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

小林隼人[†1]　小笹哲哉[†1]　渋木英潔[†2]　森辰則[†2]

†1 横浜国立大学 大学院 環境情報学府
†2 横浜国立大学 大学院 環境情報研究院
E-mail: {hayato-k,kosasa,shib,mori}@forest.eis.ynu.ac.jp

1　はじめに

近年、インターネットの普及に伴い Web 上には電子化された文書が増大していおり、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負荷が大きい。よって収集した情報を分析し、情報の判断を補助する技術が求められている。

ここで、あるトピックに注目したときに、図 1 のように、Web 上の文書にはある観点について詳細な記述と、複数の観点についてまとめてある記述が存在する。

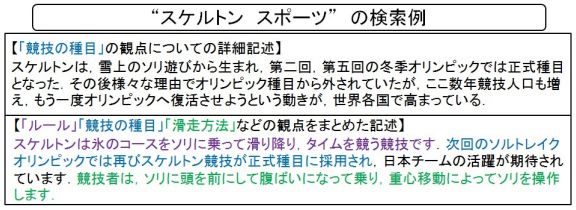

図 1　観点についてまとめた記述と詳細な記述

以上の背景に対し、Web 検索の結果を、多様な観点を網羅した文書と、その文書中の各観点について詳細な文書群に分けて提示することで、利用者が効率よく文書に記述されている内容を把握できる出力の組織化が行えると考える。そこで、本研究では、non-factoid 型質問応答システムの出力（回答候補となるパッセージ群）を対象に行われている、各回答中に現れる各観点の記述について内容関連性に基いて他の回答候補とのリンクを生成し、観点を網羅している回答候補を上位に出力するという研究 [1] を応用することを考える。

本稿では、まず、Web 検索結果の文書群と質問応答システムの出力の違いについて考察し、文書間に生成されるリンク関係を、より粒度の小さい単位で生成する手法を検討する。さらに、生成されたリンク関係を用いて、多様な観点をもつ文書と、各観点について詳細に記述された文書群との間の関係を可視化することを試みる。

2　関連研究

2.1　文書中の観点に注目した関連研究

non-factoid 型質問応答システムの出力（回答パッセージ群）を対象とする研究として、長尾ら [1] は回答候補集合の「まとめの観点」からの再順位付けを行っている。手法としては、回答候補パッセージを観点毎に区切ったテキストの断片と内容の類似している他の回答候補パッセージ間にリンクを生成する。生成されたリンク構造に対し、HITS アルゴリズム [2] に基づき各回答候補に重要度計算を行うことで、重要度による再順位付けがなされる。上位に順位付けされる多様な観点を提示し、その後各観点について詳細な記述をした回答候補を提示するという順序をつけることで回答全体をより効率よく理解できる出力の組織化を行っている。以下の図 2 は長尾らのシステムによって生成されるリンク構造である。ここで、回答候補パッセージを一つの文書とみなすと、ある文書内のテキスト断片から、(テキスト断片ではなく) 別の文書に対してリンクを生成している点に注意されたい。

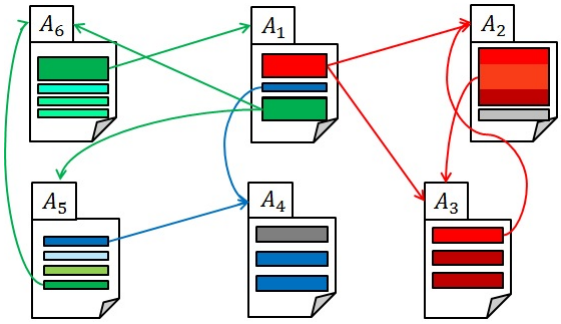

図 2　長尾らのシステムのリンク構造

また、NTCIR-11 の MobileClick タスク [3] は、与えられたクエリに対して 2 層の要約を出力するタスクであり、多くの情報を一度に見ることができないモバイルユーザにとって効果的な提示方法として、第 1 層にはクエリに関する概要的な情報とより詳細な情報へのリンクを提示し、リンクをクリックすることで遷移する第 2 層に詳細な情報の提示する。iUnit と呼ばれる重要な情報の断片を文書から抽出しランキングを行うタスクと、与えられた iUnit と文書群から 2 層の要約を生成するタスクに別れている。

2.2　本研究の位置づけ

本研究では、上記の長尾ら [1] の手法を Web 文書へ適用し、作成されるリンクによるネットワーク構造を用いて、文書群の可視化を行う。その結果は、上記 MobileClick タスク [3] で求められている 2 層の要約構造を縦横に連結したようなネットワーク構造になる。

3　長尾らの手法を Web 文書に適用する際の課題

長尾ら [1] の手法において、入力とする文書群を質問応答システムの出力（回答候補パッセージ群）から、Web 検索結果の文書群へ変更する際に考慮するべき点について述べる。

1 つには文書中に含まれる内容についての違いがある。前提として、質問応答システムよって出力される回答パッセージ群においては、各パッセージはある文書から質問の回答として抽出されているので、その内容は一貫して質問に関連している。一方で Web 上の文書を対象する場合、検索したクエリに関する情報は含まれているが、文書中の内容すべてがクエリに関する情報であるとは限らず、関係の薄い記述が含まれることもあり得る。

もう 1 つとして、文書中のテキストの長さに違いがある。質問応答システムでは、システムの出力である回答パッセージ群において、各パッセージは、ある文書から質問の回答として抽出された小さい文書断片であるため、その長さは比較的短い。一方で、Web 上の文書は、長さがまちまちであり、なおかつ、質問応答システムの出力よりもはるかに長いことが多い。

以上の 2 点を踏まえると、長尾ら [1] の手法を Web 上の文書に適用する上で大きな問題となりえるのは、ある観点を与えるテキスト断片から、その観点の詳細な記述を与える「文書全体」へリンクをはるリンク生成部分であると考えられる。テキスト断片からつながる他の文書へのリンク関係は、テキスト断片の内容に関して詳細な記述を与える文書とリンクを生成するという点に主眼をおいている。しかし、Web 検索結果の文書は、先に述べた通り、その全体がクエリに関連した記述であるとは限らず、長い文書であればあるほど、関係の薄い記述が含まれる可能性が高くなる。よって本研究では、ある文書のテキスト断片と他の文書の間にリンクを生成するのではなく、ある文書のテキスト断片と他の文書のテキスト断片の間にリンクを生成することを考える。これにより、各観点に対する詳細記述として正しい文書部分へとリンクを生成できると考える。

4　提案手法

4.1　提案手法の概要

3.1 節において検討した課題の解決方法として、長尾らの行った手法を Web 検索結果文書群に適用するために、ある文書のテキスト断片から張られたリンクの接続先を、他の文書全体から、他の文書の一部であるテキスト断片に変更する手法を採用する。

図 3 は上述のようにリンク先を変更した、目標とするシステムのリンク構造である。

本章では、長尾らの手法について簡単に述べ、提案手法と長尾らの手法の間で異なる点について詳しく述べる。また、生成されたネットワーク構造を用いて、文書群の可視化を行う。可視化については、6 章で詳しく述べる。

4.2　Web 文書間のリンクの生成

まず、各文書を観点を表す最小単位が一つの述語項構造であると考え、述語項構造を単位として分割する。これを初期のテキスト断片とする。

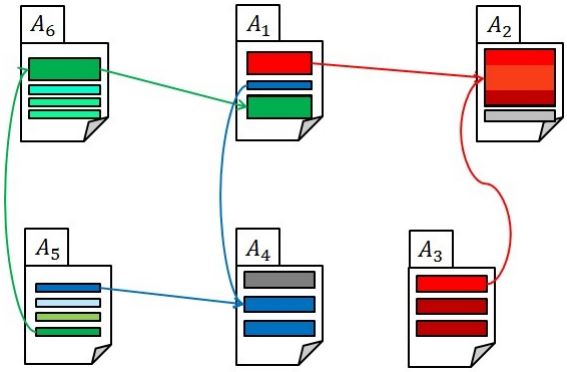

図 3　目標とするシステムのリンク構造

まずは、長尾らの方法と同様に、あるテキスト断片から別の文書へ内容関連性に基いてリンクを生成する。

文書 A_i の断片 P_{ij} と文書 A_k の内容関連性を，次の式による内容包含性により見積もる。

$$Comprehensibility(P_{ij}, A_k) = |\{w_n | w_n \in P_{ij} \& w_n \in A_k\}| / |P_{ij}|$$

この式は Runs ら [5] が文間の含意関係認識においてベースラインとして用いた尺度と同じであり，高村ら [6] も文間係数として用いている。

上記の式による内容包含性について、長尾らは閾値 α=0.6 以上のとき、リンクを生成していたが、3 章で挙げたテキスト量の違いを考慮し、本研究では閾値 α=0.67 以上の時、リンクを生成する。

4.3　テキスト断片の併合

初期のテキスト断片は、観点を表す最小の単位として分割されているため、一つの観点を表現した記述が複数のテキスト断片に分割されている場合がある。そのため、長尾らは 4.2 節で生成したリンク構造の類似性に基づき、テキスト断片を併合し、テキスト断片の大きさを適応的に変化させている。

長尾らの調査によると、同じ観点について記述されているテキスト断片は、そのリンク構造を比較すると、全く同一または、包含関係にある場合が多いことがわかった。そこで、ある文書中の連続するテキスト断片に対し、次の条件が成立するとき、 図 4 のようにテキスト断片の併合を行うことを提案している。

1. 一方のリンク構造が他方のリンク構造を包含する
2. どちらも全く同じリンク構造を持つ

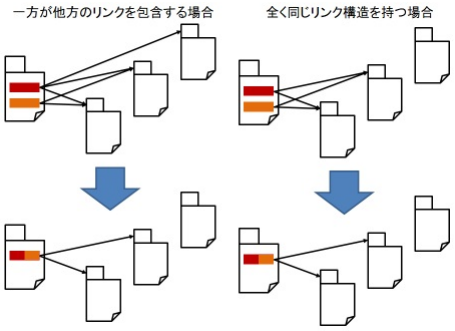

図 4　テキスト断片の併合

なお、併合処理により、各観点についての適切な大きさのテキスト断片が作成できるため、4.4 節で述べるテ

キスト断片へのリンク先の変更は併合処理の後に行う必要がある。

4.4 テキスト断片へのリンク先の変更

文書全体へのリンクをテキスト断片へとリンク先を変更する。

次の条件が成立するとき、リンク先を文書中の 1 つの観点に対して詳細である部分、つまりテキスト断片へと変更することで、リンク先の文書が長くなりすぎず、関連部分のみを詳細記述として扱うことができる。

1. リンク元のテキスト断片よりもリンク先のテキスト断片のほうが長い記述である。
2. 4.2 節で用いた内容包含性を見積もる式において閾値が β 以上である。（ここでは予備調査より β =0.6 とした）

4.5 HITS アルゴリズムに基づく文書の重要度計算

構築されたネットワークに対して，HITS アルゴリズムを適用した重要度計算を行う。具体的には，以下に示す式 2 つの式を用いて文書 A_i に対して Hub 値 ($HITS_H$) と Authority 値 ($HITS_A$) を付与する。

以下の式において、P_{ij} は A_i に含まれる j 番目のテキスト断片、$Out(P_{ij})$ は P_{ij} からリンクが張られた文書集合，$In(A_i)$ は A_i に対してリンクを張る文書集合を表す。また、$noOfFragment(A_i)$ は、A_i に含まれているすべてのテキスト断片の数を表し、$noOflinkedFragment(A_i)$ は、A_i に含まれているすべての断片の内、リンク構造を持つ断片の数を表す。このときのリンクとして、テキスト断片同士のリンク関係のみを用いることとする。

$$HITS_H^{t+1}(A_i) = \sum_{P_{ij} \in noOflinkedFragment(A_i)} \frac{max(HA_k | HA_k = HITS_A^t(A_k) A_k \in Out(P_{ij}))}{1+log(noOfFragment(A_i))}$$

$$HITS_A^{t+1}(A_i) = \sum_{A_k \in In(A_i)} \frac{HITS_H^t(A_k)}{1+\log(noOflinkedFragment(A_i))}$$

5 評価実験

5.1 実験目的

まず、リンク構造に基づく文書の順位付けが、まとめ文書を正しく上位に順位付けするかを調べる。

また、まとめ文書と詳細文書の間の関係の可視化については、テキスト断片間に生成されたリンクが正確であるかを調べる。

5.2 正解基準

まとめ文書の順位付けに関する評価のために、まとめ文書と判断する基準を定めた。評価実験にしようしない 3 つのクエリについて Web 検索の結果として得られた文書群に対して予備実験を行った。我々のまとめ文書の順位付けアルゴリズムを適用して得られた上位 20 件に対して、人出でまとめ文書であるか否かの判定を行った。その過程で客観的な判断基準として、以下の 3 つを定めた。なお、直接的にクエリに関する記述のみをクエリに関する情報、または文とする。

1. クエリに関する情報が文書全体の 5 割以上である
2. クエリに関する文が 5 文以上ある
3. 観点が複数個存在する

生成されたテキスト断片同士のリンクについて、可視化を行うことを念頭に見てみると、完全な詳細化ではなく、リンク元と同等の程度の情報であれば有用であると判断できた。よって、リンク関係の正解基準としては、リンク元のテキスト断片が表す述語項構造がリンク先のテキスト断片が表す述語項構造と一致していることとした。

5.3 使用データ

実験には以下の 5 つクエリに対して検索を行って得られる文書群のうち上位 50 件の文書群を対象とした。

表 1　検索クエリ

クエリの種類	検索に用いたクエリ
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
経済	アベノミクス

5.4 評価手法

再順位付けされた文書群に対する評価の手法として、全ての適合文書を高精度に検索するタスクのための評価尺度である、Q-measure[4] を用いることとした。具体的な計算式は以下に示す式である。

$$Q - measure = \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{cg(r)+count(r)}{cig(r)+r}$$

テキスト断片同士のリンク関係に対する評価の手法として、リンク関係に対して精度による評価を行う。

テキスト断片同士のリンク関係の評価には、再順位付けされた各クエリに対する文書群のうち、最上位の文書を用いた。

5.5 実験結果

再順位付けに対する評価実験結果、及びテキスト同士のリンク関係に対する評価実験結果を表 2 に示す。

表 2　再順位付け及びリンク関係に対する実験結果

検索クエリ	まとめ文書の順位付けの精度	リンク生成の精度
コブクロ	0.667	0.429
ビリヤード	0.719	0.261
尾瀬	0.841	0.685
なぜ空は青いのか	0.975	0.487
アベノミクス	0.864	0.469
全体	0.814	0.464

表 2 の結果より、本研究で用いた手法は、まとめの観点を持つ文書の抽出に有効であると分かった。また、リンクの生成の精度については、やや低調であると分かった。

6 文書群の可視化

提案手法によって得られたネットワーク構造を用いて、文書群を可視化するためのインターフェースを作成した。その表示例を図 5 に示す。

中央に “まとめ文書” として閲覧している文書、右側には “まとめ文書” 中のテキスト断片からのリンクが張られている “詳細文書” 群、左側には、まとめ文書らしさ、すなわち、HITS アルゴリズムによる重要度計算によって得られたスコアの高い順に文書群を並べて配置し、それぞれの文書について閲覧できるようした。

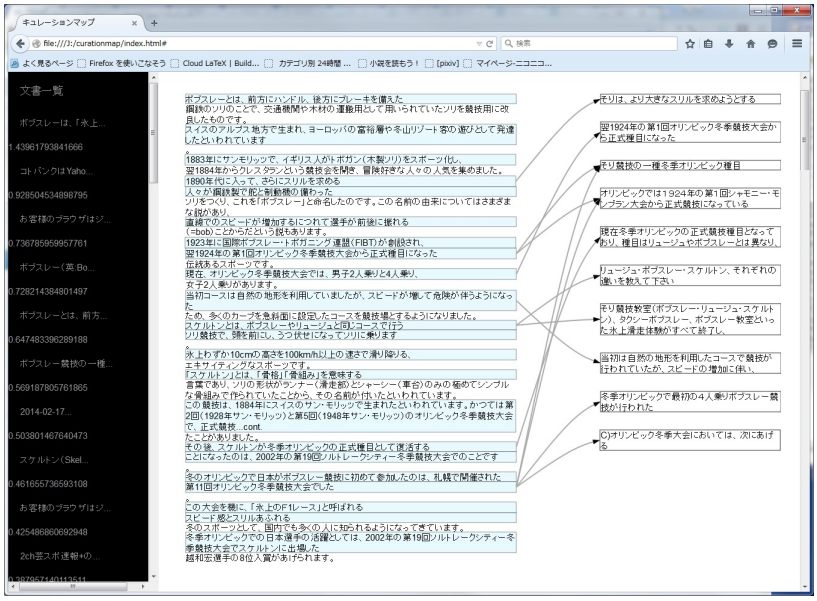

図 5　ネットワーク構造の可視化を行った結果

7　考察

実験に使用したデータはクエリに対する検索結果の上位 50 件を用いたが、まとめ文書の抽出に関する実験において好調な結果を示していたことから、上位 50 件中には多様な観点を含む文書が存在していることは確認できる。

リンク生成の精度はやや低調であり、リンクの生成基準の変更など、手法の改善が考えられる。また、生成されたリンクについては、正解となるリンクは多数あるが、それらをすべて表示してしまうと可視化結果の視認性が非常に悪くなる。そのため、表示するリンクを最小限に抑えるための、リンクの選定が必要であると考えられる。

可視化については、生成された文書間のネットワーク構造について、1 つのまとめ文書と、まとめ文書中の観点について詳細な文書群のみを出力しているが、現在の出力の可視化において提示できているリンク関係は、生成されているネットワークの一部分にすぎないので、より複雑なネットワーク構造も提示できる可視化手法の改善が考えられる。

8　おわりに

本稿では、Web 文書群に対し、観点を網羅しまとめた文書の抽出と、観点について詳細な文書群の関係の可視化を行った。今後の課題としては、表示するリンクの選定、文書群の関係の可視化手法の改善があげられる。

参考文献

[1] 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, 発表論文集, pp.93-96, 2014.

[2] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew S. Tomkins, “The Web as a Graph: Measurements, Models, and Methods,” Computing and Combinatorics, pp1-17, 1999.

[3] Makoto P.Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, Mayu Iwata, ”Overview of the NTCIR-11 MobileClick Task,” Proceedings of the 11th NTCIR Conference, December 9-12, 2014, Tokyo, Japan, pp.195-207, 2014.

[4] 酒井哲也, ”ひとつの高適合文書を高精度に検索するタスクのための評価指標,” FIT2005(第 4 回情報科学技術フォーラム) ,pp. 69-72, 2005.

[5] Masile Runs, Philip M. McCarthy, Danielle S. McNamara, Arthur C. Graesser, “A Study on Texual Entailment,” In Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence(ICTAI’05), pp. 326-333, 2005.

[6] 高村大也, 奥村学, “施設配置問題による文書要約のモデル化,” 人工知能学会論文誌 Vol.25, No.1, pp.174-182, 2010.