College of Information Sciences
and Technology

Defining Searching Sessions on Web Session Engines

Jim Jansen, College of Information Sciences and Technology, The Pennsylvania State University, jjansen@ist.psu.edu

Amanda Spink, Faculty of Information Technology, Queensland University of Technology, ah.spink@qut.edu.au

Vinish Kathuria, Infospace, Inc. – Search & Directory, Vinish.Kathuria@infospace.com

Sherry Koshman, School of Information Sciences, University of Pittsburgh, skoshman@sis.pitt.edu

PENNSTATE

1 8 5 5

Outline

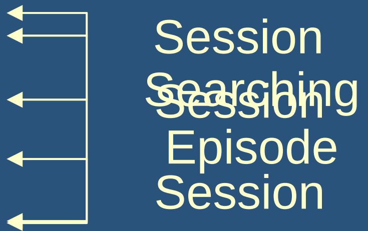
1. Introduction to the problem
2. Why is this important?
3. Research Question (defining a session)
4. Research Design (search log analysis)
5. Results (for three methods of session identification)
6. Implications of Results

Introduction to the problem

1. Searching Episode – series of interactions between a system and a searcher within a specific time period.
2. A single searching episode may be composed of more than one searching session.
3. Searching Session - series of interactions between a system and a searcher on a given information topic within a specific time period.

Example

User Id	Cookie	Time	Query
12.109.90.70	2NE8RS2A	1:34:38 PM	marathon gas station
12.109.90.70	2NE8RS2A	1:57:41 PM	department of agriculture indiana
12.109.90.70	2NE8RS2A	4:05:20 PM	ryan's restaurant group inc
12.109.90.70	2NE8RS2A	4:06:04 PM	ryan's restaurant group inc fire mountain



Issue: How does a system detect session boundaries in real time?

Why is this important?

1. Important for designing helpful searching systems, recommender systems, personalization, and targeting content to particular users.
2. These systems have a natural focus on the entire searching experience rather than algorithmic optimization at the query level.
3. In fact, **session satisfaction** (versus query) may be the defining measure for evaluating an information system with real users.

Research Question

What are the differences in results when using alternative methods for identification of Web search engines sessions?

a. IP address and cookie
b. IP address, cookie, and a temporal cut-off
c. IP address, cookie, and context changes.

Research Design

1. 4,056,374 records from Dogpile.com gathered on 6 May 2005 from 534,507 “users”.
2. Cleaned, prepared and analyzed data use methods from prior work.
3. Located the initial query by user and recreated the chronological sequence of actions by that user.

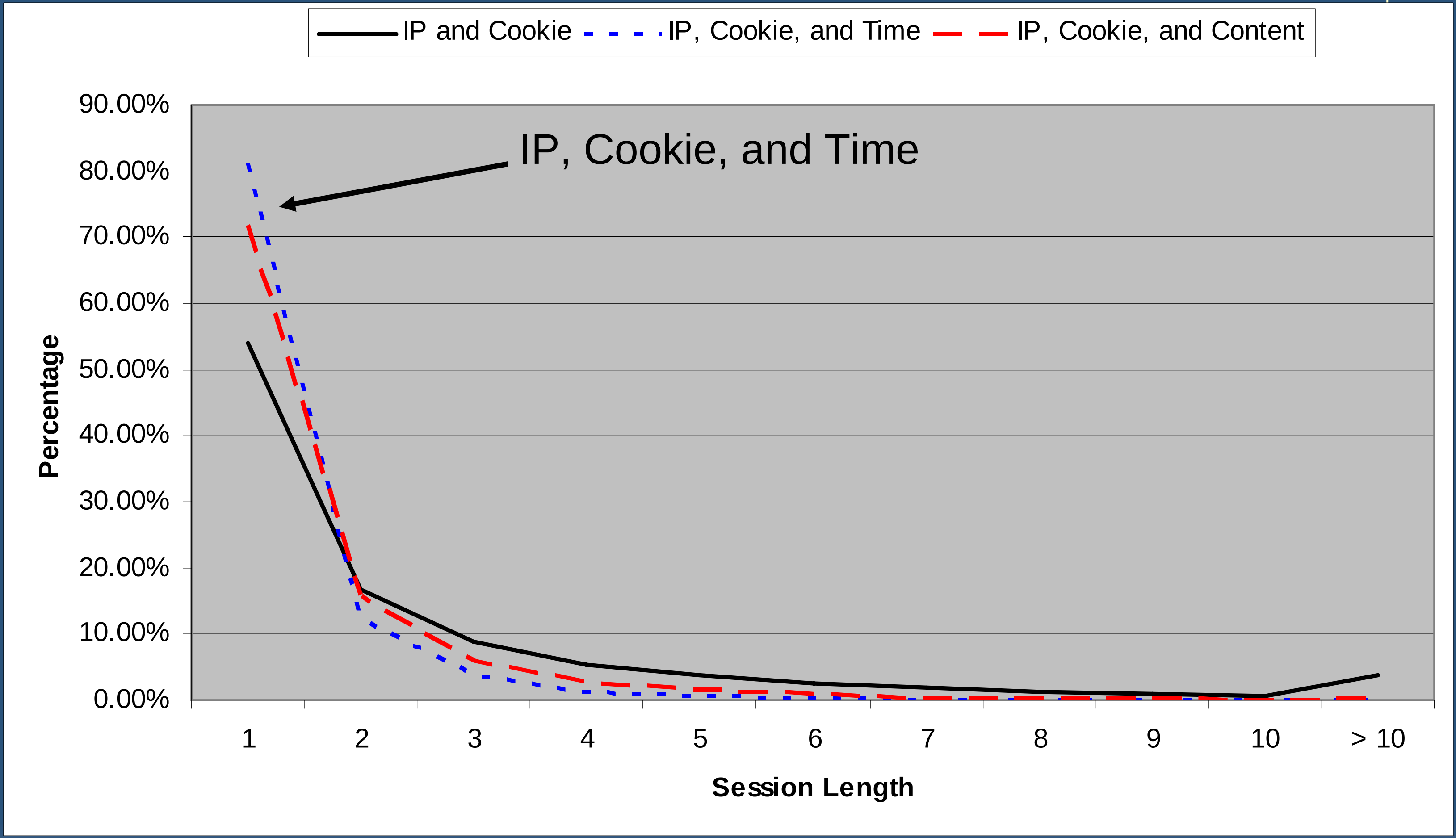
Results (Session Length)

Comparing session lengths (i.e., number of queries in a session).

	Method 1: IP and Cookie		Method 2: IP, Cookie, and 30 min. Time Limit		Method 3: IP, Cookie, and Query Content	
Session Length	Occurrences	Percentage	Occurrences	Percentage	Occurrences	Percentage
1	288,231	53.92%	533,950	81.15%	691,672	71.64%
2	88,875	16.63%	81,224	12.34%	153,056	15.85%
3	47,664	8.92%	24,840	3.78%	58,537	6.06%
4	29,345	5.49%	9,219	1.40%	27,134	2.81%
5	19,655	3.68%	3,822	0.58%	14,168	1.47%
6	13,325	2.49%	1,755	0.27%	7,745	0.80%
7	9,549	1.79%	944	0.14%	4,430	0.46%
8	7,169	1.34%	622	0.09%	2,791	0.29%
9	5,497	1.03%	442	0.07%	1,769	0.18%
10	4,130	0.77%	331	0.05%	1,193	0.12%
> 10	21,067	3.94%	871	0.13%	2,944	0.30%
	534,507	100.00%	658,020	100.00%	965,439	100.00%

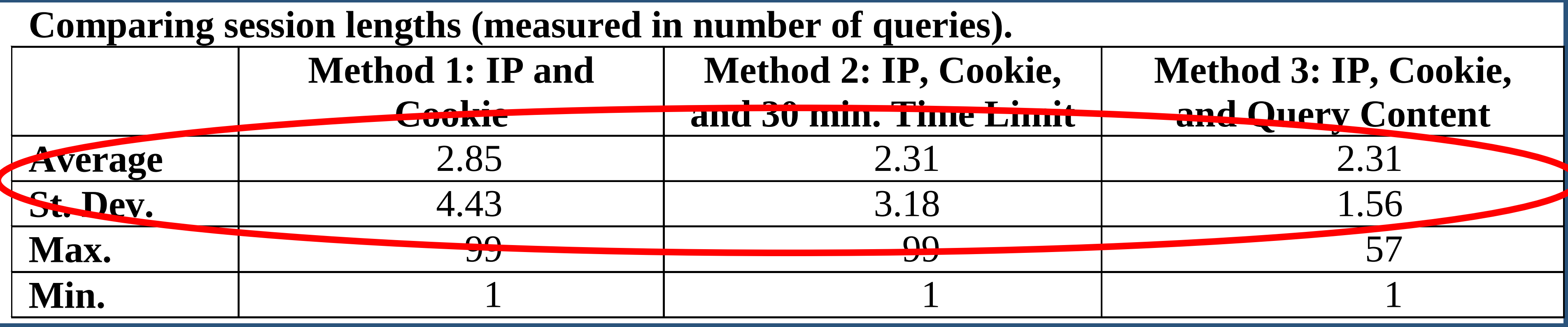
College of Information Sciences
and Technology

Results (Session Length)



PENNSTATE

Results (Session Length)



Comparing session lengths (measured in number of queries).

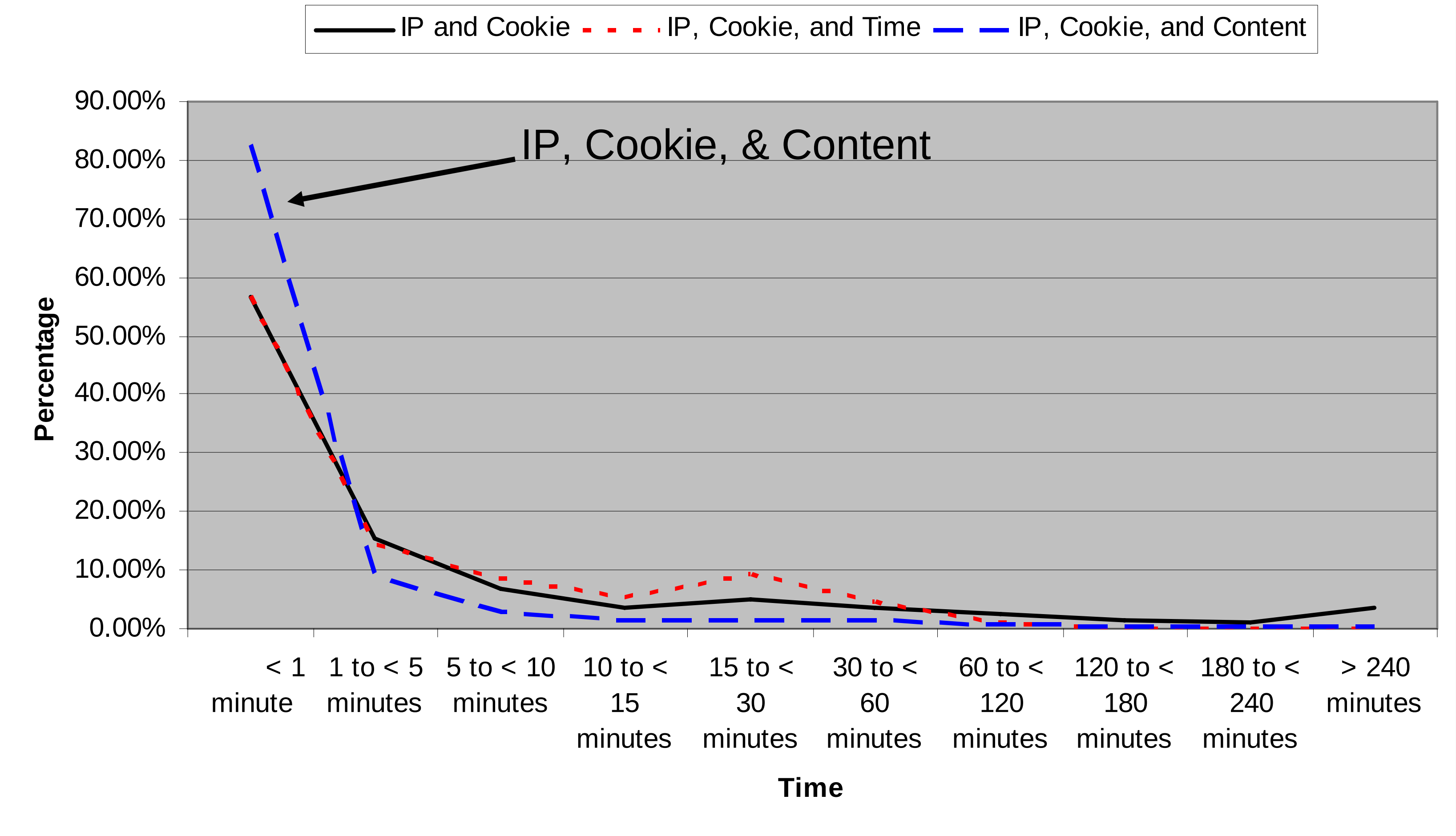
	Method 1: IP and Cookie	Method 2: IP, Cookie, and 30 min. Time Limit	Method 3: IP, Cookie, and Query Content
Average	2.85	2.31	2.31
St. Dev.	4.43	3.18	1.56
Max.	99	99	57
Min.	1	1	1

Results (Session Duration)

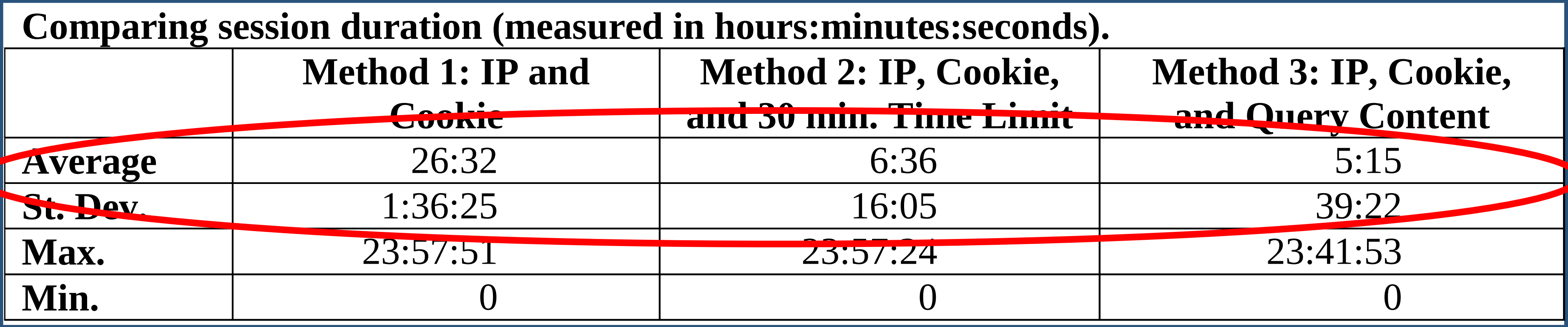
Comparing session durations (i.e., temporal length of a session).

	Method 1: IP and Cookie		**Method 2: IP, Cookie, and 30 min. Time Limit**		**Method 3: IP, Cookie, and Query Content**	
Session Duration	**Occurrences**	**Percentage**	**Occurrences**	**Percentage**	**Occurrences**	**Percentage**
< 1 minute	302,653	56.62%	372,983	56.68%	794,765	82.32%
1 to < 5 minutes	83,236	15.57%	93,251	14.17%	86,358	8.94%
5 to < 10 minutes	36,347	6.80%	55,956	8.50%	28,044	2.90%
10 to < 15 minutes	19,806	3.71%	36,020	5.47%	12,277	1.27%
15 to < 30 minutes	27,210	5.09%	61,767	9.39%	13,752	1.42%
30 to < 60 minutes	18,441	3.45%	30,790	4.68%	12,628	1.31%
60 to < 120 minutes	14,236	2.66%	6,615	1.01%	7,524	0.78%
120 to < 180 minutes	8,262	1.55%	506	0.08%	3,320	0.34%
180 to < 240 minutes	5,901	1.10%	76	0.01%	1,919	0.20%
> 240 minutes	18,415	3.45%	56	0.01%	4,852	0.50%
	534,507	100.00%	658,020	100.00%	965,439	100.00%

Results (Session Duration)



Results (Session Duration)



Comparing session duration (measured in hours:minutes:seconds).

	Method 1: IP and Cookie	Method 2: IP, Cookie, and 30 min. Time Limit	Method 3: IP, Cookie, and Query Content
Average	26:32	6:36	5:15
St. Dev.	1:36:25	16:05	39:22
Max.	23:57:51	23:57:24	23:41:53
Min.	0	0	0

Implications

- Critical for developing more supportive searching systems, especially in the more complex searching environments of exploratory searching and multitasking.
- Using content approach, Web search systems can develop systems that provide session level searching assistance to Web engine users.
- Content method presented here is advantageous for real-time system implementation.

Questions and Discussion

Jim Jansen

College of Information Sciences and Technology
The Pennsylvania State University

jjansen@ist.psu.edu

Research Funded in part by AFRL/AFOSR, BAA No. AFOSR 2005-4