


CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #17: Text - part IV (LSI)
C. Faloutsos




CMU SCS

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.

15-826 Copyright: C. Faloutsos (2017) 2




CMU SCS

Outline

Goal: 'Find **similar / interesting** things'

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

15-826 Copyright: C. Faloutsos (2017) 3




CMU SCS

Indexing - Detailed outline

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

15-826 Copyright: C. Faloutsos (2017) 4




CMU SCS

Text - Detailed outline

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

15-826 Copyright: C. Faloutsos (2017) 5




CMU SCS

LSI - Detailed outline

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

15-826 Copyright: C. Faloutsos (2017) 6




CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 7



CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 8

CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 9

CMU SCS

Information Filtering + LSI

Pictorially: concept-document matrix and...

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

15-826 Copyright: C. Faloutsos (2017) 10

CMU SCS

Information Filtering + LSI

... and concept-term matrix

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

15-826 Copyright: C. Faloutsos (2017) 11

CMU SCS

Information Filtering + LSI

Q: How to search, eg., for 'system'?

15-826 Copyright: C. Faloutsos (2017) 12

CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 13

CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 14

CMU SCS

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')

15-826 Copyright: C. Faloutsos (2017) 15

CMU SCS

LSI - Detailed outline

- LSI
 - problem definition
 - main idea
 - – experiments

15-826 Copyright: C. Faloutsos (2017) 16

CMU SCS

LSI - Experiments


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

15-826 Copyright: C. Faloutsos (2017) 17

CMU SCS

LSI - Experiments

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

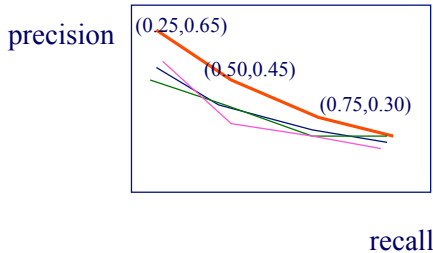
$\left\{ \begin{array}{l} ('data', 'retrieval' \dots) \\ (concept1, concept2 \dots) \end{array} \right\} \times \left\{ \begin{array}{l} \bullet \text{data} \\ \bullet \text{mining} \\ \bullet \dots \end{array} \right\}$


15-826 Copyright: C. Faloutsos (2017) 18

CMU SCS

LSI - Experiments

- Q: Who wins?

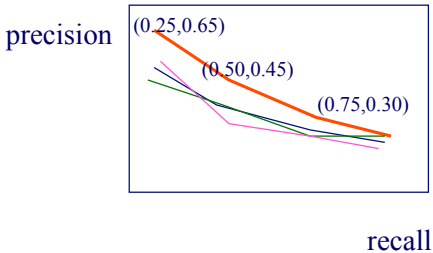



15-826 Copyright: C. Faloutsos (2017) 19

CMU SCS

LSI - Experiments

- LSI, with document-based profiles, were better



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


15-826 Copyright: C. Faloutsos (2017) 20

CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 21

CMU SCS

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

15-826 Copyright: C. Faloutsos (2017) 22