Information Retrieval Techniques for Spoken Language Processing

Lee-Feng Chien (簡立峰)

Institute of Information Science
Academia Sinica, Taiwan
Outline

- IR vs. SLP
- Conventional IR Techniques
- Web IR Techniques
- Web Mining Techniques
- Term Clustering through Web Mining
- Anchor Text Mining
I. IR vs. SLP
Information Retrieval

- a research with a long-term research goal of exploration of information storage, classification, extraction, indexing and browsing techniques for the retrieval of non-structural databases such as textual documents
Different Research Aspects

- Text IR
- Web IR
- Multimedia IR
- Intelligent IR
Different Research Aspects

- **Text IR**
  - Text indexing, searching, presentation, user study

- **Web IR**
  - Crawling, page ranking, distributed search, scalability, multi-lingual and multi-culture

- **Multimedia IR**
  - Retrieving multimedia contents such as speech, audio, music, image, video

- **Intelligent IR**
  - Advanced language and information processing topics such as question answering, cross-language, information tracking, information extraction, summarization, speech interaction, etc.
Chinese Information Retrieval

- Language issues
- Culture/geographical issues
- Chinese people issues
Chinese Information Retrieval

- **Language issues**
  - Word segmentation, term extraction, parsing, linguistic resources
  - Font display, conversion between simplified Chinese and traditional Chinese, etc.

- **Culture/geographical issues**
  - Chinese pages from world wide, language identification required, preferred topics different, etc.

- **Chinese people issues**
  - User behaviors
## Web Users and Pages (3 years ago)

<table>
<thead>
<tr>
<th>Area</th>
<th>Users</th>
<th>Web Pages</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>World-wide</td>
<td>150M</td>
<td>800M</td>
<td>7/99</td>
</tr>
<tr>
<td>China</td>
<td>4M</td>
<td>2.5~3M</td>
<td>7/99</td>
</tr>
<tr>
<td>Taiwan</td>
<td>4M</td>
<td>3M</td>
<td>7/99</td>
</tr>
</tbody>
</table>
Number of Chinese Web Pages

328,000,000 pages
Number of Chinese Web Pages

328,000,000 pages

Scalability Problem!
Number of Web Pages

The world’s largest search engine?

2,073,418,204 pages (Google)
2,095,568,809 pages (FAST)
Number of Web Pages (Cont.)
Why IR Useful for SLP?

- ASR or dictation machine: lexicon, corpus, and language model
- Voice portal: search via spoken queries
- Speech retrieval: indexing & searching
- Topic detection & tracking: document classification & clustering
IR vs. SLP

Information Retrieval

Search Engine

Q&A

IR Via Voice

Speech IR

Natural Language Processing

Parser

Dictation Machine

Speech Recognition

ISCSLP'02

L. F. Chien
### Anchor Text Mining for Query Translation

**Lu, ICDM’01**

<table>
<thead>
<tr>
<th>English</th>
<th>Traditional Chinese</th>
<th>Simplified Chinese</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sony</td>
<td>新力</td>
<td>索尼</td>
<td>ソニー</td>
</tr>
<tr>
<td>Nike</td>
<td>耐吉</td>
<td>耐克</td>
<td>ナイキ</td>
</tr>
<tr>
<td>Stanford</td>
<td>史丹佛</td>
<td>斯坦福</td>
<td>スタンフォード</td>
</tr>
<tr>
<td>Sydney</td>
<td>雪梨</td>
<td>悉尼</td>
<td>シドニー</td>
</tr>
<tr>
<td>Internet</td>
<td>網際網路</td>
<td>网络</td>
<td>インターネット</td>
</tr>
<tr>
<td>Homepage</td>
<td>首頁</td>
<td>主页</td>
<td>ホームページ</td>
</tr>
<tr>
<td>Computer</td>
<td>電腦</td>
<td>计算机</td>
<td>コンピューター</td>
</tr>
<tr>
<td>Database</td>
<td>資料庫</td>
<td>数据库</td>
<td>データベース</td>
</tr>
<tr>
<td>Information</td>
<td>資訊</td>
<td>信息</td>
<td>インフォメーション</td>
</tr>
</tbody>
</table>

Diagram illustrates the connections between English, Traditional Chinese, Simplified Chinese, and Japanese anchor texts and websites.
Cross-Language Web Search

- A Web search service allows users to query in one language and search documents that are written or indexed in another language.
II. Conventional IR Techniques
The Vector Space Model

- Measure closeness between query and document.
  - Queries and documents represented as n dimensional vectors.
  - Each dimension corresponds to a word/term.
  - Advantages: Conceptual simplicity and use of spatial proximity for semantic proximity.
## Vector Similarity

- \( d = \) The man said that a space age man appeared \( d' \)
  - Those men appeared to say their age

<table>
<thead>
<tr>
<th></th>
<th>( \tilde{d} )</th>
<th>( \tilde{d}' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>appeared</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>man</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>men</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>said</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>say</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>space</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Vector Similarity (Cont.)

- cosine measure or normalized correlation coefficient

\[
\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}
\]

- Euclidean Distance

\[
|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]
Term Weighting

- **Quantities used:**
  - $tf_{i,j}$ (Term frequency): # of occurrences of $w_i$ in $d_i$
  - $df_i$ (Document frequency): # of documents that $w_i$ occurs in
  - $cf_i$ (Collection frequency): total # of occurrences of $w_i$ in the collection
Inverted File for Keyword Matching

Hit: 2 bytes

<table>
<thead>
<tr>
<th>plain</th>
<th>cap:1</th>
<th>imp:3</th>
<th>position: 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>fancy</td>
<td>cap:1</td>
<td>imp:7</td>
<td>type: 4</td>
</tr>
<tr>
<td>anchor</td>
<td>cap:1</td>
<td>imp:7</td>
<td>type: 4</td>
</tr>
</tbody>
</table>

Forward Barrels: total 43 GB

- docid: word: 24 nhits: 8 hit hit hit hit
  - word: 24 nhits: 8 hit hit hit hit
  - null wordid
- docid: word: 24 nhits: 8 hit hit hit hit
  - word: 24 nhits: 8 hit hit hit hit
  - word: 24 nhits: 8 hit hit hit hit
  - null wordid

Lexicon: 293MB

Inverted Barrels: 41 GB

Google’s Index File Structure
Chinese IR & Indexing Unit Section

- Character-based indexing and search
  - speed/space problem
  - incorrect matching due to free combination of characters, EX: 電腦科學
- Word-based indexing and search
  - lexicon is a prerequisite and limitation
  - unknown word identification, disambiguation of word segmentation
- Csmart’s approach (Chien, SIGIR’95)
  - signature-based, feature grouping (unigram, bigram characters)
  - two-stage search, fuzzy search; suited for not too large files and demand of fuzzy search
Chinese Track in TREC’5

- Berkeley (A. Chen, SIGIR’97)
  - Indexing units
    - use dictionary to segment texts
    - obtained from public domain with 91,000 words and phrases
    - stopword list with 444 entries
  - Searching algorithm, segmentation method
    - 0.461 average precision for manual queries, 0.32 for automatic run

- CUNY (Kwok, SIGIR’97)
  - Indexing units
    - Use 2,000 words to segment texts initially
    - use a learning strategy to extend the word entries to 15,000 finally
  - Searching algorithm, segmentation method
    - 0.40 in word-based; 0.42 in word and character-based
Text Indexing

- Indexing Chinese texts
  - Indexing Units
    - Single character, Bi-character, word, term, string
  - New word identification and word segmentation problems
- Indexing English texts
  - Indexing units
    - Word, term, string (few)
  - Problems
    - Stemming, capitalization, hyphen, word sense disambiguation, typing errors
- Structure for indexing
  - Inverted file, PAT array
- Term extraction and term clustering techniques are the required key techniques
Term Extraction

- Term is a meaningful and representative unit in terms of information retrieval, e.g., name, location, proper noun, topic
- Terms are derived and most excluded in common dictionaries
- Term extraction can reduce word sense ambiguities in text retrieval and remedy weaknesses of word-based approaches, EX:
  - computer **network** (linking, net, mesh)
  - **Bank** America, current theory
- Term extraction is the first step toward concept-based IR
Chinese Term Extraction

- PAT-tree-based Approach
  - Poster Presentation Award by ACM SIGIR’98
Context Dependency

Left Context Dependency (LCD)  Right Context Dependency (RCD)

Association

Lexical Pattern

到達 訪問 新加坡

新加坡 國大 政府

usage of freedom
Example of the PAT tree

*(Chien, SIGIR’97)*

abcd, ed

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>abcd</td>
<td>(a, ab, abc, abcd)</td>
</tr>
<tr>
<td>bcd</td>
<td>(b, bc, bcd)</td>
</tr>
<tr>
<td>cd</td>
<td>(c, cd)</td>
</tr>
<tr>
<td>d</td>
<td>(d)</td>
</tr>
<tr>
<td>ed</td>
<td>(e, ed)</td>
</tr>
<tr>
<td>d</td>
<td>(d)</td>
</tr>
</tbody>
</table>

Data stream:

```
個人電腦
person
```

Semi-infinite strings:

```
bit: 1           9             17            25
個人: 10101101  11010011   10100100  ...
人: 10110001  01001000  11011001  ...
腦: 10111000  00000000  00000000  ...
個人: 10100100  01001000  00000000  ...
```

Data position

```
(24,2,1)
(0,6,1)
(4,6,1)
(5,3,1)
(8,3,2)
```

Suffix  Prefix
abcd    (a, ab, abc, abcd)
bcd     (b, bc, bcd)
cd      (c, cd)
d       (d)
ed      (e, ed)
Incremental Term Extraction

<table>
<thead>
<tr>
<th>Term length (character N-gram)</th>
<th>Number of extracted new terms</th>
<th>Number of documents with new terms extracted</th>
<th>Average number of document inputs can find new terms (A)</th>
<th>Average frequency of the extracted new terms</th>
<th>Average frequency as the term can be extracted</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>776</td>
<td>515</td>
<td>3.93</td>
<td>34.22</td>
<td>9.37</td>
</tr>
<tr>
<td>3</td>
<td>416</td>
<td>325</td>
<td>6.04</td>
<td>24.60</td>
<td>9.09</td>
</tr>
<tr>
<td>4</td>
<td>171</td>
<td>157</td>
<td>12.16</td>
<td>19.22</td>
<td>8.97</td>
</tr>
<tr>
<td>5</td>
<td>51</td>
<td>49</td>
<td>37.28</td>
<td>20.35</td>
<td>9.18</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>17</td>
<td>109.81</td>
<td>27.00</td>
<td>8.65</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>15</td>
<td>123.60</td>
<td>27.40</td>
<td>11.20</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>6</td>
<td>274.67</td>
<td>13.00</td>
<td>9.83</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>205.67</td>
<td>18.00</td>
<td>11.33</td>
</tr>
<tr>
<td>Total N-grams</td>
<td>1,455</td>
<td>814</td>
<td>2.41</td>
<td>28.95</td>
<td>9.25</td>
</tr>
</tbody>
</table>

Table 4. The detailed results for incremental term extraction when the threshold value was larger than 2 in the significance analysis; the results were obtained from a total of 1,872 political news abstracts published in July 1997.
### Incremental Term Extraction (Cont.)

<table>
<thead>
<tr>
<th>S(Y)</th>
<th>Total Extracted Terms(A)</th>
<th>No. of Correct Terms Extracted(B)</th>
<th>No. of Correct Terms Outside Dictionary(C)</th>
<th>Precision (B/A)</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1.5</td>
<td>2,291</td>
<td>1,374</td>
<td>297</td>
<td>0.60</td>
<td>0.53</td>
</tr>
<tr>
<td>&gt;2</td>
<td>1,455</td>
<td>1,135</td>
<td>258</td>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>&gt;2.5</td>
<td>723</td>
<td>593</td>
<td>172</td>
<td>0.82</td>
<td>0.23</td>
</tr>
<tr>
<td>&gt;3</td>
<td>214</td>
<td>184</td>
<td>66</td>
<td>0.86</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 3. The testing results for incremental term extraction using different threshold values in the significance analysis; the results were obtained from a total of 1,872 political news abstracts published in July, 1997.
Incremental Term Extraction (Cont.)

<table>
<thead>
<tr>
<th>S(Y)</th>
<th>Total Extracted Terms(A)</th>
<th>No. of Correct Terms Extracted(B)</th>
<th>No. of Correct Terms Outside Dictionary(C)</th>
<th>Precision (B/A)</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1.5</td>
<td>2,291</td>
<td>1,374</td>
<td>297</td>
<td>0.60</td>
<td>0.53</td>
</tr>
<tr>
<td>&gt;2</td>
<td>1,455</td>
<td>1,135</td>
<td>258</td>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>&gt;2.5</td>
<td>723</td>
<td>593</td>
<td>172</td>
<td>0.82</td>
<td>0.23</td>
</tr>
<tr>
<td>&gt;3</td>
<td>214</td>
<td>184</td>
<td>66</td>
<td>0.86</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 3. The testing results for incremental term extraction using different threshold values in the significance analysis; the results were obtained from a total of 1,872 political news abstracts published in July, 1997.

How to deal with low-frequency terms?
Characteristics of the Chinese Language

Word segmentation:
• 一台大電腦 (a set of large computer)
Characteristics of the Chinese Language

Word segmentation:
• 一台大電腦 (a set of large computer)
Characteristics of the Chinese Language

Word segmentation:
• 一台大電腦 (a set of large computer)
Characteristics of the Chinese Language

Word segmentation:
• 一台大電腦 (National Taiwan University)
Characteristics of the Chinese Language

Word segmentation:

参加一台大電腦會議 (Attend a computer science meeting in National Taiwan University)
Characteristics of the Chinese Language

New word identification:
• 台大電腦公司 (Taida Computer Inc.)
III. Web IR Techniques
Spectrums of Web Search

- Types of content
  - Text, e.g. Web text, documents, news
  - Audio, e.g. music, speech, sounds, broadcast news
  - Image, e.g. pictures, photos, graphics
  - Video, e.g. films, clips
  - Formatted Data, e.g. products

- Scopes of content
  - General or specific
  - Languages

- Scalability
  - Personal, content site, intranet, Internet
  - Thousands, millions or billions of (documents, users, queries)

- Interface
  - Web-based, WAP-based, Voice-based
An Analytical Model

User Space

X
Users

X1, X2...

Index Space

Y1, Y2...
Document Space

Seek

Y
Authors

Use

Information Need

Information Use

Short Query
Subject Terms
Real Names
Different Web Search Models

- Yahoo
  - manual recommendation in index space
- Altavista, Inktomi
  - full-text pattern matching in document space
- Google
  - citation information in document space
- Realname
  - manual real-name retrieval in user space
- DirectHit
  - collaborative analysis in user space
- AskJeeves
  - Q&A (or FAQ search) in specific domains
Hypertext on the Web
Basic Architecture of a Spider-based Web Search Engine

Scalable, e.g., 20K PCs in Google
Crawling

Authorized Pages

Out link Traverse

Duplication

Out Links

Indexed Page

Authority
Indexed Features & Page Ranking

Page Title: Academia Sinica

Anchor Text:
Highest Government Research Institution in Taiwan

Popularity

Indexed Page

abstract

Anchor Text:
Chien’s Lab

Authority
Distributed Search

Query Processor

Document Delivery
Facts and Problems I

- **Query**
  - short query problem
  - 50% are personal and company names
  - Boolean or natural language query is few

- **Browsing**
  - no more 2nd page
  - precision is more important than recall

- **Robot**
  - low coverage, deadlinks, garbage sites and pages
Facts and Problems II- Relevancy

- Who judge the retrieval relevancy
  - Users
    - What user want? What do they input?
    - Short query or NLQ?
    - HFQ、LFQ or MFQ?
  - Search engines
    - Technology limitations?
    - How many indexed pages? millions or billions of pages?
    - Ranking algorithm?
Quality vs. Quantity

Facts and Problems III - Speed

What make the retrieval speed?

Users
- Where you are and how is the bandwidth?
  - Dial-up or T1? Cache or proxy?
- What is the query?
  - HFQ, LFQ or MFQ?

Search Engines
- How far and how good the infrastructure
- Document delivery speed is the key
Important Issues

- Web user study
  - User behavior analysis
  - Query log mining
- Content indexing
  - Language identification
  - Information conversion
  - Unified indexing
  - Term extraction
  - Term clustering
- Content searching
  - English-Chinese bilingual search
  - Concept-based search
  - Personalized search
  - Cross-language search
- Content presentation
  - Concept-based term suggestion
  - Concept-based search result clustering
<table>
<thead>
<tr>
<th></th>
<th>All Chinese</th>
<th>All English</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dreamer</td>
<td>78.20%</td>
<td>19.18%</td>
<td>2.62%</td>
</tr>
<tr>
<td>GAIS</td>
<td>78.22%</td>
<td>16.90%</td>
<td>4.88%</td>
</tr>
</tbody>
</table>

Table 3  Statistics concerning what language used in each search term
## Term Length

Table 4  Statistics concerning the number of terms per query

<table>
<thead>
<tr>
<th></th>
<th>in Chinese</th>
<th>in English</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dreamer</td>
<td>3.18 characters</td>
<td>1.10 words</td>
<td>6.31 bytes</td>
</tr>
<tr>
<td>GAIS</td>
<td>3.55 characters</td>
<td>1.22 words</td>
<td>7.26 bytes</td>
</tr>
</tbody>
</table>
Information Needs by Subject Categories

-- Query categorization approach (Pu, JASIS’02)
  Obtained from analyzing about 2M queries
Term Conversion
IV. Web Mining
Web Search

Information Seeking

Web logs, texts, images, ...

Search Engine

Millions of Users
Web Mining

Knowledge Discovery

Web logs, texts, images, …

Search Engine

Millions of Users
Taxonomy of Web Mining (R. Cooley)

- Web Mining
  - Web Content Mining
  - Web Structure Mining
  - Web Usage Mining

DM
Web Content Mining

- Most focus on extraction of knowledge from the text of web pages
- Web Page Classification (Chuang & Chien’s IRWK’02)
- Text Mining
  - Web Information Extraction
  - XML/Semantic Web Mining
  - Message Understanding (NLP viewpoint)
- Multimedia Content Mining
  - Web Image Classification (Tseng’s IRWK’02)
  - Speech Archive Mining (Chien’s ISCSLP’02)
Web Page Classification Applications

- CMU Web→KB Project (1998-2000) [Craven98]

Classifying Web pages is an essential step to construct Web knowledge base
Web Usage Mining

- Data Gathering
  - Web server log, site description data, concept hierarchies

- Data Preparation
  - Distinguish among users, build sessions

- Data Mining
  - Pattern discovery & analysis
Web Structure Mining

- Google’s Page Rank

\[
PR(A) = (1 - d) + d \left( \frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)} \right)
\]

- Document Citation (siteseer)
Semantic Web Mining

- Current Web
  - Most of Web content is designed for humans to read, not for machine to manipulate meaningfully

- Semantic Web
  - XML+RDF + Ontology + Agent

- Semantic Web Mining
  - Auto-construction of Ontology
  - Case-based reasoning/inference
V. Term Clustering Through Web Mining
Term Clustering (Chuang’02)

Hierarchical clustering

- 1. 武俠小說
- 2. 補貼
- 3. eva 長榮航空 (EVA airline)
- 4. 占卜 塔羅牌 算命 紫微斗數 命理 姓名學 心理測驗 星座 愛情
- 5. 劳委會 職訓局 就業 背轉會 自傳 徵才 人力資源 104人力銀行 人力銀行 找工作 履歷表 求職 求才
Hierarchical Query Clustering

ISCLSP'02 L. F. Chien
Virus
Virus synonyms
ISCSLP’02

L. F. Chien
Search Result Clustering
Personalized Search
Hierarchical Query Clustering
The Steps

- **Feature Extraction**
  - Use co-occurred seed terms extracted from retrieved top pages

- **Term Vector**
  - Each query term is assigned a term vector
    - Record the co-occurred feature terms and their frequency values in the retrieved documents.

- **Term Similarity**
  - tf*idf-based Cosine measurement

- **Hierarchical Term Clustering**
  - Cluster popular query terms in the log into initial categories
  - Query terms with similar features are grouped into clusters.
Feature Extraction

- Use co-occurred **seed terms** extracted from retrieved top pages

![Google Image](image)

Creative Nude Photography Network -- Fine Art Nude and ...

... The Creative **Nude** and **Erotic Photography** Network is the number one net portal to the best in fine art **nude** and **erotic photography**! Over 100 CNPN Member Sites ...

Nude Places

... to be **naked**. Walking in the forest, cruising the lake in open boats, swimming, picnicking and **nude** photography are all enjoyed in the **nude**. 60 minutes $39.95. ...

A Brave Nude World

... A Brave **Nude** World! Warning: This site contains links to fine art **nude** & **erotic photography**. If you are under 18 or do not wish to view this material, You can ...

<table>
<thead>
<tr>
<th>term</th>
<th>tf/df</th>
</tr>
</thead>
<tbody>
<tr>
<td>erotic photography</td>
<td>3/2</td>
</tr>
<tr>
<td>naked</td>
<td>1/1</td>
</tr>
<tr>
<td>photography</td>
<td>2/2</td>
</tr>
<tr>
<td>art</td>
<td>3/2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Term Weighting

term weight $v_{i,j}$ is defined as:

$$v_{i,j} = \left( 0.5 + 0.5 \frac{tf_{i,j}}{\max_{t_k \in T} tf_{i,k}} \right) \log \frac{n}{n_j}$$

where $tf_{i,j}$, the term frequency, is the number of occurrences of term $t_j$ in the $v_i$'s corresponding feature term bag, $n$ is the total number of query terms, and $n_j$ is the number of query terms which contain $t_j$ in their corre-
Extraction of Basic Feature Terms

- Performance of different features: randomly selected, high-frequency, and seed terms
  - Popular queries not affected by ephemeral trends, e.g., “movie”, “basketball”, “mutual fund”, etc.
  - More expressive and distinguishable in describing a particular category
  - Two logs compared and extracted 9,709 overlapping top query terms as feature terms

<table>
<thead>
<tr>
<th></th>
<th>G-1999</th>
<th>Top 1,000 terms</th>
<th>top 20,000 terms</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-1998 Top 1,000</td>
<td>583/58.30%</td>
<td>977/97.70%</td>
<td>992/99.20%</td>
<td></td>
</tr>
<tr>
<td>Top 20,000 terms</td>
<td>914/91.40%</td>
<td>9,709/50.71%</td>
<td>14,721/76.89%</td>
<td></td>
</tr>
</tbody>
</table>
Query Clustering (Cont.)

- **Feature Extraction**
  - Use co-occurred seed terms extracted from retrieved top pages

- **Term Vector**
  - Each query term is assigned a term vector
    - Record the co-occurred feature terms and their frequency values in the retrieved documents.

- **Term Similarity**
  - TF * IDF-based Cosine measurement

- **Hierarchical Term Clustering**
  - Cluster popular query terms in the log into initial categories
  - Query terms with similar features are grouped into clusters.
Term Similarity

The similarity between a pair of query terms is computed as the cosine of the angle between the corresponding vector \((\cos \theta)\), i.e.,

\[
\text{sim}(v_a, v_b) = \frac{\sum_{t_j \in T} v_{a,j} v_{b,j}}{\sqrt{\sum_{t_j \in T} v_{a,j}^2} \sqrt{\sum_{t_j \in T} v_{b,j}^2}}.
\]
Hierarchical Term Clustering

- **Agglomerative hierarchical clustering (AHC)**
  - Compute the similarity between all pairs of clusters
    - Estimate similarity between all pairs of composed terms
    - Use the lowest term similarity value as the cluster similarity value
  - Merge the most similar (closest) two clusters
    - Complete linkage method
  - **Update** the cluster vector of the new cluster
  - Repeat steps 2 and 3 until **only a single cluster remains**
Hierarchical Term Clustering

- Agglomerative hierarchical clustering (AHC)
  - Compute the similarity between all pairs of clusters
    - Estimate similarity between all pairs of composed terms
Hierarchical Term Clustering

- Agglomerative hierarchical clustering (AHC)
  - Compute the similarity between all pairs of clusters
    - Estimate similarity between all pairs of composed terms
Hierarchical Term Clustering

- Agglomerative hierarchical clustering (AHC)
  - Compute the similarity between all pairs of clusters
    - Estimate similarity between all pairs of composed terms
    - Use the lowest term similarity value as the cluster similarity value
Hierarchical Term Clustering

- **Agglomerative hierarchical clustering (AHC)**
- Merge the most similar (closest) two clusters
  - Complete-linkage method
Hierarchical Term Clustering

- **Agglomerative hierarchical clustering (AHC)**
  - Merge the most similar (closest) two clusters
    - Complete-linkage method

### Table 1: Three well-known cluster distance functions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Distance function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-linkage (SL)</td>
<td>[ \min_{v_a \in C_i, v_b \in C_j} \text{dist}(v_a, v_b) ]</td>
</tr>
<tr>
<td>Average-linkage (AL)</td>
<td>[ \frac{1}{</td>
</tr>
<tr>
<td>Complete-linkage (CL)</td>
<td>[ \max_{v_a \in C_i, v_b \in C_j} \text{dist}(v_a, v_b) ]</td>
</tr>
</tbody>
</table>
Hierarchical Term Clustering

- **Agglomerative hierarchical clustering (AHC)**
  - Compute the similarity between all pairs of clusters
    - Estimate similarity between all pairs of composed terms
    - Use the lowest term similarity value as the cluster similarity value
  - Merge the most similar (closest) two clusters
    - Complete linkage method
  - Update the cluster vector of the new cluster
  - Repeat steps 2 and 3 until only a single cluster remains
The Clustering Algorithm

\[ \text{HAC}(v_1, v_2, \ldots, v_n) \]
\[ v_i, 1 \leq i \leq n: \text{the vectors of the objects} \]
\[ 1: \text{for all } v_i, 1 \leq i \leq n \text{ do} \]
\[ 2: \quad C_i \leftarrow \{v_i\} \]
\[ 3: \quad f(i) \leftarrow \text{false} \{f: \text{whether a cluster has been merged}\} \]
\[ 4: \text{calculate the pairwise cluster distance matrix} \]
\[ 5: \text{for all } 1 \leq i < n \text{ do} \]
\[ 6: \quad \text{choose the closest pair } \{C_a, C_b\} \text{ with } f(a) \land f(b) \]
\[ 7: \quad C_{n+i} \leftarrow C_a \cup C_b, \text{left}(C_{n+i}) \leftarrow C_a, \text{right}(C_{n+i}) \leftarrow C_b \]
\[ 8: \quad f(n+i) \leftarrow \text{false}, f(a) \leftarrow \text{true}, f(b) \leftarrow \text{true} \]
\[ 9: \quad \text{update the distance matrix with new cluster } C_{n+i} \]
\[ 10: \text{return } C_1, C_2, \ldots, C_{2n-1} \text{ with functions left and right} \]
<table>
<thead>
<tr>
<th>Cluster 44 (Airlines class, Precision: 1, Recall: 1)</th>
<th>Cluster 28 (Security software class, Precision: 0.78, Recall 0.55)</th>
<th>Cluster 38 (Job class, Precision: 0.8; Recall: 0.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eva (Airline name) tp</td>
<td>病毒 (Virus) cd</td>
<td>履歷表 (Resume) lj</td>
</tr>
<tr>
<td>長榮航空 (Airline) tp</td>
<td>norton cd</td>
<td>自傳 (Bio) lj</td>
</tr>
<tr>
<td>長榮 (EVA) tp</td>
<td>病毒 (Virus Scanning) cd</td>
<td>就業 (Employment) lj</td>
</tr>
<tr>
<td>機票 (Flight Ticket) tp</td>
<td>病毒碼 (Virus Code) cd</td>
<td>徵才 (Job Opportunities) lj</td>
</tr>
<tr>
<td>遠東航空 (FAT Airline) tp</td>
<td>防毒軟體 (Anti-virus SW) cd</td>
<td>千里馬 (Taiwan Job Online) lj</td>
</tr>
<tr>
<td>復興航空 (Formosa Airline) tp</td>
<td>防毒 (Anti-virus) cd</td>
<td>工作 (Job) lj</td>
</tr>
<tr>
<td>福航 (China Airline ab.) tp</td>
<td>電腦病毒 (Computer Virus) cd</td>
<td>找工作 (Job Finding) lj</td>
</tr>
<tr>
<td>中華航空 (China Airline) tp</td>
<td>趨勢科技 (Trends, Inc.) cg</td>
<td>人力 (Human Resource) lj</td>
</tr>
<tr>
<td>飛機 (Airplane) tp</td>
<td>pc-cillin cd</td>
<td>104 (HR site) lj</td>
</tr>
<tr>
<td>航空公司 (Airway Co.) tp</td>
<td>cih cd</td>
<td>104 人力銀行 (HR site) lj</td>
</tr>
<tr>
<td>航空 (Airway) tp</td>
<td>ssan cd</td>
<td>人力銀行 (HR site) lj</td>
</tr>
<tr>
<td></td>
<td>搞怪 cd</td>
<td>人力資源 (HR) lj</td>
</tr>
<tr>
<td></td>
<td>驚客 (Hacker) ck</td>
<td>job lj</td>
</tr>
<tr>
<td></td>
<td>駭客任務 (The Matrix) en</td>
<td>求職 (Job Hunting) lj</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Major category</th>
<th>Sub-category</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer &amp; Network</td>
<td>Security Software</td>
<td>cd</td>
</tr>
<tr>
<td></td>
<td>Software Company</td>
<td>cg</td>
</tr>
<tr>
<td></td>
<td>Hacker/Crack</td>
<td>ck</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Movie</td>
<td>en</td>
</tr>
<tr>
<td>Life Information</td>
<td>Job</td>
<td>lj</td>
</tr>
<tr>
<td>Politics</td>
<td>Local Government</td>
<td>pl</td>
</tr>
<tr>
<td>Travel</td>
<td>Airlines</td>
<td>tp</td>
</tr>
</tbody>
</table>
Clustering Results
Cluster Partition

Figure 4: An illustrative example for cluster partitioning.
Quality Function

Our definition of $QC(C)$ is a product of three components: (a) $F(C)$: A function to measure the cohesion of the clusters; (b) $S(C)$: A function to measure the isolation of the clusters; And (c) $M(C)$: A function to measure whether the number of clusters are proper, i.e., the number of clusters should be neither too few nor too many. Thus the formula of $QC(C)$ is defined as

$$QC(C) = F(C)S(C)M(C).$$
Quality Function (Cont.)

\[ F(C) = \frac{1}{n} \sum_{C_i \in C} n_i f(C_i) \]

\[ f(C_i) = \begin{cases} 
\frac{2}{n_i(n_i-1)} \sum_{v_a, v_b \in C_i \atop v_a \neq v_b} \text{sim}(v_a, v_b), & \text{if } n_i > 1; \\
1 & \text{otherwise.} 
\end{cases} \]

where \( n_i \) is the number of objects contained in \( C_i \) and \( n \) is the total number of objects in cluster set \( C \).

\[ S(C) = \frac{2}{k(k-1)} \sum_{1 \leq i < k} \sum_{i < j \leq k} \text{mindist}(C_i, C_j) \]
Quality Function (Cont.)

\[ M(C) = \sqrt{1 - \frac{(|C| - en)^2}{n^2}} \]

where \( n \) is the total number of objects contained in \( C \) and \( en \) is set as \( \sqrt{n} \).
The Clustering Partition Algorithm

HierarchicalClustering($v_1, v_2, \ldots, v_n$)
$v_i$, 1 ≤ i ≤ n: the vectors of the objects
1: $C_1, C_2, \ldots, C_{2n-1} \leftarrow \text{HAC}(v_1, v_2, \ldots, v_n)$
2: return HierarchicalPartitioning($C_1, C_2, \ldots, C_{2n-1}$)

HierarchicalPartitioning($C_1, C_2, \ldots, C_n, C_{n+1}, \ldots, C_{2n-1}$)
$C_i$, 1 ≤ i ≤ 2n − 1: the binary-tree hierarchy
1: if $n < e$ then
2: return $C_1, C_2, \ldots, C_n$
3: maxqc ← 0, bestcut ← 0
4: for all cut level $l$, 1 ≤ l < n do
5: qc ← QC(clusters($l$))
6: if maxqc < qc then
7: maxqc ← qc, bestcut ← l
8: for all $C_i \in \text{clusters(bestcut)}$ do
9: children($C_i$) ← HierarchicalPartitioning(CH($C_i$))
10: return clusters($l$)
Preliminary Experiment

Test queries
- Two sets: top 1k queries and random 1k queries
- Each of the test queries has been manually assigned according to classes

Evaluation metrics
- F-Measure
query. We then calculate the recall and precision of that cluster for each given class. More specifically, for cluster \( j \) and class \( i \)

\[
\text{Recall}(i, j) = \frac{n_{ij}}{n_i}
\]

\[
\text{Precision}(i, j) = \frac{n_{ij}}{n_j}
\]

where \( n_{ij} \) is the number of members of class \( i \) in cluster \( j \), \( n_j \) is the number of members of cluster \( j \) and \( n_i \) is the number of members of class \( i \).

The F measure of cluster \( j \) and class \( i \) is then given by

\[
F(i, j) = \frac{2 \times \text{Recall}(i, j) \times \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}
\]

For an entire hierarchical clustering the F measure of any class is the maximum value it attains at any node in the tree and an overall value for the F measure is computed by taking the weighted average of all values for the F measure as given by the following.

\[
F = \sum_i \frac{n_i}{n} \max\{F(i, j)\}
\]

where the max is taken over all clusters at all levels, and \( n \) is the number of documents.
Obtained F-Measures

(A) With variant feature sets and distance measures.

<table>
<thead>
<tr>
<th>HF</th>
<th>randterm</th>
<th>freqterm</th>
<th>coreterm</th>
<th>2,3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>.1277</td>
<td>.1181</td>
<td>.1221</td>
<td>.1421</td>
</tr>
<tr>
<td>AL</td>
<td>.5217</td>
<td>.4955</td>
<td>.4921</td>
<td>.4977</td>
</tr>
<tr>
<td>CL</td>
<td>.4757</td>
<td>.4882</td>
<td>.4794</td>
<td>.4390</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RD</th>
<th>randterm</th>
<th>freqterm</th>
<th>coreterm</th>
<th>2,3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>.1092</td>
<td>.1049</td>
<td>.1200</td>
<td>.1295</td>
</tr>
<tr>
<td>AL</td>
<td>.4370</td>
<td>.3795</td>
<td>.4394</td>
<td>.3732</td>
</tr>
<tr>
<td>CL</td>
<td>.4097</td>
<td>.2288</td>
<td>.4043</td>
<td>.3747</td>
</tr>
</tbody>
</table>

(B) With variant size of core-term feature set.

<table>
<thead>
<tr>
<th>HF</th>
<th>100</th>
<th>500</th>
<th>1,000</th>
<th>3,000</th>
<th>5,000</th>
<th>9,751</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>.3443</td>
<td>.4547</td>
<td>.4562</td>
<td>.4634</td>
<td>.4851</td>
<td>.4767</td>
</tr>
<tr>
<td>Major category</td>
<td>Sub-category</td>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer&amp;Network</td>
<td>Security Software</td>
<td>cd</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Software Company</td>
<td>cg</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hacker/Crack</td>
<td>ck</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>Movie</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Information</td>
<td>Job</td>
<td>lj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Politics</td>
<td>Local Government</td>
<td>pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel</td>
<td>Airlines</td>
<td>tp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results of Hierarchical Structure Generation

<table>
<thead>
<tr>
<th>Hierarchy depth</th>
<th>HF</th>
<th>RD</th>
</tr>
</thead>
<tbody>
<tr>
<td># 1st-level clusters</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td># 2nd-level clusters</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td># 3rd-level clusters</td>
<td>164</td>
<td>132</td>
</tr>
<tr>
<td># 4th-level clusters</td>
<td>206</td>
<td>216</td>
</tr>
<tr>
<td># 5th-level clusters</td>
<td>79</td>
<td>93</td>
</tr>
<tr>
<td># 6th-level clusters</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>F-measure</td>
<td>.6048</td>
<td>.5532</td>
</tr>
</tbody>
</table>
VI. Anchor Text Mining
Anchor Text Mining for Query Translation (Lu, ICDM’01)

<table>
<thead>
<tr>
<th>English</th>
<th>Traditional Chinese</th>
<th>Simplified Chinese</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sony</td>
<td>新力</td>
<td>索尼</td>
<td>ソニー</td>
</tr>
<tr>
<td>Nike</td>
<td>耐吉</td>
<td>耐克</td>
<td>ナイキ</td>
</tr>
<tr>
<td>Stanford</td>
<td>史丹佛</td>
<td>斯坦福</td>
<td>スタンフォード</td>
</tr>
<tr>
<td>Sydney</td>
<td>悉尼</td>
<td>互联网</td>
<td>インターネット</td>
</tr>
<tr>
<td>Internet</td>
<td>網際網路</td>
<td>網路</td>
<td>ネットワーク</td>
</tr>
<tr>
<td>Homepage</td>
<td>首頁</td>
<td>頁面</td>
<td>ホームページ</td>
</tr>
<tr>
<td>Computer</td>
<td>電腦</td>
<td>计算机</td>
<td>コンピューター</td>
</tr>
<tr>
<td>Database</td>
<td>資料庫</td>
<td>数据库</td>
<td>データベース</td>
</tr>
<tr>
<td>Information</td>
<td>資訊</td>
<td>信息</td>
<td>インフォメーション</td>
</tr>
</tbody>
</table>
Cross-Language Web Search

- A Web search service allows users to query in one language and search documents that are written or indexed in another language.
Observation of Anchor Text

Taiwan Yahoo! in USA

Observation of Anchor Text


Translation Candidates

Taiwan Yahoo  台灣雅虎  Yahoo  搜尋引擎

Anchor-Text Set

ISCSLP’02  L. F. Chien
Observation of Anchor Text


Page Authority

Taiwan Yahoo 台灣雅虎

(... in-link= 21)

(... in-link= 187)

Co-occurrence

Yahoo 搜尋引擎

in USA
Asymmetric model: \[ P(T_t \mid T_s) = \frac{P(T_s \cap T_t)}{P(T_s)} \]

Symmetric model with link information:

\[
P(T_s \leftrightarrow T_i) = \frac{P(T_s \cap T_i)}{P(T_s \cup T_i)}
\]

\[
\sum_{U_i} P(T_s \cap T_i \mid U_i) P(U_i) \approx \sum_{U_i} [P(T_s \mid U_i) + P(T_i \mid U_i) - P(T_s \mid U_i)P(T_i \mid U_i)] P(U_i)
\]

where \( P(U_i) = \frac{\# \text{in-link of } U_i}{\# \text{total in-link}} \)
Effects of Different Models

Using different models

- $M_A$: Asymmetric model
- $M_{AL}$: Asymmetric model with link information
- $M_S$: Symmetric model
- $M_{SL}^*$: Symmetric model with link information

* Training Data
  -- 109,416 anchor-text sets from 1,980,816 pages

* Test Query Set
  -- 622 English terms from 1,230 most popular English queries

<table>
<thead>
<tr>
<th>Type of Model</th>
<th>Top-1</th>
<th>Top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_A$</td>
<td>41%</td>
<td>81%</td>
</tr>
<tr>
<td>$M_{AL}$</td>
<td>44%</td>
<td>83%</td>
</tr>
<tr>
<td>$M_S$</td>
<td>51%</td>
<td>84%</td>
</tr>
<tr>
<td>$M_{SL}^*$</td>
<td>53%</td>
<td>85%</td>
</tr>
</tbody>
</table>
Performance

Top-n inclusion rates obtained with three different approaches.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Top1</th>
<th>Top2</th>
<th>Top3</th>
<th>Top4</th>
<th>Top5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor Text Mining</td>
<td>57.0%</td>
<td>68.6%</td>
<td>74.3%</td>
<td>77.9%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Dictionary Lookup</td>
<td>30.5%</td>
<td>30.5%</td>
<td>30.5%</td>
<td>30.5%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Combine Anchor Text Mining and Dictionary</td>
<td>74.0%</td>
<td>82.9%</td>
<td>86.7%</td>
<td>88.9%</td>
<td>90.2%</td>
</tr>
</tbody>
</table>
References

Overview

Text Retrieval
- Chen, A. Chinese Text Retrieval without Using a Dictionary, SIGIR’97.

Text Classification

Natural Language Processing
References (Cont.)

- **Term Extraction**

- **Web IR**

- **Speech Retrieval**

- **Information Tracking and Detection**

- **Cross-Language IR**
References

Web Mining

J. Sirvastava & R. Cooley, Mining web data for e-commerce: concepts & applications, PKDD‘01

Conferences & Workshops

Web Content Mining
D. Mladenic et al., Text Mining: What if your data is made of words, PKDD‘01
M. Hearst, Untangling Text Data Mining, ACL‘99.
(Chang et al., 2001) (s.a.) Chapter 6 Handapparat

Web Structure Mining
(Chang et al., 2001) (s.a.) Chapter 7.3 Handapparat
(Chakrabarti, 2000) s.a.

References (Cont.)

- Web Usage Mining
  (Srivastava et al., 2000) s.a.
  Spiliopoulou, M. (2000). Web usage mining for site evaluation: Making a site better fit its users. Special Section of the Communications of ACM on "Personalization Technologies with DataMining", 43(8), 127-134. Handapparat ACM Digital Library


- For more references can refer at http://www.wiwi.hu-berlin.de/~berendt/lehre/2001w/wmi/literature.html
References (Cont.)

Text and Web page categorization

Web page classification applications

Link and context analysis
- S. Brin and L. Page. The anatomy of large-scale hypertextual web search engine, WWW’98.
References  (Works in Academia Sinica)


Q&A

- Thanks!