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

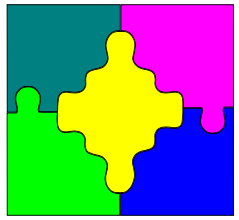
Susan Dumais

Microsoft Research

http://research.microsoft.com/~sdumais

IEEE ADL'99 - May 21, 1998

Research in IR at MS



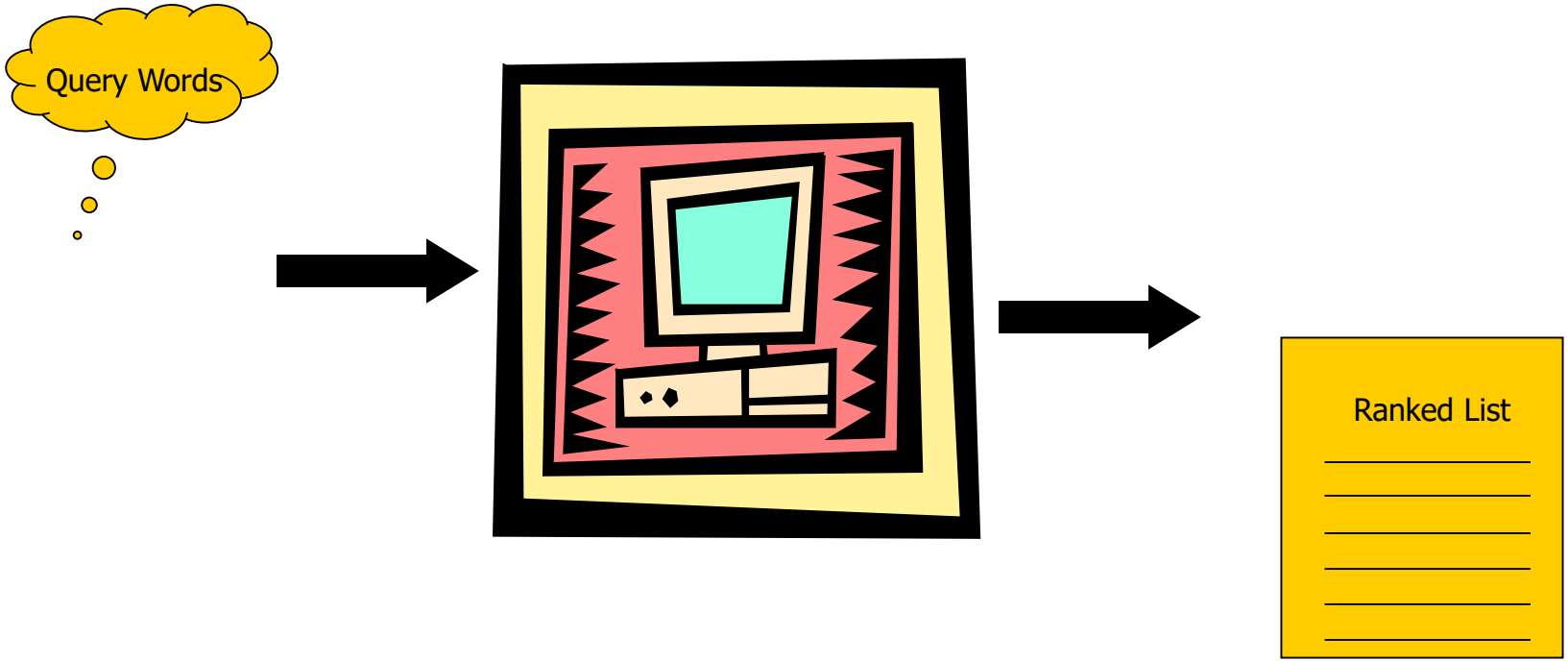
- Microsoft Research (http://research.microsoft.com)
  - Adaptive Systems and Interaction - IR/UI
  - Machine Learning and Applied Statistics
  - Data Mining
  - Natural Language Processing
  - Collaboration and Education
  - Database
  - MSR Cambridge; MSR Beijing
- Microsoft Product Groups … many IR-related

IR Themes & Directions

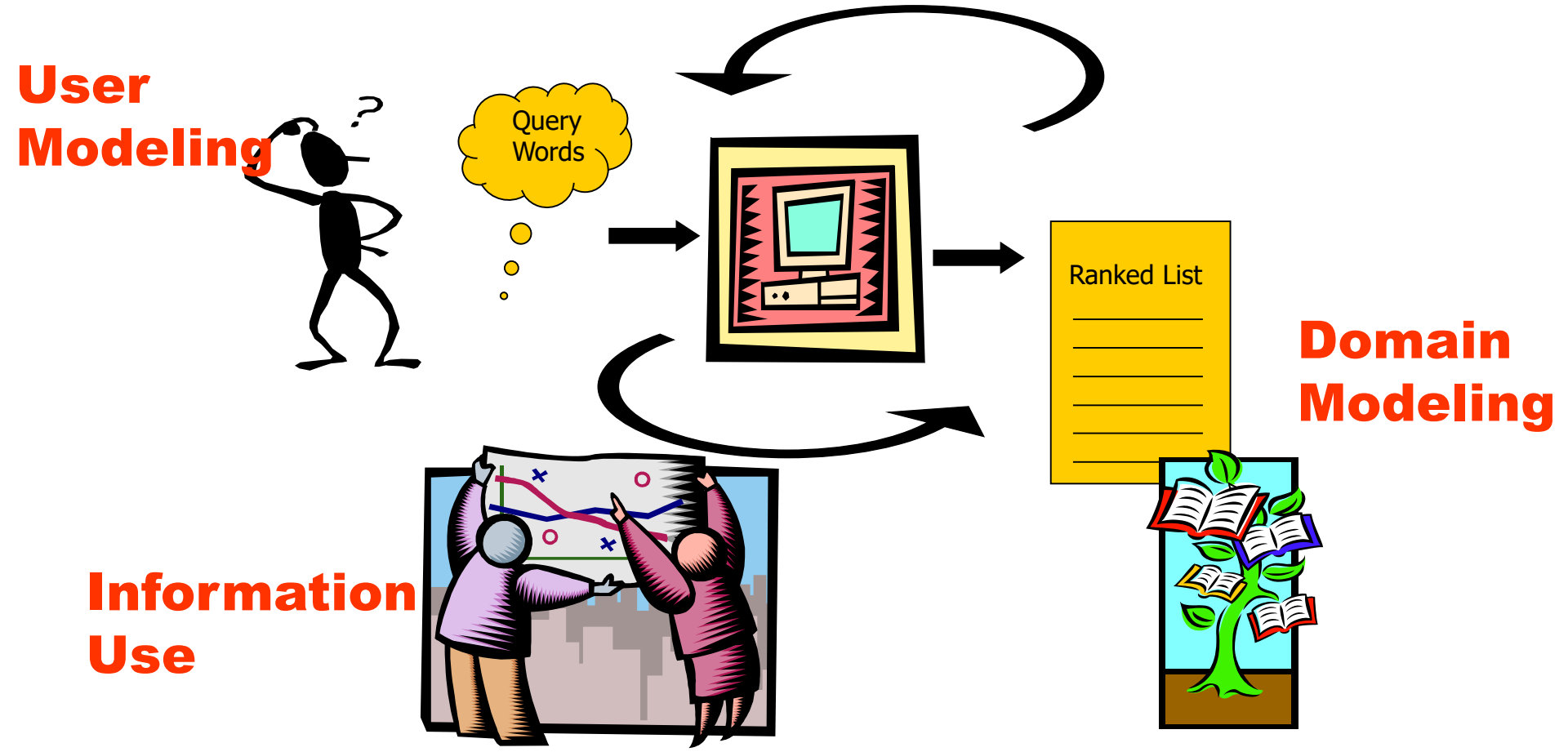
- Improvements in matching algorithms and representation
  - Probabilistic/Bayesian models
    - *p(Relevant|Document), p(Concept|Words)*
  - NLP: Truffle, MindNet
- Beyond content-matching
  - User/Task modeling
  - Domain/Object model
  - Advances in presentat

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

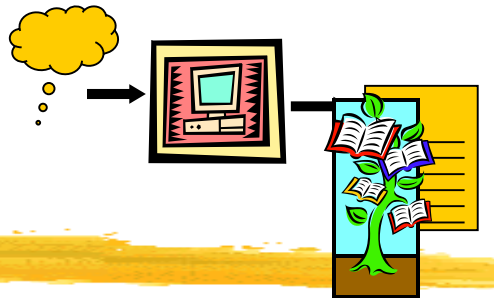
Traditional View of IR



What's Missing?

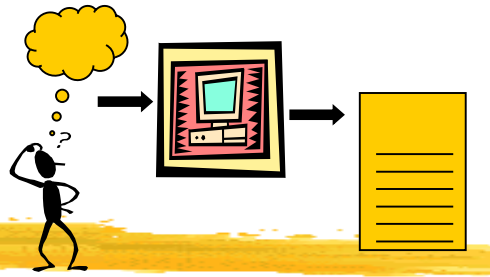


Domain/Obj Modeling



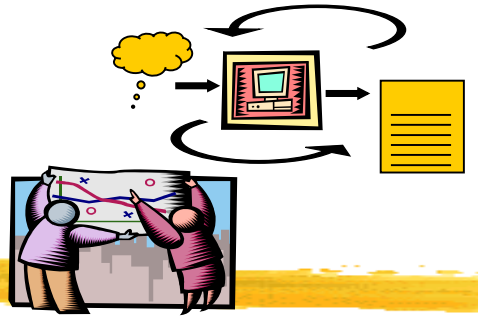
- Not all objects are equal … potentially big win
  - *A priori importance*
  - Information use (“readware”; collab filtering)
- Inter-object relationships
  - Link structure / hypertext
  - *Subject categories - e.g., text categorization. text clustering*
- Metadata
  - E.g., reliability, recency, cost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

User/Task Modeling



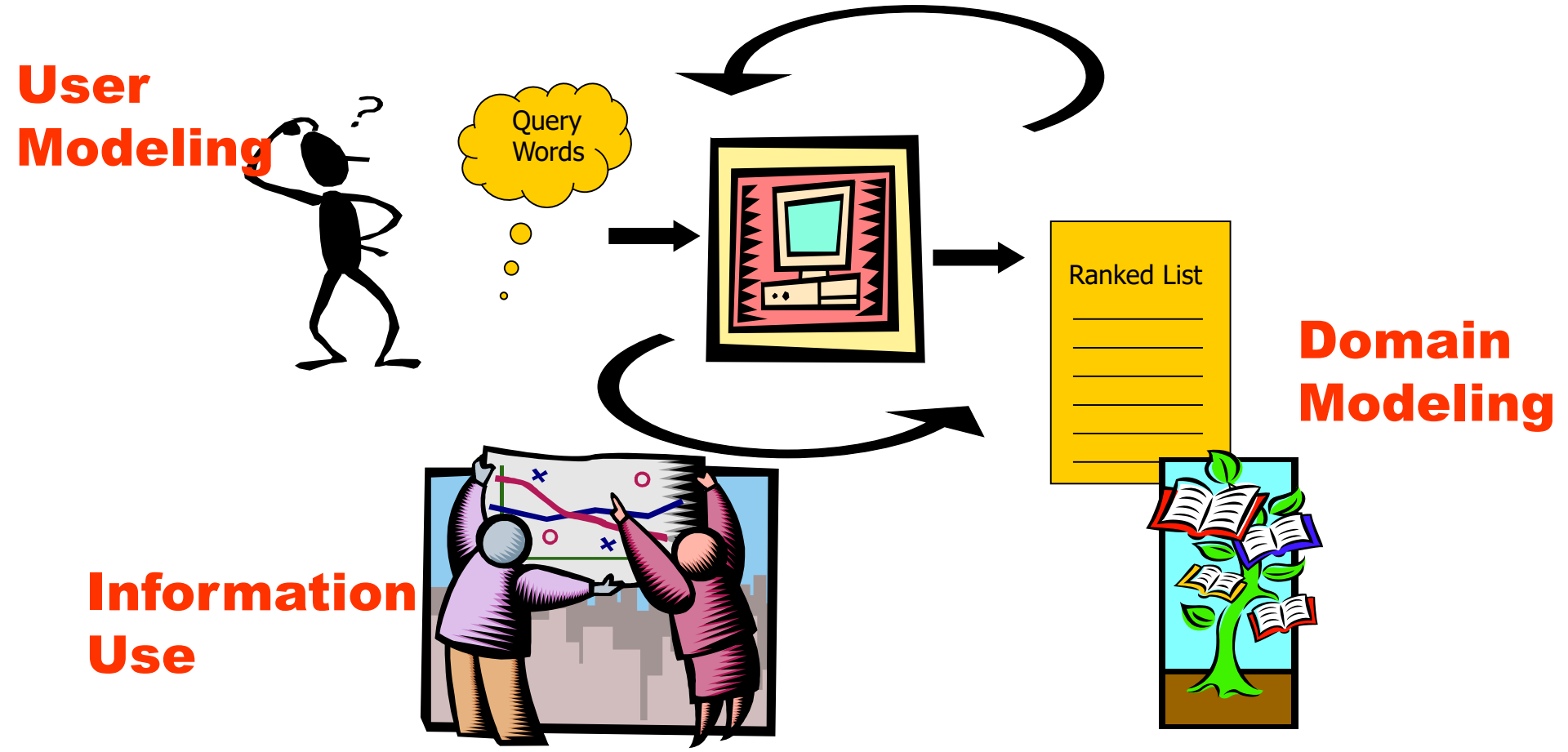
- Demographics
- Task -- What's the user's goal?
  - e.g., Lumiere
- Short and long-term content interests
  - *e.g., Implicit queries*
    - Interest model = f(content_similarity, time, interest)
  - e.g., Letiza, WebWatcher, Fab

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

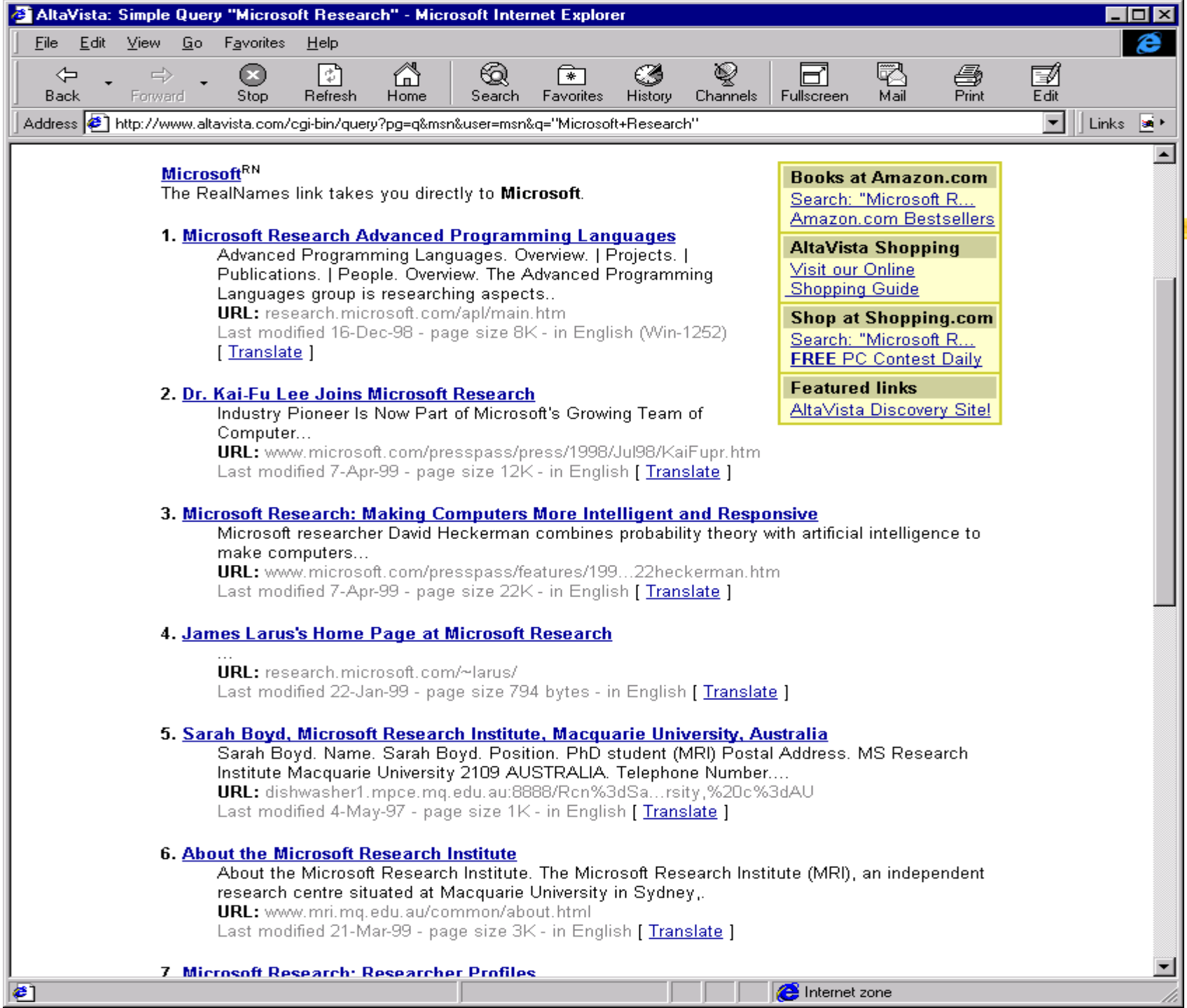


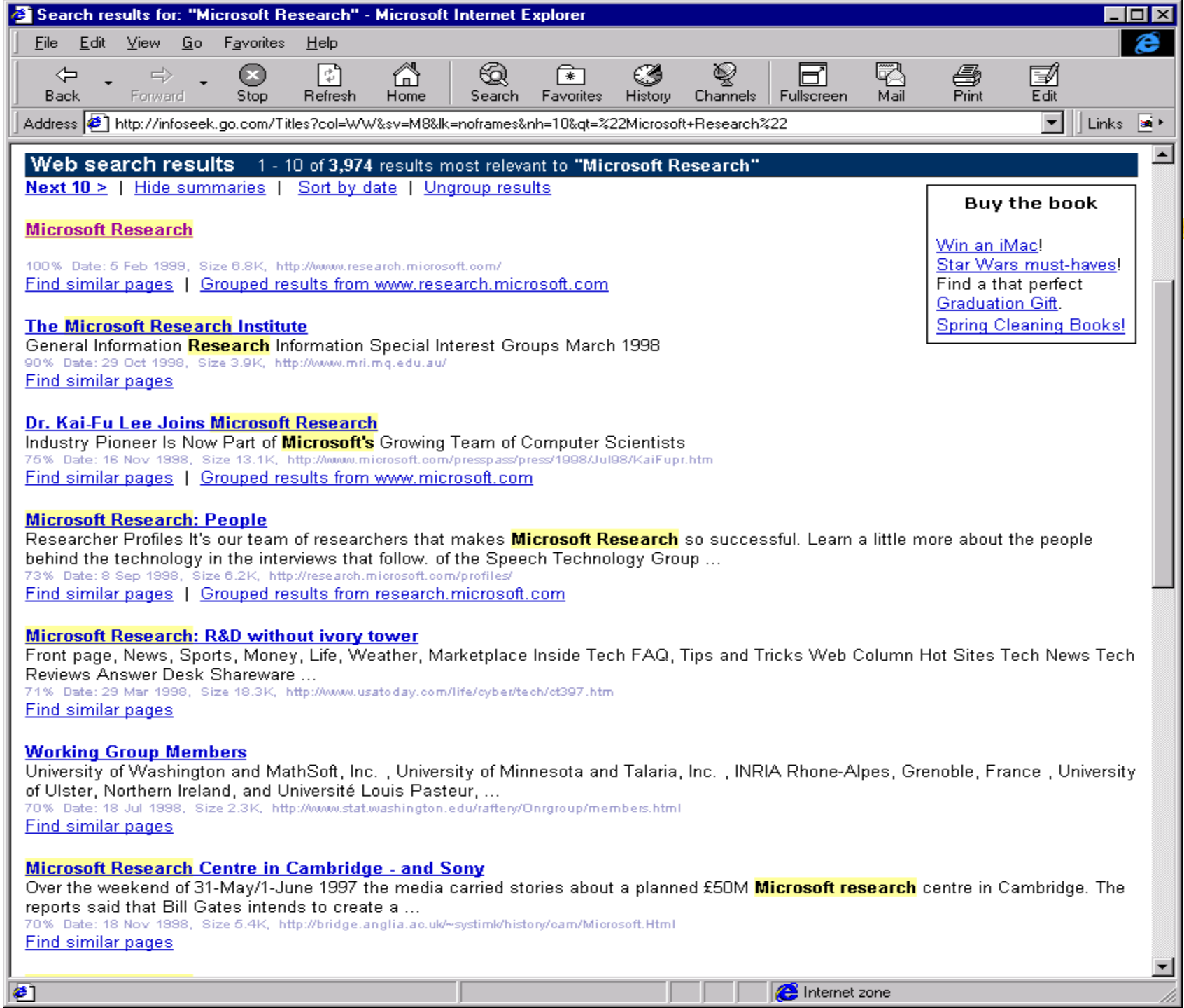
Beyond Content Matching

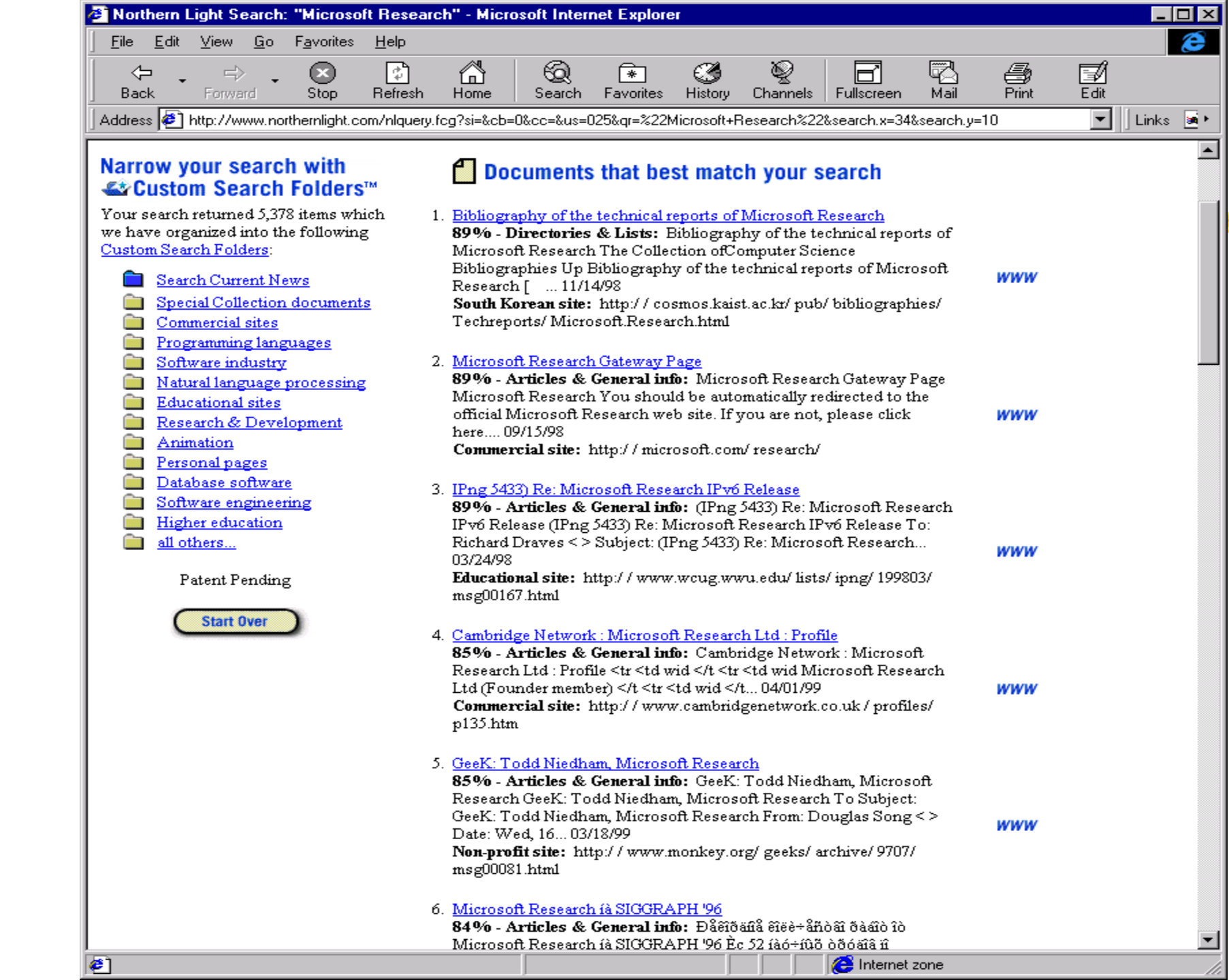
- Domain/Object modeling
  - A priori importance
  - Text classification and clustering
- User/Task modeling
  - Implicit queries and Lumiere
- Advances in presentation and manipulation
  - Combining structure and search (e.g., DM)

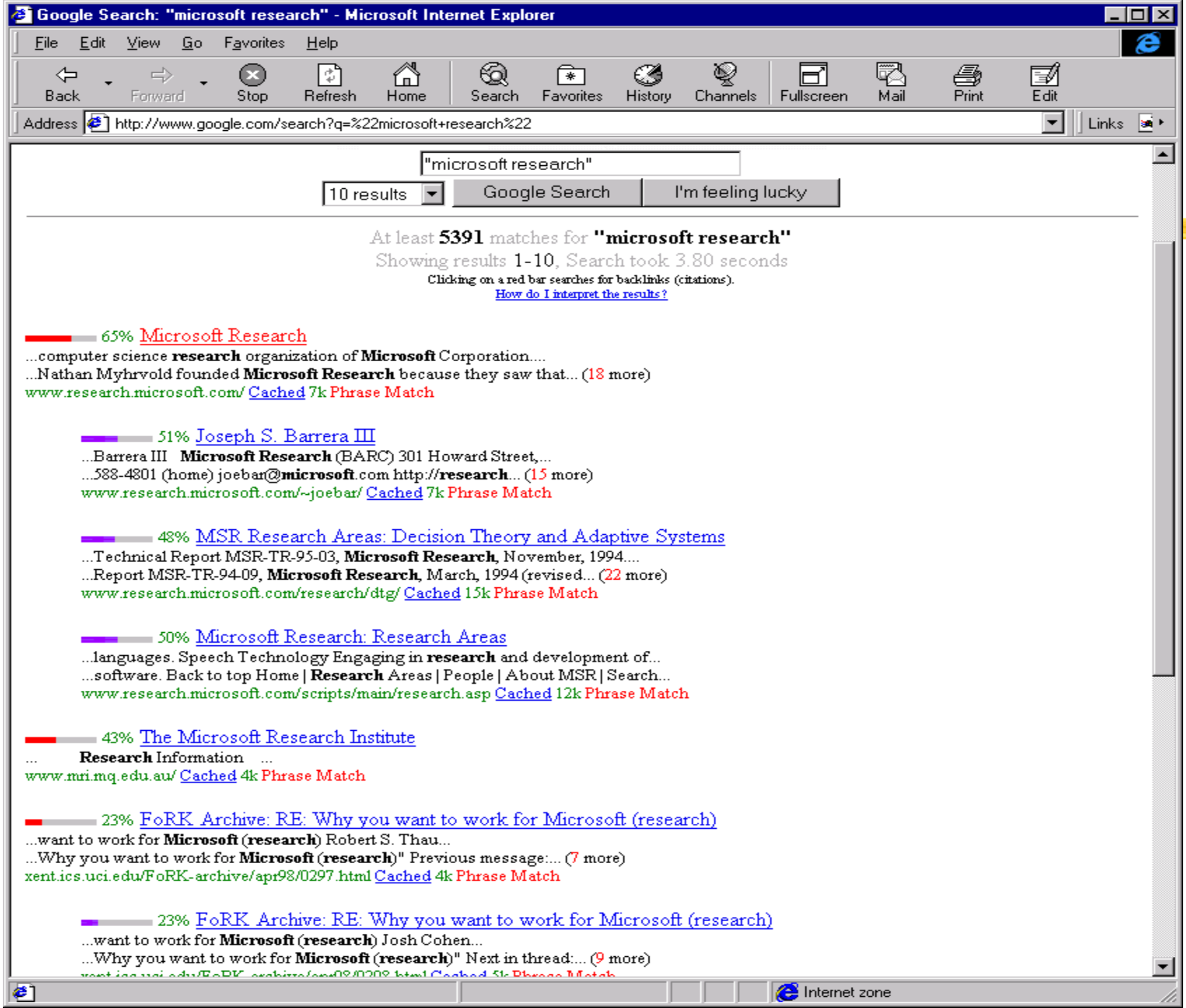
- Example:

  Web query for “Microsoft Research”


AltaVista: Simple Query "Microsoft Research" - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit
Address http://www.altavista.com/cgi-bin/query?pg=q&msn&user=msn&q="Microsoft+Research"
Links
Microsoft[RN]
The RealNames link takes you directly to Microsoft.
1. Microsoft Research Advanced Programming Languages
Advanced Programming Languages. Overview. | Projects. | Publications. | People. Overview. The Advanced Programming Languages group is researching aspects..
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...
URL: www.microsoft.com/presspass/press/1998/Jul98/KaiFupr.htm
Last modified 7-Apr-99 - page size 12K - in English [Translate]
3. Microsoft Research: Making Computers More Intelligent and Responsive
Microsoft researcher David Heckerman combines probability theory with artificial intelligence to make computers...
URL: www.microsoft.com/presspass/features/199...22heckerman.htm
Last modified 7-Apr-99 - page size 22K - in English [Translate]
4. James Larus'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:8888/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
Books at Amazon.com
Search: "Microsoft R...
Amazon.com Bestsellers
AltaVista Shopping
Visit our Online Shopping Guide
Shop at Shopping.com
Search: "Microsoft R...
FREE PC Contest Daily
Featured links
AltaVista Discovery Site!
Internet zone


Search results for: "Microsoft Research" - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit
Address http://infoseek.go.com/Titles?col=WW&sv=M8&lk=noframes&nh=10&qt=%22Microsoft+Research%22
Links
Web search results 1 - 10 of 3,974 results most relevant to "Microsoft Research"
Next 10 > | Hide summaries | Sort by date | Ungroup results
Buy the book
Win an iMac!
Star Wars must-haves!
Find a that perfect Graduation Gift.
Spring Cleaning Books!
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
90% Date: 29 Oct 1998, Size 3.9K, http://www.mri.mq.edu.au/
Find similar pages
Dr. Kai-Fu Lee Joins Microsoft Research
Industry Pioneer Is Now Part of Microsoft's Growing Team of Computer Scientists
75% Date: 16 Nov 1998, Size 13.1K, http://www.microsoft.com/presspass/press/1998/Jul98/KaiFupr.htm
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 ...
73% Date: 8 Sep 1998, Size 6.2K, http://research.microsoft.com/profiles/
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 ...
71% Date: 29 Mar 1998, Size 18.3K, http://www.usatoday.com/life/cyber/tech/ct397.htm
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, ...
70% Date: 18 Jul 1998, Size 2.3K, http://www.stat.washington.edu/raftery/Onrgroup/members.html
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/~systimk/history/cam/Microsoft.Html
Find similar pages
Internet zone


Northern Light Search: "Microsoft Research" - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit
Address http://www.northernlight.com/nlquery.fcg?si=&cb=0&cc=&us=025&qr=%22Microsoft+Research%22&search.x=34&search.y=10
Links
Narrow your search with Custom Search Folders™
Your search returned 5,378 items which we have organized into the following Custom Search Folders:
Search Current News
Special Collection documents
Commercial sites
Programming languages
Software industry
Natural language processing
Educational sites
Research & Development
Animation
Personal pages
Database software
Software engineering
Higher education
all others...
Patent Pending
Start Over
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 ofComputer Science Bibliographies Up Bibliography of the technical reports of Microsoft Research [... 11/14/98
South Korean site: http:/ / cosmos.kaist.ac.kr/ pub/ bibliographies/ Techreports/ Microsoft.Research.html
WWW
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.... 09/15/98
Commercial site: http:/ / microsoft.com/ research/
WWW
3. IPng 5433) Re: Microsoft Research IPv6 Release
89% - Articles & General info: (IPng 5433) Re: Microsoft Research IPv6 Release (IPng 5433) Re: Microsoft Research IPv6 Release To: Richard Draves < > Subject: (IPng 5433) Re: Microsoft Research... 03/24/98
Educational site: http:/ / www.wcug.wwu.edu/ lists/ ipng/ 199803/ msg00167.html
WWW
4. Cambridge Network : Microsoft Research Ltd : Profile
85% - Articles & General info: Cambridge Network : Microsoft Research Ltd : Profile <tr <td wid </t <tr <td wid Microsoft Research Ltd (Founder member) </t <tr <td wid </t... 04/01/99
Commercial site: http:/ / www.cambridgenetwork.co.uk / profiles/ p135.htm
WWW
5. GeeK: Todd Niedham, Microsoft Research
85% - Articles & General info: GeeK: Todd Niedham, Microsoft Research 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
WWW
6. Microsoft Research íà SIGGRAPH '96
84% - Articles & General info: Ðåêîðäíûå êîëè÷åñòâî ðàáîò îò Microsoft Research íà SIGGRAPH '96 Èc 52 íàó÷íûõ òðóäîâ íî
Internet zone


Google Search: "microsoft research" - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit
Address http://www.google.com/search?q=%22microsoft+research%22
Links
"microsoft research"
10 results
Google Search
I'm feeling lucky
At least 5391 matches for "microsoft research"
Showing results 1-10, Search took 3.80 seconds
Clicking on a red bar searches for backlinks (citations).
How do I interpret the results?
65% Microsoft Research
...computer science research organization of Microsoft Corporation....
...Nathan Myhrvold founded Microsoft Research because they saw that... (18 more)
www.research.microsoft.com/ Cached 7k Phrase Match
51% Joseph S. Barrera III
...Barrera III Microsoft Research (BARC) 301 Howard Street,...
...588-4801 (home) joebar@microsoft.com http://research... (15 more)
www.research.microsoft.com/~joebar/ Cached 7k Phrase Match
48% MSR Research Areas: Decision Theory and Adaptive Systems
...Technical Report MSR-TR-95-03, Microsoft Research, November, 1994....
...Report MSR-TR-94-09, Microsoft Research, March, 1994 (revised... (22 more)
www.research.microsoft.com/research/dtg/ Cached 15k Phrase Match
50% Microsoft Research: Research Areas
...languages. Speech Technology Engaging in research and development of...
...software. Back to top Home | Research Areas | People | About MSR | Search...
www.research.microsoft.com/scripts/main/research.asp Cached 12k Phrase Match
43% The Microsoft Research Institute
... Research Information ...
www.mri.mq.edu.au/ Cached 4k Phrase Match
23% FoRK Archive: RE: Why you want to work for Microsoft (research)
...want to work for Microsoft (research) Robert S. Thau...
...Why you want to work for Microsoft (research)" Previous message:... (7 more)
xent.ics.uci.edu/FoRK-archive/apr98/0297.html Cached 4k Phrase Match
23% FoRK Archive: RE: Why you want to work for Microsoft (research)
...want to work for Microsoft (research) Josh Cohen...
...Why you want to work for Microsoft (research)" Next in thread:... (9 more)
Internet zone

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 couse
  - 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

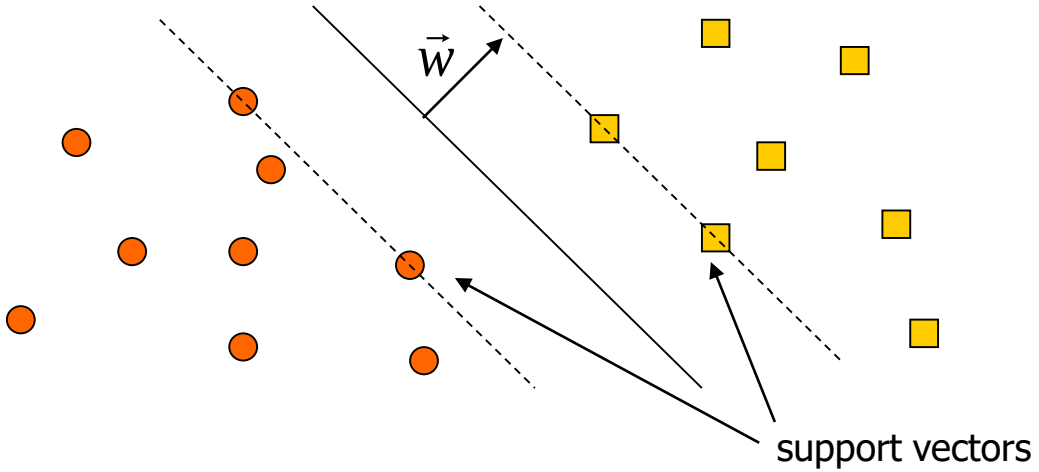
- *Text Classification*: assign objects to one or more of a predefined set of categories using text features
  - E.g., News feeds, Web data, OHSUMED, Email - spam/no-spam
- Approaches:
  - Human classification (e.g., LCSH, MeSH, Yahoo!, CyberPatrol)
  - Hand-crafted knowledge engineered systems (e.g., 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: $\min \|\vec{w}\|^2, \vec{w}\cdot\vec{x}-b \geq 1, \vec{w}\cdot\vec{x}-b \leq -1$
  - Classify new items using: $f(\vec{w}\cdot\vec{x})$



Reuters Data Set (21578 - ModApte split)

- 9603 training articles; 3299 test articles
- Example "interest" article

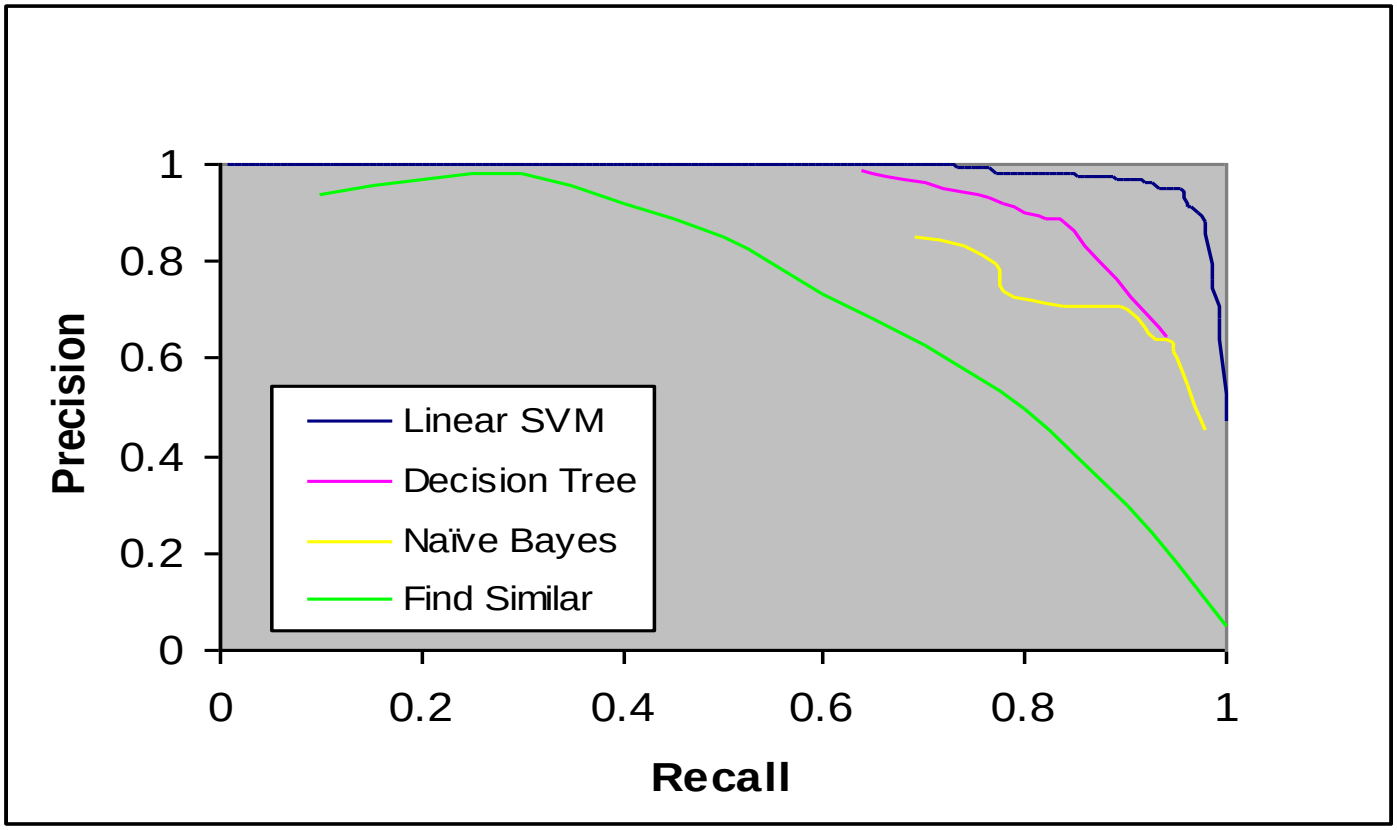
  2-APR-1987 06:35:19.50
  west-germany
  b f BC-BUNDESBANK-LEAVES-CRE 04-02 0052
  FRANKFURT, March 2
  The Bundesbank left credit policies unchanged after today's regular meeting of its council, a spokesman said in answer to enquiries. The West German discount rate remains at 3.0 pct, and the Lombard emergency financing rate at 5.0 pct.
  REUTER

- Average article 200 words long

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

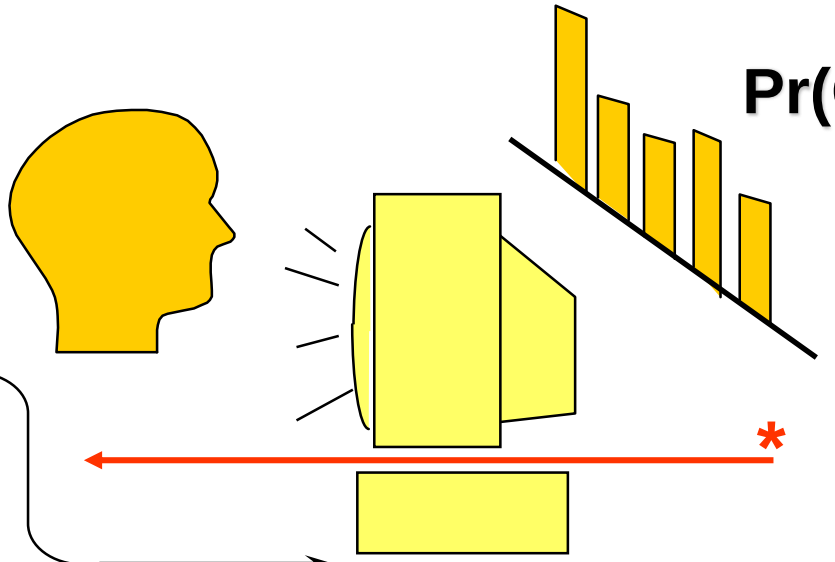
- Accurate classifiers can be learned automatically from training examples
- Linear SVMs are efficient and provide very good classification accuracy
- Widely applicable, flexible, and adaptable representations
  - Email spam/no-spam, Web, Medical abstracts, TREC

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

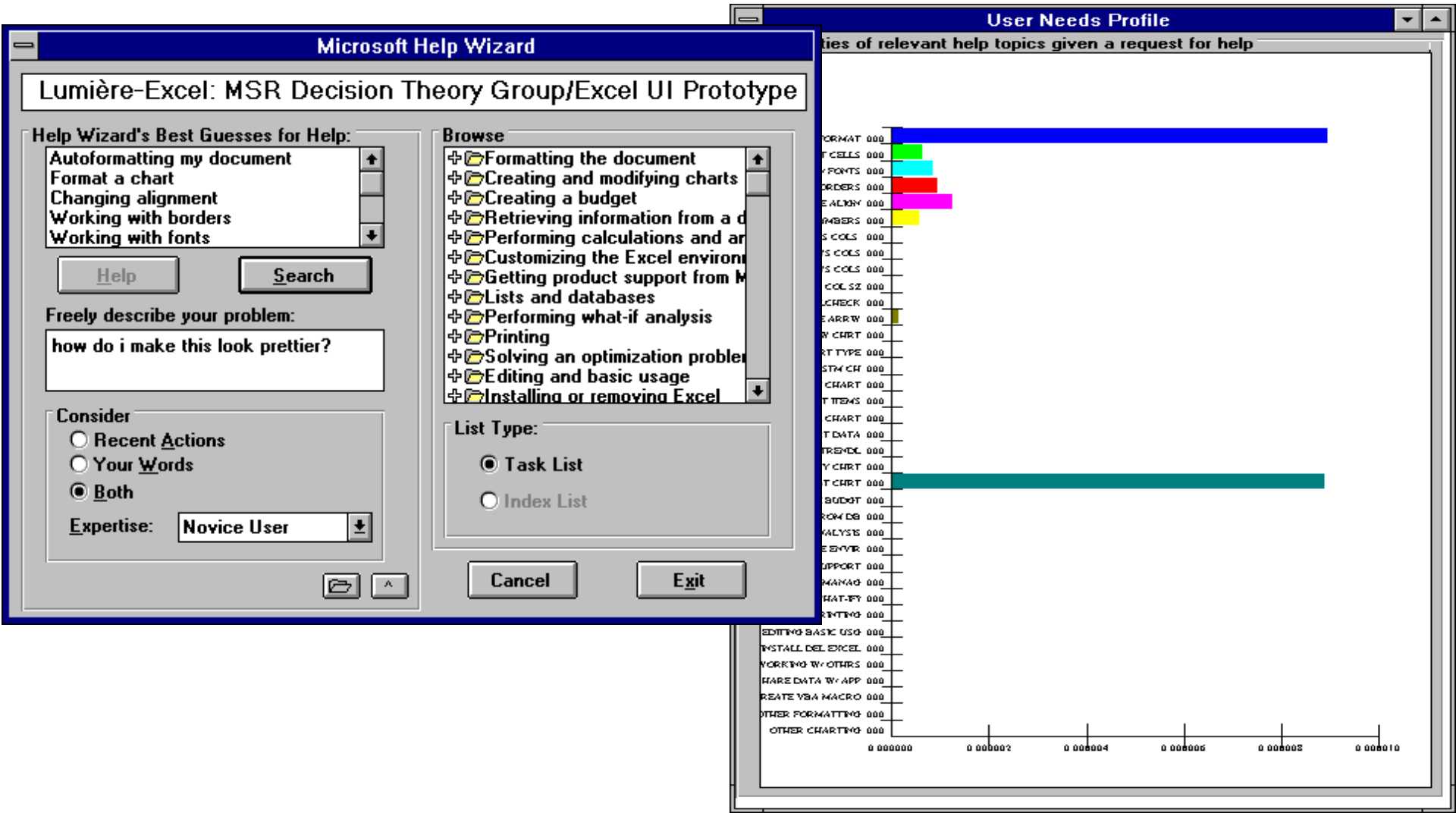
- **Inferring beliefs about user's goals and ideal actions under uncertainty**



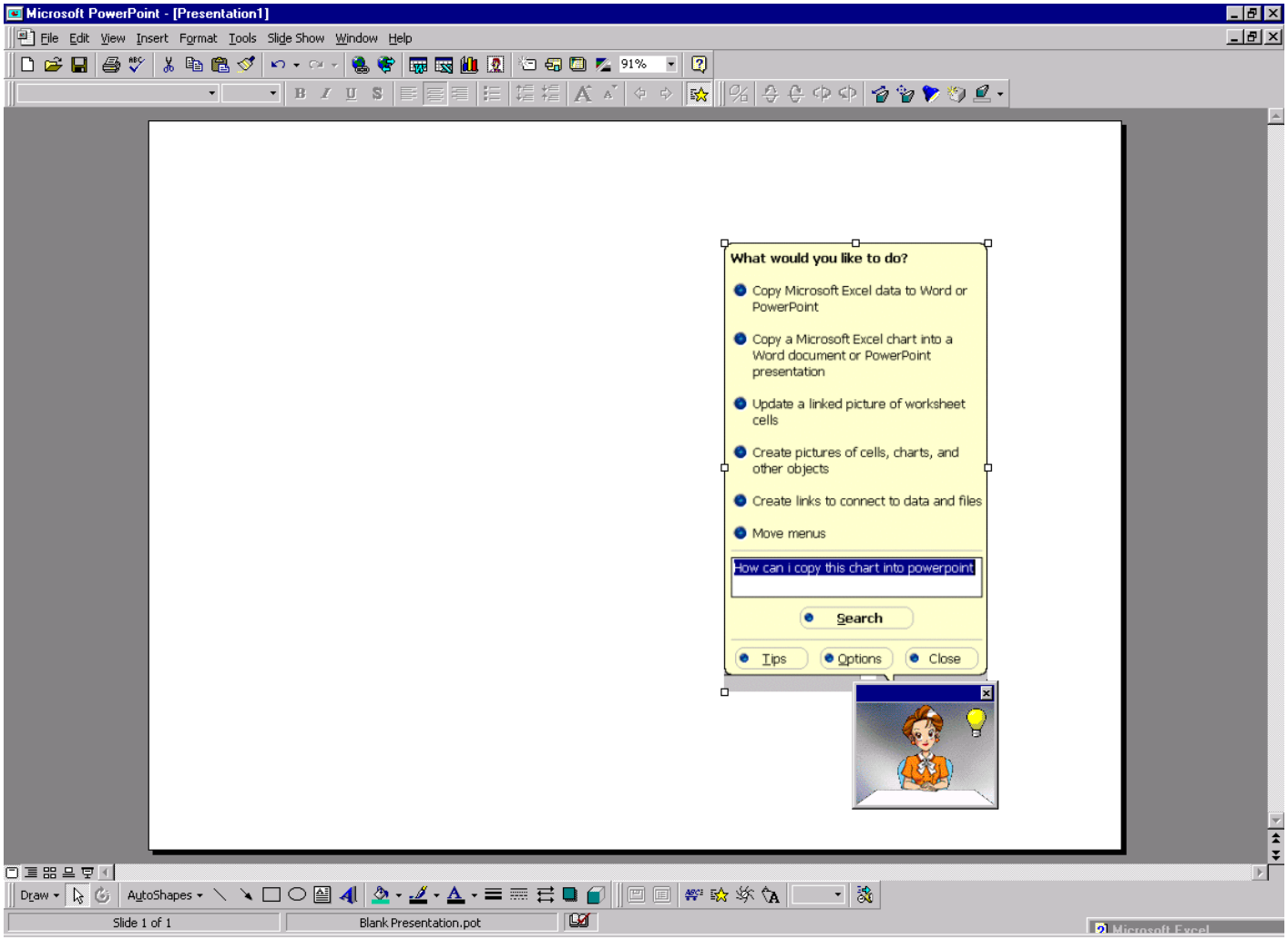
- User query
- User activity
- User profile
- Data structures

E. Horvitz, et al.

Lumiere

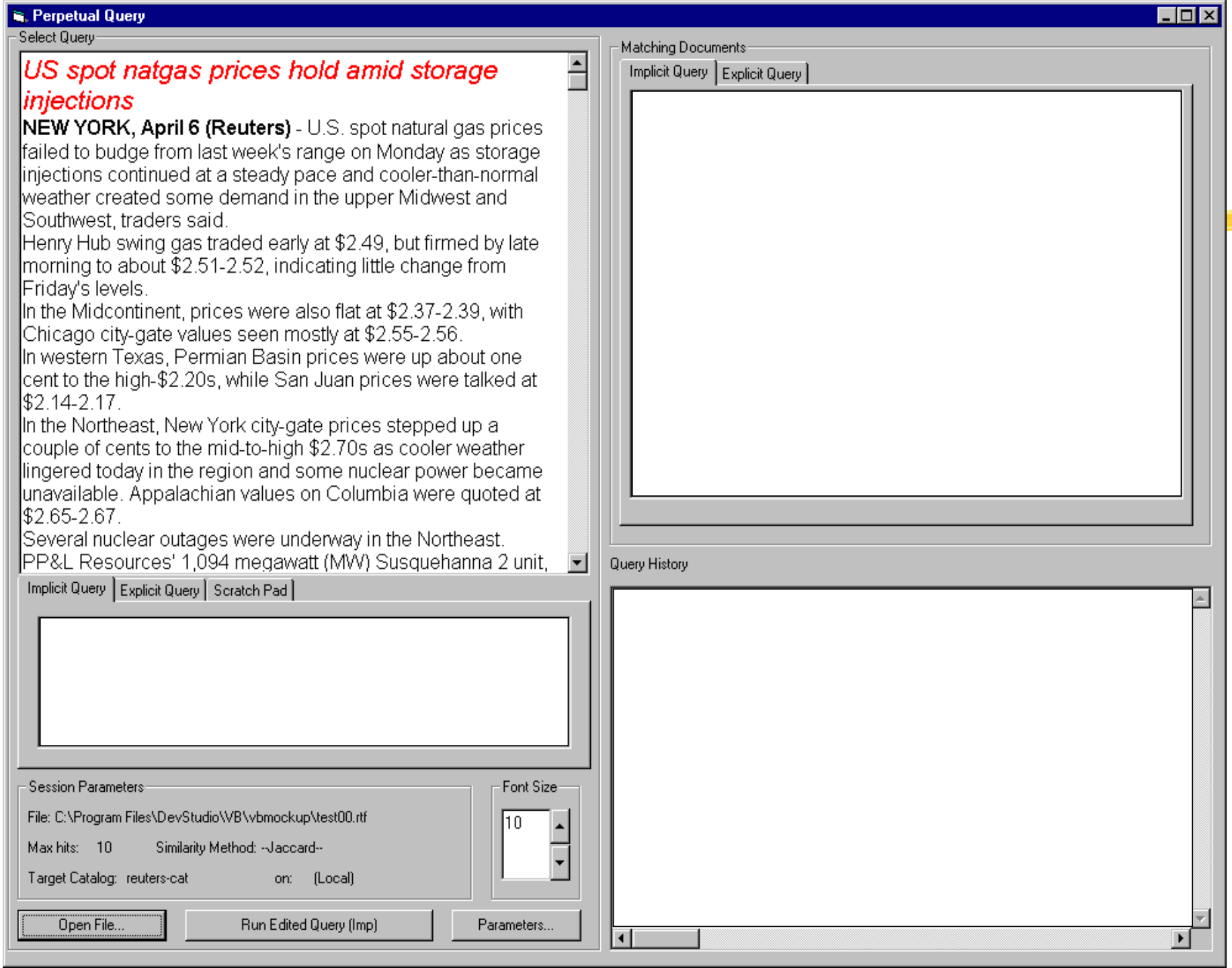

Microsoft Help Wizard
Lumière-Excel: MSR Decision Theory Group/Excel UI Prototype
Help Wizard's Best Guesses for Help:
Autoformatting my document
Format a chart
Changing alignment
Working with borders
Working with fonts
Help
Search
Freely describe your problem:
how do i make this look prettier?
Consider
Recent Actions
Your Words
Both
Expertise:
Novice User
Browse
Formatting the document
Creating and modifying charts
Creating a budget
Retrieving information from a d
Performing calculations and ar
Customizing the Excel environ
Getting product support from M
Lists and databases
Performing what-if analysis
Printing
Solving an optimization proble
Editing and basic usage
Installing or removing Excel
List Type:
Task List
Index List
Cancel
Exit
User Needs Profile
ties of relevant help topics given a request for help
0.000000
0.000002
0.000004
0.000006
0.000008
0.000010

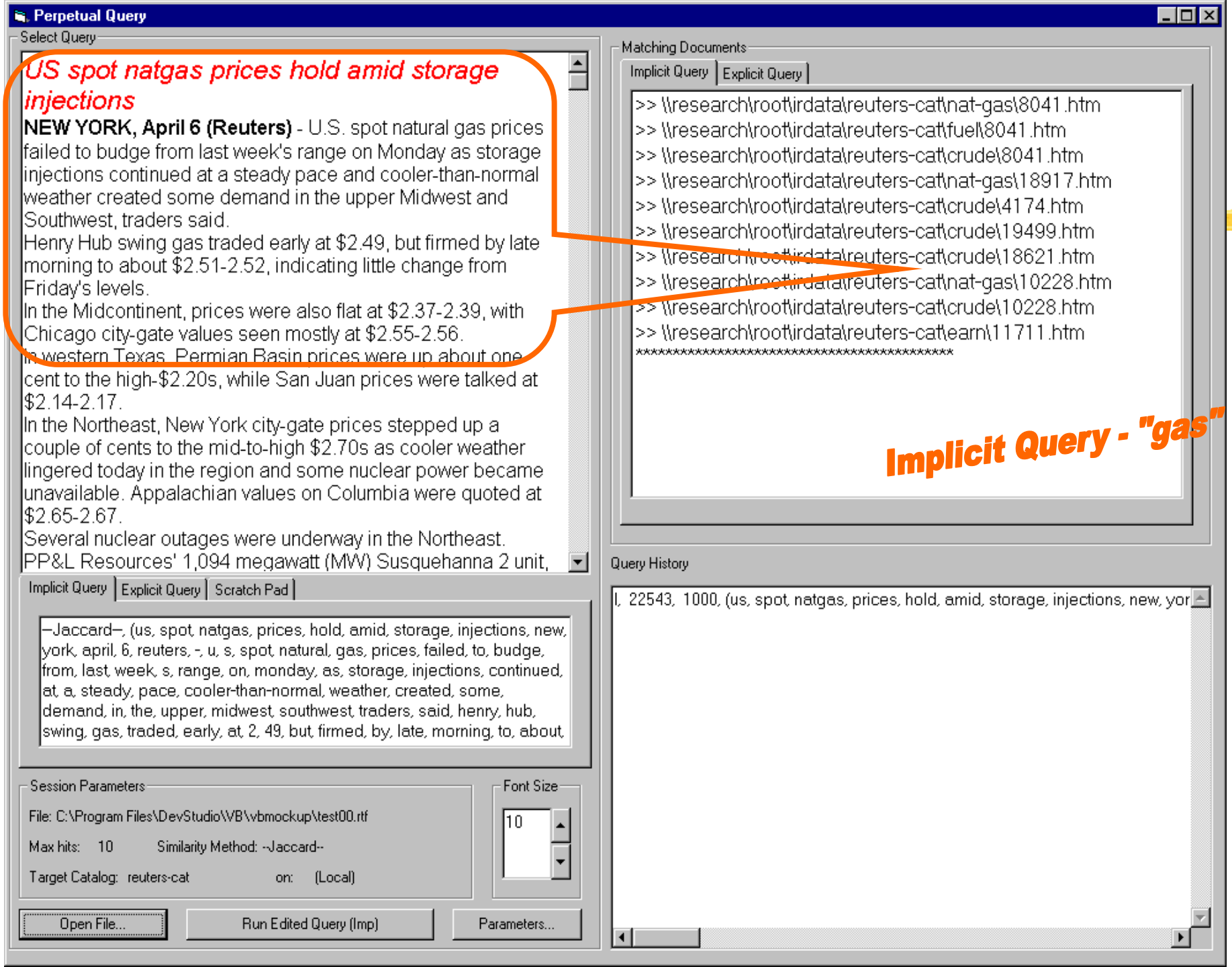
Lumiere ➡ Office Assistant



Visualizing Implicit Queries

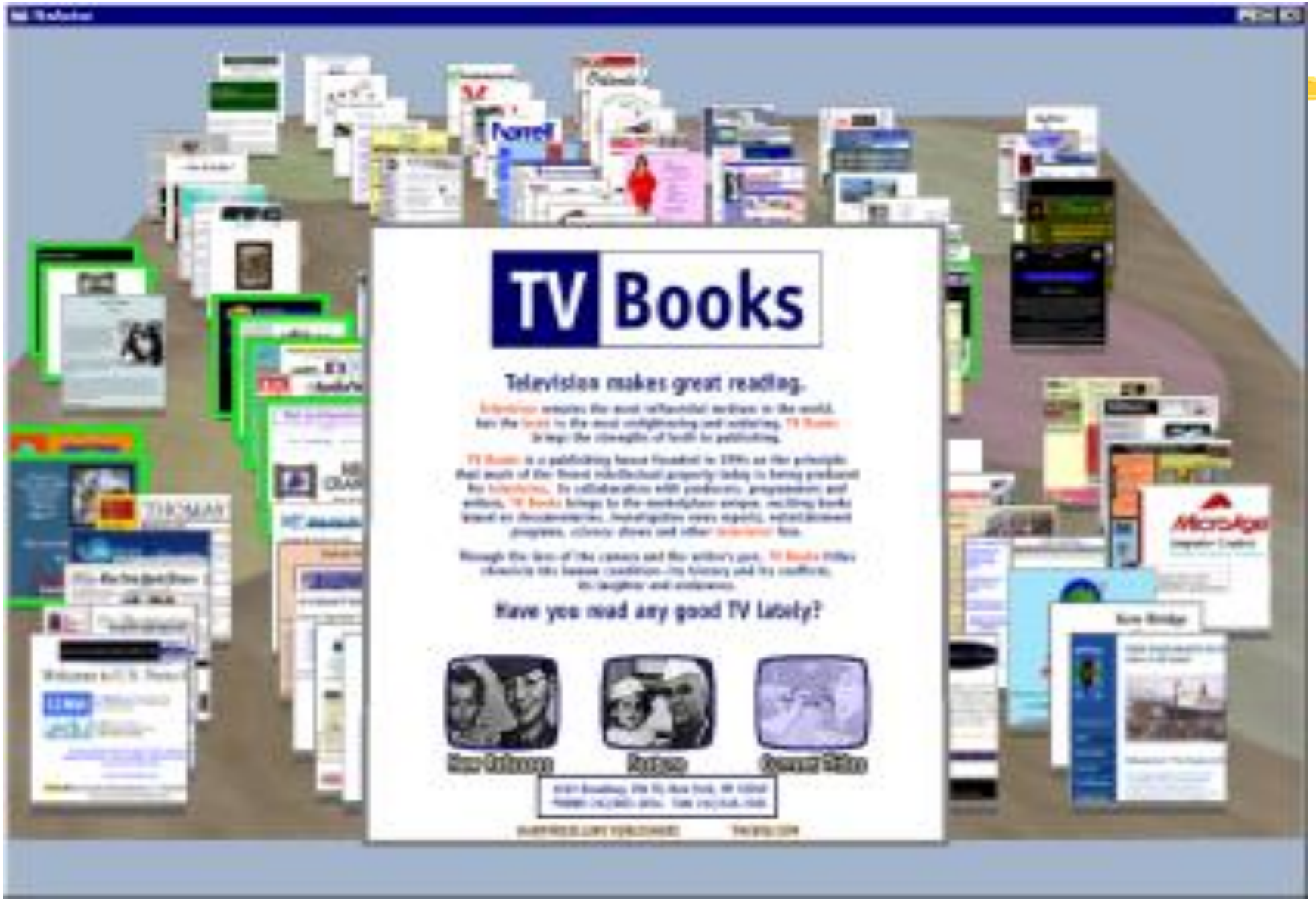
- Explicit queries:
  - Search is a separate, discrete task
  - Results not well integrated into larger task context
- **Implicit queries**:
  - Search as part of normal information flow
  - Ongoing query formulation based on user activities
  - Non-intrusive results display


Perpetual Query
Select Query
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,
Implicit Query
Explicit Query
Scratch Pad
Session Parameters
File: C:\Program Files\DevStudio\VB\vbmockup\test00.rtf
Max hits: 10
Similarity Method: --Jaccard--
Target Catalog: reuters-cat
on: (Local)
Font Size
10
Open File...
Run Edited Query (Imp)
Parameters...
Matching Documents
Implicit Query
Explicit Query
Query History


Perpetual Query
Select Query
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,
Implicit Query
Explicit Query
Scratch Pad
–Jaccard–, (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, cooler-than-normal, weather, created, some, demand, in, the, upper, midwest, southwest, traders, said, henry, hub, swing, gas, traded, early, at, 2, 49, but, firmed, by, late, morning, to, about,
Session Parameters
File: C:\Program Files\DevStudio\VB\vbmockup\test00.rtf
Max hits: 10
Similarity Method: --Jaccard--
Target Catalog: reuters-cat
on: (Local)
Font Size
10
Open File...
Run Edited Query (Imp)
Parameters...
Matching Documents
Implicit Query
Explicit Query
>> \\research\root\irdata\reuters-cat\nat-gas\8041.htm
>> \\research\root\irdata\reuters-cat\fuel\8041.htm
>> \\research\root\irdata\reuters-cat\crude\8041.htm
>> \\research\root\irdata\reuters-cat\nat-gas\18917.htm
>> \\research\root\irdata\reuters-cat\crude\4174.htm
>> \\research\root\irdata\reuters-cat\crude\19499.htm
>> \\research\root\irdata\reuters-cat\crude\18621.htm
>> \\research\root\irdata\reuters-cat\nat-gas\10228.htm
>> \\research\root\irdata\reuters-cat\crude\10228.htm
>> \\research\root\irdata\reuters-cat\earn\11711.htm
****
Implicit Query - "gas"
Query History
I, 22543, 1000, (us, spot, natgas, prices, hold, amid, storage, injections, new, yor

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

	Filing Strategy		
IQ Condition	**Semantic**	**Alphabetic**	**No Org**
IQ0: No IQ	11	3	1
IQ1: Co-occur based	8	1	0
IQ2: Content-based	10	1	0

Results: Information Storage

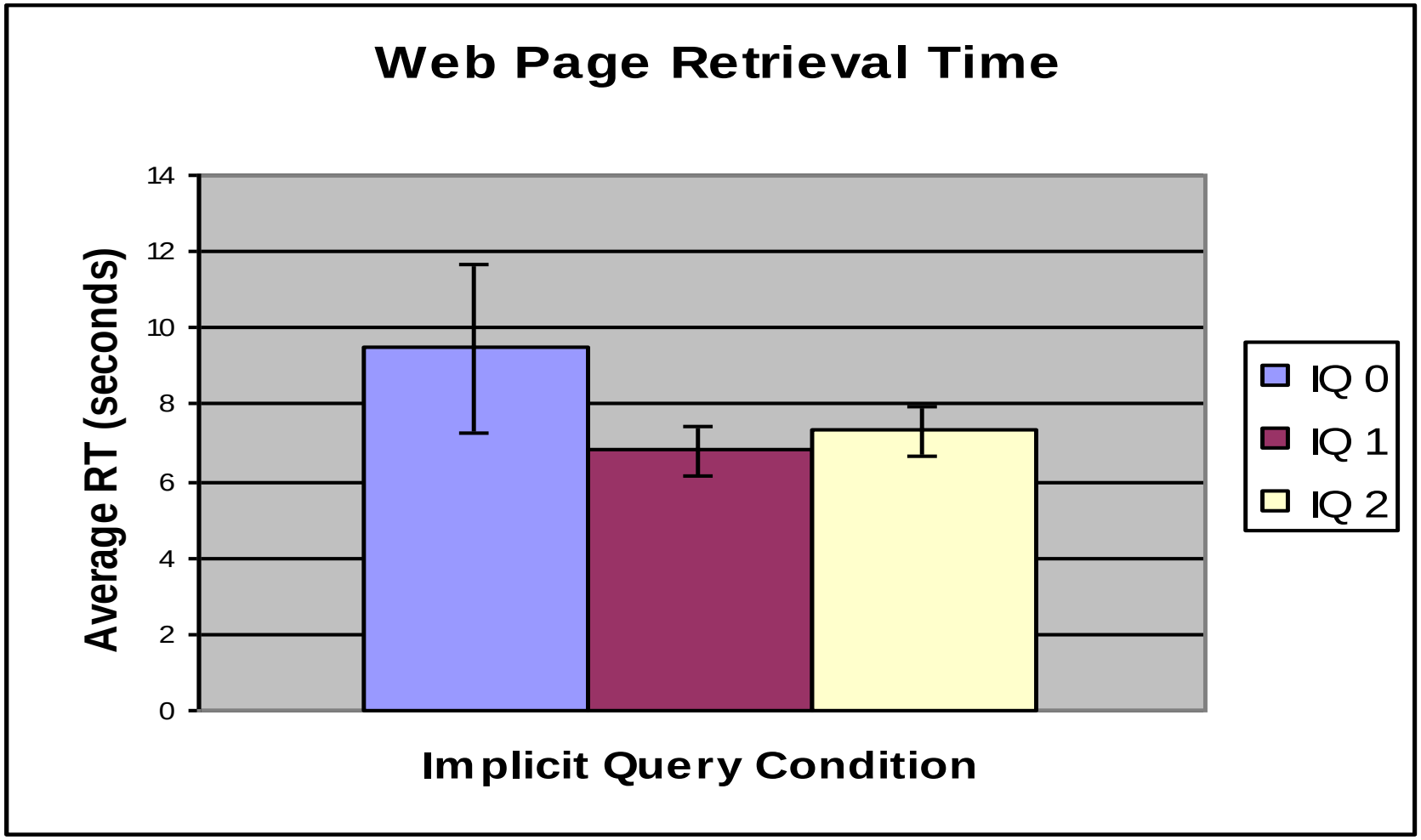
- Number of categories (for semantic organizers)

IQ Condition	Average Number of Categories (std error)
IQ0: No IQ	10.0 (3.6)
IQ1: Co-occur based	15.8 (5.8)
IQ2: Content-based	13.6 (5.9)

7 Categ

23 Categ

Results: Retrieval Time



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

The View from Wired, May 1996

- *"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]*