科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2010～2013
課題番号：　２２７０００９６

研究課題名（和文）時間指向ウェブ検索およびウェブマイニング

研究課題名（英文）Towards time-focused Web Search and Mining

研究代表者
Ａｄａｍ　Ｊａｔｏｗｔ（Jatowt, Adam）

京都大学・情報学研究科・准教授

研究者番号：００４１５８６１

交付決定額（研究期間全体）：（直接経費）　　3,000,000円、（間接経費）　　900,000円

研究成果の概要（和文）：この研究では、ドキュメントの内容が指し示す時間的表現を取り出しその時期を推定することによって、新しい時間情報検索の手法とアプローチをいくつか提案することができました。この時間情報検索をさらに向上させるために、大規模ドキュメントセットから未来に関連する集合情報を取り出す方法についても研究を行いました。
また画像からその年代を推定する方法について分析・検討するとともに、ウェブ検索においてユーザーが時間的要素をどの程度活用しているかについての調査を行いました。

研究成果の概要（英文）：This research has resulted in several novel methods and approaches for temporal information retrieval. First, we have provided methodology for estimating reference time of documents, which represents the time period to which the given document content refers to. For improving temporal information retrieval, we proposed a method for extracting collective future-related information from large document sets.
In another work, we have proposed a method for estimating age of images. Moreover, we have measured the extent to which users use time and temporal features in their search activities on the web. This online survey has been done on 110 Web users and its results were published at TempWeb2013 workshop. To further foster the research in temporal information retrieval we established Temporal Information Access task challenge in NTCiR. And we wrote a survey paper about Temporal Information Retrieval which is scheduled to appear in this year in ACM Computing Surveys journal.

研究分野：　総合領域

科研費の分科・細目：　情報学・メディア情報学・データベース

キーワード：　時間情報　時間情報解析と獲得　新鮮さによるサーチ結果のランキング

１．研究開始当初の背景

近年、検索エンジンは専門的および個人的な目的で情報を取得するために広く利用されており、また、ウェブ上で利用可能な文書コンテンツに対するデータマイニングに関する研究も多く行われています。Web 上の情報は現実の世界の出来事によって非常に迅速に変更されるため、取得した情報の時間的な側面がマッチしているかどうかを検証することは極めて重要になります。未来の出来事についての情報を得たい場合に検索エンジンもデータ･マイニング･プロセスも過去または現在の情報を返すべきではないのに、現在の一般的な検索エンジンやナレッジプロセッシングアプリケーションでは時間的要素を Web 検索における効果的な要素とみなしておらず、ユーザーも強い時間的性質を持つ情報を検索するときにそのサポートをあまり受けられていない状況にあります。

２．研究の目的

我々の研究の目的は情報の時間的な側面を考慮することによってウェブ検索やマイニング技術を向上させることです。そのため、ユーザーが異なる時間軸の情報（過去・現在・未来）をどのような方法でどのくらい検索しているかを調べたいと思いましたが、現在の情報を扱う研究は比較的多くありますので、今回は特に過去と未来についての情報を取り上げました。この研究結果を活用することで他の研究者に彼らの研究成果を検証する方法を提供することも可能になります。

３．研究の方法

⑴まず、我々は「フォーカスタイム」と呼ぶドキュメントの基準時間を推定する方法を提唱しました。フォーカスタイムとはドキュメントの内容が指し示す一定の期間を表し、ドキュメントタイムスタンプまたは文書作成時期の観念とは異なるものです。ドキュメントのフォーカスタイムを知ることは特定の期間についての情報、例えば 1930 年代のドイツ、または第二次世界大戦についてのドキュメントを検索するときに役立ちます。この手法では大規模な文書コレクションから得られた統計データを利用しており、これらのデータからある語句と特定の年代の関連性を推定することもできるようになりました。例えば、「ヒトラー」であれば 1930 年代から 1942 年までの期間に強い関連性をもっている一方で「iPhone」は 2007 年から現在までの期間と関連づけられます。このデータを利用することで、例えば日付といった任意の時間表現を持たないドキュメントについても時間の推定が可能になりました。

⑵それ以外に、それぞれの国における集合記憶の分析にも着手し、潜在的ディリクレ配分（LDA）を使ってニュースアーティクルから過去に関連する記事のメイントピックを探しました。これによってある特定の時期・国に関連するニュースの中から過去に関連する情報の総量を示し、国によって歴史がどのように記憶されているか、その違いについても明らかにすることができました。

⑶時間的要素による情報検索をより向上させるために、未来に関連する集合的な情報を大規模文書データから抽出する手法についても研究しました。未来のイベントに関する情報を探しクラスタリングすることで、起こりそうな未来のイベントを見つけてその発生確率を測定しました。さらに、未来に関連する情報とすでに起こったイベントに関するアーティクルの間のリンク検出を行うことで未来を予測できなかった情報を不確実なものとして排除し未来予測の信頼性を診断する手法も提案しました。

⑷さらに、画像の作成時期として考えられる年月日を判定するために画像だけではなく画像を取り巻くテキストもベクトル表示として採用し、画像からその年代を推定する方法を提案しました。画像によるアプローチについては、いつその画像が作成されたかがわかるような典型的な特徴を学ぶために SVM 分類を活用しました。

⑸ユーザーがどの程度に時間や時間的特徴をもつ検索行動をしているかを調査するために、日本のウェブユーザー110 人に過去・最近・未来に関する情報をどれくらいウェブ検索したかを尋ねて調査を行いました。

⑹この時間的要素による情報検索の研究をさらに進めるために、NTCIR (NII Testbeds and Community for Information access Research) に “ Temporal Information Access ” チャレンジタスクを設立しました。この計画によって同等のデータセットとタスク検索システムのパフォーマンスを評価する標準的な指針を提供することができ、時間的情報のプロセッシングと検索について現行の研究を要約した調査論文を書くことができました。このことは研究の今後の方向性を描くことにもつながりました。

４．研究成果

⑴この研究によって、時間的情報の検索に関する新しい手法とアプローチをいくつか提案することができました。まずドキュメントの内容が示している時期、ドキュメントのリファレンスタイムを測定する方法論を確立しました。この研究の成果は CIKM2011 にて発表された論文に記されています。

⑵また CIKM2011 で発表した別の論文では国によって異なる集合記憶の分析を行い、このような分析のためのフレームワークを提供することできました。

⑶この時間情報検索をさらに向上させるために、大規模ドキュメントセットから未来に関連する集合情報を取り出す方法については CIKM2011 および WI2011 にて提案しています。またウェブ上の未来志向の情報の量と性質については ICCC2013 で発表した論文で分析を行っています。

⑷また画像からその年代を推定する方法についても分析を行い、SPIR2012 でその結果を

発表しました。
⑸さらにユーザーが時間および時間的特徴をウェブ検索でどの程度利用しているかについても調査を行い、この 110 名によるオンライン調査は TempWeb2013 ワークショップにて発表されました。
⑹時間情報検索の研究をさらに発展させるために NTCIR (NII Testbeds and Community for Information access Research) に“Temporal Information Access”チャレンジタスクを設立し、時間情報検索についての調査論文を執筆しました。これは今年 ACM Computing Surveys ジャーナルにて発表される予定です。

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 15 件）

①Ricardo Campos, Gael Dias, Alipio Jorge, Adam Jatowt. Survey of Temporal Information Retrieval and Related Applications. ACM Computing Surveys, ACM Press (to appear in 2015)
②Adam Jatowt and Kevin Duh. A Framework for Analyzing Semantic Change of Words across Time, Proceedings of Digital Libraries Conference (JCDL 2014 / TPDL 2014) [Joint JCDL 2014 and TPDL 2014 conference], ACM Press, London, UK, (2014)
③Hideo Joho, Adam Jatowt, and Roi Blanco. NTCIR Temporalia: A Test Collection for Temporal Information Access Research, Proceedings of the 4th Temporal Web Analytics Workshop (TempWeb 2014), ACM Press, Seoul, Korea, pp. 845-849 (2014)
④ Adam Jatowt, Ching Man Au Yeung and Katsumi Tanaka. Estimating Document Focus Time, Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013), ACM Press, San Francisco, CA, USA, pp. 2273-2278 (2013)
⑤Hideo Joho, Adam Jatowt, and Roi Blanco. A Survey of Temporal Web Search Experience, Proceedings of the 3rd Temporal Web Analytics Workshop (TempWeb 2013), ACM Press, Rio de Janeiro, Brasil, pp. 1101-1108 (2013)
⑥ Adam Jatowt, Hideki Kawai, Kensuke Kanazawa, Katsumi Tanaka, Kazuo Kunieda and Keiji Yamada. Multi-lingual, Longitudinal Analysis of Future-related Information on the Web, Proceedings of the 4th International conference on Culture and Computing (Culture and Computing 2013), IEEE Press, Kyoto, Japan, pp. 27-32 (2013)
⑦Adam Jatowt and Katsumi Tanaka. Large Scale Analysis of Changes in English Vocabulary over Recent Time, Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), ACM Press, Maui, Hawaii, USA, pp. 2523-2526 (2012)
⑧Gael Dias, Jose Moreno, Adam Jatowt and Ricardo Campos. Temporal Web Image Retrieval, Proceedings of the 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012), Springer, Cartagena de Indias, Colombia, pp. 199-204 (2012) {acceptance rate 13/81 = 16%}
⑨ Adam Jatowt and Katsumi Tanaka. Longitudinal Analysis of Historical Texts' Readability, Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2012), ACM Press, Washington DC, USA, pp. 353-354 (2012)
⑩ Ching Man Au Yeung and Adam Jatowt. Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining, Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), ACM Press, Glasgow, UK, pp. 1231-1240 (2011)
⑪ Adam Jatowt and Ching Man Au Yeung. Extracting Collective Expectations about the Future from Large Text Collections, Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), ACM Press, Glasgow, UK, pp. 1259-1264 (2011)
⑫ Kensuke Kanazawa, Adam Jatowt and Katsumi Tanaka. Improving Retrieval of Future-related Information in Text Collections, Proceedings of the 2011 IEEE/WIC/ACM Web Intelligence (WI 2011), IEEE Press, Lyon, France, pp. 278-283 (2011)
⑬Adam Jatowt, Yukiko Kawai and Katsumi Tanaka. Calculating Content Recency based on Timestamped and Non-Timestamped Sources for Supporting Page Quality Estimation, Proceedings of the 26th Symposium On Applied Computing (SAC 2011), ACM Press, Taichung, Taiwan, pp. 1156-1163 (2011)
⑭Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. Page History Explorer: Visualizing and Comparing Page Histories, IEICE Transactions on Information and Systems, 査読有, 94-D(3), pp. 564-577 (2011)
⑮ Hideki Kawai, Adam Jatowt, Katsumi Tanaka, Kazuo Kunieda, Keiji Yamada. Query Expansion and Text Mining for ChronoSeeker - Search Engine for Future/Past Events -, IEICE Transactions on Information and Systems 94-D(3), pp. 552-563 (2011)

〔学会発表〕（計 7 件）
①Hideo Joho, Adam Jatowt, and Roi Blanco. NTCIR Temporalia: A Test Collection for Temporal Information Access Research, Proceedings of the 4th Temporal Web Analytics Workshop (TempWeb 2014), ACM Press, Seoul, Korea, pp. 845-849 (April 8, 2014)
② Adam Jatowt, Ching Man Au Yeung and Katsumi Tanaka. Estimating Document Focus Time, Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013), ACM Press, San Francisco, CA, USA, pp. 2273-2278 (October 31, 2013)
③ Ching Man Au Yeung and Adam Jatowt. Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining, Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), ACM Press, Glasgow, UK, pp. 1231-1240 (October 27, 2011)
④ Adam Jatowt and Ching Man Au Yeung. Extracting Collective Expectations about the Future from Large Text Collections, Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), ACM Press, Glasgow, UK, pp. 1259-1264 (October 27, 2011)
⑤ Kensuke Kanazawa, Adam Jatowt and Katsumi Tanaka. Improving Retrieval of Future-related Information in Text Collections, Proceedings of the 2011 IEEE/WIC/ACM Web Intelligence (WI 2011), IEEE Press, Lyon, France, pp. 278-283 (August 25, 2011)
⑥Adam Jatowt, Yukiko Kawai and Katsumi Tanaka. Calculating Content Recency based on Timestamped and Non-Timestamped Sources for Supporting Page Quality Estimation, Proceedings of the 26th Symposium On Applied Computing (SAC 2011), ACM Press, Taichung, Taiwan, pp. 1156-1163 (March 23, 2011)
⑦ Adam Jatowt, Hideki Kawai, Kensuke Kanazawa, Katsumi Tanaka, Kazuo Kunieda and Keiji Yamada. Multi-lingual, Longitudinal Analysis of Future-related Information on the Web, Proceedings of the 4th International conference on Culture and Computing (Culture and Computing 2013), IEEE Press, Kyoto, Japan, pp. 27-32 (September 16, 2013)

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

〔その他〕
ホームページ等
http://www.dl.kuis.kyoto-u.ac.jp/~adam/

６．研究組織
(1)研究代表者
Adam Jatowt（JATOWT, Adam）
京都大学・大学院情報学研究科・准教授
研究者番号：00415861