Web Experiments and Test Collections

Susan Dumais
Microsoft Research
sdumais@microsoft.com

WWW11 – May 10, 2002

Are Web Experiments and Test Collections Meaningful?

- Yes, of course. How else would you evaluate progress ?
- But, …

Are Web Experiments and Test Collections Meaningful?

- Yes, of course. How else would you evaluate progress ?
- But, …
- What you need depends on what you want to learn about or ***generalize*** to:
  - E.g., If you want to build a good ***web search service***, you'll need a live collection with all the complications (crawling, changing content, spam detection, spelling correction, duplicate detection, results presentation, etc.)
  - E.g., If you want to develop a good ***spelling correction module***, you don't need much of a collection, but you do need lots of queries
  - E.g., If you want to develop a good ***ranking algorithm***, you'll need a sizable collection of representative queries and documents. 10s of millions of documents should suffice. Ideally several such samples.

A Non-Web Example

- Study 1: An asprin a day, increases longevity by 20%
- Study 2: Asprin, does not increase longevity, and increases ulcers by 40%
- Should you take asprin?
- Questions about quality, comparability and applicability:
  - Who are the subjects (age, sex, diet, health, etc.)?
  - How many subjects?
  - Dosage of asprin?
  - Length of study?
  - …

Why Web Test Collections?

- Needed to interpret:
  - System A "90% precision in top 1"; System B "85% accuracy"
  - Existing algorithm (or system), Modify it … Does it work?
- Need comparability on:
  - Collections
    - E.g., *Size*: Higher precision-oriented scores with larger collections; Changes in graphical properties of collection
  - Queries/Tasks
    - E.g., *Types of queries*, but high variance even within type
  - Performance measures
- More confidence in the generalizability of findings if the same techniques work on different collections and queries, or if you understand why they don't
- Make your collections/queries available …

Why Web Test Collections?

- Needed to interpret:
  - System A "90% precision in top 1"; System B "85% accuracy"
  - Existing algorithm (or system), Modify it … Does it work?
- Need comparability on:
  - Collections
    - E.g., *Size*: Higher precision-or
      in graphical properties of colle
  - Queries/Tasks
    - E.g., *Types of queries*, but high v
  - Performance measures

Hawking and Robertson, 2001

Collection	QuerySet	Pr@20 Full	Pr@20 Sample	Ratio
Trec-6	Q6M	0.467	0.169	2.76
Trec-6	Q6A	0.294	0.144	2.04
Trec-3	Q3A	0.436	0.235	1.86

- More confidence in the generalizability of findings if the same techniques work on different collections and queries, or if you understand why they don't
- Make your collections/queries available …

Why Web Test Collections?

- Needed to interpret:
  - System A "90% precision in top 1"; System B "85% accuracy"
  - Existing algorithm (or system), Modify it … Does it work?
- Need comparability on:
  - Collections
    - E.g., *Size*: Higher precision-or
      in graphical properties of colle
  - Queries/Tasks
    - E.g., *Types of queries*,
  - Performance measure
- More confidence in the g techniques work on different collections and queries, or if you understand why they don't
- Make your collections/queries available …

Hawking and Robertson, 2001

Collection/QuerySet	Pr@20 Full	Pr@20 Sample	Ratio
			.76
			.04
			.86

Westerveld et al., TREC 2001

	Base	+Anchor	+URL
Topical Queries, Pr@5	0.36	0.36	
Home Page Queries, MRR	0.34	0.45	0.77

<also, Broder; Singhal and Kaszkiel>

Why Web Test Collections?

- Needed to interpret:
  - System A "90% precision in top 1"; System B "85% accuracy"
  - Existing algorithm (or system), Modify it … Does it work?
- Need comparability on:
  - Collections
    - E.g., *Size*: Higher precision-or
      in graphical properti
  - Queries/Tasks
    - E.g., *Types of queries*,
  - Performance measure
- More confidence in the g
  techniques work on diffe
  understand why they don't
- Make your collections/queries available …

Hawking and Robertson, 2001

Jin and Dumais, SIGIR 2001

Westerveld et al., TREC 2001

	Base	+Anchor	+URL	
Topical Queries, Pr@5	0.36	0.36		0.04
Home Page Queries, MRR	0.34	0.45	0.77	0.11

<also, Broder; Singhal and Kaszkiel>

Pr@10, Rel + ExRel	0.521	0.530	0.02	0.520	0.00
Pr@10, ExRel	0.176	0.168	-0.04	0.177	0.00

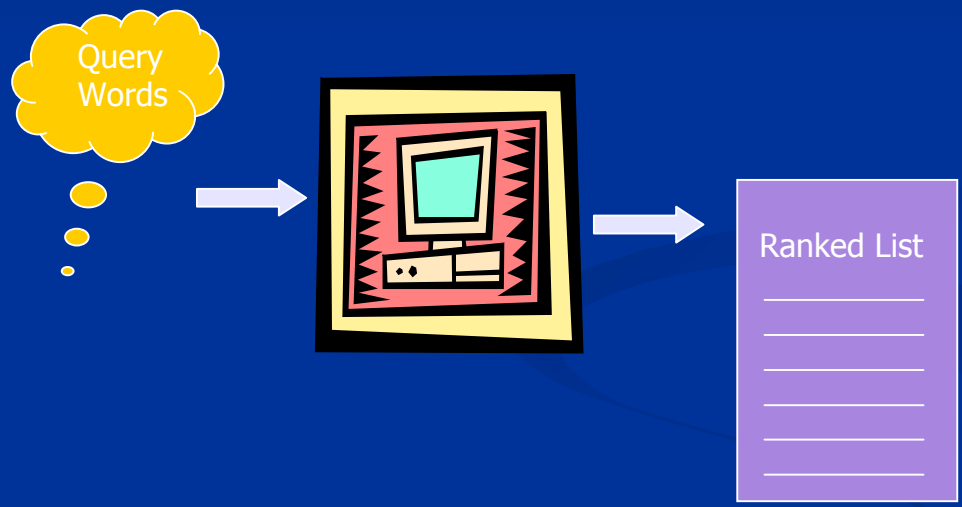
Evolution of TREC Web Collection

- *Content*: news -> web
- *Relevance judgments*: binary -> multi-valued
- *Size*: 2 million -> 18 million web pages
- *Query types*: topical -> home page finding

- … steps in the right direction, but still very much relevance and ranking focused

Web Search Evaluation

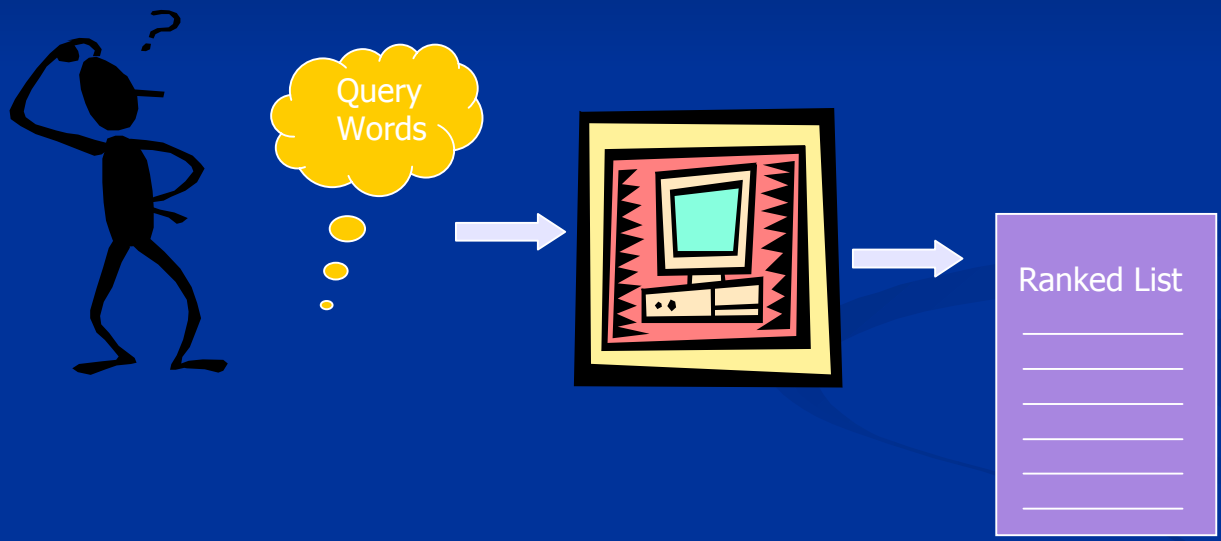
- Relevance and ranking are important, but not all …
- What's missing?



Web Search Evaluation

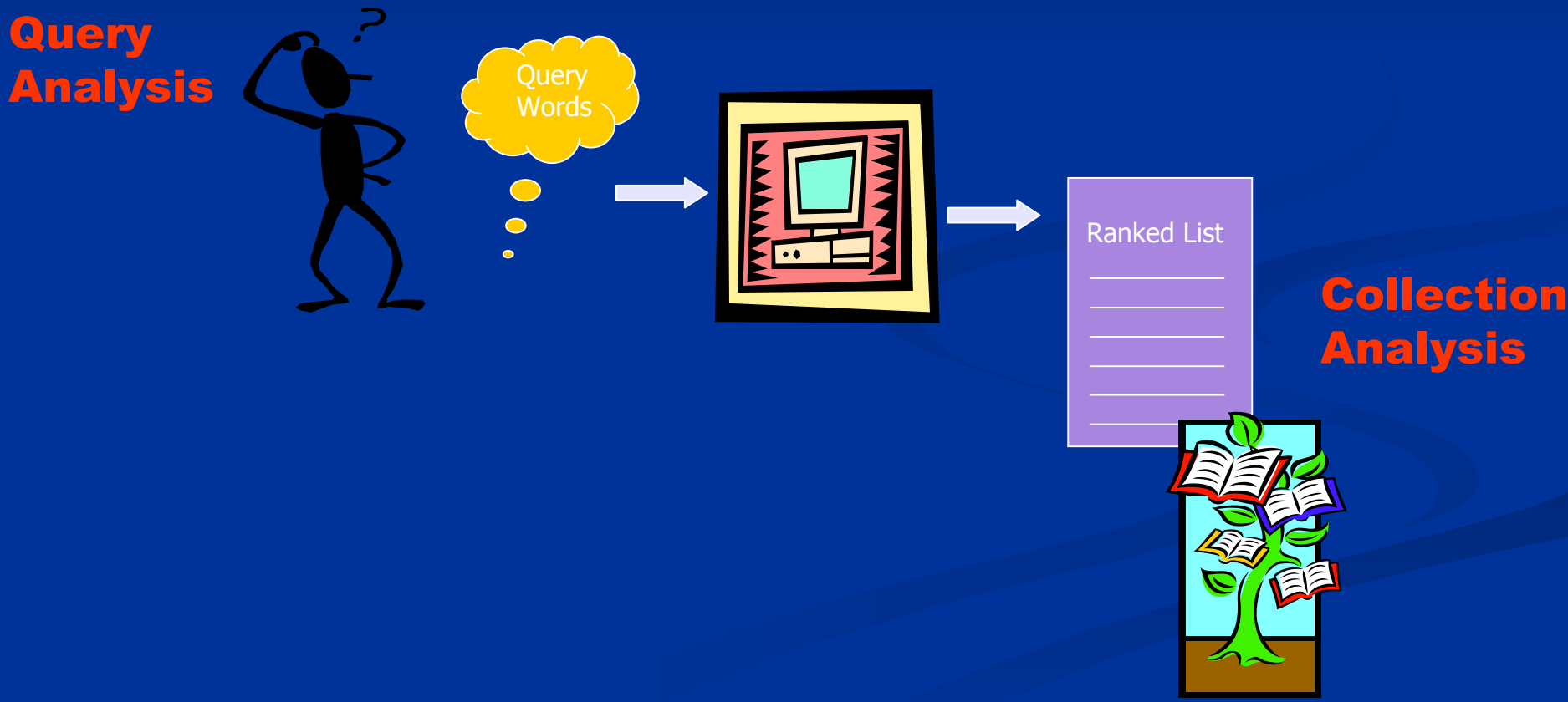
- Relevance and ranking are important, but not all …
- What's missing?

Query Analysis



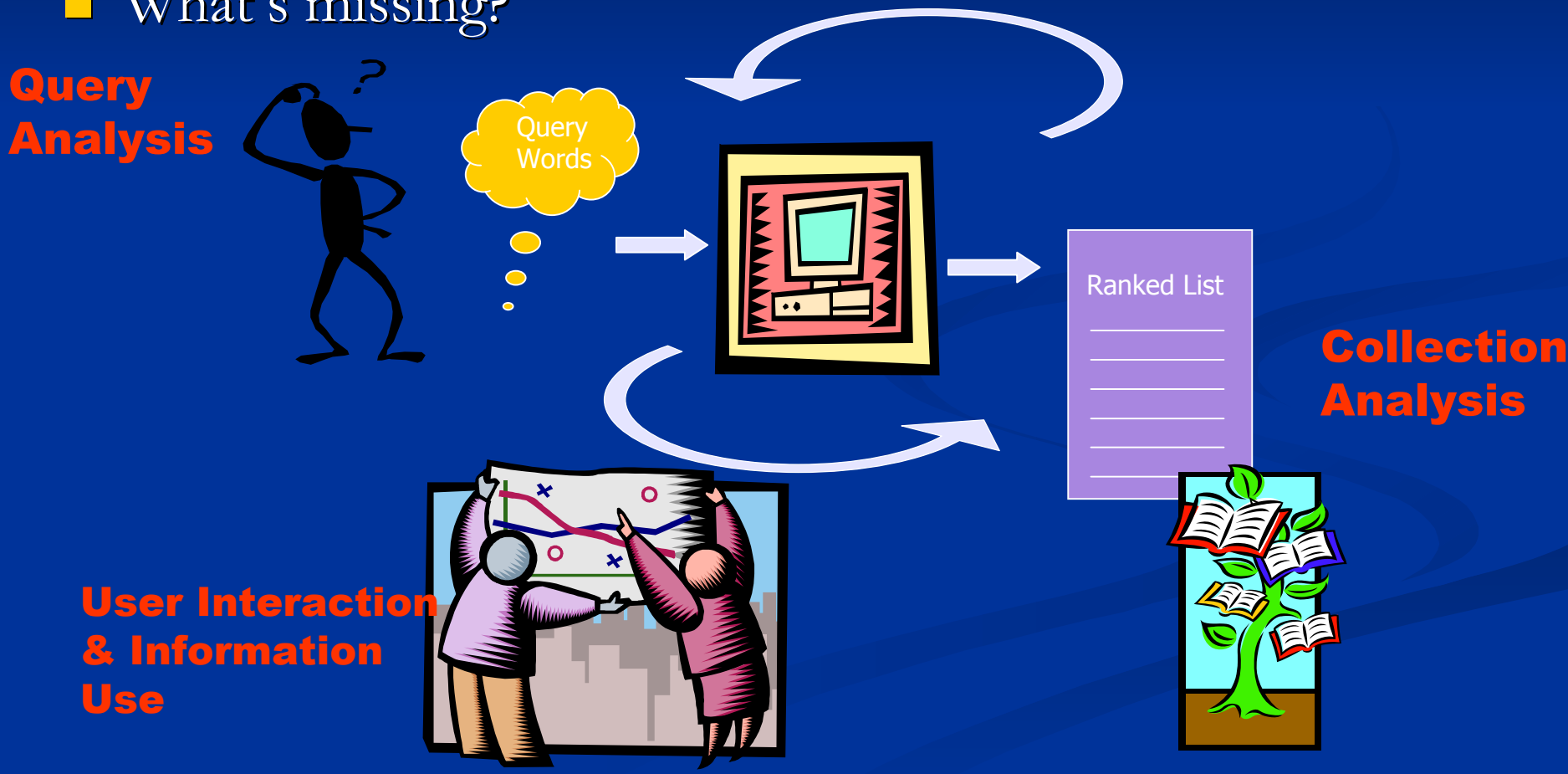
Web Search Evaluation

- Relevance and ranking are important, but not all …
- What's missing?



Web Search Evaluation

- Relevance and ranking are important, but not all …
- What's missing?



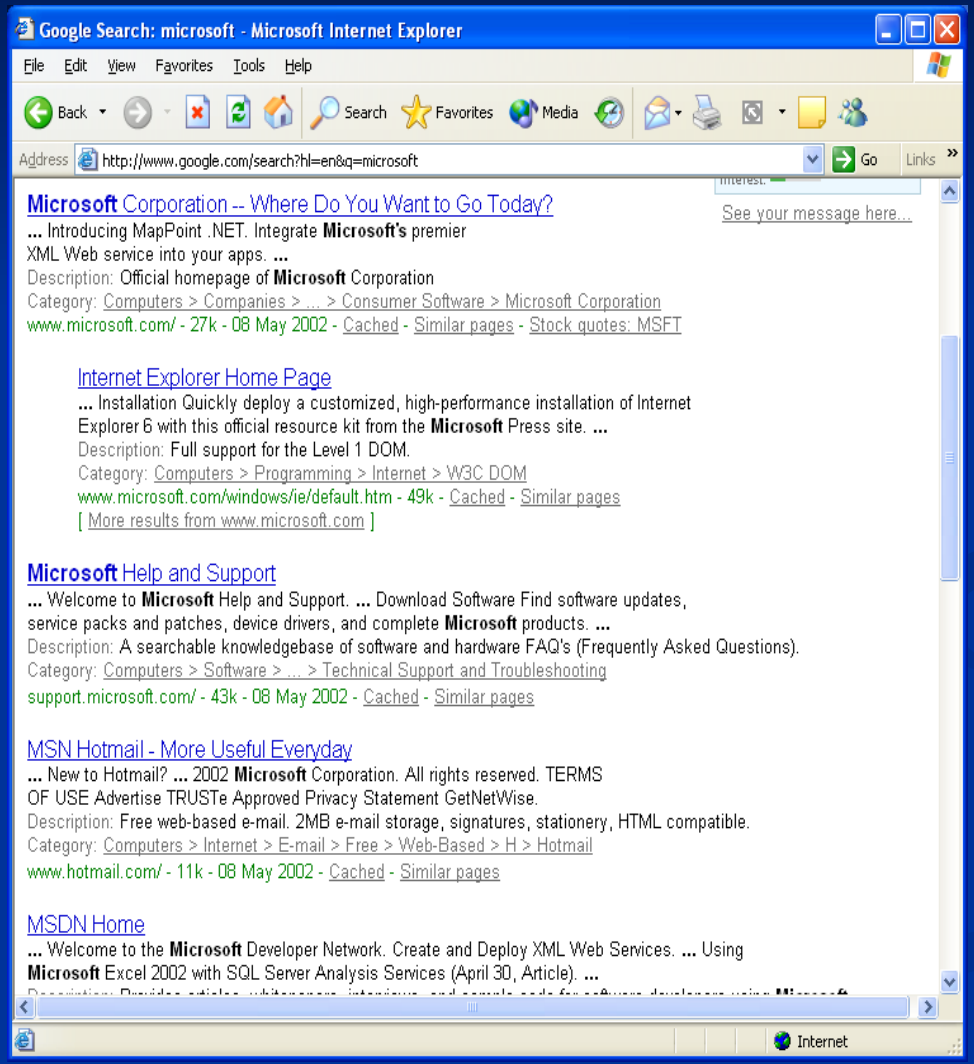
Web Search Evaluation

- Relevance and ranking are important, but not all …
- Beyond "static relevance judgments"
  - Query analysis
    - Types of queries, spelling correction, query formulation
    - Task context: topical (or other) relevance, recall/prec focus
  - Collection analysis
    - Size, graphical properties, variety of content types
  - Results presentation and user interaction
    - Breadth/variety of results -> Grouping
    - Time to find correct answer, or user satisfaction

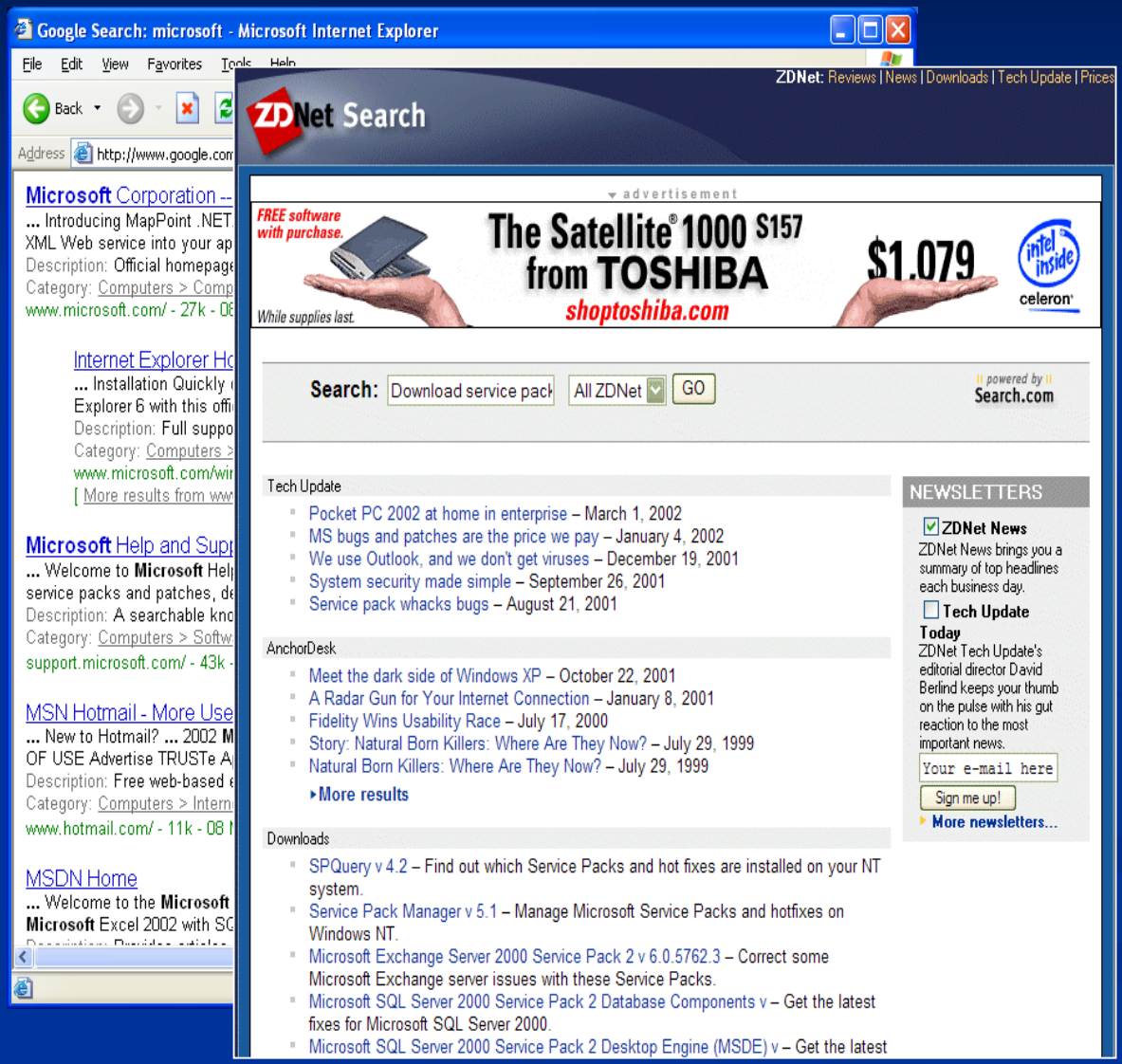
Web Search Evaluation (cont'd)

- Many evaluation techniques
  - Explicit judgments of relevance or quality
  - Implicit measures like click through data (e.g., Joachims for tuning ranking parameters)
  - Usability studies to look at satisfaction, time, etc. (e.g., Dumais et al. for presentation ideas)
  - Efficiency, Cost, Features list, Marketplace, …
- Many of these involve interacting with users, but you can study this systematically in the lab or in situ
- Can evaluate end-to-end systems, or components (ranking, spelling correction, de-dup, presentation, etc.)

e.g., Usefulness of Grouping



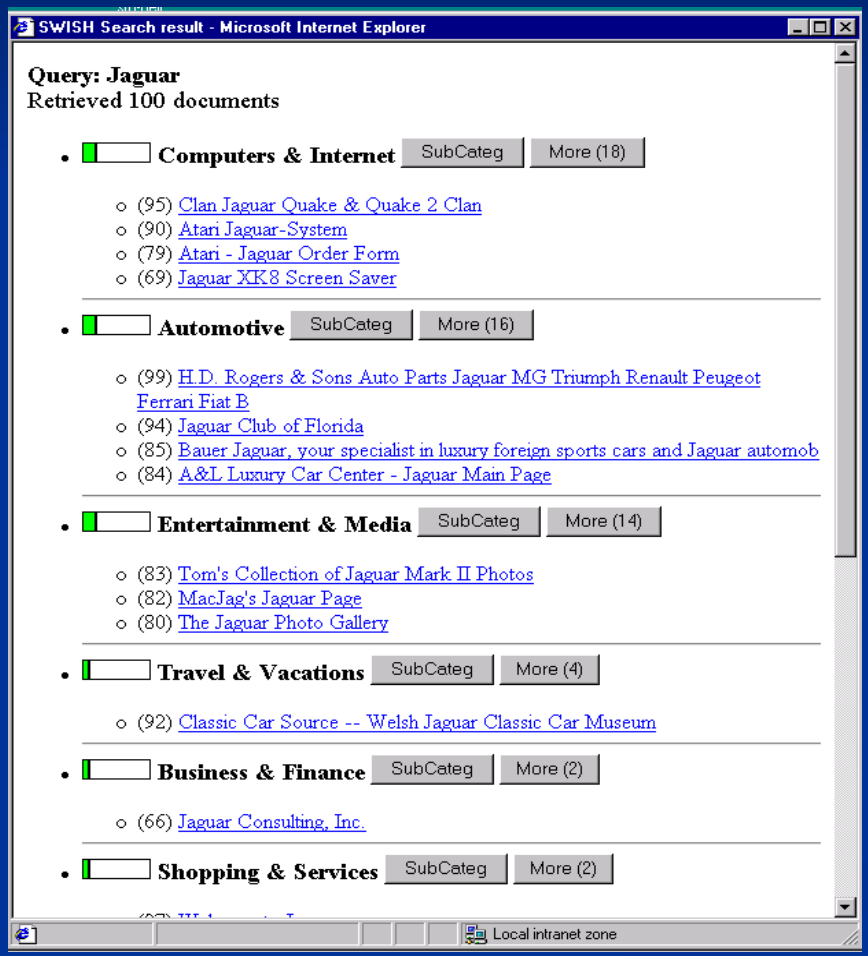
e.g., Usefulness of Grouping



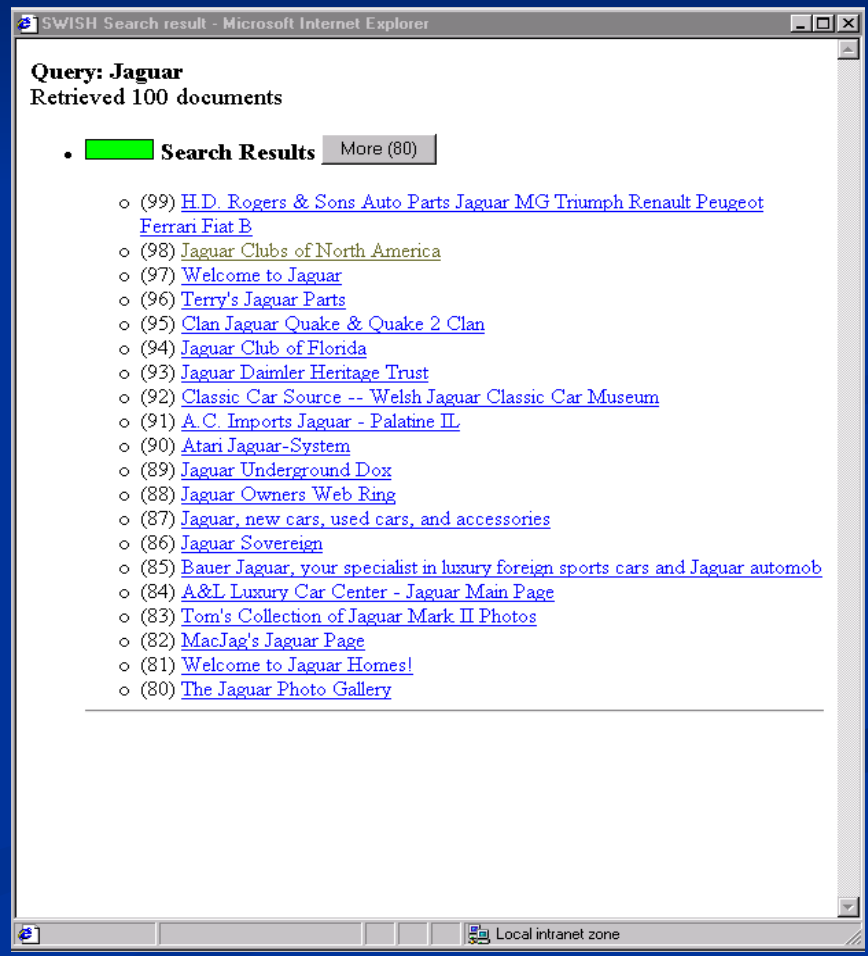
e.g., Usefulness of Grouping

Chen & Dumais (CHI'2000)

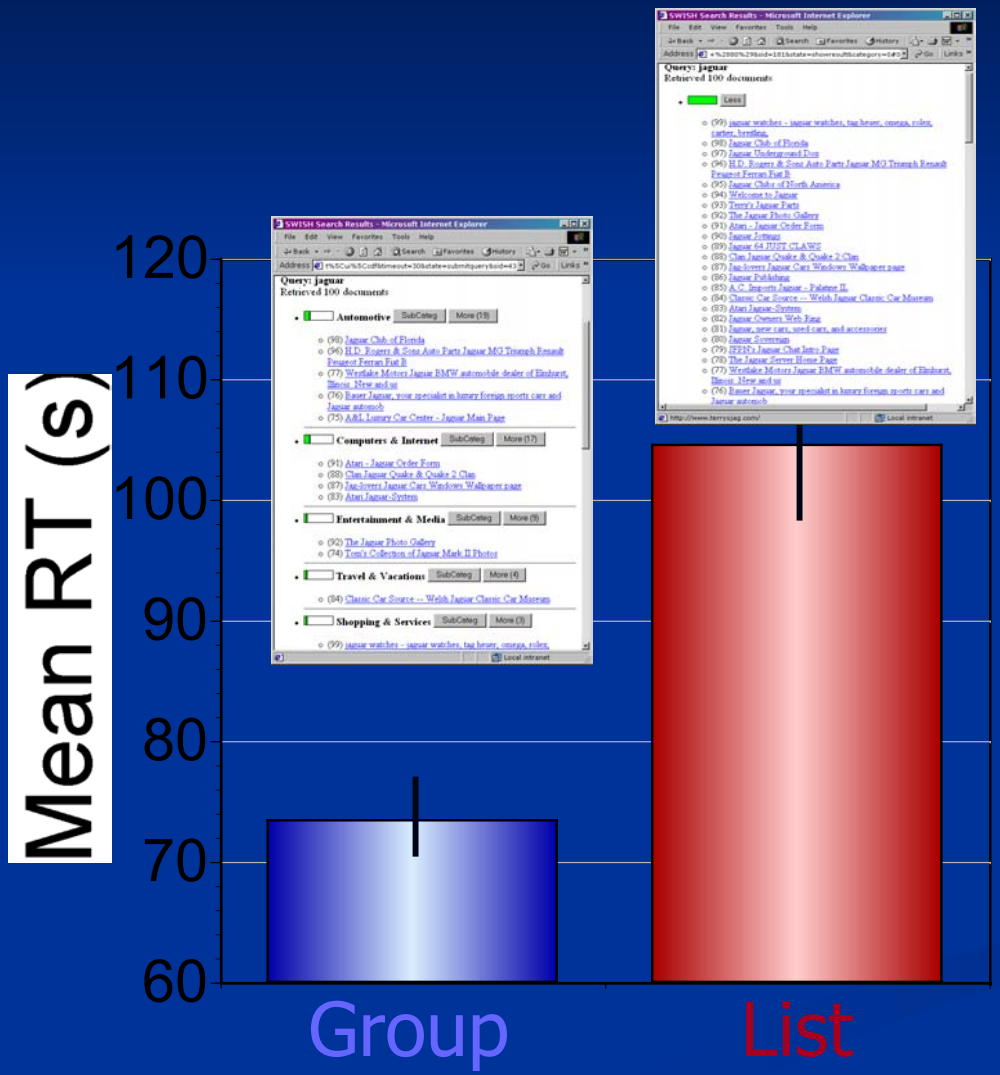
Group Interface



List Interface



Chen & Dumais (CHI'2000)



Group interface
- 42% faster
- Much preferred
  (6.4 vs. 4.2)

e.g., Usefulness of Click Through Data
Ranking and Ad Placement; Joachims (2002)

System 1 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. SVM-Light Support Vector Machines
   http://ais.gmd.de/~thorsten/svm_light
3. Support Vector Machines … References
   http://svm....com/SVMRefs.html
4. Lucent Technologies: SVM Demo App
   http://svm...com/SVT/SVMsvt.html
5. Royal Holoway Support Vector Machines
   http://svm.dcs.rhbnc.ac.uk

System 2 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. Support Vector Machines
   http://jbolivar.freeservers.com
3. An Intro to Support Vector Machines
   http://www.support-vector.net
4. Archives of Support Vector Machines
   http://www.jiscmail.ac.uk/lists/Support...
5. SVM-Light Support Vector Machines
   http://ais.gmd.dr/~thorsten/svm_light

e.g., Usefulness of Click Through Data

Ranking and Ad Placement; Joachims (2002)

System 1 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. SVM-Light Support Vector Machines
   http://ais.gmd.de/~thorsten/svm_light
3. Support Vector Machines … References
   http://svm....com/SVMRefs.html
4. Lucent Technologies: SVM Demo App
   http://svm...com/SVT/SVMSvt.html
5. Royal Holoway Support Vector Machines
   http://svm.dcs.rhbnc.ac.uk

System 2 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. Support Vector Machines
   http://jbolivar.freeservers.com
3. An Intro to Support Vector Machines
   http://www.support-vector.net
4. Archives of Support Vector Machines
   http://www.jiscmail.ac.uk/lists/Support...
5. SVM-Light Support Vector Machines
   http://ais.gmd.dr/~thorsten/svm_light

Combined Results:

1. Kernel Machines.
   http://svm.first.gmd.de
2. SVM-Light Support Vector Machines
   http://ais.gmd.de/~thorsten/svm_light
3. Support Vector Machines
   http://jbolivar.freeservers.com
4. Support Vector Machines … References
   http://svm...com/AVMRefs.html
5. An Intro to Support Vector Machines
   http://www.support-vector.net

e.g., Usefulness of Click Through Data

Ranking and Ad Placement; Joachims (2002)

System 1 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. SVM-Light Support Vector Machines
   http://ais.gmd.de/~thorsten/svm_light
3. Support Vector Machines … References
   http://svm....com/SVMRefs.html
4. Lucent Technologies: SVM Demo App
   http://svm...com/SVT/SVMsvt.html
5. Royal Holoway Support Vector Machines
   http://svm.dcs.rhbnc.ac.uk

System 2 Results:

1. Kernel Machines
   http://svm.first.gmd.de
2. Support Vector Machines
   http://jbolivar.freeservers.com
3. An Intro to Support Vector Machines
   http://www.support-vector.net
4. Archives of Support Vector Machines
   http://www.jiscmail.ac.uk/lists/Support...
5. SVM-Light Support Vector Machines
   http://ais.gmd.dr/~thorsten/svm_light

Combined Results:

1. Kernel Machines.
   http://svm.first.gmd.de
2. SVM-Light Support Vector Machines
   http://ais.gmd.de/~thorsten/svm_light
3. Support Vector Machines
   http://jbolivar.freeservers.com
4. Support Vector Machines … References
   http://svm...com/AVMRefs.html
5. An Intro to Support Vector Machines
   http://www.support-vector.net

Joachims (2002)

- Compared click through results with explicit relevance judgments (3 users, 180 queries for calibration)
  - Gave comparable results
  - Users clicked more relevant than non-relevant items
  - Users did not click on links from one engine more than the other, independent of relevance
- Evaluated a method that generates unbiased feedback about the relative quality of two search engines (or retrieval functions) … preliminary but promising

Web Collections … What Matters?

- What does the ideal web test collection look like? (size, document types, link distribution, page size, dark/light web, etc.)
- Is it possible to get relevant experimental results using small (<20GB) web snapshots?
- What are the right kind of queries for a web test collection?
- Should we be making relevance judgments differently for the web? (e.g., based on web sites or groups of pages; multi-valued relevance; etc)
- How important is the age of the collection?
- How can a test collection approach deal with the dynamic nature of the web?
- Is the test collection methodology meaningful on the web?
- Does a test collection need to have spam to be realistic?
- Experimenting on the live web forfeits direct comparability, since the web changes; is the sacrifice worth it?

Web Collections … What Matters?

- What does the ideal web test collection look like? (size, document types, link distribution, page size, dark/light web, etc.)
  - Depends on what you want to study
- Is it possible to get relevant experimental results using small (<20GB) web snapshots?
  - Some things; Graph properties vs. size
- What are the right kind of queries for a web test collection?
  - Representative
- Should we be making relevance judgments differently for the web? (e.g., based on web sites or groups of pages; multi-valued relevance; etc)
  - Little
- How important is the age of the collection?
  - Little
- How can a test collection approach deal with the dynamic nature of the web?
  - Yes
- Is the test collection methodology meaningful on the web?
  - Yes
- Does a test collection need to have spam to be realistic?
  - Yes if web (vs. site)
- Experimenting on the live web forfeits direct comparability, since the web changes; is the sacrifice worth it?
  - Loose explanatory power

Summary

- You can do systematic experiments on web data
  - Needed to evaluate progress of systems and components
  - Include TREC-style relevance judgments on static collections
  - But, move beyond that as well
- Example new directions
  - Focus on users and interaction (controlled task, or in situ)
  - Analysis of characteristics of queries and collections
  - Structured queries and content
- When you study end-to-end systems, it may be difficult to isolate source of improvements and to generalize, although you may produce a good system
- When you study components, you need to worry about how to combine them in difference contexts
  - E.g., (Good ranking + Poor spell checking) > (Ok ranking + Good spell check)