

Moving from Description to Prediction for Information Searching

Information searching: *actions (behavioral, affective, and cognitive) employed by people when interacting with an information system*

Jim Jansen

College of Information Sciences and Technology
The Pennsylvania State University

jjansen@ist.psu.edu

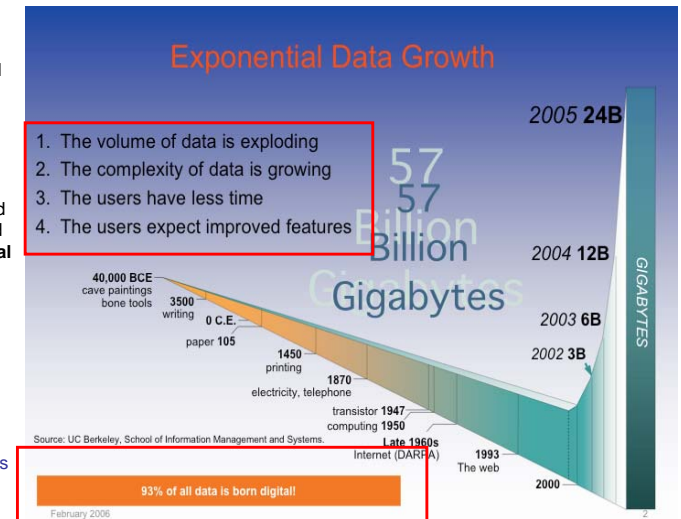
Who is Jim Jansen?

- Associate professor at College of Information Sciences and Technology, Pennsylvania State University, USA
- Active research and teaching efforts - http://ist.psu.edu/faculty_pages/jjansen/
- Funded projects: NSF, AFOSR, OSD, DITRA, USMC, ARL, and Google
- Several non-funded projects on-going

What is the information context that we are facing?

Explosion of Information - the **Zettabytes** are coming

- Moving too **'everything'** recorded and indexed
- A lot **global** but much will remain **local**
- Many bytes will **never be seen by humans**.
- **Search** (along with data summarization, trend detection, information and knowledge extraction and discovery) is **foundational technology**
- Raises issues, including:
 - **Infrastructure requirements**. How and who pays?
 - **Changes the nature of privacy and anonymity**
 - **Pressure on traditional communication channels** (Am I going to miss my local newspaper?)

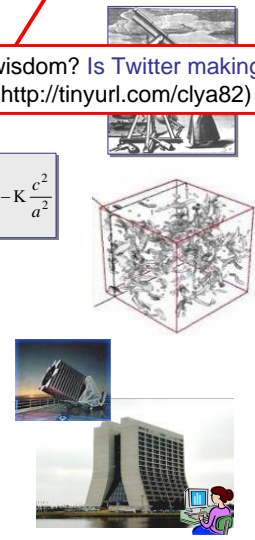


Data → Information → Knowledge

- Thousand years ago: science was **empirical** describing natural phenomena
- Last few hundred years: **theoretical** branch using models, generalizations
- Last few decades: a **computational** branch simulating complex phenomena
- Today: **data exploration** (eScience) unifying theory, experiment, and simulation
 - Data captured by sensors, instruments, or generated by simulator
 - Processed by humans and software
 - Information/ knowledge stored in computer
 - Analyzes database / collection content using data management and statistics
 - Network and Web Science

What about wisdom? Is Twitter making us smarter? (<http://tinyurl.com/clya82>)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Jim Jansen's Research

Will primarily address the algorithmic work, but end with a summary slide of affective, cognitive, and business research projects.

- Conduct **algorithmic** research, but also **affective** (emotion, mood), **cognitive** (decision making, learning), and **business** (customer relationships, keyword advertising) aspects

[Twitter](#) [search logs](#)

The State of Web Search

The Power of Search and the Web

Category	Monthly Visitors (Millions)	Avg Usage Days Per Visitor
Total U.S. Internet	167.0	18.7
Portals (Incl. Search & E-Mail)	155.2	20.8
Search/Navigation	140.9	12.2
E-Mail	123.3	12.0
Entertainment	127.7	11.6
News/Information	107.4	11.1
Instant Messengers	70.2	9.9
Retail	132.2	9.9
Directories & Resources	119.5	9.8
Community	107.0	7.5
Business/Finance	106.6	8.3
Technology	92.7	3.7

- Search is **the** top online activity
- Search drives over **5 billion monthly** queries in the U.S.
- Online activity has a **huge impact** on people's daily lives:
 - 70 minutes less with family
 - 30 minutes less TV
 - 8.5 minutes less sleep

Sources: comScore, U.S., Feb. '06, Stanford Institute for the Quantitative Study of Society, Nov. '05

Analysis of Search Marketplace

comScore Core Search Report* July 2008 vs. June 2008
Total U.S. – Home/Work/University Locations
Source: comScore qSearch 2.0

Core Search Entity	Share of Searches (%)		
	Dec-08	Jan-08	Point Change Jan-09 vs. Dec-08
Total Core Search	100.0%	100.0%	NA
Google Sites	63.5%	63.0%	-0.5
Yahoo! Sites	20.5%	21.0%	0.5
Microsoft Sites	8.3%	8.5%	0.2
Ask Network	3.8%	3.9%	0.1
AOL LLC	3.9%	3.7%	-0.2

Holding fairly stable over the last year or so

* Based on the five major search engines including partner searches and cross-channel searches. Searches for mapping, local directory, and user-generated video sites that are not on the core domain of the five search engines are not included in the core search numbers.

Analysis of Online Traffic

Top Global Web Properties
Ranked by Total Unique Visitors (000)* May 2008
Total Worldwide, Age 15+ - Home and Work Locations Source: comScore World Metrix

Property	Total Unique Visitors (000)	% Reach
Google Sites	643,809	75.5
Microsoft Sites	572,016	67.1
Yahoo! Sites	514,831	60.3
Wikipedia Sites	263,120	30.8
AOL LLC	252,394	29.6
eBay	247,791	29.0
Fox Interactive Media	169,301	19.8
Amazon Sites	159,281	18.7
Apple Inc.	140,380	16.5
CNET Networks	133,480	15.6
Ask Network	127,769	15.0
FACEBOOK.COM	123,851	14.5
Adobe Sites	107,361	12.6
Time Warner - Excluding AOL	98,000	11.5
WordPress	96,394	11.3
Viacom Digital	86,546	10.1
Baidu.com Inc.	80,201	9.4
TENCENT Inc.	77,885	9.1
Glam Media	77,391	9.1
New York Times Digital	77,172	9.0

Long tail for online traffic (i.e., a few sites with a lot of traffic and a whole bunch will little traffic)

* Excludes traffic from public computers such as Internet cafes and access from mobile phones or PDAs

Analysis of Keyword Advertising

Search Engine Marketing (SEM) Spending in North America, 2007-2011 (billions)



Note: includes paid placement, organic search engine optimization (SEO), SEM technology and paid inclusion
 Source: Search Engine Marketing Professional Organization (SEMPO), "The State of Search Engine Marketing 2007" conducted by Radar Research via IntelliSurvey, Inc., February 2008

093885

www.eMarketer.com

- **Keyword advertising**, the fastest growing advertising medium.
- **Revenue base** for major search engines such as Google and Yahoo!, as well as many content-based Web sites.
- In 2008, Google earned ~\$20 billion; more than 90% of this revenue came from keyword advertising (Google 2009).

Some of the most detailed user behavioral research current going on – almost all outside of academic and research firms!

State of Information Searching Research

- Primarily descriptive (i.e., *let me tell you what people do*)
- Examples (*search trends, popular search terms, technology uses, number of results, clicked, etc.*)
- What is lacking? Predictive research -> approaches and models that not only **describe** but can **predict** what people will do

Important for a lot of reasons – from technology development, system resource allocation, trends, extreme events, financial, and understanding users

Information Searching

- Probabilistic user modeling

- increasingly important area
- allows computer systems to adapt to users

Note: not always 'informational' anymore. Many time people are searching for 'other things'. Jansen, Booth, Spink (2008).

- Algorithmic techniques typically employ state models

- Simple Bayesian Classifier, Markov Modeling, n-grams)

- Issues – state chains break down after a couple of transitions

- Consistently supported in a variety of domains from Meister and Sullivan (1967), Penniman (1975) to Jansen (2008)

Illustration of Probabilistic User Modeling Using n-grams

User	Search State Transitions	Predictive Pattern	Next State?	Accuracy
1	ABCF	AB	C	100%
2	ABCDE	BC	D	66%
3	ABCDE	CD	E	100%
4	A	A	B	60%
5	AC	C	D	40%

Given these states ...

... how accurately can we predict these?

Example Using Search Log



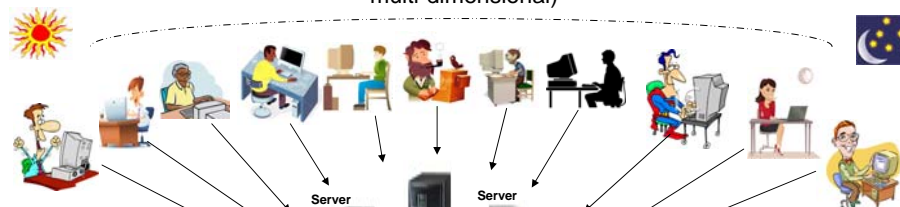
Drop out rate (folks who don't submit a query ~40%)

- ~ 965,000 searching sessions
- ~ 1,500,000 queries
- 8 states focusing on query reformulation

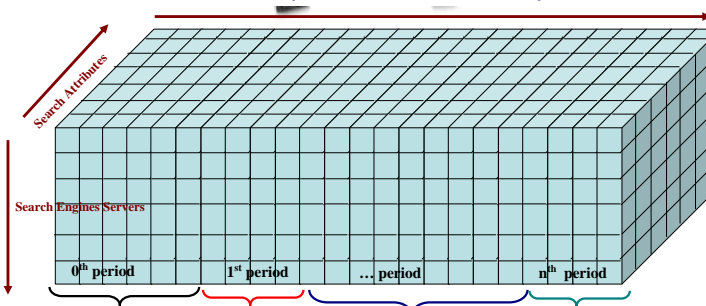
Not much better than just guessing!

- Similar results for other aspects of searching
- See - Qui (1993), Jansen (2005), Jansen & McNeese (2006)
- Maybe 'states' are not the correct paradigm?

Search engine logs as an information stream (voluminous, temporal, and multi-dimensional)

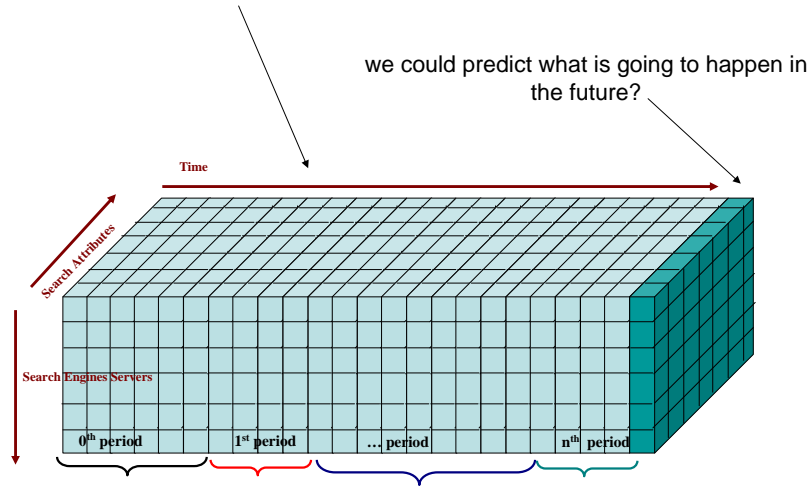


Information searching is a temporal stream (i.e., stateless)



Search Engine Logs – viewed as a temporal stream (i.e., stateless, with volume, mass, momentum, and acceleration)

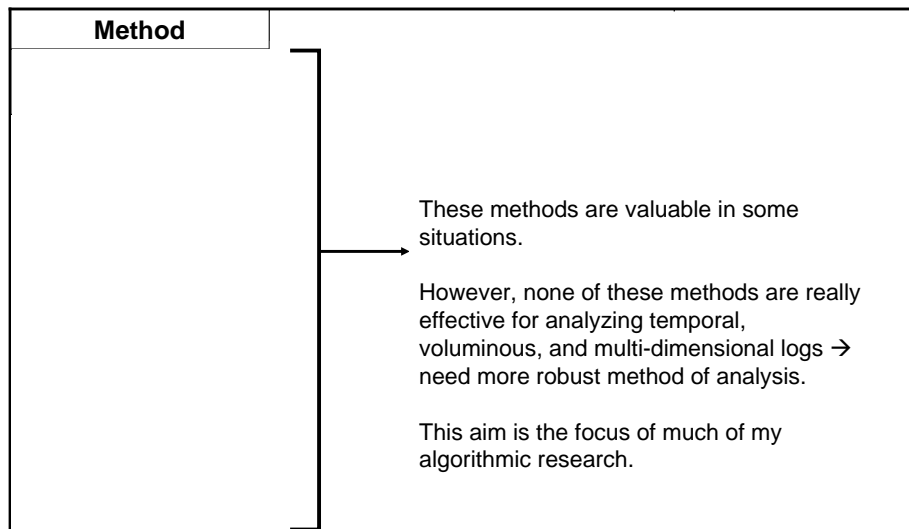
What if, based on what has happened in the past in the temporal stream, ...



Ongoing Research and Challenges

Method	Implications	Publication
N-grams	- 1 st or 2 nd order models work best	Jansen & Zhang, M. (2008)
Decision Tree	- 74% accuracy for user intent - real time	Jansen, Booth, & Spink (2008)
K-means Clustering	- 90% accuracy for user intent	Kathuria & Jansen (Working)
Time Series Analysis	- inference between query length and ranked of clicked result	Zhang, Y. & Jansen (2009)
Neural Networks	- session duration, query length, query reformulation correlate positively with future clickthrough	Zhang, Y. & Jansen (2009)
Tensor Analysis	- main and secondary trends - query reformulation, session length, and query length negatively correlated with user intent	Gopalakrish & Jansen (Under Review)

Ongoing Research and Challenges



Lets take a look at some other research work

- **User Modeling**: developing a time series analysis approach to develop an **equation** to model individual users' searching behaviors using log data (Funded by AFOSR)
- **Affective Factors**: investigating the effect of system **branding on user perceptions** of system performance using structural equation modeling and survey data (Funded by Google)
- **User Modeling**: **converting lessons learned into actionable knowledge assets** using cognitive ontology (Funded by OSD STTR Phase 1)

Lets take a look at some other work

- **Modeling Information Searching:** developing model for **predicting the underlying searching task using Bloom's Taxonomy** (Funded by AFOSR)
- **Search Engine Marketing:** **analyzing a three-year keyword advertising campaign** from an information searching perspective (In collaboration with Rimm-Kaufmann)
- **Micro-blogging for Reputation Management:** analyzing thousands of posts to Twitter using **sentiment analysis** (In collaboration with Twitter)

Research and Online Presence

- Most research papers on Website: http://ist.psu.edu/faculty_pages/jjansen/
- Blog: <http://jimjansen.blogspot.com/>
- Twitter: jimjansen
- LinkedIn: <http://www.linkedin.com/in/jjansen>

Thank you!

(open for questions and further discussion)

Jim Jansen

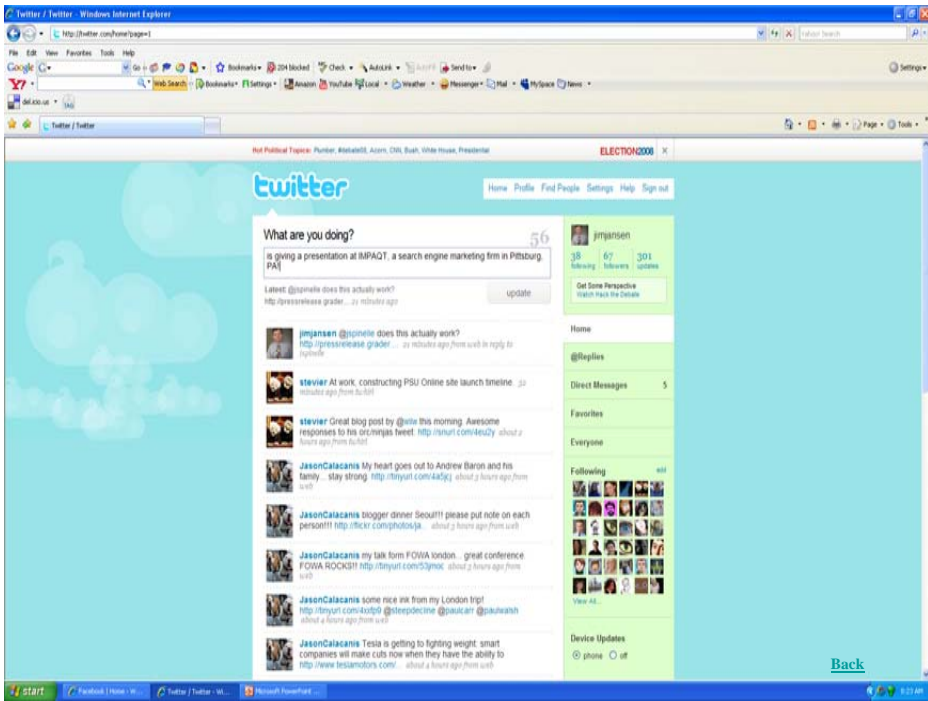
College of Information Sciences and Technology
The Pennsylvania State University

jjansen@ist.psu.edu

id	SessionID	Time	QueryID	clean_query	ResultCount	length	level_one	sp
000000e99c234b65	2006-05-10 10:56:36	6e51307996b449e	county wide home loans		16	4	new	
0000012390ee4000	2006-05-08 12:31:18	9a00444d6e5144ea	Millard Cramp		3	2	new	
000001fbc0c4d5e5	2006-05-03 07:17:40	13314701765c881f	mims com		16	1	navigational	new
0000034435e140c8	2006-05-08 21:08:54	0b55304d4be947cf	our garden gang		11	3	new	
0000039779b44df	2006-05-08 08:30:04	7396a25d5eac491f	st louis today		17	3	new	
0000042484c459a	2006-05-03 10:45:47	0f63bc0d04c4c0	Lewis Upshur		24	2	new	
0000044a63c4ac6	2006-05-12 14:09:50	6a94601aec94327	American Residential Realty		15	4	new	
0000044a63c4ac6	2006-05-12 14:10:14	6c1c1d602c5d8ea	American Residential Realty Associates		0	5	specialization	new
0000044a63c4ac6	2006-05-12 14:10:26	70b10c0d6c544585	American Residential Realty Assoc		0	5	reformulation	new

Search Logs has some common fields, such as time, queries, results, etc. We can enrich the log with additional fields.

Back



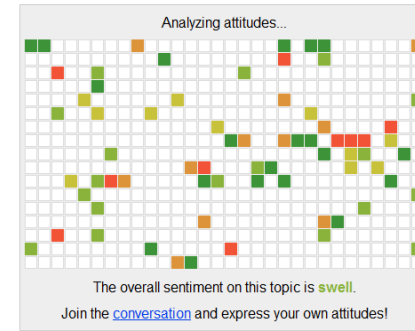
SUMMIZE LABS

Realtime Twitter Sentiment

Enter a topic in the box below. We use our [search engine](#) to find up-to-the-second tweets about this topic, then automatically analyze the attitudes expressed in those tweets.

since:2008-05-02 until:2008-05-08

Try: obama, BSG, iphone.



The colors indicate the sentiment of words and phrases found in tweets:
 great words so-so words wretched words
 swell words bad words no sentiment

[Home](#) [About Us](#) [Blog](#)

© 2008 Summize, Inc.

[Back](#)

