A user-oriented webpage ranking algorithm based on user attention time

Songhua Xu, Yi Zhu, Hao Jiang, Francis C.M. Lau

Computer Science Dept.
Yale University, USA

Computer Science Dept.
Zhejiang University, China

Computer Science Dept.
The University of Hong Kong

July 16, 2008

Chicago

Motivations

- Demand for personalized search engine
- Utilize implicit user feedback

Main Ideas

- Acquiring the user attention time on each document.
- Analyzing the user's interest implicitly reflected by the attention time.
- Re-rank the search results for a better user experience.
- Assumption: a user shall have more or less the same amount of interest towards similar documents.

Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

Previous Work I

- Implicit user feedback
  - White, Ruthven and Jose. “The use of implicit evidence for relevance feedback in web retrieval”. In Proc. of the 24th BCS-IRSG European Colloquium on IR Research. 2002.
  - Fox, Karnawat, Mydland, Dumais and White. “Evaluating implicit measures to improve web search”. ACM Transactions on Information Systems. 2005.
  - Fu, X. “Evaluating sources of implicit feedback in web searches”. In RecSys ’07: Proc. of the 2007 ACM Conference on Recommender Systems. 2007.

Previous Work II

- Query history
  - Google Web History
    (http://www.google.com/history)
- Click data [Documents clicked are considered as of more interest.]
  - Joachims. “Optimizing search engines using clickthrough data”. In KDD ’02. 2002.
  - Sun, Zeng, Liu, Lu and Chen. “Cubesvd: a novel approach to personalized web search”. In WWW ’05. 2005.
  - Radlinski and Joachims. “Query chains: learning to rank from implicit feedback”. In KDD ’05. 2005.
  - Dupret, Murdock and Piwowarski. “Web search engine evaluation using clickthrough data and a user model”. In WWW ’07. 2007.

Previous Work III

- Attention time [Collecting the time user pay on each document]
  - Kelly and Belkin. “Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback”. In SIGIR ’01: Proc. Of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001.
  - Kelly and Belkin. Display time as implicit feedback: understanding task effects. In SIGIR ’04: Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004.
  - Halabi, Kubat and Tapia. “Time spent on a web page is sufficient to infer a user’s interest”. In IMSA ’07: Proc. of the IASTED European Conference. 2007.

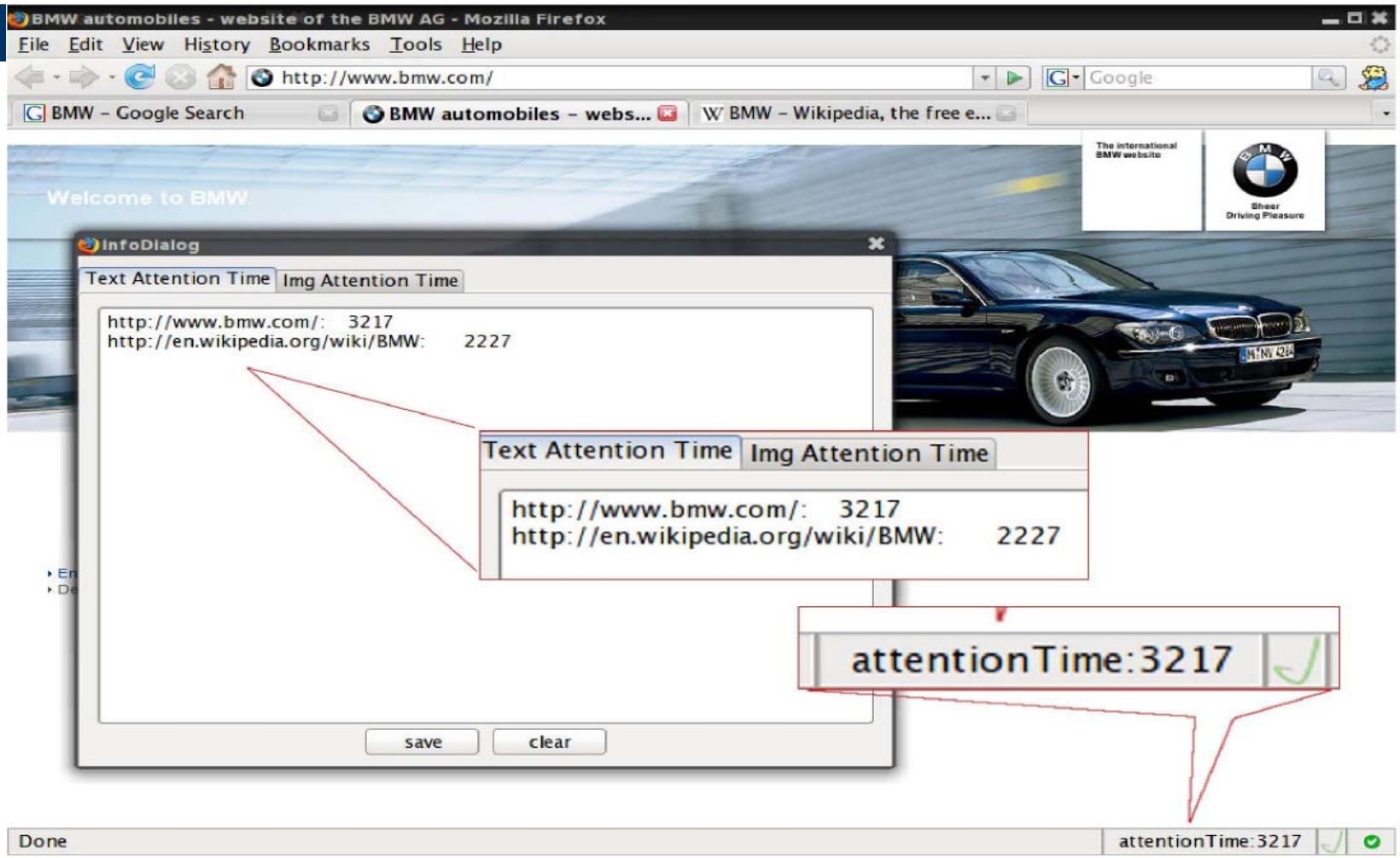
Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

User Attention Time

- For texts
  - The attention time over a document includes 1) the time a user spends on reading the summary and 2) the time in reading the actual contents.
- For images
  - The attention time includes 1) the time that a user spends on looking at the thumbnail and 2) the time on the image itself.
- We developed a customized Firefox browser to capture this type of information.

Our Customized Web Browser



Acquisition of Attention Time

- In the initial search result page, we trace the mouse or tablet pen for capturing the attention time *t1* a user spends on a certain text summary or a thumbnail image.
- For the loaded webpage containing the actual image or text content, we records the duration *t2* the document is actively displayed to the user.
- Attention time = *t1* + *t2*. To increase the accuracy, a truncation threshold, t_{max}, to represent the maximum reading time for a document of a certain length.

Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

Essential Concept

- Prediction based on the content similarity of two documents.
- We assume if the contents of two documents are sufficiently similar, then a user shall have more or less the same amount of interest towards either of them.

Estimating Document Similarities

- A good estimation of $Sim(d_0, d_1)$ plays a critical role in attention time prediction.
- For texts, we utilize the “simpack” open source package (Bernstein *et al.* 2005; Ziegler *et al.* 2006) accessible from

  http://www.ifi.unizh.ch/ddis/simpack.html
- For images, we adopt the implementation offered by the open source content based image retrieval library at

  http://www.semanticmetadata.net/lire/

Attention Time Prediction (I)

- $Sim(d_0, d_1)$ - the content similarity between document d_0 and d_1, where $Sim(d_0, d_1)$ ranges in $[0,1]$.
- $\{t_{att}(u, d_i) | i = 1, \cdots, n\}$ – acquired attention time samples by user u for documents $\{d_i\}$.
- When a new document d_x arrives, we calculate the similarity between d_x and all the documents in the training set. We then select k documents which have the highest similarity with d_x.

Attention Time Prediction (II)

- Predict the attention time for d_x by:

$$t_{att}(u, d_x) = \frac{\sum_{i=1}^{k} \left(t_{att}(u, d_i) Sim(d_i, d_x) \delta(d_i, d_x) \right)}{\sum_{i=1}^{k} \left(Sim(d_i, d_x) \delta(d_i, d_x) \right) + \epsilon}$$

where

$$\delta(d_i, d_x) = \begin{cases} 1 & \text{If } Sim(d_i, d_x) > 0.01 \\ 0 & \text{Otherwise} \end{cases}$$

Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

User-oriented Webpage Ranking (I)

- Compute a normalized attention time offset

$$t_{atten}^{offset}(i) = \frac{2\exp\left(-\kappa_d \cdot rank(i)\right)}{1+\exp\left(-\kappa_d \cdot rank(i)\right)}$$

where *rank(i)* denotes Google's webpage rank for document *i* and the parameter κ_d controls how sharp the drop-off is.

User-oriented Webpage Ranking (II)

- derive the overall attention time for *i* as

$$t_{atten}^{overall}(i) = \kappa_{overall} t_{atten}(i) + t_{atten}^{offset}(i)$$

where the parameter $t_{atten}^{overall}(i)$ is a user tunable value moderating how much he would prefer the user oriented ranking.

Prototype Search Engine

- Server Side
  - Forward user query to Google and download the first 300 records.
  - Predict the attention time for each record whose attention time is not acquired.
  - Re-rank these records through their overall attention time.
- Client Side
  - Acquire the attention time samples and periodically send to the server as well as user identification numbers.

Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

Evaluating Webpage Ranking

- We use the sum of the absolute differences of each page's rank against its rank in the user's ideal need as the error measurement for a webpage ranking result.
  - For texts, a user is asked to read the top 20 records in a ranking result and then provide his ideal ranking for these records.
  - For images, a user is asked similarly for the top 4 pages.

Text Search Results (I):“Web search technology”

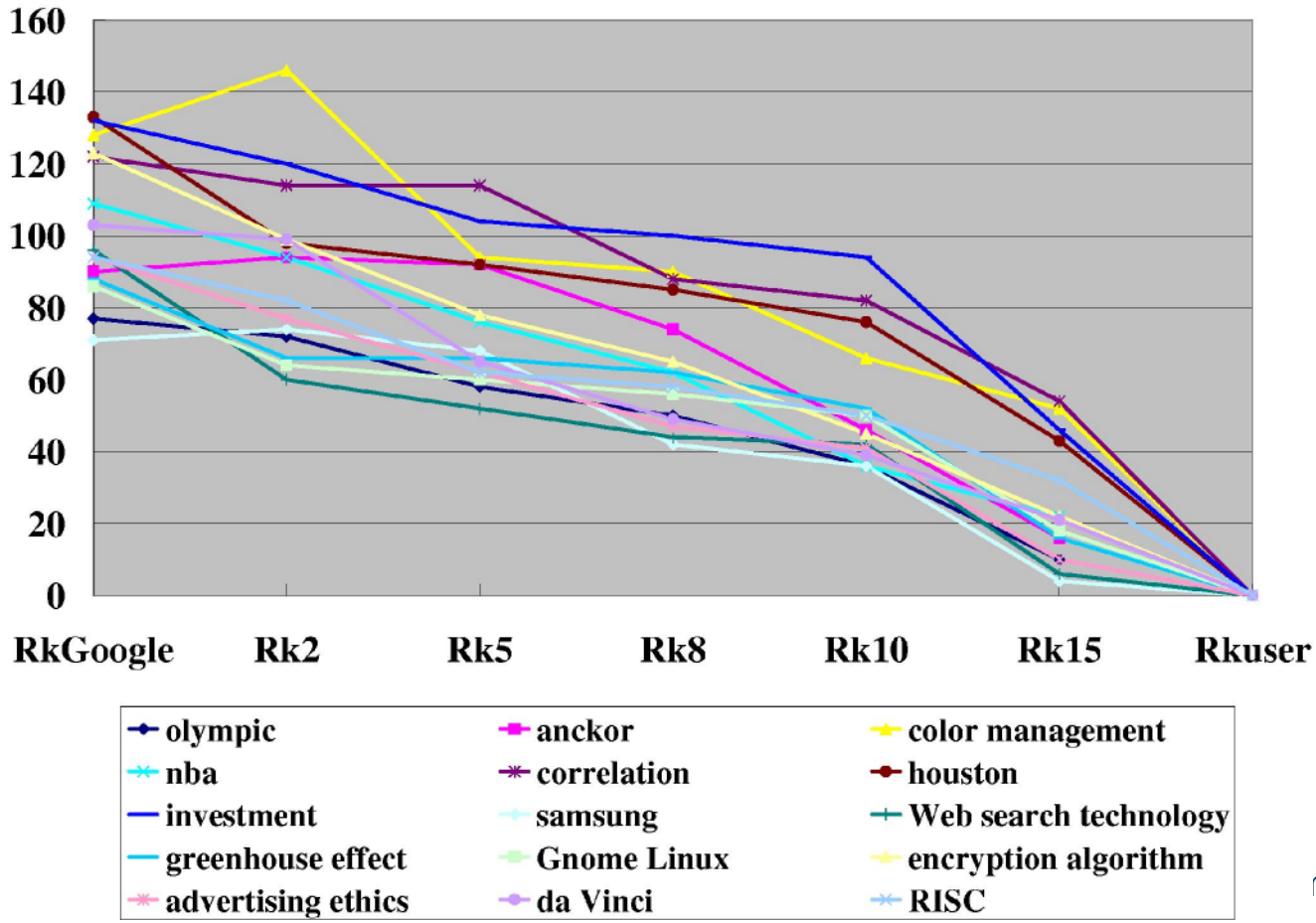
Rk_{user}	Rk_{Google}	Rk_2	Rk_5	Rk_8	Rk_{10}	Rk_{15}
6	1	1	15	13	11	7
9	4	17	16	14	12	9
1	2	2	1	1	1	1
17	3	10	17	16	16	16
2	6	3	7	2	2	2
15	5	12	9	15	15	15
16	7	13	14	17	17	17
5	8	9	4	12	10	6
11	15	13	6	6	14	11
10	14	15	13	11	13	10
14	18	16	12	9	7	14
12	16	12	11	10	9	12
3	4	4	4	3	3	3
13	11	9	10	8	8	13
8	6	5	3	4	4	8
7	5	14	8	7	6	5
4	7	8	5	5	5	4
0	**96**	**60**	**52**	**44**	**42**	**6**

Error measurement

Text Search Results (II)

Search keyword	Rk_{Google}	Rk_2	Rk_5	Rk_8	Rk_{10}	Rk_{15}
greenhouse effect	88	66	66	62	52	16
Gnome Linux	86	64	60	56	50	18
encryption algorithm	123	99	78	65	45	22
RISC	94	82	62	58	50	32
advertising ethics	94	77	62	47	41	10
da Vinci	103	99	65	49	39	21
olympic	77	72	58	50	36	10
anckor	90	94	92	74	46	16
color management	128	146	94	90	66	52
NBA	109	94	76	62	36	22
correlation	122	114	114	88	82	54
houston	133	98	92	85	76	43
investment	132	120	104	100	94	46
samsung	71	74	68	42	36	4

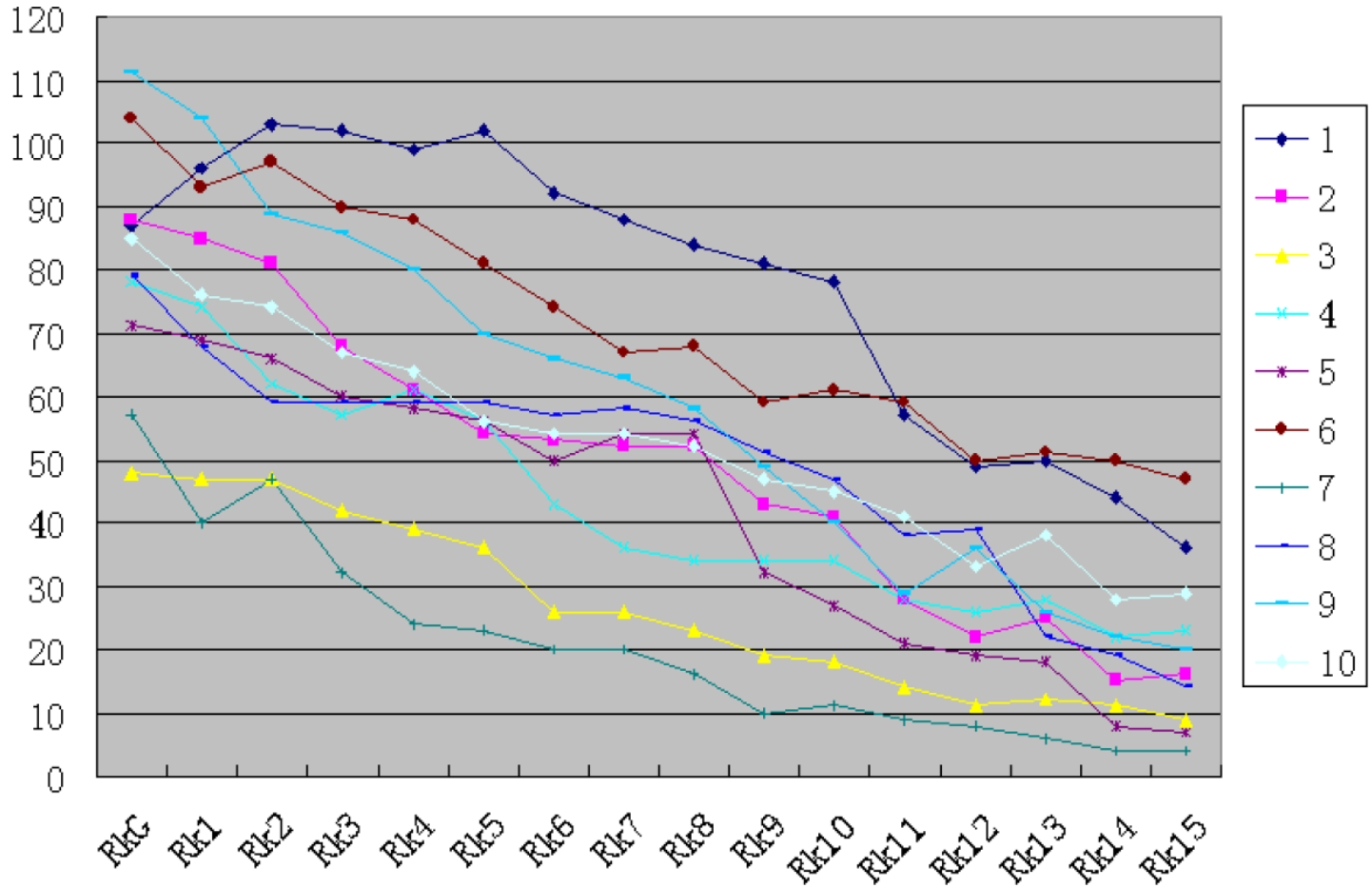
Text Search Results (II)



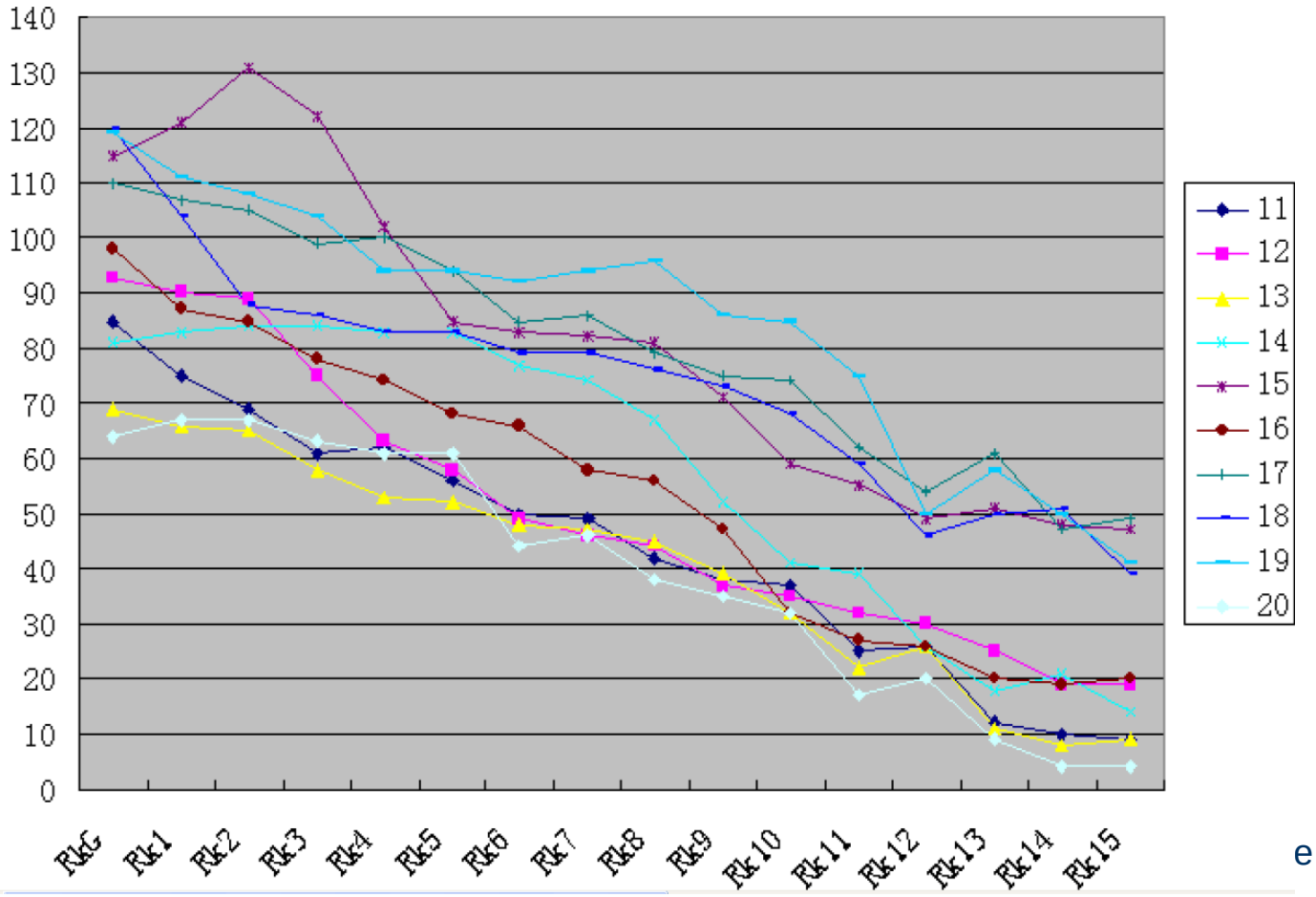
Text Search Results (III)

No.	Search Keyword	# Vs	Rk_Y	Rk_1	Rk_2	Rk_3	Rk_4	Rk_5	Rk_6	Rk_7	Rk_8	Rk_9	Rk_{10}	Rk_{11}	Rk_{12}	Rk_{13}	Rk_{14}	Rk_{15}
1	apple	20	136	114	110	73	37	32	31	31	30	31	29	29	30	30	31	31
2	car	20	80	82	81	79	70	60	42	38	47	32	30	34	15	19	8	9
3	barcelona	20	70	58	58	57	60	49	41	39	25	32	33	25	29	21	17	14
4	da vinci	20	64	63	63	50	45	48	38	25	22	33	25	27	22	18	15	15
5	ETS	20	54	49	50	34	31	36	27	22	15	11	10	10	7	2	2	2
6	gnome linux	20	62	49	49	51	44	30	35	30	18	25	18	12	13	10	9	7
7	greenhouse effect	20	34	31	30	25	23	18	25	14	17	19	11	9	9	5	7	5
8	happy new year	20	58	58	51	53	38	36	34	31	23	23	24	19	20	12	14	12
9	NBA	20	64	53	64	39	52	47	39	30	24	14	22	13	7	8	8	7
10	olympics	20	52	55	51	43	44	38	31	32	20	25	24	18	20	13	13	10
11	wow	20	104	90	91	91	76	64	61	78	63	54	54	54	49	40	38	34
12	great wall	20	77	59	69	63	45	58	50	40	31	26	33	26	22	18	12	13
13	hurricane	20	120	118	93	104	90	90	86	71	67	61	54	43	39	22	21	23
14	iron man	20	83	76	65	73	60	64	59	53	52	47	48	34	37	34	32	26
15	moon	20	67	56	63	47	47	44	49	39	33	28	32	19	16	17	16	17
16	national treasure	20	99	93	72	74	57	61	57	63	55	57	49	39	32	24	30	22
17	porsche	20	48	41	42	41	31	27	28	32	24	25	23	21	20	18	12	11
18	forbidden kingdom	20	74	76	66	62	58	56	49	36	28	27	23	21	26	20	16	15
19	tiger	20	111	112	92	79	78	78	74	71	64	70	56	48	50	42	34	35
20	west lake	20	51	51	47	34	41	34	33	18	22	17	14	6	7	6	6	5

Text Search Results (III)



Text Search Results (III)



e

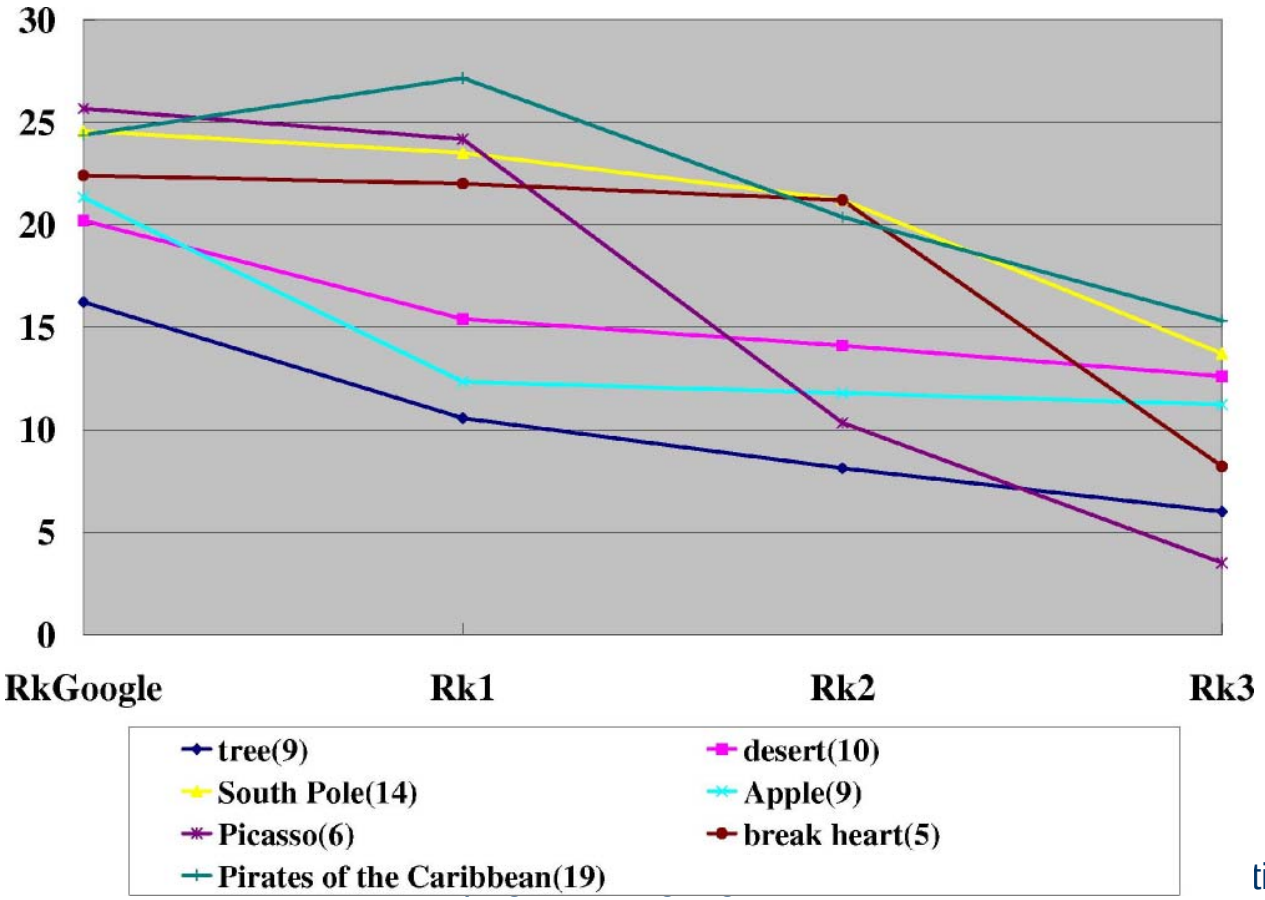
Image Search Results (I): “Web search technology”

Rk_{Google}	Rk_{1st}	Rk_{2nd}	Rk_{3rd}
9	1	1	1
16	63	3	5
17	3	2	3
23	41	15	2
41	24	37	4
48	13	4	6
25.67	**24.17**	**10.33**	**3.5**

Image Search Results (II)

Search Keyword	# Images	Rk_{Google}	Rk_{1st}	Rk_{2nd}	Rk_{3rd}
tree	9	16.22	10.56	8.11	6
desert	10	20.2	15.4	14.1	12.6
South Pole	14	24.57	23.5	21.21	13.71
Apple	9	21.33	12.33	11.78	11.22
break heart	5	22.4	22	21.2	8.2
Pirates of the Caribbean	19	24.37	27.16	20.37	15.32

Image Search Results (II)



time

Image Search Results (III)

No.	Search Keyword	# Imgs	Rk_{Google}	Rk_{1st}	Rk_{2nd}	Rk_{3rd}
1	picasso	6	25.67	24.17	10.33	3.50
2	tree	9	16.22	10.56	8.11	6.00
3	desert	10	20.2	15.40	14.1	12.60
4	south pole	14	24.57	23.5	21.21	13.71
5	apple	9	21.33	12.33	11.78	11.22
6	break heart	5	22.40	22.00	21.20	8.20
7	pirates of the caribbean	19	24.37	27.16	20.37	15.32
8	qi baishi	9	31.44	8.67	5.78	5.00
9	Victoria Harbour	7	29.29	22.00	5.14	4.00
10	eclipse	9	41.89	22.00	5.22	5.22
11	transformer	11	15.82	12.55	6.00	6.00
12	da vinci	3	31.33	19.00	9.67	2.00
13	ubuntu wallpaper	15	37.80	18.93	10.67	8.00
14	liberty	12	20.83	12.58	7.17	6.50
15	firefox	13	20.85	7.08	7.00	7.00

Organization of the Talk

- Related work
- Acquisition of user attention time samples
- Prediction of user attention time
- User-oriented webpage ranking
- Experiment results
- Future work

Future Work

- Investigate and employ algorithms that work for both text and image elements in measuring document similarity.
- Apply onto user-oriented multimedia ranking.
- Investigate and employ learning algorithms that work for document ranking, based on the user's ideal ranks provided.