Quality-biased Ranking for Queries with Commercial Intent

Alexander Shishkin
Yandex
16, Leo Tolstoy st
Moscow, Russia
sisoid@yandex-team.ru

Polina Zhinalieva
Yandex
16, Leo Tolstoy st
Moscow, Russia
bondy@yandex-team.ru

Kirill Nikolaev
Yandex
16, Leo Tolstoy st
Moscow, Russia
kvn@yandex-team.ru

ABSTRACT

Modern search engines are good enough to answer popular commercial queries with mainly highly relevant documents. However, our experiments show that users behavior on such relevant commercial sites may differ from one to another web-site with the same relevance label. Thus search engines face the challenge of ranking results that are equally relevant from the perspective of the traditional relevance grading approach. To solve this problem we propose to consider additional facets of relevance, such as trustability, usability, design quality and the quality of service. In order to let a ranking algorithm take these facets in account, we proposed a number of features, capturing the quality of a web page along the proposed dimensions. We aggregated new facets into the single label, commercial relevance, that represents cumulative quality of the site. We extrapolated commercial relevance labels for the entire learning-to-rank dataset and used weighted sum of commercial and topical relevance instead of default relevance labels. For evaluating our method we created new DCG-like metrics and conducted off-line evaluation as well as on-line interleaving experiments demonstrating that a ranking algorithm taking the proposed facets of relevance into account is better aligned with user preferences.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval

Keywords

Learning to Rank, Web Search, Relevance Measures

1. INTRODUCTION

In some information retrieval tasks the only goal of search engine is just to find the most relevant document for a given query. In the case where the set of relevant documents is small, it is likely the best response to the users needs. However, at the present time there exist many groups of user requests, which can be answered by a search engine with a large number of highly relevant results. In contrast to searching for a single correct answer, such requests involve a number of possible answers to choose from. These groups of user requests include, in particular, commercial queries, where customers often want to chose the best offer between many similar ones.

Commercial queries form highly competitive environment where position rise in search results means growth of site's incomes. Thus, in order to increase profits commercial web sites do their best to take the place in the top of search results. Webmasters optimize textual content and buy incoming links to make it easier for a search engine to find their sites and present them to the users in response to the commercial queries. As a result, in terms of textual relevance and link-based quality measures, commercial sites in the top-10 are often equally relevant. Moreover, the human judges will probably also give high relevance assessments to most competing sites, as their evaluation instructions are focused primarily on the *topical relevance* of the document[1]. Thus, any change in positions of sites in the top-10 search results will not lead to a change in ranking quality metrics.

However, giving all the power to search engine optimization, many webmasters forget about user-oriented optimization [13]. Therefore, user satisfaction with commercial web sites can vary significantly. In particular, the design, the presence of on-line feedback mechanism, user reviews of the offered products have noticeable influence on the user experience.

These observations suggest that the use of information of the sites quality in the ranking for commercial queries, which assume a lot of highly relevant answers, can significantly improve the ranking and increase user satisfaction. The quality of a document upon a commercial query, provided that the document is topically relevant, is called *commercial relevance*.

There are studies in which the authors suggest approaches to assessing the quality of the site and its integration in the ranking algorithm. For example, criteria for web page quality in terms of user behavior were described on the basis of interviews data in many papers [1, 2, 11]. Formal criteria that characterize the user-friendliness, trust, design, etc. should be constructed on the basis of the importance of certain aspects of the site quality for users. Such criteria may include the length of the text, content literacy, page titles readability, availability of maps, information about the company, easy to remember phone numbers, free shipping [7].

There are some works that suggest approaches to the use of additional knowledge about the quality of the site in the

[1]http://plg.uwaterloo.ca/~trecweb/2012.html

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

rankings [3]. For example, the aggregation of estimates from several sources, such as human judgments and click data [14] or text relevance and the time of publication [6].

In our paper we propose a new approach to quality-biased ranking which includes creation of new facets of relevance and implementation of a number of features, capturing the quality of a web page along the proposed dimensions. On the basis of several quality facets we form a cumulative rating, which is called commercial relevance. In contrast to [14] we extrapolate commercial relevance labels to the whole learning-to-rank dataset. For the topically relevant search results we define the unified relevance label as the weighted sum of topical and commercial relevance scores. Our approach allows to significantly improve off-line as well as online metrics comparing to the default ranking algorithm.

The rest of this paper is organized as follows. In Section 2 we present new relevance scale that helps us to evaluate commercial sites quality. Section 3 is devoted to our method of learning to rank with respect to the additional document quality measure. In Section 4 we describe new ranking factors, which are used for adjusting commercial relevance. In Section 5 new metrics for the method evaluation are described and finally in Section 6 our results and future work are discussed.

2. COMMERCIAL RELEVANCE SCALE

For the site quality evaluation on queries with commercial intent one can chose either human judgments or clicktrough data [9]. We have decided to use assessors' quality labels because they represent a less noisy data comparing to clickthrough or toolbar data [14].

When using clickthrough or toolbar data, it is very difficult to determine whether user is satisfied with the search result. User behavior on queries with commercial intent may vary significantly depending on product category, its price, etc. (compare pizza delivery and buying a digital camera lens). On the other hand, toolbar and clickthrough data can bring some useful information, so we made this data available to human judges during the assessment process.

In the case of a single quality label different assessors can pay their attention to the various quality aspects. Someone probably knows the site and that it can be trusted in spite of ugly design and poor usability. The other assessor will pay special attention to the presence of user reviews and so on. For the purposes of better formalization of the assessment process, we divided site quality label into several components. At the same time it provides better coverage of site features by human judgments.

Based on data from multiple studies [5, 11, 12], we defined an extended list of commercial relevance facets. Then, in order to facilitate the assessment process, we selected four quality measures, which we believe cover the most of independent quality information. It means that the site quality, defined by these measures encompasses a plenty of site features. The list of the selected site quality measures is as follows: trustability, usability, design quality and the quality of service.

We have elaborated detailed instructions of the site quality estimation for assessors. According to these instructions, the assessment consists of two stages. First, assessor should determine whether document is topically relevant for a given query. We use widespread 5-grade topical relevance scale, which include irrelevant, relevant, highly relevant, useful and vital labels.

Assessment of the site quality is much more complicated and time-consuming process than topical relevance assessment (and especially than obtaining quality information from clickthrough data). Partly it is compensated by the fact that in our method documents quality should be assessed only for the relevant ones.

We do not consider documents with the useful and vital labels assuming that they often are the only goal of a search task for a given query. As mentioned before, we focus only on the queries, that involve a choice between equally suitable results.

On the first stage of the assessment process a variety of products and services provided by the document for a given commercial query is also estimated. We distinguish three grades of the assortment variety: small, standard and large. The variety score for the query q and document d is denoted by $V(q, d)$.

During the second stage of assessment, trustability, usability, design quality and the quality of service for the whole site under review are determined. Trustability and quality of service have four degrees in our scale: spam, normal, good, and perfect.

The site will be labeled as spam if it does not allow to make a purchase or get a desired service (it is a fake site). The sites with *normal* label are not bad but do not differ from the thousands of similar commercial sites. *Good* sites provide users with a standard set of services and finally *perfect* sites are well-known market leaders. The trustability and quality of service scores for a given site s are denoted by $T(s)$ and $S(s)$ respectively. Note that these scores do not depend on the specific pair of query q and document d.

Usability and design quality have only three degrees of quality: bad, good, and perfect. Scores for these commercial relevance facets are denoted by $U(s)$ for the usability and $D(s)$ for the design quality. The values of all the above scores are from 0 to 1.

For the future use of quality information during learning to rank we aggregated four-dimensional label into one single commercial relevance score. Particularly, we have used the following expression:

$$R^c(q, d, s) = V(q, d) \cdot (2T(s) + U(s) + D(s) + 2S(s)), \quad (1)$$

where $R^c(q, d, s)$ is the commercial relevance score for given query q and document d from the site s.

The weights of the trustability and quality of service are twice as much as weights of other site quality facets. It is done for the reason that we believe that these properties are more important in terms of user satisfaction, but we do not consider this choice of parameters as the only possible.

3. LEARNING TO RANK WITH NEW LABELS

Commercial relevance assessment is a very difficult task, so at a fixed cost the number of commercial relevance labels will be much less than the number of topical relevance labels. We can not discard those topical relevance labels, which do not have corresponding commercial relevance estimates. This could result into significant reduction of the

size of the learning to rank dataset and, as a consequence, in degradation of the quality of a ranking function.

Thus, before starting the learning to rank process, we should extrapolate commercial relevance labels to the entire learning to rank dataset. This extrapolation procedure consists of two steps. First, we train a ranking function on the small dataset, which contains only commercial relevance labels. The resulting ranking function gives us an estimated value of commercial relevance score $R^c(q, d, s)$, which is denoted by $R^c_{est}(q, d, s)$.

Then we apply the ranking function from the first step to the complete dataset with topical relevance labels. It is possible because we use the same set of ranking features for both datasets. Since the only highly relevant documents will get a commercial relevance label, estimates of these labels are also calculated only for query-document pairs that have highly relevant labels on the topical relevance scale. Other query-document pairs in the learning to rank dataset will get zero commercial relevance score.

Having estimates for commercial relevance scores of all topically relevant results for queries with commercial intent in our dataset, we calculate the *unified relevance* score:

$$R^u(q, d, s) = R^f(q, d) + \alpha \cdot R^c_{est}(q, d, s), \qquad (2)$$

where $R^f(q, d)$ is the topical relevance score, $R^u(q, d, s)$ is the unified relevance score and α is a weighting coefficient.

Using this unified relevance score, we train the ranking function on the whole dataset. Weighting coefficient α is selected empirically in a such way that it maximizes the impact of commercial relevance, but still does not affect all topical relevance metrics. Finally, we obtain a ranking function that predicts the unified relevance score, which in turn implicitly includes both topical and commercial relevance scores.

4. FEATURES FOR MEASURING SITE QUALITY

For better prediction of new relevance labels, which include both topical and commercial relevance, we introduce some new features specific to commercial sites. They are new in the sense that they are nearly useless for ranking in terms of topical relevance, because topical relevance labels do not carry any information about commercial quality. But for approximating new commercial relevance these features are very helpful since they capture information about the quality of a web page.

From numerous studies on this topic [3, 11, 12] we had chosen some promising features and then supplemented them with our own features. Table 1 provides a list of some quality features used in our research. Note that most of these features are domain features that aggregate information from all documents of the commercial site. This agrees with the fact that, according to Equation 1, commercial relevance depends on the whole site quality.

Comparison of learning to rank with and without described quality features is given in the Results section.

Table 1: Features for measuring site quality.

Detailed contact information
Company's pages in social networks
Absence of advertising
Number of different product items
Verbosity of products description
Availability of shipping service
Salesclerk service (email, phone, customer feedback)
On-line consulting system
Price discounts
Readability of domain name
Average URL length
Average page title length
Consistency of page title and page content
Average depth of the URL path

5. NEW METRICS FOR THE METHOD EVALUATION

For evaluation of our results we developed two NDCG-like metrics [8] based on human judgments about commercial sites quality. First metric presents a weighted quality of the search results for a given set of commercial queries. Its value for one query q is expressed as

$$Goodness(q) = \sum_{i=1}^{10} \frac{R^c(q, d_i, s_i)}{\log_2(i+1)}, \qquad (3)$$

where $R^c(q, d_i, s_i)$ is the commercial relevance for i-th search engine result for the query q. The total value of this metric for a given set of queries is just the average value of $Goodness(q)$ among all queries in this set. The bigger this metric is, the better search engine results are.

Our second off-line metric represents the ratio of low quality search engine results for commercial queries. Similar to the first metric it is calculated for the given set of queries as an average of query-dependent values among all this set. Expression for the query-dependent value now has the form

$$Badness(q) = \sum_{i=1}^{10} \frac{(R^c(q, d_i, s_i) \le th)}{\log_2(i+1)}, \qquad (4)$$

where th is the threshold for the minimum acceptable commercial relevance value for search engine results. The smaller this metric is, the better search engine results are.

Also we use well-known A/B testing [10] and interleaving [4] on-line experiments for evaluating our results. We pay special attention to the Abandonment Rate and Clicks per Query metrics, calculated only for clicks with long dwell-time. We believe that these metrics are most valuable for queries with commercial intent.

6. RESULTS AND DISCUSSION

We have proposed new measure of document quality for commercial queries - commercial relevance. We have developed several ranking features for measuring site quality. In contrast to [14], we proposed a method of extrapolating additional relevance labels for the entire learning-to-rank dataset, which allowed us not to lose any topical relevance information during learning.

We have developed off-line DCG-like metrics and monitored their changes during the experiment with incorporating quality information into the ranking function. Figure 1 shows the variation of our Goodness metric for some time before and after the modification of the ranking function.

The horizontal axis represents the time value and the vertical axis represents the relative value of our metric.

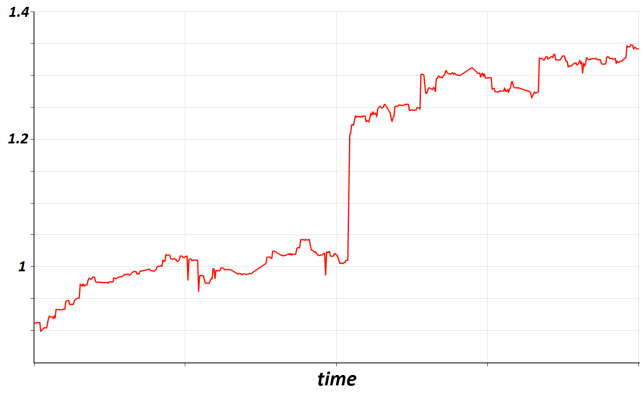

Figure 1: Goodness metric increase during the experiment.

It can be seen that this metric increased almost by 30% comparing to the initial state. Figure 2 represents the variation of our second off-line metric - Badness of search engine results. Again, the horizontal axis represents the time value and the vertical axis represents the relative value of our metric.

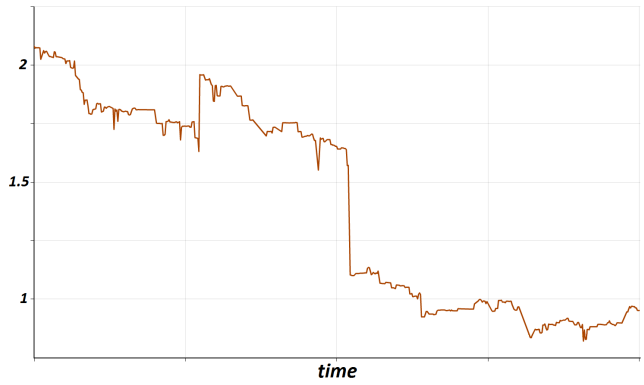

Figure 2: Badness metric decrease during the experiment.

It can be seen that Badness metric decreased almost by 70%. At the same time, classic NDCG metrics calculated using only topical relevance labels remained nearly unchanged during the experiment.

We compared our results with the learning to rank without introducing new commercial features. We observed that improvement in both Goodness and Badness metrics was almost 20% smaller than in the case where all new features were used.

Our on-line interleaving experiment showed that users chose new ranking results 1% more often than results from default ranking system. In the A/B experiment our quality-biased ranking approach demonstrated 5%-decrease in the Abandonment Rate and the Clicks per Query metric increased by 1,5%.

Future work includes using a number of relevance labels instead of single aggregated label in the process of learning to rank. Another approach to further improvement of commercial search results quality is the development of new commercial ranking features.

7. ACKNOWLEDGMENTS

The authors would like to thank Pavel Serdyukov for helpful discussions.

8. REFERENCES

[1] A. B. Albuquerque and A. D. Belchior. E-commerce websites: a qualitative evaluation. In *WWW 2002 Poster Session*, May 2002.

[2] P. Alpar. Satisfaction with a web site. *Electronic Business Engineering*, 4, 1999.

[3] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. WSDM, February 2011.

[4] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems*, 30(1), February 2012.

[5] V. Davidaviciene and J. Tolvaisas. Measuring quality of e-commerce web sites: Case of lithuania. *Ekonomika ir Vadyba*, 16, 2011.

[6] A. Dong and R. Z. et al. Time is of the essence: Improving recency ranking using twitter data. In *WWW 2010 Proceedings*, pages 331–340, April 2010.

[7] M. Ivory, R. Sinha, and M. Hearst. Empirically validated web page design metrics. In *ACM CHI*, April 2001.

[8] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20:422–446, 2002.

[9] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD'02 Proceedings*, 2002.

[10] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, 2009.

[11] G. L. Lohse and P. Spiller. Quantifying the effect of user interface design features on cyberstore traffic and sales. In *CHI 98 Conference Proceedings*, pages 211–218, 1998.

[12] M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[13] K. Nikolaev, E. Zudina, and A. Gorshkov. Combining anchor text categorization and graph analysis for paid link detection. In *WWW 2009 Poster Session*, April 2009.

[14] K. Svore, M. Volkovs, and C. Burges. Learning to rank with multiple objective functions. In *WWW 2011 Proceedings*, pages 367–376, March 2011.