

Web Experiments and Test Collections

Susan Dumais

Microsoft Research

sdumais@microsoft.com

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 aspirin a day, increases longevity by 20%
- Study 2: Aspirin, does not increase longevity, and increases ulcers by 40%
- Should you take aspirin?
- Questions about quality, comparability and applicability:
 - Who are the subjects (age, sex, diet, health, etc.)?
 - How many subjects?
 - Dosage of aspirin?
 - 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-oriented vs. higher recall-oriented in graphical properties of collections
 - Queries/Tasks
 - E.g., *Types of queries*, but high variability
 - 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 ...

Collection	QuerySet	Pr@20	Pr@20	Ratio
		Full	Sample	
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

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-orientation in graphical properties of collections
 - Queries/Tasks
 - E.g., *Types of queries*,
 - 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 ...

Hawking and Robertson, 2001				
Collection	Query Set	Pr@20 Full	Pr@20 Sample	Ratio
Westerveld et al., TREC 2001				
		Base	+Anchor	+URL
	Topical Queries, Pr@5	0.36	0.36	0.76
	Home Page Queries, MRR	0.34	0.45	0.86

<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 properties

- Queries/Tasks

- E.g., *Types of queries*,

- Performance measures

- More confidence in the generalization techniques work on different collections
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

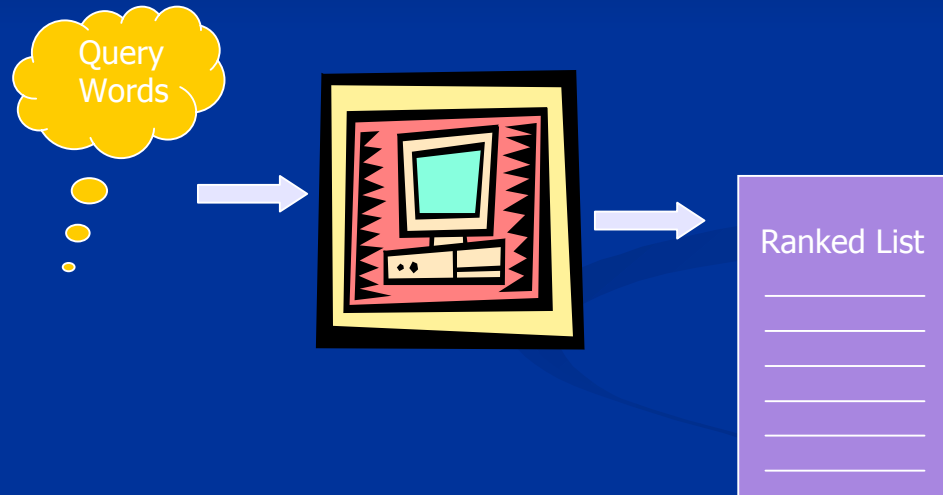
Evolution of TREC Web Collection

- *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

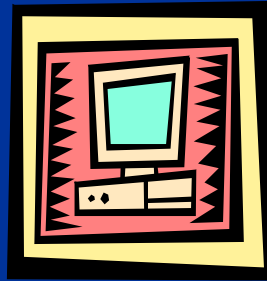
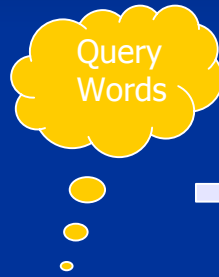
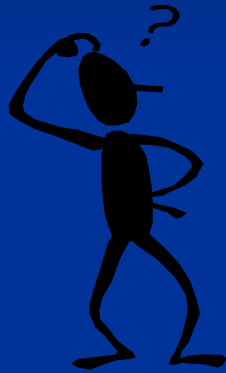
- Relevance and ranking are important, but not all ...
- What's missing?



Web Search Evaluation

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

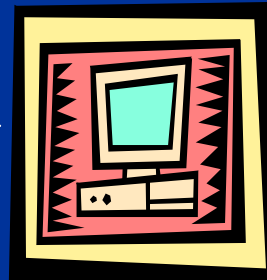
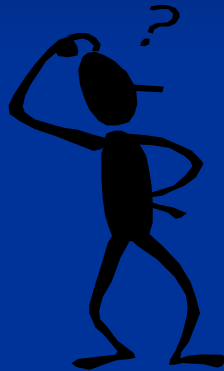
**Query
Analysis**



Web Search Evaluation

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

**Query
Analysis**



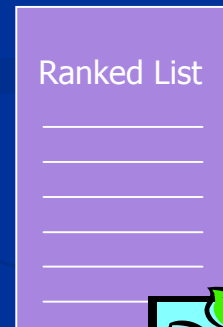
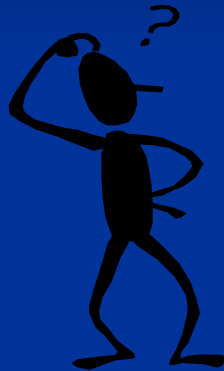
**Collection
Analysis**



Web Search Evaluation

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

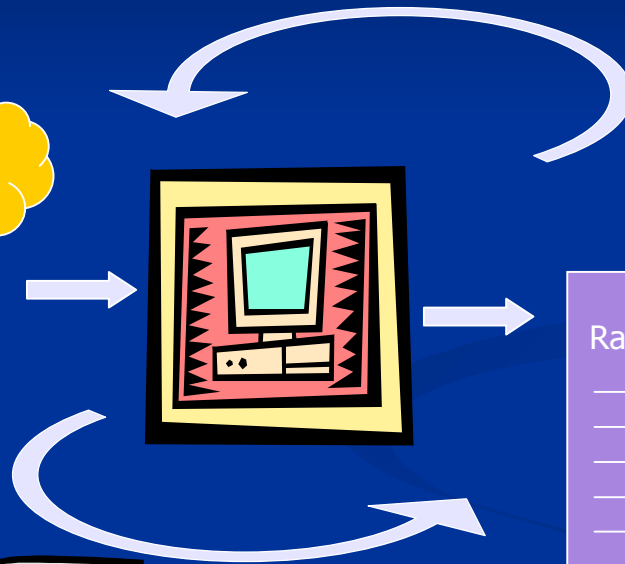
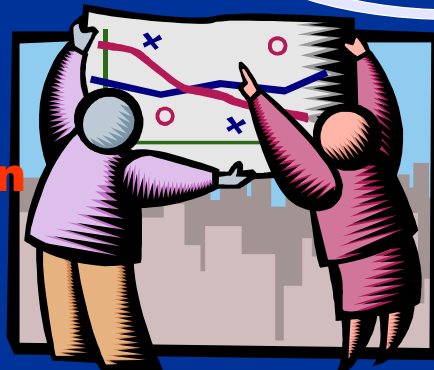
Query Analysis



Collection Analysis



User Interaction & Information Use



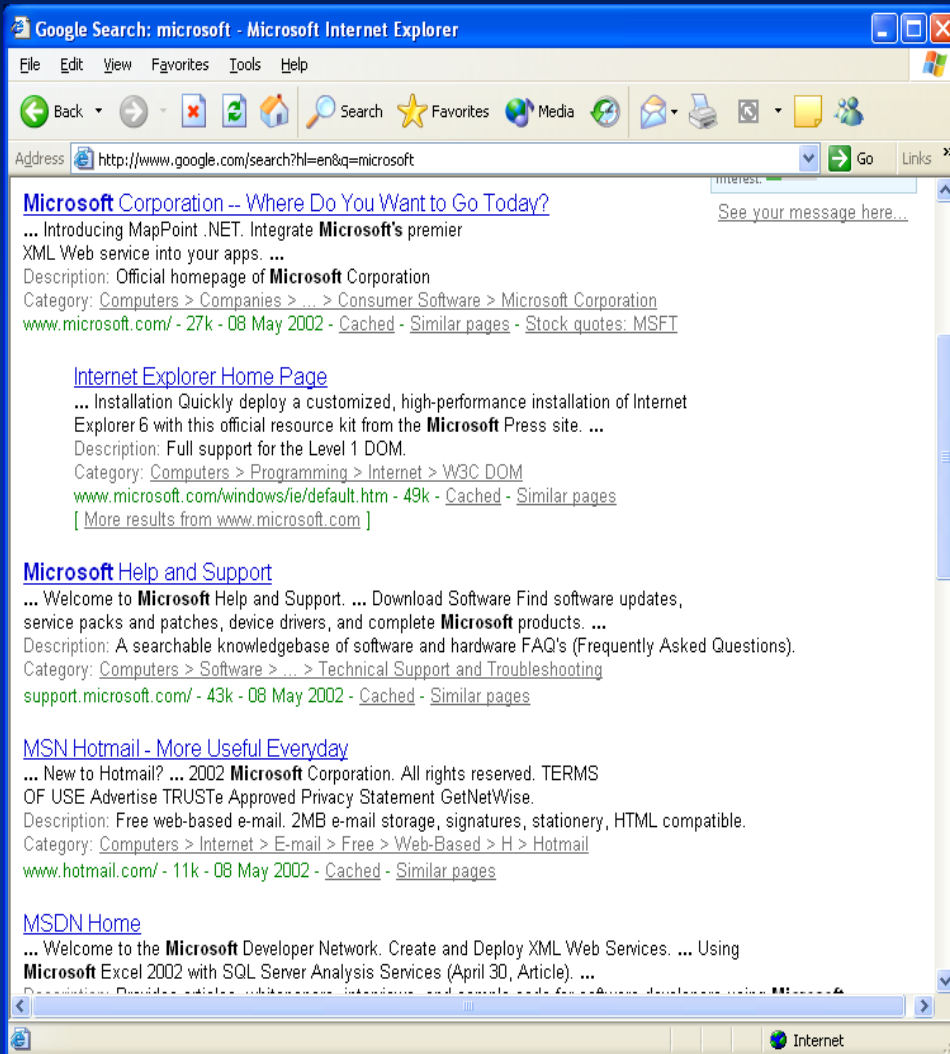
Web Search Evaluation

- 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

Google Search: microsoft - Microsoft Internet Explorer

File Edit View Favorites Tools Help

ZDNet: Reviews | News | Downloads | Tech Update | Prices

ZDNet Search

Address http://www.google.com

Microsoft Corporation ...
... Introducing MapPoint .NET
XML Web service into your ap
Description: Official homepage
Category: Computers > Comp
www.microsoft.com/ - 27k - 06

Internet Explorer Ho
... Installation Quickly
Explorer 6 with this offi
Description: Full suppo
Category: Computers >
www.microsoft.com/wir
[More results from ww

Microsoft Help and Sup
... Welcome to **Microsoft** Help
service packs and patches, de
Description: A searchable kno
Category: Computers > Softw
support.microsoft.com/ - 43k -

MSN Hotmail - More Use
... New to Hotmail? ... 2002 M
OF USE Advertise TRUSTe A
Description: Free web-based e
Category: Computers > Intern
www.hotmail.com/ - 11k - 08 1

MSDN Home
... Welcome to the **Microsoft**
Microsoft Excel 2002 with SG
Description: Provide statisti

advertisement

FREE software with purchase.

The Satellite 1000 \$157 from TOSHIBA \$1.079

shoptoshiba.com

intel inside celeron

While supplies last.

Search: Download service pack All ZDNet GO powered by Search.com

Tech Update

- Pocket PC 2002 at home in enterprise – March 1, 2002
- MS bugs and patches are the price we pay – January 4, 2002
- We use Outlook, and we don't get viruses – December 19, 2001
- System security made simple – September 26, 2001
- Service pack whacks bugs – August 21, 2001

AnchorDesk

- Meet the dark side of Windows XP – October 22, 2001
- A Radar Gun for Your Internet Connection – January 8, 2001
- Fidelity Wins Usability Race – July 17, 2000
- Story: Natural Born Killers: Where Are They Now? – July 29, 1999
- Natural Born Killers: Where Are They Now? – July 29, 1999

▶ More results

Downloads

- SPQuery v 4.2 – Find out which Service Packs and hot fixes are installed on your NT system.
- Service Pack Manager v 5.1 – Manage Microsoft Service Packs and hotfixes on Windows NT.
- Microsoft Exchange Server 2000 Service Pack 2 v 6.0.5762.3 – Correct some Microsoft Exchange server issues with these Service Packs.
- Microsoft SQL Server 2000 Service Pack 2 Database Components v – Get the latest fixes for Microsoft SQL Server 2000.
- Microsoft SQL Server 2000 Service Pack 2 Desktop Engine (MSDE) v – Get the latest

NEWSLETTERS

ZDNet News
ZDNet News brings you a summary of top headlines each business day.

Tech Update
Today
ZDNet Tech Update's editorial director David Berlind keeps your thumb on the pulse with his gut reaction to the most important news.

Your e-mail here

Sign me up!

▶ More newsletters...

e.g., Usefulness of Grouping

The screenshot shows a Microsoft Internet Explorer browser window with the address bar set to <http://www.google.com>. The page displays search results for the query "Jaguar" on the ZDNet Search engine. The results are grouped into several categories, each with a "SubCateg" button and a "More" button indicating the number of documents in that group.

Query: Jaguar
Retrieved 100 documents

- Computers & Internet** (18 documents)
 - (95) [Clan Jaguar Quake & Quake 2 Clan](#)
 - (90) [Atari Jaguar-System](#)
 - (79) [Atari - Jaguar Order Form](#)
 - (69) [Jaguar XK8 Screen Saver](#)
- Automotive** (16 documents)
 - (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
 - (94) [Jaguar Club of Florida](#)
 - (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
 - (84) [A&L Luxury Car Center - Jaguar Main Page](#)
- Entertainment & Media** (14 documents)
 - (83) [Tom's Collection of Jaguar Mark II Photos](#)
 - (82) [MacJag's Jaguar Page](#)
 - (80) [The Jaguar Photo Gallery](#)
- Travel & Vacations** (4 documents)
 - (92) [Classic Car Source -- Welsh Jaguar Classic Car Museum](#)
- Business & Finance** (2 documents)
 - (66) [Jaguar Consulting, Inc.](#)
- Shopping & Services** (2 documents)

The browser window also shows a sidebar with various links and a search box. The search box contains the text "Download service". The sidebar includes links for "Microsoft Corporation", "Microsoft Help and Support", "MSN Hotmail", and "MSDN Home".

Chen & Dumais (CHI'2000)

Group Interface

SWISH Search result - Microsoft Internet Explorer

Query: Jaguar
Retrieved 100 documents

- Computers & Internet
 - (95) [Clan Jaguar Quake & Quake 2 Clan](#)
 - (90) [Atari Jaguar-System](#)
 - (79) [Atari - Jaguar Order Form](#)
 - (69) [Jaguar XK8 Screen Saver](#)
- Automotive
 - (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
 - (94) [Jaguar Club of Florida](#)
 - (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
 - (84) [A&L Luxury Car Center - Jaguar Main Page](#)
- Entertainment & Media
 - (83) [Tom's Collection of Jaguar Mark II Photos](#)
 - (82) [MacJag's Jaguar Page](#)
 - (80) [The Jaguar Photo Gallery](#)
- Travel & Vacations
 - (92) [Classic Car Source -- Welsh Jaguar Classic Car Museum](#)
- Business & Finance
 - (66) [Jaguar Consulting, Inc.](#)
- Shopping & Services

List Interface

SWISH Search result - Microsoft Internet Explorer

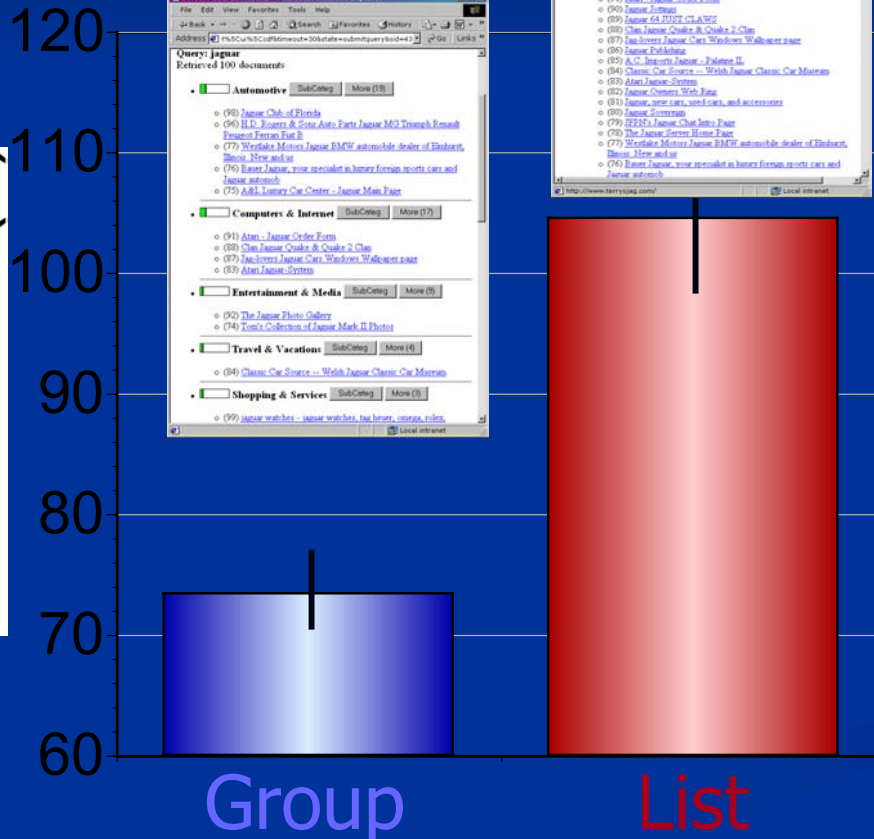
Query: Jaguar
Retrieved 100 documents

- Search Results
 - (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
 - (98) [Jaguar Clubs of North America](#)
 - (97) [Welcome to Jaguar](#)
 - (96) [Terry's Jaguar Parts](#)
 - (95) [Clan Jaguar Quake & Quake 2 Clan](#)
 - (94) [Jaguar Club of Florida](#)
 - (93) [Jaguar Daimler Heritage Trust](#)
 - (92) [Classic Car Source -- Welsh Jaguar Classic Car Museum](#)
 - (91) [A.C. Imports Jaguar - Palatine IL](#)
 - (90) [Atari Jaguar-System](#)
 - (89) [Jaguar Underground Dox](#)
 - (88) [Jaguar Owners Web Ring](#)
 - (87) [Jaguar, new cars, used cars, and accessories](#)
 - (86) [Jaguar Sovereign](#)
 - (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
 - (84) [A&L Luxury Car Center - Jaguar Main Page](#)
 - (83) [Tom's Collection of Jaguar Mark II Photos](#)
 - (82) [MacJag's Jaguar Page](#)
 - (81) [Welcome to Jaguar Homes!](#)
 - (80) [The Jaguar Photo Gallery](#)

Chen & Dumais (CHI'2000)



Mean RT (s)



- 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 Holloway 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://svm.first.gmd.de/~thorsten/svm_light
3. Support Vector Machines ... References
http://ais.gmd.de/~thorsten/svm_refs.html
4. Lucent Technologies: SVM Demo App
<http://svm...com/SVT/SVM.svt.html>
5. Royal Holoway Support Vector Machines
<http://svm.dcs.rhnc.ac.uk>

System 1 Results:

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

System 2 Results:

1. Kernel Machines
<http://svm.first.gmd.de>
2. Support Vector Machines
<http://jbolivar.freesevers.com>
3. Support Vector Machines
<http://www.support-vector.net>
4. Archives of Support Vector Machines
<http://www.jiscmail.ac.uk/lists/Support...>
5. An Intro to Support Vector Machines
<http://www.support-vector.net>

System 2 Results:

1. Kernel Machines
<http://svm.first.gmd.de>
2. Support Vector Machines
<http://jbolivar.freesevers.com>
3. 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.de/~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.freesevers.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://svm.first.gmd.de/~thorsten/svm_light
3. Support Vector Machines ... References
http://ais.gmd.de/~thorsten/svm_refs.html
4. Lucent Technologies: SVM Demo App
<http://svm...com/SVT/SVMApp.html>
5. Royal Holoway Support Vector Machines
<http://svm.dcs.rhnc.ac.uk>

System 2 Results:

1. Kernel Machines
<http://svm.first.gmd.de>
2. Support Vector Machines
<http://jbolivar.freesevers.com>
3. Support Vector Machines
<http://www.support-vector.net>
4. Archives of Support Vector Machines
<http://www.jiscmail.ac.uk/lists/Support...>
5. An Intro to Support Vector Machines
<http://www.support-vector.net>

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.freesevers.com>
4. Support Vector Machines ... References
<http://svm...com/SVMRefs.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 different contexts
 - E.g., (Good ranking + Poor spell checking) > (Ok ranking + Good spell check)