# Reciprocal Rank Using Web Page Popularity

Xenophon Evangelopoulos, Christos Makris, and Yannis Plegas

Computer Engineering and Informatics Department,
University of Patras, Greece
{evangelopo,makri,plegas}@ceid.upatras.gr

**Abstract.** In recent years, predicting user behavior has drawn much attention in the fields of information retrieval. To that extend, many models and even more evaluation metrics have been proposed, aiming at the accurate evaluation of the information retrieval process. Most of the proposed metrics, including the well-known nDCG and ERR, rely on the assumption that the probability (R) a user finds a document relevant, depends only on its relevance grade. In this paper, we employ the assumption that this probability is a function of a combination of two factors; its relevance grade and its popularity grade. Popularity, as we define it from daily page views, can be considered as users' vote for a document, and by combining this factor in the probability R we can capture user behavior more accurately. We present a new evaluation metric called Reciprocal Rank using Webpage Popularity (RRP) which takes into account not only the document's relevance judgment, but also its popularity, and as a result correlates better with click metrics than the other evaluation metrics do.

**Keywords:** Information Retrieval, Evaluation, Metrics, User Behavior, User Model.

## 1 Introduction

Designing evaluation metrics consists an important direction of information retrieval, which has gained tremendous attention over the last few years due to the expeditious evolution of information retrieval systems. Some of the best known evaluation metrics that have been developed over the years are Mean Average Precision (MAP), Precision at k (P@k), normalized Discounted Cumulative Gain (nDCG) [9], Expected Reciprocal Rank (ERR) [2] etc. A good evaluation metric should reflect the rate of relevance of the retrieved results. Furthermore, a proper user model which reflects users' interaction with the retrieval system is of utmost importance.

There are two different types of user behavior models; the *position models* and the *cascade models*. The first [7] assume that a click depends on both relevance and examination. Moreover, the probability of examination depends only on the position. As a result, position models consider each result in a page as

L. Iliadis et al. (Eds.): AIAI 2014 Workshops, IFIP AICT 437, pp. 116–125, 2014.

independent from other results and thus fail to capture interaction among them in examination probability.

On the other side, cascade models rely on the assumption that users examine all the results sequentially from top to bottom and stop as soon as a relevant document is found and clicked. The examination probability in this case depends on the rank of the document and the relevance of all the previous documents. It is showed by Chapelle and Zhang [3] that cascade models can predict click-through rates mode accurately than position models.

In this paper we induce a novel evaluation metric which incorporates a new concept called *web page popularity*. Our metric is a cascade-based evaluation metric, which means that we assume the user scans all results from top to bottom and stops when she/he finds a relevant document and clicks on it. The difference of our proposed metric lies on the definition of the probability $R$, that a user finds a document relevant. Unlike previous editorial metrics we assume that the probability a user clicks on a document depends on two factors; not only on its relevance grade, but also on its popularity grade. For each document $D_i$, we have two values, a relevance grade as it is proposed by experts and popularity grade as it is given by web traffic statistics. We then incorporate these two values in the probability $R$ and evaluate how well our proposed metric captures user behavior by comparing its performance with the performance of click-metrics.

The remainder of this paper is organized as follows. First Section 2 presents some recent related work on the field. Section 3 then describes the cascade model on which our new metric is based. Section 4 explains in detail the notion of web page popularity. Our metric is presented in Section 5 and Section 6 provides evidence of our metric's well-behavior. Finally we conclude the paper in Section 7.

## 2 Related Work

Evaluating the quality of information retrieval systems constitutes an important task in the research area of information retrieval. A great number of scientific papers have been published trying to best model, how well the search system satisfies users' search needs. As a result, a broad range of evaluation metrics which quantify system performance have been proposed, including Discounted Cumulative Gain (DCG or nDCG as it is mostly used) by Järverlin and Kekäläinen [9], Average Precision (AP) and Expected Reciprocal Rank (ERR) by Chapelle et al. [2] just to name a few. The latter belongs to a group of IR metrics which have an underlying cascade user model, which is the user model our proposed evaluation metric also uses.

The basic concept an information retrieval measures, is the concept of relevance. Most evaluation metrics use a set of relevance judgments for a set of documents as firstly induced by TREC evaluations. Relevance judgments are collected by human experts who are obliged to asses for a document's relevance to a given query and thus capture the notion of user relevance [6] [15]. Instead, we argue that an expert's opinion about a document's relevance is not sufficient enough to account for all users' opinion. Thus, we introduce a second relevance

indicator which accounts for a huge amount of real users' opinion. We call this concept web page popularity and is derived from the daily page views a web page gets.

A resultant issue which arises when evaluating information retrieval systems is the test collection incompleteness problem, where a significant amount of relevance judgments is missing from the test collection. Buckley and Voorhees [1], Sakai [14] and Chucklin et al [5] battle this problem either by introducing a new metrics or by alternative solutions. We will also contemplate this problem using the concept of web page popularity.

## 3 Cascade User Model

One of the first things one should take into consideration when developing an evaluation metric is how well it's user model reflects users' satisfaction [8]. There exist two main categories of user models: position models and cascade models. Position models [7] [13], as their name implies, are trying to express the examination probability as a function of the position. Some of the metrics relying on the position model include the DCG metric and the RBP [10] metric. However, as showed by Chapelle et al. [2] position models appear to face some serious drawbacks due to the fact that they assume that the probability of examination depends only on position.

Our proposed metric is based on the cascade model [3], assuming a linear traversal through the ranking, and that the rest of the results below a clicked document are not examined by the user. Let $R_i$ be the probability of examination of the $i$-th document. As soon as the user clicks on a result and is satisfied with it, she/he terminates the search and results below are not examined at all, no matter of what their position is.

As showed by Craswell et al. [7] $R_i$ represents the attractiveness of a result. More specifically it measures the probability of a click on a result which can be interpreted as the relevance of the snippet. We expand this by inducing a novel factor contributing in the measurement of examination probability. This factor is called *web page popularity*. A more detailed explanation and definition of web page popularity follows in section 4. In order to develop an evaluation metric which captures user behavior well, one should not only rely on experts' judgments, but also on other factors which capture user behavior. Popularity, as we define it from daily page views, can be considered as users' vote for a document and as result a value which expresses user's behavior. Here, we make the assumption that at each rank, the document's probability relies not only on its relevance, but also on its popularity. This means that when a user examines all documents from top to bottom, at each rank $r$ she/he is expected to click on a result after examining its snippet $S_r$, which finds relevant **and** its url link $L_r$, which finds popular. As a result we suggest that probability of clicks is highly correlated both with the relevance of the document and its popularity.

$$R_r = P(C_r|S_r, L_r)$$

The $R_i$ values as we will show in the next section can be set as a function of two factors. The relevance grade of document $i$ and the popularity of the url of document $i$. For a given set of $R_i$, the probability of click on the $i$-th document can thus be expressed as:

$$P(C_r = 1) = \prod_{i=1}^{r-1}(1 - R_i)R_r.$$

## 4 Web Page Popularity

In the previous section we presented the cascade user model, where we presented a novel concept named as popularity of a web page. But how can one define the popularity of a link or better of a page? Cho et al. [4] in their study define popularity *V(p,Δt)* of a page $p$ as the number of "visits" or "page views" the page gets within a specific time interval $\Delta$t. Here, we also use this notation but abstract it generally as Web Page Popularity.

**Definition 1 (Web Page Popularity).** We define the popularity *P(p,Δt)* of a web page $p$ as the number of page views ($pv$) that page gets within a specific time interval $\Delta t$.

Following from the above definition we construct a popularity grade in accordance with the relevance grade, so that the popularity of each link can be measured and compared properly. Particularly, given a number of page views $pv$ for a link $u$, we define its popularity grade as follows:

$$p_u = \left\lfloor \frac{\ln pv_u}{5} \right\rfloor \tag{1}$$

Equation (1) was developed based on traffic statistics about every web page. More specifically, pageview values $pv$ ranged from 0 (that is no page views at all) to 500.000.000 (Google's page views) per day. In order to incorporate these large numbers in a metric we decided to get the natural logarithm from each $pv$ value so that we don't lose the amount of information for each website. Moreover we mapped each natural logarithm to a 5-scale ranging from 0 to 4 in order to create a grade-scale similar to that of the relevance grades.

According to equation (1) each website received a 5-scaled grade (unknown, well-known, popular, very popular and famous) according to its daily traffic statistics. Table 1 shows the popularity grade of some websites according to their daily page views.

**Table 1.** Popularity grade for four sample websites

| WebSite | Daily Page views | Popularity Grade |
|---|---|---|
| http://google.com | 584.640.000 | 4 |
| http://wikipedia.com | 30.451.680 | 3 |
| http://ceid.upatras.gr | 11.228 | 1 |
| http://sample-site.wordpress.com | 11 | 0 |

## 5 Proposed Metric

In this section we will introduce our new proposed evaluation metric. This novel metric is based on the cascade model described in section 3. As stated before our proposed metric is described by two factors of relevance. The relevance grade of the document and the popularity grade of the document. Let $g_i$ be the relevance grade and $p_i$ the popularity grade of the $i$-th document. Then the relevance probability can be equally defined by $g_i$ and $p_i$.

$$R_i = \mathcal{R}(g_i, p_i).$$

Here $R_i$ denotes a different value equally defined by two factors and $\mathcal{R}$ consists a mapping function from relevance and popularity grades to probability of relevance. There are many different ways to define $\mathcal{R}$, but here we select it in accordance with the gain function of ERR used in[2]

$$\mathcal{R}(r) = \frac{2^r - 1}{2^{r_{max}}},$$

where

$$r = \frac{g + p}{2}, \qquad r \in \{0, ..., r_{max}\}.$$

This means that the probability of relevance can be defined equally by two factors; its relevance grade and its popularity grade. As we will prove in the next section this is a well-balanced way to define relevance due to the equal contribution of each factor. Thus, when a document is fairly-relevant ($g = 1$), but is very popular ($p = 4$), the probability that a user will click on that result is much higher than if the document was non-popular.

In order to make things more clear, consider the following scenario: a user examines all the results after posing a query in a search machine. She/he discovers two results (regardless of their position) which he considers as very relevant. If the first result is a far more popular web page than the other, then she/he will not examine the latter.

The next step is to define a utility function for our metric. A utility function $\varphi$ takes as argument the position of each document and should satisfy that at the first position it will take the maximum value 1 and as position increases $\varphi$ converges to 0. In other words, $\varphi(1) = 1$ and $\varphi(r) \to 0$ as r goes to $+\infty$. In accordance with ERR metric, we select the special case $\varphi(r) = 1/r$. As a result, our metric can be defined as follows:

$$RRP := \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r.$$

Compared to other cascade-based metrics, RRP relies not only on the documents' relevance judgment but also on the popularity grade which consists a real-time "judgment" of users. Thus, our metric captures user behavior more successfully.

## 6 Evaluation

It is known that the evaluation of a new retrieval metric is not an easy task due to the fact that there is no ground truth to compare with. However, it has been shown [12] that the quality of a retrieval system can be well estimated by click-trough data. A common technique though is comparison with click metrics. That way one can see how well does a metric approximates user behavior and as a result captures user satisfaction. In this section, we try to evaluate our novel metric by computing correlations between click metrics and editorial metrics including our proposed metric.

### 6.1 Data Collection

In order to collect click through data we developed an information retrieval system based on Indri Search Engine (Lemur Project [11]) using TREC Web Track data. We then asked users from Information Retrieval Class of the Computer Engineering and Informatics Department of Patras to perform an informational search on a number of predefined queries by TREC Web Tracks. Particularly, we used 200 queries from TREC Web Tracks 2009 to 2012 and their document results at depth 20 as returned by Indri. Each user was asked to perform search on 200 predefined queries according to a special informational need. Click-through data from each interaction with the system were collected in log files.

We intersected these click through data with relevance judgments as they were defined by TREC Web Tracks on a five-grade scale ($0 \rightarrow 4$). We also intersected click through data with popularity grades which were computed from daily page views as shown in section 4. Finally, each document was graded according to a five-grade scale from 0 to 4.

The final dataset consisted of relevance and popularity grades for each result-document of each query. For some queries of TREC Web Tracks though, all of their results-documents were graded with zero relevance judgments or with zero popularity grades. We removed these queries in order to simulate a search process which reflects reality, ending with click-through data for 167 different queries.

### 6.2 Correlation with Click Metrics

Using the dataset described in previous subsection we computed the pearson correlation between a set of click metrics and editorial metrics. Particularly, the editorial metrics used were:

- **nDCG:** Normalized Discounted Cumulative Gain.
- **AP:** Average Precision, where grades perfect, excellent and good are mapped to *relevant* and the rest to *non-relevant*.
- **ERR:** Expected Reciprocal Rank.
- **RRP:** Our proposed metric.

The click metrics used were:

- **PLC** Precision an Lowest Rank as defined by [2].
- **Max, Mean and Min RR** Maximum, Mean and Minimum Reciprocal Ranks of the clicks.
- **UCTR** Binary variable indicating whether there was a click or not in a session.

We did not include the Search Success (SS) metric induced by Chapelle et al. [2], as it uses relevances not only clicks. We also confirmed the findings of [2], that QCTR has negative or close to zero correlation with all the editorial metrics and skipped it as well. Table 2 shows correlations between all the above mentioned metrics.

From table 2 we conclude that our proposed metric shows higher scores in correlation with click metrics than other editorial metrics. Particularly, *RRP* outperforms position-based metrics such as *nDCG* and *Average Precision*; additionally *RRP* correlates better with click metrics even than the cascade-based metric *ERR*.

As click metrics are concerned, we observe that *reciprocal ranks (max,min and mean)* along with *precision at lowest rank* seem to correlate better with *ERR* and our proposed metric *RRP*. This holds due to the fact that they both use as utility function the reciprocal rank $\frac{1}{k}$. *Precision at lowest rank* uses reciprocal rank of the lowest position and as a result shows higher correlation with *ERR* and *RRP* too.

**Table 2.** Pearson Correlation between editorial and click metrics

| | PLC | MeanRR | MinRR | MaxRR | UCTR |
|---|---|---|---|---|---|
| nDCG | 0.498 | 0.497 | 0.503 | 0.445 | -0.024 |
| AP | 0.402 | 0.417 | 0.395 | 0.396 | -0.004 |
| ERR | 0.528 | 0.512 | 0.517 | 0.459 | 0.064 |
| RRP | 0.559 | 0.554 | 0.588 | 0.472 | 0.041 |

We can see though, that in most cases our proposed metric appears to score significantly better that other editorial metrics except for *UCTR*. This can be explained as *UCTR* does not account for clicks (rather than their absence) and therefore lacks the source of correlation with click metrics. Moreover, overall results of *UCTR* enhance our assertion as it seems to correlate with editorial metrics much lower (around zero) than other click metrics do.

### 6.3 Performance on Incomplete Collection

Search engines are using an enormous amount of data, which is constantly changing. Thus it is difficult to maintain complete relevance judgments even for a small proportion of the web corpus. As a result there exists a set of queries which partly or completely lacks relevance judgments.

There has been an effort to confront this problem in various ways over the years. Buckley and Voorhees [1] in their work propose a robust preference-based

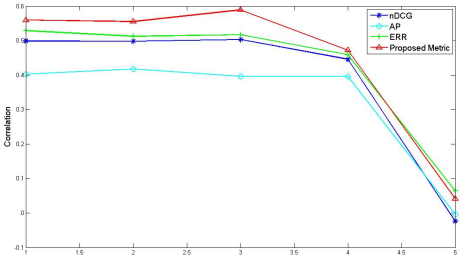


**Fig. 1.** Correlation between Editorial and Click metrics when no queries are "unjudged"

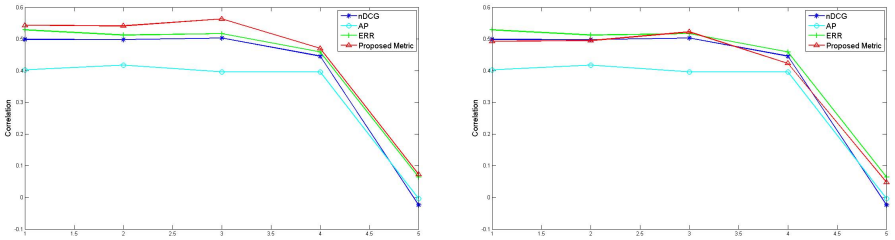


**Fig. 2.** Correlation between Editorial and Click metrics when 41 and 83 queries are "unjudged" respectively

measure called *bpref* which measures the effectiveness of a system on the basis of judged documents only. On the other side, Sakai [14] proposes an alternative solution which does not need a new metric.

Our proposed metric which accommodates two factors of relevance can battle the missing relevance judgment issue in a complete different way. In situations when there is a document with missing a relevance judgment, we employ the second factor, the popularity grade which accounts for user preference. In order to enhance our assertion we conducted experiments on how correlation between click metrics and our proposed metric was affected when some queries of our dataset lacked relevance grades. Particularly, from the initial set of 167 queries, we created 2 sets where in the first set 41 from 167 queries lacked relevance grades and in the second set 83 from 167 queries lacked relevance grades. We then computed the correlation between click and editorial metrics of each partly "unjudged" dataset.

Figures 1, 2 and 3 show the correlation diagram for the initial dataset and the two partly "unjudged" datasets. In figure 2 where a small set of the dataset is "unjudged", our metric retains the highest scores in correlation with click metrics. As we increase the amount of "unjudged" queries (half of the dataset contains no relevance grades), we observe deterioration of our metric's performance. We can easily conclude that in cases when the amount of relevance

judgments absence is not extensive, our proposed metric can still express user behavior better than the other editorial metrics.

## 7 Conclusions and Future Work

In this paper, we proposed a novel information retrieval metric called reciprocal rank with webpage popularity (RRP). This cascade model-based metric incorporates an additional relevance factor while computing the probability that a user finds a document relevant. The second relevance factor we incorporate is called popularity grade and is calculated by the number of daily page views a document gets. The number of daily page views of a web page, can be viewed as a users' vote for this web page and thus, by combining this factor with experts' relevance judgments, our evaluation metric captures user behavior better.

In order to verify our thoughts, we conducted experiments on a TREC dataset with click data collected by students' search sessions. The results showed that our evaluation metric correlates better with click-metrics and as a result expresses user behavior better. Furthermore, we showed that in situations where a significant amount of relevance judgments is unavailable, our metric still correlates better with click-metrics using only popularity grades.

Our plans for the future include the development of a novel user model which will be able to implement the notion of popularity and model accurately its reflect on user behavior. Moreover, new experiments on a dataset with greater number of queries and sessions is another thought in order to enhance the user model's findings.

**Acknowledgments.** This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

## References

1. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pp. 25–32. ACM, New York (2004)
2. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 621–630. ACM, New York (2009)
3. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 1–10. ACM, New York (2009)
4. Cho, J., Roy, S., Adams, R.E.: Page quality: In search of an unbiased web ranking. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD 2005, pp. 551–562. ACM, New York (2005)

5. Chuklin, A., Serdyukov, P., de Rijke, M.: Click model-based information retrieval metrics. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2013, pp. 493–502. ACM, New York (2013)
6. Cyril, W.: Cleverdon. The significance of the cranfield tests on index languages. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1991, pp. 3–12. ACM, New York (1991)
7. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM 2008, pp. 87–94. ACM, New York (2008)
8. Huffman, S.B., Hochster, M.: How well does result relevance predict session satisfaction? In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 567–574. ACM, New York (2007)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
10. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst., 27(1), 2:1–2:27 (2008)
11. University of Massachusetts and Carnegie Mellon University. The lemur project (January 2014), `http://www.lemurproject.org/`
12. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval quality? In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 43–52. ACM, New York (2008)
13. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: Estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 521–530. ACM, New York (2007)
14. Sakai, T.: Alternatives to bpref. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 71–78. ACM, New York (2007)
15. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, pp. 162–169. ACM, New York (2005)