

# Towards Recency Ranking in Web Search

Anlei Dong Yi Chang Zhaohui Zheng Gilad Mishne  
Jing Bai Ruiqiang Zhang Karolina Buchner Ciya Liao Fernando Diaz  
Yahoo! Inc.

701 First Avenue, Sunnyvale, CA 94089

{anlei, yichang, zhaohui, gilad, jingbai, ruiqiang, karolina, ciyaliao, diazf}@yahoo-inc.com

## ABSTRACT

In web search, *recency ranking* refers to ranking documents by relevance which takes freshness into account. In this paper, we propose a retrieval system which automatically detects and responds to recency sensitive queries. The system detects recency sensitive queries using a high precision classifier. The system responds to recency sensitive queries by using a machine learned ranking model trained for such queries. We use multiple recency features to provide temporal evidence which effectively represents document recency. Furthermore, we propose several training methodologies important for training recency sensitive rankers. Finally, we develop new evaluation metrics for recency sensitive queries. Our experiments demonstrate the efficacy of the proposed approaches.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Experimentation

## Keywords

Recency ranking, recency sensitive query classification, temporal features, recency modeling

## 1. INTRODUCTION

When a user submits a query to a search engine, he/she expects to obtain relevant search results. Classic notions of relevance focus on topical relevance. Web search introduces non-topical facets to relevance. For example, incorporating the authoritativeness of the information source in ranking can significantly improve user satisfaction. In this paper, we will argue that the freshness of documents can and should influence ranking and evaluation.

A large number of new web pages are created every day, and existing web pages are outdated with high rates. If stale documents are presented to users, they may seriously degrade search experiences. In recent years, the temporal dimension of web search has

been studied from different perspectives, such as web dynamics, crawling, temporal features, and modeling. Web page recency has been effectively taken into account for some specific applications (e.g. research publication databases and news article page ranking). However, there has not been a large scale, comprehensive study focused on recency sensitive evaluation and ranking.

In web search, *recency ranking* refers to ranking documents by relevance which takes freshness into account. Recency ranking on the large-scaled web search offers several unique and challenging research problems. First, ranking algorithms for recency sensitive queries need to satisfy both topical relevance and freshness. In order to ensure robustness, the system needs to promote recent content *only when appropriate*, so as to avoid degrading performance for other queries classes. Second, recency sensitive queries operate at different time scales. The freshness of a document depends on the time sensitivity of the query. For example, for the query “WSDM”, related pages are time-sensitive to the year. However, for a breaking-news query, the relevance of the results displayed may be sensitive to days or even hours. Third, measuring the true age of a document is hard. In general, temporal features cannot be accurately extracted from web pages; usually, only weak temporal evidence can be obtained. Finally, gathering data for training and evaluation requires temporally sensitive judgments synchronized with retrieval runs.

We propose a system which addresses recency sensitive queries using a query classification framework. That is, our system can be broken into two models: a high-precision recency sensitive query detector and a specialized recency sensitive ranker. Training a specialized ranker allows the system to model the unique data distribution and features useful for recency ranking. The main contributions of this paper include:

1. the analysis and formulation of recency ranking problem,
2. the development of a breaking-news query classifier with high accuracy and reasonable coverage,
3. the extraction of document temporal evidence, and
4. the development of new algorithms and strategies for recency ranking models.

## 2. RELATED WORK

The works that are most related to our approach include [8] and [22], which also directly improve ranking recency in web search. However, our approach provides a more generic solution than previous works, in terms of the content and query perspective. Diaz [8] proposed a solution to integrate search results from a news vertical search into web search results, where the news intent is detected by either inspecting query dynamics or using click feedback; our

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.

WSDM'10, February 4–6, 2010, New York City, New York, USA.  
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...\$10.00.

approach is to optimize ranking recency on all web pages directly, covering a more generic content perspective. Zhang *et al.* [22] proposed a ranking score adjustment method on year-qualified-queries, for which a few simple but effective adjustment rules are applied to the ranking results based on the timestamps extracted from the documents. Our approach can be applied on all time-sensitive queries, which covers a more generic query perspective.

The following prior works exploit the temporal dimension in general web search. Baeza-Yates *et al.* [2] studied the relation between the web dynamics, structure and page quality, and demonstrated that PageRank is biased against new pages. In T-Rank Light and T-Rank algorithms [3], both activity (i.e., update rates) and freshness (i.e., timestamps of most recent updates) of pages and links are taken into account for link analysis. Cho *et al.* [6] proposed a page quality ranking function in order to alleviate the problem of popularity-based ranking, and they used the derivatives of PageRank to forecast future PageRank values for new pages. Nunes *et al.* [17] proposed to improve web information retrieval in the temporal dimension by combining the temporal features extracted from both individual document and the whole web. Pandey *et al.* [18] studied the tradeoff between new page exploration and high-quality page exploitation, which is based on a ranking method to randomly promote some new pages so that they can accumulate links quickly.

Temporal dimension is also considered in other information retrieval applications. Del Corso *et al.* [7] proposed the ranking framework to model news article generation, topic clustering and story evolution over time, and this ranking algorithm takes publication time and linkage time into consideration as well as news source authority. Li *et al.* [15] proposed a TS-Rank algorithm, which considers page freshness in the stationary probability distribution of Markov Chains, since the dynamics of web pages is also important for ranking. This method is proved to be effective in the application of publication search. Pasca [19] used temporal expressions to improve question answering results for time-related questions. Answers are obtained by aggregating matching pieces of information and the temporal expressions they contain. Furthermore, Arıkan *et al.* [1] incorporated temporal expressions into language model, and demonstrated experimental improvement in retrieval effectiveness.

Recency Query Classification plays an important role in Recency ranking. Diaz [8] determined the newsworthiness of a query by predicting the probability of a user clicks on the news display of a query. König *et al.* [14] estimated the click-through rate for dedicated news search result with a supervised model, which is to satisfy the requirement of adapting quickly to emerging news event.

### 3. RECENCY DATA ANALYSIS

In this paper, we target breaking-news queries, because breaking-news queries exhibit most typical issues with regards to recency in search ranking results and where it is clear that user experience can be improved by improving the ranking. Breaking-news queries include queries about topics which are in the news at the time when the query was entered. In other words, there was a “buzz” or major influx of content in the media on that topic at the time when the query was entered by the user. The event which the query refers to may have happened several days before the date of the query, but if there is still significant news coverage for that event the query should be classified as breaking-news. There are other non-breaking-news-queries which may also have recency problems. We first study breaking-news queries in attempt to solve the recency problem within this query category. With increased maturity of the technology, we will extend similar approaches to other queries.

Recency ranking data collection is different from regular rank-

ing data collection. In regular ranking data collection, the relevance judgement for a query-url pair is usually static over time because document freshness does not affect the user satisfaction. The judgement for a recency-sensitive query-url pair, however, should incorporate the freshness of the document. For example, the epidemic *swine flu* was identified in April 2009. Therefore, for the query “swine flu” in April 2009, the web page of a news article reporting the identification of this epidemic can be appropriately labeled with the grade “excellent”. A few days later, this web page becomes outdated as there have been many more web pages reporting the latest status of this epidemic, beyond its identification; thus, the grade should be demoted from “excellent”.

We conducted an editorial test in order to confirm that there was indeed a relationship between objective recency and subjective recency. We sampled a set of queries automatically classified as recency-sensitive (details of this method are described in Section 4). From this set, editors selected only queries which were truly recency-sensitive for evaluation. For each query, we select the 20-30 top-ranked urls returned by a commercial search engine. We then employ different techniques to promote fresh urls. For example, based on the link discovery times, the fresher a page, the higher promotion score it should be given.

It is obvious that, to label recency data, a tuple  $\langle \text{query}, \text{url}, t_{\text{query}} \rangle$  needs to be provided for judgement instead of only  $\langle \text{query}, \text{url} \rangle$ , where  $t_{\text{query}}$  is query issue time. If the judging time  $t_j$  of  $\langle \text{query}, \text{url} \rangle$  is far behind query issue time  $t_{\text{query}}$ , it is impractical for the editors to do reliable judgement. Therefore, we collected sets of recency data periodically instead of collecting all the recency data for only one time. Each time we collected a set of recency data, we ask editors to judge the query-url pairs immediately, so that the query issue time and judging time are as close as possible. Collecting recency data periodically can also prevent the data distribution from being too biased towards a short period of time. For example, within one day, there are many similar queries related to the same breaking news, which are very different from the query distribution over a longer time span.

We then asked human editors for the following labels,

1. relevance judgments of query-url pairs
2. recency-sensitivities and true timestamps of a sample of urls

We apply five judgement grades on query-url pairs: perfect, excellent, good, fair and bad. For human editors to judge a query-url pair, we ask them to first grade it by non-temporal relevance, such as intent, usefulness, content, user interface design, domain authority, etc; then, the grade can be adjusted solely based on the recency of the result. More specifically, a result should receive a demotion if the date of the page, or age of the content, makes the result less relevant in comparison to more recent material or changes in time which alter the context of the query. This demotion should be reflected in the following judgment options:

- shallow demotion (1-grade demotion): if the result is somewhat outdated, it should be demoted by one grade (e.g., from excellent to good);
- deep demotion (2-grade demotion): if the result is totally outdated or totally useless, it should be demoted by two grades (e.g., from excellent to bad).

The advantages of this recency-demotion grading method include: 1) recency is incorporated into overall relevance so that the ranking model learning problem to be formulated as the optimization of a single objective function; 2) recency can also be decoupled















