# QUERY-BASED DISCOVERING OF POPULAR CHANGES IN WWW

**Adam Jatowt**
University of Tokyo
Tokyo, Japan
jatowt@miv.t.u-tokyo.ac.jp

**Khoo Khyou Bun**
University of Tokyo
Tokyo, Japan
kbkhoo@miv.t.u-tokyo.ac.jp

**Mitsuru Ishizuka**
University of Tokyo
Tokyo, Japan
ishizuka@miv.t.u-tokyo.ac.jp

**ABSTRACT**

This paper presents the method for retrieving and summarizing changes in topics from online resources. Users often want to know what are the major changes in their areas of interest. Usually, change detection applications are based on predetermined sets of web pages. User needs to provide the addresses of web pages in order to receive recent information about occurring changes. Our approach involves creation of dynamic web collection for a given area of user's interest. Such collection would contain informative and up-to-date resources. Periodically, we monitor the set of pages in search for new textual data and detect significant terms to extract sentences reflecting popular changes within every period. Since many web pages can be static over long time, we propose a method for evaluating how up-to-date a web page is in context of a given topic. Each WWW page is scored according to the frequency and contents of its changes. The most valuable pages form a base for next change summarizations. Additionally we expand the web collection to include new, valuable resources by finding pages, which have similar characteristics to the top-scored pages from the collection.

**KEYWORDS**

Change tracking, Web collection, Information retrieval, Web mining, Summarization, Text mining

## 1. INTRODUCTION

World Wide Web has become the major information source for many people. Large quantity of easily available data and the heterogeneous nature of web pages make information retrieval a challenging task. User interested in a given topic is provided with many information resources of various nature, content and level of reliability. Besides, the simplicity of publishing the information results in rapid and unpredictable change of web site contents. User is not aware which of pages contain correct, up-to-date information. Our proposal of the system would assist users in searching for new, up-to-date information about given topics.

There are several human-edited topic directories of web pages like Yahoo! or Open Directory Project. Unfortunately, such web page lists cover only a small portion of WWW and often do not produce accurate results in response to a user query. In addition, it takes much time before the changes of a web site can be

detected and the directory updated. On the other hand, it is difficult and time-consuming task for user to manually determine hundreds of topic-related pages and track each of them for meaningful changes. Our first objective is to develop a system, which provided with a query, would generate responding collection of valuable WWW pages. Since web page is a dynamic resource with varying characteristics, we should constantly examine its features. The notion of "up-to-date-ness" of a web page describes how fast textual changes appear and how common they are in comparison to the rest of documents from the collection. We assume that popular changes occurring in many WWW pages convey useful information concerning user's area of interest. Web pages, which do not change frequently enough or whose changes do not contain updated information will be gradually ignored by our system. To ensure high quality of the collection we search for new resources and periodically include them into our set. Such web pages should possibly have similar qualities as the top-scored pages from the collection. Our method involves in-links counting in a way similar to Cocitation algorithm [Dean and Henzinger, 1999]. Frequency of out-links from the group of pages linking to the top-scored pages is computed in order to find new resources.

To sum up, we would like to maintain topically related set of pages and then to monitor them periodically in order to retrieve and analyze common changes. Such set would consist of frequently updated WWW pages, which corresponds to user's topic. The collection size should also be minimal for low computational cost.

In the remainder of this study, we describe our efforts towards constructing automatic way for change extraction and summarization from online resources at regular time intervals. In next Section we review some similar existing systems as well as the work from related research areas. Section 3 discusses the methodology used to fulfill our objective. In Section 4 the results of experiments are presented. We conclude and propose future research directions in the last Section.

## 2. RELATED WORK

This research has its roots in news tracking, change detection, and web collection constructing. In last years several attempts have been made to construct systems for developing automatic reports of recent events [McKeown et al., 2002; Google News; Radev et al., 2001a]. Usually these applications search for commonly discussed topics on predetermined set of web pages like well-known newswires. Google News tracks about five thousand news sources. User can issue a query and read recent articles related to his interest. However Google News does not generate its own summaries of news but rather displays links to variety of sources discussing given event. On the other hand, other systems like Newsinessence [Radev et al., 2001a] or Newsblaster [McKeown et al., 2002] provide summaries of popular recent events. Anyway, there is a need for an application that could summarize information from any type of web pages, hence not only newswires. By maintaining web collections on specific topics one can find and include informative resources, which are devoted to a topic and discuss it in variety of ways. WebInEssence [Radev et al., 2001b] is an example of such system, attempting to provide summaries generated from web collections. Differently to this proposal, our application is focused on summarizing changes in text content of web collections.

There are several systems designed for detecting changes on WWW pages. Any features like link structure, text, pictures or layout can be tracked. Usually these systems request user to provide the addresses of web page or set of pages to be monitored [Liu et al., 2000; Changedetect]. User is expected to specify what kind of changes are to be tracked. However it may be difficult to predict the type of changes that will occur. In result these systems notify about all changes, sometimes also meaningless ones. In spite of this fact, there was almost no research done on extraction and summarization of meaningful, textual changes from collection of web pages. The exception here is ETTS system [Khoo and Ishizuka 2001].

Much research was devoted to building web collections [Carriere and Kazman 1997; Spertus 1997; Marchiori 1997; Pirolli, Pitkow and Rao 1997]. Carriere & Kazman construct collections by issuing query to search engine and expanding obtained results by exploiting link structure. Others attempt to collect pages by crawling sites starting from base seeds. Kleinberg proposed the idea of finding *authoritative* and *hub* pages in WWW where a good *authority* page is linked by many hub pages [Kleinberg 1998]. In our system we collect web pages by expanding results obtained from search engine and by finding related web pages using method similar to the Cocitation algorithm [Dean and Henzinger, 1999].

# 3. SYSTEM ARCHITECTURE

Conceptually, we can break down our methodology into three phases: (a) web collection forming, (b) change tracking and (c) collection expanding. The last two steps are repeated periodically with specified monitoring frequency.

## 3.1 Web Collection

We collect WWW pages relevant to user's interest by issuing query to search engine. First, base set of web pages is created by fetching 200 highest-ranked hits generated by search engine. Next we examine out-links of the pages. After ranking links by their frequency of occurring, we download up to 50 the most common web pages. Common links enhance probability of finding sites, which contribute to the topic of query. In the last step we repeat the process for pages two clicks away from the base set. Our system checks also for potential duplicate web pages, which should be discarded.

For tracking changes in text content, textual data must be extracted from downloaded HTML files and segregated in sentence format. The following steps are conducted:

- Extracting text from HTML files: removing links, images, meta-information, html markup language or other scripts from web pages.
- Sentence selection: only text strings starting with uppercase and ending with sentence delimiters are extracted.
- Text processing: discarding punctuation, converting text to lowercase.

Next, we remove common words via stop-list and subject remaining terms to stemming. Finally, we rate pages according to their contents. Each WWW page has a score computed by dividing its sum of term weights by the number of all terms occurring within the document. Term weights are calculated by multiplying collection frequency by the exponent of the document frequency of a term (Equation 1).

## 3.2 Change Tracking

Periodically, we extract changes from the web collection by comparing old and new versions of each page. Such extraction can be carried according to arbitrarily specified time schedule. Files containing new changes are created for the web pages with high positions in the ranking list of collection. We decided to process each time only 150 web pages, which have the highest score. The extraction of changes is done by sentence comparison. If a given sentence does not exist in a previous version of a web page then it is regarded as a new one and included into change file. The question can be raised here about the case of slightly changed sentences like: grammatical, spelling corrections or order reversing. It would be possible to loose more comparison rules for example: by disregarding word order or allowing some degree of difference in words usage. However such cases seem to seldom happen. We assume that changes usually occur in a form of new sentences added or substituted in place of old ones.

As a next step, we pick words from new sentences and subject them to stemming. Then a weight $W$ denoting a score of "popularity" for each term is computed by following weighting scheme (Equation 1).

$$W_j = CF_j \times \exp(\frac{n_j}{N})$$

Equation 1. Term weight

$CF_j$ is a collection frequency of a term $j$ ,that is the number of times a term appears in the collection. $N$ and $n_j$ are respectively a number of all documents in the collection and a number of web pages containing the term $j$ . In result we obtain ranked list of the most common terms reflecting recent changes.

## 3.3 Collection Extension and Summarization of Results

Some web pages change in short and regular time intervals with new meaningful contents being posted. Including such pages into web collection would improve quality of subsequent change summaries. Thus it is desirable to select them and exploit for information capture. To do so, we assign a dynamic score $S_t$ to each page from the web collection for every time $t$ when we track changes. As was said before, periodically, we obtain a list of common terms from all changes for one period. Therefore for a given web page, it is possible to describe the contents of changes in terms of "commonness" simply by summing weights of all new terms on the page and dividing the sum by number of these terms. However, in case of minimal changes such as the addition of one or few sentences with high scored words, we still would obtain a relatively high score. In order to minimize this effect, it is necessary to give more significance to texts containing higher number of words. We do so by multiplying total weight of changes occurring in a web page by logarithm of number of new terms $N_d$ (Equation 2).

$$S_t = \frac{\sum_{j=1}^{N_d} W_j \times (1 + \log N_d)}{N_d}$$

Equation 2. Dynamic score - weight of changes for a page

Another measure of quality of a WWW page is its frequency of changing. Therefore to formulate a model for page "up-to-date-ness" we need to take into consideration the number of times that examined web page changed during its monitoring process. Our idea was to sum weights of all changes that occurred on the page with respect to the points of time in which they were discovered. Recent additions of new information indicate higher usefulness of the page for current summaries. Therefore, the latest changes should be scored higher then changes, which occurred sometime around the beginning of the page tracking. However, the highest scores would be assigned when meaningful changes happened at all regular intervals on a page. We propose an "up-to-date-ness" function $D$ for a web page as a sum of declining in time weights of changes.

$$D = \sum_{t=1}^{T} \frac{S_t}{(T - t + 1)}$$

Equation 3. Up-to-date-ness function

$T$ is the number of time units, which passed from the beginning of monitoring process and $t$ denotes point in time when a change score $S_t$ was assigned. Below is an example of "up-to-date-ness" function (bold line) in Figure 1 calculated for scores from Table 1.

Table 1. Scores for example of up-date-ness function

| **Time t** | **Score S** | **Time t** | **Score S** | **Time t** | **Score S** |
|---|---|---|---|---|---|
| 1 | 10 | 7 | 0 | 13 | 15 |
| 2 | 0 | 8 | 0 | 14 | 0 |
| 3 | 15 | 9 | 15 | | |
| 4 | 25 | 10 | 0 | | |
| 5 | 0 | 11 | 20 | | |
| 6 | 5 | 12 | 0 | | |

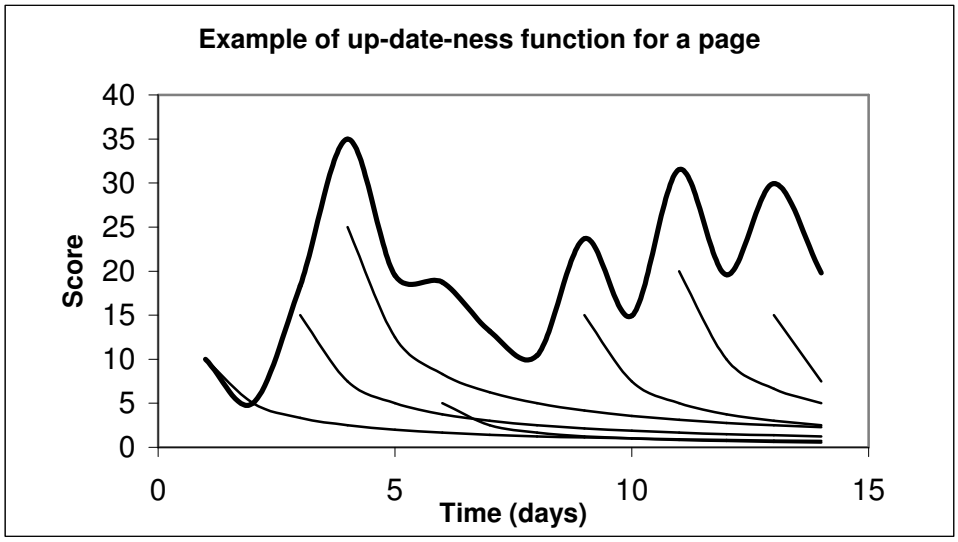


Figure 1. Example of up-to-date-ness function with scores from Table 1

Our idea of web page collection assumes that along with change extraction, new resources should be discovered and included into the collection. After the inclusion, their scores would be calculated to place the pages into appropriate positions of ranking list. In result every consecutive summary should be more accurate and based on better, up-to-date resources.

To automatically discover new informative resources, we search for web pages similar to a few top-scored pages from our ranking list. WWW pages with highest rankings in the list are examined with regards to their in-links. Common in-links are sorted according to their frequency and 200 of most frequent parent pages are downloaded and stored in a separate folder. As a next step, we extract out-links from the retrieved web pages searching for the repeated ones. Finally, the most common 10 links are fetched and included into our web collection. Consecutive change tracking phases will process also these pages. Only original pages should be admitted to the collection. Duplicate ones would not only impact results but also cause miscounting of links during next web collection expanding. Therefore we enforce originality rule by filtering duplicate pages and pages from the same sites, that is, the ones having identical domain names (ie. www.domain.com). Pages within the same site can have identical links posted ("homepage" link, "back" link etc.), which influence link counting. Thus, in result of our policy, it may happen that for the given set of the top ranked pages there will be less than 10 included pages.

After completion of change tracking, we obtain ranked list of the most common terms reflecting recent changes together with their responding weights. For every period we calculate overall weights of new sentences. Sentence weight is the sum of term scores divided by the number of terms. In this way we obtain ranked list of sentences, which convey the meaning of main changes in a given period. User is presented with the list of few most highly weighted sentences. To increase comprehension we add preceding and following sentences, which surround selected top sentences. Additionally, such extracts contain links to their host web pages to allow user read the whole document if it is needed.

As was mentioned before, a few most high-scored web pages are exploited for summarizing results. Nevertheless, we continually compute scores for the remaining pages from lower position at ranking list. Thanks to that, there is a chance to enclose them for forthcoming summaries under condition of improvement of their characteristics.

## 4. RESULTS

We present results from two experiments for queries "Iraq War" and "Harry Potter". Change tracking was performed several times during period from 18th February to 12th May for both experiments. Table 2

and Table 3 display terms with the highest weights for these respective collections at the end of change monitoring during the interval from 9th to 12th of May.

Table 2. Top terms for "Iraq War" query from 9th to 12th May

| Term | Weight | Term | Weight | Term | Weight |
|---|---|---|---|---|---|
| iraq | 567.87 | govern | 92.81 | security | 55.50 |
| post | 343.36 | country | 82. 74 | iran | 55.19 |
| iraqi | 209.41 | peace | 81.43 | official | 54.66 |
| military | 167.11 | Baghdad | 79.83 | work | 54.20 |
| american | 159.80 | international | 77.76 | sanction | 52.23 |
| force | 117.21 | president | 76.23 | city | 52.75 |
| bush | 114.91 | support | 70.45 | administration | 51.65 |
| world | 108.02 | attack | 69.50 | america | 50.05 |
| saddam | 106.25 | right | 68.72 | continue | 46.39 |
| time | 95.38 | hussein | 67.60 | know | 44.17 |
| nation | 94.00 | british | 57.67 | soldier | 43.95 |

Table 3. Top terms for "Harry Potter" query from 9th to 12th May

| Term | Weight | Term | Weight | Term | Weight |
|---|---|---|---|---|---|
| harry | 131.04 | world | 23.85 | card | 16.22 |
| potter | 127.65 | review | 23.53 | niet | 16.15 |
| book | 74.57 | movie | 21.87 | secret | 15.62 |
| rowling | 34.40 | quidditch | 20.83 | chamber | 15.48 |
| order | 29.29 | haar | 19.09 | place | 15.41 |
| site | 27.57 | game | 18.75 | timeline | 15.07 |
| hogwart | 26.51 | warner | 17.87 | magic | 12.44 |
| time | 26.39 | award | 17.15 | post | 12.27 |
| read | 25.00 | check | 16.59 | cooki | 12.05 |

In the above tables there are several words that are not related to the topic of “Harry Potter” like: “niet” or “haar”. We discovered that they have been included due to two Dutch websites about “Harry Potter” movie. For “Iraq War” query there is one term “post” which seems not to fit to the rest of terms. However, this is due to fact that “Washington Post” newspaper was widely cited in different online sources.

Tables 4 and 5 show selected sentences for changes from different time intervals. Because of space limits we present only the top-scored sentences for different time intervals.

Table 4. Top sentences for "Harry Potter” query

| Top sentences with following and preceding sentences | Weight | Page name and description | Time frame |
|---|---|---|---|
| Last summer, a Arkansas school district banned the borrowing of Harry Potter books from school libraries without the written consent of a parent. This prevoked outrage from free-speech groups and Harry Potter fans. A forth grader's parents were so miffed, they took the district to court. | 11.19 | http://thehpn.rupture.net Harry Potter Network | 29 April – 3 May |
| So insiders say even initial thoughts on casting the fourth film are still way off in the future. Bloomsbury told CBBC Newsround’s website the book – Harrius Potter et Philosophi Lapis – comes out in July as a special hardback version. At the same time, a version in Welsh will hit the shelves. | 8.10 | www.mugglenet.com Ultimate Harry Potter Site | 26 April – 29 April |
| This is undeniably a better film than the last one, which is odd, since it’s also undeniable that the second book is the weakest in the series. But perhaps it’s not so odd. The first book, after all, is at its best in the opening chapters, in which we, more or less along with Harry, discover that there’s a whole world of wizards and | 6.72 | www.imdb.com Description of “Harry potter and the Chamber of Secrets” movie | 18 February – 24 February |

| | | | |
|---|---|---|---|
| witches, that Harry is both talented in magic and an advertent celebrity, etc | | | |
| HP Fanfic author Laura Klotz has recently published a book entitled Saving the World in Your Spare Time: "The Pocket Guide to Effecting Positive Change. "This book is designed to help anyone who wants to change the world around them but isn't sure how to go about it. Although primarily geared toward adults, the chatty tone in which the book is written makes it suitable for teenagers as well. | 8.67 | www.the-leaky-cauldron.org Web Blog for Harry Potter | 3 May - 6 May |

Table 5. Top sentences for "Iraq War" query

| Top sentences with following and preceding sentences | Weight | Page name and description | Time frame |
|---|---|---|---|
| Though Saddam Hussein did not use weapons of mass destruction nor set fire to Iraq's oil fields nor attack Israel with rockets, and though the conflict was relatively short, the war has had many serious results ranging from death and destruction in Iraq to regional instability to a weakened world economy. Iraq has the world's second largest proven oil reserves. Oil industry observers predict a gold-rush of profits for the Anglo-American oil giants in the post Saddam setting. | 23.75 | www.globalpolicy.org/security/issues Iraq Crisis – Global Policy Forum | 9 May –12 May |
| Playing cards issued by the U.S. military featuring pictures of the leaders of Saddam Hussein's regime. Look for shirts, products and information to help show your support for our country and the freedom we have through a mighty god who made our nation a light to the world. | 28.62 | http://politicalhumor.about.com/cs/saddam | 3 May – 6 May |
| The Philippines has become a new front in the global war on terror, as U.S. troops head to that nation to 'disrupt and destroy' the extremist group Abu Sayyaf. In a recent lecture, Heritage expert Dana Dillon points out that although past anti-terrorism efforts in the Philippines have had some success, "Abu Sayyaf has been linked to recent terrorist bombings in the Philippines and still remains a viable threat." | 9.97 | http://www.heritage.org Heritage Foundation: Policy Research and Opinions | 21February– 24 February |
| Michael Marti told Reuters that members of a convoy returned fire after shots were fired at them from a crowd outside a U.S. command post. He said soldiers counted "potentially" two injured Iraqis. | 28.10 | www.warincontext.org War with Iraq, War with Terrorism, Middle East articles | 29 April –3 May |

Given the scope and diversity of types of pages and topics discussed for any query, it should be not surprising that results may not constitute coherent summary. Intuitively, it is very important to construct appropriate web collection with related pages. We have noticed that results are better for narrow topics, where documents tend to be more related topically with each other.

One interesting observation was that changes for "Iraq War" represent many negative opinions about the war, whose impact and negativity did not change in time even with the war being finished. Direct reason for this was the fact that there were many pages explicitly dedicated to express protests against war like www.endthewar.org, www.antiwar.com or www.againstbombing.org.

During collection expanding 68 new pages were added for query "Harry Potter" with 28 pages explicitly devoted to this topic. For "Iraq War" the numbers were respectively 57 and 88. Higher precision rate for "Iraq war" 65% comparing to "Harry Potter" query 43% can be explained by global significance and wider interest in the first topic.

Except for several newswire resources and sites devoted to both topics, there were also many message boards or blog type pages included in the collections for these experiments. "Harry Potter" collection had more percentage of such type sources added than "Iraq War" one. Although, this kind of pages can blur final results, they provide information about opinions or hot topics in opposition to information about pure facts provided by newswire sources.

Sometimes we found web pages, which instead of text, contained pictures with text hidden inside. To retrieve such contents we would need to use OCR techniques. As for now, pages of this kind are gradually excluded in subsequent steps from the summary list due to few textual changes.

## 5. CONCLUSIONS AND FUTURE WORK

We have proposed methodology for extracting and summarizing textual changes in web collections. The main idea is to score each web page according to its frequency and content of changes. The most valuable pages, according to this measure, are used for generating consecutive summaries and are exploited for finding new web pages to be included into the collection. In result we obtain set of relevant, up-to-date documents. User is periodically provided with the most popular changes in form of extracted sentences for every time frame.

Our system has several limitations on which we want to focus in the future. One problem concerns changes in the form of old and new links appearing on a web page. In case of the linked pages located on the same site, we should check them for new text and include their weights as a portion of total weight of the parent web page. Besides, since new sentences can be placed in different semantic contexts on one web page, one should take into consideration not only the content of a changed sentence but also its relation to the surrounding text. Additionally, there is a need for applying some sentence clustering and ordering technique for generating better summaries. As a next step we would like to construct a method for "history representation" of changes. Such representation would allow us to distinguish between novel and repetitive changes or to perform trend analysis.

## REFERENCES


Carriere, J. and Kazman, R., 1997. WebQuery: Searching and Visualizing the Web through Connectivity. *Proceedings of WWW6.* Santa Clara, USA.

Changedetect, http://www.changedetect.com

Dean, J. and Henzinger, M., 1999. Finding Related Pages in the World Wide Web. *Proceedings of WWW-8, the Eighth International World Wide Web Conference.* Toronto, Canada.

Google News. http://news.google.com

Khoo, K. B. and Ishizuka, M., 2001. Emerging Topic Tracking System. *Proceedings of Web Intelligence 2001*. Maebashi City, Japan.

Kleinberg, J., 1998. Authoritative Sources in Hyperlinked Environment. *Proceedings of 1998 ACM SIAM Symposium on Discrete Algorithms.* San Francisco, USA.

Liu, L. et al, 2000. WebCQ: Detecting and Delivering Information Changes on the Web. *Proceedings of International Conference on Information and Knowledge Management*. Washington, DC, USA.

Marchiori, M., 1997. The Quest for Correct Information on the Web: Hyper Search Engines. *Proceedings of WWW6.* Santa Clara, USA.

McKeown, K. et al, 2002. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. *Proceedings of the Human Language Technology Conference.* San Diego, USA.

Pirolli, P., Pitkow, J. and Rao, R., 1997. Silk from a sow's ear: Extracting usable structures from the Web. *Proceedings of ACM SIGCHI Conference on Human Factors in Computing*. Atlanta, USA.

Radev, D. et al, 2001a. NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization. *Proceedings of the Human Language Technology Conference.* San Diego, USA.

Radev, D. et al, 2001b. WebInEssence: A Personalised Web-Based Multi-Document Summarization and Recommendation System. *The Second Meeting of the North American Chapter of the Association for Computational Linguistics.* Pittsburgh*, USA*.

Spertus, E., 1997. ParaSite: Mining Structural Information on the Web. *Proceedings of the Sixth International World Wide Web Conference*. Santa Clara, USA.