様式VI

博士学位論文調査報告書

論 文 題 目　History-related Knowledge Extraction from Temporal Text Collections（テキストコレクションからの歴史関連知識の抽出）

申請者氏名　　DUAN YIJUN

最 終 学 歴　　平成２9年　9月
京都大学大学院情報学研究科　社会情報学専攻修士課程　修了

令和２年 3月
京都大学大学院情報学研究科　社会情報学専攻博士後期課程
研究指導認定見込

学 識 確 認　　平成　　年　　月　　日（論文博士のみ）

論文調査委員　京都大学大学院情報学研究科
（調査委員長）　教　授　吉川 正俊

論文調査委員　京都大学大学院情報学研究科
教　授　鹿島 久嗣

論文調査委員　京都大学大学院情報学研究科
教　授　田島 敬史

論文調査委員　京都大学大学院情報学研究科
特定准教授 アダム・ヤトフト

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	DUAN YIJUN
論文題目	History-related Knowledge Extraction from Temporal Text Collections (テキストコレクションからの歴史関連知識の抽出)		

（論文内容の要旨）

History is a record of what has happened in the past. Many useful and diverse lessons can be learnt from history. For example, contemporary people rarely associate the past with advanced technology, however, looking at technology in the past can still teach us how modern technology connects to and came from the ancient world. Thanks to the accumulated large amounts of digitized documents from the past, it is possible now to employ large scale analyses for uncovering history-related knowledge. Many documents are primary sources giving direct descriptions of events as they happened in the past, while other documents contain chronologically ordered content describing histories of entities or accounts of past events.

Average users face the following fundamental challenges when trying to learn and understand the history. Firstly, with the rapid growth of the Web, more and more history-related documents are available causing severe information overload. Secondly, to understand the past information is sometimes difficult, especially for the young generations. Thirdly, there usually exist underlying patterns embodied in historical documents. To manually grasp such high-level and informative information would also require much cognitive effort. Lastly, although learning from examples and learning from comparison are both effective strategies extensively adopted in cognition and education, they are not easy to be applied to large numbers of diverse historical accounts. To overcome the above challenges, several research interests are proposed in the thesis.

Chapter 1 outlines the thesis, including the research background of history-related studies, motivation of the research, tasks involved and an overview of the thesis.

Chapter 2 introduces a novel type of summarization task consisting of generating gists of histories of multiple entities. Four methods are proposed which utilize diverse kinds of signals such as information about documents, eras, topics and correlation between events.

Chapter 3 introduces a novel research problem of categorizing entities into history-based categories for category characterization and understanding. To solve this problem an unsupervised approach is proposed based on a concise optimization framework.

Chapter 4 introduces a special kind of summarization task - Comparative Timeline Summarization and proposes effective approaches towards solving it. The unique character of the proposed summarization allows

capturing important comparative aspects of evolutionary trajectories hidden in two sets of compared timeline documents.

Chapter 5 approaches the problem of a special kind of summarization - summarization of past news articles stored in long-term news archives based on user issued queries. The comparative character of the proposed summarization allows finding important contrastive aspects in two temporally distant time periods. Users can benefit from such novel kind of access to historical document archives for needs including analyzing trends, determining historical analogies, as well as for educational or entertaining purposes, etc.

Chapter 6 proposes a novel research problem of automatically detecting across-time typical comparable entity pairs from two input sets of entities and introduces an effective method for solving it. A concise ILP model is used for maximizing the overall representativeness and comparability of the selected entity pairs.

Chapter 7 describes the novel task of diachronic document collection periodization. To address the introduced problem a two-step framework is proposed which consists of a joint matrix factorization model for learning dynamic word embeddings, and a well-defined optimization formulation for corpus periodization.

Finally, Chapter 8 draws a conclusion of the thesis and gives the discussion on the future researches.

（続紙 ２ ）

（論文審査の結果の要旨）

過去に蓄積された大量のデジタル化文書により大規模な分析が可能になってきている．しかし，利用者が大量のデジタル化文書から歴史を学習し理解しようとする場合，次の基本的な課題に直面する．まず，Webの急速な成長により深刻な情報過負荷が発生している．また，特に利用者の生誕前の過去の情報を理解することは容易ではない場合が多い．このような高レベルで有益な情報を手動で把握するには，多くの認知的努力も必要となる．本論文では，大量の歴史的デジタル文書から知識を抽出するために以下の六つの課題に取り組み，各課第について以下の成果を上げている．

第一に，複数のエンティティの履歴の要旨を生成する新しいタイプの要約タスクを提案した．そして，文書，時代，トピック，イベント間の相関関係などさまざまな種類の情報を利用する四つの方法を提案した．第二に，カテゴリの特徴付けと理解のために、エンティティを履歴に基づくカテゴリに分類するという新しい問題を導入し，この問題を解決するために，簡潔な最適化の枠組みに基づく教師なしアプローチを提案した．第三に，比較タイムライン要約というタスクを導入し，それを解決するために二つの比較対象の時系列文書集合に隠された進化の軌跡の特質を捉える効果的なアプローチを提案した．第四に，利用者が発行した問合せに基づいて長期ニュースアーカイブに保存された過去のニュース記事を要約する問題に取り組み，二つの時間的に離れた期間において重要で対照的な側面を見つけることができる要約の比較特性を提案した．第五に，エンティティの二つの入力セットから、時間の経過に伴う典型的な同等のエンティティ対を自動的に検出する問題を提案した．また，それを解決するため，選択したエンティティ対の全体的な代表性と比較可能性を最大化する簡潔なILPモデルを使用する効果的な方法を提案した．第六に，通時的な文書収集の期間化という新しいタスクに対処するために，動的な単語の埋め込みを学習するための行列因数分解モデルとコーパス周期化のための最適化で構成される2段階の枠組みを提案した．

以上，本論文では，過去に蓄積された大量のデジタル化文書から，歴史に関連する知識を抽出する問題に取り組み，成果を上げている．この研究成果は，大量のデジタル化文書からの多角的な歴史の学習と理解に資するもので，学術上，および，実際上，寄与するところが少なくない．よって、本論文は博士（情報学）の学位論文として価値あるものと認める．また、令和2年2月17日に実施した論文内容とそれに関連した試問の結果，合格と認めた．

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　年　　月　　日以降