webページの有用性を求めるためのキーワード重み付け手法の提案

植草大輔† 安藤大幸† 藤本敬介† 中山泰一†
†電気通信大学情報工学科 †

1 はじめに

近年，多くの情報が掲示板やブログといったweb媒体を用いて広がっている．特にブログ媒体では，現在注目されている事を知るという使い方や，その事柄を他人がどう感じているかといった感想を見たいときに利用される．なお，これらの情報を知りたいときにGoogleなどの通常の検索エンジンではなく，ブログのみをwebから検索できるブログ検索という物がある．しかしブログ検索では，関連性のないタグや別の記事のキーワードが検索され，目的とする情報がなく有用でない事がしばしばある．このため，情報を集めるのに時間や手間が多くかかってしまう事が考えられる．そこで，本研究ではブログを対象としたwebページの有用性を推定するためのキーワード重み付け手法の提案を行う．

2 関連研究及びサービス

webページを評価するシステムにGoogle PageRank [1]がある．これは，webの膨大なリンク構造を用いてwebページの重要性を求めるサービスである．良質なwebページは多くのwebページからリンクされるという基本概念より重要度を0〜10の11段階の評価点で判定するシステム．

また，webページの信頼性を自動評価する研究に福島らの[2]がある．これは，信頼性の評価に影響すると思われる要因をいくつか用意し，ひとつづつ成立，非成立を判断し，webページの信頼性を計算するものである．

3 設計

本システムでは一般に使われる検索エンジンで用いるキーワードと検索結果を用いてwebページの有用性を評価する．このキーワード重み付け手法を使う事によりキーワードとURLで対象のwebページに目的とする情報があるかどうかをを数値化し調べる事ができる．

なお，webページの有用性は以下の二つの要素より定義する．

- ブログの記事の内容がユーザの求めるもの
- 記事内の目的とする話題についての情報量の多さ

この二点を両方とも満たす時，有用性の評価値が高くなる．

3.1 概要

評価は以下の様な手順で行う．キーワードを入力し，検索結果を取得する．この時にスペースで区切ったそれぞれのキーワードに重みを設定する．次に，入力されたURLからHTMLを取得し，ページの解析を行う．検索ヒット数や，形態素数，それぞれのキーワードの数からwebページの有用性の評価を行う．

システムの構築にはTechnoratiのブログ検索API [3]，形態素解析にYahoo! JAPANの日本語形態素解析サービス[4]を用いた．

3.2 評価方法

評価には単語の一般性を計算するTF-IDF法とユーザの入力したいくつかの検索キーワード$t_1, \ldots t_i$にそれぞれ重みをつけ，以下のように算出する．

$\mathrm{tf}(d, t_i)$はキーワードt_iがあるブログd中に現れる頻度を$M(d)$で割った値である．$M(d)$は文書の形態素数である．ここで，今回の重み付け手法の提案としてキーワードに重みCを設定する．なお，重みはユーザが各キーワード毎に手動で設定する．またページ内のt_iの数を$\mathrm{count}(t_i)$とする．

$\mathrm{idf}(t_i)$は，ブログ全体Aにおいてキーワードt_iが現れる文書の数$X(t_i)$に基づく値である．式は以下のように定義する．

$$\mathrm{tf}(d, t_i) = C \times \mathrm{count}(t_i)/M(d)$$
$$\mathrm{idf}(t_i) = \log(A/X(t_i)) + 1$$

なお，TF-IDF法を利用する際，対象となる文書の総数が必要であるが，Technorati JAPAN [3]のState

---

Proposal of key word weighting methods to improve utilities of Web page

Daisuke Uekusa† Hiroyuki Andou † Keisuke Fujimoto† and Yasuichi Nakayama†
†Department of Computer Science, The University of Electro-Communications†

of the Live Web というレポートによると全国で 7000 万ブログがあり，日本語のブログは全体の 37 %に及ぶ．これを考慮に入れ，本研究では対象とする日本語のブログページの総数を 1 億ページと仮定した．

検索キーワードが多いほど目的のページに近づくはずであると考えの元，ページの有用性 util(d) を以下の式にて定義した．

$$\mathrm{util}(d) = \sum_{k=1}^{i} (\mathrm{tf}(d, t_i) \times \mathrm{idf}(t_i))$$

この util(d) の値が高ければ高いほど，web ページの有用性が高いものとなる．

4 実験結果

検索キーワード t_1, t_2, t_3 を数種類用意し，Technorati JAPAN [3] のブログ検索より，web ページの URL を無作為に選び，本手法を用いて実験した．また，対象の web ページが有用であるかどうかは直観で判断した．

実験結果は以下のようになった．

検索語 1 パナソニック，社名，変更

キーワードの重みをそれぞれ
パナソニック=7，社名=2，変更=5 と設定

	キーワード数	形態素数	評価値
ページ 1	8, 6, 7	631	14.86
ページ 2	2, 1, 1	556	14.10
ページ 3	1, 1, 1	557	14.03
ページ 4	2, 1, 1	547	14.10
ページ 5	2, 6, 4	1689	14.11

検索語 2 iPod，touch，感想

キーワードの重みをそれぞれ
ipod=4，touch=10，感想=6 と設定

	キーワード数	形態素数	評価値
ページ A	10, 10, 4	4147	13.22
ページ B	1, 2, 1	631	13.22
ページ C	1, 2, 1	2723	13.05
ページ D	3, 7, 6	689	13.88
ページ E	1, 1, 8	2576	13.11

上記の表はそれぞれのキーワードの数，形態素数から対象となった web ページを評価した値をまとめたものである．

検索語 1 の結果では評価値が 14.10 程度あれば対象ページに目的とする "パナソニックの社名変更についての話題" があり，有用性があると判断できた．なお，ページ 1 での評価値が 14.86 と他のページと比べると高い数値が出ているが，このページではブログ外の web ページのパナソニックの社名変更についてのニュースや意見の一部抜粋でこのように高い値になったと思われる．

検索語 2 では，"iPod touch についての感想" についての記述がある web ページを目的のページとした．この結果，13.22 程度よりも高い評価値の対象ページにおいて有用であると直観で判断できた．しかし，ページ B のみアップル社の別製品についての感想が述べられており，有用とは言えないページであった．これは，検索語が適当でないものと思われる．ページ C やページ E のような評価値の低いページでは目的とする情報は見当たらなかった．また，検索語 1 のページ 1 のようにページ D が他のページに比べて，かなり高い数値が出ているがページ D においては，ブログの記事だけではなく，コメント欄で目的についてのさかんな意見交換が行われている事が反され，上記のような結果になったと思われる．

5 まとめ

本研究では，web ページの有用性をキーワード重み付け手法を用いる事で判定する実験を行った．実際に重み付け手法を適用したところ，検索語が適当でない場合を除き，この重み付け手法を適用する事により，有用でない web ページを判断できた．

なお，今後の課題として，本実験では，web ページの評価が他の web ページと比べて有用であるか，そうでないかを判断している．これではページを見ただけでは有用性が判断できないので，有用性の値に一定の基準値を設定し，判定することが必要であると考えられる．

参考文献

[1] Google PageRank
http://www.google.co.jp/

[2] 福島 隆寛, 内海 彰：Web ページの信頼性の自動推定, 知能と情報（日本知能情報ファジィ会誌）Vol19, No.3, pp.239-249 (2007)

[3] Technorati JAPAN
http://www.technorati.jp/

[4] Yahoo!デベロッパーネットワーク
http://developer.yahoo.co.jp/