Success Index: Measuring the efficiency of search engines using implicit user feedback

Apostolos Kritikopoulos, Martha Sideri, Iraklis Varlamis

Athens University of Economics and Business, Patision 76, Athens, Greece

apostolos@kritikopoulos.info, sideri@aueb.gr, varlamis@aueb.gr

Abstract

The problem of characterizing web content and evaluating search engine results in terms of relevance to the user's intention has been an important issue in web research. Despite the large number of manually categorized web query datasets, which can be used for testing and tuning ranking algorithms, the evaluation problem remains unsolved because of the size of the web, the diversity of the query space, and the subjective nature of user satisfaction. In this paper we study Success Index (SI), a method for evaluating the quality of ranking algorithms. Success Index takes into account a user's clickthrough data, and provides an evaluation of ranking. Through extensive user blind tests we show that the results of Success Index compare favorably to those of an explicit evaluation. We also review other existing implicit evaluation techniques and summarize the features that can be exploited for evaluating ranking.

Keywords: search engine, metric, evaluation, implicit

1. Introduction

A ranking algorithm prioritizes documents according to their importance and relevance to the user. A ranking algorithm may rank the full set of documents in the repository (this is done off-line) or the results of a query (on-line). Evaluating a ranking algorithm is an important and difficult problem. One approach is to manually tag documents and queries, and compare the results of search with the tagged corpus for precision and recall; the applicability of this approach to web search is limited by the size of the corpus and the diversity of the query space.

A second approach is to evaluate the ranking algorithm by feedback from the user. User feedback can be either implicit or explicit depending on the way user satisfaction is captured. Explicit feedback methods require from the users to grade as many results as possible and evaluate ranking based on the average user grade for all the queries. Implicit feedback methods attempt to capture users' satisfaction without the users being aware of it. In its simplest form, relevance of a document to the query based is deduced depending on whether the user clicked on the document link or not.

Although explicit methods capture user satisfaction more accurately, they are costly and often users do not cooperate.

In this paper, we overview implicit and explicit methods, and provide evidence on the behavior of search engine users against explicit ranking. Among the numerous approaches that argue on the use of implicit and explicit information in the evaluation of ranking results, and on the factors that bias user evaluation, there is not yet a widely accepted metric that captures users' satisfaction from the ranking of the query results. Our work offers a simple, well defined metric based on implicit information. In the following, we present a ranking evaluation method, called Success Index, which is based on the aforementioned implicit feedback metric. Success Index was first used in [Kritikopoulos et. al. (2005)] as a complement to an explicit evaluation measure, in order to support personalization of web search results. Success Index in was tested on two search engines, one for the Greek web and one for weblogs. The results prove that the reliability of our method is comparable to those of explicit methods and that the main bias factors indicated in the bibliography do not have any effect on our metric.

Section 2 presents an overview of related works. Section 3 presents the suggested evaluation method. Sections 4 and 5 illustrate the experimental setup and results. Finally, Section 6 summarizes the outcomes of this work and discusses the next steps.

2. Related Work

Approaches that use explicit feedback [Chirita et al. (2004)], [Kumar et al. (2001)] ask the user to give a positive grade to each result and calculate the average grade for each set of results. Approaches that use implicit feedback assume that users are presented a ranked set of links to the documents accompanied by a title, a URL and a short description and based on this information they decide and click on the most relevant link. The order of clicks, the time spent in each link, the number of clicked documents, the time spent in reading a description etc, are useful feedback information for evaluating the algorithm. Evaluation is performed without the user being aware of it.

Joachims [Joachims et. al. (2005)] introduces techniques based entirely on clickthrough data to learn ranking functions. Click data and eye movement data is recorded, as the user reads the search results and stops over some of them (eye fixation). Explicit feedback (manual results ranking) is also used as a basis for comparison. The results can be summarized in the following; a) users trust the search engine and click on the top ranked results, and b) bad ranking quality forces users to click on bottom ranked results. In [Joachims et. al. (2005)] the significance of the difference between implicit and explicit ranking, was measured by a binomial test. The agreement between the two rankings is presented in Table 1. Eye tracking measures (i.e. pupil dilation and eye saccades) have been also used by researchers [Salogarvi et. al. (2003)] in order to infer relevance judgments.

The idea that a non-clicked link, which is above a clicked one, is judged to be less relevant is introduced in [Joachims (2002)] and used by authors in [Radlinski et. al. (2005)]. Authors evaluate ranking, using clickthrough data, without managing to improve performance (see Table 1). In [Radlinski et. al. (2005)], the authors assume that a search process may contain, query revision and resubmission between any two clicks. They detect query chains for each query and from the clickthrough logs the gain implicit information on the relevance of a page to one or more queries.

Authors in [Agichtein et. al. (2006)], [Agichtein, Zheng (2006)] identify a rich list of implicit behavior features and use them as feedback for improving ranking results. Clickthrough features, page history features and query text features are employed. Authors use Precision at K, Mean Average Precision and other metrics to compare results against various ranking algorithms. Unfortunately, the research does not conclude on a metric that takes into account the whole of implicit information. In [Oztekin et. al. (2003)] a simple method for evaluating ranking methods is introduced. Their metric calculates the average position of user clicks in the results of a search. They identify several bias factors for their metric (total number of links returned, different sets of links) as well as factors that possibly affect their results (automated clicks by robots and crawlers, changes in user behaviour) and suggest solutions. However, an evaluation of the metric is missing.

In our experiments, users are not aware of the ranking algorithm selected in each set of query results, so their behaviour is not biased from the search engine selection. Moreover, crawlers and robots cannot access the engine since a log in procedure is prerequisite, Finally, since the ranking algorithm is chosen randomly in every individual query, the total number of links and the different set of links bias is not very strong, However, it can be easily avoided in a future experiment, if users are asked to evaluate all three ranking algorithms without being aware of which is which.

In [Fox et. al. (2005)] and [Fox (2003)] the clickthrough data was enriched with more behavioral data. Bayesian models are used to measure relevance between implicit and explicit judgments. Dwell time, position, scroll count, and exit type were predictive of individual page judgments.

Table 1. *Comparison of evaluation techniques that use implicit feedback*

Research work	**Input**	**Similarity to explicit**	**Test used**
Joachims(2005)	Eye movement	64 to 80	Binomial
Joachims(2002)	Clicks	Less than 90%	two-tailed binomial sign test
Radlinski(2005)	QueryChains	64 and 80	Binomial
Agichtein(2006), Zheng (2006)	Clicks, Dwell time, Query text	Mean Average Precision: 0,32	two-tailed t- test, using top-k precision, average precision

Fox et. al. (2005)	Dwell time, position, scroll count, and exit type	No evidence is provided	No evidence is provided
Success Index	Clicks	95% 0,79	t-test cosine

The review proves our intuition that implicit feedback is valuable when evaluating ranking results and can significantly improve the ranking quality when properly exploited. The implicit –behavioral- data collected by researchers comprises clickthrough data, query data, page dwell data, eye movement data etc. To our knowledge there is no formula that includes all the above mentioned implicit features. The last row in Table 1 presents the experimental evaluation of our algorithm, which is further explained in sections 4.3 and 5. Our basis for comparison is Average User Satisfaction (AUS), an explicit evaluation measure which is detailed in section 3.

However, all researchers agree on a presentation-bias of users, who tend to click on the higher ranked links. The FairPairs algorithm [Radlinski et. al. (2006)] attempts to avoid the presentation bias, by switching the order in which top ranked results are presented. All existing approaches present implicit features that can be exploited in implicitly evaluating ranking results. They study the factors that affect each feature importance but do not conclude on a metric for measuring the quality of results' ranking. Our work, contributes a measure that captures simple clickthrough data and is comparable in efficiency to those that use explicit information. In the following, we study a ranking evaluation measure (first introduced in [Kritikopoulos et. al. (2005)]) called Success Index, that uses implicit data and favors first clicks against the last ones. Our implicit measure takes into account the position of the clicked result as well as the order of the click in the click-stream. In our evaluation method we compare three different rankings and we expect the user to rank the best results higher, so we do not avoid the bias. As a result we exploit the presentation bias, in order to evaluate ranking results.

3 Success Index

In our experiments we do not assume a priori knowledge neither on the ranking of documents nor on their relevance to every possible query. We rely on implicit and explicit users' judgments in order to define the quality of ranking. Our primary aim is to evaluate user satisfaction for different ranking methods, using uninformed (blind) testing. The results presented to a user query, are ranked by one of the available ranking methods. The method was selected randomly each time. The evaluation was based on the posts selected and the order of selection. Since the users are not aware of the algorithm used for each query we are confident that the tests are totally unbiased. More details on the evaluation method are available at [Kritikopoulos et. al. (2006)]. The typical use case for our search engine is that the user of BlogWave [Kritikopoulos et. al. (2006)] enters a query and chooses between the presented posts.

The post is presented in a new window and the user is called to declare her satisfaction with a vote (a number between 1=not satisfied and 5=extremely satisfied). The user could vote many posts from the result set, although she can visit some posts without voting (we assume that the vote for these cases is 0).

We use the Average User Satisfaction (AUS) an explicit evaluation measure defined as the average of all votes:

$$\text{AUS}=\frac{\sum_{visited\_posts} vote_u}{|visited\_posts|} \qquad \textbf{Equation 1}$$

In order to further enhance the evaluation process we also use an implicit evaluation measure the Success Index (SI) metric which was presented in [Kritikopoulos et. al. (2005)]. The basic advantage of Success Index is that it does not require the user to vote for her satisfaction.

BlogWave records the posts clicked on by the user, and the order in which they are clicked. We then evaluate the user's response using Success Index, a number between 0 and 1:

$$SI = \frac{1}{n}\sum_{t=1}^{n}\frac{n-t+1}{d_t * n} \qquad \textbf{Equation 2}$$

where: n is the total number of the posts selected by the user

dt is the order in the list of the t-th post selected by the user

The SI score rewards the clicking of high items early on. The reverse ranks of the items clicked are weight-averaged, with weights decreasing linearly from 1 down to 1/n with each click. For example, suppose n = 2 and the posts ranked 2 and 10 were clicked. If 2 is clicked first, then the SI score is bigger (27.5%); if it is clicked second, the SI is smaller (17.5%). More controversially, SI penalizes many clicks; for example, the clicking order 2-1-3 has higher score than 1-2-3-4. In the absence of rating (when the user visits the post but does not provide a score) we assign zero score to the post. However, in our experiments we excluded the queries for which we have no user feedback.

Table 2. *Examples of the SI score*

Selection Order	*1*	*2*	*1*	*3*	*5*	*7*	*10*	*3*	*1*	*2*
SI score	*100%*	*42,59%*			*10,10%*			*38,88%*		

4. Experimental setup

4.1 The search engines

For the experiments, we asked our blind testers to use the SpiderWave and BlogWave search services and provide as with as many evaluation data as possible. The former indexes the Greek Web and the latter is its equivalent for weblogs Users were able either to click on a result and go (implicit feedback) or return to the results page and explicitly express their satisfaction, by giving a grade from 0 to 5 to the search result. Of course, users could browse the whole list of the results before clicking on any individual link. In an attempt to eliminate subjective bias, the experiment was double-blind since neither the individual users nor we know in advance the ranking method that is used in every query (the method was selected randomly). The use of double-blind test, allows the comparison of ranking methods against different query sets, which is the case in our experiment and generally in web search engines. This information is stored in the database and is used only for the evaluation of user satisfaction.

4.2 The experiments

Four experiments have been set up for testing SI efficiency, and similarity to explicit evaluation. In all the experiments, search results were presented in groups of ten based on a global ranking (of all pages in the set). The global ranking was computed using three PageRank variations and the SI formula was used to measure user satisfaction from the ranked results,

The first experiment was intended to evaluate Compass Filter [Kritikopoulos et. al. (2005)], a method for reordering the results of a user's search, based on the user's history and profile. More specifically the pages that matched the query criteria were ranked based on the graph formed by them and other URLs that the user has visited in the past. For this experiment we only had implicit evaluation (based on clickthrough data) since the mechanism for collecting explicit information had not been implemented. We use the PageRank algorithm as a basis for providing a global ranking for pages. However, when enough data is available on a user's browsing history, we re-rank the query results using Compass Filter.

In the second experiment we evaluated Wordrank efficiency. Wordrank is a PageRank variation that prioritizes pages with similar content by adding virtual hyperlinks between them and increasing the density of the web graph. It provides a global ranking for the pages of the set which is used when ranking the documents matching a query. Both implicit and explicit (AUS) feedback has been recorder, as the user was able to click on a result and consequently vote for her choice.

The third and fourth experiments were performed on BlogWave, a search engine for weblogs that is build upon a blog ranking algorithm (BlogRank). BlogRank assumes various virtual links in the weblogs graph (e.g. between posts of the same author,

between posts that share links to news articles, use common tags and have adjacent dates of posting) and produces a global ranking for all the weblogs in the set. The two experiments are identical, though in different time periods.

4.3 The results

Although, explicit voting was available in the latter three experiments, it was optional. This gave us an indication on the interest of users to provide explicit feedback. The degree of user satisfaction was measured using both implicit and explicit feedback. The average score for both measures (AUS and SI) and all the experiments is displayed in Table 3. The AUS score is missing from the first experiment, since the explicit scoring mechanism was not implemented on the time of the experiment. In the same experiment the number of queries in which PageRank was used is significantly bigger than this of Compass Filter. The reason for this is that only for a few users we had enough browsing history data in order to apply Compass Filter. The results were very encouraging as far as it concerns the efficiency of our ranking methods compared to PageRank. However, a more important conclusion can be drawn from Table 3: the SI scores (implicit) are analogous to the AUS scores (explicit) for all the experiments.

Table 3. *Evaluation of the different algorithms using implicit and explicit feedback*

Experiment	**Search Engine**	**Date**	**Ranking algorithm**	**Queries submitted**	**User satisfaction metrics**	
					AUS (1 - 5)	**SI Score (average)**
1	SpiderWave	02/2003	PageRank	508	-	48.58%
			Compass Filter	44	-	**57.70%**
2	SpiderWave	10/2005	PageRank	32	2.25	28.11%
			WordRank	35	**3.53**	**58.26%**
3	BlogWave	04/2006	PageRank	16	1.66	15.80%
			Extended PageRank	38	2.45	35.30%
			BlogRank	22	**3.74**	**55.30%**
4	BlogWave	12/2006	PageRank	78	1.87	13.90%
			Extended PageRank	87	2.41	34.80%
			BlogRank	88	**3.67**	**57.60%**

We further analyze this conclusion by comparing the users' evaluation in a per query basis. The methods we employ for the comparison are t-Test and cosine similarity as explained in the following paragraph. Finally, we present some useful information concerning the position of the clicked links. The following graph presents the percentage of URLs among the top-K results which are clicked (or graded). It is obvious that more than 80% of the pages visited and evaluated by the user (implicitly or explicitly) are among the top-20 positions.

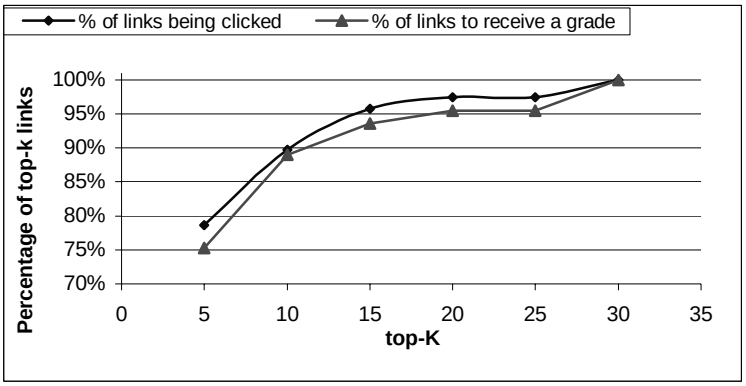

Figure 1. *The position of clicked links*

From the above, it becomes obvious that users tend to visit a result page and then continue their search without stating their satisfaction explicitly. This strengthens the need for a method that implicitly captures user feedback. Moreover, it seems that users rarely click on pages ranked below the top-10 positions [Silverstein 1998] and this sets the lower limit for our metric.

5. Evaluation of our metric – comparison to explicit feedback metrics

But does SI depict the real satisfaction expressed by the user? We are sure that the Average User Satisfaction (AUS) represents the quality of the returned URLs, because it is based on the actual votes made by the user. The SI on the other hand is an automated way to characterize the value of the results presented by a search engine. At the 2nd experiment of Table 3 we performed a t-Test on the 67 queries (32 pagerank and 35 wordrank based) ranked by the user; we compared the Average User Satisfaction (divided by 5 to normalize) and the SI score. In simple terms, the t-test compares the actual difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means). One of the advantages of the t-test is that it can be applied to a relatively small number of cases. It was specifically designed to evaluate statistical differences for small samples, and thus it has been used in many research works in this area as shown in Table 1. Our null hypothesis was that the means are almost similar with the hypothesized mean difference set to 0.1 no matter which one is higher. The p value was 44.71% which is more than 5% and that means that our null hypothesis cannot be rejected and we can safely conclude that the two metrics are significant similar.

For each of the 67 queries we have two different scores, one for SI and one for AUS. Scores may vary across queries, but this is acceptable, if the variation is analogous for the two metrics. A good and simple metric for validating the equivalence between SI and AUS irrespectively of the exact values is the cosine similarity metric. This is given by the following equation:

$$Sim(SI, AUS) = \frac{\sum_{i=1}^{n} SI_i * AUS_i}{\sqrt{\sum_{i=1}^{n} SI_i^2} * \sqrt{\sum_{i=1}^{n} AUS_i^2}}$$ **Equation 3**

As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer and that the similarity of the two metrics increases. The cosine angle is a measure of similarity between Success index scores and AUS scores for the same set of results. The angle itself is an "estimate of closeness". If the two metrics are very alike on their respective sets of results, their corresponding angle should be very small and approaching zero (cosine angle approaching 1). On the other hand, if the angle is high, let say, 90 degrees, the vectors would be perpendicular (orthogonal) and the cosine angle would be 0. For our experiment we had n=67 and Sim(SI,AUS)=0.796 which is very close to 1.

6. Conclusions – Future Work

We have proposed a method for measuring the quality of search engines' results and consequently of a ranking algorithm. This metric relies only on the order of the user's selection on the results. Our experimental results are quite encouraging. We developed and tested our method in the context of two very modest fragments of the Web. This scaled-down experimentation and prototyping may be an interesting methodology for quickly testing information retrieval ideas, and for expanding the realm of research groups, especially academic groups lacking strong industrial contacts, that are in a position to conduct search engine research.

Finally, a very challenging question (for this and many other approaches to Web information retrieval) is to develop a realistic mathematical user model, predicting on the basis of few parameters the user's needs, expectations and behavior. Such a model would help evaluate and optimize novel approaches to personalized information retrieval, and suggest more principled metrics for evaluating a search engine's performance. The next step is to add some implicit parameters and compare the results with SI metric. For example we could add the time difference between the clicks of a user, the grouping of the results (if they were presented in the first, second page etc), the dwell time etc.

References

Agichtein E., Brill E., Dumais S. T. (2006), *Improving Web Search Ranking by* Incorporating User Behavior Information, in the Proceedings of SIGIR .

Agichtein E., Zheng Z. (2006), Identifying "best bet" web search results by mining past user behavior, KDD 2006: 902-908

Chirita, P., Olmedilla, D. and Nejdl. W. PROS: a personalized ranking platform for web search. In Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Eindhoven, Netherlands, 2004.

Fox S. (2003). *Evaluating implicit measures to improve the search experience,* in SIGIR 2003 workshop

Fox S., Karnawat K., Mydland M., Dumais S., White T. (2005), Evaluating implicit measures to improve web search, ACM Transations on Information Systems 23(2):147–168.

Harman, D. (1992). Ranking algorithms. In *information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Upper Saddle River, NJ, 363-392.

Joachims T. (2002), *Optimizing search engines using clickthrough data,* In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). ACM.

Joachims T., Granka L., Pan B., Hembrooke H, Gay G., (2005), *Accurately interpreting clickthrough data as implicit feedback*, Proceedings of the 28th annual international ACM SIGIR conference.

Kelly D., Teevan J. (2003), *Implicit feedback for inferring user preference: a bibliography*, ACM SIGIR Forum.

Kritikopoulos A., Sideri M. (2005) *The Compass Filter: Search Engine Result Personalization Using Web Communities*, Lecture Notes in Computer Science, Springer, Volume 3169, pp. 229 – 240

Kritikopoulos A., Sideri M., Varlamis I. (2006) *BlogRank: ranking weblogs based on connectivity and similarity features,* Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications, Pisa.

Kumar, S., Ertoz, L., Singhal, S., Oztekin, B. U., Han, E. and Kumar V., Personalized profile based search interface with ranked and clustered display. Technical report, University of Minnesota, 2001.

Oztekin U., Karypis G., Kumar V. (2003), *Average Positive of User Clicks as an Automated and Not-Intrusive Way of Evaluating Ranking Methods*, Univ. of Minnesota - Computer Science and Engineering. TR 03-009.

Radlinski F., Joachims T. (2005), *Query chains: Learning to rank from implicit feedback*, ACM Conference on Knowledge Discovery and Data Mining (KDD).

Radlinski F., Joachims T. (2006), *Minimally Invasive Randomization for Collecting Unbiased Preferences from Clickthrough Logs*, Proceedings of the 21st National Conference on Artificial Intelligence (AAAI).

Salogarvi J., Kojo I., Jaana S., Kaski S (2003), *Can relevance be inferred from eye movements in information retrieval?* , In Proceedings of the Workshop on Self-Organizing Maps, pages 261–266.

Silverstein, C., Henzinger, M., Marais, H. and Moricz M,., Analysis of a very large Altavista query log. SRC 1998-014, Digital Systems Research Center, 1998.

SpiderWave, *http://195.251.252.44*