The Measurement of Relative Recall with Weights: a Perspective of User Feedback

Juncheng Wang*
Center for Studies of Information Resources, Wuhan University, Wuhan, China
Email: w.juncheng@gmail.com

Zhenzhen Fu, Jun Cheng, and Feicheng Ma
Center for Studies of Information Resources, Wuhan University, Wuhan, China
Email: {zhenzhenfu, chengjun116, feicheng.ma}@gmail.com

***Abstract*—For a long time, recall is a key indicator in evaluating the retrieval performance of search engines. However, with the fact that the total documents on the internet are hard to access completely and users always focus merely on the first few pages, we believe that the traditional recalls can't undertake the function of evaluation effectively any more. As a result, this paper proposes a new modified recall algorithm named as R-W(n), which not only focuses on the top N relevant results judged by users but also brings in different weights for different rankings. Meanwhile, we develop an experimental system, which is similar to a meta-search engine, to gather users' feedback and related data. And then, based on the criteria of measuring recall algorithm's effectiveness we propose and the experimental data gathered, the results between R-W(n) and traditional recall algorithms are compared. Finally, we draw a conclusion that the R-W(n) is superior to traditional recall algorithms, for it solves the weakness presented before and performs better in discerning good search engines from bad ones.**

***Index Terms*—relative recall, weights, relevant feedback, AHP**

I. Introduction

Nowadays, information around the world increases in an exponential rate, so that we always meet with the problem that how to find appropriate information as much as possible [1]. Fortunately, there are many different types of search engines offering various information retrieval services for us [2]. When we want to search for certain information, we just Google it. Meanwhile, a reality question arises: with so many search engines at hand, which performances better, and we should choose which one when meet with a specific retrieval task.

In fact, some scholars have noticed it as early as 1950s [3], and mainstream indicators and methods for retrieval performance measurement, including precision and recall, has formed. However, in order to make the two indexes more effective in differentiating retrieval efficiency of different search engines, many scholars have been making improvements [4-7]. Of them, recall is the most fascinating, which is also the focus of this paper.

Because the total relevant documents on the internet are hard to access completely, traditional recall seems imperfect [8]. Besides, based on the user-centered philosophy, we believe that the traditional recall doesn't take it into consideration fully. So, this paper will introduce a modified recall algorithm and bring forward experimental approaches to assess its effectiveness.

The modified recall algorithm named R-W(n) is to reach the goals as follows:

(1) Considering that information is overflow on the internet and the amount of total relevant documents can hardly be figured out, we bring the idea of relative recall, and develop a system to reuse the documents retrieved from meta-search engines. They are used to build a pool that is assumed to contain all relevant contents.

(2) Consider users' searching habits. We only focus on the top N relevant results, for that recent studies show most users mainly focus on the several top pages returned by search engines and ignore the others.

(3) Instead of treating all the relevant results equally, we employ users' evaluation toward the relevant results, and add weights to the top N results according to their evaluation.

Our experimental platform is similar with a meta-search engine, but the major difference is that it can record participants' feedbacks towards the results returned by various search engines. Criteria of measuring recall algorithm's effectiveness are based on two assumptions: First, measuring search engines' efficiency with different recall algorithms, the bigger significant diversity of recall scores among those search engines, the more effective this recall algorithm is, for it can do better in discerning a good search engine from a bad one. Second, the better an algorithm can represent users' opinion, the more effective the algorithm is.

The rest of this paper is organized as follows. We overview the previous researches of recall measurement in section 2. In section 3 we briefly introduce the structure of the system, which shows hits from eight search engines to users, and records the participants' "yes/no" judgments and ranks of the results according to their relevant degree.

Manuscript received April 13, 2010; revised July 11, 2010; accepted July 14, 2010.

* Corresponding author. Tel: 86-15337127482; fax:86-27-68754541

doi:10.4304/jcp.5.11.1686-1692

In section 4 we present the formula of three kinds of recall algorithm, calculate the recall scores of eight search engines in the experiment, and then judge efficiency of the recall modification by comparing the three recall algorithms referring to the criteria mentioned above in section 5. We present our conclusions in section 6 and briefly discuss our future research in section 7.

II. Backgrounds And Motivations

The research on recall and precision derived from mid-1950s when Information Retrieval (IR) just began. The concepts of recall and precision were first introduced by Perry in 1955, and Cleverdon first made them as indexes to evaluate IR system's efficiency [9]. Recall measures how efficient the system is when retrieving the relevant documents from the WWW, while precision measures the relevance of the retrieved documents to the user requirements [10]. Nowadays, recall and precision have become two key indicators of IR system's efficiency [11-13].

However, traditional Precision and Recall is imperfect under the background of the information overflow, for the total relevant document collection on the internet is hard to assess. Hence, alternative algorithms have been proposed, like relative recall (R for short). R is defined as the number of relevant document retrieved by a system divided by the total number of the relevant documents retrieved by all the selected systems [14].

Relative recall becomes a classic index for evaluating the retrieval performance of search engines. As research progresses, some scholars discover that this kind of recall algorithm still has its own limitations, that is it takes account of all relevant documents system returns. While according to research in [15], many users don't read documents after certain pages that search engine returns, thus subsequent documents which users don't concern are useless in satisfying users' needs. In this sense, the algorithm of relative recall which considers top N relevant documents is more realistic. We call this kind of recall R-N for short.

However, R-N doesn't figure out relevant degree of each result in the top N collection, in other words, R-N treats the N results with the same importance, which is inappropriate. For users don't simply grant binary weight "yes / no" to results, their answers could be in between. This reveals that there are some defects in R-N in distinguishing different search engines' efficiency.

This paper proposes a modified algorithm of recall named as R-W(n), inheriting the idea that the top N relevant results are mainly concerned, but R-W(n) ranks the top N relevant results with different weights according to users' ranking, while R-N ranks all the relevant results with the same weight of "1". We use the analytic hierarchy process (AHP) to assess weights. Details about how to calculate weight with AHP will be introduced later.

III. Experimental Method

A. Experimental model

This experimental model is built to gather users' evaluating results, for calculating recall with each recall algorithm. This system is based on the research of Gorden and Pathak (1999) [16], Can et al. (2004) [17] and Morrison (2008) [18]. In their experiment Meta Search Engine was established with adequate search engines and participants were asked to give information need according to the given topics and entered corresponding search query. Then they judged whether the results search engine returned were relevant to the information need. On the basis of their experimental platform, some improvements are put forward in this study. Our experimental model is shown in fig.1.

B. Process

The experimental procedures are given as follows: First, participants enter a query according to assigned topics (as in fig.2). The experiment system submits the query to the IR systems and retrieves in eight search engines, and then collects top 10 results from each search engine and stores them into local database. Second, the system gets data from database and shows them to participants randomly as fig.3. Participants judge whether the results are relevant or not, their judgment results will be recorded in database. Third, the system shows all of the results participants consider relevant to participants who will rank them through moving the "up" and "down" button (as in fig.4). The ranking results will be stored into the database and the top ten results in ranking list is the source of calculating top 10 relevant recall. After that, system returns the 10 most relevant results to participants in form of a ten-dimension matrix as in fig.5, and they compare relevant degree of the results in pairs. At last, weights of the top 10 results are calculated with MATLAB.

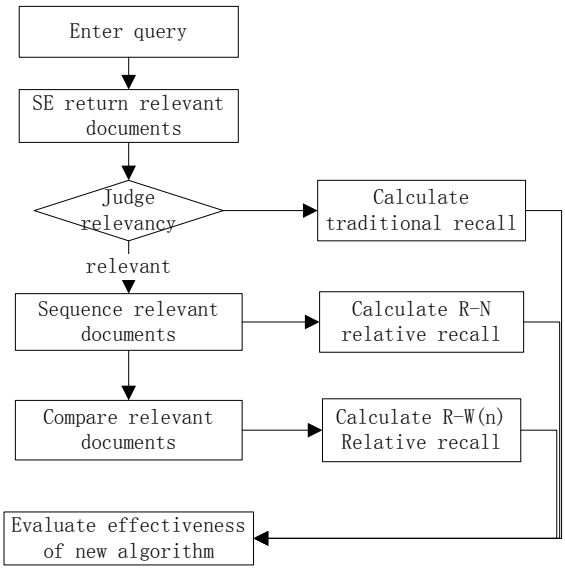

Figure 1. Relevant experiment model

Figure 2. Retrieval interface

Figure 3. Search results presentation

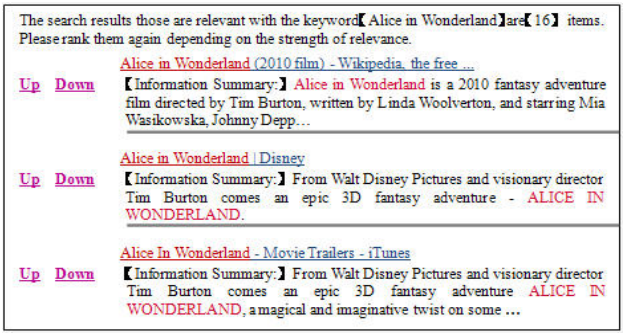

Figure 4. Rank interface

Figure 5. Comparing in pairs interface

C. *Relative subjects*

Five typical topics are selected in this experiment, including news, entertainment, research, facts and exact site (Table 1). The participants are selected from graduates in the School of Information Management at Wuhan University. Each participant is assigned two topics randomly. Moreover, the experiment is specially designed as a laboratory session of a professional course with the consideration that the participants can take it seriously and the results will be more accurate.

TABLE I.

RESEARCH OBJECTS LIST OF INFORMATION TOPICS

Topic	Explanation
News	The report on the latest facts
Entertainment	The information related to entertainment, such as movie, teleplay, music, etc.
Research	The information related to scientific research, including Social sciences, humanities, science, medicine, etc.
Facts	The required results is a short factual description without personal opinion and judgment
Exact site	Targeting to a specific document, web page or web site in order to location some information exactly

Eight Chinese IR systems that belong to directories, search engines and folksonomies respectively are chosen for this study (Table 2). For one reason, the three types of systems are mainstream services in current IR markets. For another reason, the study can be compared with Morrison's research. In the process of selection, two detailed principles are used. First, the IR system should possess its own resources and retrieval algorithm. Second, the ones that rank relatively higher in the Alex for each type are chosen. Taking these factors into account, Google Directory and Sougou are chosen for the directory, which have relatively comprehensive and accurate categories and enormous users. For the search engine, Baidu, Baidu News and Google Scholar are selected. Baidu is the largest Chinese web search engine, while Baidu News and Google Scholar are vertical search engines that could be used to compare with the comprehensive search engine [19]. 365key, Baidu Zhidao and Baidu Cang are picked out as the folksonomy. Baidu Zhidao is an online Q&A system, whose resources are provided by users. Baidu Cang and 365key are online bookmarking systems just like Delicious, in which users add interesting documents, and then tag and describe them by their own judgments [20].

TABLE II.

RESEARCH OBJECTS LIST OF IR SYSTEM

IR system	URL
Google Directory	http://www.google.com/dirhp
Sougou	http://www.sogou.com/dir/
Baidu	http://www.baidu.com
Baidu News	http://news.baidu.com
Google Scholar	http://scholar.google.cn/
365key	http://www.365key.com/
Baidu Zhidao	http://zhidao.baidu.com/
Baidu Cang	http://cang.baidu.com/

IV. Recall Calculation

With the data of participants' evaluation to the results retrieved by eight search engines, in this section, we calculate each search engine's recall referring to the three recall algorithms mentioned before.

A. Traditional relative recall (R)

Traditional relative recall focuses on all the documents labeled as relevant by users. It is the ratio of the number of relevant documents retrieved by the specified search engine to the total number of relevant ones from all search engines.

Here is a formulation of the traditional relative recall with matrix: search engines in the experiment are expressed as S_i (i=1，2…，m), and queries are expressed as X_j (j=1，2，…，n). When using the query X_j to search in the search engine S_i, the number of relevant results is denoted as a_{ij}. Therefore, the results of each search engines based on each query can be presented by matrix P:

$$P=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad (1)$$

According to the algorithm of traditional recall, the relative recall of search engine S_i corresponding to a certain query X_j can be presented as (2):

$$R_{ij}=a_{ij}/\sum_{i=1}^{m}a_{ij} \quad (2)$$

By averaging recalls of search engine S_i with different queries, we can get the recall of the search engine S_i in the end, the formulation is seen as (3):

$$R_i=\sum_{j=1}^{n}R_{ij}/n \quad (3)$$

The results of traditional relative recall of the eight search engines in the experiment are shown in table 3:

TABLE III.

Traditional Recall of the Search Engines

	Baidu	Google Directory	Baidu News	Baidu Zhidao
Recall	0.314	0.012	0.070	0.134
	Baidu Cang	Google Scholar	Sougou	365key
Recall	0.054	0.083	0.310	0.019

B. Top-N relative recall (R-N)

The R-N algorithm considers the top N relevant results ranked by participants to measure the ability of search engine in returning the top N relevant results. When calculating R-N, the matrix of relevant results is presented as (4):

$$P_1=\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix} \quad (4)$$

P1 possesses some different properties to P in (1), where b_{ij} means the number of documents that the search engine S_i returned in the top N relative documents when searching a certain query X_j, and $\sum_{i=1}^{m} b_{ij} = N$.

Note that the top N relevant results are gotten from users by ranking all of the relevant ones using the "moving up/ moving down" button as in fig.4. Here we set N=10 and get each search engine's R-N results in table 4:

TABLE Ⅳ.

R-N of The Search Engines

	Baidu	Google Directory	Baidu News	Baidu Zhidao
Recall	0.423	0.017	0.060	0.153
	Baidu Cang	Google Scholar	Sougou	365key
Recall	0.036	0.062	0.283	0.011

C. Top-N relative recall with weights (R-W(n))

As an modified algorithm based on R-N, R-W(n) combines the relative recall and weights, where W presents the weight of each top N relevant results. We use AHP to calculate the weight, and specific calculation is as follows:

1）Establish judgment matrixes

The aim of establishing a judgment matrix is to quantify users' judgment, in other words, to compare every two elements in a quantitative way. AHP scale method with span of 1-5 is used as the judging criteria, and the top N relevant results are compared in pairs by participants where N is 10 here. At last we get the 53 judgment matrixes. Now take one of the judgment matrixes for example:

$$B=\begin{bmatrix} 1 & 1 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 5 \\ 1 & 1 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 5 \\ 1/3 & 1/3 & 1 & 2 & 2 & 3 & 4 & 4 & 4 & 4 \\ 1/3 & 1/3 & 1/2 & 1 & 1 & 3 & 4 & 4 & 4 & 4 \\ 1/3 & 1/3 & 1/2 & 1 & 1 & 3 & 4 & 4 & 4 & 4 \\ 1/4 & 1/4 & 1/3 & 1/3 & 1/3 & 1 & 3 & 4 & 4 & 4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 1/4 & 1/3 & 1 & 3 & 3 & 3 \\ 1/5 & 1/5 & 1/4 & 1/4 & 1/4 & 1/4 & 1/3 & 1 & 2 & 2 \\ 1/5 & 1/5 & 1/4 & 1/4 & 1/4 & 1/4 & 1/3 & 1/2 & 1 & 1 \\ 1/5 & 1/5 & 1/4 & 1/4 & 1/4 & 1/4 & 1/3 & 1/2 & 1 & 1 \end{bmatrix}$$

2）Consistency test of each judgment matrix

The purpose of the consistency test is to judge the validity of the matrix, the indicator of which is named CR (Consistency Ratio). CR is the ratio of CI (Consistency Index) to RI (Random Index): CR=CI/RI. From the mathematical respect, if CR<0.1, the consistency of the judgment matrix is logical, or the matrix should be modified or treated invalid.

Here we use MATLAB to calculate the score of CR and skip its specific formula here. We make a consistency test for the 53 matrixes and 39 of them are logical.

3）Confirm the weight of top N relevant results

We get the weight of top N relevant results by calculating the eigenvector of valid matrixes in step 1 using the MATLAB. The weight of each top 10 relevant result is the average of that of all the valid matrixes. Their weights are shown in table 5:

TABLE V.

WEIGHT OF THE TOP N RELEVANT RESULTS

W1	W2	W3	W4	W5
0.1853	0.1639	0.135	0.1138	0.0979
W6	**W7**	**W8**	**W9**	**W10**
0.0821	0.0675	0.058	0.0526	0.0438

4) Calculating R-W(n)

R-W(n) is a modified recall algorithm of R-N with weight, so we calculate it on the basis of R-N. At first we build a matrix P_1 like (4), and then we build a matrix as (5):

$$P_2 = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{l1} & \cdots & d_{ln} \end{pmatrix} \quad (5)$$

where d_{ij} is the coding number of search engine from which the i^{th} most relevant result came under the query X_j. The coding number from 0 to 7 represents different search engines in the experiment as introduced above; and $d_{ij}=\{0,1,\ldots,7\}$, $l=\{1,2,\ldots,10\}$, for the experiment takes the top 10 relevant results as the study object.

Consider about a certain search engine m, we build a matrix P3 in (6):

$$P_3 = \begin{pmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{l1} & \cdots & e_{ln} \end{pmatrix} \quad (6)$$

where $e_{ij}=\{0,1\}$ shows whether the i^{th} most relevant result about the query X_j belongs to the search engine m, if true then e_{ij} is 1, else 0.

Then the R-W(n) of search engine m with query X_j is presented as (7):

$$R_{mj} = \sum_{i=1}^{l} w_i c_{ij} \,/\, W \quad (7)$$

where W is the sum of results' scores, and w_i is the i^{th} relevant result's weight.

Calculating R-W(n) with all queries to get the average, is the R-W(n) of search engine m. Each search engine's R-W(n) is shown in table 6:

TABLE VI.

R-W(N) OF THE SEARCH ENGINES

	Baidu	Google Directory	Baidu News	Baidu Zhidao
Score	0.4587	0.0145	0.0578	0.1443
	Baidu Cang	Google Scholar	Sougou	365key
Score	0.0308	0.0763	0.2624	0.0091

V. ANYLASIS

We have calculated the scores of R, R-N and R-W(n) of the eight search engines respectively in the experiment. The three types of scores will be compared in pairs: First, compare "R-N" with "R" to test the effects of modification that "consider the top N relevant results that users pick out"; then compare "R-W(n)" with "R-N" to test the effects of "add weight to each result". When comparing in pairs, we firstly judge whether there is significant difference between the two sets of recall scores through rank sum test, and then do Analysis of Variance for the two sets of recall scores to measure the difference degree.

A. Comparison between R-N and R

Before comparing the scores of R-N with R, we should do a rank sum test with the two sets of results to determine whether there is significant difference. Here we use SPSS software for Wilcoxon Matched-Pairs Signed-Ranks Test. Result shows that P(2-tailed)=0.727 in sign test, which is much bigger than 0.05, meaning there isn't significant difference between the two data sets. It shows that R-N is a relative algorithm of R instead of a totally different one, thus the result is acceptable. Furthermore, we determine the differences of the two data sets through comparing some variances.

TABLE VII.

STATISTICS DESCRIPTION OF R AND R-N

	Statistics description				
	N	**Mean**	**Std. Deviation**	**Minimum**	**Maximum**
R	8	0.1245	0.12181	0.012	0.314
R-N	8	0.1306	0.14884	0.011	0.423

Table 7 shows that the Std. Deviation of R and R-N is 0.1218 and 0.1488, proving that the recall discrepancy among search engines using the R-N algorithm is bigger. The maximum with R-N is bigger than that with R, and the minimum is smaller than that with R. Comparing table 3 and table 4, three search engines' recall increase in varied extents using the algorithm of R-N, of which "Baidu" increases by 34.7% and "Google Directory" by 41.7%; while other five search engines' recall decrease in varied extents, of which "365key" falls by 42%. The decrease of a search engine's recall using the R-N algorithm shows that its retrieved results belong more to the non top N relevant ones. While the increase of a search engine's recall like "Baidu" shows that the results it returned belong more to the top N relevant results. We can see that R-N algorithm can do better in judging which one is better between "Baidu" and "Sougou" on the standard of Recall. In general, R-N algorithm is better than R algorithm in making the difference between search engines more apparent, so R-N is more effective in discerning a good search engine from a bad one.

B. Comparison between R-W(n) and R-N

The purpose of comparing R-W(n) with R-N is to analyze the influence of adding weight of results into

recall algorithm, and to evaluate the effect of the modified algorithm. First of all, do Wilcoxon Matched-Pairs Signed-Ranks Test with the two sets of recall scores. Result shows that in sign test P (2-tailed) =0.289, which is bigger than 0.05, showing that there isn't significant difference between the two data sets. Thus R-W(n) is an similar algorithm instead of a totally different one, and the result is also acceptable. Then determine the different degree of the two data sets by comparing the variances.

TABLE VIII.

STATISTICS DESCRIPTION OF R-N AND R-W(N)

Statistics description					
	N	**Mean**	**Std. Deviation**	**Minimum**	**Maximum**
R-N	8	0.1306	0.1488	0.011	0.423
R-W(n)	8	0.1318	0.1566	0.009	0.459

Table 8 shows that Std. Deviation of R-N and R-W(n) is 0.1488 and 0.1566, proving that the recall discrepancy of different search engines using the R-W(n) algorithm is bigger. The maximum with R-W(n) is bigger than that with R-N, and the minimum is smaller than that with R-N, further proving that the dispersion of the data using R-W(n) is bigger than that with R-N. Two search engines' recall increase in varied extents using the algorithm of R-W(n), of which "Baidu" increases by 8.5% and "Google Scholar" by 22.6%; while other six search engines' recall decrease in varied extents, of which "365key" falls by 22.6%. The decrease of a search engine's recall shows that the results it returns don't rank in the top N relevant documents, while the increase of a search engine's recall shows that the results, which this search engine returns, rank higher in result sets. The ranking order of searching engines is the same in R-W(n) and R-N, while the difference of recall between two adjacent search engines is larger. Taking "Baidu" and "Sougou" for example, the difference of recall is 0.11 with R-N and 0.20 with R-W(n). In total, R-W(n) algorithm is better than R-N algorithm in making the difference between search engines more apparent, so R-W(n) is more effective in discerning a good search engine from a bad one.

VI. CONCLUSIONS

The R-W(n) algorithm contains two important elements: first, constraint the study object to the top N relevant documents, which refines the function of examining the ability in returning the top N most relevant documents. This modification corresponds to users' search habits in real life; second, it borrows the idea of calculating weights in AHP and adds different weights to the relevant results by users' ranking, which makes measurements of results' relevant degree more precise. To the question of how to evaluate a recall algorithm's efficiency, we propose criteria based on two considerations: how effective it is to distinguish the good search engines from the bad ones, and how effective it is to satisfy users' needs. In the experiment, we use Wilcoxon Matched-Pairs Signed-Ranks Test and Analysis of Variance to analyze the discrepancy of different search engines with different algorithms.

In the process of the experiment, the data source is mainly from participants' evaluating result, the proceeding represents the idea of "user care"; what's more, to ensure the data accuracy, this experiment is done in an experimental class and data is selected by consistency test. All of the processes can make the experiment conclusion more persuasive.

After analyzing the data using the three kind of Recall (R, R-N and R-W(n)) algorithms in pairs, we can reasonably conclude that taking the top N most relevant results as our study object and giving weights to the results by the order ranked by participants can improve the performance of the recall without deviating from the traditional algorithm.

VII. DISCUSSIONS

The present study also has some limitations which could be addressed in future work. In the experiment eight representative search engines are chosen. However, in subsequent experiments, choosing more search engines to study may do better in analyzing the different performance in recall between these search engines. In addition, when building the judgment matrix in the experiment, scale method with span of 1-5 is used, while a more precise scale method refers to the span of 1-9, so we can make use of the 9-level scale method in the further research to raise the evaluation accuracy.

On the research of Recall, many researchers propose various improved algorithm. Now we mainly discuss the influence of weights to Recall's evaluation effectiveness. Besides weight, there are many other elements to be studied, like the position of search results in the retrieval system. At the same time, the relationship between Recall and Precision is so tight that they cannot be separated. Therefore the research of other elements' influence on Recall, the relationship between Recall and Precision are the next step of our research.

ACKNOWLEDGMENT

This study was supported by the National Natural Foundation of China (Grant No. 70833005) and Self-research program for Doctoral Candidates (including Mphil-PhD) of Wuhan University in 2008. The authors would like to express our sincere gratitude to the contributing authors and to the referees for reviewing papers for this special issue.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval, " Addison-Wesley Longman Publishing Co., nc., Harlow, UK, 1999.

[2] M. Melucci and D. Hawking, "Introduction: A perspective on Web Information Retrieval," Information Retrieval, vol. 9, no. 2, pp. 119-122, 2006.

[3] A. Kent, M. M. Berry, F. U. Luehrs Jr, and J. W. Perry, “Machine literature searching VIII. Operational criteria for designing information retrieval systems,” American documentation, vol. 6, pp. 93-101, 1955.
[4] G. Salton and C. Buckley, “Improving retrieval performance by relevance feedback,” Journal of the American Society for Information Science, vol. 41, no. 4, pp. 288-297, 1990.
[5] S. P. Harter and C. A. Hert, “Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods,” Annual Review of Information Science and Technology, vol. 32, pp. 3-94, 1997.
[6] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction,” Proceedings of DARPA Broadcast News Workshop, Herndon, VA, pp. 249–252, 1999.
[7] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation,” Proceedings of the 27th European Conference on Information Retrieval, pp. 345-359, 2005.
[8] E. Yilmaz, J. A. Aslam, and S. Robertson, “A new rank correlation coefficient for information retrieval,” Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 587-594, 2008.
[9] C. W. Cleverdon, J. Mills, and M. Keen, “Factors determining the performance of indexing systems,” vol. 2, Aslib Cranfield Research Project, Cranfield, 1966.
[10] M. Buckland and F. Gey, “The relationship between Recall and Precision,” Journal of the American Society for Information Science, vol. 45, no. 1, pp. 12-19, 1999.
[11] D. Hawking, N. Craswell, P. Bailey, and K. Griffihs, “Measuring search engine quality,” Information Retrieval, pp. 33–59, 2001.
[12] B. T. S. Kumar and J. N. Prakash, “Precision and Relative Recall of Search Engines : A Comparative Study of Google and Yahoo, ” Singapure Journal of Library& Information Management, vol. 38, pp. 124-137, 2009.
[13] 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, Baltimore, MD, pp. 127-135, 1996.
[14] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," ACM Transaction on Information Systems, ACM, NY, pp. 205-229, 1989.
[15] C. Silverstein, M. Henzinger, and M. Moricz, “Analysis of a very large web search engine query log,” ACM SIGIR Forum, vol. 33, no. 1, pp. 6–12, 1999.
[16] M. Gordon and P. Pathak, “Finding information on the World Wide Web: The retrieval effectiveness of search engines,” Information Processing and Management, vol. 35, pp. 141–180, 1999.
[17] F. Can, R. Nuray, and A. B. Sevdik, "Automatic performance evaluation of Web search engines," Information Processing and Management, vol. 40, no. 3, pp. 495-514, 2004.
[18] P. J. Morrison, “Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web,” Information Processing and Management, vol. 44, no. 4, pp. 1562-1579, 2008.
[19] P. Mayr and A. K. Walter, "An exploratory study of Google Scholar, "online information review, vol. 31, no. 6, pp. 814-830, 2007.
[20] C. Xu and H. Chu, “Social tagging in China and the USA: A comparative study,” Proceedings of the American Society for Information Science and Technology, vol. 45, no. 1, pp. 1-9, 2009.

Juncheng Wang was born in Yichang City, Hubei Province, the People’s Republic of China, on January 6th, 1984. He received the B.S. and Master degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively. Now he is a Ph.D. candidate in informatics at the School of Information Management, Wuhan University. His research interests include information retrieval, information organization, and Information Lifecycle Management.

He published his first research article when he was a junior, and till now, he has published more than twenty papers, including both Chinese and English ones.

Zhenzhen Fu was born in Jingzhou City, Hubei Province, the People’s Republic of China, on May 22th, 1986. She graduated from Wuhan University to obtain the B.S. degree in 2008.

She is currently a Master student of informatics at the School of Information Management, Wuhan University. Her research interests refer to information system design, information retrieval and information policy.

Jun Cheng was born in Shanxi Province, the People’s Republic of China, on December 9th, 1988. She is currently a junior at the School of Information Resource Management in Wuhan University. Her research interests include information system and information organization.

Feicheng Ma was born in August, 1947. He received the Master degree from Wuhan University, Wuhan, China, in 1983. His research interest include informatics theory, information economy and information resource management.

He is currently a professor and Director of Information Resource Center in Wuhan University, and was Dean of the School of Information Management in Wuhan University. He has published ten books and more than 100 journal papers in the field of informatics. In addition, he is a member of the editorial board of Journal of Science.