WWWホームページからの共起語自動抽出実験

1N－4

間瀬久雄† 徳田圭世† 森本由起子† 辻 洋† 丹羽芳樹‡
†（株）日立製作所 システム開発研究所
‡（株）日立製作所 基礎研究所

1．はじめに

我々は、キーワードによる情報検索においてシソーラスの利用が有効であるとの仮定に基づいて、シソーラス管理システムを開発している[1)]。本システムは、シソーラスデータの登録，参照，個人カスタマイジング，更新機能のほか、キーワード抽出技術を用いて特定の単語と対になって出現する傾向の強い単語（以下、共起語と呼ぶ）を検索対象文書群から自動抽出し、シソーラスの一部として登録する機能を持つ。情報検索において、これらのシソーラスデータを適宜ユーザに提示することにより、検索結果の絞り込みや検索の発想を支援できると考えている。

本報では、情報検索における共起語の有効性検証の一環として、日本のWWWページ45,000件から共起語を自動抽出する実験を行ったので報告する。

2．シソーラスと共起語

シソーラスとは、単語（概念）と単語（概念）との意味的関連を体系付けたものである。従って、単語間の関連を辿ることにより、ある単語（概念）に類似する単語（概念）が得られる。しかし、シソーラスは静的なデータであり分類学的な側面が強いため、ドメインの異なる情報の検索には不向きである。また、ある単語から「連想」される単語情報を得ることはできない。例えば、「大学」という単語に関連する単語として、一般のシソーラスからは「学校」や「国立大学」などの単語を得られても、「就職」「サークル」など人間が直感的に連想する単語情報は得られない。

我々のシステムで扱う共起語は、この「連想」される単語情報に相当する。共起語は、一般シソーラスを補うデータとして、情報検索支援に有効であると考えている。

3．共起語抽出方法

3．1 共起語の考え方

本報で言う共起語とは、文書を一単位としており、同一の文書に出現する傾向の高い単語は共起性が強いとしている。ここで、単語Xが単語Aに対する共起性を表す尺度（共起頻度）は次の式で算出する。

$$\frac{\text{単語A及びXが共に出現する文書数}}{\text{単語Xが出現する文書数}}$$

上式の値が大きいほど、共起の度合が強いとしている。また上式より、単語Aに対する単語Xの共起頻度と、単語Xに対する単語Aの共起頻度は異なる。さらに上式によれば、分母の値が小さい場合、共起頻度が大きくなりやすい傾向にある。しかし、共起頻度が1/1(=1)である単語と、90/100(=0.9)である単語とでは、直感的には後者の方が共起性が高いと言える。従って、最終的に共起語を抽出する際には出現頻度を考慮する必要がある。我々のシステムでは、出現頻度に応じて単語をクラス分けし、それぞれのクラス別に共起頻度の高い単語を出力する。

3．2 共起語抽出アルゴリズム

我々の共起語抽出ツールはテキストファイル群を入力とする。共起語抽出は大きく次の3つのフェーズからなる。

（1）単語抽出（形態素解析）

単語辞書（約15万語彙）を参照して文章を単語に分割し、名詞，サ変動詞語幹，未知語（辞書に未登録の語）をキーワード候補として抽出する。このとき、名詞または未知語の連続する語については複合語としてキーワード候補に含める。

（2）キーワード選定（ノイズの除去）

キーワードとして適切でない単語を除去するために、以下の4種類のフィルタを用いる。

（a）言語の種類に基づくフィルタ

本ツールは日本語文章を対象としているので、

表１　共起語抽出における単語統計データの推移

	文書数	単語レコード数	異なり単語数
単語分割後（フィルタ処理前）	44,710 文書	5,046,546ﾚｺｰﾄﾞ	718,239 種類
言語の種類に基づくフィルタ処理後	35,206 文書	4,292,129ﾚｺｰﾄﾞ	648,662 種類
字種・品詞に基づくフィルタ処理後	35,195 文書	3,810,373ﾚｺｰﾄﾞ	604,533 種類
不要語に基づくフィルタ処理後	35,190 文書	3,412,761ﾚｺｰﾄﾞ	602,755 種類
出現頻度に基づくフィルタ処理後（共起語抽出前）	35,151 文書	2,501,136ﾚｺｰﾄﾞ	62,324 種類

「米軍基地」	「エイズ」	「教育」	「ピザ」
強制使用 (9/9)	ｅｚｎｅｔ (13/13)	教育機関(96/96)	ピザまん(5/5)
国際都市形成構想 (5/5)	sampletext (13/13)	学校教育(73/74)	手作りピザ(5/5)
代理署名 (8/9)	薬害エイズ問題 (10/10)	教育研究(58/58)	タコス(3/5)
振興開発 (5/6)	薬害エイズ (9/9)	情報処理教育センター(51/51)	サンドウィッチ(3/5)
地位協定 (8/10)	エイズ問題 (5/5)	教育学部(127/133)	週休制(3/5)
安保条約 (4/5)	薬害エイズ殺人 (19/20)	教育課程(80/106)	内業(3/5)
日米特別行動委員会 (4/5)	総合案内センター (19/21)	Education(122/166)	調理補助(3/6)
実弾砲撃演習 (4/5)	洗脳クリニック (17/19)	社会科(79/109)	男女アルバイト(3/6)
ＳＡＣＯ (6/8)	薬害 (46/52)	養護(89/128)	日勤務(3/6)
米軍用地 (6/8)	ａｓｙｕｒａ (19/22)	教科(109/172)	ソフトドリンク(3/6)
・・・・・・	・・・・・・	・・・・・・	・・・・・・

図１　共起語自動抽出結果の例

英語文章はノイズとなる。ここでは、暫定的な処理として、ある割合以上の（次章の実験では８０%）英単語を含む文書を除去している。

（ｂ）字種・品詞に基づくフィルタ

記号やひらがな，古語からなる単語や，一文字からなるサ変動詞，ある文字数以上（次章の実験では２０字）からなる単語はキーワードになりにくいと考えられるので除去する。

（ｃ）不要語に基づくフィルタ

文章の分野に関わらずキーワードとなりえない不要語（次章の実験では日本語3,287語，英単語801語）を除去する。

（ｄ）出現頻度に基づくフィルタ

文書データ全体に対して、ある一定割合以上の文書件数（次章の実験では１０%）に出現する単語を除去する。また、出現文書件数の低い（次章の実験では５件未満）単語を除去する。

（３）共起語認定 [2)]

キーワード選定で残った単語について、前章の方式により、共起語を抽出する。抽出する単語の個数および、出現頻度に基づいて生成するクラスの個数などのパラメータは予めユーザが指定できる。

４．WWWページからの共起語自動抽出実験

前章で述べた方式により、WWWページの実データ約45,000件（平成９年初め頃）を用いて共起語を自動抽出した。表１に単語統計データを、図１に出力サンプルを示す。最終的に共起語情報を出力できた単語は、62,324種類に絞り込むことができた。

本実験より、本方式による共起語抽出においては次の傾向があることが分かった。

（１）文書内容（流行，トピック性）を反映した共起語を抽出できた（「図１で「米軍基地」の共起語として「強制使用」「代理署名」を出力）。

（２）複合語は共起語として有効である。複合語の方が固有名詞性が高く、多義性もないことが多い。一方、出現頻度が小さくなるので、共起頻度の値が相対的に高くなりやすい）。

（３）頻度の高い共起語に適切なものが多い。

５．終わりに

実際の検索システムに本システムを組み込み、検索作業の効率向上を定量的に評価する予定である。

参考文献

1)西川他：シソーラス管理システムにおけるカスタマイズ機能について，情報処理学会第54回全国大会講演論文集 3-171～172，1997.

2)丹羽他：動的な共起解析を用いた対話的文書検索支援，情報処理学会自然言語処理研究会報告，96-NL-115，1996.