Knowledge Management Systems: Development and Applications
Part II: Techniques and Examples

Hsinchun Chen, Ph.D.
McClelland Professor,
Director, Artificial Intelligence Lab and Hoffman E-Commerce Lab
The University of Arizona
Founder, Knowledge Computing Corporation

Acknowledgement: NSF DLI1, DLI2, NSDL, DG, ITR, IDM, CSS, NIH/NLM, NCI, NIJ, CIA, NCSA, HP, SAP

美國亞歷桑那大學, 陳炘鈞 博士
Discovering and Managing Knowledge: Text/Web Mining and Digital Library
Knowledge

Views

- Hierarchical view (Nunamaker et al., 2001)
- Reversed hierarchy (Tuomi, 1999)
- As a state of mind, an object, a process, access to information, and a capability (Alavi and Leidner, 2001)
- Resource-based theory (Barney, 1991; Penrose, 1959; Wernerfelt, 1984; Drucker, 1995)

Classifications

- Tacit and explicit dimensions (Polanyi, 1965)
- Individual vs. collective knowledge
- Declarative vs. procedural knowledge
- Causal, conditional, relational and pragmatic knowledge

- Revealed underlying assumptions in KM
- Implied different roles of knowledge in organizations
- **Textual knowledge** - Most efficient way to store, retrieve, and transfer vast amount of information
- Advanced processing needed to obtain knowledge
  - Traditionally done by humans
  - It is useful to review the discipline of **Human-Computer Interaction** to understand human analysis needs
HCI Evolution and Milestones

**Early Developments**
- 1967 - Neisser's Cognitive Psychology
- 1977 - Lindsay & Norman, Human Information Processing

**AI and Cognition**
- 1981 - Simon: AI and cognitive science
- 1983 - Card, Moran & Newell: GOMS
- 1983 - Gentner and Stevens: Mental models
- 1983 - Anderson: experimental studies and models of cognitive processes in memory and learning

**Social Context**
- 1991 - Norman: Cognitive Artifact Analysis
- 1991 - Bødker: A Human Activity Approach to User Interface Design
- 1991 - Caroll & Rosson: Social and behavioral theory in design
- 1992 - Kuttii: Activity Theory in CSCW

**Software psychology (1960-70s)**
- 1986 - Norman: Model of task performance and stages of user activities
- 1986 - Minsky: study of mind and AI
- 1989 - Ellis: Model of academic information seeking patterns

**AI, Cognitive Modeling (1980s)**
- 1992 - Spink: Model of search process and user interaction
- 1998 - Kuhlthau: Six-stage model of the information seeking process
- 1998 - Sutcliff & Ennis: Process model of information searching
- 2000 - Choo, Detlor, Tumbull: Information Seeking on the Web

**Collaboration, Info. Seeking (1990s-Now)**

**User Modeling**

**Process Model**
Evolution of Automatic Text Processing

Early Foundations
- 1971: Salton: Vector Space model
- 1972: Sparck Jones: Statistical interpretation of term specificity
- 1975: Salton: Theory of Text Indexing
- 1979: van Rijsbergen: Probabilistic model

From Machine Learning to Data Mining
- 1992: Turtle & Croft: Inference network for IR
- 1993: Hearst: TextTiling for summarization
- 1995: Vapnik: Statistical learning theory

Information Retrieval Techniques (~1970s)
- 1981: Preece: Spreading Activation for IR
- 1983: Simon: Machine learning
- 1987: Fumas: Vocabulary problem
- 1989: Salton: Automatic Text Processing
- 1989: Church: Mutual information

Machine learning, AI, Text Clustering (1980s)

Web mining, Text summarization, Visualization (1990s-Now)
- 1996: Shneiderman: Taxonomy of Information Visualization
- 1997: Etzioni: Web mining
- 1998: Chen: Internet searching and browsing
- 1998: Kleinberg: HITS for Web structure mining
- 1998: Joachim: Text classification with SVM
- 2003: Chau & Chen: Vertical spidering algorithms

AI, Probabilistic Techniques

Web mining
Text Mining: Intersection of IR and AI

Information Retrieval (IR) and Gerald Salton

- Inverted Index, Boolean, and Probabilistic, 1970s
- Expert Systems, User Modeling and Natural Language Processing, 1980s
- Machine Learning for Information Retrieval, 1990s
- Search Engines and Digital Libraries, late 1990s and 2000s
Text Mining: Intersection of IR and AI

Artificial Intelligence (AI) and Herbert Simon

• General Problem Solvers, 1970s
• Expert Systems, 1980s
• Machine Learning and Data Mining, 1990s
• Agents, Network/Graph Learning, late 1990s and 2000s
Representing Knowledge

• IR Approach
  • Indexing and Subject Headings
  • Dictionaries, Thesauri, and Classification Schemes

• AI Approach
  • Cognitive Modeling
  • Semantic Networks, Production Systems, Logic, Frames, and Ontologies
• For Web Mining:
  – **Web mining techniques**: resource discovery on the Web, information extraction from Web resources, and uncovering general patterns (Etzioni, 1996)
    • Pattern extraction, meta searching, spidering
  – **Web page summarization** (Hearst, 1994; McDonald & Chen, 2002)
  – **Web page classification** (Glover et al., 2002; Lee et al., 2002; Kwon & Lee, 2003)
  – **Web page clustering** (Roussinov & Chen, 2001; Chen et al., 1998; Jain & Dube, 1988)
  – **Web page visualization** (Yang et al., 2003; Spence, 2001; Shneiderman, 1996)
An Automatic Text Mining Framework for Knowledge Discovery on the Web
Text Mining Techniques:

• Linguistic analysis/NLP: identify key concepts (who/what/where…)
• Statistical/co-occurrence analysis: create automatic thesaurus, link analysis
• Statistical and neural networks clustering/categorization: identify similar documents/users/communities and create knowledge maps
• Visualization and HCI: tree/network, 1/2/3D, zooming/detail-in-context
Text Mining Techniques: Linguistic Analysis

- Word and inverted index: stemming, suffixes, morphological analysis, Boolean, proximity, range, fuzzy search
- Phrasal analysis: noun phrases, verb phrases, entity extraction, mutual information
- Sentence-level analysis: context-free grammar, transformational grammar
- Semantic analysis: semantic grammar, case-based reasoning, frame/script
Natural Language Processing

- Text Tokenization
- Part-of-speech-tagging
- Noun phrase generation

Interactive navigation using semantic indexing enables information retrieval better than previously possible for diverse large collections.

# of phrase from cgi..4..

Legend:
- N - Noun
- A - Adjective
- P - Preposition

4 phrases

Interactive navigation -> A N
semantic indexing -> A N
information retrieval -> N N
diverse large collections -> A A N
Text Mining Techniques: Statistical/Co-Occurrence Analysis

- Similarity functions: Jaccard, Cosine
- Weighting heuristics
- Bi-gram, tri-gram, N-gram
- Finite State Automata (FSA)
- Dictionaries and thesauri
Co-occurrence analysis

- Heuristic term weighting
- Weighted co-occurrence analysis

1. **TP information retrieval**

Your request found 40 relevant terms. (40 terms per page)

1. 100% PM Dr. J. Allen Sears (1)
2. 77% TP IR system (1)
3. 77% TP information retrieval engine (1)
4. 77% TP speech collection (1)
5. 72% PI John Makhoul (1)
6. 68% TP widespread use (1)
7. 47% PI Bruce Schatz (1)
8. 47% PI Donna Harman (1)
9. 47% PI Kui-Lam Kwok (1)
10. 47% PI Mitch Marcus (1)
11. 47% PI Nicholas Belkin (1)
12. 47% PI Phil Hayes (1)
13. 47% PI Robert Moore (1)
14. 42% TP BBN BYBLOS speech recognition system (1)
15. 42% TP CommandTalk (1)
Text Mining Techniques: Clustering/Categorization

- Hierarchical clustering: single-link, multi-link, Ward’s
- Statistical clustering: multi-dimensional scaling (MDS), factor analysis
- **Neural network clustering: self-organizing map (SOM)**
- Ontologies: directories, classification schemes
Neural Network Analysis

- Document clustering
- Category labeling
- Optimization and parallelization

Automatic Generation of CL: Foundation from NSF/DARPA/NASA Digital Library Initiative-1
KMS Techniques: Visualization/HCI

- Structures: trees/hierarchies, networks
- Dimensions: 1D, 2D, 2.5D, 3D, N-D (glyphs)
- Interactions: zooming, spotlight, fisheye views, fractal views
Advanced Visualization

- 1D, 2D, 3D
Advanced Techniques

Automatic Generation of CL: (Continued)

- Entity Extraction and Co-reference based on TREC and MUG
- Text segmentation and summarization
- Visualization techniques and HCI
Integration of CL:

- Ontology-enhanced query expansion (e.g., WordNet, UMLS Metathesaurus)
- Ontology-enhanced semantic tagging (e.g., UMLS Semantic Nets)
- Spreading-activation based term suggestion (e.g., Hopfield net)
YAHOO vs. OOHAY:

- YAHOO: manual, high-precision
- OOHAY: automatic, high-recall
- Acknowledgements: NSF, NIH, NLM, NIJ, DARPA
Knowledge Computing Approach

From YAHOO! To OOHAY?

Object Oriented Hierarchical Automatic Yellowpage

OOHAY Y
Text and Web Mining in Digital Libraries:
AI Lab Research Prototypes
An Automatic Text Mining Framework for Knowledge Discovery on the Web
Web Analysis (1M):

Web pages, spidering, noun phrasing, categorization
Arizona DLI-2 project: "From Interspace to OOHAY?"

Research goal:
automatic and dynamic categorization and visualization of all web pages in the US (and the world, later)

Technologies:
- OOHAY techniques
- Multi-threaded spiders for web page collection
- High-precision web page noun phrasing and entity identification
- Multi-layered, parallel, automatic web page topic directory/hierarchy generation
- Dynamic web search result summarization and visualization
- Adaptive, 3D web-based visualization

Research Status
**OOHAY: Visualizing the Web**

**URL's: *** +ROCK (184) *** - Netscape**

**Related Terms:** ROCK |

- [http://www.icom.ca/cgpinc](http://www.icom.ca/cgpinc)
  **Title:** CG Publishing Inc Home Page
  **Keywords:** CG | COLLECTIBLES | ZEPPELIN | ROCK | ULTIMATE BOOTS | ZEPPELIN LIVE | STAR WARS COLLECTIBLES | WARS COLLECTIBLES

- [http://www.microserve.net/~xpander/alaskatp.html](http://www.microserve.net/~xpander/alaskatp.html)
  **Title:** ALASKA
  **Keywords:** ALASKA | PRODUCE MUSIC | POWERFUL ROCK | SYMPHONY | MEMORABLE QUOTES | SONGS LIVE | INNOCENT BYSTANDER | KEITH COPLAND | GENTLE GIANT

- [http://www4.nando.net/music/gm/Rockit](http://www4.nando.net/music/gm/Rockit)
  **Title:** Rock-it Comix
  **Keywords:** ROCK-IT COMIX | MEGADETH | ROLL | COMIC MAGAZINES
Lessons Learned:

- Web pages are noisy: need filtering
- Spidering needs help: domain lexicons, multi-threads
- SOM is computational feasible for large-scale application
- SOM performance for web pages = 50%
- Web knowledge map (directory) is interesting for browsing, not for searching
- Techniques applicable to Intranet and marketing intelligence
News Classification (1M):

Chinese news content, mutual information indexing, PAT tree, categorization
記者王家英/台北

便利商店電子商務開步走

《經濟日報》

春節剛過，便利商店業積極跨足電子商務，繼福客多推出網路購物服務，統一超商18日起也將推出網路買CD門市取貨付款服務，萊爾富16日則先以型錄購物為電子商務測試暖身。泰山集團的福客多昨（11）日起與大同集團合作推出網路購物服務，在國內便利商店介入電子商務拓得頭籌。4月福客多擬將網路品牌egrocer獨立經營，並與好鄰居生活館串聯。統一超商籌備多時的網路購物也將在18日跨出第一步，與霖揚公司合作推出網路試聽購買CD，全省2,300家門市取貨付款的服務，消費者線上訂貨三天後可就近在7-Eleven門市付款取貨。此外統一超商與博客來網路書店洽談的書籍電子商務，也進入最後敲定階段。萊爾富便利商店則擬採取三三階段策略，逐步介入電子
近年來因經濟的起飛，人口結構的老化，加上生活型態的改變，西式飲食的盛行，導致台灣地區之大腸直腸癌發生率及死亡率節節高揚。就死亡率而言，大腸直腸癌目前已是台灣地區因惡性腫瘤死亡人口的第三位，僅次於肝癌及肺癌。不論在我國或先進國家，大腸直腸癌已是今日公共衛生重要的一環。近年來有關大腸直腸癌的流行病學研究甚多，其中較具體的結論是遺傳與飲食。我們大概可以說家族一等親中若有人得到大腸直腸癌，則其一生中得到相同癌症的機會約為一般人的三倍。目前公認纖維質食物攝取太少，以及攝取太多的肉類，由於會導致大便通過大腸的平均時間長，所以致癌的機會也會大增。就大腸直腸癌病變而言，
<table>
<thead>
<tr>
<th>相關詞</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   巴勒斯坦</td>
<td></td>
</tr>
<tr>
<td>□   柯林頓</td>
<td></td>
</tr>
<tr>
<td>□   約旦</td>
<td></td>
</tr>
<tr>
<td>□   國家</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>相關分類</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   以色列</td>
<td></td>
</tr>
<tr>
<td>□   市長陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   要求陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   陳水扁方面</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>相關詞</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   巴勒斯坦</td>
<td></td>
</tr>
<tr>
<td>□   柯林頓</td>
<td></td>
</tr>
<tr>
<td>□   約旦</td>
<td></td>
</tr>
<tr>
<td>□   國家</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>相關分類</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   以色列</td>
<td></td>
</tr>
<tr>
<td>□   市長陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   要求陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   陳水扁方面</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>相關詞</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   巴勒斯坦</td>
<td></td>
</tr>
<tr>
<td>□   柯林頓</td>
<td></td>
</tr>
<tr>
<td>□   約旦</td>
<td></td>
</tr>
<tr>
<td>□   國家</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>相關分類</th>
<th>說明</th>
</tr>
</thead>
<tbody>
<tr>
<td>□   以色列</td>
<td></td>
</tr>
<tr>
<td>□   市長陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   要求陳水扁</td>
<td></td>
</tr>
<tr>
<td>□   陳水扁方面</td>
<td></td>
</tr>
</tbody>
</table>
Lessons Learned:

• News readers are not knowledge workers
• News articles are professionally written and precise.
• SOM performance for news articles = 85%
• Statistical indexing techniques perform well for Chinese documents
• Corporate users may need multiple sources and dynamic search help
• Techniques applicable to eCommerce (eCatalogs) and ePortal
Personal Agents (1K):

Web spidering, meta searching, noun phrasing, dynamic categorization
1. Enter Starting URLs and Key Phrases to be searched

2. Search results from spiders are displayed dynamically
1. Enter Starting URLs and Key Phrases to be searched

2. Search results from spiders are displayed dynamically

For project information and free download: http://ai.bpa.arizona.edu

OOHAY: CI Spider, Meta Spider, Med Spider
OOHAY: Meta Spider, News Spider, Cancer Spider

For project information and free download: http://ai.bpa.arizona.edu
3. Noun Phrases are extracted from the web ages and user can selected preferred phrases for further summarization.

4. SOM is generated based on the phrases selected. Steps 3 and 4 can be done in iterations to refine the results.

For project information and free download: http://ai.bpa.arizona.edu

OOHAY: CI Spider, Meta Spider, Med Spider
Lessons Learned:

- Meta spidering is useful for information consolidation
- Noun phrasing is useful for topic classification (dynamic folders)
- SOM usefulness is suspect for small collections
- Knowledge workers like personalization, client searching, and collaborative information sharing
- Corporate users need multiple sources and dynamic search help
- Techniques applicable to marketing and competitive analyses
CRM Data Analysis (5K):

Call center Q/A, noun phrasing, dynamic categorization, problem analysis, agent assistance
Category: Express Exchange

Your request found 47 relevant documents.

1. **Subject:** Customer’s DJ 310 failed during warranty. Dealer charged $47 to diagnose problem.
   **CustomerWords:** Customer’s DJ 310 failed during warranty. Dealer charged $47 to diagnose problem before customer could confirm his purchase date. Please express exchange this customer’s unit. Printer has been down since February. Warranty purchase date was 2-1-94.
   **ResolutionText:** Leasa N. called customer and arranged to send EE & extend warranty. Also, told cust. that she will send him 2 cartridges for his inconvenience and the fact that the

2. **Subject:** Customer would like to do an Express Exchange. His unit is still under warranty.
   **CustomerWords:** Customer would like to do an Express Exchange. His unit is still under warranty (3 yrs.). Please call--thanks.
   **ResolutionText:** The customer had sent his unit into CrSC for white lines running through the text. CrSC sent the printer back and said that the paper path had been cleared. The print quality problem was obviously still there. The customer is extremely upset with our repair quality and the fact that we don’t have an 800#. He will never buy HP again. Dick Miller arranged for an Express Exchange through CrSC. CS0# L4812-02.
Lessons Learned:

• Call center data are noisy: typos and errors
• Noun phrasing useful for Q/A classification
• Q/A classification could identify problem areas
• Q/A classification could improve agent productivity: email, online chat, and VoIP
• Q/A classification could improve new agent training
• Techniques applicable to virtual call center and CRM applications
Nano Patent Mapping (100K):

Nano patents, content/network analysis and visualization, impact analysis
Data: U.S. NSE Patents

- Top assignee countries and institutions

<table>
<thead>
<tr>
<th>Rank</th>
<th>Assignee Country</th>
<th>Number of Patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>37760</td>
</tr>
<tr>
<td>2</td>
<td>Japan</td>
<td>5637</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>5214</td>
</tr>
<tr>
<td>4</td>
<td>France</td>
<td>1617</td>
</tr>
<tr>
<td>5</td>
<td>Canada</td>
<td>1528</td>
</tr>
<tr>
<td>6</td>
<td>United Kingdom</td>
<td>709</td>
</tr>
<tr>
<td>7</td>
<td>Switzerland</td>
<td>370</td>
</tr>
<tr>
<td>8</td>
<td>Israel</td>
<td>326</td>
</tr>
<tr>
<td>9</td>
<td>China (Taiwan)</td>
<td>307</td>
</tr>
<tr>
<td>10</td>
<td>Italy</td>
<td>307</td>
</tr>
<tr>
<td>11</td>
<td>Australia</td>
<td>287</td>
</tr>
<tr>
<td>12</td>
<td>Netherlands</td>
<td>268</td>
</tr>
<tr>
<td>13</td>
<td>Republic of Korea</td>
<td>247</td>
</tr>
<tr>
<td>14</td>
<td>Sweden</td>
<td>195</td>
</tr>
<tr>
<td>15</td>
<td>Belgium</td>
<td>176</td>
</tr>
<tr>
<td>16</td>
<td>Denmark</td>
<td>99</td>
</tr>
<tr>
<td>17</td>
<td>Finland</td>
<td>81</td>
</tr>
<tr>
<td>18</td>
<td>Norway</td>
<td>55</td>
</tr>
<tr>
<td>19</td>
<td>Singapore</td>
<td>48</td>
</tr>
<tr>
<td>20</td>
<td>Austria</td>
<td>43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Assignee Name</th>
<th>Number of Patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>International Business Machines Corporation</td>
<td>1302</td>
</tr>
<tr>
<td>2</td>
<td>Xerox Corporation</td>
<td>957</td>
</tr>
<tr>
<td>3</td>
<td>Minnesota Mining and Manufacturing Company</td>
<td>807</td>
</tr>
<tr>
<td>4</td>
<td>Eastman Kodak Company</td>
<td>708</td>
</tr>
<tr>
<td>5</td>
<td>Motorola, Inc.</td>
<td>508</td>
</tr>
<tr>
<td>6</td>
<td>The Regents of the University of California</td>
<td>491</td>
</tr>
<tr>
<td>7</td>
<td>NEC Corporation</td>
<td>483</td>
</tr>
<tr>
<td>8</td>
<td>Micron Technology, Inc.</td>
<td>457</td>
</tr>
<tr>
<td>9</td>
<td>Canon Kabushiki Kaisha</td>
<td>408</td>
</tr>
<tr>
<td>10</td>
<td>E. I. Du Pont de Nemours and Company</td>
<td>367</td>
</tr>
<tr>
<td>11</td>
<td>General Electric Company</td>
<td>367</td>
</tr>
<tr>
<td>12</td>
<td>Texas Instruments Incorporated</td>
<td>366</td>
</tr>
<tr>
<td>13</td>
<td>Hitachi, Ltd.</td>
<td>335</td>
</tr>
<tr>
<td>14</td>
<td>The United States of America as represented by the Secretary of the Navy</td>
<td>330</td>
</tr>
<tr>
<td>15</td>
<td>The Dow Chemical Company</td>
<td>327</td>
</tr>
<tr>
<td>16</td>
<td>Kabushiki Kaisha Toshiba</td>
<td>317</td>
</tr>
<tr>
<td>17</td>
<td>Abbott Laboratories</td>
<td>297</td>
</tr>
<tr>
<td>18</td>
<td>Advanced Micro Devices, Inc.</td>
<td>295</td>
</tr>
<tr>
<td>19</td>
<td>Massachusetts Institute of Technology</td>
<td>271</td>
</tr>
<tr>
<td>20</td>
<td>Merck &amp; Co., Inc.</td>
<td>251</td>
</tr>
</tbody>
</table>
Top technology fields (US Patent Classification first-level categories)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Technology Field</th>
<th>Number of Patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>435: Chemistry: molecular biology and microbiology</td>
<td>7837</td>
</tr>
<tr>
<td>2</td>
<td>514: Drug, bio-affecting and body treating compositions</td>
<td>6364</td>
</tr>
<tr>
<td>3</td>
<td>424: Drug, bio-affecting and body treating compositions</td>
<td>4760</td>
</tr>
<tr>
<td>4</td>
<td>428: Stock material or miscellaneous articles</td>
<td>3847</td>
</tr>
<tr>
<td>5</td>
<td>250: Radiant energy</td>
<td>3783</td>
</tr>
<tr>
<td>6</td>
<td>530: Chemistry: natural resins or derivatives; peptides or proteins; lignins or reaction</td>
<td>3772</td>
</tr>
<tr>
<td>7</td>
<td>536: Organic compounds -- part of the class 532-570 series</td>
<td>3701</td>
</tr>
<tr>
<td>8</td>
<td>438: Semiconductor device manufacturing: process</td>
<td>3584</td>
</tr>
<tr>
<td>9</td>
<td>257: Active solid-state devices (e.g., transistors, solid-state diodes)</td>
<td>3480</td>
</tr>
<tr>
<td>10</td>
<td>427: Coating processes</td>
<td>3179</td>
</tr>
<tr>
<td>11</td>
<td>436: Chemistry: analytical and immunological testing</td>
<td>2941</td>
</tr>
<tr>
<td>12</td>
<td>430: Radiation imagery chemistry: process, composition, or product thereof</td>
<td>2883</td>
</tr>
<tr>
<td>13</td>
<td>359: Optics: systems (including communication) and elements</td>
<td>2743</td>
</tr>
<tr>
<td>14</td>
<td>356: Optics: measuring and testing</td>
<td>2556</td>
</tr>
<tr>
<td>15</td>
<td>422: Chemical apparatus and process disinfecting, deodorizing, preserving, or sterilizing</td>
<td>1665</td>
</tr>
<tr>
<td>16</td>
<td>204: Chemistry: electrical and wave energy</td>
<td>1660</td>
</tr>
<tr>
<td>17</td>
<td>252: Compositions</td>
<td>1647</td>
</tr>
<tr>
<td>18</td>
<td>524: Synthetic resins or natural rubbers -- part of the class 520 series</td>
<td>1515</td>
</tr>
<tr>
<td>19</td>
<td>546: Organic compounds -- part of the class 532-570 series</td>
<td>1503</td>
</tr>
<tr>
<td>20</td>
<td>210: Liquid purification or separation</td>
<td>1451</td>
</tr>
</tbody>
</table>
Content Map Analysis

- NSE Grant Content Map (1991 – 1995)
Content Map Analysis

- NSE Grant Content Map (1996 – 2000)

Region color indicates the growth rate of the associated technology topic. The number associated with the colors were the actual growth rate: # of grants/patents during 1991-1995 / # of grants/patents during 1996-2000 for a particular topic (region). Regions with comparable growth rate as the entire field were assigned the green color.
Sample Patent Citation Networks

Backbone citation network for the field "Chemistry: molecular biology and microbiology (all patents shown were cited by more than five times)

PI-inventors and their patents form a closely linked cluster within the largest connected component of the backbone citation network.
H1.1 Patent – Number of Cites

- H1.1 supported: PI-inventors’ patents had significantly higher number of cites measure than most other comparison groups (except IBM)
- Order of the groups: NSF, IBM > Top10, UC, US > EntireSet, Japan > European, Others
H2.1 Inventor – Number of Cites

**H2.1 supported:** PI-inventors had significantly higher number of cites measure than most other comparison groups.

**Order of the groups:** NSF > Top10, Japan, EntireSet, US, IBM > UC, European, Others

**Japanese inventors** had high number of cites measure despite the small number of cites for each patent they file.

### Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF</td>
<td>277</td>
<td>5.74</td>
<td>15.78</td>
<td>(------*-----)</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>277</td>
<td>3.10</td>
<td>6.93</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>Top10</td>
<td>277</td>
<td>4.20</td>
<td>13.20</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>277</td>
<td>1.69</td>
<td>3.85</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>277</td>
<td>2.99</td>
<td>12.48</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>EntireSet</td>
<td>277</td>
<td>3.49</td>
<td>14.07</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>277</td>
<td>3.92</td>
<td>14.39</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>European</td>
<td>277</td>
<td>1.45</td>
<td>7.23</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>277</td>
<td>1.01</td>
<td>3.94</td>
<td>(-----*-----)</td>
<td></td>
</tr>
<tr>
<td>Pooled StDev = 11.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Pooled StDev = 11.13**

0.0 2.0 4.0 6.0
Lessons Learned:

• Units of analysis: inventors, institutions, and countries
• USPTO patents are clean and comprehensive
• Content and network analyses help reveal trends and key innovations/inventors
• Patent analyses help with impact study
Newsgroup Categorization (1K):

Workgroup communication, noun phrasing, dynamic categorization, glyphs visualization
Disadvantages:

- No sub-topic identification
- Difficult to identify experts
- Difficult to learn participants’ attitude toward the community
Thread Representation

- Time
- Message
- Length of Time
- Person
People Representation

- Time
- Message
- Length of Time
- Thread
Visual Effects:

• Thickness = how active a subtopic is

• Length in x-dimension = the time duration of a sub-topic
Proposed Interface (Interaction Summary)

Visual Effects:

- Healthy sub-garden with many blooming high flowers = popular active sub-topic
- A long, blooming flower is a healthy thread
Proposed Interface (Expert Indicator)

Visual Effects:

• Healthy sub-garden with many blooming high flowers = popular sub-topic

• A long, blooming people flower is a recognized expert.
Lessons Learned:

• P1000: A picture is indeed worth 1000 words
• Expert identification is critical for KM support
• Glyphs are powerful for capturing multi-dimensional data
• Techniques applicable to collaborative applications, e.g., email, online chats, newsgroup, and such
GIS Multimedia Data Mining (10GBs):

Geoscience data, texture image indexing, multimedia content
Airphoto analysis: Texture (Gabor filter)
AVHRR satellite data: Temperature/vegetation
Lessons Learned:

- Image analysis techniques are application dependent (unlike text analysis)
- Image killer apps not found yet
- Multimedia applications require integration of data, text, and image mining techniques
- Multimedia KMS not ready for prime-time consumption yet
Knowledge Management Systems: Future
Other Emerging Categorization Challenges/Opportunities:

- Multilingual terminology and semantic issues
- Web analysis and categorization issues
- E-Commerce information (transactions) classification issues
- Multimedia content and wireless delivery issues
- Future: semantic web, multilingual web, multimedia web, wireless web!
The Road Ahead

• The Semantic Web: XML, RDF, Ontologies
• The Wireless Web: WML, WIFI, display
• The Multimedia Web: content indexing and analysis
• The Multilingual Web: cross-lingual MT and IR
For Project Information at AI Lab:

• http://ai.arizona.edu

• hchen@eller.arizona.edu