Beyond Content-Based Retrieval: Modeling Domains, Users, and Interaction

Susan Dumais
Microsoft Research
http://research.microsoft.com/~sdumais

Microsoft Research (http://research.microsoft.com)
- Adaptive Systems and Interaction - IR/UI
- Machine Learning and Applied Statistics
- Data Mining
- Natural Language Processing
- Collaboration and Education
- Database
- MSR Cambridge; MSR Beijing

Microsoft Product Groups ... many IR-related
IR Themes & Directions

- Improvements in matching algorithms and representation
  - Probabilistic/Bayesian models
    - $p(\text{Relevant} | \text{Document})$, $p(\text{Concept} | \text{Words})$
  - NLP: Truffle, MindNet

- Beyond content-matching
  - User/Task modeling
  - Domain/Object modeling
  - Advances in presentation

WWW8 Panel
Finding Anything in the Billion-Page Web: Are Algorithms the Answer?
Moderator: Prabhakar Raghavan, IBM Almaden
Traditional View of IR

Query Words

Ranked List

[Diagram showing a process from query words to a ranked list]
What's Missing?

User Modeling

Query Words

Information Use

Ranked List

Domain Modeling
Domain/Obj Modeling

- Not all objects are equal ... potentially big win
  - A priori importance
  - Information use ("readware"; collab filtering)
- Inter-object relationships
  - Link structure / hypertext
  - Subject categories - e.g., text categorization, text clustering
- Metadata
  - E.g., reliability, recency, cost -> combining
User/Task Modeling

- Demographics
- Task -- What’s the user’s goal?
  - e.g., Lumiere
- Short and long-term content interests
  - e.g., Implicit queries
    - Interest model = f(content_similarity, time, interest)
  - e.g., Letiza, WebWatcher, Fab
Information Use

- Beyond batch IR model ("query->results")
  - Consider larger task context
- Human attention is critical resource ... no Moore’s Law for human capabilities
  - Techniques for automatic information summarization, organization, discover, filtering, mining, etc.
- Advanced UIs and interaction techniques
  - E.g, tight coupling of search, browsing to support information management
The Broader View of IR

User Modeling

Query Words

Information Use

Domain Modeling

Ranked List
Beyond Content Matching

- **Domain/Object modeling**
  - A priori importance
  - Text classification and clustering

- **User/Task modeling**
  - Implicit queries and Lumiere

- **Advances in presentation and manipulation**
  - Combining structure and search (e.g., DM)
Example:
Web query for "Microsoft Research"
Microsoft
The RealNames link takes you directly to Microsoft.

1. Microsoft Research Advanced Programming Languages
   URL: research.microsoft.com/apl/main.htm
   Last modified 16-Dec-98 - page size 8K - in English (Win-1252) [ Translate ]

2. Dr. Kai-Fu Lee Joins Microsoft Research
   Industry Pioneer Is Now Part of Microsoft's Growing Team of Computer...
   Last modified 7-Apr-99 - page size 12K - in English [ Translate ]

   Microsoft researcher David Heckerman combines probability theory with artificial intelligence to make computers...
   URL: www.microsoft.com/presspass/features/1999...22heckerman.htm
   Last modified 7-Apr-99 - page size 22K - in English [ Translate ]

4. James Larue's Home Page at Microsoft Research
   ...
   URL: research.microsoft.com/~larus/
   Last modified 22-Jan-99 - page size 794 bytes - in English [ Translate ]

5. Sarah Boyd, Microsoft Research Institute, Macquarie University, Australia
   Sarah Boyd. Name. Sarah Boyd. Position. PhD student (MRI) Postal Address. MS Research Institute Macquarie University 2109 AUSTRALIA. Telephone Number....
   URL: dishwasher1.mpce.mq.edu.au:8000/Rcn%3dSa...rsity,%20c%3dAU
   Last modified 4-May-97 - page size 1K - in English [ Translate ]

6. About the Microsoft Research Institute
   About the Microsoft Research Institute. The Microsoft Research Institute (MRI), an independent research centre situated at Macquarie University in Sydney..
   URL: www.mri.mq.edu.au/common/about.html
   Last modified 21-Mar-99 - page size 3K - in English [ Translate ]

7. Microsoft Research: Researcher Profiles

8. Microsoft Research: Publications

9. Microsoft Research: Projects

10. Microsoft Research: People

11. Microsoft Research: Overview

12. Microsoft Research: Advanced Programming Languages


14. James Larue's Home Page at Microsoft Research

15. Sarah Boyd, Microsoft Research Institute, Macquarie University, Australia

16. About the Microsoft Research Institute

17. Microsoft Research: Researcher Profiles

18. Microsoft Research: Publications

19. Microsoft Research: Projects

20. Microsoft Research: People

21. Microsoft Research: Overview

22. Microsoft Research: Advanced Programming Languages


24. James Larue's Home Page at Microsoft Research

25. Sarah Boyd, Microsoft Research Institute, Macquarie University, Australia

26. About the Microsoft Research Institute

27. Microsoft Research: Researcher Profiles

28. Microsoft Research: Publications

29. Microsoft Research: Projects

30. Microsoft Research: People

31. Microsoft Research: Overview

32. Microsoft Research: Advanced Programming Languages

33. Microsoft Research: Making Computers More Intelligent and Responsive

34. James Larue's Home Page at Microsoft Research

35. Sarah Boyd, Microsoft Research Institute, Macquarie University, Australia

36. About the Microsoft Research Institute

37. Microsoft Research: Researcher Profiles

38. Microsoft Research: Publications

39. Microsoft Research: Projects

40. Microsoft Research: People

41. Microsoft Research: Overview

42. Microsoft Research: Advanced Programming Languages

43. Microsoft Research: Making Computers More Intelligent and Responsive

44. James Larue's Home Page at Microsoft Research

45. Sarah Boyd, Microsoft Research Institute, Macquarie University, Australia

46. About the Microsoft Research Institute
**Microsoft Research**

100% Date: 5 Feb 1999, Size 6.8K, http://www.research.microsoft.com/

**Find similar pages | Grouped results from www.research.microsoft.com**

**The Microsoft Research Institute**

General Information Research Information Special Interest Groups March 1998

**Find similar pages**

**Dr. Kai-Fu Lee Joins Microsoft Research**

Industry Pioneer Is Now Part of Microsoft's Growing Team of Computer Scientists

**Find similar pages | Grouped results from www.microsoft.com**

**Microsoft Research: People**

Researcher Profiles It's our team of researchers that makes Microsoft Research so successful. Learn a little more about the people behind the technology in the interviews that follow. of the Speech Technology Group ...

**Find similar pages | Grouped results from research.microsoft.com**

**Microsoft Research: R&D without ivory tower**

Front page, News, Sports, Money, Life, Weather, Marketplace Inside Tech FAQ, Tips and Tricks Web Column Hot Sites Tech News Tech Reviews Answer Desk Shareware ...

**Find similar pages**

**Working Group Members**

University of Washington and MathSoft, Inc., University of Minnesota and Talaria, Inc., INRIA Rhone-Alpes, Grenoble, France, University of Ulster, Northern Ireland, and Université Louis Pasteur, ...

**Find similar pages**

**Microsoft Research Centre in Cambridge - and Sony**

Over the weekend of 31-May/1-June 1997 the media carried stories about a planned £50M Microsoft research centre in Cambridge. The reports said that Bill Gates intends to create a ...
70% Date: 18 Nov 1998, Size 5.4K, http://bridge.anglia.ac.uk/~systmlk/histrom/cam/Microsoft.html

**Find similar pages**
Documents that best match your search

1. **Bibliography of the technical reports of Microsoft Research**
   
   89% - **Directories & Lists**: Bibliography of the technical reports of Microsoft Research
   
   The Collection of Computer Science
   
   Bibliographies
   
   Up: Bibliography of the technical reports of Microsoft Research
   
   South Korean site: http://cosmos.kaist.ac.kr/pub/bibliographies/Techreports/Microsoft.Research.html

2. **Microsoft Research Gateway Page**
   
   89% - **Articles & General info**: Microsoft Research Gateway Page
   
   Microsoft Research You should be automatically redirected to the official Microsoft Research web site. If you are not, please click here... 09/14/98
   
   Commercial site: http://microsoft.com/research/

3. **(Ping 543) Re: Microsoft Research IPv6 Release**
   
   89% - **Articles & General info**: (Ping 543) Re: Microsoft Research IPv6 Release
   
   (Ping 543) Re: Microsoft Research IPv6 Release
   
   To: Richard Draves < > Subject: (Ping 543) Re: Microsoft Research IPv6 Release
   
   03/24/98
   
   Educational site: http://www.wcuq.wvu.edu/lists/ipng/199803/msg00167.html

4. **Cambridge Network: Microsoft Research Ltd: Profile**
   
   85% - **Articles & General info**: Cambridge Network: Microsoft Research Ltd: Profile
   
   <fr > <fr > <fr > <fr > Microsoft Research Ltd: Profile
   
   (Founder member) < >
   
   Commercial site: http://www.cambridgenetwork.co.uk/profiles/p135.htm

5. **Geek Todd Niedham, Microsoft Research**
   
   85% - **Articles & General info**: Geek: Todd Niedham, Microsoft Research
   
   To Subject: Geek: Todd Niedham, Microsoft Research
   
   From: Douglas Song < > Date: Wed, 16...03/18/99
   
   Non-profit site: http://www.monkey.org/geeks/archive/9707/msg00081.html

6. **Microsoft Research fia SIGGRAPH ’96**
   
   84% - **Articles & General info**: Microsoft Research fia SIGGRAPH ’96
   
   Re: 52 fia-fâu i fia-fâu i
Estimating Priors

**Link Structure**
- Google - Brin & Page
- Clever (Hubs/Authorities) - Kleinberg et al.
- Web Archeologist - Bharat, Henzinger
- Similarities to citation analysis

**Information Use**
- Access Counts - e.g., DirectHit
- Collaborative Filtering
New Relevance Ranking

Relevance ranking can include:
- content-matching ... of course
- page/site popularity <external link count; proxy stats>
- page quality <site quality, dates, depth>
- spam or porn or other downweighting
- etc.

Combining these - relative weighting of these factors tricky and subjective
Text Classification

**Text Classification**: assign objects to one or more of a predefined set of categories using text features

- E.g., News feeds, Web data, OHSUMED, Email - spam/no-spam

**Approaches:**
- Human classification (e.g., LCSH, MeSH, Yahoo!, CyberPatrol)
- Hand-crafted knowledge engineered systems (e.g., CONSTRUE)
- Inductive learning methods
  - (Semi-) automatic classification
Inductive Learning Methods

- Supervised learning from *examples*
  - Examples are easy for domain experts to provide
  - Models easy to learn, update, and customize

- Example learning algorithms
  - Relevance Feedback, Decision Trees, Naïve Bayes, Bayes Nets, Support Vector Machines (SVMs)

- Text representation
  - Large vector of features (words, phrases, hand-crafted)
Support Vector Machine

**Optimization Problem**

- Find hyperplane, \( h \), separating positive and negative examples
- Optimization for maximum margin: \( \min \| \mathbf{w} \|^2 \), \( \mathbf{w} \cdot \mathbf{x} - b \geq 1 \), \( \mathbf{w} \cdot \mathbf{x} - b \leq -1 \)
- Classify new items using: \( f(\mathbf{w} \cdot \mathbf{x}) \)
The Bundesbank left credit policies unchanged after today's regular meeting of its council, a spokesman said in answer to enquiries. The West German discount rate remains at 3.0 pct, and the Lombard emergency financing rate at 5.0 pct.
Example: Reuters news

118 categories (article can be in more than one category)

Most common categories (#train, #test)
- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369, 119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Overall Results
- Linear SVM most accurate: 87% precision at 87% recall
**ROC for Category - Grain**

**Recall:** % labeled in category among those stories that are really in category

**Precision:** % really in category among those stories labeled in category
Text Categ Summary

- Accurate classifiers can be learned automatically from training examples
- Linear SVMs are efficient and provide very good classification accuracy
- Widely applicable, flexible, and adaptable representations
  - Email spam/no-spam, Web, Medical abstracts, TREC
Beyond Content Matching

- Domain/Object modeling
  - A priori importance
  - Text classification and clustering
- User/Task modeling
  - Implicit queries and Lumiere
- Advances in presentation and manipulation
  - Combining structure and search (e.g., DM)
Lumiere

- Inferring beliefs about user’s goals and ideal actions under uncertainty

- User query
- User activity
- User profile
- Data structures

Pr(Goals, Needs)

E. Horvitz, et al.
Lumiere
Lumiere → Office Assistant
Visualizing Implicit Queries

- **Explicit queries:**
  - Search is a separate, discrete task
  - Results not well integrated into larger task context

- **Implicit queries:**
  - Search as part of normal information flow
  - Ongoing query formulation based on user activities
  - Non-intrusive results display
US spot natgas prices hold amid storage injections

NEW YORK, April 6 (Reuters) - U.S. spot natural gas prices failed to budge from last week’s range on Monday as storage injections continued at a steady pace and cooler-than-normal weather created some demand in the upper Midwest and Southwest, traders said.

Henry Hub swing gas traded early at $2.49, but firmed by late morning to about $2.51-2.52, indicating little change from Friday’s levels.

In the Midcontinent, prices were also flat at $2.37-2.39, with Chicago city-gate values seen mostly at $2.55-2.56.

In western Texas, Permian Basin prices were up about one cent to the high-$2.20s, while San Juan prices were talked at $2.14-2.17.

In the Northeast, New York city-gate prices stepped up a couple of cents to the mid-to-high $2.70s as cooler weather lingered today in the region and some nuclear power became unavailable. Appalachian values on Columbia were quoted at $2.65-2.67.

Several nuclear outages were underway in the Northeast. PP&L Resources’ 1,094 megawatt (MW) Susquehanna 2 unit,
US spot natgas prices hold amid storage injections

NEW YORK, April 6 (Reuters) - U.S. spot natural gas prices failed to budge from last week's range on Monday as storage injections continued at a steady pace and cooler-than-normal weather created some demand in the upper Midwest and Southwest, traders said.

Henry Hub swing gas traded early at $2.49, but firm'd by late morning to about $2.51-2.52, indicating little change from Friday's levels.

In the Midcontinent, prices were also flat at $2.37-2.39, with Chicago city-gate values seen mostly at $2.55-2.56.

In western Texas, Permian Basin prices were up about one cent to the high-$2.20s, while San Juan prices were talked at $2.14-2.17.

In the Northeast, New York city-gate prices stepped up a couple of cents to the mid-to-high $2.70s as cooler weather lingered today in the region and some nuclear power became unavailable. Appalachian values on Columbia were quoted at $2.65-2.67.

Several nuclear outages were underway in the Northeast. PPL Resources' 1,094 megawatt (MW) Susquehanna 2 unit,
Data Mountain with 100 web pages.
Data Mountain with Implicit Query results shown (highlighted pages to left of selected page).
IQ Study: Experimental Details

- Store 100 Web pages
  - 50 popular Web pages; 50 random pages
  - With or without Implicit Query highlighting
    - IQ0: No IQ
    - IQ1: Co-occurrence based IQ - ‘best case’
    - IQ2: Content-based IQ

- Retrieve 100 Web pages
  - Title given as retrieval cue -- e.g., “CNN Home Page”
  - No implicit query highlighting at retrieval
Results: Information Storage

- Filing strategies

<table>
<thead>
<tr>
<th>IQ Condition</th>
<th>Filing Strategy</th>
<th>Semantic</th>
<th>Alphabetic</th>
<th>No Org</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ0: No IQ</td>
<td></td>
<td>11</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>IQ1: Co-occur based</td>
<td></td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IQ2: Content-based</td>
<td></td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
# Results: Information Storage

- **Number of categories (for semantic organizers)**

<table>
<thead>
<tr>
<th>IQ Condition</th>
<th>Average Number of Categories (std error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ0: No IQ</td>
<td>10.0 (3.6)</td>
</tr>
<tr>
<td>IQ1: Co-occur based</td>
<td>15.8 (5.8)</td>
</tr>
<tr>
<td>IQ2: Content-based</td>
<td>13.6 (5.9)</td>
</tr>
</tbody>
</table>

![7 Categ][7 Categ]

![23 Categ][23 Categ]
Results: Retrieval Time

Web Page Retrieval Time

Implicit Query Condition

Average RT (seconds)

IQ 0
IQ 1
IQ 2
Results: Retrieval Time

- Large variability across users
  - min: 3.1 secs
  - max: 39.1 secs

- Large variability across queries
  - min: 4.9 secs (NASA home page)
  - max: 24.3 secs (Welcome to Mercury Center)

- Popularity of Web pages did not matter
  - Top50: 12.9 secs
  - Random50: 12.8 secs
Implicit Query Highlights

- IQ built by observing user’s reading behavior
  - No explicit search required
  - Good matches returned
- IQ user model:
  - Combines present context (+ previous interests)
- Results presented in the context of a user-defined organization
Summary

- Improving content-matching
- And, beyond ...
  - Domain/Object Models
  - User/Task Models
  - Information Presentation and Use
- Also ...
  - non-text, multi-lingual, distributed
- http://research.microsoft.com/~sdumais
“information retrieval seemed like the easiest place to make progress ... information retrieval is really only a problem for people in library science -- if some computer scientists were to put their heads together, they’d probably have it solved before lunchtime” [GS]