Research in Information Retrieval and Management

Susan Dumais

Microsoft Research

Library of Congress   Feb 8, 1999

Research in IR at MS

- Microsoft Research (http://research.microsoft.com)
  - Decision Theory and Adaptive Systems
  - Natural Language Processing
  - MSR Cambridge
  - User Interface
  - Database
  - Web Companion
  - Paperless Office
- Microsoft Product Groups … many IR-related

IR Themes & Directions

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

Improvements: Using Probabilistic Model

- MSR-Cambridge (Steve Robertson)
- Probabilistic Retrieval (e.g., Okapi)
  - Theory-driven derivation of matching function
  - Estimate: P_Q(r_i=Rel or NotRel | **d**=document)
    - Using Bayes Rule and assuming conditional independence given Rel/NotRel

$$P_Q(r_i \mid \mathbf{d}) = P(r_i)P(\mathbf{d} \mid r_i) / P(\mathbf{d})$$

$$P_Q(r_i \mid \mathbf{d}) = P(r_i)\prod_{j=1}^{t} P(x_j \mid r_i) / P(\mathbf{d})$$

Improvements: Using Probabilistic Model

- Good performance for uniform length document surrogates (e.g., abstracts)
- Enhanced to take into account term frequency and document
  - "BM25" one of the best ranking function at TREC
- Easy to incorporate relevance feedback
- Now looking at adaptive filtering/routing

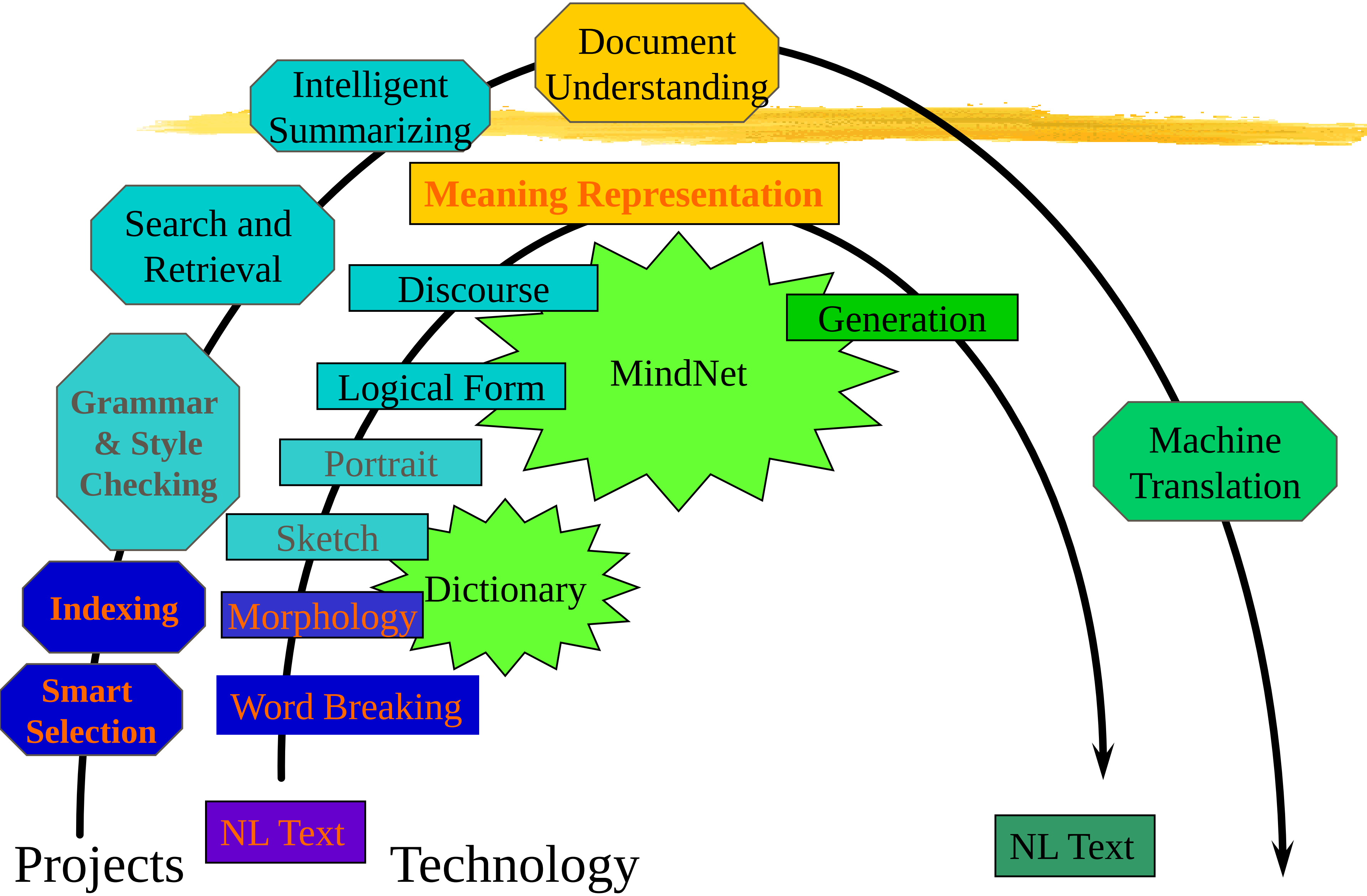
Improvements: Using NLP

- Current search techniques use word forms

- Improvements in content-matching will come from:
  -> Identifying relations between words
  -> Identifying word meanings

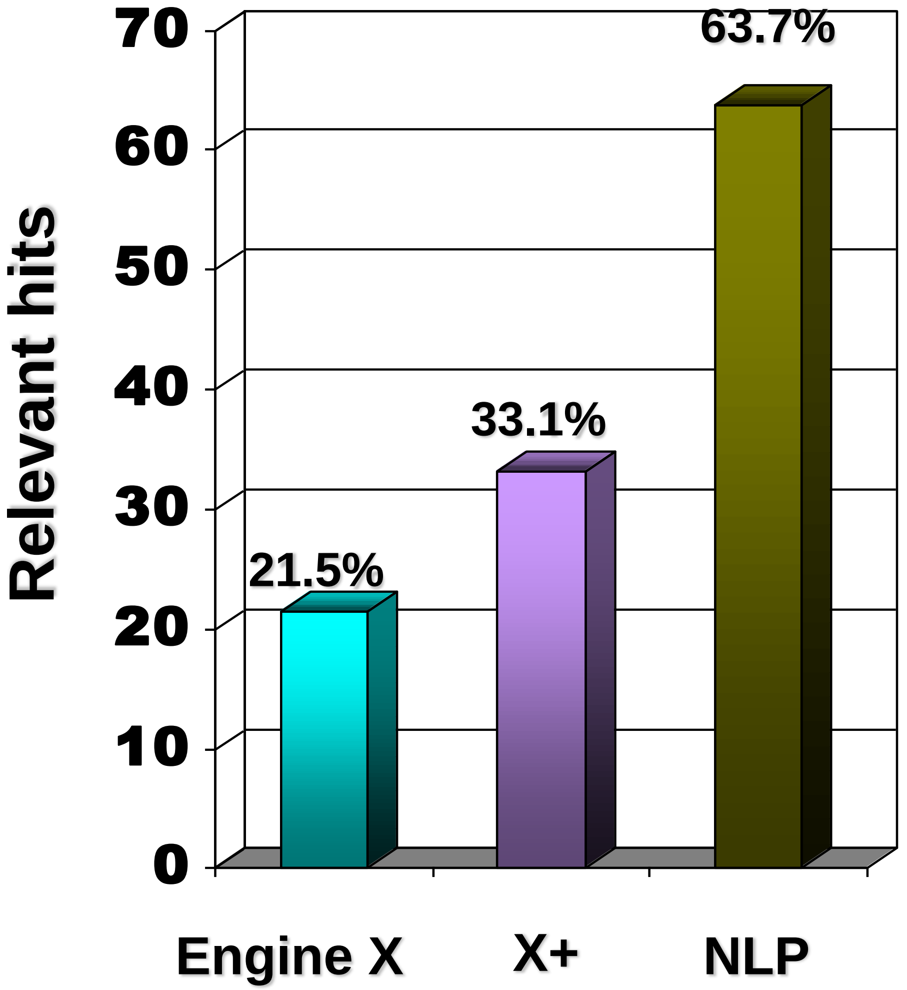
- Advanced NLP can provide these
- http:/research.microspft.com/nlp

NLP System Architecture



“Truffle”: Word Relations

% Relevant In Top Ten Docs

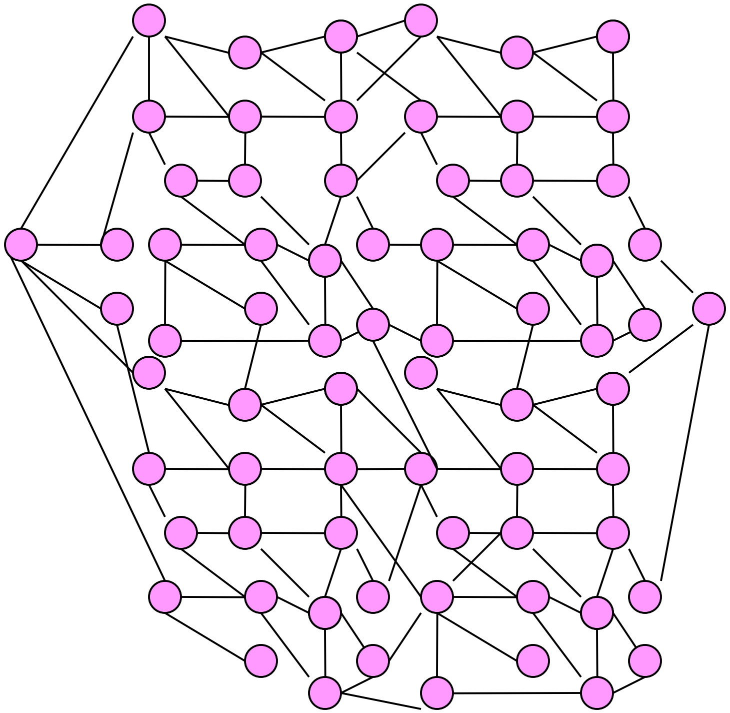


Result:

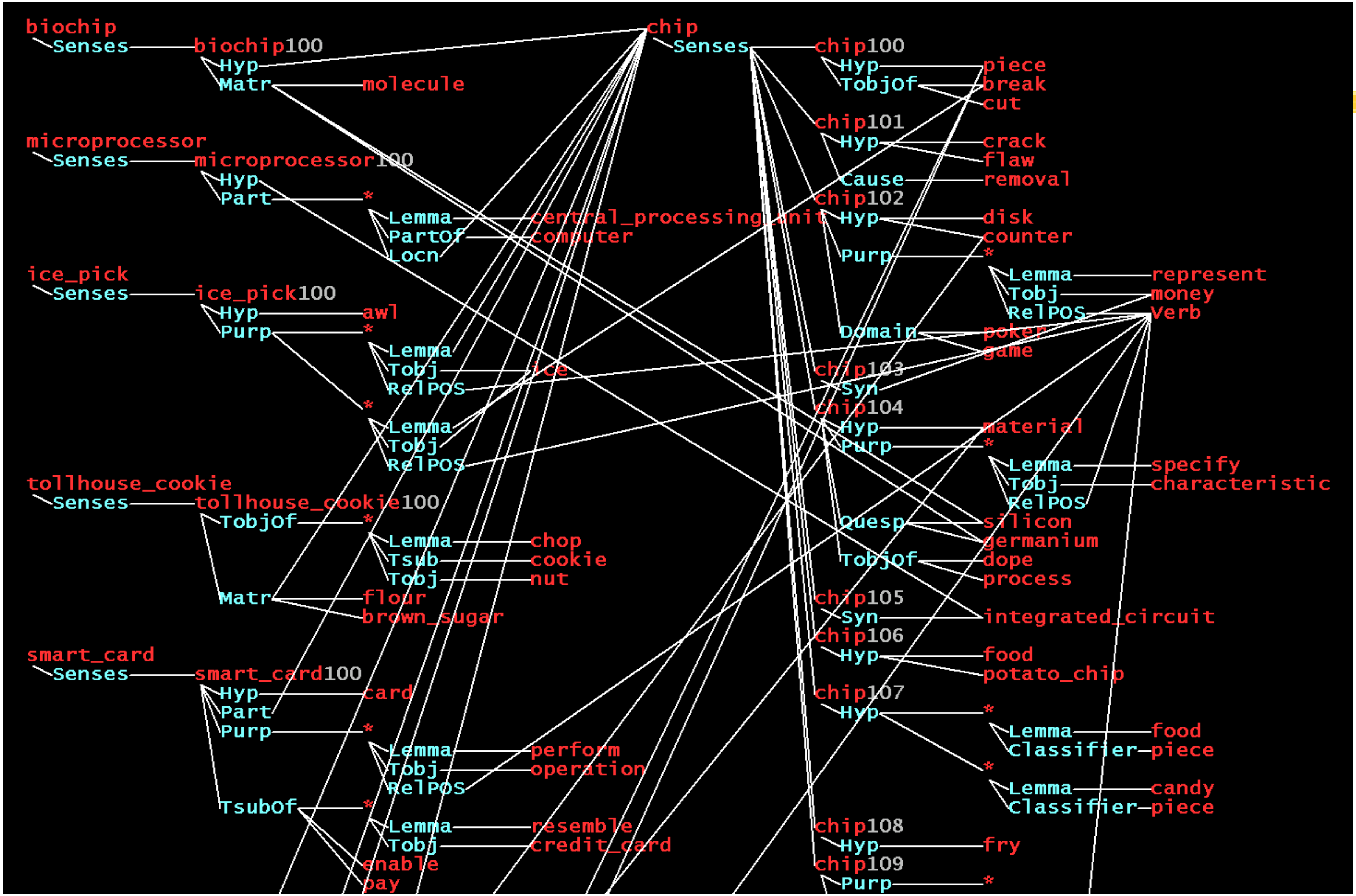
2-3 times as many relevant documents in the top 10 with Microsoft NLP

“MindNet”: Word Meanings

- A huge knowledge base
- Automatically created from dictionaries
- Words (nodes) linked by relationships
- 7 million links and growing



MindNet

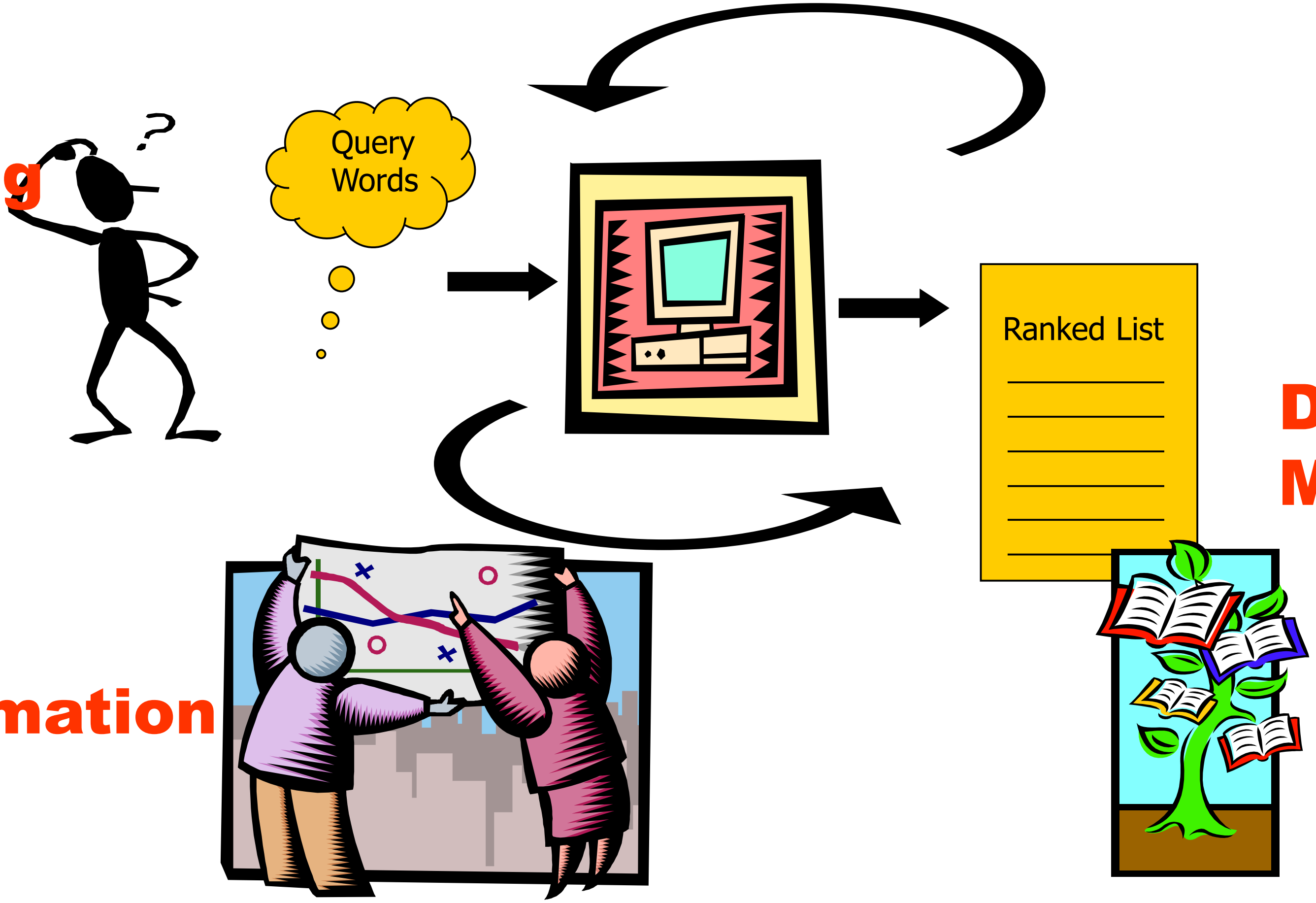

biochip
Senses
biochip100
Hyp
Matr
molecule
microprocessor
Senses
microprocessor100
Hyp
Part
*
Lemma
central_processing_unit
PartOf
computer
Locn
ice_pick
Senses
ice_pick100
Hyp
awl
Purp
*
Lemma
Tobj
ice
RelPOS
*
Lemma
Tobj
RelPOS
tollhouse_cookie
Senses
tollhouse_cookie100
TobjOf
*
Lemma
chop
Tsub
cookie
Tobj
nut
Matr
flour
brown_sugar
smart_card
Senses
smart_card100
Hyp
card
Part
Purp
*
Lemma
perform
Tobj
operation
RelPOS
TsubOf
*
Lemma
resemble
Tobj
credit_card
enable
pay
chip
Senses
chip100
Hyp
piece
TobjOf
break
cut
chip101
Hyp
crack
flaw
Cause
removal
chip102
Hyp
disk
counter
Purp
*
Lemma
represent
Tobj
money
RelPOS
Verb
Domain
poker
game
chip103
Syn
chip104
Hyp
material
Purp
*
Lemma
specify
Tobj
characteristic
RelPOS
Quesp
silicon
germanium
TobjOf
dope
process
chip105
Syn
integrated_circuit
chip106
Hyp
food
potato_chip
chip107
Hyp
*
Lemma
food
Classifier
piece
*
Lemma
candy
Classifier
piece
chip108
Hyp
fry
chip109
Purp
*

Beyond Content Matching

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

Broader View of IR

User Modeling



Query Words

Ranked List

Domain Modeling

Information Use

Beyond Content Matching

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

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

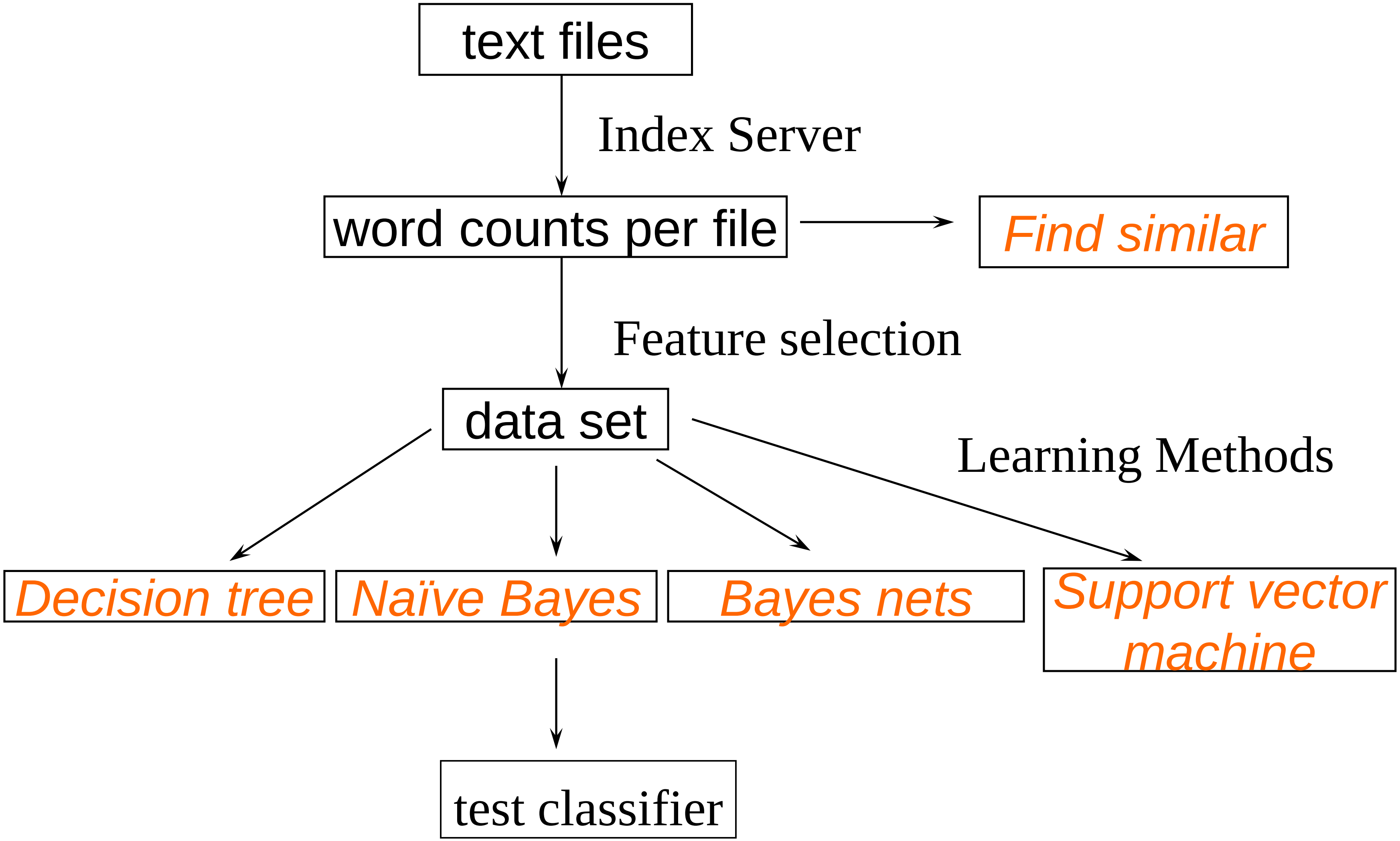
Classifiers

- A classifier is a function: f(**x**) = conf(class)
  - *from* attribute vectors, $\mathbf{x}=(x_1,x_2, \ldots x_d)$
  - *to* target values, confidence(class)
- Example classifiers
  - if (interest AND rate) OR (quarterly), then confidence(*interest*) = 0.9
  - confidence(*interest*) = 0.3*interest + 0.4*rate + 0.1*quarterly

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)

Text Classification Process



text files

Index Server

word counts per file

Find similar

Feature selection

data set

Learning Methods

Decision tree

Naïve Bayes

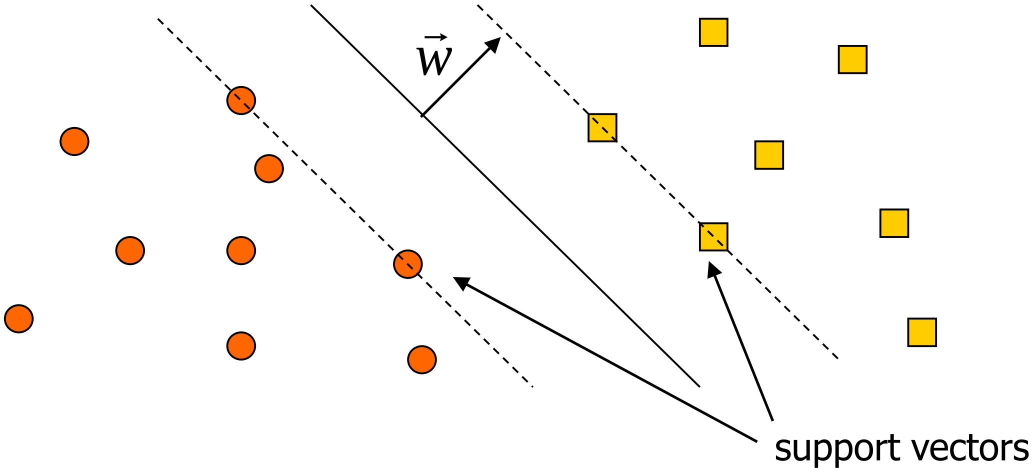
Bayes nets

Support vector machine

test classifier

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: $f(\vec{w}\cdot\vec{x})$



Support Vector Machines

- Extendable to:
  - Non-separable problems (Cortes & Vapnik, 1995)
  - Non-linear classifiers (Boser et al., 1992)
- Good generalization performance
  - Handwriting recognition (LeCun et al.)
  - Face detection (Osuna et al.)
  - Text classification (Joachims, Dumais et al.)
- Platt's Sequential Minimal Optimization algorithm very efficient

Reuters Data Set (21578 - ModApte split)

- 9603 training articles; 3299 test articles
- Example “interest” article

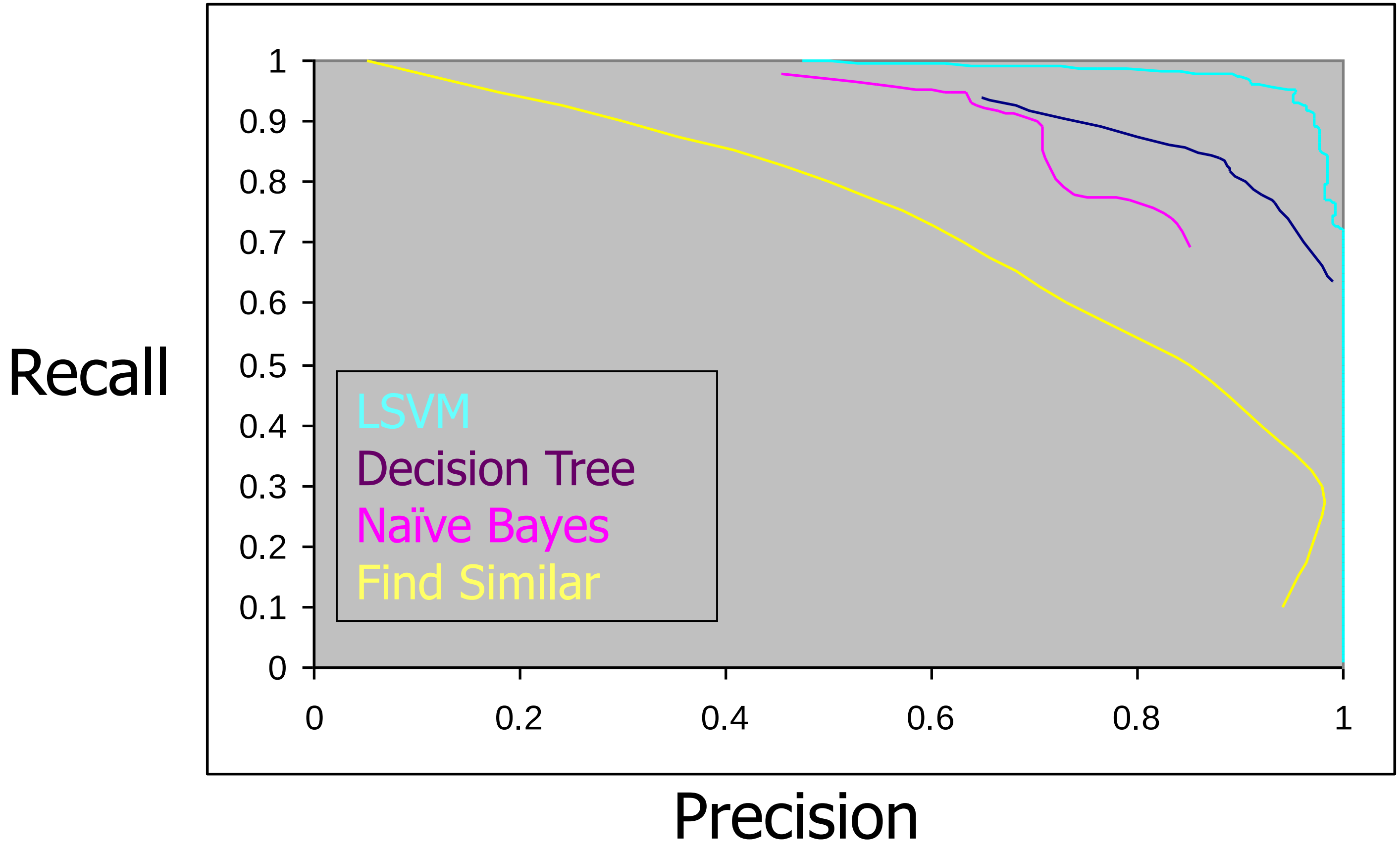
  2-APR-1987 06:35:19.50
  west-germany
  b f BC-BUNDESBANK-LEAVES-CRE 04-02 0052
  FRANKFURT, March 2
  The Bundesbank left credit policies unchanged after today's regular meeting of its council, a spokesman said in answer to enquiries. The West German discount rate remains at 3.0 pct, and the Lombard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

Reuters ROC - Category *Grain*



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

Precision: % really in category among those stories labeled in category

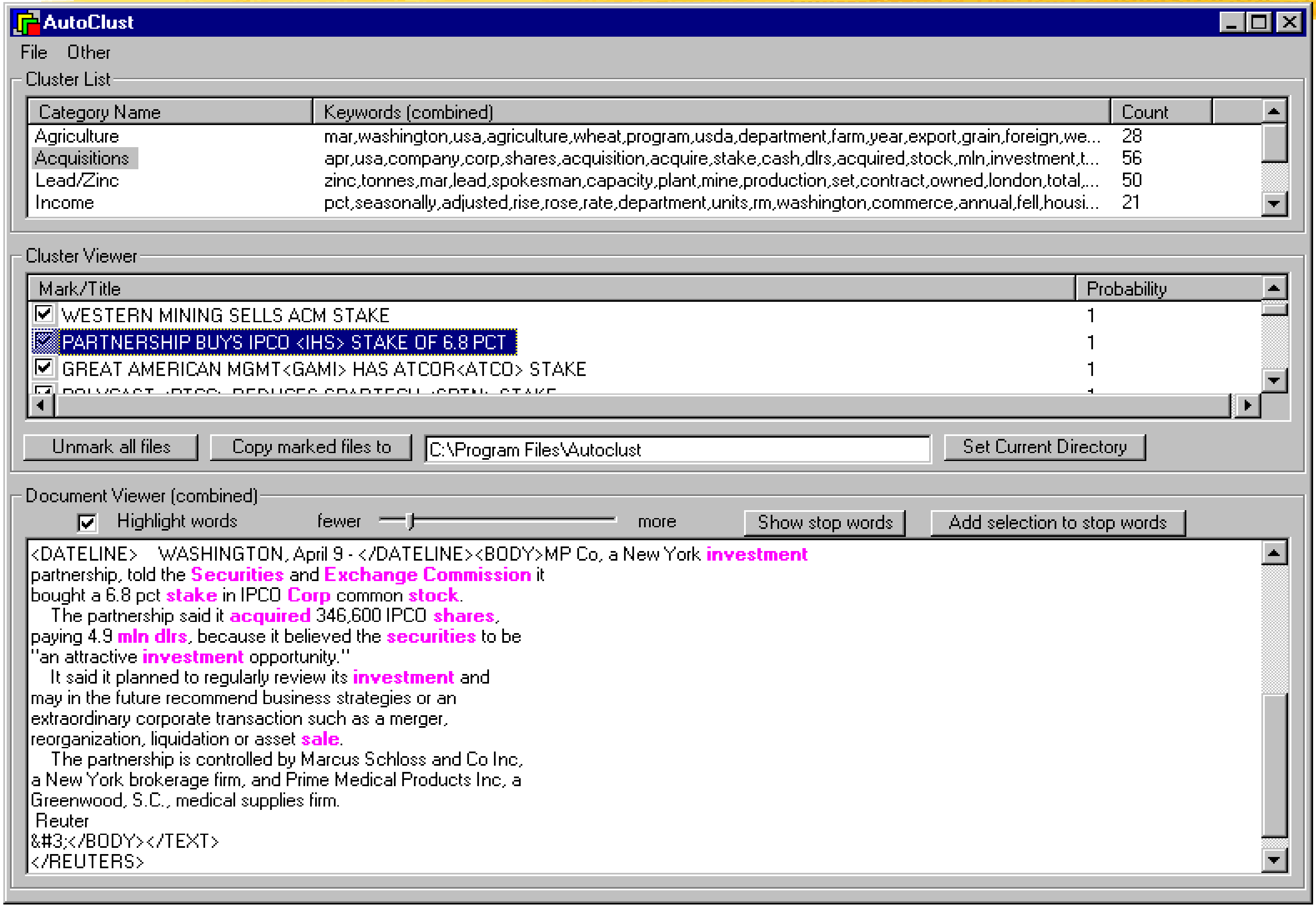
Text Categ Summary

- 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

Text Clustering

- Discovering structure
  - Vector-based document representation
  - EM algorithm to identify clusters
- Interactive user interface

Text Clustering



Category Name	Keywords (combined)	Count
Agriculture	mar,washington,usa,agriculture,wheat,program,usda,department,farm,year,export,grain,foreign,we...	28
Acquisitions	apr,usa,company,corp,shares,acquisition,acquire,stake,cash,dlrs,acquired,stock,mln,investment,t...	56
Lead/Zinc	zinc,tonnes,mar,lead,spokesman,capacity,plant,mine,production,set,contract,owned,london,total,...	50
Income	pct,seasonally,adjusted,rise,rose,rate,department,units,rm,washington,commerce,annual,fell,housi...	21

Mark/Title	Probability
☑ WESTERN MINING SELLS ACM STAKE	1
☑ PARTNERSHIP BUYS IPCO <IHS> STAKE OF 6.8 PCT	1
☑ GREAT AMERICAN MGMT<GAMI> HAS ATCOR<ATCO> STAKE	1

C:\Program Files\Autoclust

☑ Highlight words

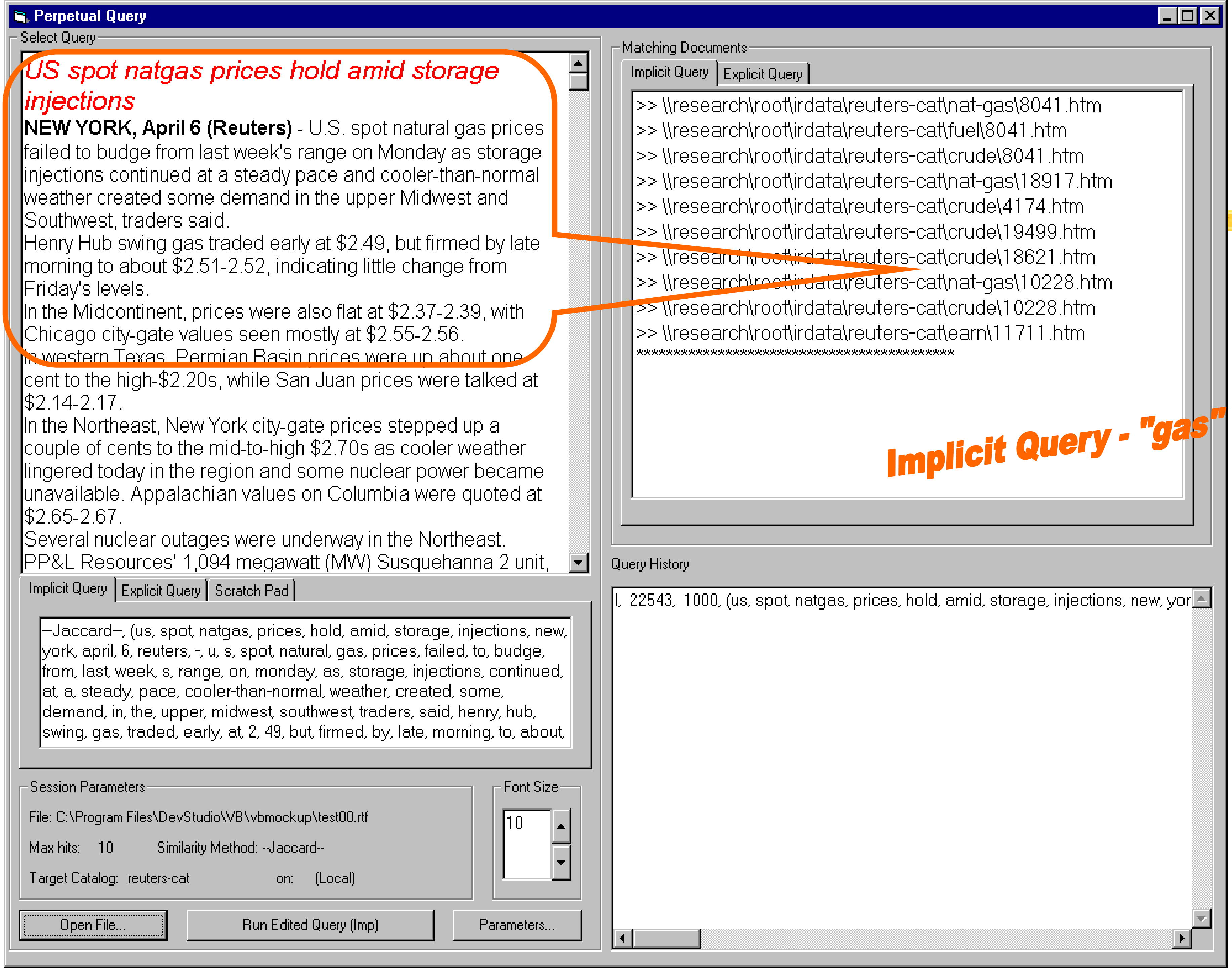
```
<DATELINE>    WASHINGTON, April 9 - </DATELINE><BODY>MP Co, a New York investment
partnership, told the Securities and Exchange Commission it
bought a 6.8 pct stake in IPCO Corp common stock.
    The partnership said it acquired 346,600 IPCO shares,
paying 4.9 mln dlrs, because it believed the securities to be
"an attractive investment opportunity."
    It said it planned to regularly review its investment and
may in the future recommend business strategies or an
extraordinary corporate transaction such as a merger,
reorganization, liquidation or asset sale.
    The partnership is controlled by Marcus Schloss and Co Inc,
a New York brokerage firm, and Prime Medical Products Inc, a
Greenwood, S.C., medical supplies firm.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Beyond Content Matching

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

Implicit Queries (IQ)

- Explicit queries:
  - Search is a separate, discrete task
  - User types query, Gets results, Tries again …
- **Implicit queries**:
  - Search as part of normal information flow
  - Ongoing query formulation based on user activities, and non-intrusive results display
  - Can include explicit query or push profile, but doesn’t require either



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,

```
>> \\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"

```
--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
```

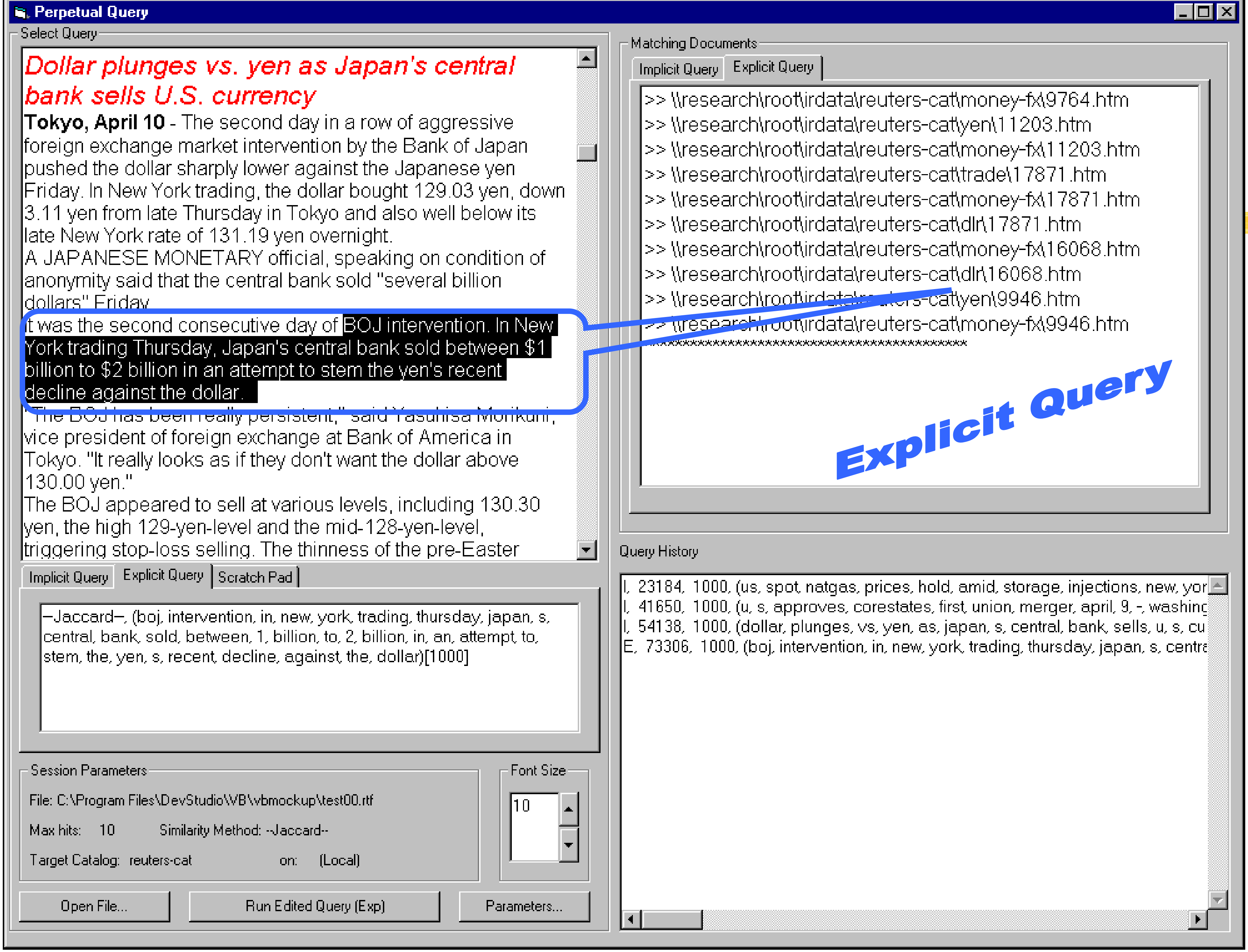
```
I, 22543, 1000, (us, spot, natgas, prices, hold, amid, storage, injections, new, yor
```

File: C:\Program Files\DevStudio\VB\vbmockup\test00.rtf

Max hits: 10 Similarity Method: --Jaccard--

Target Catalog: reuters-cat on: (Local)

Font Size: 10



Dollar plunges vs. yen as Japan's central bank sells U.S. currency

Tokyo, April 10 - The second day in a row of aggressive foreign exchange market intervention by the Bank of Japan pushed the dollar sharply lower against the Japanese yen Friday. In New York trading, the dollar bought 129.03 yen, down 3.11 yen from late Thursday in Tokyo and also well below its late New York rate of 131.19 yen overnight.

A JAPANESE MONETARY official, speaking on condition of anonymity said that the central bank sold "several billion dollars" Friday.

It was the second consecutive day of BOJ intervention. In New York trading Thursday, Japan's central bank sold between $1 billion to $2 billion in an attempt to stem the yen's recent decline against the dollar.

"The BOJ has been really persistent," said Yasuhisa Morikuni, vice president of foreign exchange at Bank of America in Tokyo. "It really looks as if they don't want the dollar above 130.00 yen."

The BOJ appeared to sell at various levels, including 130.30 yen, the high 129-yen-level and the mid-128-yen-level, triggering stop-loss selling. The thinness of the pre-Easter

```
>> \\research\root\irdata\reuters-cat\money-fx\9764.htm
>> \\research\root\irdata\reuters-cat\yen\11203.htm
>> \\research\root\irdata\reuters-cat\money-fx\11203.htm
>> \\research\root\irdata\reuters-cat\trade\17871.htm
>> \\research\root\irdata\reuters-cat\money-fx\17871.htm
>> \\research\root\irdata\reuters-cat\dlr\17871.htm
>> \\research\root\irdata\reuters-cat\money-fx\16068.htm
>> \\research\root\irdata\reuters-cat\dlr\16068.htm
>> \\research\root\irdata\reuters-cat\yen\9946.htm
>> \\research\root\irdata\reuters-cat\money-fx\9946.htm
**************************************************
```

Explicit Query

```
I, 23184, 1000, (us, spot, natgas, prices, hold, amid, storage, injections, new, yor
I, 41650, 1000, (u, s, approves, corestates, first, union, merger, april, 9, -, washinc
I, 54138, 1000, (dollar, plunges, vs, yen, as, japan, s, central, bank, sells, u, s, cu
E, 73306, 1000, (boj, intervention, in, new, york, trading, thursday, japan, s, centra
```

--Jaccard--, (boj, intervention, in, new, york, trading, thursday, japan, s, central, bank, sold, between, 1, billion, to, 2, billion, in, an, attempt, to, stem, the, yen, s, recent, decline, against, the, dollar)[1000]

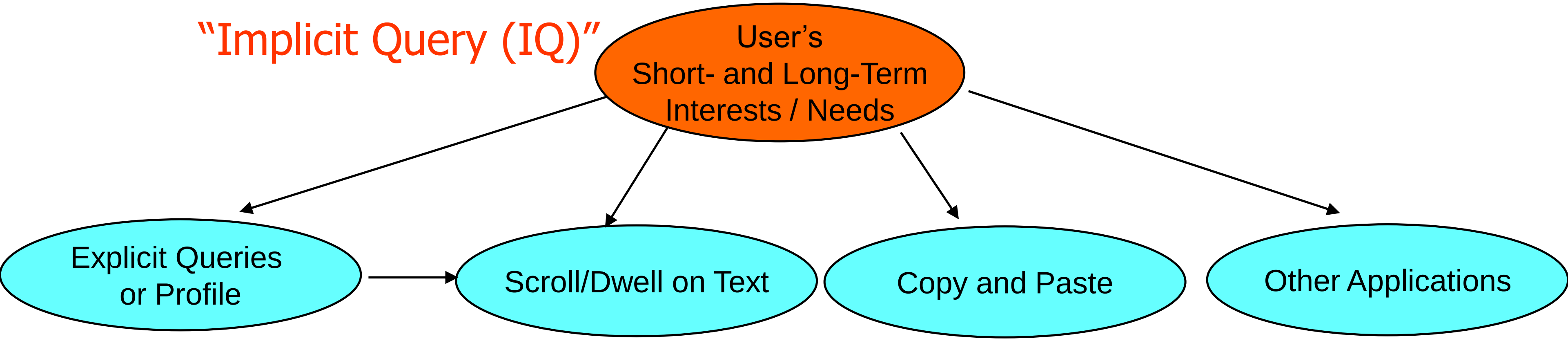
File: C:\Program Files\DevStudio\VB\vbmockup\test00.rtf

Max hits: 10 Similarity Method: --Jaccard--

Target Catalog: reuters-cat on: (Local)

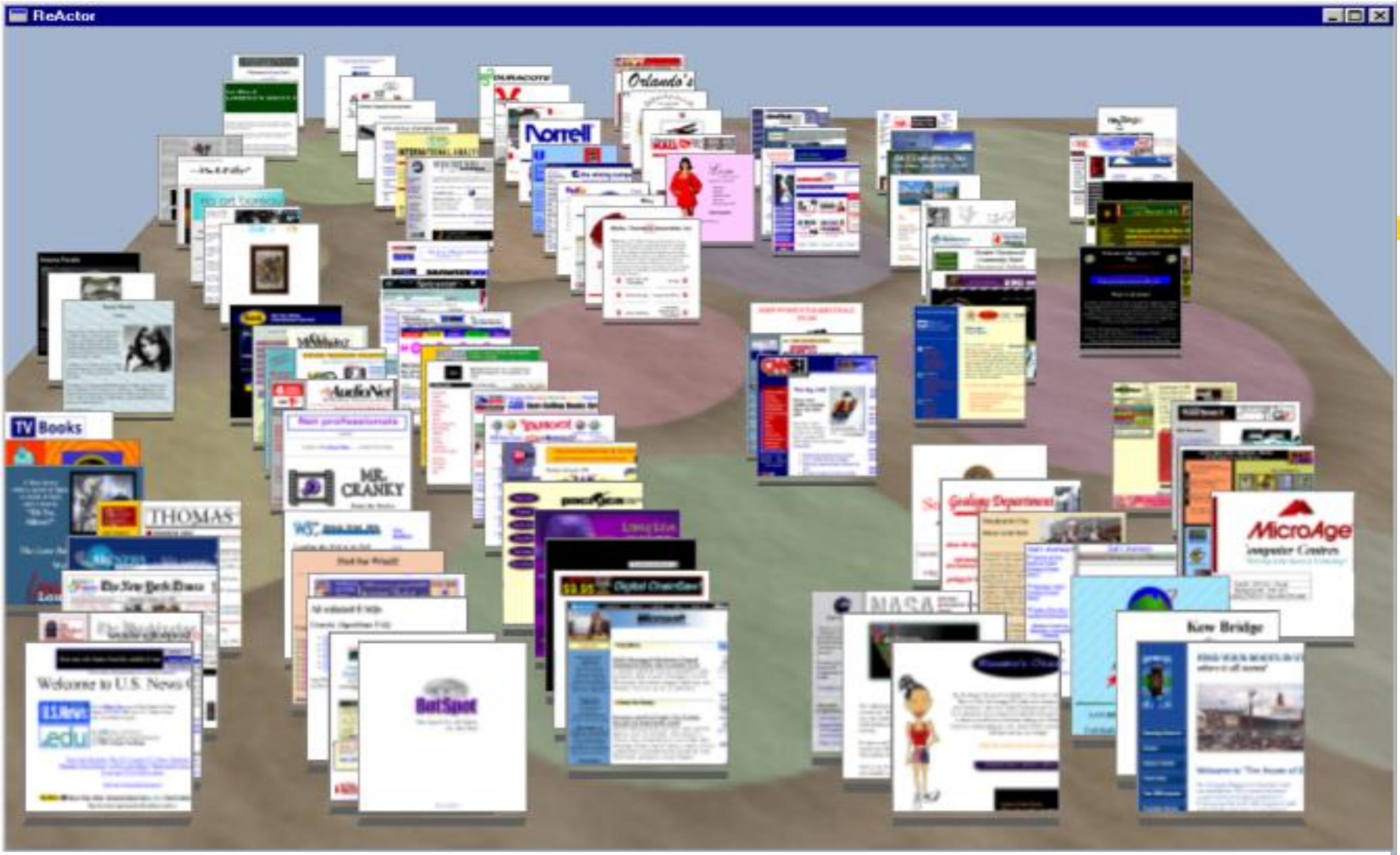
User Modeling for IQ/IR

- IQ: Model of user interests based on actions
  - Explicit search activity (query or profile)
  - Patterns of scroll / dwell on text
  - Copying and pasting actions
  - Interaction with multiple applications



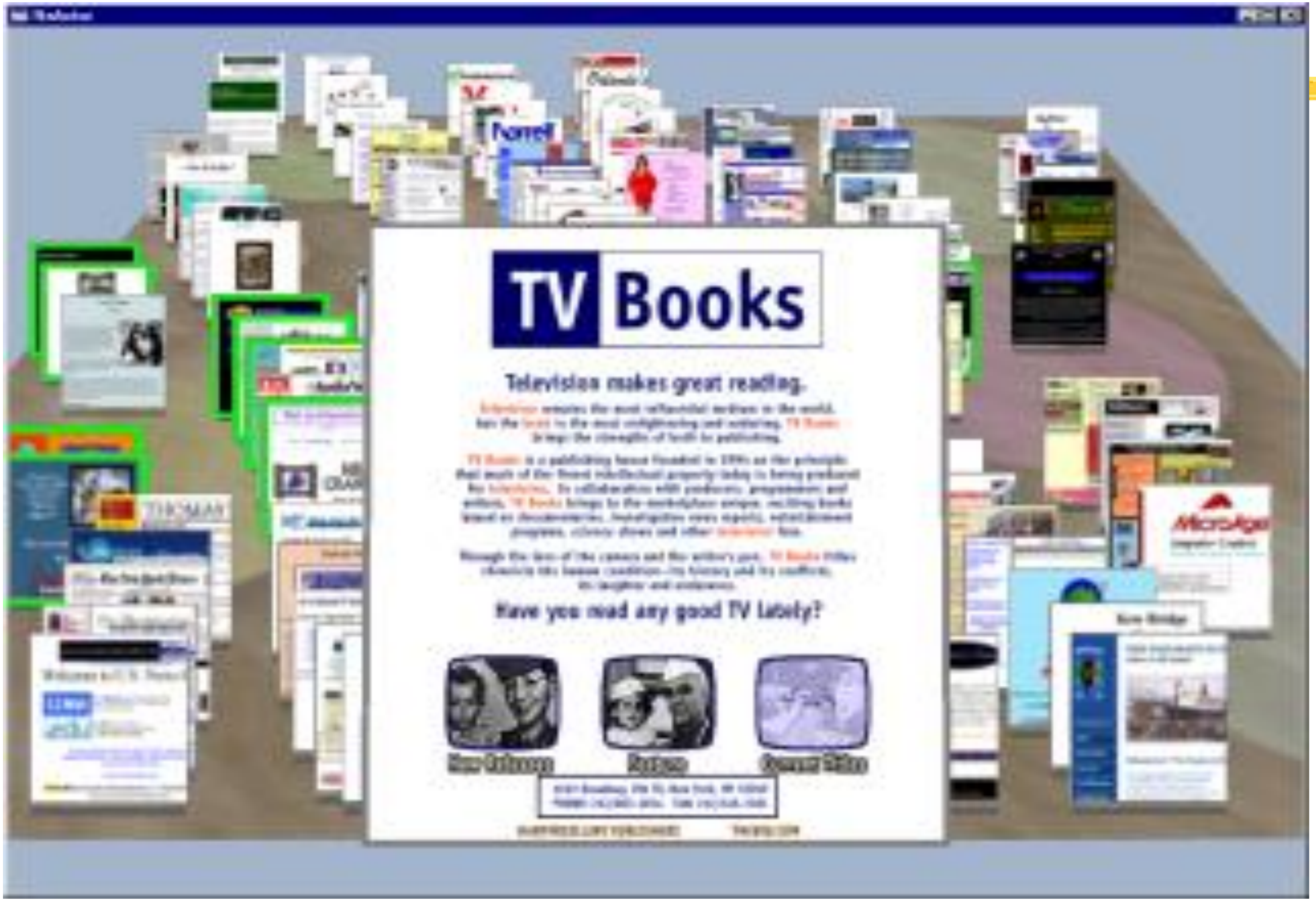
Implicit Query Highlights

- IQ built by tracking user’s reading behavior
  - No explicit search required
  - Good matches returned
- IQ user model:
  - Combines present context + previous interests
- New interfaces for tightly coupling search results with structure -- user study



Data Mountain with 100 web pages.

Data Mountain with Implicit Query results shown (highlighted pages to left of selected page).



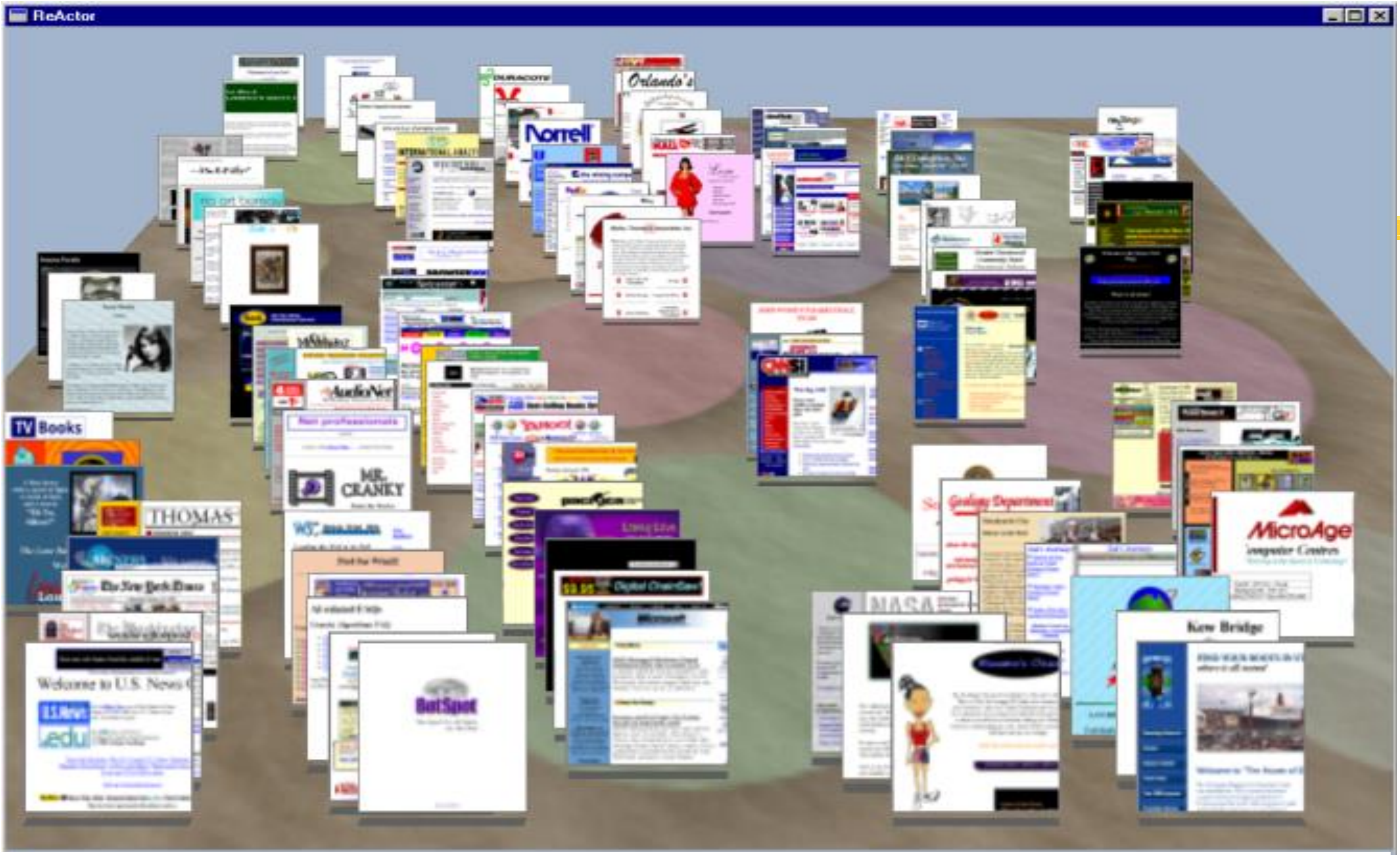
TV Books

Television makes great reading.

Have you read any good TV lately?

IQ Study: Experimental Details

- Store 100 Web pages
  - 50 popular Web pages; 50 random pages
  - With or without Implicit Query
    - IQ1: Co-occurrence based IQ
    - IQ2: Content-based IQ
- Retrieve 100 Web pages
  - Title given as retrieval cue -- e.g., "CNN Home Page"
  - No implicit query highlighting at retrieval



Find: “CNN Home Page”

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

⌘Number of categories

IQ Condition	Average Number of Categories (std in parens)
IQ0: No IQ	9.3 (3.6)
IQ1: Co-occur based	15.6 (5.8)
IQ2: Content-based	12.8 (4.9)

Results: Retrieval Time

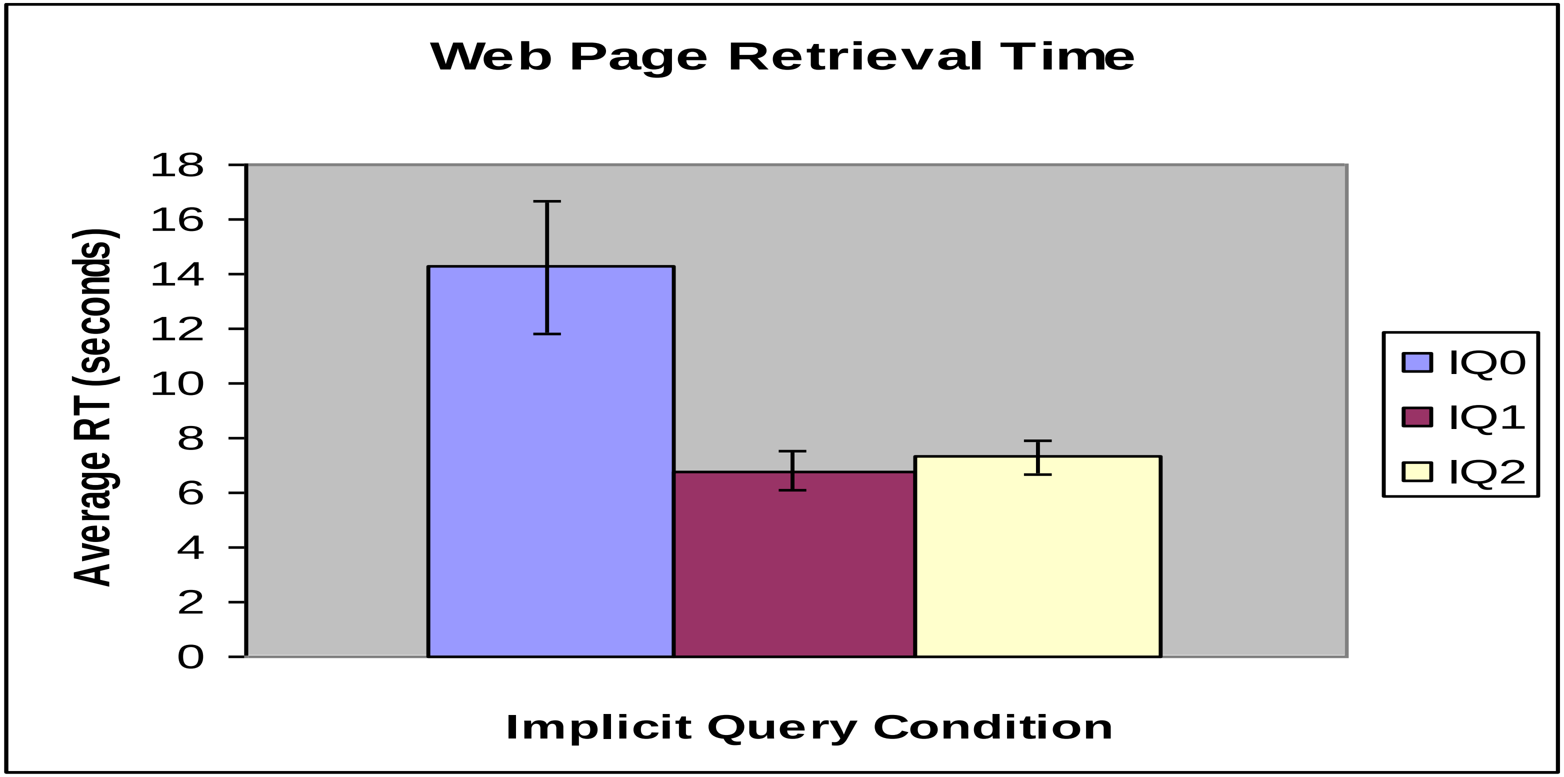
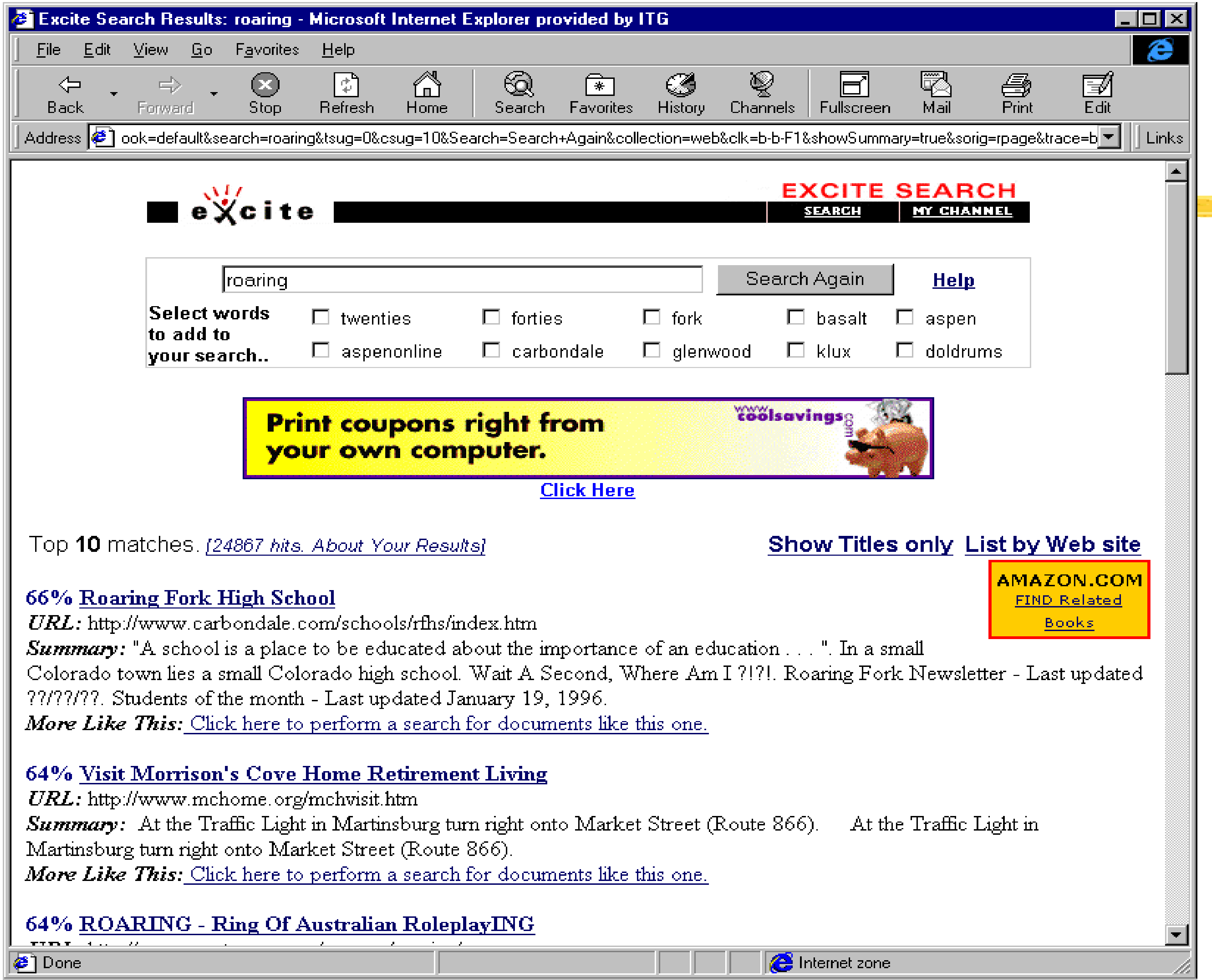


Figure 3. Average web page retrieval time, including standard error of the mean, for each Implicit Query condition.

Example Web Searches

user = A1D6F19DB06BD694 ***date = 970916*** ***excite log***

150052	lion
152004	lions
152036	lions lion
152219	lion facts
153747	roaring
153848	lions roaring
160232	africa lion
160642	lions, tigers, leopards and cheetahs
161042	lions, tigers, leopards and cheetahs cats
161144	wild cats of africa
161414	africa cat
161602	africa lions
161308	africa wild cats
161823	mane
161840	lion
161858	lion lions
163041	lion facts
163919	picher of lions
164040	lion picher
165002	lion pictures
165100	pictures of lions
165211	pictures of big cats
165311	lion photos
170013	video in lion
172131	pictureof a lioness
172207	picture of a lioness
172241	lion pictures
172334	lion pictures cat
172443	lions
172450	lions


Excite Search Results: roaring - Microsoft Internet Explorer provided by ITG
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit
Address
Links
excite
EXCITE SEARCH
SEARCH
MY CHANNEL
Search Again
Help
Done
Internet zone

ook=default&search=roaring&tsug=0&csug=10&Search=Search+Again&collection=web&clk=b-b-F1&showSummary=true&sorig=rpage&trace=b

roaring

Select words to add to your search..

□ twenties □ forties □ fork □ basalt □ aspen

□ aspenonline □ carbondale □ glenwood □ klux □ doldrums

Top **10** matches. *[24867 hits. About Your Results]*

Show Titles only **List by Web site**

66% Roaring Fork High School

URL: http://www.carbondale.com/schools/rfhs/index.htm

Summary: "A school is a place to be educated about the importance of an education . . . ". In a small Colorado town lies a small Colorado high school. Wait A Second, Where Am I ?!?!. Roaring Fork Newsletter - Last updated ??/??/??. Students of the month - Last updated January 19, 1996.

More Like This: Click here to perform a search for documents like this one.

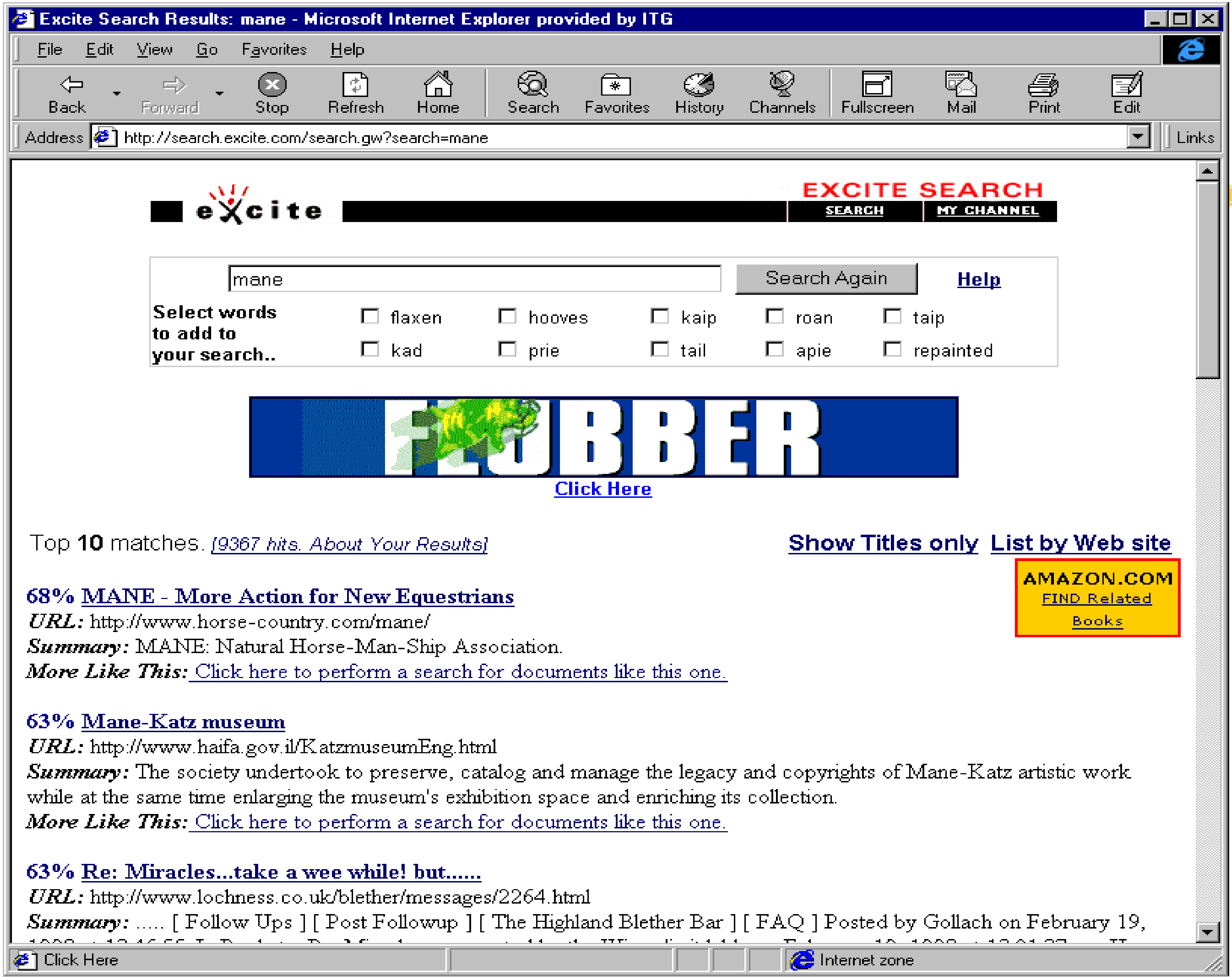
64% Visit Morrison's Cove Home Retirement Living

URL: http://www.mchome.org/mchvisit.htm

Summary: At the Traffic Light in Martinsburg turn right onto Market Street (Route 866). At the Traffic Light in Martinsburg turn right onto Market Street (Route 866).

More Like This: Click here to perform a search for documents like this one.

64% ROARING - Ring Of Australian RoleplayING



mane

Search Again

Help

Select words to add to your search..

□ flaxen □ hooves □ kaip □ roan □ taip
□ kad □ prie □ tail □ apie □ repainted

Top **10** matches. *[9367 hits. About Your Results]*

Show Titles only **List by Web site**

68% MANE - More Action for New Equestrians
URL: http://www.horse-country.com/mane/
Summary: MANE: Natural Horse-Man-Ship Association.
More Like This: Click here to perform a search for documents like this one.

63% Mane-Katz museum
URL: http://www.haifa.gov.il/KatzmuseumEng.html
Summary: The society undertook to preserve, catalog and manage the legacy and copyrights of Mane-Katz artistic work while at the same time enlarging the museum's exhibition space and enriching its collection.
More Like This: Click here to perform a search for documents like this one.

63% Re: Miracles...take a wee while! but......
URL: http://www.lochness.co.uk/blether/messages/2264.html
Summary: ..... [Follow Ups] [Post Followup] [The Highland Blether Bar] [FAQ] Posted by Gollach on February 19,

Click Here

Summary

- Rich IR research tapestry
- Improving content-matching
- And, beyond ...
  - Domain/Object Models
  - User/Task Models
  - Information Presentation and Use

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