Calculating Content Recency based on Timestamped and Non-Timestamped Sources for Supporting Page Quality Estimation

Adam Jatowt[1,2]
[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku,
606-8501 Kyoto, Japan
Phone: +81-75-7535969
[2]Microsoft IJARC Fellow
adam@dl.kuis.kyoto-u.ac.jp

Yukiko Kawai
Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku,
603-8555 Kyoto, Japan
Phone: +81-75-7052958
kawai@cc.kyoto-su.ac.jp

Katsumi Tanaka
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
606-8501 Kyoto, Japan
Phone: +81-75-7535969
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

The web is characterized by low publishing barriers and contains content of varying degrees of quality and credibility. It is often difficult for web searchers to locate high quality content in returned search results. In this paper, we propose evaluating the extent to which search results contain recent information related to user queries. Our approach is based on corroborating search results with query-related information obtained from timestamped and non-timestamped sources. It uses news articles collected from online news archives and also employs a simple search index mining process to find terms representing fresh topics. As another contribution, we show how the proposed approach can be used for estimating the focus time of web pages, that is, the time periods to which the content of pages refers. We demonstrate the proof-of-concept system that evaluates and visualizes in real time the freshness levels and focus time of web search results.

Categories and Subject Descriptors

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

General Terms

Algorithms

Keywords

web content analysis, content freshness, information recency

1. INTRODUCTION

Research aimed at evaluating web content credibility is recently becoming increasingly necessary since when the web started to influence our daily lives. Many types of credibility violations can be encountered in the web. Below we list three example types in the increasing order of their frequency and perceived social acceptance.

1. Deliberate cheating, e.g. purposely changing information, posting lies.
2. Wrongly interpreting or presenting information from biased angles, e.g. facts may be correct but their context is wrong.
3. Omitting information, e.g. to support some claims, due to the lack of knowledge or simply forgetting, or being unable to update own pages.

The third type seems to be the most frequent one. Especially, as keeping content up-to-date requires certain effort and time, it is relatively common for authors to fail updating pages despite the occurrence of new, important information related to the topics of their pages. It has been found that there are many abandoned documents in the web and that lots of pages provide obsolete information [2]. The lack of important or recent information in pages is actually often regarded as a kind of "soft" or "justifiable" credibility violation in comparison to deliberately misguiding readers by altering and manipulating information. Note, however, that despite the somewhat lax perception of this kind of behavior, its consequences can still be harmful to users.

In order to verify the freshness and completeness of information users often tend to utilize external sources, especially, high quality ones [10,12]. However, manual investigation of multiple related resources is often tiresome as it usually takes considerable time for users to correctly locate fresh and comprehensive information on particular topics. Therefore we propose an automatic approach for estimating the degree to which web pages cover most recent information related to searched topics. We focus on web search results as search engines are unquestionably the main gateways to the web[1]. Moreover, there is explicit information provided about user information need in the form of her or his search query. In

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SAC'11, March 21-25, 2011, TaiChung, Taiwan.
Copyright 2011 ACM 978-1-4503-0113-8/11/03…$10.00.

[1] Note that the proposed freshness evaluation method is not only limited to web search results but could be applied to arbitrary pages using any state-of-the-art keyword extraction algorithm.

particular, we concentrate on a set of queries representing real-world entities such as persons, places, companies, etc.

We propose two approaches based on timestamped and non-timestamped resources. The first approach uses news articles which constitute authoritative and time-stamped data, providing important factoid-type information on real-world objects. Recently, many digital news archives have appeared on the web. On the top of that, large web companies like the Google[2] or Yahoo![3] started assembling data from multiple distributed news providers. Many of such online news aggregators enable temporal search for returning news articles published within selected time frames. It is thus possible to collect documents for a user query that were published recently and that presumably contain information on latest, related events. We detect such events by clustering the collected news articles and then compare the content of clusters with web search results.

The assumption behind the above approach is that there is sufficient number of news articles or, in general, any related, timestamped documents that can be collected for user query. However, for obscure objects, one may not be able to obtain enough such documents to detect any events. Therefore, we also propose the second approach which relies on finding terms that represent recent events related to queries. This is done by mining search engine interfaces. We issue several extended queries to a search engine and we extract representative terms from the returned search results. These terms are then filtered using log-likelihood rate test based on comparison with another set of representative terms that are, in contrast, indicative of old events related to the query.

Note that pages' recency cannot be simply assessed by checking their timestamps (i.e., content creation or update time). This is because pages may not contain any explicit temporal metadata or it may be misleading. The "last-modified" dates returned by web servers are often unreliable or unavailable. Also the timestamps, if present in pages, serve only as a crude proxy and do not provide enough assurance as for the actual content freshness. This is because, even if a given content has been recently published, it may not necessarily contain the latest information related to page topics.

We demonstrate a prototype system to support users in assessing the extent to which the returned documents cover fresh topics. It visualizes calculated newness scores next to the returned search results and provides additional information for users that could be helpful in evaluating search result quality. In addition, we demonstrate how the proposed approach can be used for estimating the focus time of documents, that is, the time periods to which the document content refers. This kind of analysis could be useful for approximating document last-modification dates as well as for improving temporal IR, whose typical objective is to return documents that are related to specified time frames.

Our approach is different from previous works in recency ranking of search results [4,6,9,11] in several aspects. First, we do not detect temporal queries as we do not propose any re-ranking of search results. Instead, in our approach, freshness scores are visualized next to returned search results in order to support users in finding high quality pages. This choice is motivated by the observation that users may have very diverse intents behind their queries and it may be difficult to correctly predict them. Second, for many queries, even for news type ones, gathering timestamped documents may not be always feasible. Therefore we also propose a more general method that is based on mining the web for query-dependent fresh information. Lastly, our approach identifies pages not only containing the fresh information but rather pages that have high coverage of recent information as the latter ones should be more valuable for users.

To sum up, our contributions are three-fold: a) we propose approach for freshness evaluation of web search results using timestamped and non-timestamped sources, b) we demonstrate the potential of our method for estimating document focus times, and c) we present a proof-of-concept prototype system.

The remainder of this paper is structured as follows. In Section 2 we describe the related work. In the next section we present two methods for estimating the recency of search results. In Section 4 we demonstrate the implemented prototype system and show the experimental evaluation. Finally, we conclude the paper in the last section.

2. RELATED RESEARCH

In the information quality theory, the accuracy, authority, objectivity, currency and coverage are frequently used evaluative criteria for documents [13]. The checklist approach has been the most commonly advised strategy for users to manually evaluate the quality of information in web sites. According to the prescribed guidelines, users should investigate different criteria of encountered content, usually, by answering prepared set of questions (e.g. "does the site provide information about when the content was last posted or updated?" [13]).

In contrast to the approaches that rely solely on target content, Meola [12] and Lankes [10] argued that users often seek for commonalities and coherence among multiple information sources in order to gauge the extent to which they can rely on particular information. This kind of corroboration fits well into the web environment where plenty of different information sources are available for comparison.

Yet, despite the presence of credibility evaluation guidelines many users are unprepared, do not possess sufficient skills to properly assess the quality of online information or are reluctant to do so [12,13,17]. Consequently, few of them perform rigorous evaluation of the quality of obtained information. For example, according to the recent statistics in medical domain [15], three-quarters of online health seekers in the USA or about 85 million Americans gather health advice online without consistently examining the quality indicators of the information such as its source and date. Interestingly, 53% of the health seekers admitted that the online health information had some kind of impact on how they take care of themselves or care for someone else. This example supports the claim that automatic tools for assisting users in the judgment of web content quality are becoming increasingly necessary. It is however important to emphasize that the notion of the quality is a complex construct composed of many underlying dimensions [13,17].

[2] http://news.google.com/archivesearch

[3] http://www.yahoo.com

Currency of information is one of the key quality factors. For pages covering relatively dynamic topics, information conflict could occur between their content and the latest related events. The lack of current information is a major drawback especially for pages which make impression of containing fresh content. These could be newswire pages, pages explicitly claiming to contain latest events related to particular topics or pages for which readers have expectation of content currency. The last type of pages contains documents such as government pages, company homepages, Wikipedia[4] articles, and many types of official pages.

Some researchers have already approached the problem of web information credibility by focusing on information currency. Juffinger et al. [8] proposed calculating the coverage of news in blogs and the level of their synchronization as a measure of blog credibility. In [7] we reported preliminary results of a method for estimating news coverage of web search results with recency and importance as its two constructs. In this paper we extend our previous work by proposing a method based on processing non-timestamped sources. We also introduce the concept of page focus time and present a proof-of-concept, real-time system.

Toyoda and Kitsuregawa [18] demonstrated a measure for estimating approximate creation dates of pages based on novelty scores of linking documents. Bar-Yossef et al. [2] employed link-structure analysis on the web for identifying decayed web sites. In contrast, in this work, we approach the problem of search results' recency from the content-based viewpoint.

In response to various events occurring in the world users often issue timely queries to search engines in order to learn about these events or track their progress. Diaz [5] has recently proposed an efficient algorithm for incorporating news articles into web search results. We take a different approach as we indicate pages that are informative on recent events and provide up-to-date information.

The previous works on recency ranking [4,6,9,11] often focused on classifying queries into time-sensitive and non-time-sensitive ones [6,14]. For time-sensitive queries, the temporal relevance of documents has been computed, usually, based on modifications of language models [4,9,11]. In our work, we do not need to detect time-sensitive queries as our approach does not involve page ranking. Also, our methods are not based on language modeling.

Usually, the prior works on temporal ranking required the presence of timestamped documents such as news archives or occurrence of explicit temporal expressions in content [1,4,6,9,11]. However, for many queries there may not be enough timestamped documents available, or the page creation time cannot be extracted with high reliability. Moreover, documents may not have sufficient temporal references to calculate their temporal relevance. Dong et al. [6] used multiple recency features to represent document recency. Similar to our work the authors have used timestamped and non-timestamped documents. For the former they measured the recency by using information on document age, while for the latter they used various link-based evidences. In this work, we use clustering method for timestamped sources that helps to identify query-related events and measure their importance and recency. For general type of documents we propose web mining process based on using multiple evidences of recent query-related topics on the web.

3. METHOD

3.1 Approach for Timestamped Sources

The method based on timestamped resources works as follows. First, a user issues a query to a search engine. Next, the query is analyzed for the occurrence of named entities such as object, person, place or other names using a standalone recognition tool. The extracted named entity is then transferred to an online news archive. The news articles related to the object that were published within a pre-specified time frame are retrieved. Using this data we detect main events that occurred in the required time frame through clustering. Next, the distilled events are compared with the content of the returned web search results in order to find the degree to which the events are covered in the web pages. This information is then shown to users to support them in choosing credible search results.

3.1.1 Data Collection from News Archive

After a user submits her or his query, the query is analyzed for the occurrence of named entities. The part of the query that describes any real world object is then forwarded to an online news archive together with a predefined time period $T=[t_{beg},t_{end}]$ [5] for which the news articles are to be collected. T defines the scope of the recency analysis. When extracting news articles, T is divided into R number of continuous and non-overlapping time units, which serve as a set of temporal constraints for the query. That is, the query is issued R number of times to the news archive, each time with different temporal constraint (see Figure 1). For each such query the news archive will deliver only documents created within the time unit specified by the time constraint associated with the query. We apply this data collection process to ensure that the collected news articles come from all the time units within the required time frame T. Otherwise, the bulk of results could come only from a single sub-period of T, for example, due to a single major event that occurred within that sub-period. Or, in another case, only recent documents could be returned depending on the ranking policy of the news archive. In both cases the results would be biased and one might not obtain the correct representation of all major events related to the query within T.

We collect up to N/R results for each unit time period, where N is the pre-specified total limit of results. For efficiency and due to access restrictions, we use only snippets and titles of returned news articles instead of their whole content. The snippets are composed of content parts of news articles that include the query words and can be regarded as short domain-focused summaries. For brevity, we will call the news articles' snippets as news articles. In addition, for each news article we record its timestamp.

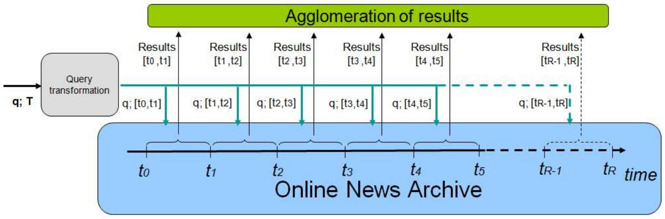


Figure 1 Data collection from online news archive.

[4] http://www.wikipedia.org

[5] By default t_{end} is equal to the query issuing time.

3.1.2 Event Detection

We detect events in our collection by applying *k-Means* - a popular partitional clustering method and assuming that a single news article covers only one event as a main topic of its content. We exploit the well-known observation that news articles on the same event are often generated around the same time. However, relying only on the temporal closeness of news articles would result in poor quality clusters. Thus, in our method, the distance function between each pair of news articles depends both on their time distance as well as on content distance.

$$\begin{aligned} Dist\,(d_i, d_j) &= (1-\alpha) * TimeDist\,(d_i, d_j) \\ &+ \alpha * ContDist\,(d_i, d_j) \end{aligned} \tag{1}$$

Time distance in Equation 1 is a normalized time difference between the timestamps of the news articles expressed as the number of days. Content distance is measured by calculating Euclidean distance between the news articles' feature vectors constructed using the bag-of-terms representation after eliminating stop words. Feature vectors are created using *term frequency – inverse document frequency* (*tf-idf*) weighting scheme, which is commonly used in web search and many IR systems [16]. We use weights *0.5* and *1* in the *tf* part to assign lower scores to term instances occurring in snippets as opposed to the ones inside news articles' titles, respectively.

By varying the mixture parameter α in Equation 1 we can control the influence of the content- and time-based distances in the clustering. For $\alpha=0$ the news articles are clustered based solely on their temporal distributions and independently of their content. In such a case, bursty events can be effectively detected; however, the method performs poorly when two or more different events occur around the same time. On the other hand, for $\alpha=1$ the clustering is done irrespectively of time. This offers the advantage of detecting events that last over longer time spans. In the current implementation, we use $\alpha=0.5$ as a default value.

The optimal cluster number, *k,* for the k-means algorithm is calculated by applying *Calinski-Harabasz method* [3]. It measures the quality of clustering results to find the cluster combination characterized by the minimum average distance between the documents within the same cluster (intra-cluster distance) and the maximum average distance between different clusters (inter-cluster distance). The method requires first setting the minimum and maximum number of clusters. Formally, it selects the number k ($k \geq 2$) of clusters by maximizing the following function:

$$\Phi(k) = \frac{\sum_{l=1}^{k} B_l / (k-1)}{\sum_{l=1}^{k} W_l / (N-k)} \tag{2}$$

B_l is the square sum of Euclidean distances between a cluster l and any other cluster, and W_l is the square sum of Euclidean distances between the members of the cluster l. Thus, B_l represents the inter-cluster distance and W_l represents the intra-cluster distance.

In our implementation we remove very small clusters (i.e., the ones having the number of members less than 3 news articles). In addition, in order to decrease the effect of noisy clusters we also remove low quality clusters. The following formula is applied to estimate the cluster quality.

$$Q_l = \frac{B_l / (k-1)}{W_l / (csize\,(l) - 1)} \tag{3}$$

Here, *csize(l)* denotes the size of the cluster *l* expressed as the number of its documents. The high quality clusters should describe topics that are different from other clusters and that also contain topically similar documents.

3.1.3 Freshness Calculation

The resulting clusters are represented in the vector space to be compared with the content of web search results. We calculate centroid vector for each cluster called event vector, v_l^{event}.

Next, we compute the vector representation of web search results. The feature vectors of web search results are calculated using the *tf-idf* weighting scheme after the removal of stop words. We measure the similarity between the web search results and all the clusters using the cosine similarity measure. Freshness score of a page is then calculated as the weighted sum of page-cluster similarities.

$$F_p^{time} = \frac{1}{k} \sum_{l=1}^{k} w_l * sim\left(v_p, v_l^{event}\right) \tag{4}$$

v_p is a feature vector of page p. w_l denotes here the weight assigned to the cluster l and is defined as:

$$w_l = \frac{csize\,(l)}{\underset{1 \leq i \leq k}{Max}\,(csize\,(i))} * e^{\lambda \frac{t_l - t_{beg}}{t_{end} - t_{beg}}} \tag{5}$$

The weight w_l depends on the age of clusters expressed as the relative time period that elapsed since the event's occurrence. The event's occurrence time, t_l, is estimated as the average timestamp of news articles belonging to the underlying cluster. We also define the event's importance and include it into the equation for calculating the weights. It is calculated as the normalized membership count of the cluster representing the event with the objective to reflect the relative news attention on the event. The assumption here is that a high quality page should contain more information on important events rather than on the less important events. Note that the importance and recency are not perfectly orthogonal dimensions as it usually takes some time for an event to become popular.

Parameter λ in Equation 5 is used to control the influence of cluster age on the weight and, hence, the impact of recency on the page scores. The default value of λ is set to *0.5*.

Equation 4 is based on summing evidences of page similarities with all the detected events. Thus a page that describes many recent events will have a high score assigned. An alternative way is to use maximum function instead of the sum. In this case, pages that cover only one event such as online news articles would receive high scores.

We think that there is certain user expectation of information's age and significance in web search. That is, for general queries, such as the names of real-world objects or persons, pages covering relatively unimportant events should appear in the top search results only if the events are novel. On the other hand, documents devoted to older news should be included in the top search results only if the referred events are of high importance. Following, web pages that mostly cover unimportant and old events should not have high visibility in search results; at least, for the general types of queries. Such pages should be ranked highly only in the case when users explicitly formulate queries in a way to retrieve the information on particular old events. The dependency of information usefulness on the event recency and importance is conceptually visualized in Figure 2.

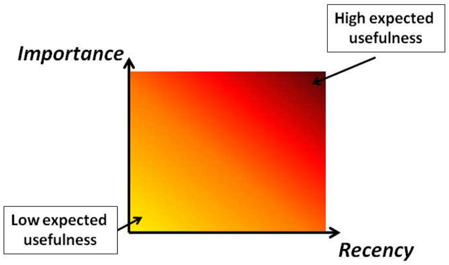


Figure 2 Information usefulness in relation to event's importance and recency (dark color indicates high usefulness).

3.2 Approach for Non-Timestamped Sources

The freshness evaluation can be also done using search engine mining process. It is based on the assumption of self-updating web according to which majority of topic-related pages contain contents that are updated as the time goes by.

In this approach, we first find terms that characterize fresh topics related to user query. We detect such terms by collecting content on the web that is not only related to the user query but is also associated with the set of freshness-indicating expressions. This set contains expressions such as "latest", "recent", "news" etc. Dates indicating periods or time points near the query issuing time could be also used here as temporal expressions. Naturally, some terms from the set cannot always guarantee the return of recent content related to query. However, on average, when considering the majority view on the web, the results should indicate recent topics related to the target query. In addition, we also use a filtering process based on obsoleteness-indicating terms such as "past", "history", etc.

The algorithm is shown in Figure 3. We describe its steps below:

1. Extended queries are formed with the user query and each term from the set of freshness-indicating expressions.
2. Extended queries are formed with the user query and each term from the set of obsoleteness-indicating expressions.
3. All the extended queries from step 1 and 2 are issued to a search engine.
4. *M* top results returned from the search engine are collected for each extended query[6].
5. Log-likelihood ratio test is applied to each term based on its occurrence frequencies in the search results obtained for extended queries from step 1 and step 2.
6. Based on the results from step 5 each term has positive and negative scores calculated that are used for estimating freshness of web search results.

We use here again document snippets as in the method for time-stamped sources. This allows building a real time freshness evaluation system. The log-likelihood ratio test mentioned in step 5 is used to compare term occurrence in the results of queries extended with freshness-indicating and obsoleteness-indicating expressions. The objective is to decrease the influence of terms that do not uniquely characterize either fresh or obsolete topics. For each term we make a 2*2 contingency table as shown in Table 1. Let F and O indicate search results obtained for queries extended with freshness-indicating expressions and for queries extended with obsoleteness-indicating expressions, respectively. f_j is the count of search results containing a term j in F, while o_j is the count of search results containing term j in O.

Table 1 2 * 2 contingency table for terms.

f_j	o_j
$\|F\|$ - f_j	$\|O\|$-o_j

The log-likelihood ratio test is calculated as follows:

$$LL_j = 2\times\left(f_j \times \log\left(\frac{f_j}{E_1}\right)\right) + \left(o_j \times \log\left(\frac{o_j}{E_2}\right)\right) \qquad (6)$$

where E_1 and E_2 are expected values estimated as below,

$$E_1 = \frac{|F|\times(f_j + o_j)}{|F|+|O|}$$
$$E_2 = \frac{|O|\times(f_j + o_j)}{|F|+|O|} \qquad (7)$$

We use the following conditions to assign positive and negative scores to terms,

$$pos_j = \begin{cases} LL_j & if \quad \frac{f_j}{|F|} > \frac{o_j}{|O|} \\ 0 & otherwise \end{cases}$$
$$neg_j = \begin{cases} LL_j & if \quad \frac{f_j}{|F|} < \frac{o_j}{|O|} \\ 0 & otherwise \end{cases} \qquad (8)$$

The freshness degree of a page is measured by the following equation,

$$F_p^{notime} = \frac{\sum_{i\in V_p} pos_i}{\sum_{i\in V_p} neg_i} \qquad (9)$$

V_p denotes here the vocabulary set of a page p. Pages that contain many terms having high positive scores and few terms having high negative scores are deemed to be high quality pages from the recency viewpoint. The weakness of the above approach is its poor ability to detect events that are periodically re-occurring. Terms that characterize such events may have decreased scores after the log-likelihood ratio test.

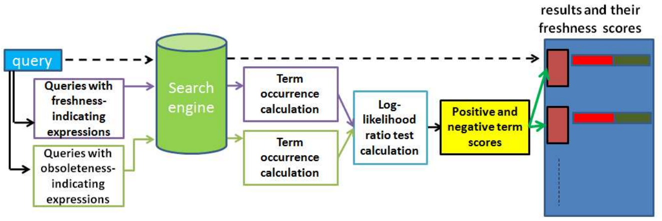

System showing search results and their freshness scores
query
Queries with freshness-indicating expressions
Queries with obsoleteness-indicating expressions
Search engine
Term occurrence calculation
Term occurrence calculation
Log-likelihood ratio test calculation
Positive and negative term scores

Figure 3 Outline of the method using non-timestamped documents.

[6] *M=100* by default.

4. DOCUMENT FOCUS TIME

By using the clusters distilled from timestamped documents it is possible to measure the *focus time* of documents.

Definition. *The* ***focus time of a page p*** *is the set of time periods determining the occurrence of events covered by the content of p.*

In other words, the focus time of a page identifies time frame to which the document's content refers. Thus, a page about the 1st and 2nd World Wars would have time focus: {[1914, 1918], [1939, 1945]}. Note that focus time is different from the page creation or modification time. We estimate it by comparing similarity of events and page's content against a predefined threshold, θ.

$$\Gamma_p = \left\{ \quad \left[t_l^{beg}, t_l^{end}\right] \quad \middle| \quad sim\left(v_p, v_l^{event}\right) \geq \theta \right\} \qquad (10)$$

Here, t_l^{beg} and t_l^{end} are the start and end time points of the time period during which an event l happened. The period [t_l^{beg}, t_l^{end}] is calculated as two standard deviations before and after the event's occurrence time, t_l. Although assuming the timestamps of the oldest and youngest news articles in the cluster as t_l^{beg} and t_l^{end} would be simpler; however, in such a case the time period of the event could be biased too much by outlier documents in the cluster. The time periods returned by Equation 10 are automatically merged if they overlap or connect so that the final set has the minimal number of elements.

The concept of page focus time serves to map page content onto a timeline and position it with respect to other events related to the query. It can be also used for estimating the time point of the latest page update. We base this estimation on the assumption that the last-modification date of a page should be equal to or older than the latest event reported in a page. In other words, we assume that the page's author can report a given event only after the occurrence of this event[7]. The expected last-modification dates are then determined as the start date of the most recent time periods within the documents' focus times. We use here the start date of the latest reported event rather than its end date as the page may have been updated for the last time after the start date of the event. Taking the left-hand side of the event's time boundary is thus a safer choice than using the right-hand side one.

In general, the focus time of documents has the potential to be useful for improving temporal IR [1]. For queries containing temporal component such as "US election 1988", including the estimated document focus time into the process of ranking search results should improve the overall relevance. For example, a page which is relevant to the 1988's US presidential election, but which does not contain the term "1988" or any other date in its content, would probably not be ranked highly by a conventional web search engine. Yet, it could be ranked highly if page focus time is used.

Note that it is impossible to correctly estimate the focus time of a page if it is outside the initial time frame *T*. It is also difficult to estimate the true focus time of a page whose content is unrelated to any timestamped documents. In such a case, other methods should be applied, such as the ones based on extracting temporal expressions in text [1].

5. EXPERIMENTS

In this section we present proof-of-concept system implementation and demonstrate how it can help users to better choose high quality search results. Next, we show the evaluation of our methods using a large online news archive.

5.1 System Implementation

We have built a prototype system in C# using Microsoft .NET framework and the Yahoo! Search BOSS API. As a source of news articles we use Google News Archive[8] which is a popular online news aggregator. According to the Google, over 4,500 popular English-language news sites are archived including major publishers and news providers.

Figure 4 shows example results returned by the system for the query "Poland" with time frame *T=[2000,2009]*. On the left-hand side, in each row, the system displays the titles, hyperlinked URLs and snippets of search results received from the Yahoo! search engine. The freshness scores are visualized using colored bars with percentage numbers on the right-hand sides of the search results. Under the bars, the system shows the focus time periods of each search result displayed on a timeline. The earliest possible last-modification dates discussed in the previous section are indicated on the timeline as vertical dashed lines. We do not show the lines that point to dates older than 1995 - the time since when the web became popular. Also, no information is shown on the timelines of search results for which no focus time was detected (as for the 2nd and 3rd result in Figure 4) or for documents with little textual content (less than 20 terms).

Some additional information for each search result is also presented on the right-hand side of the timeline. It consists of the top 5 terms characterizing the cluster which is the most similar to a given search result and the titles and links of two complementary news articles. These news articles are selected in such a way that they discuss events which are highly recent and popular, and at the same time are least similar to the page content. The idea behind the complementary news articles is to offer users additional information to particular search results, especially the information that is poorly represented in target pages. Note the difference between the concept of complementary content and "similar" or "related pages" provided by current search engines such as the Google.

At the top part of the system interface we show the aggregated information on the clusters in order to facilitate data understanding. First, there is a timeline displaying the time periods of all the distilled events related to the query. These aggregated results should help users better understand the context of individual search results and serve for comparison purposes. For example, users can see how many important events a particular search result covers by comparing its timeline with the aggregated timeline on the top of the system interface.

In addition, we also show the average quality score calculated over all the clusters, which determines the goodness of clustering. Clicking on the top part of the interface displays the auxiliary window shown in gray on the right-hand side in Figure 4. It contains the detailed information on all the clusters including their

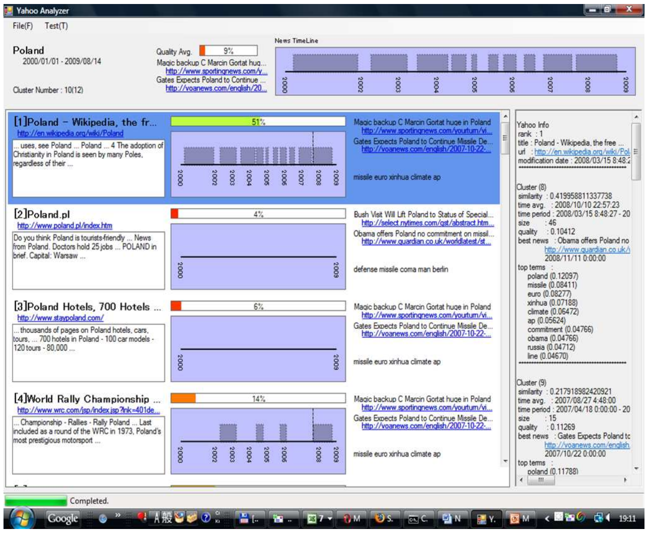
[7] We ignore cases of reporting future events by page authors.

[8] The choice is not only limited to the Google News Archive as any other online news aggregator could be used instead.

time periods, quality scores, cluster size, top terms and representative news articles.

When a user clicks on a given search result additional data related to the search result appears in the auxiliary window on the right-hand side. It contains the information on the similarity of the page with all the distilled clusters, cluster quality scores and other cluster-related information such as the top terms and representative news articles. Users can see which clusters are most similar to a given search result and how good and how large they are.

The objective of this system as well as of the whole idea of measuring freshness is not to substitute the relevance measure neither to change the search results ranking. It would be risky to claim that a page containing information on recent events is more relevant than the one without such information as it is highly dependent on the actual user needs. We propose independent metrics to be visualized next to search results for supporting users in their search experience. These metrics could be, however, also incorporated into overall page quality evaluation.



Poland
2000/01/01 - 2009/08/14

Cluster Number : 10(12)

Quality Avg. 9%
Magic backup C Marcin Gortat hug...
http://www.sportingnews.com/y...
Gates Expects Poland to Continue ...
http://voanews.com/english/20...

[1]Poland – Wikipedia, the fr...
http://en.wikipedia.org/wiki/Poland
... uses, see Poland ... Poland ... 4 The adoption of Christianity in Poland is seen by many Poles, regardless of their ...
51%
Magic backup C Marcin Gortat huge in Poland
http://www.sportingnews.com/yourturn/vi...
Gates Expects Poland to Continue Missile De...
http://voanews.com/english/2007-10-22-...
missile euro xinhua climate ap

[2]Poland.pl
http://www.poland.pl/index.htm
Do you think Poland is tourists-friendly ... News from Poland. Doctors hold 25 jobs ... POLAND in brief. Capital: Warsaw ...
4%
Bush Visit Will Lift Poland to Status of Special...
http://select.nytimes.com/gst/abstract.htm...
Obama offers Poland no commitment on missil...
http://www.guardian.co.uk/worldlatest/st...
defense missile coma man berlin

[3]Poland Hotels, 700 Hotels ...
http://www.staypoland.com/
... thousands of pages on Poland hotels, cars, tours, ... 700 hotels in Poland - 100 car models - 120 tours - 80,000 ...
6%
Magic backup C Marcin Gortat huge in Poland
http://www.sportingnews.com/yourturn/vi...
Gates Expects Poland to Continue Missile De...
http://voanews.com/english/2007-10-22-...
missile euro xinhua climate ap

[4]World Rally Championship ...
http://www.wrc.com/jsp/index.jsp?lnk=401de...
... Championship - Rallies - Rally Poland ... Last included as a round of the WRC in 1973, Poland's most prestigious motorsport ...
14%
Magic backup C Marcin Gortat huge in Poland
http://www.sportingnews.com/yourturn/vi...
Gates Expects Poland to Continue Missile De...
http://voanews.com/english/2007-10-22-...
missile euro xinhua climate ap

```
Yahoo Info
 rank  : 1
 title : Poland - Wikipedia, the free ...
 url   : http://en.wikipedia.org/wiki/Pol
 modification date : 2008/03/15 8:48:2
*****************************************

Cluster (8)
 similarity  : 0.419958811337738
 time avg.   : 2008/10/10 22:57:23
 time period : 2008/03/15 8:48:27 - 20
 size        : 46
 quality     : 0.10412
 best news   : Obama offers Poland no
               http://www.guardian.co.uk/
               2008/11/11 0:00:00
 top terms   :
     poland (0.12097)
     missile (0.08411)
     euro (0.08277)
     xinhua (0.07188)
     climate (0.06472)
     ap (0.05624)
     commitment (0.04766)
     obama (0.04766)
     russia (0.04712)
     line (0.04670)
*****************************************

Cluster (9)
 similarity  : 0.217918982420921
 time avg.   : 2007/08/27 4:48:00
 time period : 2007/04/18 0:00:00 - 20
 size        : 15
 quality     : 0.11269
 best news   : Gates Expects Poland to
               http://voanews.com/english
               2007/10/22 0:00:00
 top terms   :
     poland (0.11788)
```

Figure 4 Snapshot of the prototype system.

5.2 Evaluation

The experiments were done on the 1st November 2010. We have constructed a set of 50 test queries which represent the names of persons, countries, organizations, etc. They are shown in Table 2.

Table 3 displays the list of freshness-indicating and obsoleteness-indicating expressions used for creating extended queries in Section 3.2. The former contain also dates that indicate time near the experiment date, while the latter contain past dates and some non-temporal expressions that should return general content related to queries.

We recorded freshness scores using our system for the top 20 web search results for each query. Pages containing less than 20 terms were removed. In total, 912 pages were evaluated. The time frame for each query was set to t_{beg}=1/1/2010 and t_{end}=1/11/2010, while the maximum number of news articles was $N=100$ and the number of partitions was $R=10$. The minimum and maximum numbers of clusters were set to 3 and 15, respectively. Other parameters were set to their default values.

We also calculated the combined freshness score that merges normalized freshness scores from Sections 3.1 and 3.2 using mixing parameter ε.

$$F_p^{both} = \varepsilon F_p^{time} + (1 - \varepsilon) F_p^{notime} \qquad (11)$$

To create ground-truth data we manually assigned freshness scores to each page using the 6-point Likert scale. The score equal to 1 indicates that a page does not report any recent events related to the query that occurred within the input time frame, while the score equal to 6 means the page describes most of the recent events related to the query.

We compared the system results with the original ranking generated by the Yahoo! search engine treated as a simple baseline. By this, we wished to reject hypothesis that our method is not useful. This hypothesis would be true if search results ranking closely followed the manual judgments and achieved better results than the ones for our methods.

In Table 4 we show normalized Discounted Cumulative Gain (nDCG) scores calculated at the top 3, 5, 10, 15 and 20 results over all the queries. On average, both proposed methods produce better results than the Yahoo! search engine when one considers the recency dimension of web pages. This indicates that our method could be used to complement search engines in this aspect. The best results are obtained for the combined approach with parameter ε=0.75. It suggests that the combination of both the approaches can yield better results than the methods used alone.

Table 2 Queries.

france	microsoft	nissan	playstation 3
pakistan	apple	nvidia	xbox 360
iraq	google	amd	disneyland
israel	yahoo!	china	twitter
syria	sony	mcdonald	youtube
thailand	nec	nintendo	bill gates
india	imac	brad pitt	tom cruise
toyota	ipad	johnny depp	ipod
iran	steve jobs	panasonic	iphone
north korea	walt disney company	maria sharapova	cristiano ronaldo
germany	obama	david beckham	ronaldinho
tiger woods	fernando alonso	paris hilton	britney spears
poland	russia		

Table 3 Expressions used for extending queries in the approach that uses non-timestamped documents.

Freshness-indicating exp.	Obsoleteness-indicating exp.
recent, latest, last, lately, recently, latest new, recent news, latest events, recent events, 2010, september 2010, october 2010	overview, archive, old, past, history, historical, about, 2004, 2003, 2002, 2001, 2000

Table 4 Evaluation results.

	F^{time}	F^{notime}	F^{both} ($\varepsilon=0.75$)	*Baseline*
nDCG$_3$	0.61	0.52	**0.63**	0.45
nDCG$_5$	0.64	0.56	**0.66**	0.50
nDCG$_{10}$	0.72	0.65	**0.74**	0.59
nDCG$_{15}$	0.78	0.73	**0.80**	0.67
nDCG$_{20}$	0.80	0.76	**0.81**	0.75

6. CONCLUSIONS

The lack of strict publishing barriers and poor quality control of web content demand more refined approaches towards analyzing online information. In this paper we have proposed measuring and visualizing content freshness of web search results. We achieved this by synchronizing the information in web search results with the one in relevant news articles and by collecting fresh content associated with user query from the web. We have also demonstrated how our approach can estimate document focus time, a measure used for mapping page content on timeline. We evaluated our approach on a set of object queries and presented the proof-of-concept prototype system.

In the future, we plan to provide method for automatically detecting the length of time frame *T* that defines the scope of recency analysis. Intuitively, it should be short for fast changing topics. One possibility is to estimate its length using the results from the web mining process described in Section 3.2. In addition, we would like to investigate hierarchical clustering of web search results in order to detect the subtopics related to queries. The information on fine-grained topics would help to better estimate recency of particular pages.

7. ACKNOWLEDGMENTS

We would like to thank Yasunobu Kuwada, Soungwoong Yoon and Kyohei Seto for their help with this work. This research was supported the National Institute of Information and Communication Technology, Japan, the MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the Microsoft IJARC CORE6 Project, "Mining and Searching the web for Future-related Information".

8. REFERENCES

[1] I. Arikan, S. J. Bedathur, and K. Berberich. Time Will Tell: Leveraging Temporal Expressions in IR. *Proceedings of WSDM 2009*, 2009.

[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, A. Tomkins. Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. *Proceedings of WWW 2004*, 328-337, 2004.

[3] T. Calinski and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3(1), 1-27, 1974.

[4] W. Dakka, L. Gravano and P. Ipeirotis. Answering General Time-Sensitive Queries. *TKDE 2010*.

[5] F. Diaz. Integration of News Content into Web Results. *Proceedings of WSDM 2009*, 182-191, 2009.

[6] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, F. Diaz: Towards recency ranking in web search. *Proceedings of WSDM 2010*, 11-20, 2010.

[7] A. Jatowt, Y. Kawai and K. Tanaka. Estimating News Coverage of Web Search Results, *Proceedings of the 2010 IEEE/WIC/ACM WI 2010*, 460-463, 2010.

[8] A. Juffinger, M. Granitzer and E. Lex. Blog credibility ranking by exploiting verified content. *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW 2009)*, 51-58, 2009.

[9] N. Kanhabua and K. Nørvåg. Determining Time of Queries for Re-ranking Search Results. *Proceedings of ECDL 2010*, 261-272, 2010.

[10] R.D. Lankes. Credibility on the Internet: Shifting from Authority to Reliability. *Journal of Documentation*, 64(5), 667-686, 2008.

[11] X. Li, W. B. Croft, Time-based language models. *Proceedings of CIKM 2003*, 469-475, 2003.

[12] M. Meola. Chucking the Checklist: A Contextual Approach to Teaching Undergraduates Web-Site Evaluation. *Libraries and the Academy*, Vol. 4, No. 3. 331-344, 2004.

[13] M.J. Metzger. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *JASIST*, 58(13), 2078-2091, 2007.

[14] D. Metzler, R. Jones, F. Peng, R. Zhang. Improving search relevance for implicitly temporal queries. *Proceedings of SIGIR 2009*, 700-701, 2009.

[15] Pew Internet & American Life Project. Online Health Search 2006,http://www.pewinternet.org/~/media/Files/Reports/2006/PIP_Online_Health_2006.pdf

[16] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management: an International Journal*, 24:5, 513-523, 1988.

[17] A. Scholz-Crane. Evaluating the Future: A Preliminary Study of the Process of How Undergraduate Students Evaluate Web Sources. *RSR; Reference Services Review*, 26(3/4), 53-60, 1998.

[18] M. Toyoda and M. Kitsuregawa. What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots, *Proceedings of WWW 2006*, 233-241, 2006.