ネットワークを用いた複数テキストの要約方式の提案

豊浦 潤　津高 新一郎　瀬尾 和男
RWCP情報ベース機能三菱研究室

1 はじめに

大量文書を扱うための新たな情報処理技術の必要性が年々高まっている．我々は，この問題に対し，単語の出現頻度などの統計量に基づき文書とキーワードを空間配置し，ユーザはこれを鳥瞰しながら必要な情報にアクセスできる情報可視化散策システムの研究を行なってきた[1]．可視化システムで問題になるのは提示する情報量で，一時に多くの情報が提示できれば効率的だが，これが多過ぎるとユーザは把握理解が難しくなる．そのため，[1]のシステムでは，簡単なキーワードマッチングで表示する文書を絞っていた．

一方，ここ数年のインターネットの普及などによりアクセス可能な文書の量は爆発的に増大し，これらは，同一の情報ソースからの編集や引用を持つ文書を多く含んでいる．そのため，同一キーワードを持つという条件だけで文書検索をしても，実際に本文を参照してみると，金太郎飴の如く，どの文書も内容が殆んど変わらないという事態に陥ることが多くなっている．

こうした問題を解決するキーポイントは，増殖した文書の同位体を縮退させる技術であると筆者は考える．以下では，この縮退技術を実現するための方法として，複数文書の内容をネットワーク形式で表現し，その内容を要約的にユーザに提示する方法を提案する．

2 アプローチ

2.1 ベクトル表現

従来の文書処理システムでは，文書をキーワードの出現を表すベクトル表現で表すものが多かった．ベクトルモデルでは，文書の類似度は内積で表される．これは換言すれば，共起する単語を多く共有する文書は内容が近いという仮定に基づくものと言える．しかし，内容に踏み込んだ文書処理を実行するためには，キーワードベクトルによる文書コーディングでは情報量が不十分であることは自明である．

2.2 ネットワーク表現

ベクトル表現の限界を超えるためには，文書の文法情報や，文書間の参照関係などを織り込んだ文書コーディングが必要である．この目的を実現するために，われわれは，文書とキーワードを要素とし，これらの関係を要素間のリンクとして埋め込んだネットワークを構築し，これを利用した検索・可視化技術の研究を行っている[2],[3]．

ネットワーク表現は，文書やキーワード間の連想関係を自然に表現でき，可視化も自然な形で実現できるなどの利点がある．しかし，どのような情報を利用して，どのような方法を用いるのが良いのかといった具体性には，まだまだ欠けているため，今後の検討課題は多い．

2.3 係り受けネットワーク

現時点で自然言語処理技術により得られる文法的な情報の中で，単語間の関連性を明確に示すのが係り受け情報である．共起関係に比べ，係り受け関係がある単語間の関連性は，より確実である．

そこで，単語の係り受け関係を表現するネットワークを構築すれば，このネットワークは構造的に元の文書の内容を良く反映していると考えられる．一方，文書単位でネットワークを作らずに，複数の文書から1つの大きなネットワークを構築すれば，重複したキーワードが畳み込まれ，最初に述べた縮退が期待される．さらに，縮退が起こった部分は複数の文書で言及されており，重要な内容を含むと推察されるので，この部分の内容を用いて何らかの要約を作成できれば有用である．

3 アルゴリズム

係り受けネットワークの作成方法と要約方法について以下具体的に示す．なお以下では，前処理として文書を形態素解析，係り受け解析した結果，文節単位で係り受けが分かっているとする．

3.1 キーワードの抽出

まず，キーワードとして不適切な単語，たとえば接続詞や感動詞などからなる文節から除く．具体的には、連体詞、接頭辞、接尾辞、副詞、指示詞、接続詞、感動詞、判定詞、助動詞などが対象。ここで、助詞は文節間の関係を示すと考え残す。

次に、残った単語をルールを用いて連結する。これは主に分かれている名詞を連結して複合名詞とするものである。但し、名詞の品詞の細分類を参照し、人名と人名は連結するが、人名とサ変名詞は連結しないなど、簡単なヒューリスティクスを用いる。

最後に文節を代表する単語を決定する。これは、文節内に助詞以外に複数の単語が残っている場合に、キーワードを１つに絞るために行う。ここでもヒューリスティクスを用い、原則的に、固有名詞＞名詞＞動詞＞形容詞の優先順位とした。

3.2 ネットワークの作成

以上により、キーワード間の係り受け関係が得られる。これを用いキーワードをノードとし、係り受けを有方向のリンクとしたネットワークが作成できる。具体例を図1に示す。図に示したように、助詞はノード間の関係を示すものとしてリンクに貼られる。また、図には表示していないが、リンクには「その係り受けが有った文書のID」も貼られており、キーワードでなくキーワード間の関係から文書が参照できるようになっている。

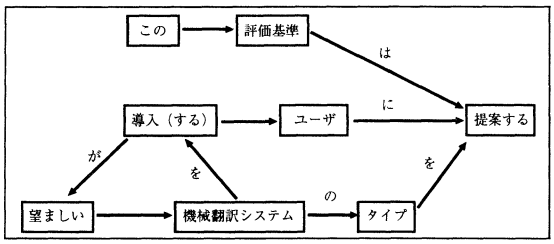

図 1: 係り受けネットワークの例

3.3 要約方法

以下で作成する要約は、洗練されたものでなく。どちらかと言えば、キーワードの羅列に毛が生えた程度のものに近い。しかし、不足情報の多くはユーザが補完することを期待している。

まず、要約を構成するための中心になるノードを探す。いまリンクの出る側のノードを葉、入る方向のノードを根と表現することにする。特に、ノードへ入る方向のリンクを１本経由する場合は、リンクの出る側を１次の葉、ノードへ入る方向のリンクを２本経由する場合は、リンクの出る側を２次の葉、などと呼ぶ。ここでネットワーク上で、入る方向のリンクが多いノード（構文木の根に対応）は話題の中心なってると考えられるので、そうしたノードを中心に要約を構成することにする。

次に、中心になるノードがいくつか選択されたあと、要約の範囲を決定する。具体的には、中心のノードから N 次の葉までを要約に含めことにして N により決まる。N が大きければ情報量は増えるが要約としては長くなり、ノイズが載ることも多くなる。

最後に選択した範囲のノードとリンクで、要約を生成する。具体的には、選択した葉のノードのうち高次の葉から中心のノードへ向けてリンクを辿ることで自動的に文を生成させる。図の例で、「提案する」を根として１次の範囲で要約すれば、「評価基準は、ユーザに、タイプを、提案する。」といった出力になる。

4 おわりに

以上、複数文書から係り受けネットワークを構築し、そのうち重要と思われる部位の要約を作成する方法を提案した。ここでは、述べなかったが、単語の出現頻度、リンクの頻度、表層格による重要度なども、提案した係り受けネットワークに埋め込み、抽出可能なので、今後、要約や検索に取り入れて行く方向で検討を進める予定である。

また、現在、NAIST の自然言語処理ツールを用いて、新聞記事１年分を対象に、ネットワークを構築して、要約を生成する実験を開始しているので、その結果についても今後報告する予定である。

参考文献

[1] H.Arita, T.Yasui and S.Tsudaka: “3D Stroller: Strolling in the self-organized information space”, *Proc.of RWC Symposium*, pp53-58, 1997.

[2] 豊浦潤, 岡隆一, “テキストの知識ベース化のための自己組織化ネットワークの提案”, 信学技報, *NLC96-59*, pp.23-30, 1997.

[3] 豊浦潤, 津高新一郎, 小中裕喜, “大量テキストのネットワーク表現と可視化手法の検討”, 情処秋季全大, *2L-01*, 1998.