Optimizing parameters of the Expected Reciprocal Rank

Yury Logachev, Pavel Serdyukov
Yandex
Leo Tolstoy st. 16, Moscow, Russia
ylogachev@yandex-team.ru, pavser@yandex-team.ru

ABSTRACT

Most popular IR metrics are parameterized. Usually parameters of these metrics are chosen on the basis of general considerations and not adjusted by experiments with real users. Particularly, the parameters of the Expected Reciprocal Rank measure are the normalized parameters of the DCG metric, and the latter are chosen in an ad-hoc manner. We suggest an approach for adjusting parameters of the ERR metric that allows to reach maximum agreement with the real users behavior. More exactly, we optimized the parameters by maximizing Pearson weighted correlation between ERR and several online click metrics. For each click metric we managed to find the parameters of ERR that result into its higher correlation with the given online click metric.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Metrics, Experimentation, Performance

Keywords

information retrieval measures, evaluation

1. INTRODUCTION

One of the most challenging problems in the field of Web Search is choosing an appropriate metric for learning and evaluating retrieval algorithms. Chapelle et al. suggested the Expected Reciprocal Rank (ERR) metric [1], which received wide recognition in the community. They calculated a correlation of the ERR and other IR measures with online click metrics to prove significance of the ERR metric. Chapelle et al. discovered that correlation of the ERR (and other measures) with click metrics varies for different types of queries (Navigational vs. Non-navigational, queries of various lengths) and different markets. It implies that for different tasks and different purposes different measures (or different parameters of the same measure) should be used. We suppose that one way to set a purpose of an IR system is to choose the target click-metric. For example, if we want our users not to visit a competitor's search engine, then abandonment rate is an adequate click metric. If the focus is on the fast satisfaction of the users, then the position of the first click is an appropriate metric.

The ERR metric has a set of parameters. Each parameter (weight) means the probability of getting completely satisfied after reaching a document with a certain relevance grade. Chapelle et al. suggested a method of setting these parameters using the gain parameters of the DCG metric: $R(g) = \frac{2^g - 1}{2^{g_{max}}}$ where $g \in \{0, \ldots, g_{max}\}$ are the relevance grades. Thereby, commonly used parameters of the ERR metric for a 5-grade scheme with grades *Perfect*, *Excellent*, *Good*, *Fair*, *Bad* are respectively $\approx 0.94, 0.44, 0.19, 0.06, 0$. The same set of parameters (also for a 5-grades scheme) was used at TREC 2010/2011 [2] and de facto became a standard. We argue that these parameters should be adjusted more accurately and depend on the purpose (target click-metric) and market. Thus we suggest a method for optimizing these parameters by maximizing Pearson correlation between ERR and a target online click metric.

2. PARAMETERS OPTIMIZATION

We followed Chapelle et al. and optimized weighted Pearson correlation. Suppose that there are N configurations (a configuration is a query and an ordered set of results). For the i-th configuration, let x_i be the value of ERR metric, y_i the value of the click metric, and n_i the number of times this configuration is present in the data set. Then, the weighted correlation is computed as following:

$$C(x, y, n) = \sum_{i=1}^{N} \frac{n_i(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{N} n_i(x_i - m_x)^2}\sqrt{\sum_{i=1}^{N} n_i(y_i - m_y)^2}},$$

where m_x and m_y are the weighted averages:

$$m_x = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} n_i x_i, m_y = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} n_i y_i.$$

x_i as the value of ERR metric may be considered as a function of five variables: $x_i = x_i(P, E, G, F, B)$, where $params = (P, E, G, F, B)$ corresponds to the weights of the

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Target click metric:	MaxRR	MinRR	MeanRR	UCTR	SS	PLC	NDCG-based
Bad	0.	0.01	0.	0.02	0.	0.0	0.0
Fair	0.21	0.21	0.20	0.10	0.03	0.20	0.06
Good	0.21	0.22	0.20	0.29	0.38	0.21	0.19
Excellent	0.26	0.30	0.28	0.29	0.38	0.27	0.44
Perfect	0.98	0.94	0.97	1.0	0.93	0.97	0.94
C_{opt}	**0.83**	**0.89**	**0.89**	**0.21**	**0.44**	**0.86**	
C_{old}	0.82	0.88	0.86	0.17	0.34	0.84	

Table 1: Results of the experiment. In each column optimal parameters for the given metric are presented. In C_{opt} and C_{old} rows presented values of correlation of target click metric with ERR with optimized and original parameters respectively.

Perfect, *Excellent*, *Good*, *Fair*, *Bad* documents respectively. Thus $C(x, y, n)$ may be considered as $C_{x,y,n}(params)$. So the optimization problem is formulated as following:

$$Q(params) \to \max_{params} \text{subject to } P, E, G, F, B \in [0; 1],$$

where

$$Q = C_{x,y,n}(params) - \sum_{i=0}^{3} a \cdot 10^{k \cdot (params[i+1] - params[i])}.$$

We added an extra summand to the target function Q to encourage the following essential requirement:

$$P \geq E \geq G \geq F \geq B.$$

This requirement follows from the nature of the parameters (P, E, G, F, B). Parameters $a = 100, k = 400$ were selected experimentally to meet the latter requirement.

Any click metric may be used to optimize the correlation with. We examined the same 6 metrics as Chapelle et al.: MinRR, MaxRR, MeanRR, UCTR, SS and PLC [1] (Section 6.2).

2.1 Data collection.

We used query logs of a popular search engine for three months period. Queries generated by search bots were filtered using a proprietary bot filtering algorithm. We followed Chapelle et al. and considered only one-query sessions (60 minutes period was used to delimit sessions), that did not have clicks on additional SERP elements (such as ads). We sampled random 13,755 unique queries and filtered such of them for which click rate on additional SERP elements is higher than 0.10. We were guided by the following reasons: if the CTR on the additional elements is high, then the cases with no such clicks (as mentioned we consider only such queries) are probably outliers. We then asked our judges to assess all result documents (with the common 5-grade system) that were actually shown to the users. As a result we got 10,134 queries (32,239 configurations) in 9,500,687 search sessions.

2.2 Optimization method.

For each configuration from the data set the value of the target click metric y_i was calculated as the average over all the sessions belonging to the given configuration. Next, for each configuration the ERR@10 measure was calculated and stored as the polynomial $x_i(params)$. For example, if the i-th configuration looks like $<$ *Perfect*, *Good*, *Bad* $>$ (we take 3 documents instead of the 10 in this example for the simplicity) then according to the definition of ERR:

$$x_i(params) = P + \frac{1}{2}(1 - P)G + \frac{1}{3}(1 - P)(1 - G)B.$$

Thus each configuration specifies the pair $(x_i(params), y_i)$. That allows to calculate $Q(params)$ as a function of *params*. Finally that function was optimized by the truncated Newton algorithm (SciPy.optimize package [1] was used). The results are presented in Table 1.

3. DISCUSSION AND CONCLUSION

We described a method of tuning parameters of the ERR measure. For each target click metric the correlation of ERR with new parameters is higher. The most noticeable improvement was obtained for the SS (Search Success) click metric. The reason is probably that this metric takes less noisy clicks into account and consequently it is easier to optimize correlation with it.

We demonstrated that for different purposes (i.e. target click metric) there are different optimal parameters of the ERR measure. It is clear that the parameters that we obtained in the experiment are not universal and depend on the market and other specifics of the search engine under study. However, we believe that it is worthwhile to tune them in each case, if online click metrics are assumed to be indicators of search engine user satisfaction. In the future, we plan to experiment with other metrics and markets, different types of queries. Besides, we are going to develop a method to optimize correlation with several click metrics simultaneously.

4. REFERENCES

[1] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. CIKM '09, 2009.

[2] C. L. A. Clarke, N. Craswell, N. Craswell, and G. V. Cormack. Overview of the trec 2010 web track. TREC '10, 2010.

[1] docs.scipy.org/doc/scipy/reference/tutorial/optimize.html