様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　３　日現在

機関番号：　１２７０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０２５４

研究課題名（和文）情報キュレーションのための能動的かつ対話的な情報抽出・要約技術に関する研究

研究課題名（英文）Study on active and interactive techniques of information extraction and automated summarization for information curation

研究代表者

森　辰則（Mori, Tatsunori）

横浜国立大学・環境情報研究科（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）　　3,800,000円

研究成果の概要（和文）：意思決定等における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。そこで、あるトピックに対し複数の情報を収集・吟味し、分析結果を付記した文章を作成、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。しかし、その作成は人手により個人技に依存している。一方、二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。そのため、注目トピックによる検索文章群から始めて、文章間の関係を解説する文章を発掘し、文章を繋いでいけば、文章を関係付けて理解するための情報複合体、キュレーションマップが得られる。
本研究では、同マップを利用者が構築していく際に必要となる各種技術の開発をおこなった。

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 creation of the information 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 set of methods that are needed for users to construct curation maps.

研究分野：　自然言語処理

キーワード：　情報キュレーション　情報抽出　自動要約　テキストマイニング　自然言語処理

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。さもないと、流言、風評に踊らされる。しかし、そのような能動的な評価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。最近では、2011 年 3 月 11 日の震災ならびに原発事故以降これらの情報が Web 上で錯そうし、様々な誤解を招いてきている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」にたどり着くことができない。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する仕組みの構築が急務である。なお、「文章」は文列であり、文書の一部である。

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。最近では、「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、我々の先行研究の知見によれば、複数の言明の成立条件を整理して解説をしている文章が Web 上に少なからず存在する。そのため、注目トピックによる検索文章群を初期情報とし、文章の間の関係を解説する新たな文章を自動的に発掘し、所与の文章を繋いでいくことを丹念に繰り返していけば、図 1-1 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を得ることができると考えられる。我々はこれを「キュレーションマップ」と呼ぶ。

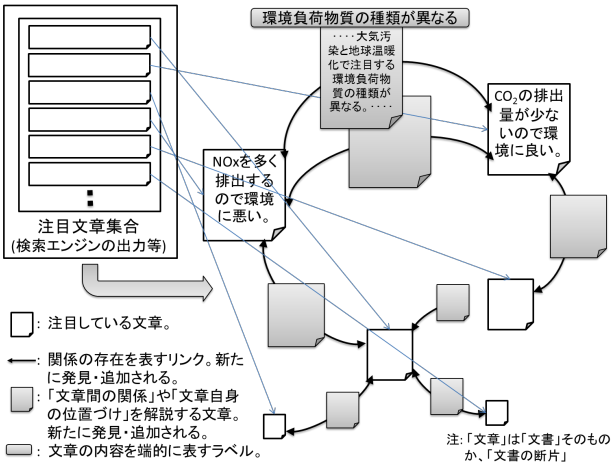

図 1-1 文章間関係を読み解くためのキュレーションマップの生成

２．研究の目的

本課題では、情報キュレーションの支援を目的として、キュレーションマップを利用者が構築していく際に利用可能な、能動的かつ対話的インタフェースを実現するために、必要となる情報抽出・要約技術を開発する。

特に、主要課題である「解説文章のマイニング」に加え、解説文章の要点を短い表現にする「解説文書へのラベル付与」を検討する。

３．研究の方法

対話的にキュレーションマップを生成するタスクの枠組みは、概略、次の段階からなると想定する。まず、準備段階に相当する、1) 注目しているトピックに関する文章群の収集(情報キュレーション過程における「集成」に対応)、2) 各文章からトピック関連の主要部を抽出(同「蒸留」に対応)、3) 蒸留された文章群に対してできるかぎり解説文章を発見して追加する初期マップの自動生成、がある。次いで、対話段階では、4)追加された解説文章に対して取捨選択等、利用者が編集を行う段階(同「蒸留」に対応)、5) 利用者が 2 文章を選ぶと、その間の関係を解説する文章を発見し、新たなノードとするとともにこのノードから選択された 2 文章へのエッジを生成する段階(同「マッシュアップ」に対応)、6) 利用者が 1 文章を選ぶと、これに関する解説文章を発見し、新たなノードとするととも

に、その解説文章に示される関係が成立する新たな文章をノードとして追加・リンクする段階(同「集成」と「マッシュアップ」に対応)、7)利用者が独自の視点から自ら解説文章を執筆し、新たなノードとするとともにこのノードから注目している文章へのエッジを生成する段階(同「上昇」に対応)等が利用者からの入力に応じて繰り返し生じる。

上記枠組みで検討すべき技術課題の要は、各段階の状況に応じた「解説文章のマイニング」と「解説文書へのラベル付与」である。前者は図 1-1 における追加文章(網掛け)とそれから張られるエッジの発見にあたる。後者は、図 1-1 における「環境負荷物質の種類が異なる」のように解説文章の要点を短い表現で表したものを生成しラベルとして解説文章に付与することである。このラベルは、インタフェース上での一覧性を高めるために、解説文章の表示代替物として機能する。

以上述べた技術課題について、本課題では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ものであった。

a. 解説文章のマイニングを実現するための一般モデルの検討
b. 解説文章に対する要点を表すラベル生成のための一般モデルの検討
c. 所与の 2 文章間の関係を解説する文章のマイニング手法の検討(比較型)
d. 所与の 2 文章間の関係を解説する文章のマイニング手法の検討(一般関係型)
e. 所与の 1 文章と関係のある文章のマイニング手法の検討
f. キュレーションマップ生成タスクにおける対話的インタフェースの検討

４．研究成果

(1) 平成 25 年度の成果

本年度は、

H25-1)解説文章のマイニングための一般モデルの検討
H25-2)解説文章の要点を表すラベル生成のための一般モデルの検討
H25-3)所与の二文章間の関係を解説する文章のマイニング手法の検討(比較型)（第一期)

を実施する計画であった。

H25-1, H25-3 については、我々が提案した新しい要約の概念である調停要約(対立する二言明について、両者の成立条件等を解説している文章を要約として見つける)を中心に検討を進め、情報信憑性判断支援のための要約生成タスクにおけるアノテーション手法[雑誌論文③]、対話型条件結論マップ生成に向けた条件と結論の抽出手法について提案[学会発表⑧]をし、査読付き学術雑誌論文等により公表した。さらに H25-1 に関して、Web 文書に対する出典抽出手法[学会発表⑩]、文脈に応じた用語解説の抽出手法[学会発表⑪]、電子掲示板における利害に関する書き手の立場の推定手法を提案し、学会大会発表により公表した。

一方、H25-2 については、計画段階では、一つの解説文章に対して、要点を表すラベルを生成するということを検討するものであったが、本年度の検討の結果、この機能を含みつつ、単一文章に対してではなく、複数の解説文章群に対して、要点を与えるより汎用な手法を検討した。あるトピックに関する解説文章「集合」があったときに、そこには複数の観点が含まれる。その中には、ある単一の観点で述べられている詳細な解説記事もあり得るし、一方で、複数の観点にわたって概要を示すような、まとめの解説記事もあり得る。そこで、まとめの解説記事を選びつつ、他の詳細解説記事との対応関係を同時に明らかにできる手法(以降、H25-2 改と記す)を考案し、対外発表により結果を公表した[学会発表⑨]。

(2) 平成 26 年度の成果

本年度は、

H26-1) 所与の 2 文章間の関係を解説する文章のマイニング手法の検討(比較型)（第二期)
H26-2) 所与の 2 文章間の関係を解説する文章のマイニング手法の検討(一般関係型)

を実施する計画であった。

研究提案時、H26-2 では、H26-1 の「比較」を含む各関係で二文書間の解説をする文章を個別に抽出する手法を検討する計画であったが、前年度の研究成果として得られた「複数の解説文章群に対して、まとめ文章を抽出し、要点を与える」手法が、複数文書間の解説をする文章をマイニングする汎用の手法として利用できると期待されたので、本年度はその拡張を更に検討することとした(H26-改)。

また、引き続き検討すべき基礎的課題として、H25-1)解説文章のマイニングための一般モデルの検討があり、これは、平成 27 年度の、H27-1)所与の 1 文章と関係のある文章のマイニング手法の検討に繋がるものである。

H26-改については、平成 25 年度に公表した「質問応答におけるまとめの観点からの回答の順位付け手法」を受け、質問応答の機能を前提とせず、一般の Web 文書を対象としてまとめ文章を発見し、文書間の構造を可視化する手法を検討した。対立した複数の意見をまとめる要約技術についても検討し、査読付き国際会議論文[雑誌論文②]に採択された。

H25-1,H27-1 については、大学入試問題を題材としたより実践的な情報アクセス手法、ならびに、文脈に応じた用語解説の抽出・分類手法について検討し、研究会や学会大会発表により公表した[学会発表④⑤⑥]。特に後者については、人工知能学会 2013 年度研究

会優秀賞を受賞し、記念招待講演を行った[学会発表⑦]。

(3) 平成 27 年度の成果
本年度は、

H27-1) 所与の 1 文章と関係のある文章のマイニング手法の検討
H27-2)キュレーションマップ生成タスクにおける対話的インタフェースの検討

を実施する計画であった。
研究提案時、H27-1 では、より基礎的課題である H25-1) 解説文章のマイニングための一般モデルの検討を行いつつ、所与の 1 文章を固定した状況において、解説文章の発見を行う手法を検討する予定であったが、これまでの研究成果として得られた「複数の解説文章群に対して、まとめ文章を抽出し、要点を与える」手法が、複数文書間の解説をする文章をマイニングする汎用手法として利用可能という知見に基づき、H27-1 と H27-2 を統合し、汎用な対話型の可視化インタフェースを検討した。
具体的には、平成 26 年度に検討した「一般の Web 文書を対象とした『まとめ』の観点からの文書の順位付け手法」を受け、これを基本エンジンとすることにより、キーワード列を入力として受け付け、関連文書を検索した後、その中で、包括的な理解に役立つまとめ文書を発見しやすくするとともに、まとめ文書を起点として文書間の対応関係を可視化し、より詳細な情報をマップ状に得られる対話型検索結果可視化インタフェースを開発した。同成果は、査読付き国際会議論文に採択され[雑誌論文①]、学会大会でも公表した[学会発表①]。
また、昨年度に引き続き、H25-1,H27-1 について、大学入試問題を題材としたより実践的な情報アクセス手法を検討した。特に、自由記述の回答となる大学入試の二次試験を対象とし、教科書や参考書等を知識源としたときの、情報抽出・要約に関する技術を検討し、学会大会等で公表した[学会発表②③]。

(4) 研究成果の概要図
主要な成果について、その概要図を示す。

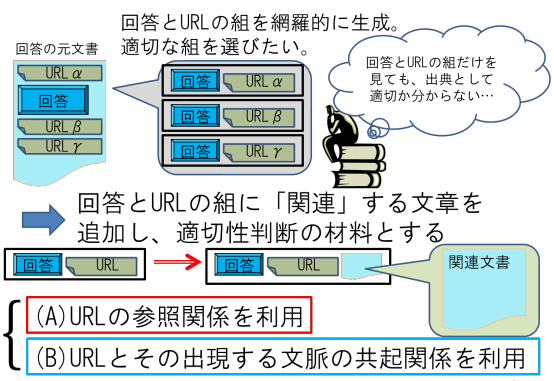

図 4-1 Web 文書に対する出典抽出手法

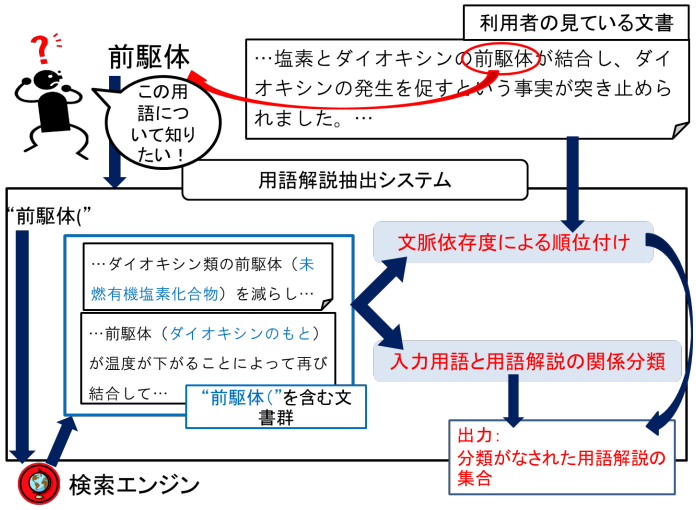

図 4-2 文脈に応じた用語解説の抽出手法

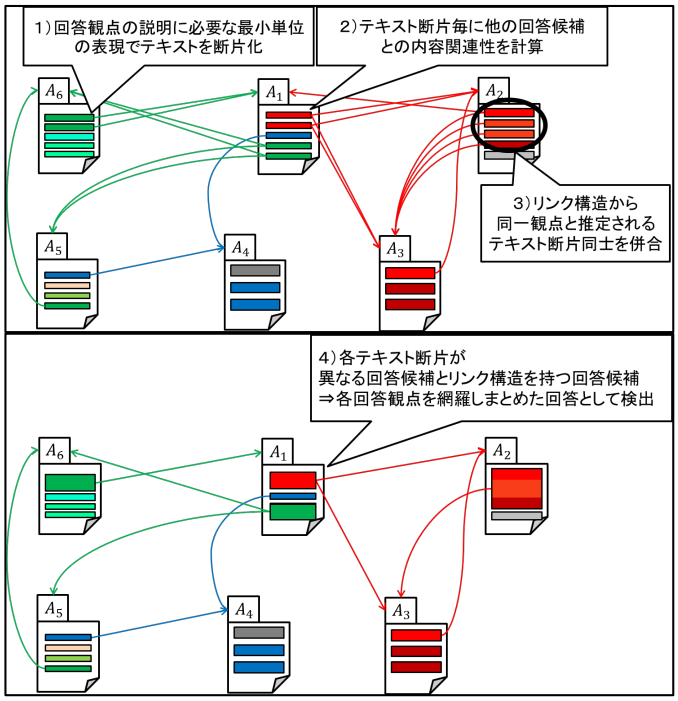

図 4-3 複数の解説文章群に対して、まとめ文章を抽出し、要点を与える手法

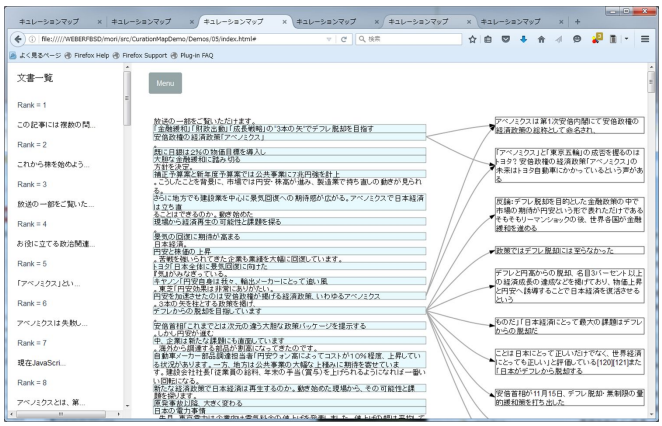

図 4-4 まとめ文章を抽出・可視化する汎用な対話型の可視化インタフェース

５．主な発表論文等

〔雑誌論文〕（計 3 件）

① Kotaro Sakamoto, Keiichi Nagao, Hayato Kobayashi, Hideyuki Shibuki, Tatsunori Mori, and Noriko Kando. Re-ranking answer candidates based on exhaustiveness of variety of answer viewpoints in non-factoid QA. In Proceedings of Web Question Answering: Beyond Factoids, WebQA'15 co-located with The 38th Annual SIGIR Conference (SIGIR 2015), (**ページ番号は付与されず**), (2015), http://sigir2015.org/content/workshops/webqa, **査読有**

②Jinlong Guo, Yujie Lu, Tatsunori Mori, and Catherine Blake. Expert-Guided Contrastive Opinion Summarization for Controversial Issues. In Proceedings of The 3rd International Workshop on Natural Language Processing for Social Media (SocialNLP 2015), pp. 1105-1110, (2015) , **査読有**

③渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則. 情報信憑性判断支援のための Web 文書向け要約生成タスクにおけるアノテーション. 自然言語処理, Vol. 21, No. 2, pp. 157-212, (2014), **査読有**.

〔学会発表〕（計 17 件）

①小林隼人, 小笹哲哉, 渋木英潔, 森辰則. 観点毎の詳細度を考慮したネットワーク構造の発見に基く Web 文書群の関係の可視化. 言語処理学会第 22 回年次大会, 2016 年 3 月 10 日, 東北大学川内北キャンパス(宮城県・仙台市).

②阪本浩太郎, 中山周, 渋木英潔, 石下円香, 森辰則, 神門典子. 東大入試世界史第 1 問(大論述問題)を解く質問応答システムの検討. 言語処理学会第 22 回年次大会, 2016 年 3 月 9 日, 東北大学川内北キャンパス(宮城県・仙台市).

③阪本浩太郎, 石下円香, 藤田彬, 渋木英潔, 狩野芳伸, 三田村照子, 森辰則. 大学入試の世界史論述問題における質問応答システムの自動評価に関する一考察. 情報処理学会自然言語処理研究会, 2015-NL-222(13), 2015 年 7 月 16 日, 首都大学東京秋葉原サテライトキャンパス(東京都・千代田区).

④阪本浩太郎, 渋木英潔, 石下円香, 森辰則, 神門典子. 大学入試の論述問題を解く質問応答システムの検討. 言語処理学会第 21 回年次大会, 2015 年 3 月 17 日, 京都大学(京都府・京都市)

⑤ Kotaro Sakamoto, Hyogo Matsui, Eisuke Matsunaga, Takahisa Jin, Hideyuki Shibuki, Tatsunori Mori, Madoka Ishioroshi, and Noriko Kando. Forst: Question Answering System Using Basic Element at NTCIR-11 QA-Lab Task. The 11th NTCIR Conference on Evaluation of Information Access Technologies, 2014 年 12 月 9 日, 学術総合センター(東京都・千代田区)

⑥ Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly~Y. Itakura, Di~Wang, Tatsunori Mori, and Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. The 11th NTCIR Conference on Evaluation of Information Access Technologies, 2014 年 12 月 9 日, 学術総合センター(東京都・千代田区)

⑦本間康允, 渋木英潔, 森辰則. 利用者の状況に応じた用語解説抽出システムの提案とその実現に向けた検討. 人工知能学会合同研究会 2014・優秀賞記念講演(**招待講演**), 2014 年 11 月 21 日, 慶應義塾大学日吉キャンパス(神奈川県・横浜市).

⑧松本拓也, 渋木英潔, 森辰則. 情報信憑性判断支援のための対話型条件結論マップ生成に向けた条件と結論の抽出. 言語処理学会第 20 回年次大会, 2014 年 3 月 18 日, 北海道大学(北海道・札幌市).

⑨長尾慶一, 渋木英潔, 森辰則. non-factoid 型質問応答におけるまとめの観点からの回答の順位付け手法の提案. 言語処理学会第 20 回年次大会, 2014 年 3 月 18 日, 北海道大学(北海道・札幌市)

⑩櫻井大平, 渋木英潔, 森辰則. 質問応答システムにおける利用者の回答選択支援のための出典抽出手法の提案. 言語処理学会第 20 回年次大会, 2014 年 3 月 18 日, 北海道大学(北海道・札幌市)

⑪本間康允, 渋木英潔, 森辰則. 利用者の状況に応じた用語解説抽出システムの提案とその実現に向けた検討. 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, SIG-AM-06-05, 2014 年 3 月 15 日, 東京大学駒場 I キャンパス(東京都・目黒区). **2013 年度人工知能学会「研究会優秀賞」受賞**

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕

６．研究組織

(1)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

(2)研究分担者

(3)連携研究者