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

阿部穣太郎 †1　渋木 英潔 †1　森 辰則 †1

†1横浜国立大学
E-mail: {jotaro,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

意思決定に Web を用いることが日常となっている一方で、情報を取捨選択するためには利用者による能動的な評価が必要とされるがその負荷は大きい。このため、近年、「NAVER まとめ[1]」など、特定のトピックに沿って複数の情報 (文章) を収集・吟味し、分析・判断結果等を付記した文章を作成し、他者と情報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、現在、その作成は人手により、投稿者の個人技に依存する。一方で、同サービス以外にも、二つ以上の情報間の関係を解説するいわゆる「まとめ文章」が Web 上に存在する。そのため、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、より詳細な観点で記された「詳細文章」を「まとめ文章」で繋いでいくことを繰り返せば、文章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、この情報複合体を**キュレーションマップ**と呼び、文献 [1] で自動生成するための手法を提案している。

文献 [1] では、あるトピックに関する文書集合が入力として与えられた場合、まず各観点に分割されたまとめ文章（文書）を提示し、分割された個々の観点からその観点に対する詳細文書と推定された文章に対してリンクが張られたネットワーク構造（キュレーションマップ）を出力する。しかしながら、文献 [1] では、観点に対応した適切な長さのテキストに分割されておらず、まとめ文章の検出精度やキュレーションマップの可読性の低下を招いていた。それゆえ、本稿では、分割されたテキストを適正化することでこの問題の解決を試みる。

2　キュレーションマップ自動生成手法の概要

キュレーションマップ自動生成手法の基本的な考え方は以下の通りである。2 つのテキストを比較して、互いのテキスト中に同じ観点から書かれた記述[2]がある場合、そのテキスト間にリンクを張る。これをあるテキスト集合に対して繰り返すと、テキストをノードとしたネットワーク構造ができる。リンクが集中しているテキストは、他の多くのテキストと観点を共有していることとなり、それらのテキスト群をまとめているテキストとみなすことができる。したがって、このネットワークに HITS[2] や PageRank[3] などのグラフベースのランキングアルゴリズムを適用することで、まとめ文章としての重要度を計算できる。ここで問題は、ある観点に対応する記述の範囲をどのように決定するかである。我々は、観点として考えられる最小単位にテキストを一度分割し、その後、同じ観点をもつテキスト断片同士をまとめることで、適切な範囲となるテキスト断片を決定した。

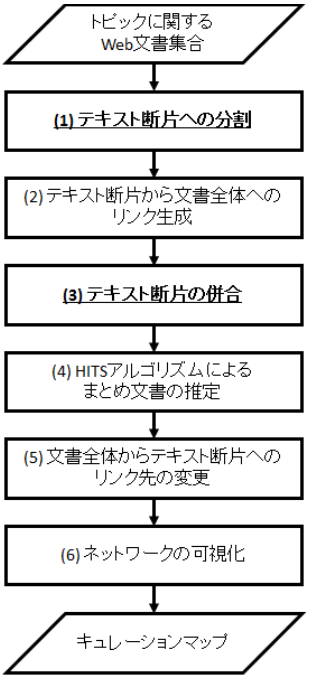

図 1: キュレーションマップ自動生成手法の流れ

図 1 に従来手法 [1] の流れを示す。まず、(1) 入力された各文書を観点を端的に表しうる最小単位（テキスト断片）に分割する。従来手法では、観点を表す端的な表現の最小単位を 1 つの述語項構造と仮定し、文書中のテキストを述語の直後で分割する。(2) 各テキスト断片から、単語の包含性に基づいて同じ観点を持つと判断された文書に対して重み 1 のリンクを張ることで、テキスト断片や文書をノードとしたネットワーク

[1] https://matome.naver.jp/
[2] 互いの記述の長さは問わない。

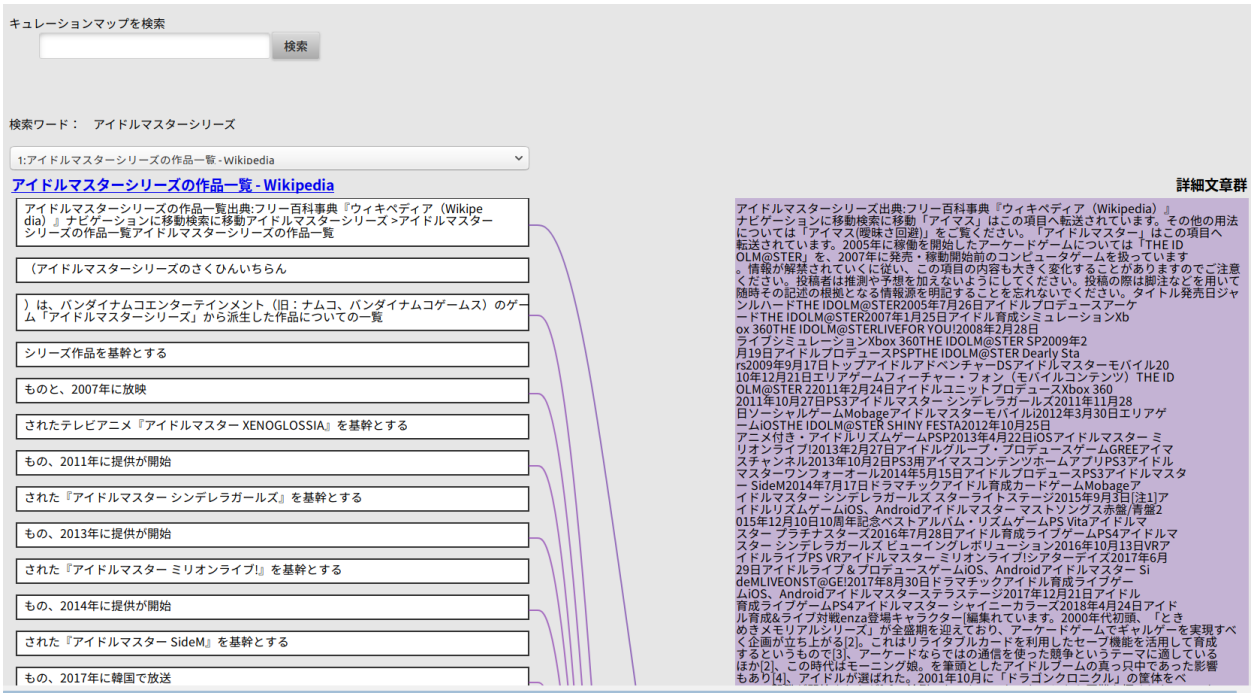

図 2: 従来手法によるキュレーションマップ（適正化前）

構造を構築する。(3) 連続した 2 つのテキスト断片を比較して、それぞれのリンク先の文書集合が同一または包含関係にある場合、2 つのテキスト断片は同じ観点を持つとして 1 つのテキスト断片に併合する。(4) 併合後のネットワークに対して HITS アルゴリズムを適用し、各ノードのまとめ文章としての重要度を計算する。(5) リンク元のテキスト断片とリンク先の文書内のテキスト断片とを単語の包含性に基づいて比較し、同じ観点を持つと判断された場合にリンク先を文書からテキスト断片に変更する。最後に、(6) ネットワーク構造を可視化することでキュレーションマップを提示する。

3 従来手法の問題点

従来手法 [1] によるキュレーションマップの例を図 2 に示す。図の例では、トピックとして「アイドルマスターシリーズ」が与えられている。図の左側には、まとめ文章が観点ごとに区切られて提示され、右側には観点ごとの詳細文章が提示されている。左側の観点ごとに区切られたテキストを見ると、必要以上に小さく断片化されており、人間にとって理解しやすいテキストであるとはいえない。

テキストの過剰な断片化は、図 1 中の下線の処理である「(1) テキスト断片への分割」と「(3) テキスト断片の併合」に原因があるため、本稿ではこれらの処理に修正を加えることでテキスト断片の適正化を行う。

4 テキスト断片の適正化

4.1 テキスト断片への分割

従来手法のキュレーションマップにおいてテキスト断片が読みにくい原因の一つとして、同じ観点について述べている一連の文章が、複数の別のテキスト断片として分割されていることが挙げられる。従来手法では、1 文に複数の観点があることを想定し、述語項構造を最小単位としてテキスト断片に分割している。しかしながら、述語項構造単位に分割されたテキストは可読性の点で問題があるとともに、述語項構造に分割することで、仮定条件の従属節と主節が別のテキストにされるなど、書き手が本来意図する内容と異なった内容になってしまう危険性がある。したがって、テキスト断片の最小単位を述語項構造から文へと変更し、より大きく読みやすい単位に変更する。この処理を**文単位化**と定義する。

4.2 テキスト断片の併合

ある程度の長さの文章を作成する際、書き手は同じ観点で書いた文を文章中に散在させるのではなく、まとめて記述することが一般的である。そのため、観点が不明なテキスト断片の前後に同じ観点で書かれた 2 つのテキスト断片が存在している場合、不明なテキスト断片を前後のテキスト断片と同じ観点であるとみなし、1 つのテキスト断片として併合する。これにより、同じ観点で書かれた文章中に内容的に関連性が低い文が混ざっていた場合でも、それらの文が 1 つのテキスト断片となることが期待できる。この処理を**挟み撃ち併合**と定義する。

表 1: トピック（検索クエリ）とその分野

分野	トピック
人物	コブクロ
スポーツ	ビリヤード
地名	尾瀬
自然言語による質問	なぜ空は青いのか
政治経済	アベノミクス

表 2: まとめ文書検出実験の結果

手法	MRR
従来手法	0.65
提案手法	0.61
文単位化なし手法	0.75
挟み撃ち併合なし手法	0.68
重複文書併合なし手法	0.76

従来手法でテキスト断片を併合する際、リンク先の文書集合が同一または包含関係であることが併合の条件であったが、その条件では、併合されるべきテキスト断片が併合されない事例があった。それゆえ、リンク先に重複する文書が存在するに条件を緩和する。これにより、例外的に他の文書にもリンクが張られていることで包含関係が成り立たなかったテキスト断片同士が併合されるようになると期待できる。この処理を**重複文書併合**と定義する。

以上の 3 つの処理を加えて生成されたキュレーションマップの例を図 3 に示す。図 2 と同じトピックであるが、左側のまとめ文章のテキスト断片が内容を理解しやすい長さで分割されており、可読性が向上している。また、ネットワークの構造も大きく変化しており、HITS アルゴリズムによる重要度計算の結果である、まとめ文章にも影響が及ぶことが予想される。

5 実験

提案手法の有効性を示すために、以下の 2 通りの実験を行う。1 つ目は、まとめ文書の検出精度がどの程度改善されたかを目的とした実験であり、まとめ文書検出実験と定義する。2 つ目は、分割されたテキスト断片が人間の観点とどの程度一致するかを目的とした実験であり、観点分割実験と定義する。

使用した文書集合は、どちらの実験でも、表 1 に示す 5 トピックを検索クエリとして BingSearchAPI[3]で取得した Web 文書であり、トピックごとに上位 50 文書を入力とした。表 1 の 5 トピックは多様な分野から選ぶようにした。提案手法と比較するために、従来手法 [1]、提案手法から文単位化を除いた手法（**文単位化なし手法**）、提案手法から挟み撃ち併合を除いた手法（**挟み撃ち併合なし手法**）、提案手法から重複文書併合を除いた手法（**重複文書併合なし手法**）の 4 手法を用いた。

まとめ文書検出実験では、HITS アルゴリズムによる重要度ランキングの上位 5 件を出力とした。複数の観点が存在する文書を正解文書として第一著者が判定し、評価尺度として MRR（平均逆順位）を用いた。

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

[3]https://azure.microsoft.com/ja-jp/pricing/details/cognitive-services/search-api/

表 3: 観点分割実験の結果

手法	P	R	F
従来手法	0.09	0.95	0.16
提案手法	0.25	0.60	0.33
文単位化なし手法	0.18	0.67	0.28
挟み撃ち併合なし手法	0.17	0.91	0.29
重複文書併合なし手法	0.22	0.62	0.32

Q は全てのトピック、$rank$ は上位 5 件において初めて正解文書が出た順位である。

観点分割実験では、まとめ文書検出実験で正解文書と判定された上位 1 件を対象として、第一著者が観点ごとに分割し、その分割点を正解情報とした。評価尺度として、以下の式で適合率 P、再現率 R、F 値 F を用いた。

$$P = \frac{\text{一致した分割点の数}}{\text{出力文書中の分割点の数}} \quad (2)$$

$$R = \frac{\text{一致した分割点の数}}{\text{正解文書中の分割点の数}} \quad (3)$$

$$F = \frac{2PR}{P+R} \quad (4)$$

分割点が一致しているかどうかの判定は、テキストの分割に個人差が存在することを考慮し、正解の分割点との差が 50 字以内であれば正解とした。ただし、正解の分割点に対応する各手法の分割点は 1 つまでとし、1 対 1 対応が保証されるようにした。

6 考察

まとめ文書検出実験と観点分割実験の結果を表 2 と表 3 にそれぞれ示す。

表 2 から、まとめ文書検出実験において、提案手法から一部の処理を除いた、文単位化なし手法、挟み撃ち併合なし手法、重複文書併合なし手法の 3 手法は、従来手法の精度を上回った。したがって、それぞれのテキスト断片の適正化処理は、まとめ文書の検出精度の改善に一定の効果があったと考えられる。その一方で、全ての適正化処理を行った提案手法は従来手法よりも精度が低下した。この理由について考察する。ネットワーク構造を構築する際、単語の包含性に基づいてリンクを張るため、長い文章の方が多くのリンクを張

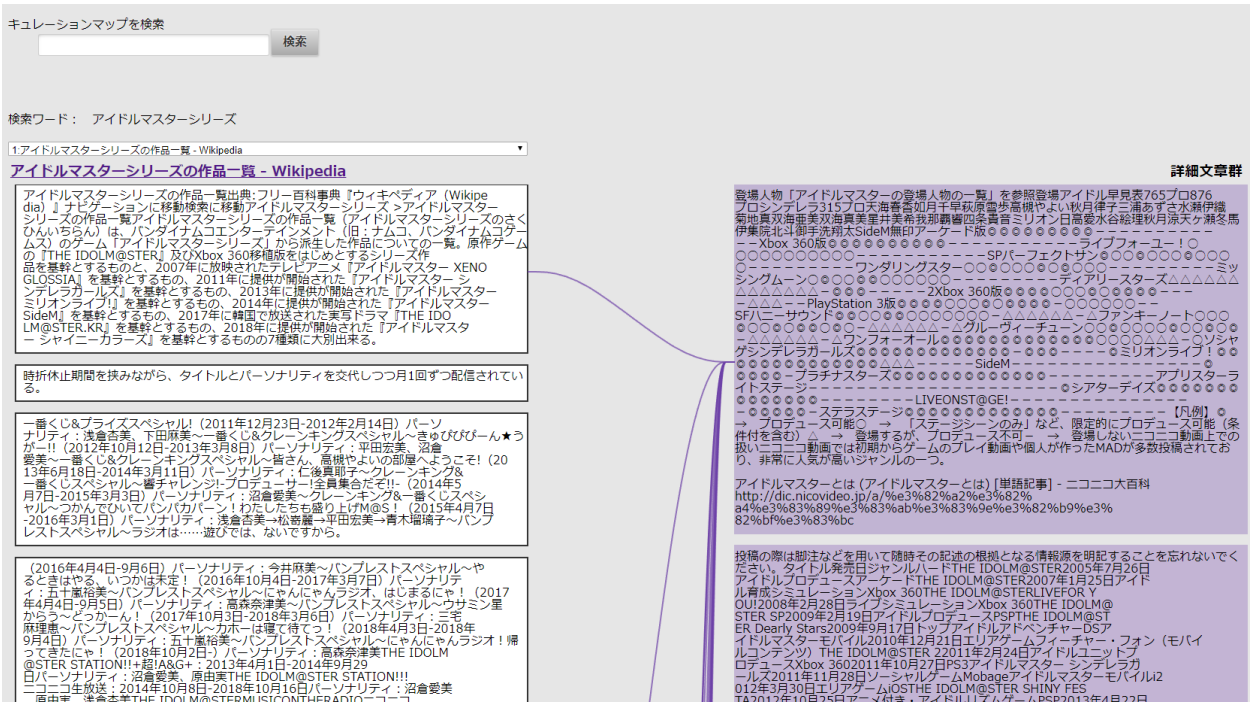

図 3: 提案手法によるキュレーションマップ（適正化後）

られやすい。また、多くの観点を含むまとめ文章は、その性質上、文章全体が長くなる傾向にある。結果として、従来手法においてテキスト断片が適切に併合されていなくとも、ある程度の精度でまとめ文書を検出できていたと思われる。一方、提案手法では、併合が促進された結果、テキスト断片の数が従来手法と比較してトピック平均で 1915.2 から 394.4 に減少した。その結果、ネットワークのリンク数が 1313.0 から 148.4 に減少し、検出精度が低下したと思われる。したがって、リンクの重みを文章全体の長さに基づいて正規化したり、併合の際にリンクの重みを変更したりすることで改善できるのではないかと考えている。

表 3 から、観点分割実験において、提案手法は従来手法と比較して、再現率が減少（0.95 → 0.60）したが適合率が大きく増加（0.09 → 0.25）し、F 値が 0.16 から 0.33 に向上した。3 つの比較手法の F 値から、文単位化、挟み撃ち併合、重複文書併合の順に効果が高かったことが分かった。挟み撃ち併合なし手法の再現率が 0.91 と従来手法と比較して減少度合が小さいことから、挟み撃ち併合が再現率を大きく低下させている。しかしながら、適合率の改善には挟み撃ち併合の貢献が最も大きかった。また、向上した提案手法の F 値も 0.33 とあまり高いものではないため、改善の余地があると思われる。重複文書併合なし手法の F 値は 0.32 と提案手法とあまり差がなく、まとめ文書検出実験で最も高い値（0.76）であったことを考慮すると、重複文書併合なし手法が総合的に最も良い結果だったといえる。

7 まとめ

本稿では、キュレーションマップ自動生成手法において、文単位化、挟み撃ち併合、重複文書併合の 3 つの処理によりテキスト断片の適正化を行うことで、まとめ文書の検出精度向上とキュレーションマップの可読性の向上を試みた。まとめ文書検出実験において、それぞれの適正化処理では一定の精度向上があったが、全ての処理を行った提案手法ではリンク数の減少から精度を改善することができなかった。観点分割実験において、F 値が 0.16 から 0.33 に向上し、文単位化、挟み撃ち併合、重複文書併合の順に効果が高かったことを確認した。今後、リンクの重みを変更するなど改善していきたいと考えている。

謝辞

本研究の一部は、JSPS 科研費 16K00296 の助成を受けたものである。

参考文献

[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 回年次大会発表論文集, pp. 1073–1076, 2016.

[2] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677, 1998.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. Proceedings of the 7th International World Wide Web Conference, pp. 161–172, 1998.