様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２８ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2020
課題番号：１７Ｈ０１８２８

研究課題名（和文）Novel Technologies for Improving Comprehension and Utilization of Historical Knowledge

研究課題名（英文）Novel Technologies for Improving Comprehension and Utilization of Historical Knowledge

研究代表者
Ａｄａｍ　Ｊａｔｏｗｔ（Jatowt, Adam）

京都大学・情報学研究科・特定准教授

研究者番号：００４１５８６１

交付決定額（研究期間全体）：（直接経費）　　11,980,000円

研究成果の概要（和文）：このプロジェクトでは、一次資料（ニュース記事など）や二次資料（ウィキペディアなど）の両方に基づいて歴史を分析・理解するための多くの新しいアプローチを開発しました。時系列での類似性の検出、共有または類似した歴史に基づいたエンティティのクラスタリング、過去に関連する質問への回答、アーカイブ文書の検索モデル、タイムラインの自動構築などの新しい研究課題を提案しています。この研究の過程で、いくつかのデータセットも開発しました。

研究成果の学術的意義や社会的意義
The project resulted in several publications in high impact conferences and journals related to artificial intelligence, data mining and natural language processing such as WSDM, CIKM, ECIR, ECAI, JCDL, IP&M. The proposed models can be e.g. implemented in museums or libraries to attract visitors.

研究成果の概要（英文）：The project has resulted in the development of many novel approaches to analyze and understand history both based on primary sources (e.g., news articles) or secondary sources (e.g., Wikipedia). We have proposed new research tasks such as detection of analogy over time, clustering entities based on their shared or similar histories, answering questions related to the past, search models for archival documents and automatically building timelines. In the process of this research several datasets have been also developed. Given the proposed methods and tools, users who are either processionals or average users can find or understand content related to history in easier and more effective way. Our society digitizes massive data from the past such as old books and news articles, however so far we had few dedicated computational modes for processing such data. With the outcomes of this project we have generated many novel directions for future research for the academic community.

研究分野：Natural language processing

キーワード：news archives　digital history　collective memory

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

Nowadays there is a lot of digitized old documents such as news article archives that can be used for finding, processing and extracting valuable knowledge on the past. Unfortunately, there is lack of tools dedicated for this kind of documents. Most of the information and knowledge retrieval approaches are dedicated to static or synchronic document collections. On the other hand, diachronic document collections such as long-term document archives require special approaches since time plays a key role in such collections and is required to be considered for proper understanding, retrieval and comparison of contained information.

Nowadays, news is one of the most important channels for acquiring high-quality information regarding our society. However, with the rapid growth of Web, more and more news articles are available causing information overload. News retrieval, processing and summarization can help to combat this problem by distilling the most important information from large amount of news articles for users. In this research we recognize the importance of news articles for our society and we especially focused on this type of documents. Our target were then news archive collections that span several decades and that contains hundred thousands of news reports on important events that happened long time ago. Utilizing this kind of heritage content is non-trivial yet is at the same time quite important in order to benefit from large amounts of accumulated data.

２．研究の目的

The objective of this research was to design search and analysis methods for extracting content from large collections of past documents (in particular news articles) that would be useful and attractive for either professional or average users. The novelty of our research lies in effectively handling the particular character of this kind of data, especially, its temporal characteristics and the unknown context for nowadays users who wish to access such collections. In particular in the context of this project we have developed the following methods and researches:

- Contrastive history-based summarization of entity groups or news article collections
- Automatic timeline generation for effectively summarizing temporal news data
- Ranking news articles in order to retrieve content that is interesting to current users – here we focused on surprise effect by trying to find surprising content from the past that would be against our assumptions or knowledge
- Ranking news articles in order to retrieve content that is relevant to present day – we proposed the notion of present relatedness which means retrieving past news articles that have strong relation to the present
- Finding temporal analogs which are similar past entities to the current entities (e.g., Walkman and ipad)
- Investigating the history of words and their semantic evolution by mining massive textual diachronic corpora such as Google Books
- Detecting influential documents that impacted future – here we focused on research articles and we tried to find those scientific papers that were innovative in some way at the tie of their publishing.

３．研究の方法

We have used a range of natural language processing tools such as timeline summarization, information retrieval, information extraction or, in general, machine learning methods applied for textual data such as clustering or word embedding representations to realize our research objectives.

４．研究成果

We have developed methods for temporal summarization of history related content in contrastive way (called comparative timeline summarization). For the above-mentioned research direction, we proposed a novel task of query-based across-time comparative news summarization that generates query-dependent contrastive summary between two news collections from different time periods (typically, one representing the present time and the other being some period in the past). The output summary is in the form of pairs of corresponding news contents. we propose a novel summarization system to address the proposed task. First for this objective, we formulated the measurement of sentence correspondence by finding corresponding terms and aligning different vector spaces. We adopted the distributed vector representation to represent the context vectors of words; then we learnt linear and orthogonal transformations between two vector spaces of input collections for learning word correspondence. Furthermore, we measured the correspondence between two candidate news as the minimal amount of distance that the words of one news need to "travel" to match the words of another news, suggested by the commonly-used Word Mover Distance (WMD) algorithm. Secondly, inspired by the popular Affinity Propagation algorithm, we propose a concise joint integer linear programming framework which detects diverse and representative news (which we call exemplars) and at the same time generates correspondent pairs from the detected exemplars. Based on this formulation, exactly optimal solution can be obtained.

We next introduced a novel task of multiple timeline summarization of news article collections. The main concept is to output a series of timelines instead of a single timeline, especially in the case of a heterogeneous news article collection that contains multiple topics and central entities.

In another research we have proposed a novel retrieval method for news archives in which we put focus on the interestingness and contextual relevance of news articles for present day users. In particular, we ranked news articles not only by their relevance to user queries but also by the contextual factors such as the extent to which they invoke surprise in the readers. The surprise was measured by the level of dissimilarity to the current events. Another criteria for reranking news articles is based on measuring relevance of the articles to the current times (contemporary relevance). This is done by finding named entities from the past that are still currently important as well as topics which are currently common. Such features would then be considered in measuring news article relevance to the present times.

The next line of research was devoted to finding terms from the past that are similar or analogical to current entities. For example, Walkman played similar role to iPad currently. This task is not simple since it requires transformation of vector spaces for measuring across-time similarity of entity descriptions. We have achieved this by optimization formulation based on automatically proposed sets of training pairs.

Another task was devoted for measuring the evolution of word meaning over time. We have carried this task by comparing word representations over time. The result of our efforts in this direction are demo system in the form of online interactive service that can take any word and return the information on the evolution of that word's meaning. Finally, we have organized two international workshops and edited one book on the topic of computational approaches to measuring word evolution.

Finally, we proposed methods for finding influential research papers from temporal scholarly datasets such as dataset of scientific publications on NLP (ACL anthology). We have also proposed method for detecting which user queries issued to search engines are related to past events.For this research task, we addressed the shortcomings of traditional scientometric approaches by proposing a novel method that utilizes a classifier for predicting publication years based on latent topic distributions. We then calculated real-number innovation scores used to identify potential breakthrough papers and turnaround years. The proposed approach can complement existing citation-based measures of article importance and author contribution analysis; it opens as well novel research direction for time-based, innovation-centered research scientific output evaluation. In our experiments, we focused on two corpora of research papers

published over several decades at two well-established conferences: The World Wide Web Conference (WWW) and the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), containing around 3500 documents in total. We indicated significant years and demonstrate examples of highly-ranked papers, thus providing a novel insight on the evolution of the two conferences. Finally, we compared our results to citation analysis and discuss how our approach may complement traditional scientometrics.

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　9件／うち国際共著　9件／うちオープンアクセス　0件）

１．著者名 Yating Zhang, Adam Jatowt, Sourav S Bhowmick, Yuji Matsumoto	４．巻 ACM Press
２．論文標題 ATAR: Aspect-based Temporal Analog Retrieval System for Document Archives	５．発行年 2019年
３．雑誌名 The 12th International Conference on Web Search and Data Mining (WSDM 2019)	６．最初と最後の頁 762-765
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3289600.3290613	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Yijun Duan, Adam Jatowt	４．巻 ACM Press
２．論文標題 Across-Time Comparative Summarization of News Articles	５．発行年 2019年
３．雑誌名 The 12th International Conference on Web Search and Data Mining (WSDM 2019)	６．最初と最後の頁 735-743
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3289600.3291008	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Min Min Chew, Sourav S. Bhowmick, Adam Jatowt	４．巻 ACM PRess
２．論文標題 Ranking Without Learning: Towards Historical Relevance-based Ranking of Social Images	５．発行年 2018年
３．雑誌名 The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)	６．最初と最後の頁 1133-1136
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3209978.3210100	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yasunobu Sumikawa, Adam Jatowt, Marten During	４．巻 ACM Press
２．論文標題 Digital History meets Microblogging: Analyzing Collective Memories in Twitter	５．発行年 2018年
３．雑誌名 The ACM/IEEE Joint Conference on Digital Libraries (JCDL 2018)	６．最初と最後の頁 213-222
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3197026.3197057	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Sumikawa Yasunobu、Jatowt Adam	4．巻 10772
2．論文標題 Classifying Short Descriptions of Past Events	5．発行年 2018年
3．雑誌名 ECIR 2017	6．最初と最後の頁 729～736
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-76941-7_69	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Duan Yijun、Jatowt Adam、Tanaka Katsumi	4．巻 ACM Press
2．論文標題 Discovering Typical Histories of Entities by Multi-Timeline Summarization	5．発行年 2017年
3．雑誌名 HT 2017	6．最初と最後の頁 105-114
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3078714.3078725	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Sumikawa Yasunobu、Jatowt Adam、D?ring Marten	4．巻 ACM Press
2．論文標題 Analysis of Temporal and Web Site References in History-related Tweets	5．発行年 2017年
3．雑誌名 WebSci 2017	6．最初と最後の頁 419-420
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3091478.3098868	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Jatowt Adam、Kawai Daisuke、Tanaka Katsumi	4．巻 ACM Press
2．論文標題 Timestamping Entities using Contextual Information	5．発行年 2017年
3．雑誌名 SIGIR 2017	6．最初と最後の頁 105-108
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3077136.3080762	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Savov Pavel、Jatowt Adam、Nielek Radoslaw	4．巻 ACM Press
2．論文標題 Towards Understanding the Evolution of the WWW Conference	5．発行年 2017年
3．雑誌名 WWW 2017	6．最初と最後の頁 835-836
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3041021.3054252	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Zhang Yating、Jatowt Adam、Tanaka Katsumi	4．巻 ACM Press
2．論文標題 Temporal Analog Retrieval using Transformation over Dual Hierarchical Structures	5．発行年 2017年
3．雑誌名 CIKM 2017	6．最初と最後の頁 717-726
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3132847.3132917	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Jatowt Adam、Campos Ricardo	4．巻 ACM Press
2．論文標題 Interactive System for Reasoning about Document Age	5．発行年 2017年
3．雑誌名 CIKM 2017	6．最初と最後の頁 2471-2474
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3132847.3133166	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　2件）

1．発表者名 Yijun Duan
2．発表標題 Across-Time Comparative Summarization of News Articles
3．学会等名 The 12th International Conference on Web Search and Data Mining (WSDM 2019)（国際学会）
4．発表年 2018年

１．発表者名
Yasunobu Sumikawa

２．発表標題
Digital History meets Microblogging: Analyzing Collective Memories in Twitter

３．学会等名
The ACM/IEEE Joint Conference on Digital Libraries (JCDL 2018)（国際学会）

４．発表年
2018年

１．発表者名
Adam Jatowt

２．発表標題
Temporal Analog Retrieval using Transformation over Dual Hierarchical Structures

３．学会等名
CIKM 2017

４．発表年
2017年

１．発表者名
Yijun Duan

２．発表標題
Discovering Typical Histories of Entities by Multi-Timeline Summarization

３．学会等名
HT 2017

４．発表年
2017年

１．発表者名
Adam Jatowt

２．発表標題
Timestamping Entities using Contextual Information

３．学会等名
SIGIR 2017

４．発表年
2017年

１．発表者名 Adam Jatowt
２．発表標題 Interactive System for Reasoning about Document Age
３．学会等名 CIKM 2017
４．発表年 2017年

１．発表者名 Yasunobu Sumikawa
２．発表標題 Classifying Short Descriptions of Past Events
３．学会等名 ECIR 2017
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	澄川　靖信 (Sumikawa Yasunobu) (70756303)	東京都立大学・大学教育センター・助教 (22604)	
研究分担者	Ｚｈａｎｇ　Ｙａｔｉｎｇ (Zhang Yating) (30793559)	国立研究開発法人理化学研究所・革新知能統合研究センター・特別研究員 (82401)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計3件

国際研究集会 The 1st Text2Story workshop (Text2Story 2018) in conjunction with 40th European Conference on Information Retrieval Conference (ECIR 2018)	開催年 2018年～2018年
国際研究集会 The 4th International Workshop on Computational History (HistoInformatics 2017) in conjunction with the 26th International Conference on Information and Knowledge Management (CIKM 2017)	開催年 2017年～2017年

国際研究集会	開催年
The 1st workshop on User Interfaces for Spatial-Temporal Data Analysis (UISTDA 2018) in conjunction with the ACM Intelligent User Interfaces Conference (IUI 2018)	2018年～2018年

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関