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

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

Doc 1
Laptop
Portable

LSI Dimension 1

Doc 2

LSI Dimension 2

Doc 3
Laptop
Portable
Computer
Display

Doc 1

Doc 2

Doc 3

Portable

Laptop
Evaluations of LSI for Information Retrieval

- Information Science Test Collections
  - Text objects ("documents")
  - Queries and relevance judgments
  - Evaluation
    - Precision: # relevant retrieved / # retrieved
    - Recall: # relevant retrieved / total # 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

MED: Precision-Recall Curve

- Precision
- Recall

- VO
- LS-99
- Term
- SMART
## LSI Evaluations

### Summary Results

<table>
<thead>
<tr>
<th>Test Collection</th>
<th>LSI</th>
<th>keyword</th>
<th>measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Med-e</td>
<td>.66</td>
<td>.51</td>
<td>average precision</td>
</tr>
<tr>
<td>Med</td>
<td>.52</td>
<td>.46</td>
<td>average precision</td>
</tr>
<tr>
<td>Cran</td>
<td>.39</td>
<td>.29</td>
<td>average precision</td>
</tr>
<tr>
<td>ADI</td>
<td>.29</td>
<td>.26</td>
<td>average precision</td>
</tr>
<tr>
<td>Cisi</td>
<td>.11</td>
<td>.11</td>
<td>average precision</td>
</tr>
<tr>
<td>News</td>
<td>.61</td>
<td>.55</td>
<td>average precision</td>
</tr>
<tr>
<td>TM</td>
<td>.40</td>
<td>.35</td>
<td>average precision</td>
</tr>
<tr>
<td>TREC</td>
<td>.30</td>
<td>.26</td>
<td>average precision</td>
</tr>
<tr>
<td>TREC</td>
<td>8676</td>
<td>8043</td>
<td>number relevant</td>
</tr>
<tr>
<td>Toefl</td>
<td>53.5</td>
<td>29.5</td>
<td>number correct</td>
</tr>
</tbody>
</table>
Latent Semantic Indexing (LSI)
Best matching document does not contain “laptop”
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. **Foldin** 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 (secretaire 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 deputes, l’anniversaire de la proclamation de la Charte Canadienne des droits et libertes 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’égalite 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)

<table>
<thead>
<tr>
<th></th>
<th>Eng-&gt;Fr</th>
<th>Fr-&gt;Eng</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-LSI</td>
<td>98.3%</td>
<td>98.5%</td>
<td>98.4%</td>
</tr>
<tr>
<td>no-LSI</td>
<td>47.4%</td>
<td>49.5%</td>
<td>48.6%</td>
</tr>
<tr>
<td>No Overlap</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-LSI</td>
<td>98.7%</td>
<td>99.1%</td>
<td>98.9%</td>
</tr>
<tr>
<td>no-LSI</td>
<td>.1%</td>
<td>.1%</td>
<td>.1%</td>
</tr>
</tbody>
</table>
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 représentants 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
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

LSI Dimension 1

LSI Dimension 2

Doc 1

Doc 3

Laptop

Portable

Computer

Display

Doc 2
Using LSI for Information Filtering

Sample Results

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

- **Results**

<table>
<thead>
<tr>
<th></th>
<th>average pr</th>
<th>pr at 10</th>
<th>number rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>word filter</td>
<td>.288</td>
<td>.462</td>
<td>6252</td>
</tr>
<tr>
<td>rel docs filter</td>
<td>.374</td>
<td>.672</td>
<td>6878</td>
</tr>
<tr>
<td>word + rel docs</td>
<td>.379</td>
<td>.682</td>
<td>7078</td>
</tr>
</tbody>
</table>
Simple Filtering Topic

Number: 106
Domain: Law and Government
Topic: U.S. Control of Insider Trading
Description:
Document will report proposed or enacted changes to U.S. Laws and regulations designed to prevent insider trading.

Narrative:
A relevant document will contain information on proposed or enacted changes 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.

Concept(s):
1. insider trading
2. securities law, bill, legislation, regulation, rule
3. Insider Trading Sanctions Act, Insider Trading and Securities Fraud Enforcement Act
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

The graph shows the comparison of different methods for assigning reviewers, with LSI (Latent Semantic Indexing) performing better than random assignments. The y-axis represents the mean relevance rating, while the x-axis shows the number of papers per reviewer. The legend includes various methods for assignment, such as perfect match of reviewers, LSI on all datasets, and human experts. The LSI methods consistently outperform the random assignments.
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**

<table>
<thead>
<tr>
<th></th>
<th>median rank of true match</th>
<th>75th percentile rank of true match</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>keyword</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>max (LSI+key)</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
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/lsi.html
http://superbook.bellcore.com/~std/LSI.papers.html