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　　元 年　　６ 月 １２ 日現在

機関番号：１２７０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００２９６

研究課題名（和文）情報キュレーションマップ構築手法の検討と高度情報アクセスタスクへの応用

研究課題名（英文）Study on an automated method for generating information curation map and its application for advanced information access tasks

研究代表者

森　辰則（MORI, Tatsunori）

横浜国立大学・大学院環境情報研究院・教授

研究者番号：７０２１２２６４

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：意思決定におけるWebの利用が日常となる一方で、情報を取捨選択する負荷は大きい。ある主題に対し情報を収集・吟味し、分析結果を付記して、他者と共有するキュレーションサービスが注目されているが、人手に依存している。一方、文章間の関係を解説する文章がWeb上に少なからず存在する。我々は、文章間の関係を解説する文章を発掘し、文章を繋いでいくことにより得られる、文章を関係付けて理解するための情報複合体をキュレーションマップと呼ぶ。
本課題では、情報ナビゲーションを可能とするキュレーションマップの自動構築手法を開発するとともに、高度情報アクセスタスクの一例として大学入試論述問題の自動解答手法の開発を行った。

研究成果の学術的意義や社会的意義
一般のWeb検索等を想定し、キュレーションマップ生成に基づき、まとめ回答を順位付けして示すとともに、まとめ回答に張られたリンクにより詳細回答へ至る俯瞰的な可視化が行える仕組を提案し、その有効性を示した。また、現実的で複雑な質問応答を目指して、大学入試問題の自動解答を検討した。同マップは論旨構造を考慮した要約とみなせるので、論述問題を解く際の情報編纂過程に関連すると考え、各種知識源を編纂し世界史論述問題に自動解答する手法を検討し、その有効性を示した。
これらの成果は、利用者の能動的判断を支援するために、Web上の情報を効率よく整理しナビゲートする仕組を提供する点において、学術的・社会的意義がある。

研究成果の概要（英文）：Although Web is widely used for decision-making in everyday life, users have to make great effort to select useful information in Web. Curation service, which gathers information on a certain topic, selects good one, compiles a summary with useful comments, and promotes users to share the information, has drawn a great deal of attention. However, the compilation highly depends on the curators' skill. On the other hand, there exist not a few texts that describe the relation between two other texts. Therefore, by mining such texts, we expect that we can obtain a curation map, which makes us easily understand the relation among information pieces.
In this project, we study a method for generating information curation map, and a method for answering university entrance exams as an example of advanced information access task.

研究分野：自然言語処理

キーワード：情報キュレーション　自動要約　情報抽出　世界史論述問題　自然言語処理

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

さまざまな情報を World Wide Web (以下、Web)から得て、状況判断や意思決定を行うことが日常となっている。一方で、Web 上の情報は玉石混交であり、利用者による能動的な評価が必要とされるが、それは利用者に過大な労力を強いることが普通であるため、検索エンジンの出力する上位数件で、その情報を十分に吟味せず判断することも珍しくない。2011 年 3 月の震災、原発事故においては Web 上で情報が錯そうし、様々な誤解を招いた事実がこれを裏付けている。他方、Pariser が著書“The filter Bubble”(邦題「閉じこもるインターネット」)で指摘しているように、Google 等の検索エンジンや Facebook 等の SNS においては、利用者の個人情報に「合わせて」システム側が提供する情報を背後で選別をする際の、行き過ぎたパーソナライゼーションが問題になり始めている。すなわち、「読みたい文章(情報)」が優先されることにより、「読むべき文章」にたどり着けない。そのため、利用者による能動的な判断を支援するために、Web 上の情報を効率よく整理しナビゲートする仕組みの構築が急務である。

このような背景の下、Web 上の情報について、利用者各自が、広い視野の下で中立の立場から様々な情報を比較し、論理的・合理的に選別・分析できるように、批判的思考(Critical thinking)を促進し、意思決定の支援を行うシステムに対する期待が高まっている。その先駆的研究には WISDOM と情報信頼性判断支援システムがある。後者は我々が参加をした共同研究の成果である。ある言明(一つの命題に対応する文)の真偽を判断したい場合、機械自身にはその判断はできないという立場から、判断主体である利用者の支援をすることを目的とし、何が機械的にできるのかという観点から研究が行われた。その経験によれば、利用者にとって判断の際に本当に役に立つ情報とは、「人の判断」、すなわち、いま注目している話題に纏わる複数の言明について、他の人がどのようにして総合的に勘案して真偽判断を下したのか、であった。

このように、人が行った情報の整理や判断の結果を広く共有するための試みが行われている。古くは図書館における「パスファインダ」であり、話題毎に、利用者がそれを知る際に調べると良い文献が整理されている。近年では、「NAVER まとめ」や「Togetter」など「キュレーションサービス」が注目を集めている。この文脈におけるキュレーションとは、Web 上の情報を特定のトピックに沿って「人手」で収集・吟味し、分析・判断結果等を付記した文章を作成することであり、これを公開し他者と情報共有することで新たな価値を創出する。しかしながら現状では、文章の作成は投稿者の個人技に依存するところが大きい。一方で、次項で述べる調停要約の研究の知見によれば、複数の言明を整理して解説をしている、いわゆる「まとめ文書」がキュレーションサービス以外にも Web 上に少なからず存在する。そのため、注目トピックにおける検索結果などの複数文書間に、内容の類似性に基づく参照リンクを自動的に張ることにより、より詳細なトピックについて記された「詳細記述文書」を「まとめ文書」で繋いでいくことを繰り返せば、図 1 のような複数の情報を関係付けて理解するための情報複合体が得られる。我々はこれを「キュレーションマップ」と呼ぶ。

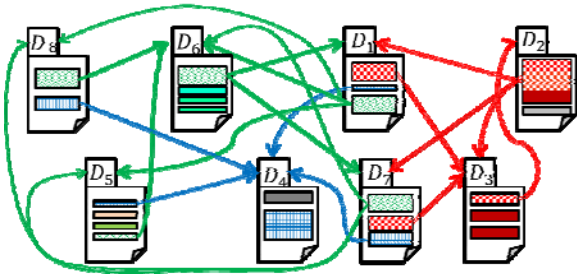


図 1 情報キュレーションマップ

２．研究の目的

本課題では、情報ナビゲーションを可能とするキュレーションマップの構築手法を考察するとともに、高度情報アクセスタスクへの応用を検討する。

(1) 本課題の前身となる研究プロジェクト(基盤研究(C)(平成 25～27 年度))においては、質問応答の結果文書に対し、複数のサブトピックを俯瞰する「まとめ文書」を上位に再順位付けする手法を提案している。この手法で得られた文書間のネットワーク構造自身が、情報キュレーションマップとして解釈可能であるが、多数の文書ノードが多数のリンクで結ばれた煩雑な構造となるので、そのままでは利用者に提供することは難しい。そのため、情報ナビゲーションを可能とする、情報キュレーションマップ構築手法を検討する。

(2) 現実的で複雑な質問応答を想定した取り組みとして、「ロボットは東大にはいれるか」プロジェクトや NTCIR QALab における社会系科目の入試問題を解くタスクがある。我々は両者に参画しているが、その過程で、情報キュレーションマップを、論旨構造を考慮した要約結果であるとみなしたときに、論述問題を解く際の情報編纂過程と関連があると考えた。そのため、世界史の論述問題等の高度情報アクセスタスクの解法についても検討する。

３．研究の方法

主要課題である a)「情報ナビゲーションを可能とする、情報キュレーションマップ構築手法」に加えて、b)「世界史の論述問題等の高度情報アクセスタスクへの応用」を検討する。これら

の課題について、次に示す各部分課題群への取り組みにより解決を試みる。
(1) a-1. 情報ナビゲーションを可能とする、ネットワーク構造の縮約手法の検討
(2) a-2. 情報キュレーションマップの精緻化手法の検討
(3) b-1. 情報の編纂と俯瞰的可視化への応用
(4) b-2. 情報要約への応用

４．研究成果

(1) a-1. 情報ナビゲーションを可能とする、ネットワーク構造の縮約手法の検討

Web 文書や質問応答の結果として得られた文書群を対象として、情報キュレーションマップの生成に関する検討を行った[学会発表①]。

図２に示すとおり、各文書内の文書断片、すなわち文章群の各々から、別文書に向けて、両者の間の内容包含度に基づき、自動的に参照リンクを張ったネットワーク構造において、各文書は、リンク元として見た時には、Hub ノードとして解釈され、リンク先として見た時には、ある特定話題を詳細記述した Authority ノードとして解釈される。ここで、リンク重みを持つネットワークに対する HITS アルゴリズムを適用すると、各文書について、Hub 値と Authority 値が得られる。Hub 値は、Hub ノードとしての良さを表すので、我々はそれを「まとめ文書」としての良さの尺度に利用する。なお、Authority 値は、多くの Hub ノードから参照されるノードに対して値が大きくなり、そのノード自身が有用な情報を有しているか否かを表すので、「詳細記述文書」としての良さの尺度に利用できる。

このネットワーク構造の生成手法において、ノードとリンクを取捨選択する手法を検討した。具体的には、類似するノード群をまとめ、代表ノードを提示する手法を検討した。ノード間の類似度として、被リンク構造の類似度と内容関連度の両者を考慮する手法を検討した。

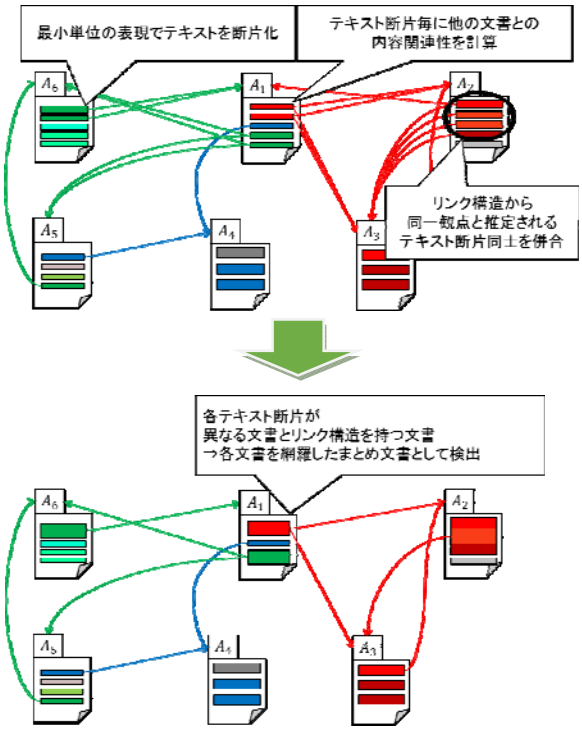


図２ 情報キュレーションマップの生成

(2) a-2. 情報キュレーションマップの精緻化手法の検討

(1)の生成手法において、同一観点の説明記述が別の文章であると認識され、そこから別のリンクが張られてしまうという、文章の過剰な断片化が観察されたため、その抑制手法の検討を行った。節を最小セグメントとし、最小セグメントから始め、リンク構造が類似するセグメントを漸進的に併合しセグメントを拡張するベースライン手法に対し、最小セグメントの再選定、リンク構造の類似判定条件の改善、セグメント併合条件の改善を行うと、図３に示すように観点の一貫性がある適切なセグメントが得られることが確認された[学会発表①]。

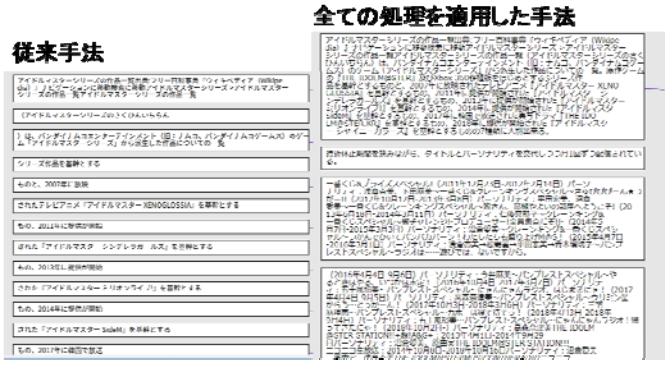


図３ 生成される文章断片の精緻化の様子

(3) b-1. 情報の編纂と俯瞰的可視化への応用

一般の Web 検索や質問応答を想定し、(1)(2)で述べた、キュレーションマップ生成に基づき、「まとめ回答」を順位付けして示すとともに、まとめ回答に張られたリンクにより「詳細回答」へ至る俯瞰的な可視化が行える可視化システムの検討と実装を行った。同システムの出力画面を図４に示す。まとめ回答を観点毎に動的に切り分け、各観点の文章からその「詳細回答」と

なる文書へのリンクを自動的に張ることができる。ある「詳細回答」文書を選択すると、それを「まとめ回答」と再解釈して、より詳細な観点に切り分け、再帰的により詳細な文書を得ることもできる[学会発表⑤]。

(4) b-2. 情報要約への応用

次の部分課題を検討するとともに、世界史論述問題解答器をオープンソースとして公開をした[雑誌論文③] [学会発表⑨⑭]。また、一般読者向けの解説書を出版した[図書①]。

b-2-1) 正解情報の整備

世界史論述問題については、人間作成の模範解答と、教科書等の知識源との間の対応付けを精密に記したコーパスを整備した[学会発表⑫]。多言語情報要約を想定して、複数のトピックにわたって日中英のツィートを収録し、感情分析等の情報を付与したコーパスを整備した[雑誌論文④]。

b-2-2) 各種知識源の利活用に関する検討

知識源から問題に関連する文群を取り出す際の再現率向上をめざし、深層学習等を用いて、問題文には明示されていないが、解答するにあたって非常に重要な重要語句を見出す手法について検討を行い、有効性について論じた[学会発表③⑧⑬]。

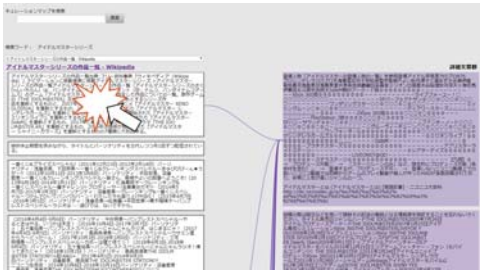
また、解答の際の知識源の一つである世界史用語集の語釈文の活用をめざし、見出し語を語釈文に適切に埋め込み整形する手法について、機械学習に用いる特徴量の精緻化ならびに擬似訓練事例の自動獲得の観点から精度向上を行うことを検討した[学会発表②⑥]。

b-2-3) 「まとめ文書」の利活用に関する検討

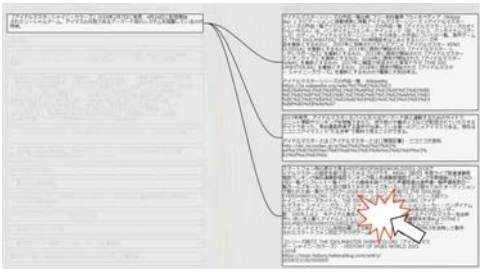
論述記述の骨組みを与える情報源として、ある国における重要イベントについて年代を追って記した「各国史」を「まとめ回答」として活用する方法を提案した。具体的には、問題に記された要求に照合する各国史の記述を「まとめ回答」としてイベントオントロジーの整合性の観点から抽出し、それを検索質問として、教科書の記述を検索し論述記述の素材となる文を得る手法を検討し、その有効性について論じた[学会発表④]。

b-2-4) 論述問題の解答に対する自動評価に関する検討

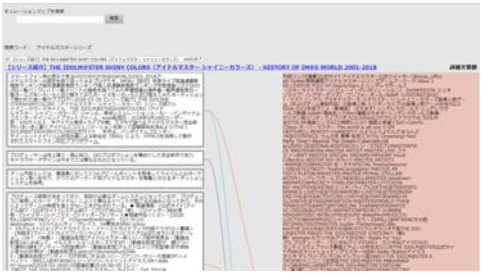
論述問題の解答に対する評価に関して、歴史分野に関する情報整合性(時間的・地理的整合性)に注目した評価尺度を検討し、有効性について論じた[雑誌論文①③] [学会発表⑦⑩⑪]。



(i) 左が、ある「まとめ文書」で、観点毎の文章に分けられている。右は紐づけられた「詳細文書」。「まとめ文書」中の、ある観点を説明する文章をクリックすると…



(ii)紐づけられた「詳細文書」のみがハイライトされる。ある「詳細文書」をクリックすると…



(iii) その「詳細文書」を「まとめ文書」として解釈し、左側に表示し、その中の各観点に紐づく「詳細文書」を提示する。同様にして、深掘りをしていける。

図 4 情報キュレーションマップ可視化システム

5．主な発表論文等

〔雑誌論文〕(計 4 件)

① Kotaro Sakamoto, 他 7 名 7 番目. Automatic Evaluation of World History Essay Using Chronological and Geographical Measures. Proceedings of 8th International Workshop on Evaluating Information Access (EVIA 2017), pp.20-23 (2017) 査読有

② Kotaro Sakamoto, 他 6 名 5 番目. FelisCatusZero: A world history essay question answering for the University of Tokyo's entrance exam. Proceedings of Open Knowledge Base and Question Answering Workshop at SIGIR (OKBQA 2017), pp.45-46 (2017) 査読有

③ Hideyuki Shibuki, 他 7 名 7 番目. Chronological and Geographical Measures for Evaluation of World History Essay QA in University Entrance Exams. In Proceedings of Open Knowledge Base and Question Answering Workshop at SIGIR (OKBQA 2017), pp. 35-40 (2017) 査読有

④ Yujie Lu, Kotaro Sakamoto, Hideyuki Shibuki, and Tatsunori Mori. Construction of a Multilingual Annotated Corpus for Deeper Sentiment Understanding in Social Media. Journal of Natural Language Processing, Vol. 24, No. 2, pp.205-266 (2017) 査読有

〔学会発表〕（計 14 件）

① 阿部穣太郎, 渋木英潔, 森辰則. キュレーションマップ自動生成手法におけるテキスト断片の適正化. 言語処理学会第 25 回年次大会 (2019)
② 大矢康介, 阪本浩太郎, 渋木英潔, 森辰則. 世界史用語集のゼロ代名詞の表層格推定における自動生成された擬似訓練データの利用. 言語処理学会第 25 回年次大会 (2019)
③ 飯塚章裕, 福原優太, 阪本浩太郎, 渋木英潔, 森辰則. 世界史大論述問題解答の自動生成に向けた指定語句による検索結果の分析. 言語処理学会第 25 回年次大会 (2019)
④ 福原優太, 阪本浩太郎, 渋木英潔, 森辰則. 世界史大論述問題を解くための質問応答システムにおける各国史の利用. 言語処理学会第 25 回年次大会 (2019)
⑤ 阿部穣太郎, 渋木英潔, 森辰則. キュレーションマップにおけるインタラクティブ UI の提案. 人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (2019)
⑥ 大矢康介, 阪本浩太郎, 渋木英潔, 森辰則. 世界史用語集の語釈文における見出し語に照応するゼロ代名詞の表層格の推定. 言語処理学会第 24 回年次大会 (2018)
⑦ 渋木英潔, 他 6 名 6 番目. NTCIR-13 QA Lab-3 における大論述問題の事例分析. 言語処理学会第 24 回年次大会 (2018)
⑧ Yuanzhi Ke, 他 6 名 6 番目. Deep Learning Method to Extract Implicit Keywords for Historical Essay Questions. 言語処理学会第 24 回年次大会 (2018)
⑨ Kotaro Sakamoto, 他 6 名 6 番目. Forst: Question Answering System for Second-stage Examinations at NTCIR-13 QA Lab-3 Task. The 13th NTCIR Conference on Evaluation of Information Access Technologies (2017)
⑩ Hideyuki Shibuki, 他 6 名 6 番目. Overview of the NTCIR-13 QA Lab-3 Task. The 13th NTCIR Conference on Evaluation of Information Access Technologies (2017)
⑪ 渋木英潔, 他 6 名 6 番目. 世界史論述問題の自動評価のための時間的・地理的情報の利用. 言語処理学会第 23 回年次大会 (2017)
⑫ 福原優太, 阪本浩太郎, 渋木英潔, 森辰則. 世界史論述問題における模範解答-知識源の対応を表すアノテーションの検討. 言語処理学会第 23 回年次大会 (2017)
⑬ 阪本浩太郎, 他 6 名 6 番目. 大学入試世界史論述問題における非指定重要語句生成に関する検討. 言語処理学会第 23 回年次大会 (2017)
⑭ Kotaro Sakamoto, 他 8 名 7 番目. Forst: Question Answering System for Second-stage Examinations at NTCIR-12 QA Lab-2 Task. The 12th NTCIR Conference on Evaluation of Information Access Technologies (2016)

〔図書〕（計 1 件）

① 新井 紀子, 東中 竜一郎(編). 第 3 章 森 辰則, 星野 力 他. 東京大学出版会. 人工知能プロジェクト「ロボットは東大に入れるか」: 第三次 AI ブームの到達点と限界. p. 296 (2018)

〔産業財産権〕

○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕

FelisCatus Zero-multilingual : オープンソース世界史論述問題解答器
https://github.com/ktr-skmt/FelisCatusZero-multilingual

６．研究組織