

Recency Ranking by Diversification of Result Set

Andrey Styskin, Fedor Romanenko, Fedor Vorobyev, Pavel Serdyukov
Search Quality Group, Yandex
119021, Leo Tolstoy 16
Moscow, Russia
{styskin, fedor, melton, pavser}@yandex-team.ru

ABSTRACT

In this paper, we propose a web search retrieval approach which automatically detects recency sensitive queries and increases the freshness of the ordinary document ranking by a degree proportional to the probability of the need in recent content. We propose to solve the recency ranking problem by using result diversification principles and deal with the query's non-topical ambiguity appearing when the need in recent content can be detected only with uncertainty. Our offline and online experiments with millions of queries from real search engine users demonstrate the significant increase in satisfaction of users presented with a search result generated by our approach.

Categories and Subject Descriptors:

H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval.

General Terms:

Algorithms, Measurement, Performance, Experimentation.

Keywords:

Recency ranking, diversity, web search.

1. INTRODUCTION

Modern web search engines face the need to consider different non-topical facets of relevance when ranking web documents in response to user queries. While the need in a certain document feature, besides the topical relevance, might be expressed only implicitly in the query, it is still important to recognize its presence in order to adequately satisfy the underlying information need. However, the quality of such recognition cannot be perfect in all cases. Consequently, it brings a certain level of non-topical ambiguity to the queries, which must be taken into account when generating a search result.

One of the most popular non-topical facets of relevance is document freshness. In this paper, we demonstrate how to deal with query ambiguity surrounding the need for recent information. Such ambiguity typically appears during

short periods of time, when users become increasingly interested in one newsworthy aspect, typically an event, related to a well-known entity: a person (e.g. [Michael Jackson]) or a location (e.g. [Japan]). For example, consider the query [michael jackson]. There has been constant and continuing interest in the biography and discography of Michael Jackson, which can be satisfied even by non-fresh documents, but users issuing the query [michael jackson] on (June 25, 2009) were often looking for the freshest news related to the Michael Jackson's death. But both intents (news and discography) continued even during this period.

Recently, a number of techniques for search result diversification has been proposed in order to compensate for topical ambiguity of queries and increase the chance to satisfy the user. We propose to follow the same principles to deal with recency sensitive queries and their non-topical ambiguity. Our approach aims to maximize the probability that the average user finds some useful information among the search results on recency sensitive queries by blending necessary amount of recent results into the result set. To the best of our knowledge, our paper describes the first attempt to tackle the problem of non-topical query ambiguity with a result set diversification technique.

The main contributions of this paper include: 1) the approach to recency ranking by means of search result set diversification, 2) a thorough offline and online evaluation of the proposed approach in terms of a search result quality metric and the overall user satisfaction.

The remainder of the paper is organized as follows. We first review the related work in Section 2. Section 3 describes our machine learning based approach to obtaining a smooth probability of the need in recent content for a query. Section 4 explains how we utilize that probability to diversify ordinary document rankings with fresh documents. Section 5 presents the results of evaluation and Section 6 concludes the paper and outlines the research questions left for future work.

2. RELATED WORK

There are only a few papers focused exclusively on recency ranking. The pioneering work in this area proposed to learn a ranking function which is trained using a subset of features that help to infer the recency of page content [3]. The follow-up work extended that subset to include features extracted from the micro-blogging data stream [4]. Our approach differs in a number of aspects. For one, we try to deal with temporal ambiguity of queries and balance the number of fresh and ordinary relevant documents in the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...\$10.00.

result set, based on the smooth probability of the need in recent content. As a result, while the aforementioned works focus entirely on improving the ranking for one specific case (breaking news queries, 1-2% of search engine's traffic), we aim to affect the ordinary ranking by explicitly increasing its freshness for any query with non-zero probability of the need in recent content. The inference of query's recency sensitivity plays an important role in recency ranking. The aforementioned works detected only highly recency sensitive queries using a linear combination of a few features. In a similar way, Arguello et. al. [5] used a number of features to find verticals relevant for a query, including the News vertical. While those methods focused on binary classification of a query, our work is rather based on regression to obtain and utilize the precise estimate of the probability of the need in recent content.

Our work is also largely based on the principles of search result diversification. In [2], a framework for evaluation that systematically rewards diversity was proposed. In [1], a systematic approach to diversifying results that aims to minimize the risk of dissatisfaction of the average user was presented. A number of follow-up papers were published recently which we do not cover here due to space constraints. Our work complements the research in this area by demonstrating that diversification principles and algorithms are also helpful to increase the chance to satisfy the user in the presence of non-topical query ambiguity.

3. RECENCY SENSITIVE QUERY CLASSIFICATION

In order to quantify the ambiguity of a recency sensitive query, we learned a regression model which predicted the level of interest in recent documents for a particular query and a particular time slot. We used around 30 different features (including their minor modifications) previously described in the papers dealing with similar problems (see Section 2). We do not provide a thorough analysis of feature importance due to space constraints. The most valuable features were the probabilities of queries to be generated by language models of recent content from different sources, including the query, social and news data streams, as well as the probability of a click on a news item.

To train the regression model we asked annotators to provide labels of recency sensitivity for a set of queries. In order to preselect a list of candidate queries for assessment, we defined a small threshold on each feature used to learn our regression function and filtered out all queries that did not have at least one feature value exceeding the corresponding threshold. As a result, we collected judgments for a set of 4000 unique queries issued to Yandex (www.yandex.ru) web search engine (the major russian search engine) over the period of three weeks. On each day during this period, judges were presented with the queries submitted by search engine users on that particular day and were asked to determine whether these queries express an interest in upcoming or ongoing events for which web search users would prefer recent content. Labeling queries basically represented the manual assignment of the probabilities that a particular query is recency sensitive. So, if a query was strictly about a recent event it received the probability of 0.95 (e.g. [flood in thailand] on the day of the event). If a query's primary interest was related to a recent event, but many users would also like

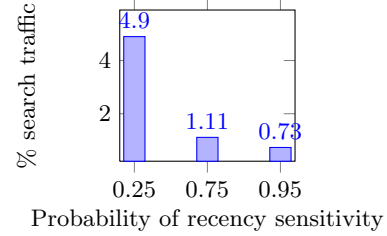


Figure 1: Recency sensitive queries traffic coverage

to see just topically relevant results, it was labeled with 0.75 (e.g. the query [oscar] on the day of the ceremony). If the query's primary interest was not likely to be focused on a particular event, but there was some chance that users issuing such a query would look for some fresh content, assessors assigned the probability of 0.25 (e.g. it always makes sense to present users with some recent content in response to the query [britney spears]). Otherwise, a query was assigned zero probability to be recency sensitive. Each query was labeled by 3 assessors. Average Cohen's kappa coefficient between all pairs of assessors was 0.76, which is considered a substantial agreement.

We learned the regression model to assign smooth probability of the need in fresh content to any query using Gradient Boosted Regression Trees (GDBT) [8]. Recency sensitive queries traffic coverage by these types based on 4000 human made judgments is illustrated on Figure 1.

4. DIVERSIFICATION OF THE SEARCH RESULT WITH FRESH DOCUMENTS

4.1 Diversification method

To produce a search result for recency sensitive queries we follow the search result set diversification principles. Namely, we aim to maximize the utility of the diversified search result expressed in terms of the Expected Reciprocal Rank measure, which we extended to include an abandonment probability and to handle multiple query intents. Both extensions are proposed by Chapelle et. al. [9] and we combine both of them in this work. We call this metric Intent Aware Expected Reciprocal Rank with Abandonment (ERR-IAA) and regard as the objective we aim to maximize:

$$\text{ERR-IAA} = \sum_{i=1}^r p\text{Break}^r * \sum_t P(t|q) * \prod_{i=1}^{r-1} (1 - R_i^t) * R_r^t, \quad (1)$$

where $P(t|q)$ is taken from the distribution over two classes of information needs t (the need in fresh topically relevant documents and the need in any topically relevant documents) for the given query q . Each document is assigned the probability R_r^t to satisfy the information need of type t at position r . We take into account the probability that the documents at the previous $r - 1$ positions have not satisfied that need. We also assume that any user always may stop (abandon the search result) at rank r with the abandonment probability $p\text{Break}^r$ due to accumulated frustration ($p\text{Break}$ is empirically set to 0.85 in our experiments).

We assume that the optimal search result page is the one which maximizes the ERR-IAA measure. In order to maxi-

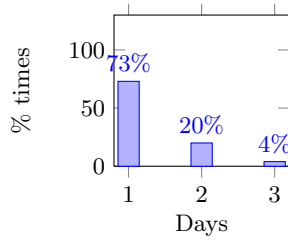


Figure 2: Cumulative share of query instances submitted since the day of the first query

mize it, we follow the greedy approach described by Agrawal et. al. [1] and select the document, whose selection leads to the maximum increase of the objective at each step of the selection process.

4.2 Aggregation of ordinary and fresh results

We also assume that any web document is fresh only for 3 days since the time of its creation or the last update. We use a proprietary algorithm to extract the correct and the most relevant timestamp from document content. Our choice of the number of days is motivated by the following observations. First, according to the studies by Dong et. al. [3], assessors, who judged the freshness of a set of documents, found out that the 1–4 days old documents are the ones most likely containing fresh content. Second, the peak of interest in new events lasts for three days in average according to our analysis of 100,000 spiky and long-tail (so, previously unseen) queries submitted by users of Yandex search engine in January, 2011. Figure 2 shows average share of total query volume for each such a query per for each day since the 1st day they become first known to the search engine until the 3rd day when their popularity falls off almost completely.

Guided by our definition of a fresh document, we produce the ranking of fresh topically relevant documents by simply removing the outdated documents from the initial ranking. As a result, we have two document rankings which we use to maximize the ERR-IAA measure, described by Equation 1: the one containing any topically relevant documents and the one containing only fresh topically relevant documents.

However, in order to proceed with maximization of our objective, we still needed to determine the probabilities of relevance R_i^f . In order to be independent of specific retrieval scores, which may significantly vary over queries and bear a relative, rather than absolute meaning, we turn document ranks into their probabilities of relevance using the internal search engine’s statistics about the probability to encounter a relevant document at the specific position. So, since we fixed the probabilities of document relevance for each of two aggregated rankings, the final ordering of documents depended only on the probabilistic output of our classifier of query’s recency sensitivity.

5. EXPERIMENTS

5.1 Offline results

The research questions we aim to answer in this subsection are how our search quality objective (ERR-IAA) changes while we aggregate two result sets: ordinary and fresh, and

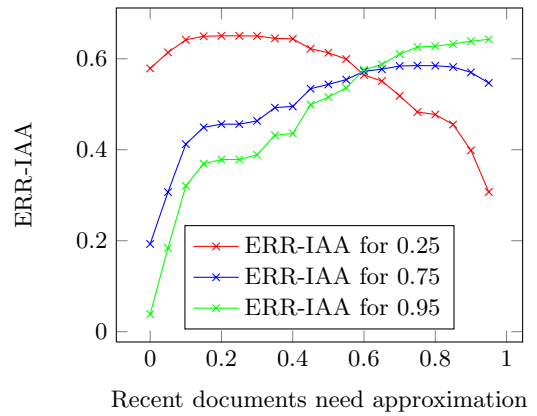


Figure 3: ERR-IAA for queries with different “true” probabilities of need in fresh content

how the quality of our classifier of query’s recency sensitivity affects the quality of such aggregation.

Figure 3 demonstrates how ERR-IAA changes as the estimation of the probability of the need in recent documents deviates from its true value for the three different true values assigned by our assessors: 0.25, 0.75 and 0.95. As we see, while minor errors in the probability estimation do not significantly affect the quality of the aggregated ranking, it is evidently important to keep the errors low as the ranking quality drops quite rapidly with their increase.

Further, we analyze the quality of the aggregation for the recency sensitivity classifier that we use in this work (Figure 4). We split our queries judged by their recency sensitivity (see Section 3) into two parts (training and test, 2000 queries each) and conduct the evaluation via two-fold cross-validation. We train our classifier on the training set of queries and evaluate how its accuracy affects the quality of the aggregation on the test set. Note, that in a real setting that we simulate in this experiment, the initial ranking naturally contains some number of fresh documents. As a result, ERR-IAA measured on the ordinary ranking starts to grow as the recency need probability approaches 1.0, since the queries with such high probability of the need in recent content are typically the queries that are unambiguous: highly descriptive and possessing enough discriminating power to retrieve very relevant content. For example, for the query [europe alert icelandic ash cloud], both the ordinary result set and the fresh result set are quite similar on the day of the infamous volcano eruption. The major gain from applying the diversification comes for the queries with probabilities of the need in recent content from 0.3 to 0.8. This is to be expected, as our approach to recency ranking focuses on the cases of temporal query ambiguity, in contrast to the previous approaches, which aim to maximize the quality of ranking for the queries with no temporal ambiguity (see [3, 4] for more details).

5.2 Online results

In order to test our approach in terms of web search engine metrics measuring user satisfaction, we conducted an A/B test [7]. Some users of Yandex search engine were always presented with ordinary search results (control bucket), which were never diversified with fresh documents, and some

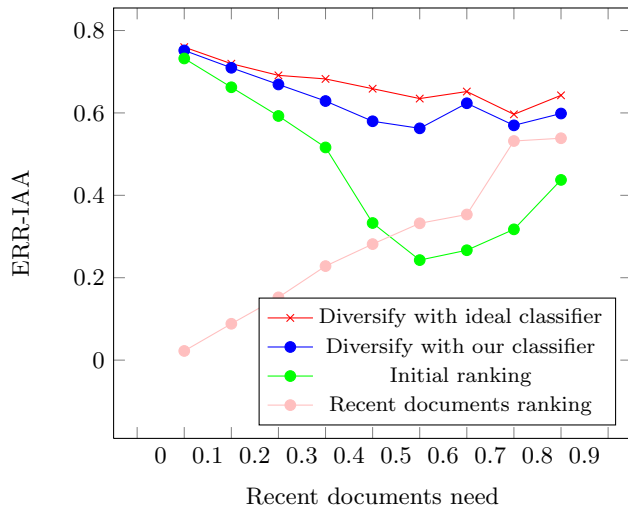


Figure 4: ERR-IAA for queries with different level of ambiguity, diversified and non-diversified rankings

users were presented with diversified search results (treatment bucket). We ran the experiment for 13 days in March 2011 and that involved about 10 million queries in each bucket (issued by real users as we filtered out bots and spammers). We measured user satisfaction using metrics suggested by Radlinski et. al. [10], as they are known to correlate with search result quality. Final results for the control and the treatment are listed in Table 1.

Table 1: User behavior metrics for the control and the treatment buckets

Metric	Meaning	Contr	Treat
Abandonment Rate	% of queries with no results clicked	33.65	32.77
Time to first click	Time to first click on any result (in sec)	10.95	10.76
1st CTR	% of queries with 1st position clicked	44.22	45.31
2nd CTR	% of queries with 2nd position clicked	14.88	14.92
First Click Position	Position of first click	1.91	1.87

All metrics in the treatment are significantly different from metrics in the control (Mann-Whitney U test, $\alpha = 0.01$) and all these differences indicate the increase of search result quality after diversification. In other words, the decrease of abandonment rate means less cases when users could not find any relevant result, the decrease of the mean first click position indicates that top results became more relevant, the decrease of the mean time to first click also indicates that relevant results received higher ranks and hence could be spotted faster, and the increase of CTRs of the URLs at the first two positions also indicates their increased relevance.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present an approach to improve recency ranking, while preserving the overall relevance of the ordinary search result. We developed a multi-grade recency sensitive queries classifier that predicts the degree of the need in recent documents. We further demonstrated how to diversify the ordinary search result with fresh documents by maximizing the search quality measure which takes the query's temporal ambiguity into account. We demonstrated the behavior of our diversification model in different cases using a set of judged queries. We finally confirmed the intuitions behind our approaches by a large-scale online experiment involving millions of queries from real users.

While we consider a fixed time window to determine if documents are fresh, it definitely makes more sense to determine time window which takes the essence of the information need expressed in the query into account. We also need to more systematically handle the challenge of score normalization to obtain the probabilities of document relevance generated according to each possible definition of relevance. In this regard, we look forward to exploit the techniques of results merging developed in the area of Distributed Information Retrieval [6].

The diversification based approach to recency ranking can be also useful to aggregate documents from a set of relevant verticals (videos, images or shopping items). However, the danger of over-diversification is not well studied yet. It is not clear if users would prefer too many results of different kinds blended into one search result page. Our long term goal is to develop a unified approach to deal with several kinds of query ambiguities: topical and non-topical.

7. REFERENCES

- [1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, Samuel Jeong. Diversifying Search Results. In *WSDM'09*
- [2] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Buttcher, Ian MacKinnon. Novelty and Diversity in Information Retrieval Evaluation In *SIGIR'09*, 2009.
- [3] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, Fernando Diaz. Towards Recency Ranking in Web Search In *WSDM'10*, 2010.
- [4] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng. Time is of the Essence: Improving Recency Ranking Using Twitter Data In *WWW'10*, 2010.
- [5] Jaime Arguello, Fernando Diaz, Jamie Callan, Jean-Francois Crespo. Sources of Evidence for Vertical Selection. In *SIGIR'09*, 2009.
- [6] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*.
- [7] Ron Kohavi, Randy Henne, and Dan Sommerfield Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. In *KDD '07*, pages 959–967, 2007.
- [8] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine In *Annals of Statistics*, 29, pages 1189–1232, 2001.
- [9] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630, 2009.
- [10] Filip Radlinski, Madhu Kurup, Thorsten Joachims How Does Clickthrough Data Reflect Retrieval Quality? In *CIKM'08*, pages 26–30, 2008.