文書に対する大衆の興味の強さの推定

沢井 康孝[†]・山本 和英[†]

ある入力文書が多くの人にとってどの程度興味や関心を持つかを算出する指標を提案する。各個人の興味や関心は多種多様であり、これを把握することで情報のフィルタリング等を行う研究は知られているが、本研究では不特定多数すなわち大衆が全体でどの程度の興味を持つかについて検討を行った。このような技術は、不特定多数に対して閲覧されることを想定している Web サイトにおける提示文書の選択や表示順の変更など、非常に重要な応用分野を持っている。我々は大衆の興味が反映されている情報源として順位付き文書を使用した。本手法ではこれを学習データとして利用して、文書に含まれる語句及び文書自体に興味の強弱を値として付与する手法を構築した。興味を値として扱うことで、興味の強弱を興味がある・ないの2値ではなく興味の程度を知ることや興味発生の要因分析を行うことが可能である。提案手法は、文書に含まれる語句を興味判別する素性として扱い、内容語、複合名詞、内容語及び複合名詞の組み合わせの3種類について比較、議論した。評価は、ニュース記事のランキングを対象にして、実際の順位とシステムの順位を比較した。その結果、順位相関に基づいた評価値は0.867であり、手法の有効性を確認した。さらに、ほぼ興味を持たれない記事に対して抽出精度0.90を超える精度で弁別できることを実験で確認した。

キーワード：大衆の興味、順位情報、順位相関、テキストマイニング

Estimating Level of Public Interest for Documents

YASUTAKA SAWAI[†] and KAZUhide YAMAMOTO[†]

We propose a new measure to estimate level of public interest given a document. Although personal interests is of great variety, public interest, that is collection of personal interests, has consistency to some extent regardless of time difference. The task here is not to know whether a given document has interest or not, but to know how much interest a given document has, that expects enabling deep interest analysis by use of our measure. This problem has many applications such as display control of documents on the Web, that is assumed to be seen by public. We use in this paper document collection with ranking information in terms of public interest. We estimate level of interest for each word, and then for each document by utilizing the ranking information. As feature set we use three kinds: content words, compound words, and the combination of them. In the evaluation we use newspaper ranking as a source, and evaluate the performance by comparing our output to the real ranking. The results illustrates that the extended rank coefficient of these two rankings is 0.867. We also show that more than 0.90 accuracy is attained for rejecting little interest documents.

[†] 長岡技術科学大学 電気系, Department of Electrical Engineering, Nagaoka University of Technology

Key Words: *Public interest, Ranking, Rank coefficient, Text mining*

1 はじめに

近年、Web の普及や様々なコンテンツの増加に代表される不特定多数の情報の取得や不特定多数への情報の発信が容易になったことで、個人が取得できる情報の量が急激に増大してきている。個人が取得できる情報量は今後さらに増え続けるだろう。このような状況は、必要な情報を簡単に得られるようにする一方で不必要な情報も集めてしまう原因になっている。

この問題を解決する方法として大量の情報の中から必要な情報だけを選択する技術が必要で、これを実現する手段として検索、フィルタリング、テキストマイニングが挙げられる。このような技術はスパムメールの排除や Web のショッピングサイトの推薦システム等で実際に使われている。

本論文では大量の情報の中から必要な情報を取得する手段として人間の興味に着目し、文書に含まれる語句及び文書自体に興味の強弱を値として付与することを提案する。本論文では、不特定多数の人がどの程度興味を持つかに注目した。すなわち不特定多数を全体とした大衆に対する興味の程度である。興味の強弱を語句及び文書自体に付与することにより人間の興味（文書の面白さ、文書の注目度）の観点で情報を選別することが可能となるだけでなく、

興味の強弱を値として与えることで興味がある・ないの関係ではなく興味の強さの程度を知ることができる。また、文書に含まれる語句に与えた興味の強弱の値から文書のどの部分が最も興味が強いか明らかになるため、文書のどの部分が興味の要因となるのか分析を行うことが可能である。

このように語句の興味の強弱自体を明らかにすることは、例えばタイトル作成や広告等において同一の意味を示す複数の語句の中から興味強い語句を選択する際の基準として利用できるため、興味を持ってもらえるように文書を作成する支援となることが期待できる。さらに、Web 上でのアクセスランキングなどはアクセス数の集計後に知ることのできる事後の情報である。本論文の文書自体に付与する興味の強弱の値を利用することでこの順位を事前に予測することが可能となり、提示する文書の選択や表示順の変更などをアクセス集計前に利用することが期待できる。

大衆の興味が反映されているデータに注目することでこのような大衆の興味を捉えることが出来ると考える。また興味を持つことになった原因と持たれない原因を分析する手がかりになると期待できる。

本論文では、多くの人が興味を持つ文書を判断するため、まず興味の判断に必要な素性を文書から抽出する。次に抽出した素性に興味の強弱を値で推定して付与する。さらに興味の強弱の値が付与された素性から文書自体の興味の強弱を推定する。

2 章にて本論文で対象とする興味, 3 章にて関連研究, 4 章で順位情報の詳細, 5 章で提案手法について述べ, 6 章で評価実験及び考察を行う. さらに 7 章で提案手法の拡張について述べ, その評価を 8 章にて行う.

2 興味について

2.1 大衆の興味

興味はある対象に対して面白いと感じる又は魅かれている状態を示すものであり, 個人によって興味を持つ対象は様々である. 我々は個人の興味の集合を大衆の興味とした. 大衆の興味を引きやすいということはより多くの人に興味を持たれるものであり, このようなものを大衆の興味を得られやすい情報として扱っている. 前述のように各個人の興味は多種多様であり, 1 つの尺度では表せないものである. しかし不特定多数の人がいる中で興味を持つ人がどの程度いるかという大衆の興味として扱うことで興味の傾向などが得られると考えている.

2.2 対象とする興味

我々が定義した大衆の興味には次に示す 2 種類の興味が含まれていると考える.

- 時間変動を含んでいる興味
時間変動を含んでいる興味はトレンドや流行と呼ばれ, 今現在や過去のある時点で興味を集めているもの
- 対象自体が持っている興味
対象 (語句) 自体が普遍的に持つ「興味を発生させる強さ」を指し, 時間で変動せずに対象自体が持っている普遍的な興味の度合い

我々は語句自体が持つ時間変動を持たない後者の興味の強さに注目し本論文の対象とした. 例えばトレンド分析の結果「収賄の疑いで逮捕」が抽出された場合, これは現在注目を集めている特定の話題として抽出されたものである. 本論文ではこのような時間で変動するトレンドではなく, 「逮捕」という語句そのものが普遍的に持つ興味があると考え, その強さを明らかにするためにその語句に対して興味の強さを値として与えた.

3 関連研究

興味の分析や抽出を行う先行研究には, 多数の販売記録から対象者が興味を持つ商品の推薦を行うものや, 時系列分析を行って単語の出現傾向から現在注目されているイベントや単語を検出する研究などがある. さらに最近では文書の書き手が個人である Blog などを用いることで主観情報の分析や個人の興味, テレンド分析を行う研究が存在する.

奥村らは Web 上の文書を時間情報が含まれた document stream で扱い, burst 検出手法に基づいて分析を行い注目されている話題を検出する手法を提案している (藤木, 南野, 鈴木, 奥村 2003). このシステムでは Blog などに対する書き込み時間を基に単語の出現間隔が短くなっている箇所に注目すべき話題があるとして注目されている話題の検出を行っている. このように現在注目を集めている話題, 及び過去に注目を集めていた話題を検出するのはトレンド分析の 1 種であり, 興味を分析する研究の 1 つとして挙げられる.

また西原らは興味を持ってもらえるタイトルやアブストラクトの作成を目標に, 興味を引く研究発表のタイトルを作成する支援システムを提案している (西原, 砂山, 谷内田 2007). システムは論文タイトルを入力にとり, 興味を値として推定しこの値によって入力されたタイトル群に順位を付与している. 興味値としてタイトルに含まれる名詞の分かりやすさと, 単語の組み合わせの斬新さという 2 つの尺度からタイトルの面白さを評価している.

興味自体の分析を行っている研究として福原らの KANSHIN の開発が挙げられる (福原, 村山, 中川, 西田 2006). このシステムは Blog 記事を大量に収集し時系列分析することで社会の「関心動向」(トレンド)を把握することを可能としている. 収集した記事から単語が出現する頻度の推移を分析し, 興味の発生を「関心パタン」として周期型, 漸次増加型, 突発型, 関心持続型, その他の 5 つに分類している. さらに単語と気温の関係のような出現単語と現実で起こる現象の関係についても議論を行っている (Fukuhara, Murayama, and Nishida 2005).

興味を持つ商品や楽曲の推薦を行う研究として協調フィルタリングを使用した研究が挙げられる. 協調フィルタリングは複数のユーザが評価したデータが存在する中で, 対象者と嗜好等の類似度によって求めたいいくつかの類似性の高いユーザの情報を選択し, 使用者へ推薦する情報の選別に利用する. 個人の興味を対象としプロフィールやユーザー間の類似度を利用して情報の選択を行う. Resnick らはニュースの記事の集合から, 対象者の好みに合う記事を推薦するシステムを開発した (Resnick, Iacovou, Suchak, Bergstrom, and Riedl 1994). 個人がニュース記事に付与した 5 段階の評価値を利用している. Upendra らは楽曲のプレイリスト生成システムを考案している (Shardanand and Maes 1995). これは多数の人が作成したプレイリストから対象者にプレイリストを提供するシステムである.

これらの研究では単語の出現する頻度を時系列で解析することや, 出現した単語の斬新さ等の尺度を定義することによって興味を捉えている. これに対し本論文では文書の内容自体と大衆の興味が反映されているデータを基に興味の強さを推定した. 対象とする興味はトレンドのような時間で変動する興味ではなく, 語句自体が持つ時系列で変化しない興味の強さである. このような興味を対象に, 大衆の興味が反映されているデータから興味の強弱を値として推定する研究は従来行われていない.

4 順位情報付き文書

人の興味の情報が得られるものとして Web 掲示板, Blog, Web ページのアクセスランキング, アンケートなどが存在する. 我々は大衆の興味を含んでいる情報源として順位情報に注目し, Web で公開されているニュースランキングを選択した. ランキングは不特定多数の利用者の興味が反映されており, 大衆の興味を含むデータとして有効である.

4.1 順位情報と興味

本論文で扱う順位情報について述べる. 本論文における順位情報とは, ある対象を一定の基準で並び替えを行い, 昇順または降順に付与した順番の事である. 順位を付与する際に用いられる基準には様々な種類が存在する. 以下に順位情報と順位を決定した基準についていくつか示す.

順位情報と順位の決定基準

- アクセス数ランキング: アクセス数
- 選択式のアンケート: 投票数
- 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: クエリの使用回数

例として示した順位情報は人に選択されるほど上位の順位が付与される. そのため, 順位情報は多数の人の興味が集合した結果であり, 大衆の興味を強く反映していると考ええる.

順位を決定する基準であるアクセス数や投票数は人々に選択された数そのまま反映され直接的に人の興味に関係する. 順位情報自体はこの基準の上下関係である. この意味で順位情報は大衆の興味と間接的に関係している情報と考えられる. そこで, 本論文では大衆の興味という情報をスコアの形で推定するために, 順位情報をアクセス数やダウンロード数に変換することを考えた.

4.2 順位情報の扱い

4.1 節で述べたように順位情報は興味に対して間接的な関係を示している値である. 本論文で集めた順位情報はアクセス数によって順位が決定されている. そこでまず順位情報をアクセス数の値に変換することを考える. これは順位からアクセス数を推定することで行い, 順位情報の示す上下関係から順位付けの決定基準に変換することに相当する. この変換によって順位が持っている興味に対する間接的な関係をアクセス数のように興味に対して直接関係する値に変換することが可能になる.

アクセス数と順位の間にある関係は経験則でべき乗の法則に従うことが知られている (Adamic and Huberman 2002). 本論文ではこの経験則を利用して順位を興味に直接関係する値として変

換する．順位の変数 r とアクセス数の変数 h の関係をべき乗の法則に従うとし，両対数上で直線となる特性から式 (1) が仮定できる．式 (1) に順位とアクセス数の関係式を示す．

$$\log(h) = C_1 - C_2 \cdot \log(r) \quad (1)$$

式 (1) において C_1 及び C_2 は任意の定数を示している．本研究では式を簡単にするため $C_1 = C_2 = 1$ として扱う．この結果順位 r からアクセス数 h を推定する変換式 $Hit(r)$ を式 (2) に示す．

$$Hit(r) = h = 10^{-\log(r)} = \frac{1}{r} \quad (2)$$

式 (2) によって文書に付与されている順位 r はアクセス数に変換される．式 (2) で表される推定アクセス数は 0 から 1 までの値をとり，最大のアクセス数が 1 となるように正規化されている．

5 提案手法

本節では文書に興味の強弱を値として付与する方法について提案する．提案手法の概略を図 1 に示す．

文書に含まれる語句に興味の強弱を推定するために，まず文書に含まれる形態素から内容語

図 1 提案手法の大まかな流れ

を興味判別の素性として抽出した。なぜなら文書が興味を持たれるかはその文書の内容に依存しており、人が文書の興味を判断する際に文法的機能を示す機能語では直接的に興味を引く対象にはならないと考えたためである。

次に、今注目している素性が興味を引くかどうかを興味の強弱の値で表すことによって推定する。興味を値で推定する方法として学習データを利用した2つの尺度を用いた。1つ目の尺度として学習データにおいて順位付き文書に出現する回数と順位無し文書に出現する回数の割合を使用した。これは、興味を持たれやすい文書はランキングに出現しやすく、それに伴って文書に含まれる内容語がランキングの中で出現する回数が増えると考えられるためである。2つ目の尺度は出現順位である。興味が持たれやすいならば高い順位に出現することが多くアクセス数も高い値を持つと考える。ここでは順位から推定したアクセス数の平均を利用した。以上2つの尺度を本論文では興味を推定するために用いる。

最後に文書自体に興味の強弱を値で付与する。文書の興味の強弱は文書に含まれる素性に付与された興味の強弱の値によって決定する。以上の3つのステップで本システムは興味の強弱に関する値を文書に付与する。

(1) 興味を推定するための素性の抽出

入力された文書を形態素解析し内容語を抽出する。

(2) 素性の興味強度の推定

順位付き文書を利用して素性に興味の強弱を値で付与する。

(3) 文書の興味強度の推定

文書に対して興味の強弱を値で付与する。

次節より各ステップの詳細について説明する。

5.1 興味を推定するための素性の抽出

入力されたそれぞれの文書から興味の判別を行う際に重要であると考えられる素性を抽出する。形態素解析器に茶筌¹を用いた。人が文書に対して興味があるかどうか判断する場合、助詞、助動詞、接続詞などの機能語に注目することは考えにくく、文書の内容が表れている内容語が重要であると考えられる。また出現した単語の順序よりも、どのような単語が出現しているかが重要であると考えられる。そこで、文書を内容語の集合 (bag-of-words) で表現することにした。本論文で抽出する内容語の品詞を茶筌 (IPADIC) の品詞体系に従って次に示す。

使用する品詞

- (1) 名詞 (一般, 固有名詞, サ変接続, 形容動詞語幹)
- (2) 動詞 (自立)

¹ 形態素解析器 ChaSen, Ver. 2.3.3, 奈良先端科学技術大学院大学 松本研究室, <http://chasen.naist.jp/hiki/ChaSen/>.

(3) 形容詞 (自立)

ただし、以下に示す 8 形態素は出現頻度が非常に多く、また他の語句の補助的な機能を持つ語句であることから興味の強弱の判断材料として不向きであるため、ストップワードとする。

ストップワード

する、ある、よる、いる、なる、いう、みる、できる

抽出処理の例を例 1 に示す。

例 1 内容語の抽出処理

入力：共産社民両党に国会を混乱させたことをわびた。

抽出単語：共産、社民、党、国会、混乱、わびる

本節で述べた品詞、条件を持つ内容語について次節以降の処理を行う。

5.2 素性の興味強度の推定

対象となる素性に学習データを利用して興味の強弱を値として与える方法について述べる。学習データはあらかじめ用意しておいた順位情報が付与されている文書、及び順位情報が付与されなかった文書の集合である。文書自体の興味の強弱が内容語に左右されたとしたときに、内容語自身が持つ興味への影響の強弱は、注目している内容語が学習データの順位付き文書において出現する回数及びその順位に関係していると考えられる。

興味に関係する要素

- 順位情報付き文書に出現した文書数
- 順位情報無し文書に出現した文書数
- 順位別出現文書数

入力文書から抽出した素性には、興味の強弱に基づいた値として興味スコアを付与する。興味スコアを付与するために以下の処理を行う。

- (1) 学習データにおける対象素性の各出現回数を数える。出現回数は順位情報付き文書に出現した文書数、順位情報無し文書に出現した文書数、各順位における出現文書数の 3 種類である。
- (2) 対象が学習データ内で一度も出現していない場合、興味スコアの付与は困難であるため興味スコアの付与は行わずに素性の集合から取り除く。
- (3) 各種出現回数から興味スコアを付与する。

本論文では興味を 2 つの尺度で評価する。まず平均アクセス数の形で興味を評価する方法について述べる。

入力文書 D に含まれる素性を w として w が順位 r に出現した文書数 $Rank_DF_r(w)$ 、素性 w が

順位情報付き文書に出現した文書数を $Ranked_DF(w)$, 素性 w が順位情報無し文書に出現した文書数を $UnRanked_DF(w)$ と表す. このとき素性 w について平均アクセス数 $Average_Hit(w)$ を式 (3) によって求める. 順位 r をアクセス数に変換する関数 $Hit(r)$ は式 (2) を用いる. 式 (3) における M は順位付き文書に付与される順位で最も低い順位を示す.

$$Average_Hit(w) = \frac{\sum_{r=1}^M (Hit(r) \cdot Rank_DF_r(w))}{Ranked_DF(w)} \quad (3)$$

次に, 興味を持たれる要素ならば順位付き文書に出現する回数が多いという観点から素性 w が全文書に対して順位付き文書に出現する割合を $Ranked_Ratio(w)$ で示し式 (4) を用いる.

$$Ranked_Ratio(w) = \frac{Ranked_DF(w)}{Ranked_DF(w) + UnRanked_DF(w)} \quad (4)$$

以上 2 つの尺度から素性 w の興味スコア $Interest(w)$ を式 (5) によって算出する.

$$Interest(w) = Average_Hit(w) \cdot Ranked_Ratio(w) \quad (5)$$

それぞれの素性について式 (5) を使って興味スコアを決定する. この時, 興味スコアが算出できなかった素性は除外する. こうして入力文書 D に対して値が決定できた興味スコア列を F で示し文書の興味推定に使用する. 興味スコアの作成例を例 2 に示す.

例 2 興味スコア列の作成

入力 D_0 : 南極観測隊の越冬交代式が 1 日, 昭和基地で開かれた.

抽出単語: {南極, 観測, 越冬, 交代, 昭和, 基地, 開く}

素性	$Ranked_DF$	$UnRanked_DF$	$Rank_DF_r$			$Interest$
			$r = 1$	$r = 2$	$r = 3$	
南極	2	3	1	0	1	0.266
観測	1	3	0	1	0	0.125
越冬	0	3	0	0	0	0
交代	0	0	0	0	0	—
昭和	2	2	0	2	0	0.25
基地	0	0	0	0	0	—
開く	1	9	0	0	1	0.033

興味スコアの計算例 (南極):

$$\begin{aligned} Interest(\text{南極}) &= Average_Hit(\text{南極}) \cdot Ranked_Ratio(\text{南極}) \\ &= \frac{\frac{1}{1} + \frac{1}{3}}{2} \cdot \left(\frac{2}{2+3} \right) = 0.266 \end{aligned}$$

興味スコア列: $F_0 = \{0.266, 0.125, 0, 0.25, 0.033\}$

5.3 文書の興味強度の推定

入力文書 D から作成した興味スコア列 F を利用して興味の強弱を値とした文書興味スコアを文書に付与する方法について述べる. 入力文書に含まれる素性それぞれが, 順位に対して影響を与えるとしたとき, 興味スコアの低い素性は順位を上げる要素であり, 興味スコアの低い素性は順位を下げる要素である. そこで, 入力文書 D に含まれる素性から作成した興味スコア列 F を使用して文書興味スコアを式 (6) によって算出する.

$$Interest_Doc(D) = \frac{1}{N} \sum_{val \in F} val \quad (6)$$

式 (6) において, F は入力文書 D から作成した興味スコア列を示し, N は F の興味スコア列の要素数を示している. 例 2 に示した入力文書 D_0 の興味スコアは例 3 に示すように計算される.

例 3 文書興味スコアの算出

$$\begin{aligned} Interest_Doc(D_0) &= (0.266 + 0.125 + 0 + 0.25 + 0.033) \div 5 \\ &= 0.132 \end{aligned}$$

本研究では入力文書に文書興味スコアを計算し順位を付与するシステムを構築した. 入力文書の順位付けは各入力文書について式 (6) を使用して文書興味スコアを付与し, スコアの高い順に順位を与える. システムは順位の高い順に出力を行う.

6 評価実験

6.1 評価方法

本論文の評価はシステムが付与したランキングの上位文書がより多くの人に興味を持たれているかを確認する事である. 実際の順位が付与された文書を含む文書群を入力とし, システムが付与した順位と実際の順位を比較することで評価を行った. このとき実際に付与されている順位を正解として扱っている.

本論文ではシステムの出力順位と実際の順位が近似しているほど高精度であると考え, 評価式は相関係数を求める Spearman の順位相関係数 (Kendall and Gibbons 1990) を基に式を拡張して用いた. Spearman の順位相関係数 γ を式 (7) に示す. 式 (7) において n は入力順位の総数, d_i は i 番目に入力された 2 つの順位 x_i, y_i の差分を示している.

$$\gamma = 1 - 6 \cdot \frac{\sum_{i=1}^n (d_i)^2}{n \cdot (n^2 - 1)} \quad (7)$$

$$d_i = x_i - y_i \quad (8)$$

本論文における精度はシステムの出力した順位と実際の順位が類似するほど精度が高いとする。

本評価では式 (8) を式 (9) の様に変更して評価式として扱う。式 (9) において m 番目にシステムが出力した順位と正解順位の対を $Rank_Sys_m, Rank_Ans_m$ と示す。

$$d_m = \begin{cases} Rank_Sys_m - Rank_Ans_m, & \text{if exist } Rank_Ans_m \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

$$Rank_Sys_m = m \quad (10)$$

式 (9) を使った場合式 (7) において, n は入力文書の総文書数を示し, i はシステム出力の順位, d_i はシステムの順位と実際の順位との差を示している。ただし順位の差を計算する上で実際の順位とシステムの順位の両方の値を持つ文書のみに対して順位差 d_i を計算を行う。これは正解の順位が正確に得ることが出来ないためである。評価式は以下に示す情報が反映されている。

- 実際の順位とシステムの順位の差
- 興味を持たれなかった記事がシステムの上位に出現する数

評価式において上記の項目が大きい値を持つとき評価値を下げることになる。評価式は -1 から 1 までの値をとり, 評価式の結果が 1 に近づくほどシステム出力の順位と実際の順位が近いことを示す。評価値が 1 に近づくことはシステムが大衆の選択した文書に高い順位を与え, さらにシステムの出力した順位と正解の順位が近いことを示す。また評価の結果が 1 の時, システムの出力した順位と実際の順位に差が全くないことを示している。評価式を使った計算を例 4 に示す。

例 4 様々な正解に対する評価値の計算

入力文書数を 10 文書とし正解が 1 位, 2 位, 3 位の 3 つの場合の計算を示す。空欄は順位が無いことを示している。

システム順位	1	2	3	4	5	6	7	8	9	10
正解順位 1	1		2	3						
正解順位 2	3		1				2			
正解順位 3	1	2	3							
正解順位 4								3	2	1
正解順位 5								1	2	3

各評価値の計算を以下に示す。

正解 1: $1 - 6 \cdot (0 + 1^2 + 1^2) \div (10 \cdot (10^2 - 1)) = 0.99$

正解 2: $1 - 6 \cdot (-2^2 + 2^2 + 5^2) \div (10 \cdot (10^2 - 1)) = 0.81$

正解 3: $1 - 6 \cdot (0 + 0 + 0) \div (10 \cdot (10^2 - 1)) = 1$

正解 4: $1 - 6 \cdot (5^2 + 7^2 + 9^2) \div (10 \cdot (10^2 - 1)) = 0.07$

正解 5: $1 - 6 \cdot (7^2 + 7^2 + 7^2) \div (10 \cdot (10^2 - 1)) = 0.11$

6.2 学習データの収集

本論文ではランキングに付与されている順位情報を用いるために順位情報が付与されている文書を収集した。順位付き文書の収集源として我々はニュースランキングに着目した。順位情報が付与されている文書が新聞記事であるため、形態素解析などの処理が容易であり、また閲覧数を基にランキング付与を行っているという点において大衆の興味への関係性が高いと考えたからである。

本論文では順位付き文書を朝日新聞社の「アクセス Top30」²を対象に収集した。このランキングは Web 上で公開されており、掲載される記事のジャンルは、社会、スポーツ、ビジネス、暮らし、政治、国際、文化、芸能、サイエンスと広範囲の記事が含まれている。収集時期において朝日新聞社が公開していたニュースランキングは、順位を決定する基準として 0 時から 24 時までの 24 時間分の総アクセス数を使用し、アクセス数の多い順番で順位を決定している。また、このニュースランキングでは全ての記事の順位が公開されているのではなく、アクセス数上位 30 位以内の記事が順位と共に公開されている。アクセス数は公開されていない。順位の決定に 24 時間分の総アクセス数を用いているので、ランキングの中にはその時間内に掲載された記事だけではなくアクセス数の収集期間よりも以前に掲載された記事も含まれている。よって、ランキングに複数の日で出現する記事が存在するため、記事には複数の順位情報が付与されることもある。記事の収集については順位付きの記事だけではなく、順位が付かなかった記事についても収集を行い、順位付き文書と順位無し文書の両方を学習データとして用いている。

実際に記事を収集した期間は 2004 年 1 月 1 日から 2004 年 12 月 31 日までの 1 年分について収集を行った。表 1 に収集データの詳細を示す。

6.3 評価対象

評価対象として入力する文書は同一の掲載日付でまとめた記事の集合である。システムは各記事に文書興味スコアを付与し、このスコアによって順位を付与して出力する。この入力文書の掲載日付に対して次の日に掲載されるランキングを参照し、システムの順位と実際の順位の対応を作成しこの 2 つの順位の比較を行う。

評価用データとして収集する文書は学習データと同様に朝日新聞社の「アクセス Top30」とし

表 1 収集した学習データの統計データ

文書の種類	異なり記事数	1 日平均記事数	文数
順位付き	8830	30	68869
順位無し	25587	71.9	238259

² アサヒ・コム アクセス Top30, <http://www.asahi.com/whatsnew/ranking/>.

表 2 収集したテストセットの統計データ

記事の種類	全記事数	1 テストセットの平均記事数	文数	割合
順位付き	720	24	7946	0.246
順位無し	2238	74.6	20471	0.754

た．抽出した記事から順位情報付き文書と順位情報無し文書を掲載日付で分割したテストセットを作成した．収集期間は学習データと異なる 2005 年 3 月の 1 ヶ月を収集の期間とし，テストセットを 30 セット作成した．この時のテストセットに含まれている順位付き文書を正解として扱う．

また，朝日新聞社のニュースランキングは順位付き文書を 30 位まで取得することができるが，ランキングに含まれる文書は同一の掲載日付のみではないため 1 つのテストセットに含まれている順位付き文書数は 30 未満になる場合もあり，順位付き文書数はテストセットによって左右する．作成したテストセットについての詳細なデータを表 2 に示す．

1 つのテストセットごとに評価式 (7) 及び式 (9) を使用して評価を行い全テストセットの平均値をシステム全体の評価値とする．

6.4 システムの評価

実験では入力文書に順位を付け正解順位との比較を行うことを評価している．システムの出カ例を付録に示す．評価値の比較のため次の 2 つのモデルを考える．

- ランダム

入力記事群にランダムに順位を付与する．

- IDF を用いた手法

対象素性 w の興味の強さを $Interest(w)$ の代わりに文書に含まれる語句の特徴性を表す idf と対象素性が順位付き文書に出現する割合 $Ranked_Ratio(w)$ を使用して素性の興味スコアを計算する．計算式を式 (11) に示す．

$$Interest_{idf}(w) = Ranked_Ratio(w) \cdot idf(w) \quad (11)$$

式 (11) において $Ranked_Ratio(w)$ は学習データから求め， $idf(w)$ については別にコーパスを用意し，各形態素の idf を算出している． idf 辞書を作成するために使用したコーパスは日本経済新聞³の 1990 年から 2004 年までの 14 年分である． idf をカウントする際の文書の単位は 1 記事としている． $Interest_{idf}(w)$ を $Interest(w)$ の代わりに使用して興味スコアを算出し，文書興味スコアの推定に用いる．

³ 日本経済新聞全記事データベース 1990 年～2004 年，日本経済新聞社．

システムが出力した順位と実際の順位との評価値を図 2 に示す。図 2 には各評価対象の評価値の平均、最大及び最低の評価値をエラーバーで示している。

図 2 において提案手法はランダムと比べて平均の評価値で 0.17 向上している。IDF と比べて提案手法は平均の評価値で 0.08 向上している。提案手法は平均、最低値、最高値のそれぞれにおいて比較対象より良好な結果を得られることが分かった。

次に実際のランキングの 30 位以内に出現した記事を正解記事として、システムの上位 10 記事分を取り出した場合の正解記事の抽出精度を調査した。表 3 にシステムの出力上位 10 記事に注目した抽出精度を示す。

比較対象として先ほど述べた 2 つに加え SVM⁴ による分類器を追加した。SVM の学習は内容を素性に順位付き文書を正例、順位無し文書を負例として行っている。使用したカーネルは線形カーネルである。SVM の抽出精度については 10 記事分ではなく分類の結果正例として出力された中で正解記事が含まれている割合を示している。

SVM は評価条件が異なるため正確な比較ではないが、提案手法は他に定義した比較して、ランダムに比べて約 26 ポイント、IDF に比べて約 24 ポイント、SVM に比べて約 13 ポイント向上している。

図 2 内容語を素性とした文書の興味推定の評価

⁴ SVM 学習ツール TinySVM, Ver. 0.09, 奈良先端科学技術大学院大学 松本研究室, <http://chasen.org/~taku/software/TinySVM/>.

6.4.1 品詞別に分けた評価

内容語の中でどのような語が興味の強弱の判別に貢献しているか調査するため内容語を品詞別に分けて評価実験を行った. 表 4 に品詞別の評価を示す. 表において使用した品詞は○, 未使用の品詞は空欄となっている.

表 4 より以下の事が分かった.

- 全ての品詞を使った場合の評価値が 0.862 であり名詞のみを使った場合の評価値が 0.858 であることから, 名詞の貢献する割合が大きく名詞は興味判別のために重要である.
- 動詞, 形容詞は文書の興味の判別に対する貢献は小さいが全体の精度を下げるものではない.

影響が最も大きかった名詞をさらに品詞詳細に分けて評価を行った結果を表 5 に示す.

以上より興味を判別する際に重要となる語句は名詞—一般, 名詞—サ変, 動詞, 名詞—固有
名詞, 形容詞の順で重要であることが分かった.

6.4.2 学習データ内の出現回数と精度

次に, 興味を判定する素性である内容語が学習データ内の出現回数によって評価値とどのような関係があるか調査を行った. これは出現回数が少ない素性に付与した興味スコアが不安定であることが評価値にどの程度影響を与えるのか調べるためである. また出現回数の多い素性は文書を興味の強弱で判別することに不向きであるかどうかを確認することができる. 実験は学習データ内に出現した回数で閾値を設定し閾値以上又は閾値以下の出現回数であった内容語は文書の興味の推定の時に扱わないようにして行った. 出現回数が閾値以下の素性を削除して

表 3 興味を持たれる記事の抽出精度

評価対象	平均	最低値	最高値
ランダム	0.26	0	0.6
IDF	0.28	0	0.8
SVM	0.39	—	—
提案手法	0.52	0.2	0.9

表 4 品詞別に分けた評価実験結果

使用品詞			評価値
名詞	動詞	形容詞	
○	○	○	0.862
○	○		0.861
○		○	0.860
○			0.858
	○	○	0.830
	○		0.826
		○	0.709

表 5 名詞を詳細別に分けた評価実験結果

使用品詞	名詞—一般	名詞—固有名詞	名詞—サ変
評価値	0.850	0.808	0.840

実験を行った結果を図3に示す。

図3より学習データにおいて頻度が少なかった素性を削除した結果、特に「出現回数1回以下を削除」から「出現回数5回以下を削除」において精度の低下は見られなかった。この原因は学習データにおいて出現回数が少ない語が評価対象として入力する文書にも出現することが少ない事、また出現回数が少ない素性に付与する興味スコアの信頼性が低い事が挙げられる。出現回数が少ない素性は1回の出現回数の増加で興味スコアが大きく変化することからスコアが不安定であり信頼性を低いとしている。

次に出現回数が閾値以上の素性を削除して実験を行った結果を図4に示す。

図4より学習データにおいて閾値より頻度が多い素性を削除した結果、閾値を下げていくと閾値が1,000までは精度の向上が見られた。出現頻度が高い語を削除しても精度に変化が表れなかったことから、これらは興味を判別する要素として重要ではないと考えられる。

以上の結果から評価値の低下が見られなかった「出現回数5回以下を削除」及び「出現回数1,000回以上を削除」の両方を利用した。その結果、評価値で0.865であり、閾値が無い場合の評価値(表4の最大値0.862)と比較して精度が上昇したことが確認できた。これにより「出現回数5回以下」及び「出現回数1,000回以上」は全体として興味判別に貢献していないと考える。

6.4.3 素性に付与されたスコア

興味スコアの値が付いた素性の観測結果について述べる。素性に付与した興味スコアについ

図3 出現回数が閾値以下の素性を削除した評価値の変動

図 4 出現回数が閾値以上の素性を削除した評価値の変動

て並び変えを行い、スコアの上位と下位からいくつかの内容語を例として例 5 及び例 6 に示す。

例 5 高い興味スコアが付いた内容語

乱造, 落葉, 撲殺, 偏愛, 鉢巻き, 作画, 倦怠, 怪死, 往還, 討伐

例 6 低い興味スコアが付いた内容語

調べ, 調査, 入る, 予定, 述べる, 会社, 関係, 求める, 発表, 東京

興味スコアが高い値の語句の特徴としては低頻度の語が多く、低い値の語句の特徴としては高頻度の語、経済のジャンルに属する語句及びスポーツのジャンルに属する語句が多いという傾向が分かった。これはシステムが出力する順位の結果においてもスポーツ記事全般と経済に関する記事の興味スコアが低いという同様の傾向が得られている。よって今回の順位付けは大きく分けて経済記事、スポーツ記事、その他の記事の 3 種類の特徴を捕らえているようになっていることが考えられる。これは記事の特性上、スポーツ記事では記事間で同様の単語が使われることが多いため、それらの単語が順位付き文書で使用される割合が減少し、この結果興味スコアが低くなったと考える。また、経済の記事は全経済記事に対して順位付き記事に表れる割合が少ない。以上の傾向は大衆の興味をとらえるという点では妥当な結果と考えられる。

また例 7 に示すような動詞は 6.4.1 節で述べたように興味推定の貢献度は少なかった。これは内容語単体では興味を推定するのに情報が不十分であることが考えられる。これは、一部のサ変名詞にも言えることである。例えば例 7 にある「結婚」という語句については「結婚」自体

が興味を持たれるのでは無く、それと共に出現する語句を同時に扱うことで情報に限定性が付与され興味の対象となると考える。例えば「結婚」と「タレント+結婚」では興味を判別するという観点から語句に含まれる情報が違うと考える。さらに、「タレント」の部分は固有名詞まで限定できた方が興味の判別要素としては適切なのかもしれないが、この処理を行うのは容易ではない。固有名詞は数多くあって統計的扱いが困難なためである。

例7 単体では興味の判別要素にならない語句

入る, 述べる, 求める, 発表, 結婚, 調査

複合名詞の問題も存在する。例えば「首相官邸」などの語が「首相」と「官邸」に分けられた状態でしか扱っていないため語句が示す意味が実際とは異なってしまうことが挙げられる。複合名詞になって初めて興味を判断できるとするならば改善が必要である。

以上の考察から次節以降では素性がもつ情報量を増やす方向で拡張を試みる。

7 処理単位の拡張

前節で述べた内容語を素性とした興味判別は記事の内容を特定する情報が少なく、判別する要素として十分ではないことが問題として分かった。この問題に対応するため、処理の単位の拡張を行った。提案する処理単位を以下に示す。

- 1 複合名詞を素性として抽出
- 2 内容語及び複合名詞の組み合わせを素性として抽出

以降では、内容語及び複合名詞の組み合わせを「複合素性」と呼び、興味の推定に使用する。次節より、複合名詞を素性として抽出する方法、複合素性を抽出する方法、拡張した素性を使って文書に興味のスコアを付与する方法について順に述べる。

7.1 複合名詞の抽出

入力されたそれぞれの文書から興味の判別を行う際に重要であると思われる素性を複合名詞として抽出する。複合名詞の研究は盛んに行われており、本論文では専門用語の抽出(中川, 湯本, 森 2003)を参考に複合名詞を作成した。複合名詞の作成対象となるのは名詞, 接尾辞, 接頭詞の連続部分である。

複合名詞作成にあたって以下に示す出現回数をコーパスよりカウントする。

- 2 単語の文中における共起回数をコーパスよりカウントする
- 2 単語の接続回数をコーパスよりカウントする
- 2 単語それぞれに接続する名詞の種類数をコーパスよりカウントする

本研究では複合名詞作成のコーパスとして日本経済新聞 1990 年から 2004 年までの 14 年分

を使用した. 接続した2つの形態素を「 ij 」とした場合, i と j が文中において共起する回数を $Count_{co}(i, j)$, i の後に j が接続する回数を $Count_{bigram}(i, j)$, 前方の語 i の後に接続する名詞の種類数を $Count_{var}(i)$, 後方の語 j の前に接続する名詞の種類数を $Count_{var}(j)$ としたとき以下の条件のいずれかに一致するものを複合名詞とした.

名詞の結合条件

- 条件 1:

$$Count_{bigram}(i, j) \div Count_{co}(i, j) > 0.01 \quad (12)$$

- 条件 2:

$$\sqrt{(Count_{var}(i) + 1)^2 + (Count_{var}(j) + 1)^2} > 100 \quad (13)$$

3 語以上の複合名詞はそれぞれの形態素間について同様の方法を用いて判定を行う. 複合名詞の作成例を例 8 に示す.

例 8 複合名詞の判定

入力を「競売入札価格の決定」としたとき名詞が連続している部分は「競売, 入札」と「入札, 価格」の部分である. この2つについて複合名詞の判定について行う.

対象	条件 1	条件 2	判定
競売, 入札	0.913	368.5	結合
入札, 価格	0.195	1191.8	結合

「競売, 入札」と「入札, 価格」の両方が複合名詞として扱う条件を満たしているため「競売入札価格」を複合名詞として扱う.

以上の手順で複合名詞を作成し, 興味の判別素性として用いる. また複合名詞にならないものは興味判別の素性として扱わない.

7.2 複合素性の抽出

入力されたそれぞれの文書から興味の判別を行う際に重要であると思われる素性を内容語及び複合名詞を組み合わせた複合素性として抽出する.

そのため 7.1 節で述べた複合名詞の処理も行う. 複合名詞とならなかったものは内容語として扱う. 文書の形態素解析を茶釜で行い, 文書が内容語のみによって構成されているとし, まずは接続している名詞に対して複合名詞処理を行い, 複合名詞と内容語の集合を作成する. 入力文書と複合名詞処理後の結果の例を例 9 に示す. 内容語を抽出する条件は 5.1 節と同様である.

例 9 複合名詞と内容語の抽出

入力：インド北部で厳しい冷え込みが続いている。

抽出単語：インド北部，厳しい，冷え込み，続く

例 9 に示した内容語と複合名詞の集合から次に示す条件で組み合わせを作成する。

組み合わせの作成条件

- 内容語及び複合名詞の出現順は考えない。
- 組み合わせの作成は 1 つの文中から探す。
- 2 要素からなる組み合わせで行う。
- 組み合わせを出現文書数でカウントする。
- 学習データ内で出現文書数 5 以下 1,000 以上の組み合わせは扱わない。

出現文書数の閾値は 6.4.2 節の内容語を素性とした評価で最も良好な結果を得られた下限閾値 5 及び上限閾値 1,000 をそのまま適用した。例 9 の内容語及び複合名詞から複合素性を作成した例を例 10 に示す。

例 10 複合素性の作成

{インド北部，厳しい}，{インド北部，冷え込み}，..., {冷え込み，続く}

複合素性は、2 要素からなる素性集合の全部分集合である。本節で説明した複合素性を 1 つの素性として扱い、5.2 節で述べた方法で興味スコアを付与し、これを利用して文書興味スコアの付与を行う。

7.3 拡張した素性を使った文書興味スコアの計算方法

入力文書 D_i に対して内容語、複合名詞、複合素性の 3 種類の素性集合が抽出できる。それぞれに対して式 (3) 及び式 (5) を使って興味スコア列を作成する。値が付与できなかった素性はこの時点で除去される。作成した興味スコア列をそれぞれ、内容語を F_i 、複合名詞を C_i 、複合素性を T_i として、全ての素性を利用するときの文書興味スコアを式 (17) に示す。

$$F_i = \{Val_1, Val_2, \dots, Val_{N_F}\} \quad (14)$$

$$C_i = \{Val_{C_1}, Val_{C_2}, \dots, Val_{N_C}\} \quad (15)$$

$$T_i = \{Val_{T_1}, Val_{T_2}, \dots, Val_{N_T}\} \quad (16)$$

$$Interest_Doc'(D_i) = \frac{\sum_{Val \in F_i, C_i, T_i} Val}{N_F + N_C + N_T} \quad (17)$$

式 (17) において N_F 、 N_C 、 N_T はそれぞれの興味スコア列の要素数を示している。この評価式を使って文書興味スコアを文書に与え、スコアが高い順に文書に順位を付ける。

8 拡張した素性の評価

評価方法は6.1節と同様の方法を用いる．内容語を素性とした場合では6.4.2節で最もよい評価値を得られた下限閾値5回及び上限閾値1,000回の結果を採用している．使用する素性を内容語，複合名詞，複合素性，内容語と複合名詞，内容語と複合素性，内容語と複合名詞と複合素性の6種類について評価を行った．評価結果を表6に示す．表において文書興味スコアの算出に使用した素性は○，未使用の素性は空欄となっている．

表6より拡張した素性である複合名詞，複合素性だけを使用した場合の精度は内容語のみを使用した場合より精度が低下している．しかし内容語の素性と同時に使用する場合において精度は向上する事が分かった．複合名詞のみ及び複合素性のみを使用した2つの評価では値を付与できない記事，つまり値を付与できる素性が1つもない記事が発生した．これは，素性の情報量を増やすことで使用できる素性が減少し評価値及び抽出精度が落ちたと考えられる．複数種類の素性を使用した場合の精度の向上については，内容語を素性とした方法が他の2つを素性とした場合のスパースな部分において補完が行われ，さらに内容語だけでは得られなかった情報が追加されたことになるため精度が向上したと考える．最も評価値が上昇したのは内容語と複合名詞と複合素性の全てを素性として使用する場合であり，評価値で0.867，抽出精度で0.57という結果が得られた．

8.1 システム出力の観察

使用する素性を変化させたときにシステムの出力にどのような変化が起こっているのか確認する．システムが出力した順位の上位10位，上位30位，下位10位，下位30位について実際に順位が付与されている記事（正解記事）が含まれている数について調査を行った．システムの上位及び下位における正解記事数を表7に示す．表7の「全て」は内容語と複合名詞と複合素性を素性として使用することを示す．

表7より拡張した素性を加えるとシステム上位10記事の結果では順位が実際に付与されてい

表6 拡張した素性を使った評価実験

使用品詞			評価値	抽出精度
内容語	複合名詞	複合素性		
○			0.865	0.53
	○		0.842	0.51
		○	0.859	0.55
○	○		0.865	0.54
○		○	0.866	0.57
○	○	○	0.867	0.57

表 7 システム出力に含まれている正解記事数

	全記事数	正解記事数		変化
		内容語	全て	
上位 10 記事	300	160	171	+11
上位 30 記事	900	374	387	+13
下位 10 記事	300	11	10	-1
下位 30 記事	900	76	75	-1

る記事数が 160 記事から 171 記事に増加した。またシステム下位 30 記事の結果では正解 76 記事から 75 記事に減少した。内容語、複合名詞、複合素性の全てを用いた場合システムの出力において上位の変動が 11 件確認でき下位の変動は 1 件であった。上位の変動が下位より大きく表われ、精度の向上は主にシステム上位部分で発生している事が分かった。

システムの出力において正解記事が出現する偏りの傾向について述べる。表 7 に示したシステムの上位 30 位と下位 30 位に含まれる正解数から、順位付きの記事が含まれる傾向は上位 30 記事に 387 記事、下位 30 記事に 75 記事であった。従って上位には下位より約 5 倍ほど正解記事が出現しやすく正解記事がシステムの上位に強く偏る傾向が見られる。

次にシステム出力の下位の記事を興味を持たない記事として抽出することに利用することについて述べる。興味を持たない記事は評価用のテストセット内で順位無し記事であり、テストセットにおける興味を持たない記事の割合は 0.754 である。表 7 より、システム出力の下位 10 記事を抽出した場合、300 記事の中で $289 (= 300 - 11)$ 記事が順位無し記事（ほとんど興味を持たれない記事）である。従って興味を持たれない記事の抽出精度としては $0.97 (= 289 \div 300)$ という精度であった。また下位 30 記事を見た場合においても興味無し記事の抽出精度は $0.92 (= 1 - 76 \div 900)$ である。

システムの下位 30 記事を抽出する場合、900 記事が対象であるため全記事の 2,958 記事に対して約 3 割 $(= 900 \div 2985)$ の記事に対して抽出精度 0.90 を超える精度で興味を持たれない記事の抽出が期待できる。

8.2 順位を大きく誤った記事に対する考察

評価実験の結果、文書興味スコアから付けた順位が正解と離れている文書についてその原因を考察する。観察対象は、システム出力の下位から 10 文書を取り出したときに含まれている順位付き記事である。この記事は興味ありを興味無しと判断した誤りである。

システムが付けた順位の下位から 10 文書に出現した順位付き記事を例 11 に記事タイトルで示す。

例 11 システムの下位に出現した正解記事

- (22) 量的緩和解除, 静かな幕開け 福井総裁は国会答弁 5 時間
 - (26) 佐々木が 2 位, 日本人最高順位タイ スキー W 杯男子回転
 - (26) チーム青森, 常呂中破り決勝に カーリング日本選手権
 - (22) 将棋名人戦, 谷川九段が挑戦権を獲得 羽生三冠を破る
 - (17) 中村が活躍, セルティック V スコットランドサッカー
 - (29) 大久保が同点ゴール スペイン 1 部サッカー
 - (14) 日本のチーム長野は 5 連敗 カーリングの女子世界選手権
 - (16) 朝青龍 16 度目の優勝 決定戦で白鵬下す 大相撲春場所
 - (26) 共済年金の上乗せ給付, 10 年新規加入者から廃止へ
 - (15) 安保理拡大, 日本が 21 カ国案の提出断念 米の支持なし
- *() 内の数字は実際の順位である.

例 11 に示した順位付きの記事は低い順位の記事が占めている事が分かる. これらの記事が下位に出現した原因として未知語処理に関する問題が挙げられる. 学習データに出現しない語句はシステムにおいて除外しているため興味スコアに反映されない. 例えば例 11 に示した「チーム青森, 常呂中破り決勝に カーリング日本選手権」では「カーリング」という語句は抽出できていない. スポーツ記事が占める割合が高いのは, 6.4.3 節で述べた類似した記事が大量に存在することで素性に付与される興味スコアが低下しやすい状況にあるからだと推測する.

次に, システム出力の上位から 10 文書を取り出したときに含まれている順位無し記事を対象にする. その記事は興味無しを興味ありと判断した誤りである. システムの出力で上位から取り出した 10 文書に出現した順位無し記事について記事のタイトルを例 12 に示す.

例 12 システムの上位に出現した不正解記事

登校中の中学生切られる 男が逃走 愛知・美和町
 スーパーにライトバン突っ込み 7 人重軽傷 山形・川西
 東京の自宅マンション放火事件, 中 2 少年を家裁送致
 集団自殺か, 男女 3 人が車内で死亡 弘前・岩木山ろく
 架線にビニールひも付着, 東北新幹線に遅れ
 男女 4 人が集団自殺, 車内に練炭 静岡市の山間部
 東武伊勢崎線普通列車が停車予定駅を通過
 東京・豊島のマンションに男性の遺体 強盗殺人で捜査
 川崎の小 3 男児転落死 マンション 15 階にランドセル
 女子中学生の顔にスプレー 鹿児島・薩摩川内

例 12 の特徴としては事件記事が多く含まれていることが分かる．提案したシステムでは事件ジャンルに属する記事のスコアを高く与える傾向があり，例 12 にあるような「強盗」「自殺」「放火」等の語句が高いスコアを持って記事の中に含まれているため，高い順位が付与される．なお，不正解記事の中には他の正解記事の関連記事や，前日の正解記事の続報である場合もあり，完全な間違いと言い切れない記事も存在した．

8.3 拡張した素性の観察

内容語を使って求めた興味スコアと複合名詞を素性とした興味スコアを比較し，素性を拡張した影響の調査を行った．

内容語及び複合名詞の素性から求めた興味スコアの差を「プロ」で始まる語句を例に表 8 に示す．同様に複合素性を素性とした結果と内容語を使った場合との差を「株価」の語句を含むものを例に表 9 に示す．表では組み合わせた 2 つの内容語又は複合素性を「:」で分割して表記する．表 8，表 9 において興味スコアは複合名詞や複合素性から求めた場合の興味スコアを示し，内容語による興味スコアは複合名詞や複合素性に含まれる内容語から求めた興味スコアである．

複合名詞から算出した興味スコアと複合名詞に含まれる内容語から算出した興味スコアについて相関値を算出した結果，相関値は 0.698 となった．同様に複合素性と内容語による興味スコアの相関値を算出すると 0.551 であった．両者共に相関係数は高いものであるが，複合名詞を素性とした場合の方が高い相関値を持っている．このことから複合素性の興味スコアよりも複合名詞の興味スコアが内容語の興味スコアにより類似していると考ええる．そのため評価値も同程度得られると考えられるが，素性の情報量を増やした事でそれぞれの出現回数が減少することによるスパースの問題だけ評価値が下がったと考察する．反対に複合名詞にすることで興味を判別する素性として良くなった例を挙げる．表 8 にある「プロアイスホッケー」では内容

表 8 複合名詞の興味スコア

複合名詞	興味スコア	内容語による興味スコア
プロデビュー	0.142	0.121
プロ野球巨人	0.121	0.083
プロボクシング	0.091	0.074
プロ野球	0.062	0.103
プロ	0.045	0.045
プロ棋士	0.030	0.073
プロ選手権	0.002	0.064
プロアイスホッケー	0.000	0.045
プロ選手	0.000	0.045

表 9 複合素性の興味スコア

複合素性	興味スコア	内容語による興味スコア
株価：ライブドア	0.159	0.131
可能性：株価	0.112	0.077
大量：株価	0.098	0.087
経営：株価	0.068	0.057
上場：株価	0.063	0.065
株価：株	0.051	0.052
発表：株価	0.042	0.051
株価：推移	0.038	0.042
株価：低迷	0.034	0.071
株価：下落	0.022	0.043
株価	0.022	0.022
大手：株価	0.020	0.055
株式：株価	0.018	0.041
株価：金融	0.014	0.045
株価：終値	0.014	0.032
株価：東証	0.013	0.046
事件：株価	0.010	0.072
株価：情報	0.009	0.070
売却：株価	0.000	0.058

語のみを使用した興味スコアの推定では「プロ」の語句だけで値が付与されている。これを複合名詞として扱うことで「アイスホッケー」を考慮した興味スコアとなり、「プロ」の興味スコア (0.045) と比較して大きく異なる値 (0.000) となった。また複合名詞にすることで素性として良くなった別の例では、一般的な名詞同士の複合名詞「日本＋経済」, 「日本＋代表」のような語句に興味スコアの差が表れるようになった。「日本＋経済」, 「日本＋代表」については内容語を素性とした場合、両者の興味スコアは共に 0.035 であるが、複合名詞の場合では「日本＋経済」のスコアが 0.009, 「日本＋代表」のスコアが 0.03 という結果になっており、後者の興味がはるかに強いことが分かる。

次に複合素性にすることで興味を判別する素性として良くなった例を挙げる。表 9 にある「株価：売却」と「株価：経営」では内容語による興味スコアにほぼ差が存在しないが複合素性では「株価：売却」の興味スコアが 0.009, 「株価：経営」の興味スコアが 0.068 となっている。このように素性の情報量を増やすことで興味スコアに差を得られるようになった。このことが抽出精度が上昇した要因だと考える。

次に残された問題点について述べる。複合素性の問題点として、複合素性を作成すると大量の組み合わせが作成されてしまう。本論文では出現頻度 5 以下を削除しているが、削除された

中には実際には有効なデータも含まれていると考える。しかし判断要素として使えない対が有効なデータよりもはるかに大量に含まれているため、現在の方法では精度を下げる要因にしかない。

システムが付与した興味スコアの高い素性と低い素性の一部を例 13 及び例 14 に示す。

例 13 高い興味スコアとなった複合素性

{駐車場：突っ込む}, {男性：聴く}, {関東：低気圧}, {ダイヤ：乱れ}, {殴る：現行犯逮捕}, {車内：確認}, {NHK：現場}, {メンバー：人気}, {団体：中止}, {県警：車内}, {救急隊員：駆けつける}, {女子生徒：分かる}, {現行犯逮捕：調べる}, {揺れ：最大}, {原因：容疑}, {レギュラー：番組}, {通行人：110 番通報}, {女子：盗む}, {男性：任意}, {駆けつける：女性}

例 14 低い興味スコアとなった複合素性

{ダウ工業株：平均}, {ニューヨーク外国為替市場：円相場}, {東京株式市場：日経平均株価}, {承認：向ける}, {状況：厚生労働省}, {人：救済}, {世界：棒高跳び}, {制裁：再開}, {奪う：優勝}, {農業：求める}, {判断：輸入}, {粉飾：担当}, {保護：批判}, {補助金：決める}, {目標：抑える}

これらの素性に付与された値の妥当性を個別に判断するのは困難であるが、我々は例 13 の素性の方が例 14 よりも興味を持たれるのではないかと感じた。

8.4 学習データによる影響

学習データの量と精度の関係を調査するため、学習量を変化させて評価を行った。評価実験は最も良い結果が出た内容語、複合名詞、複合素性の全てを使用した。6.2 節において収集した全ての学習データに対して、データ量を減少させて評価を行った。学習データを変化させて評価した結果を図 5 に示す。

図 5 より学習データの増加に対して評価値の向上が見られる。学習データの増加に対する評価値の向上は今回用意したデータの 0.3 倍程度（順位付き文書約 2,500 文書、順位無し文書約 7,500 文書）のところで精度向上が緩やかになっている。このデータが描く曲線の様子から、データ量を現在の全データよりも 2 倍、3 倍と増やした場合の評価値は 0.875 程度であると予測する。

次に学習データの違いによってどの程度精度に影響するかを調査するためいくつかの学習データを作成した。6.2 節で収集した学習データを順位付き文書と順位無し文書の比率を変えないように 2 分割し、一方を学習データ A、残りを学習データ B とした。学習データ A 及び B から無作為にそれぞれ半分ずつ取り出して合わせたものを学習データ C とした。従って AB 間のデータの重複はなく、AC 間及び BC 間の重複は 50 % である。この 3 種類の学習データに対して評

図 5 学習データの量を変化させた評価実験

表 10 異なる学習データと評価値の関係

学習データ	A	B	C
評価値	0.859	0.864	0.859

価実験を行った。評価結果を表 10 に示す。

表 10 から学習データを変更させても同程度の精度を得られることが分かった。以上の結果より収集期間に関係なく同じ量の文書を学習データとして用意した場合、同程度の精度を得ることができ、学習データは多いほど良い精度を得ることが期待できる。しかし今回調査した学習量の増加と評価値の増加の 2 つの関係から、データ量を極端に増やしたとしても精度の向上は評価値で 0.875 ほどが上限であると予測する。

以上の学習データの量を変化させた評価実験と、異なる学習データによる評価実験より、約 1 万件以上の記事を用意した場合、収集した期間と関係なくほぼ同等の精度が得られることが期待できる。これは、ある程度のデータ量があれば収集した日付の違いによる影響は小さいことを示している。

9 おわりに

本論文では、順位情報を利用して、文書の興味推定を行うモデルを提案し、大衆の興味が反映されているデータをテストデータとして用いて評価実験を行った。文書の興味推定のための素性として、内容語、複合名詞、複合素性の3種類について実験を行った結果、相関係数を基にした評価値で0.867、システムが付けた順位で上位10記事を出力した場合に、実際のランキングで30位以内の記事が含まれる割合は0.57となった。さらにほぼ興味を持たれない記事を抽出する精度が0.90を超えることが期待できる結果を得た。システムの出力から経済記事が出力下位に非常に多く表われやすい傾向を持つことや実際の順位付き記事が上位に表われやすい傾向から今回付与した興味スコアは大衆の興味を捕らえるのに概ね妥当なものであると確認した。

今回はニュースランキングの文書を中心に実験を行ったが、収集が可能であれば他のランキングについても同様の実験を行いランキングが異なることによる特性の違いなどが得られれば興味の判別に役立つと考えている。また本論文では文書に興味の強弱を表すスコアを付与することのみを行ったが今後の課題として文書の中において興味を持たれる要因の自動抽出を考えている。

参考文献

- 福原知宏, 村山敏泰, 中川裕志, 西田豊明 (2006). “Weblog から社会の関心を探る.” 人工知能学会全国大会論文集 (CD-ROM), **20** (3D2-1).
- Fukuhara, T., Murayama, T., and Nishida, T. (2005). “Analyzing concerns of people using Weblog articles and real world temporal data.” *Proc. of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.
- 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学 (2003). “document stream における burst の発見.” 情報処理学会研究報告. 自然言語処理研究会, **160** (23), pp. 85–92.
- 中川裕志, 湯本紘彰, 森辰則 (2003). “出現頻度と接続頻度に基づく専門用語抽出.” 自然言語処理, **10** (1), pp. 27–45.
- Adamic, L. A. and Huberman, B. A. (2002). “Zipf’s law and the Internet.” *Glottometrics*, **3**, pp. 143–150. RAM-Verlag, <http://www.ram-verlag.de/>.
- Kendall, M. and Gibbons, J. (1990). *Rank Correlation Methods*. Oxford University Press.
- 西原陽子, 砂山渡, 谷内田正彦 (2007). “多くの人の興味をひく研究発表タイトルの作成支援.” 人工知能学会全国大会論文集 (CD-ROM), **21** (2F4-5).
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). “GroupLens: An Open Architecture for Collaborative Filtering of Netnews.” *Proceedings of ACM 1994 Conference*

on Computer Supported Cooperative Work, pp. 175–186.

Shardanand, U. and Maes, P. (1995). “Social information filtering: algorithms for automating “word of mouth”.” *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 210–217.

付録

付表 1～付表 3 にシステム出力の上位から記事を並べた結果を示す。

付表 1: 最も評価値が良い出力結果 (0.93)

(ys:システムの順位, Int:興味のスコア, Ans:正解順位)

Sys	Int	Ans	記事タイトル
1	0.1	11	警察官が逃走車両に拳銃 5 発を発砲 東京・江東区
2	0.09	28	JR 中央線, 四ツ谷駅で信号故障 ダイヤ乱れる
3	0.08	4	警官に偽装, 交通取り締まり 神奈川県警が男を書類送検
4	0.08	24	行方不明の女子中学生を 10 年監禁 米ペンシルベニア州
5	0.08	15	ロレックスなど 1 億円相当盗まれる 埼玉・川越の百貨店
6	0.07		酸素不足アラームに気づくのが遅れ乳児死亡 埼玉・川口
7	0.06	8	タレントの松本竜助さんが緊急入院
8	0.06	2	永田氏「メール仲介者は西澤氏」 懲罰委で公表
9	0.06	18	松井秀がソロ本塁打 大リーグオープン戦
10	0.06	29	「学校現場の混乱回避」民間出身の大阪府立高校長が辞任
11	0.06	1	焦点の名称, あっさり公表 永田氏が懲罰委で
12	0.06	21	辛容疑者の補助工作員, 日本人複数を拉致対象として調査
13	0.05		前阪神の久慈, プレーブスでコーチ修行
14	0.05	20	初出場の八重山商工, 高岡商を破る 選抜高校野球
15	0.05	19	今世紀末, 氷河融解で海面数メートル上昇
16	0.05		合否判定でミス, 1 人を追加合格 岐阜県立斐太高
17	0.05		ぼやの新橋演舞場, 25 日正午に再開
18	0.05		WBC 組が活躍, クルーンは連投くプロ野球オープン戦
19	0.05	26	非正社員の月給, 正社員の 6 割 厚労省調査
20	0.05	7	西澤孝氏の証人喚問, 4 月 4 日で調整 衆院懲罰委
21	0.05	23	F1 日本 GP, 「富士」で 07 年 10 月に 30 年ぶり
22	0.05	25	総合力でロッテ, 雪辱狙うソフト パ・リーグ 25 日開幕
23	0.05	17	政府に困惑広がる 志賀原発の差し止め判決
24	0.05		第 22 回織田作之助賞に松嶋ちえさん 大阪文学振興会
25	0.05		カンボジアで 5 人目の死者 鳥インフルエンザ
26	0.05	22	養殖なのに「岩のり」と表示 公取委が 9 者に排除命令
27	0.05		日航の手抜き検査, 国交相「安全意識が不十分」
28	0.05		次は「日本一」 ソフトバンク王監督がチームに合流
29	0.05		日本航空ジャパンに排除命令 特便割引の表示巡り公取委

30	0.05	13	志賀原発の運転差し止め命じる 金沢地裁判決
31	0.05		松本被告弁護団, 3 月中にも控訴趣意書提出へ
32	0.05		一関学院一岐阜城北, 初出場対決始まる
33	0.05		ベラルーシ当局, 野党支持者を拘束
34	0.05		斎藤隆がマイナーへ ドジャースが通告
35	0.04		新宿のビル賃料巡る訴訟, 高島屋と貸主が「痛み分け」
36	0.04		官房長官「報告受けて検討する」 原発運転差し止め判決で
37	0.04	3	トルシエ氏がイスラム教に改宗 名は「オマル」に
38	0.04	12	お得意さまは日本人 アラスカでオーロラ観光が人気
39	0.04		ヤフーショッピング加盟店の顧客情報 8,223 件流出
40	0.04		防衛庁, 日本無線など 2 社に水増し額の返還請求
41	0.04		日本からカネヒキリなど 9 頭が挑戦 競馬ドバイ国際競走
42	0.04		仏の雇用紛争, パリなどで暴徒 420 人拘束
43	0.04		「非常勤公務員の再任拒否は無効」東京地裁が初判断
44	0.04		ヤフー, 東芝のネット子会社買収へ
45	0.04		国連広報センター所長に幸田チャーミン氏
46	0.04		原田最後のジャンプ, 伊藤杯順延 「あす精いっぱい飛ぶ」
47	0.04	30	北陸電力「ただちに控訴」 志賀原発差し止め判決
48	0.04		全国学力調査, 結果は市区町村単位で公表せず 検討会議
49	0.04		サッカー W 杯主審の選考会終了 31 日に決定
50	0.04		国民投票法案「今国会で」 民主・渡部国対委員長
51	0.04	14	永年勤続職員の旅行券「違法」, 神戸市長に支払い命令
52	0.04		横浜は川角が先発 履正社は魚谷 24 日第 2 試合
53	0.04		王子製紙, 新社長に篠田氏
54	0.04		坂東三津五郎さんら 10 人に芸術院賞
55	0.04		阪急百貨店が出店交渉方針 新博多駅ビル核テナント
56	0.04	9	コンビニの 24 時間営業, 山間地で制限も 長野県が条例
57	0.04		巨人・岡島, 日ハムへ 実松・古城とトレード
58	0.04		エアバス, 中国で工場建設検討 受注の一部を現地生産
59	0.04		人権条例を無期限凍結 鳥取県議会
60	0.04		医療用医薬品にバーコード 取り違え防止, 今夏に導入
61	0.04		韓国, 初の女性首相誕生へ 大統領, 韓元女性相を指名
62	0.04		情報公開ランク, 鳥取が 1 位 議会と警察, 低い傾向
63	0.04		「目標は銅以上. あ, 金です」福原, 世界卓球 V 宣言
64	0.04		教科書の特殊指定廃止に反対声明 出版労連
65	0.04		公務員純減, 気象庁など「重点分野」でもゼロ回答相次ぐ
66	0.04		PSE なし中古家電, 実質販売 OK に 「後で検査」前提
67	0.04	10	「代表選手は定期合宿を」 プロ野球宮本選手会長が提言
68	0.04	27	WBC 韓国代表にボーナス 李鍾範は 2,500 万円
69	0.04		宇都宮市発注工事談合, 41 社に 6 億円課徴金納付命令
70	0.04		米通信機器大手と仏社, 合併向け交渉
71	0.04		NEC と住商, ネット事業で提携へ ビッグローブ独立
72	0.03		日興コーディアル, 東京スター銀行と買収交渉
73	0.03		白鵬, 大関昇進を確実に 北の湖理事長「高成績で安定」

74	0.03	インドネシア, 駐豪大使召還 パプア出身者のビザに抗議
75	0.03	防衛庁長官, 施設庁解体方針を首相に報告
76	0.03	新聞特殊指定の堅持求める意見書, 北海道議会も可決
77	0.03	新聞特殊指定, 官房長官「公取委は国民利益の確保を」
78	0.03	イラク各地でテロ 56 人死亡
79	0.03	マレーシア外相, スー・チー氏と会えず 予定短縮し帰国
80	0.03	ワコール, 14 億 8 千万円の申告漏れ
81	0.03	全頭検査容認求め米農務省を提訴へ 米の牛肉加工会社
82	0.03	あおぞら銀, 4 月から普通銀行に 長期信用銀行が姿消す
83	0.03	ベラルーシに EU が制裁発動へ 首脳会議で決議
84	0.03	武装勢力が衝突, 60 人死亡 ソマリア首都
85	0.03	日本は 3 勝 8 敗で 10 位 カーリング女子世界選手権
86	0.03	キャッツ粉飾決算, 公認会計士に有罪判決 東京地裁
87	0.03	片山は 24 位発進 ザ・プレーヤーズ選手権
88	0.03	インド首相「究極には平和条約を」 印パ問題で意欲
89	0.03	「イングランド代表監督の夢, かなえたい」 カペロ氏
90	0.03	米国, 再考要請 日本企業の対イラン油田投資
91	0.03	中田浩がフル出場 バーゼルは引き分け
92	0.03	織田, 4 位に後退 フィギュア世界選手権
93	0.03	EU, 共通エネルギー政策検討で合意
94	0.03	パート時給平均賃上げ 13. 6 円 68 組合で妥結
95	0.03	05 年産米の第 10 回入札, 6 割以上売れ残り
96	0.03	米 GM, 不動産ローン部門を売却
97	0.02	イラン核問題, 安保理の議長声明採択は来週に
98	0.02	国の借金 800 兆円突破 05 年末, 前年比 62 兆円増
99	0.02	浅越が 2 回戦へ テニスのナスダック 100 オープン
100	0.02	朝青龍, 白鵬ともに 1 敗守る 大相撲 13 日目
101	0.02	東京円, 米利上げ継続観測から円安進行
102	0.02	金融債も利上げ相次ぐ みずほ銀など
103	0.02	東京株が反発 円安好感, 輸出関連買われる
104	0.01	NY ダウ, 終値は 47 ドル安の 1 万 1,270 ドル
105	0.01	日経平均株価, 終値は 71 円高の 1 万 6560 円
106	0	NY 円, 1 ドル = 117 円 81 ~ 91 銭

付表 2: 平均的な評価値の出力結果 (0.86)

(Sys: システムの順位, Int: 興味のスコア, Ans: 正解順位)

Sys	Int	Ans	記事タイトル
1	0.14	9	北陸で降雪強まる見込み 低気圧が北日本に接近
2	0.09		警察官を連れ去った男, 包囲した民家にはおらず 横浜
3	0.09	6	「劇団青い鳥」の 4 人, オートバイにはねられる 名古屋
4	0.08		東京の自宅マンション放火事件, 中 2 少年を家裁送致
5	0.08		川崎の男児転落死, 殺人の疑い 投げ落とされた可能性

6	0.07	18	フジテレビ, ライブドアに 345 億円の損害賠償の催告書
7	0.07		皇后さまら演奏会鑑賞 紀子さまも 1 カ月半ぶり公の場に
8	0.06	26	西澤氏代理人が喚問通知を受理 メール問題
9	0.06	3	長男の起訴で本社社長がコメント発表
10	0.06	2	総研, ホテル側に「指導料返還しない」 耐震強度偽装
11	0.06	21	手配の NTT データ元社員を逮捕 偽造カード事件
12	0.06	7	関西一早稲田実は延長 15 回引き分け, 再試合に
13	0.06		東京のマンション耐震偽装 姉齒元建築士が構造計算
14	0.06		「GyaO」相互リンク, USEN・ライブドア提携第 1 弾
15	0.06	20	黄砂, チベットからも 偏西風で日本に到達か
16	0.06		呼吸器外し 外科部長「別れの時間作ろうと」
17	0.06		東ティモールで暴動多発 治安, 急速に悪化
18	0.05		多比良・東大教授, 論文 4 本を取り下げ
19	0.05		福島, 宮城で震度 3 震源は福島県沖
20	0.05	30	住基ネット操作手引が流出, パスワードも 北海道斜里町
21	0.05	12	新十両に出羽鳳と皇牙 大相撲・番付編成会議
22	0.05	16	USEN との経営統合・合併も視野 ライブドア平松社長
23	0.05		毎日新聞記者, 選抜高校野球の先発投手を漏らす
24	0.05	8	比女性の子の国籍認める 東京地裁, 国籍法は違憲と判断
25	0.05	1	第 2 東京タワー, 2011 年墨田区に 世界最高 610M
26	0.05		「ダ・ヴィンチ・コード」文庫版, 13 日目で百万部突破
27	0.05		中国当局, 中国人映像作家を拘束 国際団体が批判
28	0.05	11	八重山商工, 猛追及ばず 横浜が 7 - 6 で逃げ切る
29	0.05	5	京大大学院教授を懲戒解雇, 業者から 4500 万円受領
30	0.05		糸山氏, 所有 JAL 株の半数売却 個人筆頭株主
31	0.05		タイ首相辞任求め集会 バンコク繁華街の商業施設閉鎖
32	0.05		バスに爆弾, 17 人負傷 比南部ミンダナオ島
33	0.05		アフリカなどで皆既日食を観測
34	0.05		ドコモとローソンが資本・業務提携
35	0.05		オリエンタルランド, 3 億円の所得隠しの指摘
36	0.05		米同時テロ, ホワイトハウスへも計画 共謀罪の被告供述
37	0.04		イチロー, 3 試合ぶりオープン戦で 5 打数 1 安打
38	0.04	24	NTT データ元社員, カード偽造の疑い 被害 3 千万円
39	0.04		全日空が再び上方修正 3 月期決算
40	0.04		猪木さんの肖像権を独占取得 新日本プロレスの親会社
41	0.04		朝鮮総連施設の税, 90 自治体が減免 総務省が注意喚起
42	0.04		日中友好 7 団体, 30 日訪中 胡主席と会談へ
43	0.04		ワインバーガー米元国防長官死去
44	0.04		ジーコ監督, エクアドル戦へ「しっかりとした試合を」
45	0.04		尊厳死の問題「法制化加速を」 超党派議連の幹事長
46	0.04		高 1 用に「発展的な学習の内容」初登場 教科書検定
47	0.04		ライブドア, 西京銀とのネット銀行設立を断念
48	0.04		アーセナル先勝, バルサは分ける サッカー欧州 CL
49	0.04		韓国・現代自に不正資金疑惑 本社搜索, 系列社長も逮捕

50	0.04		第 16 回南方熊楠賞に文化人類学者の岩田慶治さん
51	0.04		簡保団体割引で自治会に名簿要求 千葉市の郵便局
52	0.04		ロッテ, 2 連勝で勝率を 5 割に <29 日のパ・リーグ>
53	0.04	25	強制連行訴訟, 元炭鉱労働者の訴え退ける 福岡地裁判決
54	0.04		広瀬民放連新会長「持ち株会社, 一案」
55	0.04		小泉首相, GW の外遊はガーナなど 3 カ国訪問
56	0.04		在留資格を不正取得の疑い, 中国籍容疑者逮捕 警視庁
57	0.04	27	岩国基地・滑走路移設, 談合情報通りの業者が応札
58	0.04	10	都内の眼科医, 医業停止中も診察し脱税 容疑 1 億 6 千万円
59	0.04		関西一早稲田実戦始まる 選抜高校野球第 3 試合
60	0.04		世界最多認定, こいのぼり 5,400 匹舞う 群馬でまつり
61	0.04		サッカー W 杯公式球, タイで生産佳境に
62	0.04		六ヶ所村の核燃再処理工場, 31 日にも試運転
63	0.04	17	ソニー, 韓国に液晶パネルの新工場 サムスンと協業維持
64	0.04		宮里, ウィーと同組に クラフト・ナビスコ選手権
65	0.04		イスラエル, カディマが第 1 党 連立交渉, 難航の予測
66	0.04		認知症グループホーム, スプリングラー設置義務化盛る
67	0.03		国家公務員採用「2~4 割抑制」 10 年度まで政府方針
68	0.03		ガ大阪, 大宮下し 4 位浮上 <J1 第 4 節>
69	0.03		リベリア前大統領, ナイジェリアで拘束
70	0.03		イラン, ウラン濃縮拡大 分離器の試験運転準備
71	0.03	19	雇用デモ騒然, 仏首相さらにピンチ 大統領が政治決断も
72	0.03		所得税も脱税容疑 千葉県の人材派遣会社
73	0.03		地方交付税「6 兆円削減可能」 竹中総務相が試算
74	0.03		智弁和歌山一岐阜城北, 試合始まる 高校野球第 1 試合
75	0.03	4	中国が反論, 死刑囚の臓器提供「ごく一部」
76	0.03		「有識者会議」, 国有財産全体の活用策検討委に発展へ
77	0.03		パレスチナ自治評議会がハマス新内閣を信任
78	0.03		北大津一日本文理戦が始まる 選抜高校野球
79	0.03		国交省 OB らを任意で聴取, 水門談合で公取委
80	0.03		稲田, 5 月に米で暫定王座決定戦 WBC ライト級
81	0.03		民主の公務員制度改革, 人件費「3 年で 2 割減」
82	0.03		福島県歯科医師会元会長も有罪 「日歯連」で東京地裁判決
83	0.03		JP モルガン, 過怠金 2,500 万円 作為的な相場で東証
84	0.03		給付見直し案を提示 年金一元化で政府・与党協議会
85	0.03		三菱ふそう, 小型トラックを再リコール
86	0.03		米産牛肉, 輸入再開へ日本側理解 専門家会合
87	0.03		米議会, 新たな対中制裁法案
88	0.03		中小企業の経常利益と売上高 従業員規模で 2 極化
89	0.03		グアム移転「5 割負担が上限」 関係 5 閣僚が一致
90	0.03		国際競争力確保へ東大, 早大と提携 鉱物資源機構
91	0.03	15	安保理拡大, 日本が 21 カ国案の提出断念 米の支持なし
92	0.02		米利上げ, 4. 75 %に 04 年以来 15 回連続
93	0.02		連覇狙う千葉, 広島に競り勝つ <サッカー・ナビスコ杯>

94	0.02	預金の定期金利, 全種別で平均年利が上昇
95	0.02	日経平均, 午前終値は 131 円高 今年の最高値を更新
96	0.01	NY 原油, 2 カ月ぶり高騰 終値は 66 ドル台
97	0.01	日経平均株価, 終値は 248 円高の 1 万 6,938 円
98	0.01	NY 円, 1 ドル = 117 円 89~99 銭
99	0	NY ダウ, 終値は 95 ドル安の 1 万 1,154 ドル

付表 3: 最も評価値が悪い出力結果 (0.73)

(Sys: システムの順位, Int: 興味のスコア, Ans: 正解順位)

Sys	Int	Ans	記事タイトル
1	0.1	15	車内から男女 6 人の遺体, 集団自殺か 埼玉・秩父
2	0.09		集団自殺か, 男女 3 人が車内で死亡 弘前・岩木山ろく
3	0.08	9	放火容疑の中 2 長男「けがをしないよう逃げた」
4	0.08		スーパーにライトバン突っ込み 7 人重軽傷 山形・川西
5	0.08		幼稚園バスと路線バスが衝突, 園児 12 人軽いけが 大阪
6	0.08	22	発砲は誤射, 警官を業過致傷容疑で書類送検 埼玉県警
7	0.08		ロボットの「ホテルマン」登場 千葉・浦安
8	0.07		同僚にわいせつ行為, 警部補に停職 6 カ月 青森県警
9	0.07		浅沼建築士, 数値操作認める 北海道の聴取に
10	0.07		「鄭容疑者の殺意明確」2 園児殺害の罪で起訴 長浜事件
11	0.07	12	インスリン大量注射で殺人未遂容疑, 中国人の妻逮捕
12	0.07		生徒ら 22 人を人質, 元教員が高校立てこもり 仏西部
13	0.07		安藤美姫が中京大中京高を卒業 五輪で 10 日遅れの証書
14	0.07		相談所が一時, 長男を保護 世田谷のマンション放火
15	0.07	24	「首絞め遊び」米で問題に 「死者 50 人以上」の指摘
16	0.06		「経験豊富な人を」加藤, 山崎両氏が「福田待望論」
17	0.06		「痴漢」で略式命令, 会社員に無罪判決 地裁八王子支部
18	0.06		民主, 26 歳女性県議擁立へ 衆院千葉 7 区補選
19	0.06		バンタム級王者, バクティン容疑者を解雇 協栄ジム
20	0.06		「強度不足で補強」フジタなど検討 静岡のごみ施設
21	0.06		平城京の羅城, 簡略版か 十条大路があったのは確実
22	0.06		東急建設から 1,600 万円恐喝容疑, 月刊紙代表を逮捕
23	0.06		ノート並みの「ウルトラモバイル」 マイクロソフト発表
24	0.06		帝王切開ミス, 県立病院の医師を起訴 福島地検
25	0.06	19	「SED, 絶対にやめない」 撤退観測巡りキャノン社長
26	0.06	14	格安液晶テレビ, ベスト電器も販売
27	0.06		ハウスの中で新茶つみ 三重・四日市
28	0.05	7	自筆原稿流出, 村上春樹さん「一種の盗品」と問題提起
29	0.05	8	過払い利息の返還求め一斉提訴, 「アイフル」相手どり
30	0.05		IT 生かし中国語教材 早稲田大と富士通が共同開発へ
31	0.05		「理解深めたい」皇太子さま, メキシコ訪問前に会見
32	0.05	23	西村真悟衆院議員の辞職勧告決議案を提出 自民・公明

33	0.05		開削 400 年, 春の「保津川下り」始まる 京都
34	0.05		NHK 会長, 国際放送への CM 導入を否定 衆院総務委
35	0.05	6	永田氏念頭に「政治家は腹切る覚悟いる」 民主・渡部氏
36	0.05		乳幼児用機内食に米産牛 JAL ウェイズ, 自粛徹底せず
37	0.05	13	「右翼来るので会場貸せぬ」東京商工会議所, 9 条の会に
38	0.05		横浜市幹部, 前港北区長のパーティー案内状を職員に送付
39	0.05	3	中 2 の「妹」奮闘, 3 勝 2 敗で「姉」と対戦 カーリング
40	0.05	1	「逮捕されなければ対処できた」 公判でウィニー開発者
41	0.05		偽装マンション建て替え, 初の合意 川崎・溝の口
42	0.05	2	イラク陸自, 帰国後 3 人自殺 防衛庁「原因特定できぬ」
43	0.05	18	知られざる資料, 「徳川文庫」で大量に確認
44	0.04		高松食肉センター訴訟, 住民側が逆転敗訴 最高裁判決
45	0.04		仏で捨て猫急増 鳥インフルの感染きっかけに
46	0.04		伊藤元長官を地検に告発 収支報告書の寄付不記載巡り
47	0.04		強制連行, 賠償認めず 共同不法行為は認定 長野地裁
48	0.04		よみがえる「源氏物語絵巻」 東京文化財研究所など復元
49	0.04		死刑囚の面会・手紙を拡充「権利の制限を軽減」 閣議決定
50	0.04		バスが地雷踏み, 26 人死亡 パキスタン南西部
51	0.04		カラフルな色使いが特徴 パラリンピックのメダル公開
52	0.04	5	浅田真央は 2 位, 連覇ならず フィギュア世界 Jr.
53	0.04		国交省職員に無罪判決 用地買収巡る詐欺事件で大津地裁
54	0.04		フーリガン対策, 英が捜査チームを独 W 杯に派遣へ
55	0.04		診療報酬明細書の訂正不可, 市民の請求棄却 最高裁判決
56	0.04		「メール問題調査, 今月いっぱいかかる」鳩山氏
57	0.04		三井物産, 245 億円でロンドン金融街の大規模ビル買収
58	0.04		ガス田の中国提案, 尖閣諸島周囲の日本領海含まず
59	0.04		金融商取法案を閣議決定 罰則強化し投資家保護
60	0.04		金利変動, パンプに記載せず あいおい損保, 2 万件販売
61	0.04		臓器移植目的の海外渡航, 腎臓 151 人・肝臓 199 人
62	0.04		関東つくば銀, 茨城銀と合併見送り のれん代償却困難
63	0.04	25	汚泥談合, 落札率は 9 割超に 応じぬ会社に勧誘工作
64	0.04		高校生の就職内定率, 83.5% 1 月末現在
65	0.04		カネミ油症被害で与党チーム 救済法の今国会提出目指す
66	0.03		米港湾管理進出, UAE 企業が撤退 米議会の猛反発で
67	0.03		イオン, 銀行業参入へ 個人向けに特化, 住宅ローンも
68	0.03	30	日立は前年以下, 三菱ほぼ満額 春闘一時金交渉
69	0.03		公取委, 11 社を集中調査 汚泥談合
70	0.03	4	ドミニカが 2 次リーグへ, 練習試合で日本敗れる WBC
71	0.03		米の改正愛国法が発効 ブッシュ大統領が署名
72	0.03		虐待事件のアブグレイブ刑務所, 米軍が閉鎖へ
73	0.03		「日銀首脳陣に敬意」量的緩和終結で NY 連銀総裁
74	0.03	17	国連分担率「常任理事国に下限を」 中口意識し日本提案
75	0.03		国民年金の未納対策盛る 社保庁改革関連法案を閣議決定
76	0.03	16	量的緩和解除, 静かな幕開け 福井総裁は国会答弁 5 時間

77	0.03	26	佐々木が2位, 日本人最高順位タイ スキー W 杯男子回転
78	0.03		「イランはテロの中央銀行」 米国務長官が批判
79	0.02		綱とり狙う栃東, 初日は新小結・露鵬 大相撲春場所
80	0.02		イラク議会の開会, 1 週間延期 ポスト巡る協議難航
81	0.02		個人向け国債, 金利 1 % 超す 量的緩和解除の効果か
82	0.02		三洋電機を産業再生法の対象に認定 経産省
83	0.01		日経平均終値, 78 円高の 1 万 6115 円
84	0		NY ダウ, 33 ドル安の 1 万 0972 ドル
85	0		東京円, 1 ドル = 118 円 39 ~ 41 銭 (午後 3 時)

略歴

沢井 康孝: 2008 年長岡技術科学大学大学院工学研究科修士課程電気電子情報工学専攻修了. 在学中はテキストマイニングの研究に従事. 言語処理学会学生会員. sawai@nlp.nagaokaut.ac.jp

山本 和英: 1996 年豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了. 博士 (工学). 1996 年 ~ 2005 年 (株) 国際電気通信基礎技術研究所 (ATR) 研究員 (2002 年 ~ 2005 年客員研究員). 1998 年中国科学院自動化研究所国外訪問学者. 2002 年より長岡技術科学大学電気系, 現在准教授. 言語表現加工技術 (要約, 換言, 翻訳), 主観表現処理 (評判, 意見, 感情) などに興味がある. 言語処理学会, 人工知能学会, 情報処理学会, 各会員. yamamoto@fw.ipsj.or.jp

(2007 年 9 月 13 日 受付)

(2007 年 11 月 20 日 再受付)

(2007 年 12 月 8 日 採録)