Estimating News Coverage of Web Search Results

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Yukiko Kawai
Kyoto Sangyo University
Kyoto, Japan
kawai@cc.kyoto-su.ac.jp

Katsumi Tanaka
Kyoto University
Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

***Abstract*—The abundance of content on the web and the lack of quality control require more refined approaches in analyzing online information. In this paper, we propose evaluating the extent to which web search results cover important and recent news related to real-world objects. Our method allows for identifying search results that provide comprehensive overviews of major events related to user queries or that contain most recent information.**

Keywords-information credibility, content freshenss, news coverage

I. INTRODUCTION

The web has started to play an important role for many people in delivering information related to their personal and professional lives. However, the lack of publishing barriers and poor quality control call for more refined information analysis of online content. Many pages on the web contain only partial or incomplete information. Important information may be missing due to various reasons such as authors' lack of necessary knowledge, time constraints, or even may be purposely omitted.

Also, as keeping content up-to-date requires certain effort and time, it is relatively "easy" for authors to fail updating content despite the occurrence of new, important news related to the topics of their pages. The lack of important or recent information in pages is actually often regarded as a kind of "soft" or "justifiable" quality violation in comparison to deliberately misguiding by altering and manipulating certain information. Note, however, that despite the somewhat lax perception of this quality violation, its consequence can be still quite harmful to users.

The solution to this problem is an automatic evaluation of the correspondence of search results to relevant news events as a form of support for users in their information verification and gathering task. The method that we propose calculates the extent to which web search results cover important and recent events related to user queries. This kind of approach bears resemblance to the user behavior known as corroboration [5]. In order to verify given information users tend to utilize external sources, especially, the high quality ones, rather than solely evaluating internal qualities of web sites.

Our assumption is that pages about real-world objects which cover many significant events related to these objects (e.g. Wikipedia[1] articles) are often high quality documents. Readers of such pages can receive the complete overview of the objects' current states, often, according to multiple different aspects. Although some web search engines already catch news queries and display news articles in the rankings [3], there is still no way to find exhaustive web pages that would provide comprehensive overview of salient events related to the searched objects.

In information quality research, validity and correctness of information are closely related to its freshness. For pages covering relatively dynamic topics, information conflict could occur between their content and the latest related events. It has been found that there are many abandoned documents in the web and that lots of pages provide obsolete information [1]. Therefore in our approach we also evaluate the information currency of web search results. Note that pages' currency cannot be simply assessed by checking timestamps as they may be unavailable and may not necessarily indicate the temporal validity of the content.

The remainder of this paper is structured as follows. In the next section we discuss the related work. In Section 3 we present our method for estimating the news coverage of search results. In Section 4 we show the experimental evaluation. Finally, we conclude the paper in the last section.

II. RELATED WORK

In information quality theory, accuracy, authority, objectivity, currency and coverage of information are the most frequently used evaluative criteria [6]. For example, the objectivity involves determining whether the web content represents facts or opinions, while currency is the measure of how up-to-date the content is. 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 are required to investigate different criteria of encountered content, usually, by answering fixed set of questions (e.g. "does the site provide information about when the content was last posted or updated?").

In contrast to the approaches that rely solely on target content, Lankes [5] argued that users often seek for commonalities and coherence among multiple information

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

sources in order to gauge the extent to which they can rely on particular information. This behavior called corroboration fits well into the web environment where plenty of different information sources are available for comparison.

The problem of the content quality has been recently especially investigated in collaborative environments such as Wikipedia. The usual approach relies on the calculation and management of trust between collaborating participants [8].

From the documents' freshness viewpoint, Toyoda and Kitsuregawa [7] proposed a novelty measure for estimating approximate creation dates of pages. Their approach is a recursive one, similar to the *PageRank* and exploits novelty scores of linking pages. With an opposite objective, Bar-Yossef et al. [1] employed link-structure analysis on the web for identifying decayed web sites. Juffinger et al. [4] proposed calculating the coverage of news in blogs and the level of their synchronization as a measure of blog credibility. However, to the best of our knowledge none of the related works proposed verifying information coverage and currency in web pages using news articles.

III. METHOD

Figure 1 shows the rough outline of our approach. When a user issues a query to a search engine, it is analyzed for the occurrence of named entities such as object, person place or other names using a standalone recognition tool. The extracted object name 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 then retrieved. Using this data we detect main events that occurred within the required time span. Next, the distilled events are compared with the content of the returned web search results in order to find the degree to which they are covered in the web pages. The calculated page scores can be then shown to users to support them in choosing high quality search results.

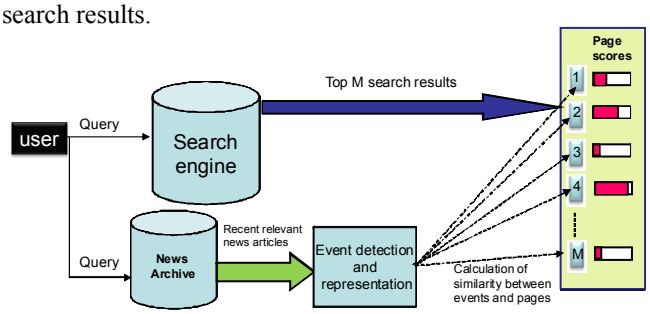

Figure 1. Outline of approach.

A. Data Collection

First, a user's query is sent to a conventional search engine from which M results are downloaded. The part of the query that describes an object is forwarded to an online news archive. It is there transformed into a series of sub-queries spanning a predefined time period $T=[t_{beg},t_{end}]$ (by default, t_{end} is equal to the query issuing time), each with a temporal constraint. The initial time period T is partitioned into R number of continuous and non-overlapping time units, which serve as temporal constraints for the sub-queries.

We receive up to N/R results for each sub-query, where N is the pre-specified maximum number of results from the news archive. The returned results are composed of news articles' snippets and their titles. For efficiency and due to access restrictions, we use only the snippets and titles here. For brevity, from now on, we simply call the news articles' snippets as news articles.

B. Event Detection

Next, we extract textual content and the attached timestamps of collected news articles. Then, events are detected through clustering. We assume here that a single news article covers only one event as a main topic of its content. For clustering, we use the *k-Means* method. The distance is measured by calculating Euclidean distance between the news articles' feature vectors constructed according to the bag-of-terms concept after a prior elimination of 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:

$$tfidf_{a,j} = \frac{n^{title}_{a,j} + 0.5 * n^{content}_{a,j}}{size(d_j)} * \log\left(\frac{N}{N_a} + 1\right) \quad (1)$$

Here, $n^{title}_{a,j}$ and $n^{content}_{a,j}$ are, respectively, the numbers of occurrences of a term a in the title and in the content of a given news article d_j; $size(d_j)$ denotes the total number of terms in d_j; N_a is the number of documents which contain the term a within N where N is the total number of collected news articles. We use weights *0.5* and *1* to give lower scores to the occurrences of terms in news articles' content when compared to the ones inside their titles.

The optimal cluster number, k, for the k-means algorithm is calculated by applying *Calinski-Harabasz* method [2]. 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). Formally, the method selects the number k ($k \geq 2$) of clusters by maximizing the following function,

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

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 drop very small clusters (i.e., the ones having the number of members less than 3 news articles). In addition to decrease the effect of noisy clusters we remove low quality clusters using the formula derived from Equation 2. According to it, the high quality clusters

describe topics that are different from other clusters and also contain topically similar documents.

The resulting clusters are represented by centroid vectors in order to be compared with the content of web search results.

C. News Coverage

In this section we describe the way in which we calculate the news coverage of web search results. First, we estimate the importance and recency of the detected events. *Event importance* is measured as the relative news attention concerned with the event and is modeled as the size of its underlying cluster.

Definition 1. *The **importance of event l** related to query q is the number of news articles that support l.*

We express the event's importance as the normalized membership count of the cluster representing the event.

$$\mathrm{Im}_{l} = \frac{csize(l)}{\underset{1 \le i \le k}{Max}(csize(i))} \quad (3)$$

On the other hand, *event recency* refers to the relative age of the event from the viewpoint of query issuing time.

Definition 2. *The **recency of event l** related to query q is the time period between the event occurrence time and the query issuing time.*

We express the event recency 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 its underlying cluster.

$$\mathrm{Re}_{l} = \frac{t_l - t_{beg}}{t_{end} - t_{beg}} \quad (4)$$

We then combine *event importance* and *recency* into a single *event score*.

$$Score_{l} = \beta * \mathrm{Im}_{l} + (1 - \beta) * \mathrm{Re}_{l} \quad (5)$$

Next, we construct feature vector of each web search result based on its downloaded content after removing stop words. The feature vector is computed using the standard *tf-idf* weighting scheme.

The event scores calculated in Equation 5 are then used as weights in measuring the similarity between the page's content and the detected events represented by clusters. We use here cosine similarity, $\cos(v_p, v_l^{event})$ between the page's feature vector, v_p, and the event vector, v_l^{event}.

$$Sim(p, l) = Score_{l} * \cos\left(v_p, v_l^{event}\right) \quad (6)$$

Finally, we define the notion of *news coverage* of a page.

Definition 3. *The **news coverage of page p** in relation to query q is the degree to which the page covers important and recent events related to q.*

We express the news coverage by summing the cosine similarity scores between the page and all the detected events.

$$EC_{p} = \frac{1}{k} \sum_{i=1}^{i=k} Sim(p, i) \quad (7)$$

Pages with the high value of news coverage are deemed to reflect salient events in their content. Note that depending on parameter β we can control the extent to which the news coverage is determined by either important or recent events. If the newness of information is of high importance, as common in rapidly changing topics, then β should be set to a small value. On the other hand, if one is interested in pages containing important events related to the query then β should be close to one. The importance and recency of information on the web are crucial factors of information quality.

IV. EXPERIMENTS

In this section we discuss the evaluation of our approach. We have built a system using Microsoft .NET. We used Yahoo! Search for obtaining web search results and Google News Archive[2] as a source of news articles.

As a set of test queries we have used the top Google Zeitgeist News queries[3] issued from January 2006 to April 2007. From this set we have selected queries containing the names of real world objects or concepts (i.e., persons, countries, companies, products, teams, groups, etc.). We then arranged these queries into three groups: "persons", "places" and "others" (i.e., objects, companies, products, concepts). The numbers of queries in the groups were 60, 14 and 19, respectively.

We recorded news coverage scores using our system for the top 10 web search results for each query. In this process pages containing less than 20 terms were omitted. In total, 828 pages had been tested (561 from "persons" group, 117 from "places" and 150 from "others"). The evaluation was done on 14/8/2009 and the time frame for each query was set to t_{beg}=1/1/2000 and t_{end}=14/8/2009. The maximum number of news articles was *N=100* and the number of partitions was equal to *R=5*. The system calculated news coverage scores for *β=0* (recency) and for *β=1* (importance).

Next, two users evaluated the pages returned in search results. They were allowed to consult external information sources during the evaluation. For each page importance and recency scores were assigned using the 5-point Likert scale. For example, the importance score equal to 1 indicates that a page does not report any important events related to the query that occurred within the input time frame, while the recency score equal to 5 means the page describes most of the recent events related to the query. The inter-evaluator

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

[3] http://www.google.com/press/zeitgeist_monthly.html

correlation of scores was 0.84 for the importance and 0.74 for the recency scores.

We checked how the scores generated by our system correlated with the average scores assigned by the evaluators using the correlation coefficient. Table 1 shows the results for different query groups.

TABLE I. CORRELATION BETWEEN THE SYSTEM SCORES AND THE AVERAGE EVALUATION SCORES.

	Persons	Places	Others	***Avr.***
Importance ($\beta=1$)	0.59	0.71	0.52	***0.59***
Recency ($\beta=0$)	0.60	0.62	0.47	***0.57***

The correlation values are relatively high indicating the positive correspondence of the system's scores with the manual scores. The results for recency are slightly lower than the ones for the importance especially for the "places" query group. We think that this may be because the information on important events related to the queries may be more popular and well-known than the one on recent events; hence, its evaluation may be easier.

Another issue related to the recency is that many pages with high recency scores show only the latest news that occurred within the last several days or weeks, while the time frame of the score calculation was set to less than 10 years.

The results related to the category "others" were also weaker than the ones related to the "persons" and "places" categories. This may be because it is relatively difficult to determine the recent or important news for general type objects (e.g., wii) or concepts (e.g., global warming) when compared to the events on particular persons or places. Also, many search results on objects (e.g., ferrari, xbox) contain much multimedia and little textual content.

We then compared the system results with the original ranking generated by Yahoo! search engine treated as a simple baseline. Obviously, web search engines use multiple features to rank search results. However, we did not know any other similar system that could serve as a baseline here. In addition, we wished to reject the hypothesis that our method is not useful, as it would be in the case when search results closely followed the evaluators' judgment.

Since web search engines do not return any numerical scores associated with search results, we have used here the Spearman Correlation Coefficient which measures the correlation between two lists of ranked items. For this, we had to convert the evaluator scores to ranks. Table 2 shows the results. On average the proposed system produces better results than Yahoo! search engine considering the recency and importance dimensions of web pages.

TABLE II. COMPARISON OF THE CORRELATION OF SEARCH ENGINE RANKS (BASELINE) AND THE SYSTEM-PRODUCED RANKS (SYSTEM) WITH THE MANUAL EVALUATION RANKS.

	Persons	Places	Others	***Avr.***
Importance (system)	0.16	0.27	0.29	***0.20***
Importance (baseline)	0.08	0.11	0.05	***0.08***
Recency (system)	0.30	0.50	0.31	***0.33***
Recency (baseline)	0.19	0.30	0.1	***0.19***

V. CONCLUSIONS

In this paper we have proposed calculating news coverage of web search results as a measure for their quality evaluation. We achieved this by synchronizing the information in web search results with the one in relevant news articles. We introduced two measures of: importance and recency of events described in the content of web search results. Naturally, the approach presented in this paper cannot fully cover all the aspects and dimensions of quality judgment. Therefore it should be regarded as one of supporting tools in the evaluation of page quality constructs.

ACKNOWLEDGMENT

This work has been supported by Microsoft IJARC CORE6 project, "Mining and Searching Web for Future-related Information" (the first author is Microsoft IJARC Fellow) and by the National Institute of Information and Communications Technology, Japan.

REFERENCES

[1] 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*, pp. 328-337, 2004.

[2] T. Calinski and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, vol. 3, no.1, pp.1-27, 1974.

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

[4] 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),* pp. 51-58.

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

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

[7] 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*, pp. 233-241, 2006.

[8] B. Vuong, E-P. Lim, A. Sun, M-T. Le, and K. Chang. On Ranking Controversies in Wikipedia: Models and Evaluation. *Proceedings of WSDM 2008*, Stanford, California, USA, pp. 171-182, 2008.