2E1-04

連想検索を利用したローカル文書管理システムの構築

Proposal of Local Document Management System based on Associative Search

傘木英将 [1]　山口亨 [1]　高間康史 [1]

Yoshimasa Kasagi　Toru Yamaguchi　Yasufumi Takama

[1]東京都立科学技術大学

Tokyo Metropolitan Institute of Technology

Abstract: As the information that can collect from the web to local database is increasing, we propose the system that can suggest related local documents when new document arrives. We build an association dictionary using web search engines for similarity calculation.

1. はじめに

従来のような、ユーザが必要な時に検索を行うという能動的な検索ではなく、文書を保存するときに検索を行うという受動的な検索を行うことで、ユーザに文書を提示する機会を増やし、今までにない発想や知識を得ることを支援可能な文書管理システムを提案する。

また、現在主流のローカル検索システムでは入力した単語と完全にマッチするファイルしか表示されず、文書間の関連度を計算することに関して不十分である。

連想検索が可能なシステムにおいても、シソーラス拡張を行ったり、ユーザがあらかじめ連想語句の関係などを入力しなければならないなど手間のかかる部分が多かった[1][2]。

本研究では、ウェブ検索エンジンを使用して作成した連想検索辞書を使うことにより、文書間の関連性をより広く計算できるようにする。

2. 文書共起度

単語同士の関連度を表す数値として文書共起度がある。数値が高いほど関連性があるとみなされる。本編では、文書共起度をもとにした単語 W_A と単語 W_B の関連度 $r(W_A, W_B)$ を

$$r(W_A, W_B) = \frac{2 \times fc}{fa + fb} \quad (1)$$

とする[3]。式(1)において fa 、 fb はそれぞれ単語A、Bの検索結果数を表し、 fc は単語 W_A 、単語 W_B を同時に入力したときの検索結果数を表す。本研究では検索エンジンとしてGoogle(http://www.google.co.jp)を使用した。形態素解析ツール茶筌(http://chasen.aist-nara.ac.jp)を用いて文書中から名詞を抽出する。

名詞の選定の方法は固有名詞など特定性の強いもののみに限定し、連想検索として意味をなさない名詞はあらかじめストップワードリストを作成しておく。図１に単語の抽出例を示す。

連絡先：東京都立科学技術大学高間研究室，〒191-0065 東京都日野市旭が丘六丁目６番地，Tel: 042-585-8629, kasagi@krectmt3.tmit.ac.jp

文書の内容	紅葉の名所京都で、約 2000 といわれる寺社が膨大な落ち葉の処理を巡り悩み多い師走を迎えている。昨年４月の廃棄物処理法改正で廃棄物の野外焼却が原則禁止になり…
抽出した名詞	名所, 京都, 寺社, 落ち葉, 師走

図１. 文書と抽出した名詞の例

単語 W_A と文書Dとの関連度 $R(W_A, D)$ は式(2)で表される。

$$R(W_A, D) = \frac{1}{M} \sum_{W_D \in D} r(W_A, W_D) \quad (2)$$

W_D は文書Dから抽出された名詞とし、 M は文書Dから抽出された名詞の数とする。

さまざまなジャンルからなる 10 個のローカル文書に対して、『京都』との関連度を計算すると、図 2 のようになる。ここで横軸は各ローカル文書とそのジャンルである。単語と文書との関連度に関して文書共起度は有効であることがわかる。

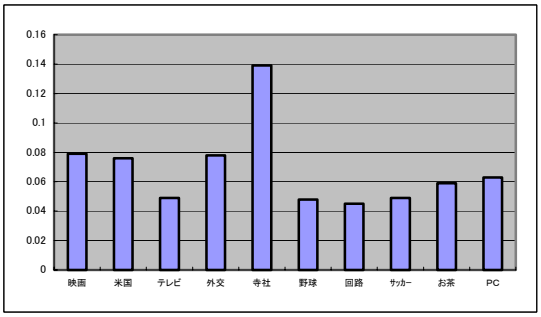

図 2.「京都」と各ローカル文書の関連度

単語と各文書の関連度を数値化し、表１のように辞書ファイルとして保存する。辞書ファイル作成にあたっては、各単語の検索エンジンでの検索結果数をあらかじめデータベースとして保存しておき、それをもとに作成する。

表 1. 辞書ファイルの内容（抜粋）

	文書 01	文書 02	文書 03	文書 04
映画	0. 242	0. 069	0. 088	0. 064
政府	0. 068	0. 12	0. 03	0. 198

3. ローカル文書管理システム

図３はプログラムのインタフェース画面である。左のテキストエリアに文章を貼り付けて保存することで、文書自体が保存される。

保存時に文書と辞書ファイルのデータを比較することによって、文書共起度に基づいた関連度の高い順に、文書が右側ウインドウに表示される。ここで、保存した新規文書とローカルにある各文書の関連度は、新規文書から前述の方法で抽出した名詞について、辞書ファイルを参照して文書共起度を合計して求める。

保存したファイルはこの時点では辞書に登録されていないため、検索対象とはなっていないが、辞書登録を実行することによって、辞書ファイルに情報が追加され、次回から関連度が計算され表示されるようになる。

新しい文書を追加したとき、文書中から新たに単語を登録して、ウェブ上からデータを集めるとなると膨大な時間がかかり非現実的であるため、辞書データの中にすでにある単語の抽出だけを行う。現在の段階で約 350 単語のデータがあり、ニュース記事などで使われる一般的な単語においては十分なレベルと考えた。

4. 実行例

図３はすでにニュース記事、日記、大学の情報など 50 程度の文書がデータベースとしてローカルに保存してある場合に、野球関連のニュースを新たに保存したときの実行例である。この文書との関連度が計算された結果、右側のウインドウにローカル文書が関連度順に一覧となって表示される。

上位にくるのはスポーツ関連のニュース記事などで、それ以外にも以前保存した個人的な野球の日記のようなものも出てくる(図４)。

この様に、ジャンルを考えることなく次々に保存することができるため、過去に保存した様々な文書が関連度をもとに表示される。

5. おわりに

今回の研究ではサンプル文書数が 50 程度であった。もし該当文書がなかった場合に上位にくるのはネット関係の言葉、あるいは広い意味に取れる言葉を多く持っている文書である。これは検索エンジンの検索結果を文書共起度計算のデータとして使っているためであると考えられる。

今後の発展として、文書から名詞を抽出する際に形態素解析ツールを使用せずに文章の流れなどから単語を取り出すことにより、新しい単語や特徴のある単語に対応でき、検索エンジンを使った連想検索をより直感的にできるものと考える。

図３. プログラムの実行結果

図４. プログラムの実行結果

6. 参考文献

[1]平田、村上、西田：連想表現と分身エージェントを用いたコミュニティにおける知識共有支援, 人工知能学会誌, vol. 15 No, 6, pp. 1117-1124, 2000.

[2] 砂山、大澤、谷内田：ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス, 人工知能学会論文誌 Vol. 16 No. 2 pp. 225-233, 2001.

[3]R. Mandala, T. Tokunaga, H. Tanaka, A. Okumura, andK. Satoh： Ad Hoc Retrieval Experiments Using Word Net and Automatically Constructed, NEC, 東京工業大学, TREC7 No48.