15-826: Multimedia Databases and Data Mining

Text - part IV (LSI)
C. Faloutsos

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

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
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

Q: How to search, eg., for ‘system’?

A: find the corresponding concept(s); and the corresponding documents
Information Filtering + LSI

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

<table>
<thead>
<tr>
<th>Concept</th>
<th>DBMS concept</th>
<th>Medical concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>system</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>retrieval</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>lung</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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’)

LSI - Detailed outline

- LSI
  - problem definition
  - main idea
  - experiments

LSI - Experiments

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

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

Observations

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

How exactly SVD works?

Indexing - Detailed outline

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

References