A Mathematical Formula for the Search Engine Ranking Efficiency Evaluation Tool S.E.R.E.E.T

Moiez Tapia[1], Wadee S. Alhalabi[2], Miroslav Kubat[3]
[1,3]Electrical and Computer Engineering Department, University of Miami, USA
[2]Computer Technology Department, College of Technology, KSA
[1]mtapia@mimai.edu,[2]w.alhalabi@umiami.edu,[3]mkubat@miami.edu

Abstract- The rapid developments in the field of internet search engines underline the need for reliable method to evaluate its performance. So far, the vast majority of researchers have relied on the "precision" and "recall" measures known from the field of Information Retrieval Unfortunately, both of them fail to assess how successfully they rank the returned documents according to their relevance. In this paper, we discuss this issue in some detail, and then propose a new mechanism for the evaluation of the quality of search-engine rankings.

Index Terms—SEREET, Search Engine Ranking, Information Retrieval Evaluation

I. INTRODUCTION

THE Internet revolution gave rise to the search engine, the only tool capable of identifying among the billions of web sites those that are relevant to the user's needs. Starting from mid 1990s, hundreds of companies specializing on these tools have appeared. Many of them have gone out of business; others have merged, and yet others have joined this thriving market only recently, seeking either to outperform their predecessors, or to fit previously unexplored niches.

The principle of this tool is simple. Upon the entry of user's query, the search engine analyzes its repository of stored web sites and returns a list of relevant hyperlinks ordered by the relevance of the web sites to what the user needs. Many mechanisms to assess this relevance have been exploited, among them keyword frequency, page usage, link analysis, and various combinations of these three. Each of the multitude of alternative ranking algorithms leads to a different hyperlink ordering. Hence it becomes necessary to determine as to which of these algorithms yields the best results in terms of offering the most realistic set of hyperlinks to an average user query.

A two-pronged strategy is necessary if the question is to be answered in a satisfactory manner. First, we need appropriate experimental procedures that submit to the machine well-selected testing queries to which the relevant answers are known. Second, we need performance criteria to evaluate the quality of the search engine responses to the testing queries.

In this paper, we focus on the latter aspect. As discussed in the next section, the previous research has predominantly used the current classical performance metrics of precision and recall that are commonly used in the field of Information Retrieval. However, the utility of these metrics for search engine evaluation is limited: precision and recall establish whether the returned list contains the predominantly relevant links, and how many relevant links are missing. What they ignore is whether more relevant links find themselves high up in the list.

We begin by an extensive survey of related work in Section 2. Section 3 addresses our method accompanied by examples and comparison with existing algorithms. In section 4, we present the discussion and conclusions of our work.

II. RELATED STUDIES

Precision and Recall is the most widely used tool to evaluate an information retrieval system. It is used by scientists to evaluate retrieval information systems. Zhang and Dong [8] present a review of many ranking algorithms and discuss the deficiencies in the existing techniques. The authors propose an algorithm with a multidimensional technique and claim an improvement in the ranking result. Their algorithm produces more relevant documents and better precision. Shafi and Rather [20] use Precision and Recall to evaluate the performance of five different search engines. Chu and Rosenthal [11] present the same evaluation criteria for retrieval performance as the work proposed by Shafi and Rather [20]. However, they use precision and response time instead of precision and recall. Li and Danzig [23] introduce a new ranking algorithm. They argue that their technique is much better in space and time complexity. The authors claim that their system has a better precision and recall than the existing algorithms.

New approaches evolved in ranking algorithms with new ideas, but Precision and Recall was used to evaluate the retrieval system. Eastman and Jansen [5] explore the impact of query operators on web search engine results. The authors use coverage, relative precision and ranking as questions trying to answer in their research. Goncalves et al. [15] present an algorithm to measure the effectiveness of a retrieval system as an overall. It measures how much a document is relevant to the query, but it does not compare two retrieval systems. It does not show if a rank of a retrieval system is efficient enough. The authors use Precision and Recall as an evaluation tool. Yuwono et al. [4] explore the relevance feedback in effecting the retrieved documents. The authors use Precision and Recall as a tool to evaluate the ranking efficiency. Hawking et al. [7] tried to answer the question "Can link information result in better PageRank?" The authors discuss the effectiveness of a search engine and its performance by measuring its precision and recall. Yuwono and Lee [3] provide four different ranking algorithms: Boolean Spread Activation, Most-cited, TFxIDF and Vector Spread Activation. The authors use different queries to compare these four algorithms with each other. Their ranking evaluation was based on Precision and Recall.

The hypertext algorithm was a new approach proposed by Brin and Page [22] to improve the ranking of retrieved web pages. The authors claim that this approach would improve the search result by having high precision rank. Baeza-Yates and Davis [16] show that link attribute of a Web Page can improve the ranking by improving the precision of the system. Trotman and O'Keefe [2] use precision to evaluate the ranking algorithm. They depict how a weight is awarded to each document.

Pay per performance (PPP) search engine is a different approach in search engine ranking. Goh and Ang [9] discuss this approach and use precision and recall to evaluate the ranking performance. Ljosland [14] presents a comparison between three search engines: Atavista, Google and Alltheweb. The author uses precision to evaluate the performance of each engine. Bifet and Catillo [1] explore the top web pages appearing in the rank. They also explore the shifted ones. The authors use precision to calculate the efficiency of the rank.

Precision and Recall was used in most ranking evaluation as we saw in previous works. However, many scientists use different evaluation tools. Precision and Recall can evaluate the retrieval system, but they cannot precisely evaluate the efficiency of the rank. Any change in the order of the retrieved documents does not necessarily affect the precision and the recall. This variable (the order of the retrieved documents) cannot be measured using Precision and Recall method.

Clarke and Cormack [6] introduced a new approach toward ranking evaluation. Their work was to evaluate each document and give a specific weight to the document according to all other retrieved documents. They are interested in documents' weight according to other documents. Their method would change the order of the retrieved documents. But it does not evaluate the rank itself. Algorithms for ranking retrieved documents such as these introduced in [22], [12] and [19] were used to rank web pages; however, they still do not measure the ranking algorithm and its efficiency. Kamvar et al. [21] explore many PageRank scheme and provide two algorithms. They present the Adaptive PageRank and the Modified Adaptive PageRank. The authors have not discussed the ranking evaluation in their work. White et al. [18] present an evaluation to encourage user to interact with the search result. They showed how their approach improves the PageRank. However, their paper does not show any tool to numerically evaluate a PageRank.

New evaluation tool other than Precision and Recall were introduced. Losee and Paris [17] oppose the use of Precision and Recall as a measure to evaluate search engine ranking performance. The authors suggest a probability method and proved that their proposed solution would result in a much better evaluation. The authors present the Average Search Length (ASL). ASL finds the average position of the retrieved document. This method is much better than precision in evaluating the ranking performance. However, as the authors mention, a small number of relevant documents in the top of the rank may represent a superior performance. They present the Expected Search Length (ESL) as an alternative approach. This method counts only the non-relevant documents. In this evaluation the system must minimize the ESL value for better performance. The authors [17] advocate our approach in finding an alternative method to measure the performance of a ranking system. Haveliwala [24] compares two different ranking by measuring the degree of similarity. He calculates the degree of overlap between the top URLs of the two ranking lists. Our approach is to find a numerical evaluation for each ranking list rather than comparing the two different ranks.

III. PROPOSED MECHANISM FOR RANKING

Precision and Recall is used to evaluate the efficiency of a retrieval system. A large number of relevant documents and a few irrelevant ones give a high system precision. Precision is calculated according to formula (1). It is the ratio of the relevant documents retrieved to the total number of retrieved documents. The recall of a retrieval system is the ratio of the relevant documents retrieved to the relevant documents in the database of the system. It is infeasible to accurately calculate the number of documents in a database of a search engine. It appears that Precision and Recall has some limitations in calculating ranking efficiency.

$$precision = \frac{\#\text{of r.d.r}}{(\#\text{of r.d.r} + \#\text{of i.d.r})} * 100\% \qquad (1)$$

where r.d.r: relevant documents retrieved
i.d.r: irrelevant documents retrieved

$$recall = \frac{\#\text{of r.d.r}}{\#\text{of r.d.db}} * 100\% \qquad (2)$$

where r.d.r: relevant documents retrieved
r.d.db: relevant documents in database

In all cases, we might have the tool to find the number of relevant documents retrieved, and the number of irrelevant documents retrieved. However, the number of relevant documents that exist in a database cannot be found. Therefore, we are certainly unable to calculate the recall value precisely. The problem with precision is presented in example 1, 2 and 3.

a. www.aa.com
b. www.amazon.com
c. www.book.com
d. www.dell.com
e. www.ebay.com
f. www.google.com
g. www.ibm.com
h. www.miami.edu
i. www.overstock.com
j. www.sony.com

Fig1. The retrieved web pages

Example 1: Suppose we have the retrieved web pages in the order shown in figure 1. Suppose the following web pages are the only relevant documents to the query and the other documents (web pages) are irrelevant.

a. www.aa.com
b. www.amazon.com
e. www.ebay.com
g. www.ibm.com
i. www.overstock.com
j. www.sony.com

Then

$$precision = \frac{\#\text{of r.d.r}}{(\#\text{of r.d.r} + \#\text{of i.d.r})} * 100\%$$

$$precision = \frac{6}{6+4} * 100\% .$$

$$= 60\ \%$$

Example 2: Suppose the following documents are the only relevant documents and the other documents are irrelevant.

b. www.amazon.com
c. www.book.com
e. www.ebay.com
f. www.google.com
i. www.overstock.com
j. www.sony.com

Suppose the search engine ranks the web pages in the order shown in Figure1:

Then

$precision = \frac{6}{6+4} * 100\%$ = 60%.

Example 3: Suppose the following web pages are the only relevant documents and the other documents are irrelevant:

a. www.aa.com
b. www.amazon.com
c. www.book.com
d. www.dell.com
e. www.ebay.com
f. www.google.com

Suppose the search engine ranks the web pages in the order shown in Figure1:
Then

$precision = \frac{6}{6+4} * 100\%$

= 60 %.

With three different sequences, we found that the precision does not change. Using precision tells us that the efficiency of the three retrieval systems is the same in the three systems. However, examples 1, 2 and 3 show that we have three systems. These systems have totally different sequences and they should have different ranking efficiencies.

We propose the *Search Engine Ranking Efficiency Evaluation Tool* (S.E.R.E.E.T.) to distinguish among any ranking systems. The purpose of this algorithm is to numerically evaluate the efficiency of the search engine rank.

Definition

Let there be m≥0 hits and n≥0 misses, (m+n)≥1.

Let i, $1 \le i \le m + n$

represents the position of a website name on the search output list that is hit or missed, and let the position of the j_{th} hit,

$1 \le j \le m$

be given by h_j Obviously $1 <= h_j <= (m+n)$, for all j.

Let the W_i denote the weight of the i_{th}. website name, where

$1 \le j \le m + n$

* Define W_i as follows:
$W_i = m + n + 1 - i$, if the i_{th} name is a hit.
$W_i = 0$, if the i_{th} name is a miss.

Then the efficiency of ranking of search engine is given by formula (3):

$$E = \sum_{i=1}^{i=m+n} Wi * \frac{2}{(m+n)*(m+n+1)} * 100\% \quad (3)$$

Example 4:

Consider there are m=5 hits and n=4 misses in the order shown below:

1. h1
2. m1
3. m2
4. h2
5. h3
6. h4
7. h5
8. m3
9. m4

w1= m+n+1-1= 5+4 =9
w2=0
w3=0
w4= m+n+1-4 =6
w5= m+n+1-5=5
w6= m+n+1-6=4
w7= m+n+1-7=3
w8=0
w9=0

Then

$$E = (\sum_{i=1}^{i=m} Wi) * \frac{2}{(m+n)*(m+n+1)} * 100\%$$

$$= (9+6+5+4+3) * \frac{2}{(5+4)*(5+4+1)} * 100\%$$

$$= 27 * \frac{2}{9*10} * 100\%$$

$$= \frac{54}{90} * 100\%$$

= 60 %

Lemma 1: For integer m > 0,

$$\frac{m*(m+1)}{2}$$

$1+2+...+m$ =

Proof:

Let S(m) = 1+2+ … + m (L1.1)
We can rewrite

S(m) = m+ m-1+ … +1 (L1.2)

Adding both the sides of (L1.1) and (L1.2),

we get 2*S (m) = m * (m+1)

Hence

$$S(m) = \frac{m*(m+1)}{2}$$

Hence we have proved the lemma.

Theorem 1:
If there are no misses, the efficiency of the search engine is 100%.

Proof:
If there are no misses, n=0.
Hence Efficiency (E) :

$$E = (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m+n)*(m+n+1)}] * 100 \ \%$$

$$E = (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m)*(m+1)}] * 100 \ \% \quad (T1.1)$$

Also, $Wi = m+1-i$

Hence from (T1.1),

Efficiency

$$E = (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m)*(m+1)}] * 100 \ \% \quad (T1.1)$$

$$= (m+1-1+m+1-2+...+m+1-m)*[\frac{2}{(m)*(m+1)}]*100\%$$

$$= (m+m-1+m-2+...+1)*[\frac{2}{m*(m+1)}]*100\%$$

$$= \frac{m*(m+1)}{2} * \frac{2}{(m)*(m+1)} * 100\% \quad \text{from Lemma 1.}$$

= 100%.

Theorem2:
Let there be m hits, m≥0, and n misses, n≥0, (m+n)≥1, for two website name lists W_1 and W_2. W_1 is such that it has its hits only positions i , 1 ≤ i ≤ m. W_2 is such that it has one miss in position M, 1≤ M ≤ m, all other hits in position i , 1 ≤ i ≤ m and one hit in position H, m < H ≤ (m + n). Then E1 >E2
where E_1 and E_2 are efficiencies for W_1 and W_2, respectively.

Proof:
We have

$$E_1 = \sum_{i=1}^{i=m} Wi * \frac{2*100}{(m+n)+(m+n+1)}$$

$$= K.[\sum_{\substack{i=1 \\ i \neq M}}^{i=m} W_i + W_M]$$

$$\text{where } K = \frac{2*100}{(m+n)+(m+n+1)}$$

$$\therefore E_1 = K[\sum_{\substack{i=1 \\ i \neq M}}^{i=m} Wi + (m+n+1-M)] \quad (2.1)$$

$$E2 = K * \sum_{\substack{i=1 \\ i \neq M}}^{i=m+n} Wi$$

$$= K[\sum_{\substack{i=1 \\ i \neq H}}^{i=(m+n)} W_i + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i \neq H}}^{i=m} Wi + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i \neq M \\ i \neq H}}^{i=m} Wi + W_M + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i \neq M \\ i \neq H}}^{i=m} Wi + 0 + (m+n+1-H)]$$

$$= K[\sum_{\substack{i=1 \\ i \neq M \\ i \neq H}}^{i=m} Wi + (m+n+1-H)] \quad (2.2)$$

Observe that H > m > M by hypothesis (2.3). Hence comparing (2.1) and (2.2) in view of the fact (2.3),
we have

$E_1 > E_2$

Theorem 3

Let there be m hits, m ≥ 1 and n misses, n ≥ 0, (m+n)≥1 for two website name lists W_1 and W_2. W_1 has all the hits in positions i, $1 \le i \le m$,
and all the misses in positions j, $m \le j \le (m+n)$
W_2 has p misses in positions k , $1 \le k \le m$
and $0 \le p \le m$
and (n-p) hits in positions l , $1 \le l \le m$

Then $E_1 > E_2$.

<u>Proof</u>

We will prove this theorem by induction.

1. By theorem 2, the hypothesis is true for p=1.
2. Assume the hypothesis is true for p = e, e ≤ n

let E_e be the efficiency of this website name list denoted by W_e

then $E_1 > E_e$ (3.1)

let us exchange the positions of one hit at position s, s<m and one miss at t , t>m.

Thus now we have a new website name list W_{e+1} such that it has

e +1 misses in positions K , $1 \le k \le n$

Let E_{e+1} be the efficiency of W_{e+1}

then

$$E_{e+1} = \sum_{i=1}^{m+n} Wi \; * \; \frac{2}{(m+n)+(m+n+1)} * 100$$

$$= \; k \; [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + Ws \; + Wt \;]$$

Where k = $[\frac{2*100}{(m+n)*(m+n+1)}]$

Hence

$$E_{e+1} = \text{K} \; [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + 0 + (m+n+1-t)]$$

$$= \; \text{K} \; [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + Ws \; + Wt \;]$$

Since s<m and t>m , we have s<t (3.3)

Also , E_e = k $[\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + Ws \; + Wt \;]$

$$= \text{k} \; [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + Ws \; + Wt \;]$$

$$= \text{k} \; [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \; + 0 + (m+n+1-s)] \qquad (3.4)$$

Comparing (3.2) and (3.4) and using (3.3), we have

$E_e > E_{e+1}$ (3.5)

From (3.1) we have $E_e > E_{e+1}$

Hence $E_e > E_{e+1}$

Hence the hypothesis is true for p=e+1
Hence by induction, we have proved the theorem.

IV. 4. Discussion and Conclusions

With a close look at many of the evaluation tools used in search engine ranking, we found that those tools do not address the need for ranking evaluation. They only look at document level, but not at the ranking evaluation level. Precision and Recall does not precisely measure ranking efficiency; they rather measure the percentage of good and bad documents in the retrieved bag.

Search Engine Ranking Efficiency Evaluation Tool (S.E.R.E.E.T.) introduced in our earlier work provides a unique tool to precisely measure the ranking efficiency. In this paper we show the mathematical formula behind the SEREET tool. If we have two different ranking algorithms with the same precision, we still can favor one over the other by using SEREET.

REFERENCES

[1]Albert Bifet and Carlos Catillo. *An Analysis of Factors Used in Search Engine Ranking.* Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[2] Andrew Trotman and Richard A. O'Keefe. *Identifying and Ranking Relevant Document Elements*. In INEX '03, 2003.

[3] B. Yuwono and D.L. Lee. *Search and ranking algorithms for locating resources on the World Wide Web*. 12th International Conference on Data Engineering (ICDE'96) p. 164

[4] Budi Yuwono, Savio L. Y. Lam, Jerry H. Ying and Dik L. Lee. *A World Wide Web Resource Discovery System*. Proceedings of the Fourth International World Wide Web Conference, Boston, MA., Dec. 1995.

[5] Caroline M. Eastman and Bernard J. Jansen. Coverage, Relevance and Ranking: *The Impact of Query Operators on Web Search Engine Results.* ACM Transaction on Information Systems, Vol. 21, No. 4, October 2003, Page 383-411.

[6] Charles L. A. Clarke and Gordon V. Cormack. *Shortest-Substring Retrieval and Ranking*. ACM Transactions on Information Systems, Vol 18, No. 1, January 2000, Pages 44-78.

[7] David Hawking, Nick Craswell, Paul Thistlewaite and Donna Harman. *Results and Challenges in Web Search Evaluation*. Computer Networks, 31(11-16): 1321-1330, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[8] Dell Zhang and Yisheng Dong. *An efficient algorithm to rank Web resources*. Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking, Amsterdam, The Netherlands, Pages: 449 – 455, 2000 .

[9] Dion H. Goh and Rebecca P. Ang. *Relevancy Rankings – Pay for Performance Search Engines in* the Hot Seat. Online Information Review, Volume 27, Number 2, 2003 , pp. 87-93(7).

[10] Helen Ashman and Paul Thistlewaite, editors. Proceedings of the Seventh International World Wide Web Conference, volume 30 of Computer Networks and ISDN Systems. The International Journal of Computer and Telecommunications Networking, Amsterdam, April 1998. Elsevier. Brisbane, Australia.

[11] Heting Chu and Marilyn Rosenthal. *Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology.* ASIS 1996 Annual Conference Proceedings, October 1996.

[12] Jon Kleinberg. *Authoritative sources in a hyperlinked environment*. Technical Report RJ 10076, IBM, May 1997.

[13] Miguel Costa and Mario J. Silva. *Optimizing Ranking Calculation in Web Search Engines:* a Case Study, 2004.

[14] Mildrid Ljosland. *Evaluation of Web Search Engine and the Search for Better Ranking Algorithms*. SIGIR99 Workshop on Evaluation of Web Retrieval (1999)

[15] Pedro Goncalves, Jacques Robin, Thiago Santos, Oscar Miranda and Silvio Meira. *Measuring the Effect of Centroid Size on Web Search Precision and Recall*. In Proceedings 8th Annual Conference of the Internet Society (INET'98). Geneva, Switzerland,July,1998. http://www.isoc.org/inet98/proceedings/1x/1x_8.htm.

[16] Ricardo Baeza-Yates and Emilio Davis. *Web Page Ranking using Link Attributes*. In Alt. track papers & posters, WWW Conf., pp. 328-329, New York, NY, USA, 2004.

[17] Robert M. Losee and Lee Anne H. Paris. *Measuring Search Engine Quality and Query Difficulty: Ranking with Target and Freestyle*. Journal of the American Society for Information Science archive Volume 50 , Issue 10 (1999), Pages: 882 – 889, 1999 .

[18] Ryen W. White, Ian Ruthven and Joemon M. Jose. *Finding Relevant Documents using Top Ranking Sentences: An Evaluation of Two Alternative Schemes*. Proceedings of the 25th Annual International ACM SIGIR Conference (SIGIR 2002). Tampere. Pages 57-64. 2002.

[19] S. Lawrence and C. L. Giles, "*Inquirus, the NECI meta search en-gine,*" in Seventh International World Wide Web Conference, (Bris-bane, Australia), pp. 95–105, 1998.

[20] S. M. Shafi and Rafiq A. Rather. *Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology*. Webology , Volume 2, Number 2, August, 2005 .

[21] Sepandar Kamvar, Taher Haveliwala and Gene Golub. *Adaptive Methods for the Computation of PageRank*. Numerical Solution of Markov Chains, 2003, pages 31-44.

[22]Sergey Brin and Lawrence Page, "*The anatomy of a large-scale hypertextual Web search engine,*" Computer Networks and ISDN Systems archive, Volume 30 , Issue 1-7 Pages: 107 – 117, 1998.

[23] Shih-Hao Li and Peter B. Danzig. *Precision and Recall of Ranking Information Filtering Systems*. Journal of Intelligent Information Systems 1-22, Kluwer Academic Publishers, Boston. 1996.

[24]Taher H. Haveliwala. *Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search*. IEEE Transactions on Knowledge and Data Engineering, 15(4):784-796. IEEE, August 2003.

Moiez Tapia is a Professor in the Department of Electrical andComputer Engineering at the University of Miami. His research interests are in the field of Multi-valued logic and calculus, real-time systems, machine learning, fault-tolerant computing.

Wadee Alhalabi, received his MSECE and PhD from the University of Miami in 2004 and 2008. He is now the director of the Virtual Reality Therapy and Rehabilitation Research Center in KSA.

Miroslav Kubat is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Miami. His research interests are in the field of machine learning, data mining and artificial neural networks.