因子組み合わせによる順位付け文書からの興味因子判別

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

1 はじめに

近年 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。このことから、情報を有効に活用するために大量の情報から必要な情報を取得する技術が重要となってきている。特にテキストから人の興味や関心などの情報を得るために、Web 掲示板や Blog などの主観情報が含まれやすい文書を用いてトレンド分析や流行・話題の発見といった研究が多く行われている。[2,3,4]

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 (1) を選択した。ニュースランキングは、アクセス数によって記事に順位付けを行い、順位と共に記事が公開されている。アクセス数が多いほど大勢の人が興味を持った記事であることから、順位情報と興味には深い関係性があると考える。

そこで、順位情報に着目し、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を判別する手がかりにすることができると考える。そのため、文書に含まれる内容語を因子として扱い、因子の組み合わせを作成することで、順位情報から興味に影響している要素の判別を試みた。

また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、興味の予測に繋がると考える。

2 興味判別処理の問題

文書に含まれる内容語を因子として扱い、順位情報を用いて興味に影響している因子の判別を行う。(例 1)

例 1) 影響する因子の判別

文書:
インド北部で厳しい冷え込みが続いている。... 観測史上２番目の寒さを記録した（最低記録は３５年の零下０．６度)。... 凍死など寒さが原因で死亡している。...

興味因子: インド、冷え込み、凍死、etc

興味の判別を行うことの問題点として、順位情報を有効に扱う点と、因子の処理単位をどのように考慮すべきかが挙げられる。

我々は以前の研究 [5] で処理単位を単語単体 (例 2-1) として興味の推定を行った 。しかし因子単体では処理単位の持つ情報が少なすぎたと考え、処理単位に因子の組み合わせである複合因子 (例 2-2) を用いて興味の判別を試みた。

例 2) 処理の単位

1： 因子単体：インド、北部、厳しい、冷え込み

2： 複合因子：{ インド、北部 }、{ インド、北部、冷え込み }、など

3 提案手法

興味因子を判別するための手法は、大きく分けて「順位情報から複合因子の興味の強さを推定する」と「複合因子を用いて文書の興味の強さを推定する」の 2 つから構成される。また、因子の組み合わせから興味の予測を行い実際のランキングとの比較を行う。

3.1 ニュースランキングについて

ニュースランキングはアクセス数によって文書の並び替えを行っている。これは不特定多数が興味を持った文書を選択した結果が反映されている。すなわち、ランキング上位の文書は大勢が興味を持った文書であると考えられ、興味と深く関係していると言える。

3.2 因子について

文書の形態素解析を茶筌 (2) で行った。文書が内容語のみによって構成されているとし、文書から内容語を取り出す。文書は例 3 に示すようになり、内容語を因子としている。

例 3) 因子集合

原文：トヨタ、レクサスの最上級セダンを全面改良。

因子集合：トヨタ、レクサス、最上級、セダン、全面、改良

次に因子を組み合わせた複合因子の作成条件を示す。

- 因子の出現順は考えない。
- 組み合わせはひとつの文中から探す。
- 2〜5 までの組み合わせで行う。
- 組み合わせは文書内に複数出現しても出現数は１として扱う。

複合因子は、2 要素以上、5 要素以下からなる因子集合の全部分集合である。本研究では複合因子で処理を行い、因子単体については扱わない。この複合因子を用いて次節以降の興味の推定の処理を行う。

3.3 複合因子の興味スコア

複合因子の興味の強さを興味スコアとして付与する。興味スコアはあらかじめ収集したランキングに出現した文書、ランキングに出現しなかった文書の両方を学習用データとして用いて推定する。

複合因子の興味スコアを推定する手順を次に示す。

(1) 順位情報をアクセス数に変換する。順位情報が付与されていない文書に対しても別処理を行い、アクセス数を付与する。

(2) 学習用データ内で複合因子が出現する文書数を数える。同時に出現文書に付与されているアクセス数を獲得する。

(3) 複合因子の出現文書数とアクセス数を用いて、興味スコアを計算する。

以上の手順を通して、すべての複合因子に対して興味スコアを計算する。

3.3.1 順位情報の扱い

ランキングに付与されている順位情報は、ランク付けする際に使用された要素の上下関係を示した値である。例えば本研究で使用しているニュースランキングはアクセス数によってランキングされている。

そこでまず順位情報をアクセス数に変換することを考える。アクセス数に変換することで順位情報が上下関係という値からランク付けする際に使用した要素に近づけることができる。それにより興味に間接的に関係する値から興味に直接的に関係する値に近づくことになる。

本研究では順位とアクセス数は Zipf's の法則 [1] に従うとして順位情報をアクセス数の値に変換することを試みた。

$$hit(Rank) = \frac{1}{Rank} \tag{1}$$

$Rank$:順位、$hit(R)$:順位 R のアクセス数

式 (1) により順位情報をアクセス数に変換した。

次に順位情報が付与されていない文書に対する処理について述べる。順位情報は必ずしも全ての文書に付与されておらず、ランキング外という文書集合が存在する。このランキング外の文書については正確な順位情報は付与されていない。本研究で使用したニュースランキングでは 30 位まで順位が付与されていて、31 位以降の記事には順位情報は付与されていない。この 31 位以降のランキング外に対してもアクセス数を付与するために、全文書数から全アクセス数を推定した。そして推定した全アクセス数に従って、ランキング外の文書には均等な値を振り分けた。

$$hit_{rankout} = \frac{sum(Doc_{day}) - sum(Ranking)}{Doc_{day} - Ranking} \tag{2}$$

$$sum(N) = \sum_{num=1}^{N} hit(num) \tag{3}$$

Doc_{day}:一日に発表される記事数
$Ranking$:ランキング付与数
$hit_{rankout}$:ランキング外アクセス数
$sum(N)$:1 から N までのアクセス数の合計

式 (1), 式 (2) を用いて全ての文書にアクセス数の情報を付与した。

3.3.2 複合因子の興味スコア付与

複合因子に興味の強さを興味スコアとして付与する。興味スコア付与のため学習用データ内における複合因子の出現文書数とアクセス数を取得し、2 つの値を用いて興味スコアを計算する。また、すべての複合因子に値を付与すると膨大な量になってしまうため興味スコアは出現文書数で閾値を設けて数を制限している。本研究では出現文書数が 10 以下の複合因子は除外した。

$$Score(w) = \frac{access(w)}{df(w)} \tag{4}$$

w:複合因子
$Score(w)$:複合因子 w の興味スコア
$access(w)$:複合因子 w の総アクセス数
$df(w)$:複合因子 w の出現文書数

アクセス数は式 (1) 及び式 (2) を用いるため、複合因子の興味スコアは 0〜1 の範囲の値を取る。

以上の処理を用いて作成した複合因子と興味スコアのリストを 3.4 節以降の文書の興味の強さを求めるために使用する。

3.4 文書の興味スコア

未知の文書を入力し、その中に含まれている複合因子から文書の興味の強さを推定し、興味スコアとして付与する。複合因子の興味スコアは 3.3.2 節で求めた値を用い、複合因子は値が付与されているものだけを対象とする。

文書の興味スコア付与手順を示す。

(1) 複合因子に興味スコアが付与されているものを文書中から探し出す。

(2) 探し出した複合因子全てを利用して文書の興味スコアを付与する。

複合因子は 3.2 節と同様の方法で作成する。複合因子を構成する要素の重複を許し、出現順序は考えていない (例 4)。

例 4) 複合因子の作成
入力文：確認した際には、異変はなかったという。
複合因子：{ 確認、際、異変 }、{ 確認、異変 }、{ 際、異変 }、{ 確認、異変 } など

3.4.1 文書の興味スコアの付与

文書の興味スコアは文書に含まれる複合因子を利用して付与する。複合因子の興味の強さがどの程度順位に影響するか求めるため、興味の強さによってグループ分けし、文書内に出現した複合因子の数をスコアグループごとに数える。スコアグループは複合因子の興味スコアの値を 0.01 刻みの 100 グループに分割した。スコアグループごとの出現数とスコアグループの重みを用いて文書の興味スコアを計算する。

$$Score_{text} = \frac{\sum_N Score_{group}(N)}{w_{all}} \tag{5}$$

$$Score_{group}(N) = G(in(N))W(N) \tag{6}$$

$in(N)$: スコアグループ N の複合因子集合
w_{all}:入力文書に含まれる複合因子の総数
$G(A)$:入力文書に複合因子集合 A が出現した総数
$W(A)$:興味スコアグループ A の重み

興味スコアグループ重みは学習用データを使って求めた。1 つのグループに属する複合因子の全てについて一位とランク外の文書に出現する回数を数え決定した。

$$W(group) = \frac{df_1(wg(group))}{df_{out}(wg(group))} \tag{7}$$

$W(A)$:スコアグループ A の重み
$wg(A)$:スコアグループ A に含まれる複合因子集合
$df_1(w)$:順位 1 位の文書群で w の出現文書数
$df_{out}(w)$:順位外の文書群で w の出現文書数

以上の式 (5)〜式 (7) から文書の興味スコアを算出する。

複合因子の興味スコアが高いものは興味に対して影響が大きい因子である。文書が持つ複合因子を興味スコアで並び替えを行い、上位の複合因子を興味因子とした。

4 実験

4.1 使用データ

本研究では Web 上のニュースランキングを利用した。実験で使用したニュースランキングは、朝日新聞社の「アクセス Top30」(1) である。学習用の文書として収集期間は順位情報付きデータを 2004 年 4 月から 12 月までの 9ヶ月間、順位外のデータも同様に 2004 年 4 月から 12 月までの 9ヶ月間の文書を収集した。評価用の文書も同様に朝日新聞社の「アクセス Top30」から収集を行っている。評価用の文書については学習と時期が被らない 2ヶ月分 (2005 年 5 月、6 月) のデータを収集した。

収集した順位付き記事数と全記事数を表 1 に示す。

表 1: 収集文書数

	学習用データ	評価用データ
順位付	8830	1830
順位無	25587	5334

4.2 作成した因子の組み合わせ

学習用データから複合因子に興味スコアを付与した数を表 2 に示す。

表 2: 複合因子数

構成要素数	複合因子数
2	148049
3	132282
4	144760
5	165749

複合因子の興味スコアは 0〜0.62 の範囲に分布している。

5 評価

5.1 上位３０記事までの出力結果

入力は一日の間に掲載される記事を未知の文書集合 1 セットとして用い、文書の興味スコアを付与した後に並び替えを行い出力した。出力順で上位 30 記事分を取り出し、実際のランキングに出現した記事数を調べた。図 1 は、出力の上位 30 記事中において、実際のランキング内に入った記事数の分布を示している。なお、1 セットの平均入力記事数は約 100 記事であり、60 セット実験を行った。

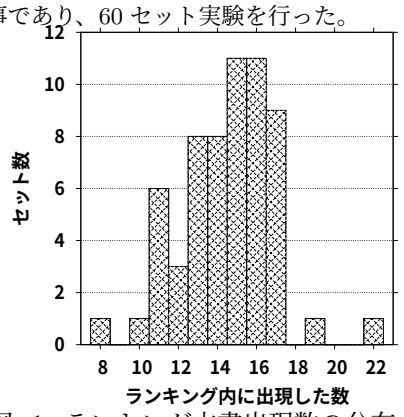

図 1: ランキング文書出現数の分布

文書の興味スコアによって降順で並び替えた上位 30 記事に、実際のランキング内に入った記事数の平均値は、14.52 記事である。抽出精度は平均で約 48 %となった。

5.2 順位相関

出力の順位と実際の順位を比較するために、順位が付いている 30 記事のみを入力とした。入力した文書に興味スコアを付与し、並び換えを行い、出力の順位と実際の順位で順位相関係数を求めた。順位相関はスピアマンの順位相関係数により算出した。順位相関係数の結果を表 3、相関係数の分布を図 2 に示す。

表 3: 出力と正解の順位相関値

	複合因子	因子単体 [5]
平均順位相関値	0.22	0.20
最大順位相関値	0.73	0.54

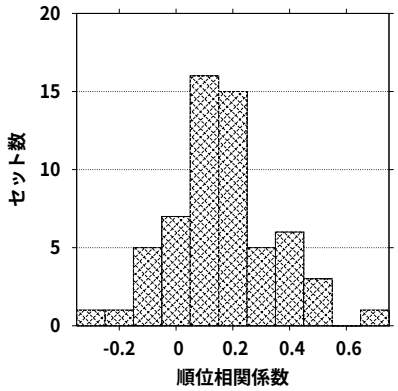

図 2: 出力と正解の順位相関の分布

順位相関値は評価用データセット 60 件中、正の値を示したものが 51 件、負の値を示したものが 9 件であった。

5.3 興味因子

記事の中で複合因子の興味スコアが高いものは記事の順位を上昇させる因子であり、これを興味因子群とした。複合因子を興味スコアで並び替えを行い、上位の 3 つまでを取り出した結果の一部を例 5 に示す。例 5-5 は複合因子が 2 つしか取り出せなかったため 2 つのみ表示している。出力結果を観測すると記事の概要を示す単語が含まれる結果が多く存在した。

例 5) 「記事タイトル」
{ 興味スコアで上位の複合因子 }
1.「辺見えみりさんと木村祐一さんが婚姻届」
{{ テレビ、番組 },{ タレント、東京 } 、{ 婚約、出す }}
2.「幼児連れ去り、容疑の夫が単独で計画」
{{ みる、長女 }、{ 誘う、容疑 } 、{ みる、わかる、事件 }}
3.「ミツバチ４０万匹死ぬ　いたずらの疑い 」
{ 同署 発見 },{ 男性、被害 } { 男性、発見 }}
4.「畳６００枚分、巨大たこフワリ 」
{{ 小学校、東京 }{ 最大、史上 }{ 大きい、日本 }}
5.「プジョー、「猫顔」の新型４０７を発売 」
{{ 全国 発売 }、{ 発売 新型 } }

6 考察

6.1 精度

文書の興味スコアによって並び替えた上位 30 位が実際のランキング内に出現する数について、 単語単体を処理単位とした場合その記事数は平均で約 10 文書であった [5]。複合因子で処理した結果は約 15 文書であるため記事判断の精度は上昇している。また、順位相関値の平均値による評価

は単語単体で処理した場合とほぼ変わらない 0.22 という結果だった。

6.2 興味因子が示す情報

5.3 節の結果を観察すると、記事が興味を持たれているという主な要素として記事の分野を示している語が多く含まれていた (例 5)。 記事特有の特徴ではなく、記事を大雑把に示した語が含まれている。そのため、文書の興味スコアは大雑把にとらえた記事の特徴、例えば「誘拐、警察、芸能、株価」などの分野によって文書が興味を持たれるかどうかの可能性を示している。また、文書の大雑把な特徴を捉えており、大勢の興味としては良好な結果に近づいていると考える。しかし、興味が持たれるのは記事の分野を示すような語だけではないため固有名詞などを捕えられるようにしていくことが検討課題である。

また、仮に高い興味スコアの複合因子を含んでいる文書でも、低スコアの複合因子が大量に含まれることでスコアを下げることがある。実際は文書の興味の強さが強いのに弱いと判断される文書が存在した。これについて現在は捕えることは難しく、別手法が必要となる。

6.3 未知データの重要性

学習時に学習できなかった未知語については本研究では無視している。しかし、未知語が文書の中心となると複合因子として一切捉えることが出来なくなるため改良が必要である。

例 5-1 については「タレント」という単語を捕えているが、固有名詞が興味の中心であるなら「タレント」が誰なのかが選択される方が望ましい。このような固有名詞を得ることができない原因として、学習用データに存在しない場合と、固有名詞となると学習時に出現回数が減り、判断材料として興味スコアが正確に付与されないことが挙げられる。

固有名詞を中心に記事を捕えることで新たに興味の原因を捕えることが望める。例えば、「訃報」の記事であっても固有名詞の認知度などにより興味の強さが変化する。しかし、このような固有名詞自体に差異を求め興味の尺度として用いるのは、非常に難しい問題である。

6.4 因子の組合せについて

本研究では「会社」という因子は、「証券会社、関連会社」に含まれ、組み合わせの時には分かれて複合因子を作成する。「証券会社」と「関連会社」の 2 つの間で興味の判断をする時、同一の因子「会社」は効果的ではない。そのため複合名詞を 1 つにした後、組み合わせを行えば判断材料として有効なものがあるのではないだろうか。例えば「関連会社による虚偽事実」から { 関連、会社、虚偽 } のように因子の組み合わせを作成する場合と “関連会社” のように複合名詞にすることで {“関連会社”、虚偽 } とする場合を考える。発生する組み合わせは異なったものとなり、因子の違いが明確になるため興味の判断が出来るようになる。複合名詞を取り入れて処理する問題点としては出現数が減少し複合因子の興味スコアをうまく付与できなくなる事が考えられる。

また、文書の興味スコア推定において、文書から複合因子を探索するとき、複合因子を構成する要素の重複を許している。同じ要素が何度も出現するため学習データが過剰に作用し精度を下げている可能性が有る。このため一度使用した因子は使わないようにし、ラティス構造を探索するような処理が有効かもしれない。これによって複合因子に付与されている興味スコアが最も高いものを選択しながら決定できる。

6.5 興味のパタン

本研究では記事の興味の強さを処理するモデルを 1 種類で行っている。しかし、実際の順位に出現する記事には興味を持たれる特徴に複数のパタンがあると思われる。特徴を捕えられない文書に対してはいくつか違う処理を行い、複数の結果から文書に興味スコアを付与することが課題である。

興味が発生するパタンをいくつか次に記述する。

- 時系列による発生
- 文書内容の特殊性 (非常識性、特殊共起)
- 固有名詞+述部による発生

以上のパタンの興味は本研究では捕えることができない。また記事の特殊性や固有名詞 (または未知語など) 自体に違いを与えることは難しい問題である。

7 まとめ

順位情報を利用して因子の組み合わせに興味の強さを付与した。これを用いて記事の興味判断を行い 30 記事取り出した中で、平均 14.52 記事が正解であった。また、文書に含まれている複合因子の観察から文書の大雑把な特徴を捉えているという結果は大勢の興味を捕えるということに近づいていると考える。

さらに興味を捕える為の課題として未知語に対する処理や、複数モデルによる興味スコアの付与などが挙げられる。

謝辞

本研究の一部は、平成 17-19 年度 総務省 戦略的情報通信研究開発推進制度 (SCOPE) の支援によって実施した。

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] 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)

[3] 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)

[4] 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)

[5] 沢井 康孝, 峠 泰成, 山本 和英:順位付け文書からの影響因子マイニング, 情報処理学会研究報告,NL163-23(2004)