User Personal Evaluation of Search Engines
– Google, Bing and Blekko

Bing Liu
Technical Report, May 8, 2011
Department of Computer Science
University of Illinois at Chicago
liub@cs.uic.edu

Abstract

In many ways, search engines have become the most important tool for our information seeking. Due to competitions, search engine companies work very hard to improve the quality of their engines. Evaluating the search quality is, however, still a difficult problem. Although many evaluations have been conducted to assess the quality of different search engines, they mainly used fixed queries, and assessed the degree of relevance of each returned page by human judges [1, 3, 4, 5, 6, 9, 10, 11, 12]. However, the queries were not originated from the human judges, but were sampled from the queries issued by search engine users. This evaluation method is by no means ideal because relevance does not mean user satisfaction. User satisfaction can only be assessed by the user based on his/her queries and the returned results from search engines. An ideal evaluation is a personal evaluation of the actual users of search engines. In this article, I describe such an evaluation of three search engines, Google, Bing and Blekko, which was performed by 35 undergraduate students from the Department of Computer Science, University of Illinois at Chicago, in Spring 2011. To make it more interesting, students were also asked to look for spam links and content farms in the search results. As expected, the evaluation shows that in terms of user satisfaction, Google is still the best, Bing is close behind, and Blekko, which is a new engine, still needs some work. In terms of filtering out spam or content farms, all three engines are excellent. However, I also found something rather surprising and unexpected, which warrants further investigation.

1. Background of the Evaluation

The evaluation was conducted as a research project in my CS376 class (*Practicum in Computer Science Presentations*). CS376 is a presentation course which teaches undergraduate students how to make technical presentations. The course comes with three assignments. For each assignment, every student needs to make an oral presentation in class. This search evaluation project is for the third assignment. For this assignment, each student needs to evaluate the three search engines and then deliver a research presentation in class based on the evaluation results. Due to the fact that presentations are time consuming, the class has two sections. S*ection A* (or *sA*) has 20 students and meets at 3:00pm on Monday. *Section B* (or *sB*) has 15 students and meets at 5:00pm on Monday. The evaluation started on February 21, 2011 and ended on March 20, 2011. It lasted for exactly 4 weeks. Since it was a course project, the evaluation was not sponsored or funded by anyone, and had no connection with any search engine company.

2. Evaluation Setup

I tried to design the evaluation process so that it can minimize biases that may be introduced into the process. The details of the evaluation setup are given below. The whole evaluation was conducted in two phases:

Phase 1 (individual engine evaluation): This phase lasted for 3 weeks. In each week, a student evaluates a single engine. To avoid any bias from the sequencing of the engines, students in each section were divided into three equal-sized groups. Each group evaluated the engines in a different sequence over the 3 weeks (see Table 1 and Table 2 for the sequences).

Table 1. Section A evaluation sequences

	Week 1	**Week 2**	**Week 3**
Group 1	Google	Blekko	Bing
Group 2	Blekko	Bing	Google
Group 3	Bing	Google	Blekko

Table 2: Section B evaluation sequences

	Week 1	**Week 2**	**Week 3**
Group 1	Google	Bing	Blekko
Group 2	Bing	Blekko	Google
Group 3	Blekko	Google	Bing

Phase 2 (comparative evaluation): This phase lasted for 1 week. In this week, all three engines are evaluated at the same time, i.e., for each query, the student searches all three engines and records his/her level of satisfaction with the result of each engine. Again to reduce possible biases, the three groups of students in each section were asked to search the three engines in the sequences in Tables 3 and 4.

Table 3: Section A search sequences

	1	**2**	**3**
Group 1	Google	Blekko	Bing
Group 2	Blekko	Bing	Google
Group 3	Bing	Google	Blekko

Table 4: Section B search sequences

	1	**2**	**3**
Group 1	Google	Bing	Blekko
Group 2	Bing	Blekko	Google
Group 3	Blekko	Google	Bing

Reduce Other Biases: To reduce the bias due to the prior usage of a search engine, I ask students to try their best to be fair and not to let their past experiences to affect the evaluation in anyway. I also told them that they should not bias against or towards Bing because my name is Bing too!

Queries: The evaluation had no given queries. The students were asked to perform their daily searches as usual based on their daily information needs with no change. The only requirement was that they needed to stick to the same search engine for the week and only to use another search engine if the first search engine did not give good results.

Number of Queries: Each student was asked to come up about 40 queries to search in each week based on their information needs, not to use the same queries every week, and not to make up queries or to copy queries from another source. The students also agreed that 40 queries per week were reasonable for them.

Type of Queries: In the literature, two main types of queries have been identified [2, 8], *navigational queries* and *information queries*.

A *navigational query* is one that usually has only one satisfactory result, or there is a unique page that the user is looking for. For example, the user types "CNN" and expects to find the Web site of CNN.com, or types the name of a researcher and expects to find his/her homepage.

An *informational query* can have a few or many appropriate results, with varying degrees of relevance or utility and varying degrees of authority. Many times the user may need to read a few pages to get the complete information. For example, the user types the topic "search engine evaluation" and expects to be provided with pages related to the topic.

Tasks of Students: For each query that a student issues to a search engine, he/she needs to record four pieces of information,

1. *Type of query*: whether it is an informational query or a navigational query.
2. *Type of search*: whether the search is a general Web search, an image search or a video search.
3. *Degree of satisfaction*: For each returned results page, the student needs to evaluate it and record his/her degree of satisfaction, *completely satisfied*, *partially satisfied* or *not satisfied*. Each student decides for him/herself the criteria for *completely satisfied*, *partially satisfied* and *not satisfied*. The evaluation is thus completely free and has no given constraints or restrictions.
4. *Spam and content farms*: For each returned results page, the student also checks whether there are spam links and/or content farms.

Overall Ratings: Before and after the evaluation, the students were also asked to give their overall ratings for each search engine.

Prior Search Engine Usage: All students use Google as their primary search engine except one, who uses Bing. Before the evaluation, all students tried Bing before, but nobody had heard of Blekko.

3. Evaluation Results

I now present the evaluation results. I will first show the level of user satisfaction, which basically indicates the search quality. I will not give separate results for the general Web search, image search and video search but only give the aggregated results of them all as the number of image searches and video searches is small. I will also list the user ratings before and after evaluation. Finally, I report results about spam and content farms. In reporting the results, I will give both the combined results of the two sections and also the individual results of each section. From these, we will see something quite surprising and unexpected.

Note also that I only report the raw results without doing any statistical significance test simply due to lack of time (search evaluation is not my search area). But based on my past data mining experiences, most of the differences should be statistically significant due to the large number of queries used in the evaluation.

3.1. Search Quality Results

Combined Results of Both Sections for All 4 Weeks: Table 5 gives the overall results of both navigational queries and information queries over the four weeks for both student sections. "No." stands for "the number of queries", "%" for "the percentage", "sA" for "section A", "sB" for "section B", "w3" for "the first three weeks" and "w4" for "week 4" (the last week). Thus, sA+sB-w3+w4 represents the overall results of sections A and B over all 4 weeks.

Table 5: Results of sections A and B of all 4 weeks: sA+sB-w3+w4

Navigational Query	**Google**		**Bing**		**Blekko**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Completely Satisfied	605	90.7%	618	87.7%	512	74.4%
Partially Satisfied	40	6.0%	46	6.5%	67	9.7%
Not Satisfied	22	3.3%	41	5.8%	109	15.8%
Total	667	100%	705	100%	688	100%
Informational Query						
Completely Satisfied	1679	82.7%	1549	73.9%	1089	58.4%
Partially Satisfied	253	12.5%	342	16.3%	391	21.0%
Not Satisfied	99	4.9%	204	9.7%	384	20.6%
Total	2031	100%	2095	100%	1864	100%

From Table 5, we can see that for navigational queries, Google and Bing are very close, while Blekko is far behind. But for informational queries, Google is way ahead of both Bing and Blekko. Blekko is still rather weak at the moment. If we compare these results with those of Google and Bing (MSN and LIVE) in 2006 and 2007 [7], we notice that they both have improved tremendously, especially for informational queries.

Results of Individual Sections: Table 6 shows the overall results of section A (i.e., sA-w3+w4), and Table 7 shows the overall results of section B (i.e., sB-w3+w4). We see something rather unexpected. Although both sections of students agreed that Google is better than Bing and Bing is better than Blekko, the gaps of their differences for section B are much larger than those for section A.

I have no solid explanation for this, but based on my experiences with the students in the whole semester and their scores for the two previous presentations, I do notice the following differences of the two sections.

1. Most students (9/15) in section B had jobs, which explained why they ended up in the 5pm section. Section A met at 3pm. Only 3 (/20) students in section A worked outside. Although for privacy reasons I told the students not to give me their actually search queries, some of them did give me most of their (non-sensitive) queries. I noticed that many queries from students in section B are related to their jobs. This was also confirmed by them.
2. On average, students in section B were older, and tend to be more conservative and had strong views of likes and dislikes, while students in section A were more open to different ideas, e.g., slashtags of Blekko and pretty images of Bing in its search page, which most students in section B did not like.
3. For some reason, on average the students in section B (sB) were weaker than those in section A (sA). As an evidence of that, I was impressed with the first presentation of most students in section A, but had to ask the students in section B to re-do it because most of them did not do well. Their final grades also reflected that.

Could the larger gaps of section B be due to the fact that students in section B searched many queries related to their jobs and they were so familiar with Google's results for these or similar queries in their daily work and saw anything different as less than satisfactory? Could it be that Google is indeed better than others for such queries? Students in section A search different things each week following their course works. Or, does the stronger and the weaker articulation abilities of section A and section B have something to do with this? I do not know the answer. As we will see later, the larger gaps of section B are also reflected in their ratings given to the engines.

Table 6: Section A results of all 4 weeks: sA-w3+w4

Navigational Query	**Google**		**Bing**		**Blekko**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Completely Satisfied	378	89.6%	424	88.5%	325	80.2%
Partially Satisfied	29	6.9%	25	5.2%	37	9.1%
Not Satisfied	15	3.6%	30	6.3%	43	10.6%
Total	422	100%	479	100%	405	100%
Informational Query						
Completely Satisfied	886	80.3%	874	75.8%	588	59.0%
Partially Satisfied	158	14.3%	201	17.4%	232	23.3%
Not Satisfied	59	5.3%	78	6.8%	177	17.8%
Total	1103	100%	1153	100%	997	100%

Table 7: Section B results of all 4 weeks: sB-w3+w4

Google		**Bing**		**Blekko**	
No.	**%**	**No.**	**%**	**No.**	**%**
227	92.7%	194	85.8%	187	66.1%
11	4.5%	21	9.3%	30	10.6%
7	2.9%	11	4.9%	66	23.3%
245	100%	226	100%	283	100%
793	85.5%	675	71.7%	501	57.8%
95	10.2%	141	15.0%	159	18.3%
40	4.3%	126	13.4%	207	23.9%
928	100%	942	100%	867	100%

Results of First Three Weeks and Last Week: I now compare the results of the first three weeks and those of the last week (week 4). Recall that in the first three weeks, each engine was evaluated individually, but in the fourth week all three engines were evaluated at the same time using the same queries. Tables 8 and 9 show their results for section A, and Tables 10 and 11 show their results for section B. Surprisingly, we see that Google and Bing both recorded better results for both navigational and informational queries in week 4. Blekko recorded similar or worse results. This is interesting. I asked the students. They offered some explanations. First, when they compared results side-by-side, their expectation for each engine changed. For example, when they saw that Google and Bing produced the same or very similar results, they tended to be satisfied even though the results did not completely meet their information needs because they thought that those results were the best on the Web. Second, in week 4 they were less strict in judging the results compared to the first three weeks. Blekko's results did not go up in week 4 because when the engines were compared, Blekko was shown clearly weaker.

Combined Results of Sections A and B from First Three Weeks and Last Week: These results are given in Table 12 and Table 13. Similar trends as above are observed, and I will not discuss them further.

Table 8: Section A results of the first three weeks: sA-w3

Navigational Query	Google		Bing		Blekko	
	No.	%	No.	%	No.	%
Completely Satisfied	187	85.8%	238	86.5%	166	82.6%
Partially Satisfied	20	9.2%	14	5.1%	19	9.5%
Not Satisfied	11	5.0%	23	8.4%	16	8.0%
Total	218	100%	275	100%	201	100%
Informational Query						
Completely Satisfied	432	76.2%	444	72.0%	287	62.3%
Partially Satisfied	94	16.6%	124	20.1%	101	21.9%
Not Satisfied	41	7.2%	49	7.9%	73	15.8%
Total	567	100%	617	100%	461	100%

Table 9: Section A results of week 4: sA-w4

Google		Bing		Blekko	
No.	%	No.	%	No.	%
191	93.6%	186	91.2%	159	77.9%
9	4.4%	11	5.4%	18	8.8%
4	2.0%	7	3.4%	27	13.2%
204	100%	204	100%	204	100%
454	84.7%	430	80.2%	301	56.2%
64	11.9%	77	14.4%	131	24.4%
18	3.4%	29	5.4%	104	19.4%
536	100%	536	100%	536	100%

Table 10: Section B results of the first three weeks: sB-w3

Navigational Query	Google		Bing		Blekko	
	No.	%	No.	%	No.	%
Completely Satisfied	106	89.8%	80	80.8%	105	67.3%
Partially Satisfied	8	6.8%	13	13.1%	21	13.5%
Not Satisfied	4	3.4%	6	6.1%	30	19.2%
Total	118	100%	99	100%	156	100%
Informational Query						
Completely Satisfied	395	83.7%	324	66.7%	234	56.9%
Partially Satisfied	49	10.4%	75	15.4%	77	18.7%
Not Satisfied	28	5.9%	87	17.9%	100	24.3%
Total	472	100%	486	100%	411	100%

Table 11: Section B results of week 4: sB-w4

Google		Bing		Blekko	
No.	%	No.	%	No.	%
121	95.3%	114	89.8%	82	64.6%
3	2.4%	8	6.3%	9	7.1%
3	2.4%	5	3.9%	36	28.3%
127	100%	127	100%	127	100%
398	87.3%	351	77.0%	267	58.6%
46	10.1%	66	14.5%	82	18.0%
12	2.6%	39	8.6%	107	23.5%
456	100%	456	100%	456	100%

Table 12: Sections 1 and 2 results of the first three weeks: sA+sB-w3

Navigational Query	Google		Bing		Blekko	
	No.	%	No.	%	No.	%
Completely Satisfied	293	87.2%	318	85.0%	271	75.9%
Partially Satisfied	28	8.3%	27	7.2%	40	11.2%
Not Satisfied	15	4.5%	29	7.8%	46	12.9%
Total	336	100%	374	100%	357	100%
Informational Query						
Completely Satisfied	827	79.6%	768	69.6%	521	59.7%
Partially Satisfied	143	13.8%	199	18.0%	178	20.4%
Not Satisfied	69	6.6%	136	12.3%	173	19.8%
Total	1039	100%	1103	100%	872	100%

Table 13: Sections 1 and 2 results of week 4: sA+sB-w4

Google		Bing		Blekko	
No.	%	No.	%	No.	%
312	94.3%	300	90.6%	241	72.8%
12	3.6%	19	5.7%	27	8.2%
7	2.1%	12	3.6%	63	19.0%
331	100%	331	100%	331	100%
852	85.9%	781	78.7%	568	57.3%
110	11.1%	143	14.4%	213	21.5%
30	3.0%	68	6.9%	211	21.3%
992	100%	992	100%	992	100%

3.2. Overall Ratings

As noted earlier, students were also asked to give ratings (1-10) to each engine before and after the evaluation. Since most students were not familiar with Bing or Blekko, I did not force them to give a rating to Bing or Blekko. If a student did not provide a rating for Bing or Blekko before the evaluation, it is not counted in computing of the average rating. Table 14 shows the average ratings of section A (sA), section B (sB) and both sections (all (sA+ sB)). We can clearly see that before and after the evaluation the ratings for Google are about the same for both section A (sA) and section B (sB), but for Bing and Blekko, there are some increases after the evaluation. Bing has the largest increase. We again see that students in section B seem to be less positive about Bing and Blekko, but more positive about Google than students in section A.

Table 14: Average overall ratings before and after the evaluation: 1 (worst) -10 (best)

Overall Impression Rating	**Google**		**Bing**		**Blekko**	
	Before-eval	After-eval	Before-eval	After-eval	Before-eval	After-eval
sA	8.8	8.9	6.0	8.4	5.0	6.3
sB	9.1	9.3	4.8	7.5	4.6	5.0
All (sA+sB)	8.9	9.0	5.5	8.0	4.9	5.4

3.3. Spam and Content Farms

Table 15 gives the results of spam and content farms. The “No.” column gives the number of queries with spam or content forms in their returned results. “%” gives the percentage of queries that have spam or content farms. We see that Google and Bing are similar, and Blekko is slightly better, which is expected as Blekko tries to use only authoritative sites. But all three engines are really quite good at filtering out spam pages and content farms.

Table 15: Query results with spam or content farms

Spam or content farms	**Google**		**Bing**		**Blekko**	
	No.	%	No.	%	No.	%
sA	93	6.1%	83	5.1%	59	4.2%
sB	15	1.3%	29	2.5%	20	1.7%
All (sA+sB)	108	4.0%	112	4.0%	79	3.1%

4. Conclusions

This article described a search engine evaluation carried out by 35 undergraduate computer science students over 4 weeks. They evaluated Google, Bing, and Blekko based on their daily searches. We can say that the results reflect the opinions of the students based on their personal information needs and assessments.

It is clear that Google is still the best search engine, but Bing is not too far behind. Blekko is new and is behind the other two engines. Although during the evaluation process, I kept reminding the students not to bias against or towards any engine, I believe that biases were inevitable. As a case in point, there was one student who actually used Bing as his day-to-day primary search engine. He believed that Bing was better than Google. He said that his belief might be because he was familiar with the behavior of Bing and knew how to search it and use it. Then, could the same be said about Google for the other students who used Google as their main engine?

One main shortcoming of this evaluation is that the segment of population is narrow, i.e., undergraduate students of computer science. Future evaluations should involve people from all walks of life.

Finally, there are still a lot of qualitative results from the students that I have not analyzed such as pros and cons of each engine, user interface, ads relevance, speed, and applications associated with each engine. There are a lot of interesting things there. I will analyze them when I have time.

Acknowledgements

I would like to thank the 35 students in my CS376 class for their participation in the evaluation in Spring, 2011. I also thank Lei Zhang (my PhD student) for helping me convert the raw data in Excel from the students to a text file suitable for my Lisp program to do the analysis.

References

1. J. Bar-Ilan, Methods for measuring search engine performance over time, *Journal of the American Society for Information Science & Technology*, 53(4), p.308-319, 2002.
2. A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2), 2002.
3. V. R. Carvalho, M. Lease, E.Yilmaz, Crowdsourcing for Search Evaluation, Workshop report of The Crowdsourcing for Search Evaluation Workshop (CSE 2010), Geneva, Switzerland, in conjunction with SIGIR-2010, July, 2010.
4. H. Chu, and M. Rosenthal. Search engines for the World Wide Web: a comparative study and evaluation methodology. *Proceedings of the 59th annual meeting of the American Society for Information Science*, 1996.
5. M. Gordon and P. Pathak, Finding information on the World Wide Web: the retrieval effectiveness of search engines, *Information Processing and Management: an International Journal*, 35(2), p.141-180, March 1999
6. D. Hawking, N. Craswell, P. Bailey, K. Griffihs. Measuring search engine quality, *Information Retrieval*, 4(1), 2001.
7. B. Liu, Personal evaluations of search engines: Google, Yahoo! and MSN. 2006 and 2007, http://www.cs.uic.edu/~liub/searchEval/SearchEngineEvaluation.htm, and http://www.cs.uic.edu/~liub/searchEval/2006-2007.html.
8. D. E. Rose and D. Levinson. Understanding user goals in Web search. *WWW-04*, 2004.
9. L. T. Su, H. Chen, and X. Dong. Evaluation of Web-based search engines from the end-user's perspective: a pilot study. *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, 1998.
10. M-C. Tang, and Y. Sun. Evaluation of Web-based search engines using user-effort measures. *Libres* 13.2, 2003.
11. L. Vaughan. New measurements for search engine evaluation proposed and tested, *Information Processing and Management: an International Journal*, 40(4), 2004.
12. E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (*SIGIR-2008*), pages 603–610, New York, NY, USA, 2008