社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－ 76　(23)
2004－ＮＬ－163　(23)
2004／9／17

順位付け文書からの影響因子マイニング

沢井 康孝, 峠 泰成, 山本 和英
長岡技術科学大学 電気系
E-mail:{sawai,touge,ykaz}@nlp.nagaokaut.ac.jp

アクセスランキングなどの順位が付与された文書群を情報にして高順位となる文書の影響因子のマイニングを試みた。まず、文書の順位情報をアクセス数などに相当する興味と直接関係のある値に変換した上で、各単語ごとに高順位への影響度を計算し、この情報から文書全体の興味の強さを推定する。以上のモデルを用いて出力の順位と実際の順位を比較した結果、出力上位 30 記事中 10.6 記事が実際に上位記事であり、順位相関係数の平均は 0.20 であった。

Effect elements mining from document ranking

Yasutaka SAWAI, Yasunari TOUGE, Kazuhide YAMAMOTO
Department of Electrical Engineering, Nagaoka University of Technology
E-mail:{sawai,touge,ykaz}@nlp.nagaokaut.ac.jp

This paper describes a method to analyze effect elements by a text ranking such as Web access ranking. We first convert ranking information into the interest value, which corresponds to Web access frequencies. We then estimate the effect of each word. We finally calculate the interest score by accumulating the effect of each word in the target text. The experiment demonstrates that 10.6 texts out of top 30 outputs are ranked in real top 30, and average correlation coefficient score attains 0.20.

1　はじめに

現在、流行や話題の発見といった人の興味に関する研究が多く行われている。Web の普及により誰でも大量の情報を容易に取得できるようになった。このことより、大量の情報の中から、必要な情報のみを取得するマイニング技術が、情報を有効に活用するための一つの重要な方法となっている。

テキストから人の興味や関心といった情報が多く得られるものとして、Web 掲示板や Blog などの感性情報を直接示す文書や、Web ページのアクセスランキングやインターネット投票などの順位付き文書が考えられる。本研究では、多くの人の興味を反映している情報源として、Web 上で公開されているニュースランキングを選択した。この情報は、ニュース記事それぞれに対してのアクセス数をカウントし、順位付き文書として公開されている。順位情報と興味には深い関係性があり、この情報に着目することによって、多くの人の興味を捉えることができるようになる。さらに順位付き文書から順位に影響する要素を得ることができれば、興味に対する分析を行うことや、多くの人が興味を持つ要素と持たない要素を判別する手がかりとすることができる。

多くの人が興味を持つ文書と持たない文書を判断するとき、話題性で興味が持たれているもの以外の文書については出現している単語のみで判断していると考えられる。時系列データを使用せず解析を行うことで、興味が集まっている文書を見つけるだけではなく、これから興味が集まりそうな文書も発見することが期待できる。

本研究では、多くの人が興味を持つ文書を判断するために、順位付き文書から判断に必要な影響因子を推定する。その影響因子を利用して文書の興味の大小の予測することで、人の持つ興味をマイニングするための手法を提案する。

以下、第 3 章では、影響因子に付いて述べ、第 4

章では提案した手法を述べ、第 5 章以降で本研究の実験結果と考察を述べる。

2　関連研究

現在、話題や流行に関する研究が多く行われている [2,3,6]。 時系列分析によって現在注目をあびているキーワードを抽出する研究、社会的要因から流行語の分析を行っている研究等 [3]、様々なアプローチで行われている。特に最近では Web を対象としたマイニングが広がり、日記のような存在である Blog や掲示板は人の意見が直接反映されている場を利用して、話題を捉える研究がある [2]。いずれも興味と関係のある研究であるが順位情報は使用していない。本研究では時系列を使用しないことで興味予測の可能性についても考えている。

また、ランキングを使用した研究 [4] もあるが、ランキングの順位情報は使用していない。本研究では順位情報に着目し、順位情報から興味へ直接関係のある値を推定して、興味の大小を予測した。

3　影響因子について

3.1　順位と興味

売り上げやアクセス数、投票数に付与されている順位は、閲覧数やダウンロード数等の数値について上下関係で示したものである。この順位は以下に示すように様々な要素によって構成されている。

順位付きのデータの例

- ニュースのアクセス数ランキング (アクセス数)
- Blog のページビューランキング (アクセス数)
- 選択性のアンケート (投票数)
- ソフトウェアダウンロードランキング (ダウンロード数)
- 検索キーワードランキング (使用回数)

() 内は順位を構成する要素

いずれの構成要素も、直接的に人の興味に関わっていると考える。そのため順位情報も人の興味に大きく関わっていることが推定できる。そこで我々は順位情報から、興味の強さを推定することで、興味に直接関わる値として使用する。

アクセス数などの順位を構成する要素と順位の間にある関係は、経験則であるべき乗の法則に従うことが知られている。両対数のグラフにおいて直線で表されるグラフを、Zipf の法則 (図 1) と呼び、単語の頻度分布、アンケートの得票数などにも現れる経験則である。

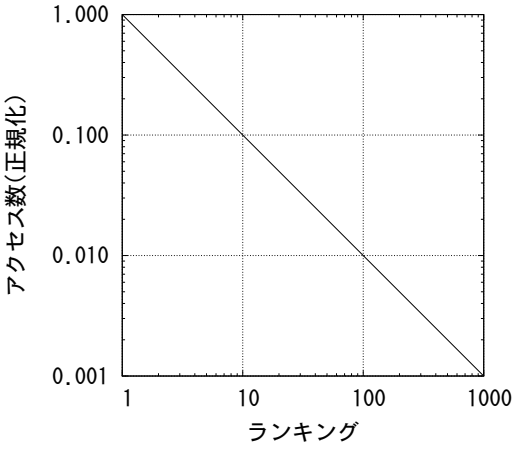

図　1: 順位とアクセス数の関係 (Zipf の法則)

3.2　影響因子の種類

順位を上げる影響因子として、ニュース記事の文書に注目して、記事の内容が興味に影響を与える要素を考えた場合、次のような要素が考えられる。

1. 新出要素

新出である情報は人の興味に強く作用し、順位を上げやすい。

- 今までになかった記事
- 新製品・新商品等に関する記事

2. 珍しい要素

珍しいという事自体が人の興味に強く作用し、順位を上げやすい。

- 珍しい記事(1)
- 記事になり難いようなものが記事として出現

3. 話題による要素

一度人の興味の度合が高くなった文書は、その関連性のある文章も興味も引きやすやすくなる。

- 過去に高い順位となった記事の続報
- 過去に高い順位となった関連記事

4. 絶対的な要素

ある事象に関する事がいかなる時でも興味を引く。

- 時期的な記事 (花粉情報など)
- 特定の固有名詞に対する記事 (有名人、企業名)

(1) 例えば以下のような記事が挙げられる。
・４３円盗んだ疑いで緊急逮捕
・透きとおったタコ、深海から迷い出る？

3.3 単語単位での影響

影響因子により高い興味を持つことで、文書は大勢の人に閲覧される。その結果、閲覧された回数に大きな差が出る。この差を上下関係のみで捉えた情報が順位情報である。

文書全体における影響因子の単位を単語として考えると、単語自体の影響の度合は、注目している単語が順位付き文書内での出現回数に関係する。また順位付き文書に現れたときの順位にも関係してくると考えられる。

本稿では影響の大きさを単語で捉えて、単語自体の影響の強さを過去の文書から決定するモデルを考える。これによって対象の文書に含まれる単語によりその文書が興味を持たれるか否かを判定することが可能となる。

4 提案手法

4.1 全体の流れ

単語単位で文書の興味について判定するため、対象記事より過去の文書から単語自体の影響の強さを求め、これを利用して文書に興味に対するスコアを与える。

具体的には、文書に含まれている異なり単語の全てについて、順位付き文書及び順位が付いてない文書の両方から、順位が付いている文書に出現する頻度、出現したときの順位、順位が付いてない文書に出現する頻度を求め、順位情報を興味の強さに変換した値として利用する。

4.2 入力と出力

本研究ではニュースランキングを利用した。使用したニュースランキングは Web 上で公開されている、朝日新聞社の「アクセス Top30」(1) である。

朝日新聞社のニュースランキングは、0 時から 24 時までのアクセス数によって記事に順位を付け、30 位までのランキングとその記事が発表されている。そのため、ランキング内にはその時間内に発表された記事だけではなく、それ以前の記事も出現する。よって、順位付き記事には同じ記事が含まれることがあるが、このことも興味へ関係してくるため、そのまま扱った。

実際には、ランキング内の順位付きの記事とそれ以外の順位情報がない記事を学習データとして利用し、この両方の記事から各単語が興味へ与える影響の期待値を推定した。

入力データは一日に出現する記事全てを使用した。記事に現れる異なり単語について、学習データから推定した興味へ与える影響の値を用いて、記事全体の興味を推定してスコアを付加した。入力データに含まれる全ての記事に対してスコアを付加した後、その値によって降順で記事を並び替えて出力した。このときの出力順を興味に対して影響が大きい順として扱い、実際にアクセス数で順位付けられた文書と比較を行っている。

4.3 順位情報から興味強度への変換

順位付き文書の順位情報は、そのままでは順位を構成する要素の上下関係を示した値でしかない。順位情報をアクセス数のような興味を直接反映している値へ近づけるため、順位とアクセス数の関係から、興味の強さという値へ変換する必要がある。

順位とアクセス数の関係は、経験則によりべき乗の法則に従う。両対数のグラフ上で直線となる特性から順位情報に対し、次に示すような操作を行った。

$$\log(Hit) = C_1 - C_2 \cdot \log(r) \qquad (1)$$

r:順位　Hit:アクセス数
C_1, C_2:定数

式 (1) より定数を $C_1 = C_2 = 1$ として、興味強度は式 (2) を利用して推定する。

$$P_R(r) = 10^{-log(r)} = \frac{1}{r} \qquad (2)$$

$P_R(r)$:順位 r の興味強度

順位情報を式 (2) により興味の強さの値とした。

4.4 各単語の影響度の算出

入力データの文書に出現した各単語には、順位付き文書より得た順位情報を順位への影響の強さとして扱うため、以下のようなを操作を行った。単語には順位付き文書内の出現回数に応じて順位情報を複数持っている。用いた単語は名詞、動詞、形容詞を異なりで扱った。形態素解析には「茶筌」(2) を用いた。

1. 順位情報から影響度を算出

順位情報付き文書に出現した単語は全てに興味強度に基づいた値を付与する。

$$E_1(w) = \frac{\sum_{\{r|r \subseteq R_w\}} P_R(r)}{C_R(w)} \qquad (3)$$

w:単語　$E_1(w)$:単語 w の影響度
$C_R(w)$:順位付き文書中の w の頻度
R_w:対象の単語 w に付いている順位情報の集合

2. 順位が付いてない文書への考慮

単語 w の興味強度 $E_1(w)$ は、順位情報に基づいているため、順位付き文書外で出現した場合を考慮していない。よって式 (3) の値は、単語がランキング内に出現した場合の影響の強さである。そのため出現する文書全てに適

応する値とするには、ランキング外で出現した場合を考えて値の修正が必要である。単語が順位付き文書内での頻度と、順位が付いてない文書での頻度を求め、単語の順位付き文書内出現確率を求めた。順位付き文書内出現確率の値により、影響度 $E_1(w)$ の値を $E_2(w)$ として補正した。

$$E_2(w) = E_1(w) \cdot \frac{C_R(w)}{C(w)} \tag{4}$$

$C(w)$:順位付き文書と順位が付いてない文書中の w の頻度

3. 出現分布に対する考慮

出現分布が広すぎる単語は、文書を比較するための参考とならない。従って、スコアを低くするように値を考慮した。また、その中でも特に広い分布を持つものは 0 とした。

$$E_3(w) = -E_2(w) \cdot log\left(\frac{C(w)}{N}\right) \tag{5}$$

但し $\frac{C(w)}{N} > 0.5$ の時 $E_3(w) = 0$

N:順位付き文書と順位が付いてない文書数

入力データの記事に出現した全単語に対し順位への影響値 $E_3(w)$ を付加し、これを単語が順位に与える影響度として扱った。

4.5 各文書の興味スコア

入力文書 D に含まれる異なり単語の影響度 $E_3(w)$ から、入力文書全体の興味スコア $S(D)$ を推定する。文書の興味を単語から推定する場合、入力文書 D に含まれる異なり単語数は文書の長さにより一定ではない事から、文書の長さの考慮が必要である。

1. 文書の興味スコア

文書に含まれる単語それぞれが、順位に対して影響を与えるとしたとき、文書が持つ順位に対する影響値は、各単語の影響度より次のように決定した。

$$S_E(D) = \sum_{\{w|w \subseteq W_D\}} E_3(w) \tag{6}$$

W_D:入力文書 D に含まれる異なり単語集合
(但し $E_3(w) = 0$ の単語は除く)

2. 文書の長さへの考慮

式 (6) では文書が長い程、影響が大きいということになってしまう。そのため、文書の長さを考慮するため式 (7) により変換した。このときの平均単語影響度は全ての単語影響度の平均である。

$$S_A(D) = |W_D| \cdot A_w \tag{7}$$

A_w:平均単語影響度

$$A_w = \frac{1}{|E_w|} \sum_{\{w|w \subseteq E_w\}} E_3(w) \tag{8}$$

E_w:全文書において出現した単語影響度の集合

式 (6)、(7) より、文書 D に付加する興味スコア $S(D)$ を以下のようにした。

$$S(D) = \frac{S_E(D)}{S_A(D)} \tag{9}$$

式 (9) によって、求められた値 $S(D)$ を文書 D の興味に対するスコアとして扱った。各文書に興味スコアを付加した後、値の降順により並び替えを行い、出力とした。

5 実験と評価

5.1 学習データ

Web 上で公開されている朝日新聞社のニュースランキング「アクセス Top30」を用いた。朝日新聞社のニュースランキングの記事及び一日に出現する記事を収集し、学習データ及び、入力データとした。収集期間は順位情報付きデータを 2004 年 4 月から 7 月までの 4ヶ月間、全記事のデータも同様に 2004 年 4 月から 7 月までの 4ヶ月間の記事を収集した。収集した記事数は順位情報付き文書数：3630、全文書数：14856、のべ文書数：18486 となった。

一日に取得できる順位付き記事数と全記事数を表 1 に示す。

表 1: 一日分の順位付き記事数と全記事数

全記事数	約 130
ランキングに入る記事	約 26

ニュースランキングでは、実際には順位付き文書を 30 位まで取得することができる。しかし、このランキングは前日に多くアクセスされた記事のランキングであるため、前日の記事のランキングではない。すなわち、2 日以上前の記事が継続してランキングに入る場合がある。

実験では入力データを一日分の記事に限定するため、入力対象として決めた日付の記事だけを扱うこととした。そのため同日に現れた記事だけを利用するので、平均して表 1 に示したような値となる。よって一回の実験においては約 130 記事を入力とし、その中に約 26 記事の順位内記事が含まれていることになる。

実験で学習データとして利用したのは、2004 年 4 月から 6 月までのデータである (順位情報付き記事数：2730、全記事数：11916)。

5.2 ランキング内記事の出現数

各記事に興味スコアを与え、興味スコアによって降順に並び替えて出力を行った。結果の一部を付録に示す。

出力順で 30 記事分を取り出し、その 30 記事中に実際のランキング 30 位以内に入った記事数を調べた。表 2 は、出力の上位 30 記事中において、実際のランキング 30 位以内に入った記事数を数え、これを 30 回分行った時の分布を示している。

表 2: ランキング文書出現数

出力上位 30 記事が ランキング 30 位以内に入った記事数	件数
6	1
7	0
8	4
9	2
10	8
11	5
12	4
13	5
14	1
合計	30

スコアによって降順で並び替えた上位 30 記事に、実際のランキング 30 位以内に入った記事数の平均値は、10.6 記事であることから、精度は平均で約 3 割となった。

5.3 順位付き記事の取得率

一日に出現する記事数及びその中に現れるランキング 30 位以内の記事数は入力データセットにより異なる。そのため一日ごとに値を正規化した結果を示す。図 2 において横軸は、一日に出現する記事数と出力記事数の割合を示している。縦軸は出力した文書中に実際のランキング 30 位以内の記事が含まれている数とランキング 30 位以内の記事の総数の割合を示している。この曲線は左上に近づくほど良いこと意味している。

なお、直線で示されている線はランダムに取った場合の理論値である。

興味スコアで並び替えた記事を、一日に出現した記事数の半分の数を出力した場合、その記事には実際のランキング 30 位以内に入った記事の約 7 割を含んでいる。

5.4 順位相関

出力の順位と実際の順位の比較のために、順位が付いている記事のみを使用して、興味スコアを付加し、並び換えた。比較には出力の順位と実際の順位

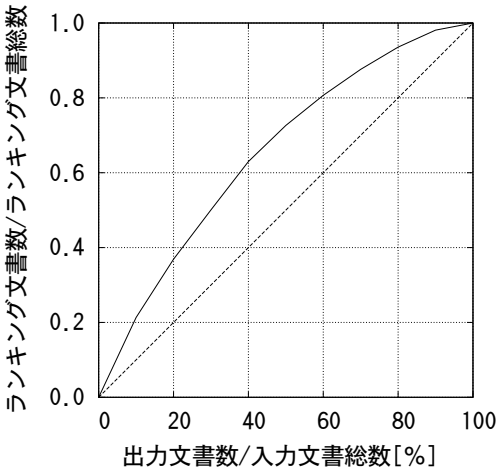

図 2: ランキング内文書の取得率

で相関値を算出した。相関の強さはスピアマンの順位相関係数を式 (10) により算出し、結果を表 3 及び図 3 に示す。

$$\gamma = 1 - 6 \cdot \frac{\sum_{x=1}^{n} (d_x)^2}{n \cdot (n^2 - 1)} \tag{10}$$

d_x:入力された二つの順位の差　n:順位の総数

表 3: 出力と正解の順位相関値

平均順位相関値	0.20
最大順位相関値	0.54

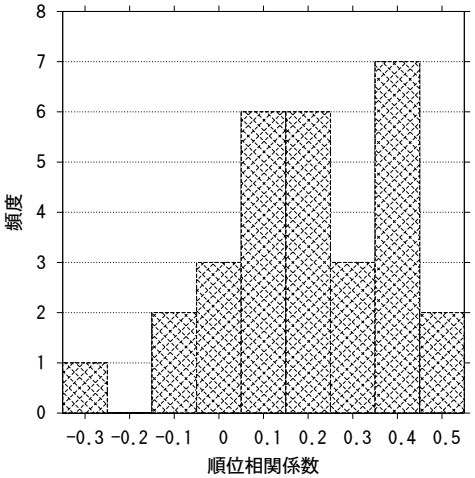

図 3: 出力と正解の順位相関の分布

相関の分布は、1 件が負の相関 (-0.2 以下)、ほぼ相関なし (-0.1〜0.1)11 件、正の相関 (0.2 以上) が 18 件となった。これより、程度は弱いが正の相関に傾いていることがわかった。また正の相関を示しているもののうち 0.4 以上の 9 件についてやや強い相関を示している。

5.5 上位の順位付き文書について

順位と記事の興味の強さがべき乗の法則に従うことで、上位順位の記事と下位順位の記事の興味の強さの差は大きい。一方、下位順位の記事が保持する興味の値は低い値でなおかつ下位順位の記事同士の興味の強さの差は小さい。よって興味の値が大きい10位までにしぼって5.2節と同様の実験を行った。出力の上位10記事中において、実際のランキング10位以内に入った記事数を数え、これを30回分行った時の分布を表4に示す。

表 4: ランキング文書出現数

出力上位10記事が ランキング10位以内に入った記事数	件数
0	2
1	8
2	9
3	5
4	5
5	1
合計	30

スコアによって降順で並び替えた上位10記事に、実際のランキング10位以内に入った記事数の平均値は、2.2記事であった。

6 考察

6.1 出力について

全体的には偏りが現れているものの、その範囲は大きく、出力の上位30件中で、約10記事程度が実際のランキング30位以内に入り、精度は平均で3割であった。

出力を観察すると、スポーツ記事全般のスコアが低い。また、スコアの全体的な傾向では、株に関する記事も同様にスコアが低い。よって株に関する記事、スポーツ記事、その他の記事の三種類に分割され、株に関する記事もスポーツ記事同様にスコアが低く上位に現れにくい結果となった。これは記事の特性上、スポーツ記事は同様の単語が使われる傾向があり、単語当たりのランキング内の使用率が減少したためスコアが低くなってしまったと考える。株に関する記事も土日以外で毎日出現していることから、スポーツ記事と同様の出現数の多さが記事内で使用されている単語の影響度の低下の原因となっていると考える。

実際の順位では株関連の記事がランキングに入ることはほとんどないため、株関連の興味が低いということは実際の傾向をうまく表している。しかし、スポーツ関連の記事はランキング上位に入ることがあり、対処が必要となる。出現頻度が多く、似た単語が使われている文書を判別するには、単語の影響のみだけでは判別する要素が少ないので、共起性等を考慮した方法を使用することが考えられる。

6.2 単語の影響度について

順位情報から推定した単語の影響度は、文書の順位を上昇させるものと下降させるものがある。

個人の興味では、個々の尺度の違いにより、順位を上昇させる語と下降させる語の存在がある。しかし多くの人の興味を捉えた場合、この二つの取扱いをどうするかが問題である。

取得した単語の影響度のデータを付録に示す (表5)。

単語の影響度の平均値は式 (7) の結果0.022となった。これは順位を上昇させる単語と、下降させる単語の境目である。多くの単語が平均値より上に存在する。文書の興味スコアは、文書に含まれる単語の影響度がどれだけ大きい値を持っているかに大きく左右された。影響度はある程度頻度を持つ単語においては0.01〜0.1までの範囲に収まる。また低頻度の語については値の範囲が広がっており、大きな影響を持っている。

6.3 全体の傾向と可能性

図4は出力された出力順位の各範囲において実際のランキング30位以内の文書が出現する回数の平均値を示している。

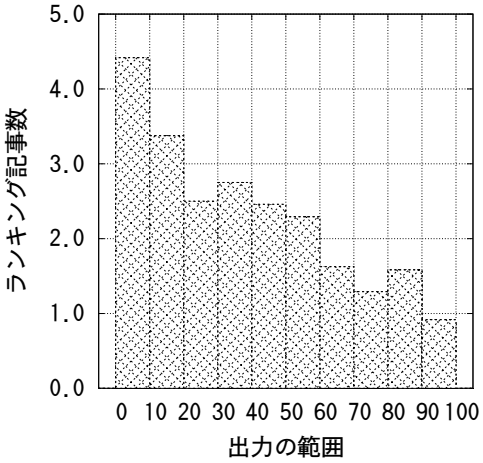

図 4: 出力順位の範囲に対する平均正解記事数

図4において、横軸が出力順位の範囲であり、その区間内にランキング内の記事数がどれだけ出現しているかを縦軸に示している。

例えば出力された20位から30位までの記事中にはランキング30位以内の記事が平均2.5記事含ま

れている。

グラフより、出力の上位ほど実際にランキング 30 位以内に出現した記事が含まれる数が多いことが分かる。この値とサンプル数では十分とは言えないが、順位情報によって興味の強い文書に偏りが発生しているため、順位情報を興味の値に近づけた興味強度は、文書を興味の大小で選択できる可能性がある。

6.4 ニュースランキング固有の影響要素について

ニュースランキングはアクセス数により、ランキングされている。0 時から 24 時までの間のアクセス数で順位付けを行っていることや、Web 上に存在するという条件により次のようなことに影響を受けている。

- 記事の発表時間に影響
  発表時間に関わらず、0 時から 24 時までの間で順位を決めているため、記事の掲示時間が同じではない。また新しく発表された記事はトップページ等に掲載されるため、アクセス数が増える時間帯等の影響も考えられる。
- リンク数の影響
  記事へのリンクが必ずしも同じ数だけであるとは限らない。ニュースサイト等の存在により、一部の記事のアクセス数が突出する場合も考えられる。また、ニュースサイトにリンクされるというのは、ニュースサイトの管理者にとって興味があるということも言えるため、個人の興味ではあるが、リンク数が興味示す場合もある。

6.5 今後の課題

本研究では、順位情報を用いて単語が順位に与える影響を推定することで興味の大小を考えている。しかし、実際には共起によっても影響があり、順位を左右させる。例えば小学生が中心の話題の記事というものがあった場合、小学生というだけで興味を持つかどうか判断するよりも、そこに、殺人や誘拐等の共起性を調べて、「小学生+殺人」または「小学生+誘拐」という要素として扱うことで精度が良くなると考えられる。さらに、ある固有名詞が中心の記事があった場合も、それにある単語との共起によって興味の度合を推定することが可能と考える。

今回はニュース記事を使用したが、他のランキングでも同様のことが可能か試していないため、今後行うべき課題の一つである。

7 おわりに

多くの人が興味を持つ文書と持たない文書を判別する手がかりとして、順位情報の利用について試みた。興味と言うパラメータは一つの要素だけで捉えられるものではないが、順位情報を興味に直接関係する値へと近似し、興味の強度として使用した結果、多少の傾向が現れる。現段階での精度は 3 割程度であるが、さらに共起性や、興味に関するパラメータを増やすことで精度の向上が期待できる。

使用した言語資源及びツール

(1) アサヒ・コム アクセス Top30, http://www.asahi.com/whatsnew/ranking/

(2) 形態素解析器「茶筌」,Ver2.3.3, 奈良先端科学技術大学院大学, 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

参考文献

[1] Lada A.Adamic, Bernardo A.Huberman :Zipf's law and the Internet, *Glottometrics*,vol.3,pp.143-150, http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf(2002)

[2] 石井 恵, 中渡瀬 秀一, 富田 準二 :名詞句と単語の勢いを用いた話題抽出手法, 情報処理学会研究報告,NL160-12(2004)

[3] 金田 重郎:現代用語辞書を用いた流行コンセプト予測, AFIIS シンポジウム, 学術フロンティア「知能情報科学とその応用」プロジェクト, http://afiis.doshisha.ac.jp/meeting/symposium_02/(2002)

[4] 武田 賢士:WWW を用いた時事的な話題の分析, 広島市立大学 情報科学部 卒業論文, http://www.nlp.its.hiroshima-cu.ac.jp/(2003)

[5] 長野 徹, 武田 浩一, 那須川 哲哉:テキストマイニングのための情報抽出, 情報処理学会研究報告,FI60-5(2000)

[6] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学:document stream における burst の発見, 情報処理学会研究報告,NL160-13(2004)

付録

以下に出力の結果を示す。入力データは 2004 年 7 月 14 日のデータである。

- [実際の順位 -は順位なし], スコア, 記事のタイトル

出力上位 10 記事のデータ

(1) [1] 2.26 交番前で暴行、大けが目撃者「警官は見ていた」埼玉

(2) [7] 2.10 「イラクの現実、忘れるのは無理」高遠さん、決意新た

(3) [2] 2.08 曽我さん一家がインドネシア出発、チャーター便で東京へ

(4) [17] 2.02 ジェンキンス氏を訴追の方針は譲らず米政府

(5) [-] 2.00 曽我さん一家がジャカルタを出発、チャーター便で東京へ

(6) [-] 1.95 19 人連続殺害容疑者を逮捕韓国の警察当局

(7) [25] 1.93 回転ドア、制動距離把握せず六本木ヒルズ事件の製造元

(8) [29] 1.91 佐世保市で御手洗怜美さんとのお別れの会級友ら献花

(9) [-] 1.84 「北朝鮮からミサイル技術購入」パキスタン元首相証言

(10) [-] 1.78 福井豪雨で死者２人不明２人、避難勧告３万９０００世帯

出力下位 10 記事のデータ

(1) [-] 1.23 女子野球世界大会、日本がカナダに勝つ

(2) [-] 1.22 全英オープン最終日丸山は通算４オーバー、２８８

(3) [-] 1.22 高津が６セーブ目アスレチックス戦

(4) [30] 1.21 松井稼５打数２安打フィリーズ戦

(5) [-] 1.19 松井秀４打数無安打タイガース戦

(6) [24] 1.19 ジョーンズが女子２００M棄権陸上全米選手権

(7) [-] 1.18 巨人、サヨナラ勝ちで４連勝

(8) [-] 1.15 バスケ日本女子、ロシアに敗れる欧州遠征初戦

(9) [-] 1.09 仙台育英、愛媛ユなど勝つ

(10) [-] 1.09 近鉄、本塁打攻勢でサヨナラ勝ち

単語の影響度の一部を示す。

表 5: 単語の影響度

単語	E_3	頻度
自殺	0.092	69
暴行	0.089	82
中学校	0.083	57
同級生	0.082	58
侵害	0.079	84
人権	0.079	88
ジョンイル	0.076	57
女児	0.070	104
小学	0.061	52
誘拐	0.060	68
爆弾	0.059	141
テロ	0.046	484
学会	0.042	66
リコール	0.041	145
選挙	0.031	409
銀行	0.030	332
イングランド	0.030	82
工作	0.030	51
支店	0.025	61
告発	0.025	87
報酬	0.025	104
議長	0.025	241
予防	0.025	61
宇宙	0.019	53
日経	0.009	172