語の共起に基づく文書検索と情報抽出

中谷　資隆　　田中　譲
北海道大学大学院工学研究科，知識メディアラボラトリー

1 はじめに

現在、Web 上には数十億ページに及ぶ膨大な文書が蓄積されている。コンピュータや Web 上の文書に対して検索を行う機会は増えており、ユーザが求める情報を効率よく素早く検索をする必要性が高まっている。文書検索においてもっとも一般的な手法であるキーワード検索では、ユーザが求めていない情報が大量に検索され、必要な情報にたどり着くまでに時間がかかる問題がある。その問題を解決するための一つの手法として語と語の共起を利用した検索がある。あるキーワードを含む文書集合に対して語と語の共起度を計算し、その結果を可視化することにより、その文書集合においてキーワードに関連が深い語を一目で把握することができる。これにより検索をより素早く直観的に行うことができる。このように、語と語の共起を利用して検索を行う方法が研究されている[1]。

従来の語と語の共起を利用した検索手法では、共起語は単語の頻度などに代表される重み付けによって計算されている。しかし、共起語としてみなされない、文書中に一度しか出現していないような重み付けの値が小さい語であってもその文書を特徴付ける重要な語であり、検索をする際の指針となる場合もある。また、重み付けの手法によっては一般的な語も共起語として出現し、適当な情報抽出ができないこともある。

本研究では、このような従来の検索手法では網羅しきれない特徴語を大量の文書集合の中から拾い上げる操作や共起語の中から適当な特徴語のみを抽出する操作を、双方向的かつ動的に行う Multi-Space Search と呼ぶ新しい検索手法とそれに適応する重み付けを用いて、文書検索や情報抽出を効率良く行うことを目的とする。

2 Multi-Space Search

本論文で提案する新しい検索の手法を”Multi-Space Search (MSS)”と呼ぶ。MSS では複数の検索情報を同時に用いて検索を行う。

Text Retrieval and Data Extraction based on Co-occurrent Words
Toshitaka Nakaya, Yuzuru Tanaka
Graduate school of Engineering and Meme Media Laboratory, Hokkaido University

従来の手法では、検索を行う場合には検索情報としてキーワードを含む文書のタイトルリストやキーワードに対する共起の語のグラフなどがそれぞれ一つずつ用いられた。MSS では、リストやグラフは一つだけに限らず、キーワードや重み付けの計算方法を変更することにより複数の異なる検索情報が提示することができる。検索情報はその他にも、Web のハイパーリンクの構造を利用した検索や、文書の書かれた年月日を利用した検索など他にも様々なものがある。MSS ではこれらの検索情報を相互に利用する、統合する、差をとるなどといった操作をユーザが直観的に双方向的に行い、それを繰り返すことで文書検索や情報抽出をすることを実現する。

本論文の基本的な考えは、MSS により、異なるキーワードや重み付けによって共起語のグラフを複数表示し、それらを統合する、差をとるなどの再計算を施すことにより、従来の手法では抽出することが困難だった特徴語や共起語のグラフを得て検索を容易に行うことである。

3 MSS を利用した検索

3.1 MSS のための共起度の計算

本システムでは、あるキーワードを入力するとそのキーワードを含む文書を検索し、任意の数の検索上位の文書集合の中でキーワードに対する共起語を計算し、結果を共起グラフとして可視化する。

共起語を抽出する上で採用している重み付けの計算方法を挙げる。文書正規化の計算方法として、コサイン正規化とピボット正規化[2]を採用している。また、共起語自体の計算方法としては、キーワードと他の語が共に出現している文書数の頻度を数え、それを単語頻度(*tf*)とみなす。そして、代表的な重み付けの方法である *tf*・*idf*、*log*(1+*tf*)・*idf*、*idf* を採用して計算している。これらの重み付けによって性質や結果の異なる共起グラフを抽出することができる。

3.2 Inverse Class Frequency

MSS に適用可能な新しい重み付けを一つ提案する。それは共起語の検索情報を複数個利用して共起語の中から特徴語を抽出するための重み付けであり、Inverse Class Frequency (*icf*)と呼ぶ。

それぞれの共起語の検索情報を一つのクラスと見なし、N個のクラスがあるとき、クラスの集合を $\mathbf{C}=\{C_1,C_2, \ldots ,C_N\}$とする。それぞれのクラスは共起語の集合から成り立つ。$\mathbf{C}$ の中で t を含むクラスの個数を $cf(t,\mathbf{C})$と定義すると、$icf(t,\mathbf{C})$は以下のように記述することができる。

$$icf(t,\mathrm{C}) = \log \frac{N}{cf(t,\mathrm{C})}$$

本論文では複数のキーワードに関連する共起語に比べて一つのキーワードにのみ共起する語の方がより特徴的な語であるという前提に立ち、*icf* を考案した。この重み付けでは語 t を含むクラスの個数が少ないほど、語 t の共起度が増す。特に同概念の語のクラス集合に対して *icf* を用いると、クラスに特有の特徴語の共起度が高められ、検索情報を改善することが可能である。

3.3 システムの構成

本研究で構築したシステムはサーバ・クライアント型のシステムである。Web サーバでは、計算や検索を高速に行うために、検索する文書集合のインデックスをあらかじめ作成する。インデックス作成には GETA[3]を使用している。GETA では WAM と呼ばれる疎行列を効率的に扱うデータ形式を用いてインデックスを作成しており、共起度の計算や類似文書検索を高速に行うことができる。これにより、双方向的かつ動的な検索を行うことが可能である。ユーザは検索インターフェースを用いて Web サーバに検索を要求する。Web サーバは WAM のインデックスを保持しており、内部で共起度を計算し、その検索結果を返す。

3.4 検索インターフェースと検索例

本システムの検索インターフェースと検索例の一つを示す（図１）。インターフェースでは重み付けの方法や共起語を出現させる数、重み付けの閾値を選択することができ、双方向的かつ動的な操作が可能である。さらに複数の検索情報を一度に表示することができ、MSS を行うことができる。この例ではコーパスとして毎日新聞の記事半年分、約 14,000 文書を使用している。

このシステムにおいて本論文で提案した *icf* を用いた検索例を一つ紹介する。キーワードとして「トヨタ」と「ホンダ」の二つのキーワードを送信し共起語の検索情報を抽出する。この二つのクラスに対して本論文で提案する *icf* の手法を適用する。「トヨタ」と「ホンダ」はどちらとも自動車会社であり、同概念の語であるため、*icf* の重み付けを用いるとそれぞれにのみ出現する単語の共起度が増し、結果としてこの例では「トヨタ」のみに関係する特徴語のみを抽出することができる。さらにこの共起グラフとタイトルリストを組み合わせて文書検索を行うことも可能である。このように、複数の検索情報を利用する MSS とそれに適応する重み付けである *icf* を用いて、より精度の高い特徴語を抽出し、文書検索を行うことができる。

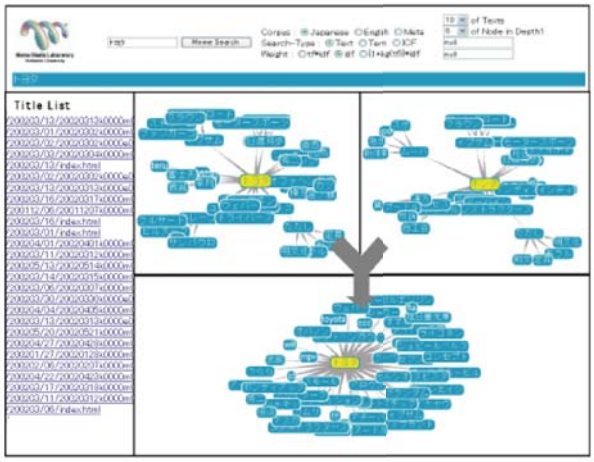

図１　検索インターフェースと検索例

4 まとめ

本研究では、従来の手法では抽出することが困難な特徴語を、複数の情報空間を統合するなどの操作を双方向的かつ動的に行う MSS とそれに適応する *icf* という重み付けによって、語と語の共起の観点から情報抽出や文書検索を行うことを示した。今後は MSS に適応する重み付けや検索インターフェースを改善し、さらに共起語以外の観点からの検索も適用して情報抽出、文書検索を行えるようにする。

参考文献

[1]Yoshiki Niwa, Makoto Iwayama, Toru Hisamitsu, Shingo Nishioka, Akihiko Takano, Hirofumi Sakurai, and Osamu Imaichi. : Interactive Document Search with DualNAVI. : First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 1999.

[2]A. Singhal, C. Buckley, and M. Mitra. : Pivoted document length normalization. : In *proceedings of SIGIR '96*, 1996.

[3] 高野明彦,西岡真吾,今一修,岩山真,丹波芳樹,久光徹,藤尾正和,徳永健伸,奥村学,望月源,野本忠司 : 汎用連想計算エンジンの開発と大規模文書分析への応用 : http://geta.ex.nii.ac.jp/