日本語Webページを対象としたタイトルタグの傾向と分析

秋岡明香* 下山剛司† 村岡洋一‡
電気通信大学/早稲田大学 早稲田大学 早稲田大学

1 はじめに

ブログの普及やネットワークサービスの大衆化に伴い，HTML文書が爆発的に増えている．例えばNetCraft社のWeb Server Surveyによると，2007年12月時点で世界のWebサーバ数は1.5億台を超えるとされている[1]．Lawrenceらの調査によると，1999年2月現在に発信されていたWebページは約8億ページであり[2]，NetCraft社の調査では当時のWebサーバ数が約430万台であることから，単純な比例計算によると，現在では約280億ページが公開されている計算になる．このような背景を踏まえ，HTML文書を知識データベースとして利用する研究も盛んに行なわれているが，膨大なデータをいかに効率良く処理し必要な情報のみを取り出すかは，重要な研究課題である．

一方でHTMLにはタイトルタグが定義されており，そのページで最も重要なキーワードや，そのページの要約を含むことが期待されている．したがって，タイトルタグに含まれるキーワードの分析を行ない傾向を正しく把握することは，昨今のWebページの傾向を把握するために役立つだけでなく，Webコンテンツから必要な情報のみを取り出す知識フィルタを作成する上でも多いに有用である．そこで本稿では，特に日本語で記述された1億URLを超えるWebページを対象に，タイトルタグの傾向と分析を行なった結果を報告する．

2 タイトルタグの全体的な傾向調査

本稿では，2007年11月に収集した，日本語で記述されたHTML文書110,925,216件のタイトルタグに対して，MeCab[3]とLingua::JA::Summarize[4]による形態素解析を適用し，抽出したキーワード15,649,393（重複を含めて349,827,579）語の全体的な傾向を示す．なお，Lingua::JA::Summarizeでの抽出単語は最大5とした．

*Sayaka Akioka, The University of Electro-Communications/Waseda University
†Takeshi Shimoyama, Waseda University
‡Yoichi Muraoka, Waseda University

図1は，タイトルタグに含まれるキーワード数の集計結果である．タイトルタグに含まれるキーワード数で最も一般的なのは5語であり，このようなページは全体の約27%を占め，次いで多かったのは2語の約20%であった．また，タイトルタグから有効なキーワードを抽出できなかったURLが約1%存在した．

図2に，キーワードの出現回数の分布を示す．また，表1は，出現頻度が上位10位までのキーワードの一覧である．表1の上位8語は出現頻度が100万回を超える．しかし，このように頻出するキーワードは極めて少数であり，抽出したキーワードのうち93%は出現頻度が10回未満である．またこれらの頻出キーワードには一般的なキーワードが多く，コンテキストマイニング等の目的において重要な単語であるとは言い難い．さらに，約6%のキーワードは出現頻度が10回以上100回未満であり，全体として99%の語は出現頻度が100回未満である．

なお本稿で抽出したキーワードは厳密な意味での単語に区切られたものではない．例えば，「連続放火」，「連続殺人犯」，「連続掲載」といったフレーズは，各々独立のキーワードとして扱われており，このような単語の切り出し方が出現頻度100回未満のキーワードを増加させた一因であると考えられる．しかしHTMLを分類・分析する上で，これらの語を関連づけることはあっても，全く同種の文書として扱われるケースは稀であると推測され，ここでは大きな問題はないと考える．

3 出現キーワードに関する考察

本章では，第2章で切り出したキーワードのうちで出現頻度が100回以上のキーワード213,290件について，いくつかの事件等と関連を考察する．

本稿で用いたデータは2007年11月現在での収集データであるが，この時期には防衛省事務次官と山田洋行の癒着疑惑が発覚した．そこでこの事件に関連するキーワードに注目すると，守屋（1022回），防衛省（723回），山田洋行（439回），次官問題（432回），守屋氏（248

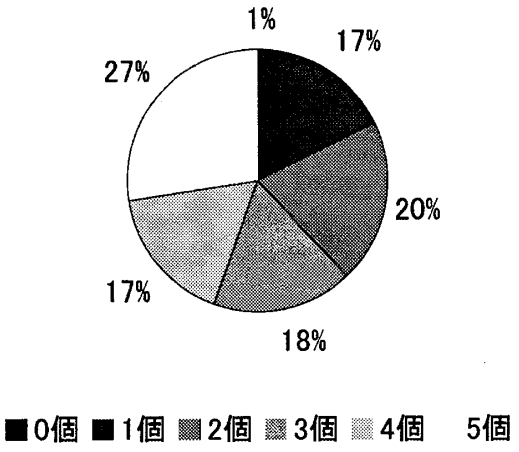

図 1: タイトルタグに含まれるキーワード数の傾向

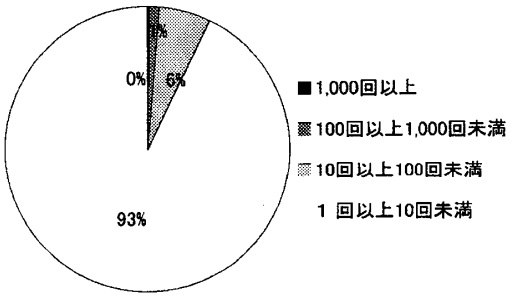

図 2: キーワードの出現回数の傾向

回)，東京地検 (108 回) などが含まれていた．

一方で約半年前の 2007 年 6 月には，ミートーホープ社の食肉偽装事件が起こっているが，2007 年 11 月時点の収集データにもこの事件に関連すると思われる偽装 (879 回)，食肉 (767 回)，ミンチ (250 回)，ミートホープ (156 回) などが含まれていた．

これらの他に，2007 年流行語大賞である「ハニカミ王子」は 203 回，2006 年流行語大賞である「イナバウアー」は 337 回含まれていた．

まとめると，社会現象となる流行語や大きな事件に関連する単語は，100 回以上の出現頻度で概ね捉えることができると予想されるが，このような語は本稿で抽出したキーワードの 1%程度であり，極めて少数部分であると言える．また，今回得た知見を知識フィルタリング等に応用するためには，この 1%程度の文書を如何に効率良く切り分けるかが大きな課題になると考えられる．さらに，今回は 2007 年 11 月のみの収集データを用いたが，時間経過とキーワードの出現頻度の変化の動向にも着目するべきであり，さらに大規模なデータの解析が必要であると思われる．

表 1: 出現頻度が上位のキーワード 10 件

順位	キーワード	出現回数
1	修正	7,169,906
2	登録データ	7,100,549
3	削除	6,006,136
4	パスワード認証画面	2,411,132
5	amp	2,253,723
6	管理者	1,831,204
7	blog	1,328,197
8	通知画面	1,247,247
9	search	985,481
10	日記	979,606

4 まとめと今後の課題

本稿では，日本語で記述された 1 億 URL を超える HTML 文書を対象に，タイトルタグに出現するキーワードについての傾向調査と分析を行なった．今後は，出現キーワードの時系列分析等のさらに詳細な調査を行ない，知識フィルタ開発等へ応用する予定である．

謝辞

本研究は文部科学省リーディングプロジェクト e-Society 「基盤ソフトウェアの総合開発」プロジェクトのサブプロジェクト「インターネット上の知識集約を可能にするプラットフォーム構築技術」の支援により実施された．

参考文献

[1] Netcraft: December 2007 Web Server Survey, http://news.netcraft.com/archives/2007/12/29/december_2007_web_server_survey.html.

[2] S. Lawrence and C. K. Giles, "Accessibility of Information on the Web", Science, Vol. 400, pp. 107 – 109, 1999.

[3] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

[4] Lingua::JA::Summarize, http://search.cpan.org/~kazuho/Lingua-JA-Summarize-0.07/.