

15-826: Multimedia (Databases) and Data Mining

Lecture #15: Text - part IV (LSI)

C. Faloutsos

Must-read Material

- Foltz, P. W. and S. T. Dumais (Dec. 1992).
"Personalized Information Delivery: An
Analysis of Information Filtering Methods."
Comm. of ACM (CACM) 35(12): 51-60.



Outline

Goal: ‘Find **similar / interesting** things’

- Intro to DB
- ➔ • Indexing - similarity search
- Data Mining



Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- ➔ • text
- SVD: a powerful tool
- multimedia
- ...



Text - Detailed outline

- text
 - problem
 - full text scanning
 - inversion
 - signature files
 - clustering
 - information filtering and LSI





LSI - Detailed outline

- LSI



- problem definition
- main idea
- experiments



Problem

- Given a stream of documents
- How to express my interests ('data', 'mining')
- So that I get the 'interesting' ones (including 'machine', 'learning')





Conclusion

- Given a stream of documents
- How to express my interests ('data', 'mining')
- So that I get the 'interesting' ones (including 'machine', 'learning')



A: LSI: automatic 'thesaurus' construction

Information Filtering + LSI

- [Foltz+, '92] Goal:
 - users specify interests (= keywords)
 - system alerts them, on suitable news-documents
- But: how to avoid false dismissals, eg.



'text' 'data'



'information' 'retrieval'

'network' 'security'

'giraffe' 'zoo'

Information Filtering + LSI

- [Foltz+, '92] Goal:
 - users specify interests (= keywords)
 - system alerts them, on suitable news-documents
- Major contribution: LSI = Latent Semantic Indexing
 - latent ('hidden') concepts
 - From a collection of documents, find such 'concepts' (= co-occurring strings)



LSI - Detailed outline

- LSI
 - problem definition
 - ➔ – main idea
 - experiments

Information Filtering + LSI

Main idea

- map each document into some ‘concepts’
- map each term into some ‘concepts’

‘Concept’ :~ a set of terms, with weights, e.g.
– “data” (0.8), “system” (0.5), “retrieval” (0.6) -
 > DBMS_concept

Information Filtering + LSI

Pictorially: term-document matrix (BEFORE)

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	1	1		
TR3				1	1
TR4				1	1

Information Filtering + LSI

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	1	1		
TR3				1	1
TR4				1	1



	'DBMS- concept'	'medical- concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

	'DBMS- concept'	'medical- concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1

Information Filtering + LSI

Pictorially: concept-document matrix and...

	'DBMS- concept'	'medical- concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

Information Filtering + LSI

... and concept-term matrix

	'DBMS- concept'	'medical- concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1

Information Filtering + LSI

Q: How to search, eg., for ‘system’ ?

Information Filtering + LSI

A: find the corresponding concept(s); and the corresponding documents

	'DBMS-concept'	'medical-concept'
data	1	
→ system	1 ↑	
retrieval	1	
lung		1
ear		1

	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

Information Filtering + LSI

A: find the corresponding concept(s); and the corresponding documents

	'DBMS-concept'	'medical-concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1

	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

Information Filtering + LSI

Thus it works like an (automatically constructed) thesaurus:

we may retrieve documents that DON'T have the term 'system', but they contain almost everything else ('data', 'retrieval')

Information Filtering + LSI

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	$\frac{1}{1} \rightarrow 0$	1		
TR3				1	1
TR4				1	1



	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	$\frac{1}{1} \rightarrow 0.8$	
TR3		1
TR4		1

	'DBMS-concept'	'medical-concept'
data	1	
system	$\frac{1}{1} \rightarrow 0.6$	
retrieval	1	
lung		1
ear		1

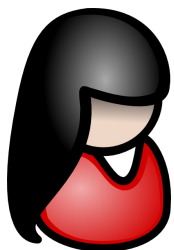
Information Filtering + LSI

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	$\frac{1}{2} \rightarrow 0$	1		
TR3				1	1
TR4				1	1



	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	$\frac{1}{2} \rightarrow 0.8$	
TR3		1
TR4		1

	'DBMS-concept'	'medical-concept'
data	1	
system	$\frac{1}{2} \rightarrow 0.6$	
retrieval	1	
lung		1
ear		1



‘system’

Information Filtering + LSI

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	↔ 0	1		
TR3				1	1
TR4				1	1



	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	↔ 0.8	
TR3		1
TR4		1

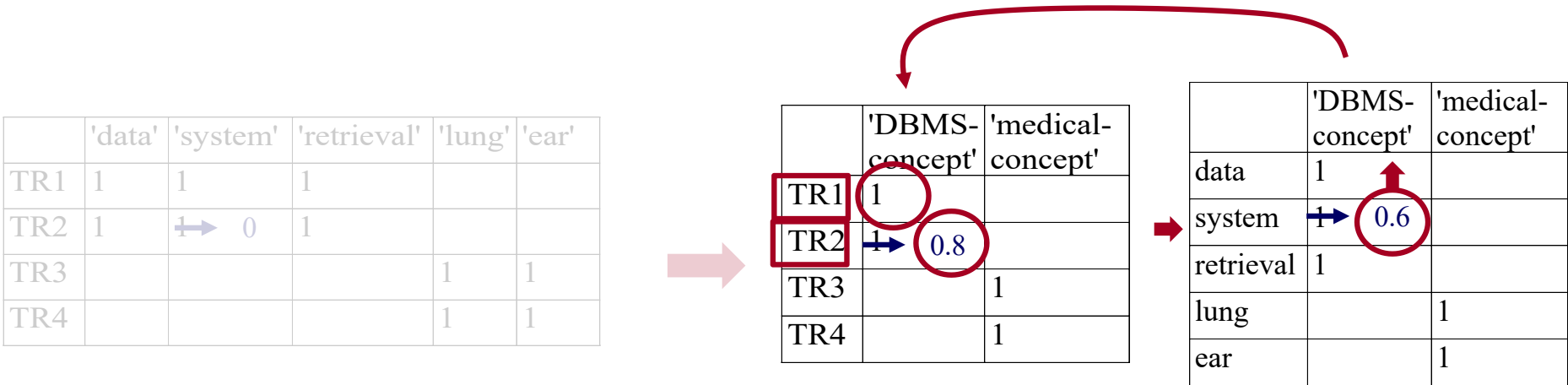
	'DBMS-concept'	'medical-concept'
data	1	
system	↔ 0.6	
retrieval	1	
lung		1
ear		1



‘system’

Usual approach: TR1 only

Information Filtering + LSI



‘system’

With LSI: both TR1 and TR2



LSI - Detailed outline

- LSI
 - problem definition
 - main idea
 - experiments



LSI - Experiments

- 150 Tech Memos (TM) / month
- 34 users submitted 'profiles' (6-66 words per profile)
- 100-300 concepts

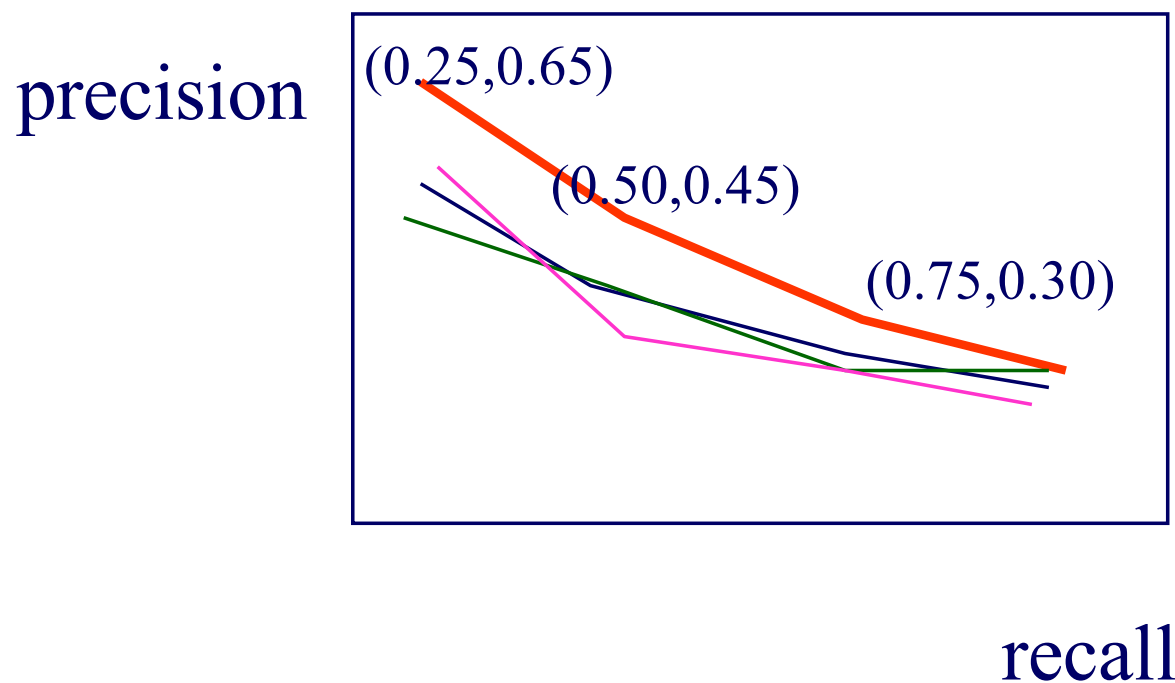
LSI - Experiments

- four methods, cross-product of:
 - vector-space or LSI, for similarity scoring
 - keywords or document-sample, for profile specification
- measured: precision/recall

$$\left\{ \begin{array}{l} ('data', 'retrieval' \dots) \\ (concept1, concept2 \dots) \end{array} \right\} \times \left\{ \begin{array}{l} \bullet \text{ data} \\ \bullet \text{ mining} \\ \bullet \dots \end{array} \right\} \left\{ \begin{array}{c} \square \end{array} \right\}$$

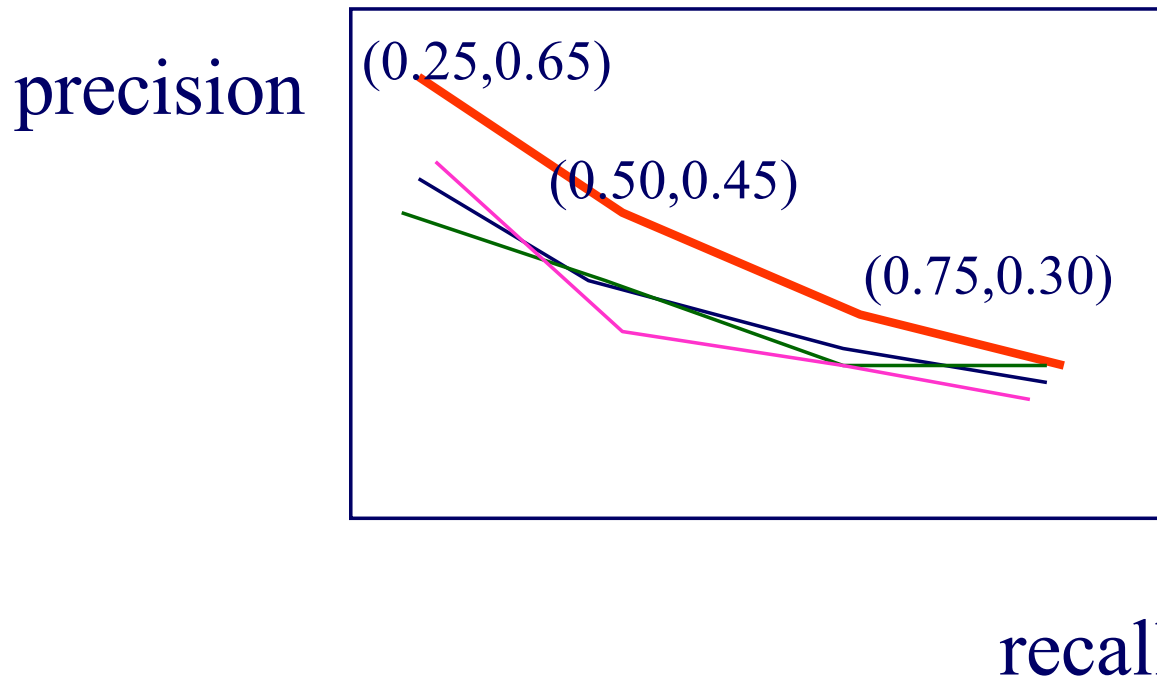
LSI - Experiments

- Q: Who wins?



LSI - Experiments

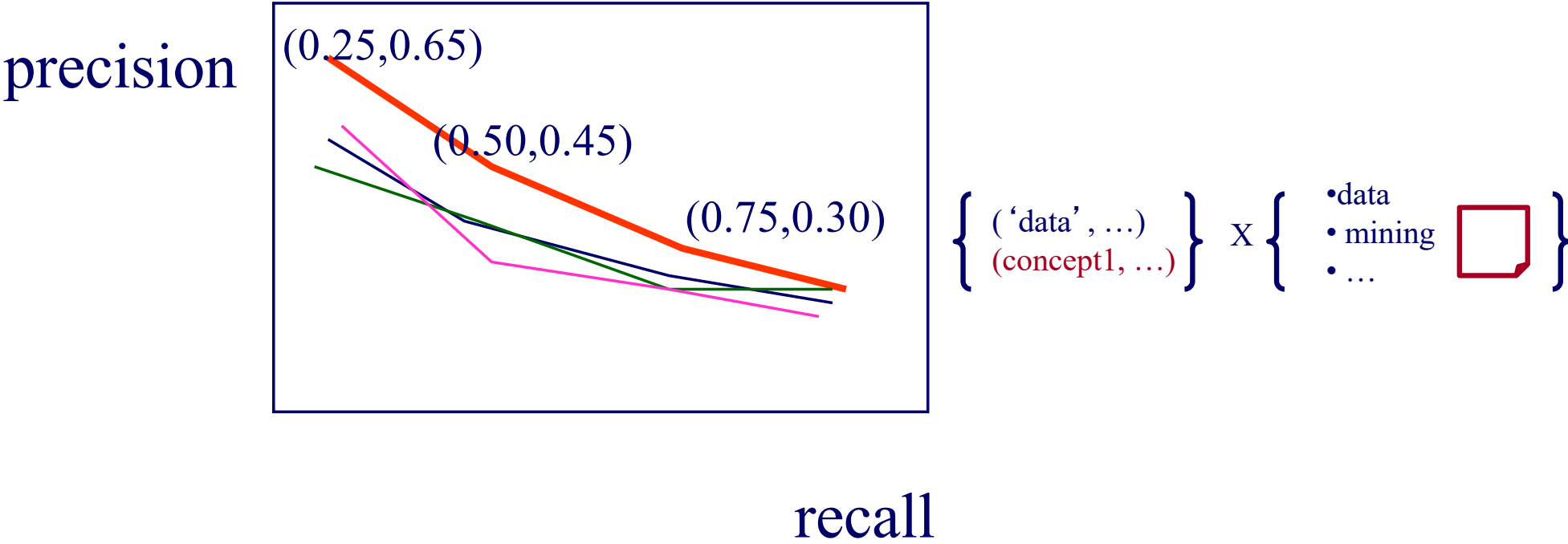
- LSI, with document-based profiles, were better



$$\left\{ \begin{array}{l} ('data', \dots) \\ (\text{concept1}, \dots) \end{array} \right\} \times \left\{ \begin{array}{l} \bullet \text{data} \\ \bullet \text{mining} \\ \bullet \dots \end{array} \right\} \square$$

LSI - Experiments

- LSI, with document-based profiles, were better



LSI - Discussion - Conclusions

- Great idea,
 - to derive ‘concepts’ from documents
 - to build a ‘statistical thesaurus’ automatically
 - to reduce dimensionality
- Often leads to better precision/recall
- but:
 - Needs ‘training’ set of documents
 - ‘concept’ vectors are not sparse anymore

LSI - Discussion - Conclusions

Observations

- Bellcore (-> Telcordia) has a patent
- used for multi-lingual retrieval

How exactly SVD works? (Details, next)

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	1	1		
TR3				1	1
TR4				1	1



	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

	'DBMS-concept'	'medical-concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1



Conclusion

- Given a stream of documents
- How to express my interests ('data', 'mining')
- So that I get the 'interesting' ones (including 'machine', 'learning')



A: LSI: automatic 'thesaurus' construction