1U-3

Webコンテンツマイニングシステム
Web Contents Mining System

本田 哲也† 久保田 光一†
中央大学大学院 理工学研究科 情報工学専攻†

要約: 本研究では, Web ページのテキストデータを形態素解析し, Web ページの内容を特徴付けるような単語を抽出する手法を提案する. また Google で検索した結果の Web ページから特徴語を抽出する. 次に各ページの重要度を測り, 重要度順に Web ページを並べ替える. そして元の Google の検索順と比較し差異があるかを調べる. これらを行うためのシステムを構築し, 実行時間と実験例について報告する.

キーワード: Web マイニング, 形態素解析

1 はじめに

インターネットと WWW の高度成長により, 我々は膨大な情報に触れることができるようになった. それに伴い, Web 上にある個人や企業の情報ソースから顧客のニーズなど企業にとって有益な情報や知識を抽出する Web マイニングの技術が注目されている.

そのなかでも本研究ではテキストデータを扱う Web コンテンツマイニングに焦点を当てる. 奈良先端科学技術大学の松本研究室が公開している, 文章を単語に分割する形態素解析ツール “茶筅” [2] を用いて, Web ページを解析する.

解析方法としては, (1) 指定した Web ページに含まれる単語からその Web ページを特徴付けると考えられるもの (特徴語) を抽出すること,(2) 指定した検索語 (クエリー) に関連する Web ページを検索し, 抽出した特徴語からその Web ページの重要度を測ることの 2 手法について考察する.

以下ではまず, 特徴語の抽出方法について説明し, 次に Web ページの重要度を測る手法について述べる. そして作製した Java クラスと実行時間, 計算機実験について言及する.

2 特徴語の抽出

ページ内のテキストデータを形態素解析し, 出現頻度が高い単語をそのページを特徴付ける一指標であると考え, それらの単語を本研究では特徴語と位置づける. 特徴語となりうる品詞は, 名詞, 動詞, 形容詞の 3 つから任意に指定できるものとし, それによって得られた単語群を特徴語の候補とする.

2.1 単語の重み付け

単語の重要度の評価は，二つの指標を掛け合わせた TF/IDF 法と呼ばれる手法が一般的に用いられている.

- TF（Term Frequency）法 ⋯ 特徴語の出現個数.
- IDF（Inverted Document Frequency）法 ⋯ 特徴語が全ページにどのように分布しているか.

$$TF = W_c \tag{1}$$

$$IDF = 1 + \log(\frac{N}{n}) \tag{2}$$

W_c: 特徴語 W の出現個数 , N : 全ページ数
n : 特徴語 W を含むページ数

ある特徴語が多くのページ中に現れる普遍的な単語だった場合には IDF は小さくなり, 逆に特定のページにしか現れない場合には IDF は大きくなる.

一般的にページの文書が長くなるにつれ, 特徴語が多く現れる. そこで文書の長さによる影響を減らすために, コサイン正規化を行う.

$$W_s = TF \cdot IDF / \sqrt{\sum_{i=0}^{m} {W_{it}}^2} \tag{3}$$

W_s : 特徴語 W のスコア, m : 特徴語の総数
W_{it} : 特徴語 W_i の TF/IDF 値

ここではスコアが大きい特徴語候補を順に, 上位 100 個を特徴語とする.

2.2 名詞句の結合

名詞句は 2 つ以上連続して現れることが多いため, 結合して一つの名詞句とする.

2.3 ノイズの除去

文章中に普遍的に現れる, “こと”, “それ”, “する”, “ある”, などの形態素はページを特徴付けないノイズと判断して無視する.

3 特徴語による Web ページの重要度判定

GoogleWebAPI[3] を用いて, あるクエリーについて検索し, それらの Web ページから特徴語を抽出して, 特徴語を多く含む順に Web ページを並べ替え, もとの検索結果を比較する.

3.1 GoogleWebAPI

GoogleWebAPI とは Google 社が 2002 年に公開した, Google のデータベースを利用できる API であり, Google の持つ検索の仕組みを外部アプリケーションから利用できるようにしたものである.

Web Contents Mining System
†Tetsuya HONDA Koichi KOBOTA, Information and System Engineering Course, Graduate School of Science and Engineering, CHUO University

3.2 Web ページのスコアリング

得られた特徴語により各 Web ページの重要度を測るためにスコアリングをする. Web ページ内の含まれる特徴語のスコアの総計をそのページの文字数で割る.

$$P_w = (\sum_{i=0}^{r} W_{iw})/P_l * 100 \qquad (4)$$

P_w : ページ P の重要度, P_l : ページ P の文字数
W_{iw} : ページ P に現れる特徴語 W_i のスコア
r : ページ P に現れる特徴語の個数

3.3 処理の流れ

以下に大まかな処理の流れを記す.

(1) クエリーを GoogleWebAPI に渡し, URL を取得. (2)HTML ファイルをプレーンテキスト化して茶筅で形態素解析する. (3) 指定した品詞に該当する形態素を特徴語候補とする. (4)TF・IDF 法で各特徴語候補の重みを測り, 上位 100 個を特徴語とする. (5) ページをスコアリングする.

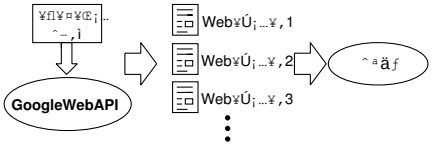

図 1 処理の流れ (前半)

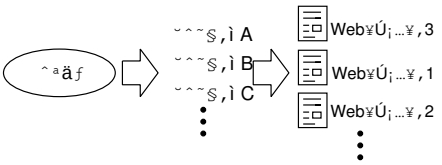

図 2 処理の流れ (後半)

4 特徴語抽出システムの構築

4.1 Java クラス “WCM”

特徴語を抽出する Java クラスを作製した. 主な機能としては, クエリーを渡すと Google の検索結果の Web ページ群から特徴語を抽出し, 各 Web ページの重要度を測定する. また, 直接 URL を指定して特徴語を抽出することができる.

4.2 実行時間

OS が Windows2000 で, CPU が 1.6GHz の計算機での実行時間の例を表 1 である. 実行時間のほとんどが茶筅による形態素解析によるものである.

表 1 実行時間の一例

	総文字数	ページ数	実行時間 (秒)
実験 1	12711	10	17.5
実験 2	166760	50	131.4

5 計算機実験

クエリーを “Java”, Google の検索結果上位５０ページを対象とした特徴語が表 2 である. なお, 特徴語とする品詞は名詞のみである. 表 3 は表 2 と同条件で Web ページの重要度を測り, Google, MSN の結果とを比較したものである. 表中の数字は上位何件目かを, MSN の列の “-”は, 上位 300 ページにその URL が無かったことを示している.

クエリーに関係の深い単語が上位に現れている. また, 既存の検索エンジンとの比較では顕著な違いがあることが見て取れる.

表 2 特徴語 上位 10 個 (クエリー=“Java”)

順位	特徴語	順位	特徴語
1	JavaSolution	11	作製
2	理解	12	動作
3	プログラミング	13	ツール
4	サーブレット	14	サーバ
5	紹介	15	Eclipse
6	利用	16	必要
7	解説	17	簡単
8	クラス	18	ページ
9	開発	19	Web アプリケーション
10	Java アプリケーション	20	プラットホーム

表 3 Google, MSN との比較 (クエリー=“Java”)

提案手法	Google	MSN
1	4	6
2	38	66
3	45	-
4	30	34
5	49	65
6	44	-
7	28	13
8	23	-
9	29	30
10	27	78

6 まとめ

本研究では形態素解析による Web ページの特徴語抽出, Web ページの重要度判定, プログラムの実装と実行例について報告した. 特徴語抽出という手法で Web ページのランキングを行えば, 既存の検索エンジンと大きく違う結果になることが分かった.

最後に, 本研究を進めるにあたり, 適切な助言を下さった同研究室の皆様に, 心よりお礼を申しあげます.

参考文献

[1] G.Chang, M.J.Healey, J.A.M.McHugh, J.T.L.Wang, 武田善行, 梅村恭司, 藤井敦, **Web マイニング**, 共立出版, 東京, 2004.

[2] “形態素解析システム茶筌”, http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[3] “Google Web APIs(beta)”, http://www.google.com/apis/