon. Erik Nielsen (vice-premier ministre et ministre de la Defense Nationale):

Monsieur le president, nous sommes en communication constante avec nos representants consulaire en Libye. D'apres nos informations, la situation est en train de se stabiliser, et les Canadiens ne sont pas immediatement menaces. Par consequent, mes reponses d'hier, dont le representant a du prendre connaissance, n'ont pas change.

Latent Semantic Indexing (LSI) Cross-Language Retrieval of Yellow Page Categories

French Query “boulangerie”
Matches English Categories:
bakeries retail
bakers retail
cake & pie bakers retail
etc.

Fully automatic & general analysis
No translation required

The screenshot shows a Netscape browser window titled "Netscape: LSI: Query". The search interface includes a menu bar (File, Edit, View, Go, Bookmarks, Options, Directory, Window, Help), a search bar with "LSI : Query" and "Searching in SIC2-XLang", a "Keyword Query" field containing "boulangerie", and a "Submit Query" button. Below the search bar, there are options for "Return" (set to 10), "best matching", "Documents", and "Number of Dimensions" (set to 314). The results section is titled "Results" and shows "Words Found: boulangerie boulangeries". Under "Matching Documents", there is a table with columns "Score" and "Title". Each row has a checkbox to the left of the score. The table lists the following results:

Score	Title
79	<input type="checkbox"/> bakeries retail
68	<input type="checkbox"/> bakers-retail
67	<input type="checkbox"/> cake & pie bakers retail
56	<input type="checkbox"/> bakeries baking & selling retail
54	<input type="checkbox"/> frozen bakery product manufacturing except bread
52	<input type="checkbox"/> bakeries selling only retail
50	<input type="checkbox"/> bread & bakery products except cookies & cracker manufacturing
40	<input type="checkbox"/> cookie shops retail
40	<input type="checkbox"/> cookie & cracker manufacturing

Information Filtering

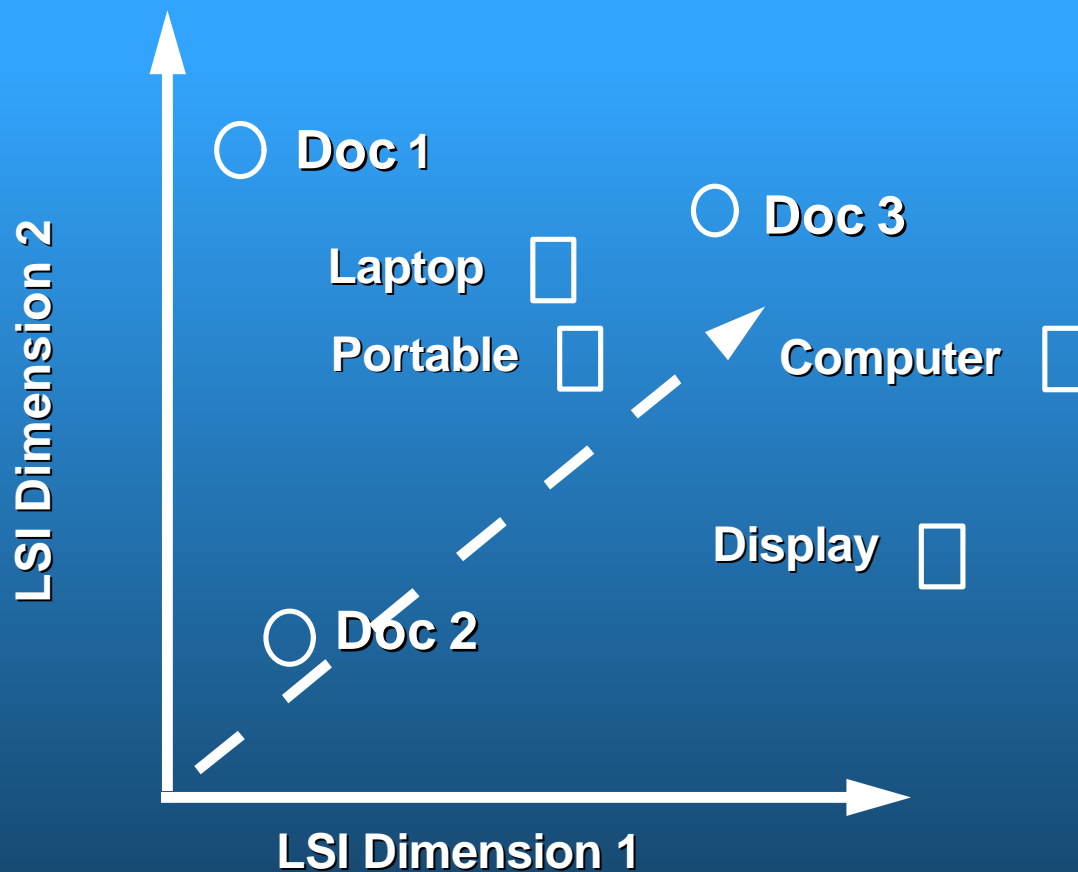
- **Information Retrieval (IR)**
 - Relatively stable database
 - Rapidly changing ad hoc user queries
- **Information Filtering (IF)**
 - Relatively stable information needs
 - Rapidly changing data stream
 - aka: information routing, selective dissemination of information, electronic clipping services, current awareness, classification, push technology

Using LSI for Information Filtering Overview

- **Build representative LSI space**
- **Represent user's interest as vector in LSI space**
- **Add new documents to LSI space**
- **If new document similar to interest, return to user**

- **Describing users' interests**
 - **Free text (like ad hoc queries)**
 - **Known relevant documents**
 - **Various combinations**

Using LSI for Information Filtering User Profile



Using LSI for Information Filtering

Sample Results

- TREC-3 filtering task
 - 50 topics of interest
 - 336,306 new documents
- Results

	average pr	pr at 10	number rel
word filter	.288	.462	6252
rel docs filter	.374	.672	6878
word + rel docs	.379	.682	7078

Simple Filtering Topic

<num> Number: 106

<dom> Domain: Law and Government

<title> Topic: U.S. Control of Insider Trading

<desc> Description

Document will report proposed or enacted changes to U.S. Laws and regulations designed to prevent insider trading

<narr> Narrative:

A relevant document will contain information on proposed or enacted changed to U.S. laws and regulations, including state laws and stock market rules, which are aimed at increasing penalties or closing loopholes in existing institutional discouragements to insider trading. NOT relevant are reports on specific insider trading cases, such as the prosecutions and settlements related to the Boesky - Milken - Drexel Burnham Lambert scandal, unless the report also contains specific information on legal or regulatory change.

<con> Concept(s):

- 1. insider trading**
- 2. securities law, bill, legislation, regulation, rule**
- 3. Insider Trading Sanctions Act, Insider Trading and Securities Fraud Enforcement Act**
- 4. Securities and Exchange Commission, SEC, Commodity Futures Trading Commission, CFTC, National Association of Securities Dealers, NASD**

Using LSI to Assign Reviewers

Overview

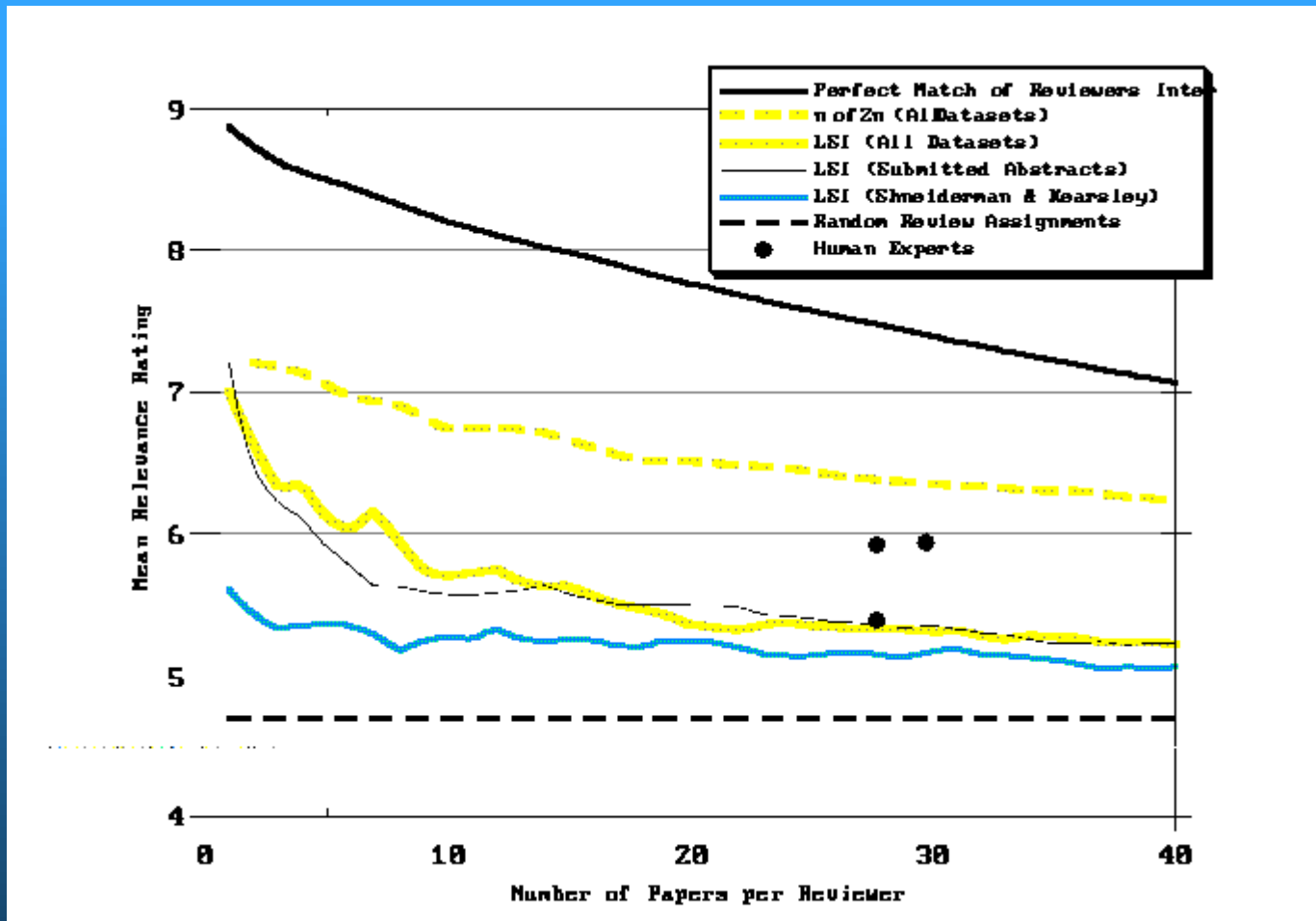
- **Matching submitted manuscripts to reviewers**
 - Difficult when there are many simultaneous submissions and many reviewers
 - E.g., ACM: CHI'94 - 350 submission; 250 reviewers; 7 reviews/submission
- **Using LSI to streamline and improve the process**
 - Represent domain
 - Represent submissions
 - Represent reviewers
 - Compute similarities
 - Load balancing and coverage
 - <20 minutes to do the assignments

Using LSI to Assign Reviewers

Evaluation

- Systematic evaluation for ACM Hypertext'91
 - 117 submissions; 25 reviewers
 - Judgments of their appropriateness as reviewers for every submission
- LSI Domain Analysis
 - Compared several collections - submitted abstracts; reviewer's descriptions; on-line texts and bibliographies
- Represent Reviewer's Interests
 - Text descriptions of interests (mean 3.3)
- Represent Submissions
 - Abstract, title, authors
- Match Submissions to Reviewers
 - top n
 - n of 2n

Using LSI to Assign Reviewers Evaluation



Using LSI to Assign Reviewers

Summary

- LSI method can be used to automate the assignment of reviewers to submissions
- n of 2n method resulted in better mean ratings than human assignment
- Tested w/ Hypertext'91 and CHI'92
- Used for CHI'93 and CHI'94
- More generally applicable - e.g., grants, RFPs

Using LSI to Find Experts

Overview

- **Bellcore Advisor (Streeter & Lochbaum)**
- **Matches request for information with appropriate technical organizations**
 - **Organizations characterized by representative texts (e.g., work descriptions, memos)**
 - **Ad hoc queries**
- **Used within Bellcore and by Technical Recruiting**

Using LSI to Find Experts Evaluation

- LSI analysis of 1500 Bellcore Technical Memos
- Represent 480 departments
- Match queries to relevant groups
- Results

	median rank of true match	75th percentile rank of true match
LSI	2	9
keyword	4	23
max (LSI+key)	1	5

Summary

- **Keyword matching results in surprisingly poor retrieval**
- **LSI can improve access to external information**
- **LSI fully automatic and widely applicable**
- **<http://superbook.bellcore.com/~std/lisi.html>**
- **<http://superbook.bellcore.com/~std/LSI.papers.html>**