様式 C-19

科学研究費補助金研究成果報告書

平成 2009 年 6 月 2 日現在

研究種目：若手研究(B)
研究期間：2006～2008
課題番号：18700111
研究課題名（和文） 情報検索とウェブアーカイブにおけるマイニング

研究課題名（英文） Information Retrieval and Mining in Web Archives

研究代表者
Adam Jatowt
京都大学 情報学研究科 特定助教
研究者番号：00415861

研究成果の概要：

One of the most important characteristics of web pages is their capability to change content and structure in time. Many popular web pages continuously change, evolve and provide new information. Recently, Internet community recognized the need to store and preserve past content of the Web so it could be available for future use. This research project aims at utilizing web archives for inferring knowledge about temporal characteristics of web pages, for knowledge discovery and for improving user experience on the Web. State-of-the-art techniques for search, browsing and knowledge discovery from web archives are still rather limited and difficult for users to use. One of main objectives is to use information hidden in web or news archives for improving Web search, browsing and other common tasks.

交付額

（金額単位：円）

	直接経費	間接経費	合 計
2006 年度	1,300,000	0	1,300,000
2007 年度	1,000,000	0	1,000,000
2008 年度	1,300,000	390,000	1,690,000
年度			
年度			
総 計	3,600,000	390,000	3,990,000

研究分野：総合領域
科研費の分科・細目：メディア情報学・データベース
キーワード：Web 履歴, 新しい情報, Web アーカイブ

１．研究開始当初の背景

Web archives have not been sufficiently utilized by researchers, despite that past data of web documents provide valuable and informative temporal context that can significantly improve search, mining and information retrieval. Most of the research activity so far centered mostly on storage and preservation of past web page contents and little work has been done on analysis and processing of data extracted from web archives.

２．研究の目的

User-friendly searching and browsing interfaces as well as the tools for mining in web archives can be used by researchers from other fields and by users who are not computer specialists. By this research I hope to stimulate the interest of Web community in the exploitation of web archive data for personal as well as for commercial purposes. In addition, I would like to provide the foundations for improving information retrieval in the current Web by utilizing past histories of web pages.

３．研究の方法

I have used data from the largest online Web Archive, the Internet Archive, which stores past versions of Web pages crawled since 1996. Besides, I have used online news archives and archives of social bookmarks to Web pages. This data was processed and visualized for users by employing various mining, searching and visualizing techniques. Among others, I have used such methods as change detection, page clustering, binary search, burst detection, lexical pattern retrieval and graph-based visualization.

４．研究成果

(1)As a result of this project, I have investigated and provided interaction methods for page histories. The first interaction method is browsing page historical content using Past Web Browser. In short, such a browser displays past versions of a page in a slide show mode and animates detected textual changes so that new content added over time is appearing and old, deleted content is disappearing from displayed pages. In addition, users can search for the versions containing certain keywords and follow past links.

Figure 1 Past Web Browser.

(2) As the next interaction method I have built a system for visualizing page evolution as a series of snapshots of past versions displayed in 2-dimensional frame in which time is one dimension and change degree is another one (Fig. 2). Moreover, frequent terms appearing on a page over time are shown as a term cloud. This kind of visualization provides a macro-scale overview of page evolution as opposed to the micro-scale view provided by the Past Web Browser.

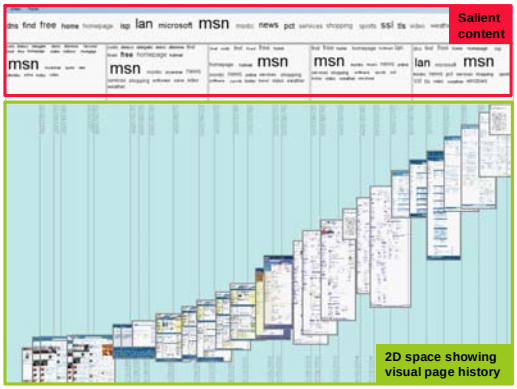

Figure 2 Example of history-based summary of an example page.

(3)I have also provided a method for detecting age of page content. For a given page on the Web my system retrieves its previous versions from online Web archives. Then these past versions are compared with the present version. By using binary

search on the series of retrieved versions it is possible to estimate the approximate time periods when particular content parts appeared on the page for the first time.

(4) Web users often re-visit pages over time. I have proposed a method that displays the content changes on the page that appeared since the last user visit. In addition, the proposed system annotates links with the degree of fresh, unseen content that is contained in the linked pages. This kind of freshness-based navigation support in Web sites helps users to understand the changes and easily find novel content.

(5) I have participated in creating a system for detecting and visualizing changes in coordinate terms over time. Coordinate terms indicate peer or rival type relationships among objects. By mining past news gathered from online news archives it is possible to detect such terms for a given query. Then the evolution of such terms is shown as a dynamic graph (Fig. 3).

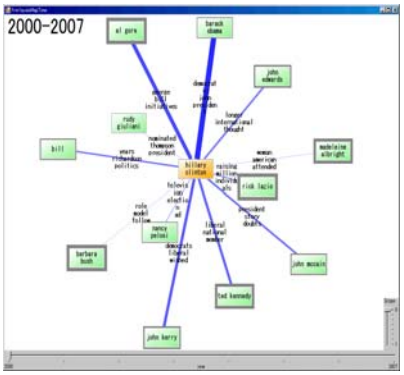

Figure 3 Visualization of changes in coordinate terms over time.

(6) I have investigated the way in which social bookmark collections can be used for improving Web search. It was revealed that social bookmarks have superior temporal characteristics and can be used for retrieving fresh but valuable pages that are still poorly visible on the Web due to their relatively short age. I have participated in building a system that combines popularity and relevance scores of pages calculated from social bookmark collections with the ones estimated by a standard way, that is, by analyzing the content and link structure of Web pages.

(7) I have also helped to create a system for evaluating trustworthiness of factual expressions on the Web. This system analyzed the temporal evolution of candidate expressions that are similar to user input. By comparing the popularity changes over time it could decide which factual expressions are still valid and which are obsolete.

(8) In my last work, I have provided method for detecting and summarizing future events based on their future-related references in online news archives or on the web. The proposed system detects context around future dates in documents related to particular real-world objects. Next it agglomerates the extracted information by clustering and burst detection. The final result is in the form of a visual output that enables users to clearly notice the forthcoming events related to the objects and their expected time periods of occurrence (Fig. 4).

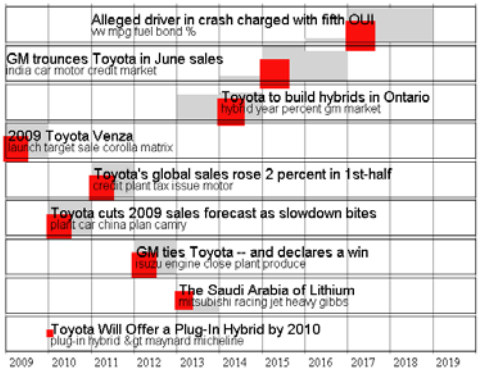

Figure 4 Visualization of future events for user query (Toyota).

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）

1. 山本 祐輔, 手塚 太郎, Adam Jatowt, 田中 克己: WebAlert: Web情報の印象集約を利用した閲覧ページ内容に対する反対意見提示. 日本データベース学会論文誌, Vol.7, No.1, 2008年6月（査読有）.

2. 山家 雄介, 中村 聡史, Adam Jatowt, 田中 克己: ソーシャルブックマークの特性分析とそれに基づくWeb検索の再ラ

ンキング手法，情報処理学会論文誌：データベース，Vol.1 No.1 (TOD 38)，2008年6月(査読有)．

3. 山本 祐輔，手塚 太郎，Adam Jatowt，田中 克己：ページ特性を考慮したWeb検索結果の集約とページ生成時間分析による知識の信頼性判断支援．電子情報通信学会論文誌，Vol.J91-D，No.03，pp.576-584，2008年3月(査読有)．

4. Adam Jatowt and Katsumi Tanaka: Towards Mining Past Content of Web Pages, New Review of Hypermedia and Multimedia, Special Issue on Web Archiving (NRHM), Taylor & Francis, 13(1), technical note, pp. 77-86 (2007)

5. 山本 祐輔，手塚 太郎，Adam Jatowt，田中 克己：ほんと?サーチ：検索結果の集約とページ生成時間分布解析によるWeb情報の信用度評価．日本データベース学会Letters，Vol.6，No.1，2007年6月(査読有)．

6. 山家 雄介，中村 聡史，Adam Jatowt，田中 克己：ソーシャルブックマークの特性を利用したweb検索のランキング精度向上，日本データベース学会Letters(DBSJ Letters) vol. 6, No. 1, 2007年6月(査読有)．

〔学会発表〕(計 14 件)

1. Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama and Katsumi Tanaka: Supporting Analysis of Future-related Information in News Archives and the Web, Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009), ACM Press, Austin, USA (2009)

2. Adam Jatowt, Yukiko Kawai, Hiroaki Ohshima and Katsumi Tanaka: What Can History Tell Us? Towards Different Models of Interaction with Document Histories, Proceedings of 19th ACM Conference on Hypertext and Hypermedia (HT 2008), ACM Press, Pittsburgh, USA, pp. 5-14 (2008)

3. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Visualizing Historical Content of Web Pages, Proceedings of the International World Wide Web Conference (WWW 2008), ACM Press, poster, Beijing, China, pp. 1221-1222 (2008)

4. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Using Page Histories for Improving Browsing the Web, Proceedings of the 8th International Web Archiving Workshop (IWAW 2008), Aarhus, Denmark (2008)

5. Hiroaki Ohshima, Adam Jatowt, Satoshi Oyama and Katsumi Tanaka: Visualizing Changes in Coordinate Terms over Time: An Example of Mining Repositories of Temporal Data through their Search Interfaces, Proceedings for the International Workshop on Information-explosion and Next Generation Search (INGS 2008), IEEE CS Digital Library, Shenyang, China, pp. 61-68 (2008)

6. Yusuke Yamamoto, Taro Tezuka, Adam Jatowt and Katsumi Tanaka: Supporting Judgment of Fact Trustworthiness Considering Temporal and Sentimental Aspects, Proceedings of Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Browsing Assistant for Changing Pages, In: Nguyen N.T., Jain L.C. (Eds.): Intelligent Agents in the Evolution of Web and Applications, Springer-Verlag, pp. 137-160 (2009)the 9th International Conference on Web Information Systems Engineering (WISE 2008), Springer LNCS 5175, Auckland, New Zealand, pp. 206-220 (2008)

7. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Detecting Age of Page Content, Proceedings of the 9th ACM International Workshop on Web Information and Data Management (WIDM 2007), ACM Press, Lisbon, Portugal, pp. 137-144 (2007)

8. Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Satoshi Oyama and Katsumi Tanaka: Trustworthiness Analysis of Web Search Results, Proceedings of the 11th European Conference on Research and Advanced technology for Digital Libraries

(ECDL 2007), Springer LNCS 4675, Budapest, Hungary, pp. 38-49 (2007)

9. Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura and Katsumi Tanaka: Towards Improving Web Search by Utilizing Social Bookmarks, Proceedings of the 7th International Conference on Software Engineering (ICWE 2007), Springer LNCS 4607, Como, Italy, pp. 343-357 (2007)

10. Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura and Katsumi Tanaka: Can Social Bookmarking Enhance Search in the Web?, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007), ACM Press, Vancouver, Canada, pp. 107-116 (2007)

11. Nimit Pattanasri, Adam Jatowt and Katsumi Tanaka: Context-Aware Search inside e-Learning Materials Using Textbook Ontologies, Proceedings of a joint conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management (APWeb/WAIM 2007), Springer LNCS 4505, HuangShan, China, pp. 658-669 (2007)

12. Yusuke Yamamoto, Taro Tezuka, Adam Jatowt and Katsumi Tanaka: Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis, Proceedings of a joint conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management (APWeb/WAIM 2007), Springer LNCS 4505, HuangShan, China, pp. 253-264 (2007)

13. Adam Jatowt, Yukiko Kawai, Satoshi Nakamura, Yutaka Kidawara and Katsumi Tanaka: Journey to the Past: Proposal of a Framework for Past Web Browser, Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (HT 2006), ACM Press, Odense, Denmark, pp. 134-144 (2006)

14. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Personalized Detection of Fresh Content and Temporal Annotation for Improved Page Revisiting, Proceedings of the 17th Conference on Database and Expert Systems Applications (DEXA 2006), Springer LNCS 4080, Krakow, Poland, pp. 832-841 (2006)

〔図書〕(計 3 件)

1. Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura and Katsumi Tanaka: Social Bookmarking and Web Search, In: San Murugesan (Ed.): Handbook of Research on Web 2.0, 3.0 and X.0: Technologies, Business, and Social Applications, IGI Global, (to appear)

2. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Browsing Assistant for Changing Pages, In: Nguyen N.T., Jain L.C. (Eds.): Intelligent Agents in the Evolution of Web and Applications, Springer-Verlag, pp. 137-160 (2009)

3. Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Utilizing Past Web for Knowledge Discovery, In: Krol D., Nguyen N.T. (Eds.): Intelligence Integration in Distributed Knowledge Management, IGI Global, pp. 283-301 (2008)

６．研究組織

(1)研究代表者

Adam Jatowt

京都大学・大学院情報学研究科・特定助教

研究者番号：**00415861**