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Lastly, PHE enables also users to input arbitrary keywords for calculating the top co-occurring terms inside historical content of pages.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.

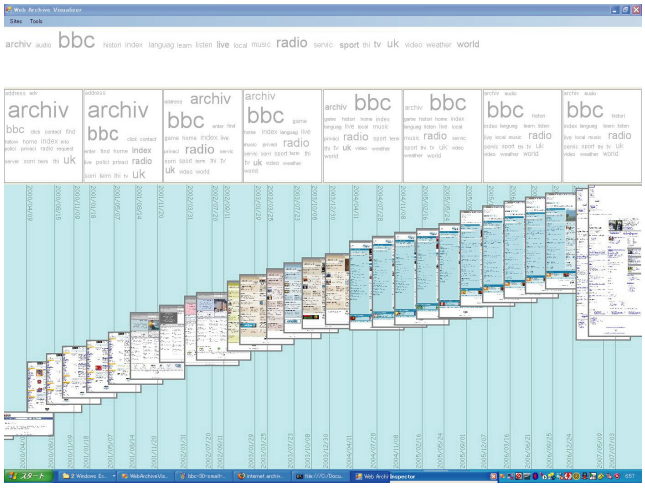

Figure 1 The history view of BBC homepage (www.bbc.co.uk).

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, #18700110).

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.