Extracting News-Related Queries from Web Query Log

Michael Maslov, Alexander Golovko, Ilya Segalovich, Pavel Braslavski
Yandex
Vavilova 40
119991 Moscow, Russia
{maslov, algo, iseg, pb}@yandex-team.ru

ABSTRACT

In this poster, we present a method for extracting queries related to real-life events, or *news-related queries*, from large web query logs. The method employs query frequencies and search over a collection of recent news. News-related queries can be helpful for disambiguating user information needs, as well as for effective online news processing. The performed evaluation proves that the method yields good precision.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Verification.

Keywords

Query Log Analysis, Web Search.

1. INTRODUCTION

The query log of a web search engine is a rich source of valuable information about user preferences, search strategies, etc. There is a large body of publications on web query log analysis, which became a powerful technique for improving retrieval effectiveness and end-user interaction with a search engine. A number of scientific publications focused on *temporal* query log analysis have appeared recently. For example, study [1] discovers topical variations of queries throughout the course of the day. Chien and Immorlica [2] cluster semantically related queries based on similar temporal behavior of their popularities.

In this poster we present a method for extracting queries related to recent, ongoing, or upcoming real-life events reflected in the news, or *news-related queries*. Our approach is complementary in a certain sense to the method described in [3]. Whereas Henzinger et al. extract queries from current TV broadcast transcripts, thus enabling 'query-free news search', we aim at extracting queries related to real-life events from a general-purpose web search engine log, using relative query frequencies and validating them against current news feeds.

News-related queries can be used for disambiguating user information needs (e.g. prompting the user with a link to an online news service), as well as for highly effective online news processing, including news clustering, summarization, and ranking.

The poster briefly describes the method for extracting news-related queries implemented in Yandex News service (http://news.yandex.ru) that uses the query log and infrastructure of the general-purpose search engine Yandex (www.yandex.ru).

2. EXTRACTION METHOD

Query significance in a time interval as against to another interval is defined as a ratio of respective query frequencies:

$$S(q,\Delta_1,\Delta_2)=\frac{F(q,\Delta_1)}{F(q,\Delta_2)},$$

where $F(q, \Delta)$ is the frequency of q in time interval Δ.

Momentary query novelty can be defined as query significance in the last hour as against to the preceding day:

$$MQN(q) = S(q,\ \Delta_{last\_int}, \Delta_{prec\_day}).$$

To suppress some hourly irregularities (like weather queries in the morning and porn queries in the night), we define *hourly query novelty* as query significance in the last hour compared to the same hour of the day averaged over the preceding week:

$$HQN(q) = S(q, \Delta_{last\_int}, \Delta_{prec\_week}),$$

where Δ_{prec_week} is a union of time intervals corresponding to seven hours in the preceding seven days.

Final *query novelty* is defined as the minimum of *momentary* and *hourly query novelty:*

$$QN(q)=min\{MQN(q), HQN(q)\}.$$

To detect *novel queries,* very rare queries are removed; the rest is normalized (this step includes stemming, capitalization, and removal of some characters, e.g. quotation marks). Queries with novelty score exceeding the pre-defined threshold are considered to be *novel*.

News-related queries are a subclass of *novel queries*. To extract them, first, very broad queries are removed (i.e. queries with more than 0.1% of relevant documents in Yandex web database). Second, there must be relevant news that arrived within the three-hour time window around the query timestamp. For queries with more than 0.01% of relevant documents in Yandex web database, the search over news collection is restricted to headings only.

3. RESULTS

The described procedure extracts up to tens of thousands of *novel queries* and subsequently tens to hundreds of *news-related queries* from about one million queries hourly. We estimate that the fraction of news-related queries is about 0.01-0.1% (however, the fraction grows tenfold when an important event occurs). We intentionally extract a narrow class of queries from the stream for reasons of productivity, aiming at achieving high precision rather

WWW 2006, May 23–26, 2006, Edinburgh, UK.
ACM 1-59593-323-9/06/0005.

than high recall. Extracted queries have a number of interesting features.

First, news-related queries have different length distribution compared to general web queries. Figure 1 represents the length distribution of queries related to Beslan tragedy on September 1-3, 2004 (however, this kind of distribution is typical for news-related queries in general). The web query statistics originate from a one-hour log of one of the Yandex front-end machines. The majority (33%) of web queries are single word queries, whereas the main share (37%) of news-related queries consists of bigrams. There are 81% multi-word queries in the total of news-related queries.

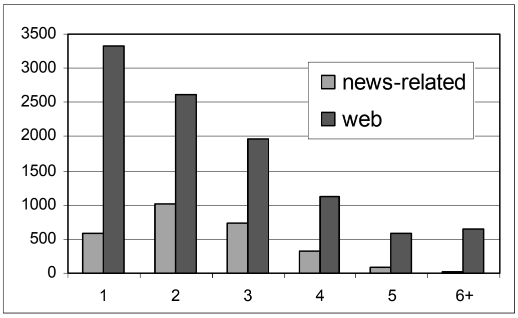

Figure 1. News-related / web queries length

News-related queries are not simply longer: they are very condensed event descriptors, often tying together important aspects of an event (e.g. location, date, actors, or type of event). This feature is illustrated by three pairs of sample queries related to three events (President Putin's press conference, Oscar nominees announced, and computer virus warning) (see Table 1). News clustering implemented in Yandex News enables to cluster lexically different queries related to an event. Hence, by using a relatively straightforward technique we can obtain important additional information that makes news processing more robust and precise. Extracting multi-word tokens from news stories alone would require exhaustive linguistic methods.

Table 1. Sample extracted event-related queries (Jan-31-2006)

Original query	English equivalent
пресс-конференция путина	*putin's press conference*
пресс-конференция в кремле	*press conference in kremlin*
компьютерный вирус 3 февраля	*computer virus february 3*
вирус пухет	*пухет virus*
горбатая гора энга ли	*brokeback mountain ang lee*
номинанты на оскар	*oscar nominees*

Moreover, the intensity of the extracted queries is a good indicator for the current user information needs. As proved by our results, internet users' interest (seemingly provoked by TV) is more reactive than Internet news sources' response, which makes news-related queries a useful parameter for news ranking.

4. EVALUATION

Yandex processes about one million queries per hour in daylight hours, which makes manual evaluation of a significant portion of the log unfeasible. We chose four one-hour intervals between 10 am and 7 pm in two consequent workdays in December 2005. The test sample included all queries automatically detected as news-related plus randomly selected 2% of the remaining queries within the respective intervals. The test sample contained 831 queries, 244 (30%) of which were automatically detected as news-related.

The test sample was presented to an assessor who evaluated queries in sequence. The assessor answered the question: "Is it safe to presume that the vast majority of the users making the query at the given time were interested in current news?" The results of the evaluation are summarized in Table 2 (the number of misses is multiplied by 50; due to lack of space we do not refer to frequency-weighted recall and precision values that are about 10% higher). The data allow us to make some observations.

First, the agreement between the assessor and automaton grows gradually from the first evaluated portion to the last one. This fact can be explained by an increase in the assessor's competence during the evaluation, since moving from early queries further the assessor is getting a more complete view of the relevant events.

Second, in both cases, recall for morning queries is considerably lower than for evening queries. This can be explained by the design of the algorithm. For productivity reasons, novel queries are detected based on query statistics for the whole preceding day (see Section 2). Thus, if an event occurred yesterday and there are both related articles and queries dated yesterday, then the algorithm often fails to detect the queries as news-related next morning.

Table 2. Evaluation Results

	Dec 7: 1 pm	**Dec7: 6 pm**	**Dec 8: 10 am**	**Dec 8: 3 pm**
Misses	9*50	7*50	7*50	5*50
TruePos	122	130	101	145
FalsePos	30	25	12	27
Precision	0.80	0.84	0.89	0.84
Recall	0.21	0.27	0.22	0.37
F1	0.34	0.41	0.35	0.51

5. CONCLUSION

The presented methodology for extracting news-related queries from a general-purpose search engine yields good precision. The extracted queries can be effectively used for improving user interaction with the search engine and in online news processing.

The results outlined in Section 4 suggest that news-related query detection in the morning can be improved by comparing query statistics not with the preceding day but with a sliding interval.

Additionally, we are going to experiment with processing intervals of about 15 minutes in order to increase the sensitivity of the method, which is important for the news processing tasks.

6. REFERENCES

[1] Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., and Frieder, O. Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *SIGIR'04*, July 25–29, 2004, Sheffield, South Yorkshire, UK, 321-328.

[2] Chien, S. and Immorlica, N. Semantic Similarity Between Search Engine Queries Using Temporal Correlation. In *WWW2005*, May 10-14, 2005, Chiba, Japan, 2-11.

[3] Henzinger, M. et al. Query-Free News Search. In *WWW2003*, May 20-24, 2003, Budapest, Hungary, 1-10