E-18

Webページの主題推定

Estimation of Web page subject

佐藤 慎哉† 山村 毅‡ 工藤 博章† 松本 哲也† 竹内 義則† 大西 昇†

Shinya Sato Tsuyoshi Yamamura Hiroaki Kudo Tetsuya Matsumoto Yoshinori Takeuchi Noboru Ohnishi

1. はじめに

近年、日々増加する膨大な情報源の中から必要な情報をすばやく見つけることが困難になってきている。そのため多くの検索エンジンには、必要な情報をすばやく見つけるために多くの工夫がされている。それらは、キーワードを含む文を提示したり、2次キーワードを用いて検索結果を絞り込んだりするものである。確かに、キーワード情報は情報検索をする上でユーザに大変有益な情報を与える。しかし、低品質な情報があふれているwebページの検索ではあまり参考にならないことも多い。あるキーワードが含まれているからといって必ずしもそれに関連した内容のページであるとは限らない。そこで、webページに書かれている内容をキーワードに関連のある情報とともに提示できれば、より効率のよい検索ができると考えられる。本稿では、webページの内容を解析してページの主題を推定する手法について述べる。

2. 概要

webページは、複数の話題について書かれていることが多いため、本手法ではページを表示上のまとまりに分割して[1]、それぞれのまとまりに対して個別に内容の解析を行う。ページの分割にはHTMLタグを利用する。各まとまりに見出しが付いているものもあるが、文字サイズ拡大など別の目的のために使われていることがあるなど、必ずしもそれがそのまとまりを正しく表しているわけではないので、見出しを直接抜き出して主題とするのは不適切な場合がある。そこで、単語の出現頻度と助詞などの表層情報を用いて重要個所を求め[2,5]、それを見出しのように表示上強調されている個所と比べ、内容的にふさわしい方をページ中の各まとまりの主題とする。最後にそれぞれのまとまりの主題を統合してページ全体の主題とする。

3. 提案手法

3．1 ページの分割

webページは大きく分けるとテキスト、見出し、図、表、リンクの5種類を組み合わせて構成されていると考えることができる。本手法では<table>、<map>等のHTMLタグを用いてページをこの5種類の要素に分割する。この時、見出し要素は直後の要素に付けられたものと考えられるので直後の要素と統合して1つの要素と考え、その要素の主題を推定するのに利用する。

3．2 テキスト要素の主題推定

ここでは、テキスト要素の主題の推定方法について述べる。まず始めにテキスト中に現れる名詞の出現頻度を求める。次に各名詞の上位概念の出現頻度を各名詞の出現頻度と単語間の類似度[3]を考慮して求める。そして、各名詞の上位概念の出現頻度と助詞などの表層情報をもとにテキスト要素中の名詞句の評価値[2,5]を求め、最も評価値の高い名詞句を抜き出す。最後に、テキスト要素に見出しが付いていればその見出しと抽出された名詞句との類似度を求め、類似していたら見出しを、類似していなかったら抽出した名詞句をテキスト要素の主題とする。

3．2．1 テキスト要素の名詞の出現頻度

形態素解析器「Chasen」[4]を用いてテキストを形態素解析して、テキストに含まれる名詞のうち接頭詞や接尾詞のようにそれ自体では意味をなさない名詞と「人」、「もの」のように抽象度が高い名詞を除く全ての名詞の出現頻度を求める。

3．2．2 テキスト要素の名詞の上概念の出現頻度

EDRのシソーラス辞書(tree構造)を用いて3.2.1で求めた全ての名詞(葉)からのノードの距離が2以下の上位概念を求める。そして、次式を用いて、求まった全ての上位概念jの出現頻度c_jを計算する。

$$w_{ij} = \frac{d_j}{d_i + d_j} \tag{1}$$

$$c_j = \sum_{i=1}^{N} w_{ij} n_i \tag{2}$$

ここで、d_i、d_jは、名詞iの*root*からの距離、上位概念jの*root*からの距離である。また、Nは、テキスト中の名詞の種類数、n_iは名詞iの出現頻度である。なお、(1)で各名詞

†名古屋大学大学院工学研究科

‡愛知県立大学情報科学部

の上位概念となっていない上位概念とのw_{ij}の値は0とする。また、見出しがついている場合は見出しに含まれる名詞(見出し語)の上位概念と等しい上位概念は内容を推定する上で重要な概念であると考えられる。そこで、(1)、(2)式から求まった上位概念jの出現頻度c_jに上位概念jの見出し語からのノードの距離に応じた値(現在は 距離が1の概念：3、距離が2の概念：2、その他：1)をかける。

3．2．3　名詞句の評価値

ここでいう名詞句とは、「名古屋大学工学部」のように名詞が連続しているもの、「名古屋大学の学生」のように名詞(の連続)が助詞の「の」でつながったもの、「忙しい学生」のように名詞(の連続)に形容詞がついているもののことを指す。名詞句の評価値は(3)式によって求める。

$$NP_i = W_i \sum_{j=1}^{M} \omega_j mc_j \qquad (3)$$

W_iは名詞句iに付いている助詞の種類による重み、ω_jは名詞句中での名詞の位置に対しての重み、mc_jは名詞句に含まれる名詞jの上位概念のうち最も頻度の高かった上位概念の出現頻度、Mは名詞句に含まれる名詞数を表す。

W_iとしては、文中における助詞の働きを考慮して、例えば次の順序[3]で名詞句を重み付けする。

は/には ＞ が/も/だ/なら/こそ
＞ を/に/、/。 ＞ へ/で/から/より　(4)

ω_jは、現在は全て1にしているが、接頭詞などでは小さく、名詞句の骨格となる最後の名詞（主辞）では大きくするといった使い方がある。

3．2．4　見出しとの類似度

3.2.3で求めた評価値の最も高い名詞句を抜き出し、次にこの名詞句と見出しとの類似度を求める。これは、内容に関連した見出しが付いているかどうかを調べるためである。もし、類似度が高ければ見出しは内容に関連があると考え見出しをテキスト要素の主題とし、類似度が低ければ見出しは内容に関連がないとして名詞句を主題とする。類似度の計算方法は以下のようになる。

$$sim1 = \frac{1}{M}\left(\sum_{i=1}^{M} \max_{1\le j\le N}(\omega_{ij})\right) \qquad (5)$$

$$sim2 = \frac{1}{N}\left(\sum_{j=1}^{N} \max_{1\le i\le M}(\omega_{ij})\right) \qquad (6)$$

ここで、Mは見出しに含まれる名詞数、Nは名詞句に含まれる名詞数、ω_{ij}は見出しに含まれる名詞n_iと名詞句に含まれる名詞n_jの式(1)により求まる類似度である。(5)、(6)式のうち大きいほうの値が設定した閾値(現在は0.7)以上なら見出しを主題とし、それ以外は抽出された名詞句を主題とする。

4.　実験結果

提案手法を用いて、図1のテキスト要素[6]の主題を推定した。まずテキスト中からは最も評価値が高い名詞句として「隠れ場所」が抽出される。このテキスト要素には「生息場所」という見出しが付いているので「隠れ場所」と「生息場所」の類似度を調べると*sim1*、*sim2*共に0.86となるためこの見出しは内容に関連があると考えられるのでテキスト要素の主題は「隠れ場所」ではなく「生息場所」となる。このように、提案手法ではwebページの比較的品質の低い文章からでも精度よく主題の推定ができる。

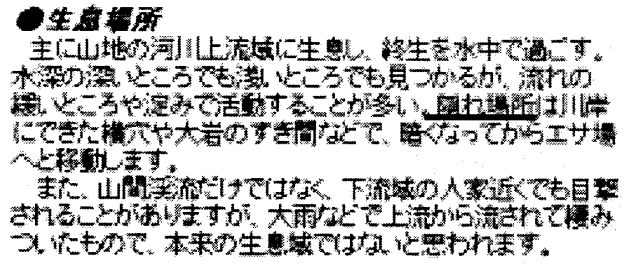
●生息場所
主に山地の河川上流域に生息し、終生を水中で過ごす。水深の深いところでも浅いところでも見つかるが、流れの緩いところや淀みで活動することが多い。隠れ場所は川岸にできた横穴や大岩のすき間などで、暗くなってからエサ場へと移動します。
また、山間渓流だけではなく、下流域の人家近くでも目撃されることがありますが、大雨などで上流から流されて棲みついたもので、本来の生息域ではないと思われます。

図1．実験で用いたwebテキスト

5.　おわりに

本稿では、webページの主題を推定する方法の概要とテキスト要素から主題を推定する方法について述べた。現在は、テキスト要素の主題推定に取り組んでいるが、今後はその他の要素から主題を推定する方法や各要素の主題を統合する方法についても検討していきたい。

6.　参考文献

1) 山田洋志,福島俊一,松田勝志,webページからのタイプ別情報抽出・分類方式,情報処理学会研究報告,Vol.2000 No.29(FI-57 NL-136) pp.143-150
2) 奥村学,難波英嗣,テキスト自動要約に関する研究動向,自然言語処理 july Vol.6 No.6 pp.1-26 1999
3) 長尾真,岩波講座　ソフトウエア科学15　自然言語処理,岩波書店
4) 茶筌:http://chasen.aistnara.ac.jp/index.html.ja
5) K.Zechner, Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences,Proc.16th International Conference on Conputational Linguistics vol.2 pp.986-989 1996
6) http://www.bob-24.com/oosan/nazo.html