Visualizing Historical Content of Web Pages

Adam Jatowt
Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
Phone: +81-75-7535969
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
Recently, along with the rapid growth of the Web, the preservation efforts have also increased. As a consequence, large amounts of past Web data are stored in Web archives. This historical data can be used for better understanding of long-term page topics and characteristics. In this paper, we propose an interactive visualization system called Page History Explorer for exploring page histories. It allows for roughly portraying evolution of pages and summarizing their content over time. We use a temporal term cloud as a structure for visualizing prevailing and active terms appearing on pages in the past.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation (e.g., HCI)**]: Hypertext/Hypermedia.

General Terms
Algorithms

Keywords
page history visualization, web archive, past web, history summarization

1. INTRODUCTION
While browsing, users create mental images of visited pages. However, the current content of pages may not necessarily provide answers to the questions about their typical topics and temporal characteristics. On the other hand, the retrospective data stored in Web archives such as the Internet Archive[1] offers the possibility for a better understanding of pages. Utilizing this data could improve various tasks on the Web such as page authoring, browsing or bookmarking. For example, page authors could better estimate the long-term relevance, freshness and usefulness of pages to which they wish to create links.

In this paper, we demonstrate Page History Explorer (PHE) - an application for providing comprehensive overviews of historical content of pages and their long-term characteristics (Figure 1). It downloads past snapshots of pages from selected Web archives and visualizes them on 2D space in which one dimension represents time and another represents the cumulative change degree. PHE also summarizes historical content of pages by indicating common and active content occurring over time. In addition, users can input arbitrary terms in order to observe page histories from keyword-related viewpoints.

In our previous study [2], we proposed Past Web Browser – a browsing application for page histories similar to traditional browsers and VCR players. It animated changes between consecutive versions of a page drawing user attention to the changed content by presenting it as appearing or disappearing effects depending on the type of change. The application helped also users to navigate through the historical content of pages. In [1] the authors demonstrated an interesting tool for visualizing popular tags of images stored in flickr[2] repository over time. Viegas et al. [3] proposed a history flow mechanism for observing the evolution of multi-authored pages such as Wikipedia pages. Their system allows for noticing the content contributed by different users and its persistence over time. In contrast to these proposals, PHE provides summarized views of the histories of arbitrary pages indicating their typical persisting content, characterizing common changes and portraying their visual and temporal characteristics.

2. SUMMARIZING PAST CONTENT
PHE downloads data from the Internet Archive for a specified URL and a time period. In order to decrease the cost, we let a user select the number of past page snapshots to be fetched. The snapshots sampled from the archive should be possibly uniformly distributed over time. Naturally, the more downloaded snapshots there are, the higher is the precision of the generated summary.

Page History Explorer uses a term cloud as a means of providing aggregate information on historical page content. We distinguish two types of such a term cloud, prevalence and activity term clouds. The former represents top prevailing terms over time, that is, terms which frequently occurred in the page in the past. We estimate the prevalence score of a term as its average frequency calculated over all the fetched page snapshots within the specified time period (Equation 1).

$$S^{prev}(a;T) = \frac{1}{T}\sum_{i=1}^{k-1}(t_{i+1} - t_i) * TF_a(t_i) \qquad (1)$$

In Equation 1, $TF_a(t_i)$ is the frequency of a term a in the page snapshot that was crawled at t_i, T is the user-specified time period, while k is the number of fetched page snapshots. The frequency of the term in every snapshot is also multiplied by the time distance between this snapshot and the next available snapshot. Terms are also filtered through stop list and stemmed.

The calculation of prevalence scores is independent of the pattern of page updates over time. If a term always appeared in the static (unchanging) content of a page (e.g., a term "copyright") it may

WWW 2008, April 21--25, 2008, Beijing, China.

[1] Internet Archive: http://www.archive.org

[2] Flickr: http://flickr.com

be of little interest to users. On the other hand, terms that are frequently added and deleted from pages may indicate important content. Thus, PHE provides also activity term cloud based on term activity scores. The activity score of a term is computed as a probability that the term occurred in added or deleted content within the time period *T*. It is estimated as the combination of the probability of a change occurrence and the probability of the term appearing in this change (Equation 2). Changes are found by comparing the content of consecutive page snapshots.

$$S^{active}(a;T) = \frac{M}{T} * TF_a^{chan}(T) \qquad (2)$$

$TF_a^{chan}(T)$ is the average frequency of a term *a* in added and deleted content over all page snapshots, while *M* is the number of snapshots that have any content change when compared to the neighboring snapshots.

In addition, we calculate term clouds for a user-specified *R* number of unit time periods within *T* in order to allow finer analysis of page history. The calculation of term scores is derived from a well-known *tf*idf* weighting scheme.

$$S_{unit}^{X}(a;T_w) = X(a;T_w) * \log\left(\sum_{j=1}^{R}\left[\frac{X(a;T_w)}{X(a;T_j)+1}\right]+1\right) \qquad (3)$$

$X(a,T_w)$ denotes here either the prevalence or activity score of a term *a* inside a given time frame T_w ($R*|T_w|=|T|$).

Lastly, PHE enables also users to input arbitrary keywords for calculating the top co-occurring terms inside historical content of pages. The equation for the calculation of the co-occurrence scores of terms is adapted from the Jaccard coefficient.

$$S_{cooc}^{X}(a,b;T) = \frac{X(a,b;T)}{X(a;T)+X(b;T)-X(a,b;T)} \qquad (4)$$

$X(a,b;T)$ is either prevalence or activity score of the common occurrence of terms *a* and *b* within *T*.

3. VISUALIZATION

PHE displays clouds of 20 top-scored terms over the specified time period in the top frame (Figure 1). In addition, term clouds for *R* number of short time periods are shown below. For every such unit term cloud up to 20 top-scored terms are displayed. The font sizes of the shown terms reflect their prevalence or activity scores depending on the user's selection.

In addition, page snapshots are converted to thumbnail images and visualized on a 2D space in which the horizontal axis represents time distance while the vertical one represents the cumulative degree of added change. Thus, the horizontal distance between a snapshot s_i and snapshot s_{i+1} depends on the temporal distance between their timestamps. The vertical distance, on the other hand, reflects the size of content that occurs in s_{i+1} while being absent from s_i. With this kind of visualization, a user can grasp the evolution of a page and spot time periods when large content additions occurred. It is also possible to roughly compare the outlook of page content and page sizes at different time points.

Clicking on any thumbnail activates a pop up window with the corresponding page snapshot. In addition, users can zoom in or out the visualized snapshots.

Upon inputting a keyword, users obtain the overview of page history from a keyword-based viewpoint. In this case, the vertical axis represents the cumulative frequency of the input keyword within added content over time. This allows for observing changes in the amount of fresh content relevant to the keyword in the past. Additionally, a term cloud consisting of top co-occurring terms is shown above.

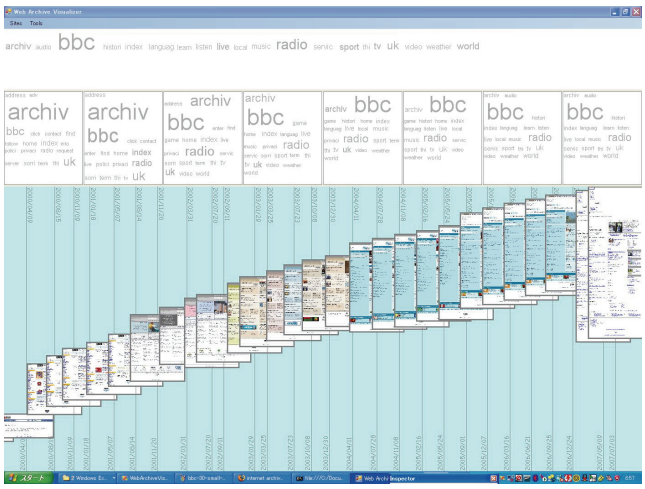

Figure 1 The history view of BBC homepage (www.bbc.co.uk).

4. CONCLUSIONS

In this paper, we proposed an approach for visualizing summarized page histories based on data extracted from Web archives. Users can obtain an overview of the long-term content and characteristics of pages. This sort of visualization can help with better understanding of pages and can add temporal context to their current content. In future, we plan to conduct user evaluation and add functionality for comparison between the histories of different pages.

5. ACKNOWLEDGMENTS

This research was supported by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless Search for Information Explosion (#18049041, Representative: Katsumi Tanaka), the Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society (Representative: Katsumi Tanaka) and by the MEXT Grant-in-Aid for Young Scientists B (#18700111, Representative: Adam Jatowt; #18700110, Representative: Yukiko Kawai).

6. REFERENCES

[1] Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P. and Tomkins, A. Visualizing tags over time. *Proceedings of the15th World Wide Web Conference*, Edinburgh, Scotland, 2006, 193-202.

[2] Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y. and Tanaka, K. Journey to the past: proposal for a past Web browser. *Proceedings of the 17th Conference on Hypertext and Hypermedia*, Odense, Denmark, 2006, 134-144.

[3] Viégas, F., Wattenberg, M. and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the Conference on Human Factors in Computing Systems*, Vienna, Austria, 2004, 575-582.