語の属性を用いた全文検索の高機能化[*1]

瀬戸口光宏[*2]　稲垣陽一[*3]　中村 隆宏[*4]　相澤 弘[*5]
株式会社シーエーシー[*6]　株式会社小学館[*7]

1 はじめに

本研究では、語の属性を用いて検索を行なうことで、連想的な検索などの全文検索機能の高度化を試みた。この有効性を検証するため、茶筅で用いられている IPA 辞書[*8]の品詞分類を利用し、プロトタイプシステムの開発、ウェブ上の文書データを用いた検証を行なった。

検証の結果、知識発見におけるこの手法の有効性が示された。さらに高度な検索を可能とするために、体系的な意味カテゴリ情報の利用とより効果的な検索インターフェースの提案を行なう。

2 プロトタイプシステムの実装

今回実装したシステムは、コーパス索引作成部とコーパス検索部の 2 つのサブシステムから構成される。

索引作成部は、形態素解析された文書群を入力として、索引データとオンメモリ用のコーパスデータを作成する。形態素解析には茶筅を用いている。

コーパス検索部は、キーワードとカテゴリ属性の組をユーザから受けとり、キーワードと共起する語のうち、指定されたカテゴリ属性に含まれる語のみを、コーパスデータを利用して抽出する。抽出された語はキーワードとの共起度によってランキングされ、検索結果となる。ユーザはまた、検索結果に対して、実際の出現コンテキストを KWIC 形式で確認することができる。

検索を高速化するため、索引作成部で生成されたデータはメモリ上にロードされる。共起検索もコンテキスト抽出もすべてオンメモリで実行され、ディスクアクセスは一切行なわない。

2.1 カテゴリ属性について

システムは、キーワードと共起する語を語の属性によってフィルタし、求めたい属性をもつ語のみを抽出する。このための属性として、今回の実験では、IPA 辞書の品詞分類を利用した。

IPA 辞書における品詞分類では、細分類まで含めると 90 弱の品詞が存在する。このうち、『名詞-固有名詞-人名』（及びその下位区分）、『名詞-固有名詞-組織』、『名詞-固有名詞-地域』（及びその下位区分）を、語の属性として利用した。これにより、人名・組織名・地域名に属する語を選択的に抽出することが可能となる。

2.2 共起度の計算

抽出された共起語のランキング方式として、頻度数／t スコア／M.I. ／ LogLog[*9]の 4 種を実装した。ユーザは検索要求に応じて、いずれかを指定することができる。

またユーザは共起の対象範囲をウィンドウサイズ（キーワードの前後 5〜50 形態素）として指定することができる。検索範囲をキーワードの前後に限定する点は、ファイル全体の特徴量をみる成分比較と大きく異なる。対象となる文書の量が膨大になればなるほど、検索範囲についての配慮が重要となると考える。

3 評価

本件では、対象コーパスとして、ウェブ上の文書データを使用した。拡張子が「html, htm, txt」の文書データを対象として、文書数 40 万強（のべ形態素数 3 億強）を回収し、システムの評価を行なった。

ここでは、「『小泉純一郎』氏と関係の深い人物について調べたい」という検索要求例を元に、その結果を記し考察を行なう。

キーワードを『小泉純一郎』、検索カテゴリを『人名』、ウィンドウサイズを 50 としたところ、表–1 の結果が得られた。

頻度数では、自身の姓が最も上位に現れる。次点の『安倍晋（三）』は、『小泉内閣メールマガジン』の最後に必ず

総編集長：内閣総理大臣　小泉純一郎
編集長：内閣官房副長官　安倍晋三

とあることが原因と考えられる。

M.I. では稀な共起の結果が計算される。昭和 63 年／平成元年の内閣の構成一覧にて現れる『坂野重信』や、

[*1] Advanced Text Search Engine Based on Semantic Attributes of Morpheme
[*2] Mitsuhiro Setoguchi
[*3] Yoichi Inagaki
[*4] Takahiro Nakamura
[*5] Hiroshi Aizawa
[*6] CAC Corporation
[*7] Shogakukan Inc.
[*8] 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座が配布している形態素解析器である茶筅と共に公開されている日本語辞書システム, http://chasen.aist-nara.ac.jp/
[*9] 文献 [1] における”Salience” のこと。BNCWeb の開発にて Sebastian Hoffman が LogLog と命名。

表–1: 出力結果

順位	頻度数	t スコア	M.I.	LogLog
1	小泉（純一郎）(188.0)	小泉（純一郎）(13.7)	森山真弓 (13.3)	安倍晋（三）(66.0)
2	安倍晋（三）(94.0)	安倍晋（三）(9.7)	坂野重信 (12.9)	（森山）眞弓 (44.4)
3	（平沼）赳夫 (22.0)	（平沼）赳夫 (4.7)	カリモフ (12.7)	森山真弓 (44.2)
4	（森山）眞弓 (21.0)	（森山）眞弓 (4.6)	上野公（成）(11.8)	（平沼）赳夫 (43.1)
5	（福田）康夫 (18.0)	（福田）康夫 (4.2)	古川貞二郎 (11.4)	片山虎之助 (40.3)
6	福田康夫 (17.0)	福田康夫 (4.1)	田中真紀子 (11.0)	（片山）虎之助 (39.8)
7	ブッシュ (15.0)	ブッシュ (3.9)	チョクトン (10.9)	カリモフ (38.1)
8	（片山）虎之助 (14.0)	（片山）虎之助 (3.7)	片山虎之助 (10.9)	福田康夫 (38.1)
9	片山虎之助 (13.0)	片山虎之助 (3.6)	森山眞弓 (10.7)	小泉純一郎 (31.6)
10	小泉純一郎 (12.0)	小泉純一郎 (3.5)	塩川正十郎 (10.5)	金正日 (30.9)

小泉内閣メールマガジンにて記述のあったシンガポールの『チョクトン』首相の名が出てくるのが興味深い。

以下にその他の検索例とその結果を記す。

- 要求: 『川崎市』と関係の深い組織は？
  - キーワード: 川崎市
  - 検索カテゴリ: 名詞-固有名詞-組織
  - 検索結果: 1. 川崎 (313.0), 2. 富士通 (170.0), 3. 富士電機 (148.0)[頻度数]
- 要求: 『金魚』とゆかりのある地域を知りたい
  - キーワード: 金魚
  - 検索カテゴリ: 名詞-固有名詞-地域
  - 検索結果: 1. 熊本 (34.0), 2. 日本 (31.0), 3. 大和郡山 (14.0)[頻度数]

今回開発したプロトタイプシステムは、「人名」「組織」「地域」の 3 カテゴリに限定されたものであり、また、対象コーパスの規模も限られたものであるが、検証の結果は本手法の有効性を十分に示唆しているものと考える。

4 拡張提案

より利便性を高めるためプロトタイプシステムの発展として以下の提案を行なう。

4.1 体系的な意味カテゴリ情報の利用

体系的な意味カテゴリを用いて、多様な種別の概念を対象とした検索を可能にする。

例えば『作物』というカテゴリを作成し、『芋』『トマト』『葡萄』などのエントリに『作物』というカテゴリ情報を付与する。

これにより、「ポリフェノールが多く含まれている作物は？」という検索要求に応えることができる。

4.2 例によるカテゴリ指定

ユーザは検索対象カテゴリを明確に意識しているとは限らない。そこで検索対象カテゴリを直接指定せず、例を入力することで検索対象を指定するようなインターフェースが考えられる。

このインターフェイスを用いた場合、「抗酸化作用を促す物質にはポリフェノール以外に何があるか？」という検索要求に対して、キーワードとして『抗酸化作用』、カテゴリ例として『ポリフェノール』を指定することで、要求に沿った検索が可能となる。

5 おわりに

本研究では、語の属性を用いた全文検索の高機能化手法を提案した。IPA 辞書の品詞分類を利用し、プロトタイプシステムの開発を行ない、提案手法の有効性を確認した。また、さらなる拡張について提案を行なった。

今後は、前節で提案した拡張を行なうとともに、特化領域向けの検索エンジンや、テキストチャットから利用するためのインターフェース拡張、アイデアをマイニングする機能の研究開発などを行なっていきたい。

参考文献

[1] A.Kilgariff and D.Tugwell: “WASP-Bench: an MT Lexicographer’s Workstation Supporting State-of-the-art Lexical Disambiguation”, Proc. MT Summit XIII, 2001, pp. 187-190.

[2] T.Nakamura and Y.Tono: “Lexical Profiling Using the Shogakukan Language Toolbox”, ASIALEX2003 Proceedings, The Third Asialex International Congress, August, Meikai Univ., Japan, 2003, pp170-176