User-Centric Content Freshness Metrics for Search Engines

Ali Dasdan
Yahoo! Inc.
Sunnyvale, CA 94089, USA
dasdan@yahoo-inc.com

Xinh Huynh
Yahoo! Inc.
Sunnyvale, CA 94089, USA
xinh@yahoo-inc.com

ABSTRACT

In order to return relevant search results, a search engine must keep its local repository synchronized to the Web, but it is usually impossible to attain perfect freshness. Hence, it is vital for a production search engine continually to monitor and improve repository freshness. Most previous freshness metrics, formulated in the context of developing better synchronization policies, focused on the web crawler while ignoring other parts of a search engine. But, the freshness of documents in a web crawler does not necessarily translate directly into the freshness of search results as seen by users. We propose metrics for measuring freshness from a user's perspective, which take into account the latency between when documents are crawled and when they are viewed by users, as well as the variation in user click and view frequency among different documents. We also describe a practical implementation of these metrics that were used in a production search engine.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement

Keywords

Crawling, document age, freshness, latency, metrics, monitoring, search engine

1. INTRODUCTION

In an ideal world, the search results seen by a user are based on the entire content available on the Web at the time of the query. In practice, a search engine has a crawler which continually crawls and re-crawls the web, fetching and storing web pages in a local repository. Due to limited resources, not all of the local copies are up-to-date. In addition, it takes time for the crawled content to be indexed and stored into searchable indexes. When a query comes in, documents in these partially stale indexes are retrieved and ranked by the runtime system, and a small subset of those documents are displayed to the user.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Since the freshness of the search indexes does impact the quality of search results as seen by users, it is important for a production search engine to measure and monitor its freshness from a user perspective. Not only would such a metric capture the effectiveness of the synchronization policy, but it would also detect any system issues or bugs along the rest of the search engine pipeline.

Most previous measures of freshness were devised in the context of developing a better synchronization policy in the crawler. Thus, they have focused on the freshness of documents as they exist in the crawler. Two common metrics are *freshness* and *age* [1]. The *freshness* of a crawler is simply the fraction of documents that are fresh, i.e., have not changed since last time they were crawled. The *age* of a document quantifies how stale the local copy is; the *age* of a crawler is the average of the ages of its documents.

The article [3] does bring in the user perspective in its "top-k freshness" metric, which looks only at documents that appear in the first page of search results. This metric is defined for a result set, however, rather than for an entire search engine.

Another article [2] introduces the concept of a weighted freshness metric, whereby documents contribute unequally to the overall freshness depending on their "importance". They point out that importance can be related to the frequency at which a document is associated with a query, although without elaborating how to quantify it.

In this article, we propose a new way of measuring freshness from a user perspective. Our contribution is in devising metrics that: (1) expand on the idea of limiting the documents to ones that users actually see, hence, user-centric; (2) build on the concept of a weighted metric by using the number of clicks or views as weights; and (3) account for the latency involved in indexing a document after it has been crawled. We also describe a practical implementation of these metrics that were used successfully in a leading web search engine.

2. USER-CENTRIC FRESHNESS METRIC

Fig. 1 illustrates our user-centric metrics. A local page was crawled (or sync'ed) at time 1, indexed at time 2, and clicked by a user at time 6. Its age reflects the fact that the local copy has been stale for 3 days, ever since the web copy was first modified after the last sync. If the web copy had not been modified, the local page would have been fresh, with an age of zero, even though the local copy has been sitting in the index for 4 days at the view time. Hence, our metrics measure the staleness of a page with respect to if and when it is clicked by users.

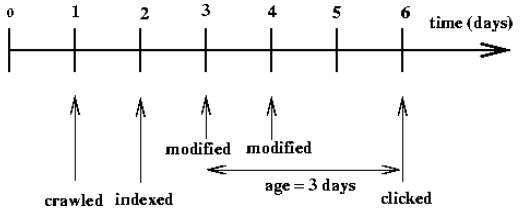

Figure 1: User-centric age of a page at time 6.

In the following definitions, assume the repository S (of size $|S|$) refers to the set of crawled pages and S_c to the set of clicked pages. The definitions for viewed pages or pages of any other category, e.g., news pages, are analogous. The weighted metrics consider the fact that pages are clicked in the search results with varying frequency. As such, a stale page clicked on by many users can have a bigger impact on the perception of freshness of a search engine index than many pages that rarely show up in user search results. The unweighted metrics ignore the frequency impact. The metrics are defined in an average sense but can easily be converted into histograms.

Definition 1. The *freshness* of a local page p at time t is defined as $F(p,t) = 1$ if p is up-to-date at time t (i.e., not modified since its last sync), 0 otherwise. The *age* of a local page p at time t is $A(p,t) = 0$ if p is up-to-date at time t, $t - t_{mod}$ otherwise, where t_{mod} is the time of the first modification after the last sync of p.

Definition 2. The *(basic) freshness* of S at time t is defined as $F(S,t) = \frac{1}{|S|}\sum_{p\in S} F(p,t)$. Similarly, the *(basic) age* of S at time t is defined as $A(S,t) = \frac{1}{|S|}\sum_{p\in S} A(p,t)$. These definitions respectively refer to the *unweighted freshness* F_u and *unweighted age* A_u when S is replaced by S_c.

Definition 3. Let $nclicks(p,t)$ be the number of times users have clicked page p since it was first modified after the last sync. The *weighted freshness* of S_c at time t is

$$F_w(S_c,t) = \frac{\sum_{p\in S_c} F(p,t) * nclicks(p,t)}{\sum_{p\in S_c} nclicks(p,t)}, \qquad (1)$$

and the *weighted age* of S_c at time t is

$$A_w(S_c,t) = \frac{\sum_{p\in S_c} A(p,t) * nclicks(p,t)}{\sum_{p\in S_c} nclicks(p,t)}. \qquad (2)$$

3. EVALUATION AND DISCUSSION

We implemented the user-centric freshness metrics described above to monitor a production search engine. The metrics were updated periodically. Due the search index size, it was impractical to measure all documents. Instead we sampled documents from the search engine's query logs. To take into account the variation in queries over time, we took a new sample periodically from query logs, and tracked those documents over multiple periods. After being tracked for a pre-determined number of periods, documents were removed from the sample.

The *freshness* and *age* metrics require knowledge about when a document is modified on the web. In order to collect this information, we have set up a separate crawler that synchronizes sampled documents periodically at a fixed rate. Thus, we knew modification times of these documents up to the resolution of the refresh period.

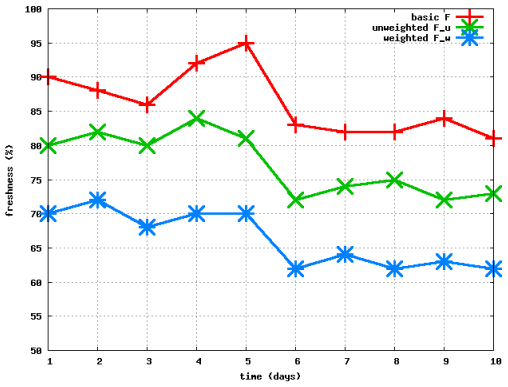

Figure 2: Comparison of freshness metrics.

We obtained daily click and view statistics from the search engine's query logs. The freshness metrics were re-calculated periodically for the local copies of the sample residing in the search indexes.

Due to the confidential nature of the data, we will not show the actual data collected. Instead, Fig. 2 shows a representative example to illustrate the value of the user-centric metrics. In this example, the fact that "user_unweighted" is consistently higher than the "basic" metric implies that the search engine is fresher for documents that are actually viewed or clicked by users. The fact that "user_weighted" is even higher implies that the search engine is even fresher on a per-click or per-view basis. In addition, the sudden drop in freshness on day 6 alerts us of a possible problem in the search engine pipeline.

One limitation of the methodology is the time resolution limit due to the refresh rate of the separate crawler and the sample rate from query logs. However, this limitation can be overcome by simply increasing those rates of data collection.

4. CONCLUSIONS

We proposed metrics for measuring freshness from a user's perspective. These metrics account for the latency added by the full search engine pipeline, and capture the variation in user click and view frequency among different documents. We described a practical implementation geared at measuring and monitoring freshness in a production search engine.

5. REFERENCES

[1] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. *SIGMOD Rec.*, 29(2):117–128, 2000.

[2] J. Han, N. Cercone, and X. Hu. A weighted freshness metric for maintaining search engine local repository. In *Proc. of Int. Conf. on Web Intelligence (WI)*, pp. 677–680. IEEE, 2004.

[3] Q. Tan, P. Mitra, and C. L. Giles. Designing clustering-based web crawling policies for search engine crawlers. In *Proc. of Conf. on Info. and Knowledge Management (CIKM)*, pp. 535–544. ACM, 2007.