研 究 報 告 書

「Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents (集合記憶の分析および歴史文書からの知識抽出手法の開発)」

研究タイプ:通常型
研究期間: 平成 23 年 10 月～平成 27 年 3 月
研 究 者: Adam Jatowt

1. 研究のねらい

History as a representation of the past has many functions. It helps to create meaning, helps to settle foundations of our nations, identities and memories, and is one of the fundamental subjects taught from elementary schools onwards. Nowadays, due to the increasing activities in digitizing and opening historical sources, the field of history can greatly benefit from the advances of computer and information sciences. New computer science techniques can be applied to help verify and validate historical assumptions based on text analytics and through comparison of multiple perspectives.

This project has **two objectives**. First, it aims at providing **computational framework for analyzing collective memory** focusing on questions such as how much the past matters for our society, how people remember and forget the past and how they find and access history-related information. Collective memory refers to the memory that members of the society share about the past. Among the aspects that we research in the study of collective memory are the patterns of remembering/forgetting, of past events, the types of remembered information and the characteristics and access statistics of past-related information.

The second objective is to provide **automatic tools for studying and understanding historical documents and documents about the past**. Here we aim to provide methods to automatically estimate temporal characteristics of documents such as when a particular document has been written and about what time period it is. Other aspects that we work on are the analysis of the scope of the language evolution, design of computational techniques for studying changes in word meaning across time and design of methods for extracting knowledge from temporal document collections and document archives.

2. 研究成果

(1)概要

The following achievements have been obtained In the progress of this research:

Collective Memory Analysis. We have offered methodology that applies topic modelling for analyzing collective memory decay over time and for investigating which past events are remembered and which are forgotten in the context of different countries. Our approach has been implemented on the collection consisting of 2.4 million news articles about diverse

countries. We then extended our analysis to history-related data found in the English Wikipedia, which can be considered as comprehensive source of our memory of the past. Novel approach was applied in this case due to different character of data in the Wikipedia.

Temporal Document Analysis. We have designed framework for estimating two types of times related to documents: creation time and focus time. The former is the publication date of a document and the latter is the time period about which the given document is. The proposed approach has been also extended to find the approximate creation time of images based on machine learning.

Language Evolution Study. We have provided framework for detecting and visualizing changes in meaning of words over time based on the largest available historical corpus, Google Books, that contains over 1TB of text data (about 0.3 trillion words). We also have solved the temporal correspondence problem by proposing method based on deep learning for finding objects in the past that correspond to present objects (e.g., Walkman in 1980s and iPod in 2010s). This method can be applied in automatic timeline construction and can support users searching in longitudinal archives by suggesting names of past entities.

Extracting Knowledge from Archives. We have designed system for visually comparing data derived from documents created in different times. It represents documents retrieved from two different time periods as two side-by-side tag clouds or word graphs. In another work we have built system for automatic summarization of product evolution (e.g., evolution of walkman or phone).

Studying how Users Search for Past Information. We studied how users retrieve information related to the past from the Web, what techniques or tactics they use and what problems they encounter. The study was first conducted on 110 users as online questionnaire and then on 30 searchers in controlled laboratory settings. The findings were compared with those obtained for searching topics that relate to the present or the future.

Visualization Tool for Analyzing Past References in Twitter. Multi-facet visualization system has been created to portray the scope of temporal attention of Twitter users and the particular topics they mention when they refer to the past or the future. We have used two datasets for presenting the results: 31M Japanese tweets and 198M USA tweets collected over time periods of several months.

Community Service & Building. We have organized two international, interdisciplinary workshops related to computational history (called "HistoInformatics") together with history science researchers, as well as we have organized NTCiR-11 task related to temporal information access. Similar task is scheduled to be continued in NTCiR-12. We have published survey on temporal IR and co-organized ICADL2014 and SocInfo2013 conferences that are partially related to the topics of Sakigake grant.

(2)詳細

Investigating Collective Memory. The research problem related to this topic is to provide

efficient tools for analyzing *collective memory*. Previous studies of collective memory were limited to manual interrogations of subjects. Given the abundance of information in digital form nowadays it is possible to use computational approaches. For making this possible, we have focused on two data sources that we believe reflect social memory well: news articles and Wikipedia pages.

We approached the first data source, i.e. *news articles*, by applying topic modelling and statistical analysis. Based on the collection of 2.4M news articles about different countries we portrayed the pattern of memory decay over time (exponential forgetting) and we detected topics remembered in context of particular country for particular year. The study was then extended by biasing topics by additional variable, publication year of news articles. This allowed finding topics from a past year remembered at a particular year (e.g., recurring memories like anniversaries) to demonstrate that collective memory also changes over time.

The *Wikipedia* constitutes extensive source of historical data. In particular, the link structure and access patterns of Wikipedia articles about historical topics can be utilized to show how much the past matters to us and how much it is remembered. With this objective, we have analyzed the connectivity of Wikipedia pages about historical persons with those about the current persons as well as we measured the access frequency of Wikipedia pages on historical persons. The results show the decay in the amount of information, the decay of connectivity to the present and the decrease in the accumulated interest of Wikipedia visitors, the more time ago a particular historical person lived. The results of this work has been submitted to Web Science Track of WWW2015 conference [1].

Temporal Document Analysis. Most of evidences about the past are in the form of text documents. To study the collective memory and history in general, one first needs to understand documents that refer to the past or are from the past. In order to facilitate such historical and memory analysis we provide series of methods for estimating document temporality. In particular, we estimate two key types of time for any input text document in digital form: the *creation time* and *focus time* [2]. The former is the publication date of a document and the latter is the time period to which the given document refer. Below, we describe our approach to these two estimation problems.

Document creation time is estimated by aggregating time distributions of n-grams extracted from the document ({1,2,3,4,5}-grams). The frequency distributions of n-grams over time are obtained from Google Ngram dataset that has been constructed from Google Books datasets and dates back to 17th century. It contains over 1TB of text data (about 0.3 trillion words). Based on this approach we have built online service that outputs estimated document age for any input document and also visualizes age of particular words contained in this document similarly to the heat map visualization. The purpose of such visualization is to indicate portions of the document that may have been copied from some older documents and to provide more informative analysis to any researchers who wish to know or verify the true age of their documents. Online service that estimates document date and visualizes age of its content will be soon available.

Besides the method which analyzes text documents, we have also designed machine learning based approach for estimating the approximate age of images that are input in digital format [3]. Our approach achieves F-measure equal to 0.51 using SVM-based classification of 1000 street images into one of 5 mutually exclusive photography eras from 1826 to 2011 (baseline random classifier achieves F-measure equal 0.2).

Document focus time [2] is understood as time period to which document content refer. For example, document that is about World War II would have focus time from 1939 to 1945. In Fig. 1 we show example of a document whose content is mapped to focus time.

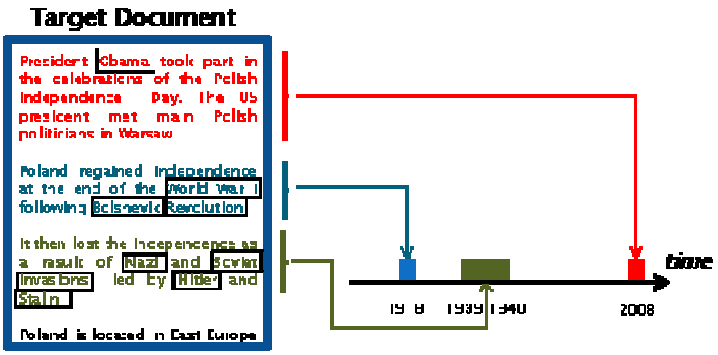

Figure 1 The concept of focus time shown as example of hypothetical document that is mapped to timeline (framed words help to position the document in time).

The estimation of document focus time is easy when underlying documents contain many temporal expressions such as dates. However, state-of-the-art approaches and tools fail to find such time in the absence of temporal expressions. We have solved this problem by building sets of automatic methods for document focus time estimation which utilize statistical co-occurrences of words and years within large document collections. The experiments on datasets created from historical textbooks, history-related web pages and Wikipedia pages on past events indicate that our method outperforms baselines, especially when documents have few dates achieving average error of 15 years of focus time detection (the total timeline length is 110 years). More detailed overview of this work has been conditionally accepted to appear in the special issue on "Time and IR" in the Information Processing and Management Journal [4].

Language Evolution Study. Our language is in the constant process of evolution. This is evident for anyone who reads old texts, especially, ones created more than few decades ago. Knowledge of the word semantic change over time should help to better understand past documents and could improve OCR algorithms. At the same time, word meaning evolution is interesting to many people as evidenced by lots of popular science books on language change.

We first attempted at *quantifying selected aspects of language evolution* such as usage of words [5] or changes in readability of old documents over time [6]. We then provided framework for *analyzing and visualizing the change in meaning of words* using the largest available historical corpus, Google Books [7]. Using three different representations of word context the proposed framework determines the change in word usage, change in sentiment and the change in similarity to related words for a given input word. Due to large size of the

datasets we needed to use Kyoto University Supercomputer to handle the processing. Our method has been also implemented on the Corpus of Historical American English. We plan to offer online service in 2015 to let users investigate the evolution of interesting words.

In the latest work we have solved the *temporal correspondence problem*, i.e., finding objects in the past that correspond to present objects. Given the input object name (e.g., iPod) and the target time period (e.g. 1980s) the task is to find the counterpart object that existed in the target time. The knowledge of temporal counterparts can help alleviate terminology gap problem for users searching within temporal document collections such as archives, can help automatic timeline construction or temporal summarization or can have usage in education. The main challenge of the temporal correspondence task comes from the change of the context (in Latin: "omnia mutantur", i.e., "everything changes") that results in low overlap of context terms. Thus it is difficult to find corresponding objects by directly employing distributional semantics to capture object meaning and by comparing their context vectors. We propose an unsupervised approach in which we first apply deep learning to represent the meaning of words within different time periods. We then transform the representation of terms in different temporal spaces to match terms semantically similar yet syntactically different.

Retrieving Knowledge from Archives. Many tasks in history involve comparison of large number of documents published at different times. We have designed system for *comparative knowledge extraction from temporal document collections* like archives. It compares the sets of documents retrieved from two time periods using two side-by-side panes displaying either comparative tag clouds or comparative word graphs. The system allows inputting a given entity (i.e., Japan) and comparing its context over two different time periods (e.g., 1980s and 2000s). To spot differential knowledge it detects similarities and differences related to the queried entity in different time periods by term cloud annotation and graph synchronization.

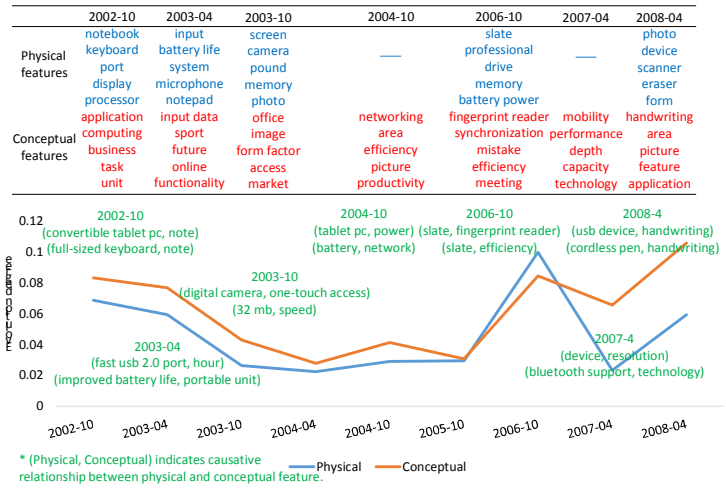

Figure 2 Example of product timeline visualization: "Tablet" 2002 – 2008.

In another research we have built automatic system for portraying the evolution of products and technology in general. Given an input query representing product type such as Walkman this system extracts terms that characterize milestones in the evolution of the product and calculates evolution degree of representative past product models. The resulting visualization

is in the form of timeline portraying conceptual and physical evolution of the product (see Figure 2 for example). We have submitted this research results to DASFAA'15 conference [8].

Studying how Users Search for Past Information & Improving Temporal IR. To learn about the past, users often search information on the Web. We studied how they can succeed in retrieving such information, what techniques or tactics they use, which web sites they visit and what problems they encounter. The study was first carried as online questionnaire [9] and then conducted in controlled laboratory settings on 30 users [10]. The obtained findings were compared with those obtained for search topics that relate to the present or the future.

We have also organized NTCiR-11 data mining challenge about temporal information access [11,12]. The task is called Temporalia and is composed of two subtasks. The first one aims to detect temporal orientation of search queries, while the objective of the second one is to retrieve documents related to past, present, future or atemporal documents for given sets topics. This has been the first test bed fully dedicated for temporal IR. 9 teams from 7 countries have participated in this task. The official datasets from Temporalia will be released in 2015. Our proposal for NTCiR-12 data challenge has been accepted as a core task.

Visualization Tool for Analyzing Past References in Twitter. Understanding *the way in which users refer to the past in their daily lives* is also related to the topic of collective memory and its development. We have constructed visualization system to portray the scope of temporal attention of Twitter users and the topics they mention when they refer to the past (memories) or to the future (predictions). We have used two datasets for presenting the results: 31M Japanese tweets and 198M USA tweets collected over 6 and 4 months, respectively. Since there no temporal taggers are available for Japanese language, we have built our own dedicated system for detecting temporal expressions in Japanese tweets. The goal of the visualization system is to effectively compare past- and future-related references in tweets. Tweets that contain temporal expressions related to the past and the future are visualized on exponential timeline in the form of a heat map, and the representative terms are collected for different timeframes of temporal attention (see Fig. 3 for overview). Users can interactively investigate the results by selecting different timeframes of attention or different publication dates. The results of this work has been accepted at WWW2015 conference [13], while the online service is provided for public use.

Community Service & Community Building. Besides organizing the Temporalia task at NTCiR-11, I have organized two international, interdisciplinary workshops related to computational history, called "HistoInformatics" [14] and I served as PC co-chair of two international conferences (SocInfo2013 and ICADL2014) that are partially related to the topics of digital humanities. Finally, we have published an extensive survey on Temporal IR and related applications including the description of the computational history field in [15].

In general, the temporal document analysis and the language evolution study were the most

challenging problems. For the former, novel techniques had to be used which can determine document focus time or its publication date in the absence of temporal expressions within document content and its timestamp. The latter problem involved usage of high speed processing architecture to tackle big data to capture the language used at different times.

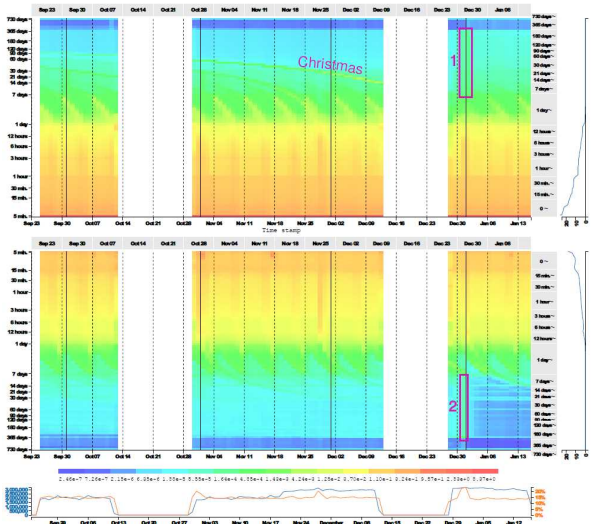

Figure 3 Visualization of temporal attention of Twitter users in US dataset (future on the top and past on the bottom parts of the graph, boxes 1 and 2 indicate new year effect).

[1] A. Jatowt, Daisuke Kawai, K. Tanaka: How Much the Past Matters? Analyzing Historical Social Networks in Wikipedia, {paper submitted to Web Science Track of WWW2015 conference}

[2] A. Jatowt, C.M. Au Yeung, K. Tanaka: Estimating document focus time. CIKM 2013: 2273–2278

[3] G. Dias, J. G. Moreno, A. Jatowt, R. Campos: Temporal Web Image Retrieval. SPIRE 2012: 199–204

[4] A. Jatowt. C.M. Au Yeung and K. Tanaka. Generic Method for Detecting Content Time of Documents. Information Processing and Management Journal (IPM) {under revision after conditional acceptance}

[5] A. Jatowt and K. Tanaka: Large Scale Analysis of Changes in English Vocabulary over Recent Time, Proceedings CIKM 2012, ACM Press, Maui, Hawaii, USA, pp. 2523–2526 (2012)

[6] A. Jatowt and K. Tanaka: Longitudinal analysis of historical texts' readability. Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012), ACM Press, pp. 353–354 (2012)

[7] A. Jatowt and K. Duh: A Framework for Analyzing Semantic Change of Words across Time, Proceedings of Digital Libraries Conference (JCDL 2014/TPDL 2014), IEEE Press, London, UK, pp. 229–238 (2014)

[8] Y. Zhang, A. Jatowt, and K. Tanaka: Object Evolution Summarization based on Analyzing Physical and Conceptual Changes. {Submitted to DASFAA2015 Conference}

[9] H. Joho, A. Jatowt, R. Blanco: A survey of temporal web search experience. Proceedings of the 3rd TempWeb2013 workshop, WWW (Companion Volume) 2013, ACM Press, pp. 1101–1108

[10] H. Joho, A. Jatowt, and R. Blanco: Temporal Information Searching Behaviour and Tactics, Information Processing and Management Journal (IPM) (to appear in 2015)

[11] H. Joho, A. Jatowt, R. Blanco: NTCIR Temporalia: a Test Collection for Temporal Information Access Research. Proceedings of the 4th TempWeb2014 workshop, WWW (Companion Volume) 2014, ACM Press, 845–850

[12] H. Joho, A. Jatowt, R. Blanco, H. Naka and S. Yamamoto: Overview of NTCIR-11 Temporal Information Access (Temporalia) Task, Proceedings of NTCiR-11, Tokyo, Japan, (2014)

[13] A. Jatowt, E. Antoine, Y. Kawai, T. Akiyama: Mapping Temporal Horizons, Analysis of Collective Future and Past related Attention in Microblogging. Proceedings of WWW 2015, ACM Press, (full paper), (to appear in May 2015)

[14] A. Jatowt, G. Dias, M. Duering and A. van Den Bosch: The HistoInformatics2014 Workshop, Socinfo2014 Workshop Proceedings, Springer LNCS (to appear in 2015)

[15] R. Campos, G. Dias, A. M. Jorge, A. Jatowt: Survey of Temporal Information Retrieval and Related Applications, ACM Computing Surveys, Vol. 47(2), ACM Press, pp. 1-41, 2014

3. 今後の展開

We plan to actively continue working on topics of this project. Our future plans embrace the following topics:

Studying Concept Evolution over Time. We plan to build system for analyzing evolution of meaning of entire concepts rather than evolution of words. This necessarily requires detection of sets of words that commonly constitute concepts. For example, to detect the evolution of the concept of "programming" we need to detect the evolution of many specific programming languages instead of looking only at the evolution of the term "programming languages" alone.

Temporal Correspondence. Our next focus is on considering user intent by suggesting different facets of queried objects and letting users choose any of them for biasing the temporal counterpart search. For example, to find temporal counterpart of current politician one can focus on different aspects of that person such as opinions on economy, international politics or even his persona life. Depending on different facets used one can detect different counterparts of the entity in the past. This study will be also extended to finding similar or related news events in the past to a given present news. Such extension should support prediction tasks on the future progress of developing events.

4. 評価

(1)自己評価

(研究者)

The initial plans of this research assumed exploratory character since the research area is still quite new. The scope of our work has then become quite large. In the process of conducting the project we have been finding interesting and useful research directions which we pursued. Our work is thus not limited to one or few narrow domains but has rather broad character. Thanks to this, solutions and ideas from one topic can support and complement those in the other topic while the broad perspective allows for selecting most interesting and promising directions. We managed to publish the results of core subtopics in well-respected conferences or journals, and the results of other subtopics are to be published later. In general, there are still many open avenues for further research within this context that are exciting and

useful, and which we plan to continue exploring in the near time.

Many of proposed approaches required processing and managing large data such as large collections of news articles, Google Ngrams based on results of Google Books projects or crawled tweet collections. We have thus necessarily spent much effort on creating efficient infrastructure. Thanks to the large scale data approach our results are reliable and have more impact.

Collaboration with history science researchers is an important issue. Interdisciplinary collaboration requires not only networking but also understanding requirements, expectations and possibilities of both sides. We have teamed with several historians, who are interested in computer-based solutions to the history field, to establish a series of interdisciplinary workshops in conjunction with international conferences on social informatics. The interdisciplinary character of such workshops assumes that each paper must be evaluated by professionals from computer science and from history science. Through these initiatives we had many chances to learn about interdisciplinary collaboration and to discuss our research ideas with history and social science researchers. We plan to continue more intensive collaboration in the future.

We want to emphasize that during the progress of this project, balanced effort went into conducting own research geared towards scientific publications and into offering community service. The latter consisted in preparing datasets for research community, constructing online services, co-organizing conferences and workshops, and completing survey article.

When it comes to administering the funding and maintaining the human network, both of these tasks went smoothly and without any bigger obstacles. The established human connections will be maintained in the future.

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

（研究総括）

新しい技術が生まれると多くの研究者は未来を描く。しかし、過去に思いを馳せる研究者が何人かいてもよい。本研究は Web や SNS を資料と捉え、その分析に基づく新しい歴史学を構想するものである。研究の目的は 2 つあり、一つは集団の記憶を分析することで、もう一つは過去について書かれた文書の分析ツールを開発することである。研究の過程で多くの技術が生まれている。例えば、文書が作成された時期と書かれた内容の時期を抽出する技術、文書間のリンク構造を調べ過去の影響を分析する技術、事物の呼称がどのように変遷してきたかを分析する技術などである。分析対象は、Wikipedia や Google Books などで、文書の規模は 1TB にものぼる。さらに、自ら研究するだけでなく、異分野の研究者からなるコミュニティを育ててきたことは高く評価できる。HistoInformatics と呼ばれる国際ワークショップが生まれ、そのコミュニティからこの分野で初めてのサーベイ論文が出版されている。さきがけに相応しい新しい研究テーマで、今後の展開が楽しみである。

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. A. Jatowt and K. Duh: A Framework for Analyzing Semantic Change of Words across Time, Proceedings of Digital Libraries Conference (JCDL 2014/TPDL 2014), IEEE Press, London, UK, pp. 229-238 (2014)
2. A. Jatowt, C.M. Au Yeung and K. 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)
3. A. Jatowt, E. Antoine, Y. Kawai, T. Akiyama: Mapping Temporal Horizons, Analysis of Collective Future and Past related Attention in Microblogging. Proceedings of the 24th International World Wide Web Conference (WWW 2015), ACM Press, (full paper), (to appear in 2015)
4. A. Jatowt and K. 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)
5. H. Joho, A. Jatowt, and R. Blanco: Temporal Information Searching Behaviour and Tactics, Information Processing and Management Journal (IPM) (to appear in 2015) (collaborative with NTCiR-11 "Temporalia" task organizers)

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. *External News Report*: L. Hoffmann, "Looking Back at Big Data", Communications of ACM (CACM)（April 2013）
   http://cacm.acm.org/magazines/2013/4/162509-looking-back-at-big-data/fulltext
   (this report provides introduction to computational history and mentions our research work on collective memory analysis as main example)
2. *Edited book (conference proceedings)*: A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. J. Flanagin, B. Tian Dai (Eds.): Social Informatics - 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013, Proceedings. Lecture Notes in Computer Science 8238, Springer 2013, ISBN 978-3-319-03259-7 (2013)
3. *Edited book (conference proceedings)*: K. Tuamsuk A. Jatowt, E. Rasmussen (Eds.): "The Emergence of Digital Libraries - Research and Practices" The 16th International Conference on Asia-Pacific Digital Libraries, ICADL 2014, Chiang Mai, Thailand, November 5-7, 2014, Proceedings. Lecture Notes in Computer Science 8839, Springer 2014, ISBN: 978-3-319-12822-1 (2014)

4. *Edited book (workshop proceedings)*: A. Nadamoto, A. Jatowt, A. Wierzbicki, J. L. Leidner (Eds.): “Social Informatics – SocInfo 2013 International Workshops, QMC and HISTOINFORMATICS”, Kyoto, Japan, November 25, 2013, Lecture Notes in Computer Science 8359, Springer 2014, ISBN 978-3-642-55284-7 (2014)