様式 C-19

科学研究費補助金研究成果報告書

平成２１年　６月２４日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：１９７００１５４
研究課題名（和文）　風評被害対策に向けた情報の重要度を決める要因の抽出・分析と重要度の自動推定
研究課題名（英文）　Analysis of Factors that Determine the Degree of Importance of Information and Automatic Estimation of the Degree for Preventing Harmful Rumor
研究代表者
村田　真樹（MURATA MASAKI）
独立行政法人情報通信研究機構・知識創成コミュニケーション研究センター言語基盤グループ・主任研究員
研究者番号：５０３５８８８４

研究成果の概要：近年のインターネット時代では、日々の生活で多数の人々がウェブ上の重要度の情報を探そうとしており、重要な情報を自動的に取り出し表示する方法を切望している。そこで、本研究では、各情報の重要度を推定する方法に関する研究を行った。新聞社や一般の多くの人が重要と考える情報の重要度を約 90%程度で推測する方法を構築した。個々の個人の興味情報が、その個人がどういう情報を重要と考えるかと相関があることも確認した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	2,100,000	0	2,100,000
２００８年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,300,000	360,000	3,660,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：情報の重要度、新聞データ、被験者、教師有り機械学習

１．研究開始当初の背景

(1)情報検索の分野では、ユーザの入力したキーワードと関連する情報を取り出して提示するものがある。本研究はそれに比べると情報自体の重要度も考慮するものとなっている。

(2)Google の検索結果ではユーザの入力したキーワードの他に、web ページのリンクの情報に基づいて多くのページから参照されるページをよりよいものとする検索結果の序列化をしている。リンク情報に基づいてどの情報が重要かを調べているものであり、文章の内容を利用しているものではない。本研究では、自然言語処理の技術も利用して文章の内容や言語表現（言葉の言い回し）なども、情報の重要度を調べる手がかりとして利用する。

(3)自然言語処理の分野では、新聞記事の情報と株価の関係を調べる先行研究がある。しかし、先行研究では個々の企業の株価との関係を調べることに着目しており、例えば小さい会社の情報なら情報の重要度は小さく、大きい会社の情報なら情報の重要度は大きいな

どといったことはしていない。この点で本研究と異なる。本研究では情報の内容や情報に記載されている表現も利用して情報の重要度を分析するものであり、大きい会社名を含む情報であるかも含めて情報の重要度を分析するものである。

(4)本研究では、会社や株価に研究の対象をしぼるものではなく、新聞記事に現れるすべての現象、例えば、株価とは関係のない、災害や交通事故や外交問題などの幅広い分野において、情報の重要度に関わる要因を抽出、分析を行う。

２．研究の目的

(1)本研究では情報の重要度を決める要因を明らかにし、その知見に基づいて情報の重要度を自動推定するにはどのようにすればよいかを明らかにする。自然言語処理に基づく情報検索、情報抽出技術を利用してこの問題を明らかにする。例えば、会社の倒産のニュースがあったとして、売上規模が大きな会社の倒産のニュースの方が、売上規模の小さい会社の倒産のニュースよりも、社会的影響が大きいために重要な情報である、と一般的には考えられる。一方で、小さな会社の倒産のニュースであっても、その会社の利害関係者にとっては重大なニュースである。このように同じ倒産に関する情報でも情報の受け手によって重要度が異なる。このような重要度を決める要因を明らかにすることが本研究の目的である。

(2)情報の重要度には､(a)多くの人にとって重要とされる一般的な情報の重要度と､(b)個々の分野または個々の人にとって重要とされる個別の情報の重要度の二種類がある。本研究ではその両方を扱う。

３．研究の方法

(1)一般的な情報の重要度の研究を行う。新聞の一面に書かれている記事は他の面の記事よりも重要であり、また長い記事は短い記事よりも重要であると思われる。それを手がかりに情報の重要度を決める要因を明らかにする。また、どちらの記事の方が重要かを問う被験者実験を行い、それを手がかりとした分析も行う。分析には種々のデータマイニング手法を利用する。例えば、新聞の構造、被験者実験に基づいて、重要な記事の事例とそうでない記事の事例に分類整理し、それらを弁別する教師あり機械学習を行い、それらの事例にどのような違いがあるかを明らかにする。

(2)個別の情報の重要度（個々の分野や人にとっての情報の重要度）の研究を行う。この個別性の研究では、先に収集した重要な記事の事例とそうでない記事の事例のデータを、分野や人の嗜好に基づいて分類してから機械学手法などのデータマイニング手法を利用することで分析を行う。これにより個々の分野での情報の重要度を決める要因を明らかにする。

４．研究成果

(1)一般的な情報の重要度に関する研究を行った。新聞の一面に書かれている記事は他の面の記事よりも重要であり、長い記事は短い記事よりも重要であると思われる。それを手がかりとした新聞データを利用した教師あり機械学習の実験を行った。新聞データを学習データとして利用した場合、入力された二つの新聞の記事のうち、いずれが１面の記事であるかを９割以上の精度で推定できた。これにより新聞の面情報を基準とした重要度、すなわち、新聞社が考える情報の重要度は計算機で比較的容易に学習できることがわかった。また、どちらの記事の方が重要かを問う被験者実験を行った。300 人以上の被験者により 560 組のデータ（以降被験者データと呼ぶ）を作成した。それを手がかりとした実験も行った。被験者データを学習データとして、2 つの入力記事のうち、被験者の 7 割以上の人が一致して重要と考える方の記事を機械学習により推定したところ約９割の精度で推定できた。このことから多くの人が重要と考える情報は高精度に特定できることがわかった。また、テキストマイニング技術や教師あり機械学習の学習過程で得られるパラメータの情報から新聞記事中のどのような単語が記事の重要度に寄与しているかを調べた。その結果、「年金」「殺人」「事件」「政府」「事故」といった単語が重要度の大きいものとわかった。これらの単語が示すものは情報の重要度に大きな寄与をしているものと思われる。この知見は今後の重要度推定システムの構築に役立つものである。情報の重要度の推定処理の高度化を目指して受身文を能動文に変換する研究も行った。

(2)ユーザ個人が考える情報の重要度に関する研究を行った。ユーザ同士の判断の一致度を知るために、記事ペアにおいてどちらの記事が重要であるかのユーザによる判定について Kappa 値を計算した。Kappa 値は 0.08 という非常に低い一致度の値が得られた。このことからどういう情報を重要と考えるかは

東証	0.634	反発	0.592
中間	0.634	利益	0.592
代表	0.630	委員	0.589
民主	0.621	ぶり	0.588
海域	0.619	野球	0.588
官民	0.618	国連	0.588
落札	0.614	把握	0.587
投資	0.613	議員	0.587
決算	0.612	ジャパン	0.586
工事	0.604	監視	0.585
トヨタ	0.603	金利	0.585
協議	0.601	株式	0.582
主導	0.598	会議	0.580
工場	0.597	廃止	0.580
商業	0.594	法案	0.577

表１　男性の重要単語

訴訟	0.658	支援	0.594
採択	0.627	大阪	0.594
消費	0.626	提供	0.592
不明	0.622	高校	0.587
時代	0.609	９月	0.585
対象	0.606	売却	0.584
義務	0.606	大学	0.580
解明	0.604	最高裁	0.580
さん	0.601	幕開け	0.580
めぐみ	0.600	和歌山	0.580
最終	0.600	予定	0.579
皇室	0.599	出産	0.579
建物	0.599	紀子	0.579
世論	0.599	懐妊	0.579
吸水	0.595	さま	0.579

表２　女性の重要単語

中部地方		関西地方	
工場	0.176	選挙	0.189
対応	0.169	中国	0.188
交渉	0.166	強化	0.180
設備	0.165	東証	0.179
賛成	0.162	友好	0.178
住宅	0.159	株式	0.168
政治	0.159	障害	0.167
トヨタ	0.157	新株	0.166
取得	0.154	資金	0.163
首相	0.153	投資	0.158
共同	0.153	訪中	0.155
政府	0.153	財界	0.155
過半数	0.152	被告	0.155
工事	0.152	訴訟	0.154
ホンダ	0.152	発行	0.153

表３　中部地方と関西地方の重要単語

人によって異なることがわかった。教師有り機械学習法を用いた実験により、個々のユーザが二つの記事のうちどちらが重要であると判断するかを 65%前後の精度で予測できることがわかった。アンケートにおいて答えてもらったユーザ個人の興味情報と、教師有り機械学習により得られた各個人が重要と考える事柄の一致具合を検証した。興味情報が機械学習で重要とされた上位 500 個の単語の方と有意に重なりが多かった被験者は 53 人で、下位 500 個の単語の方が重なりが多かった被験者は 2 人であった。53 人と 2 人は検定で有意差があるため、ユーザ個人の興味情報が、そのユーザの重要な記事の判断と相関があることがわかった。教師あり機械学習の学習過程で得られるパラメータの情報から、具体的に男性、女性がどのようなことを重要と考えているかを分析した。男性、女性がそれぞれ重要と考えている単語を表１、表２に示す。表の数字はその単語の重要度を示す。男性は「トヨタ」「野球」を女性は「出産」「懐妊」という事柄を重要と考えていることがわかった。また、中部地方と関西地方の人が考える重要な事柄も調査した。その結果を表３に示す。中部地方の人は、トヨタ、ホンダな

関東地方		九州地方	
戦後	0.168	安部	0.194
原点	0.164	地検	0.178
認識	0.162	拡大	0.175
教育	0.161	輸出	0.174
返還	0.160	決算	0.174
拡大	0.158	提携	0.171
政府	0.153	幕開け	0.167
対策	0.152	貨物	0.166
安保理	0.148	ホテル	0.165
利益	0.148	一部	0.163
行方	0.148	大学	0.161
不明	0.147	緊急	0.161
方向	0.147	国連	0.160
入学	0.147	断念	0.160
申告	0.147	検査	0.160

表４　関東地方と九州地方の重要単語

どの地域性を反映した単語を重要としている。近畿地方の人は、「株式」「資金」などの金銭的なものが重要であると考えていることがわかった。表４に中部地方と関西地方の人が考える重要な事柄も調査した結果を示す。関東地方では,「戦後」「教育」「政府」という一般的に重要と思われる事柄が重要とされている。九州地方では，安倍元首相の地元の下関に近いことから「安倍」が重要とされていると思われる。これらの単語が示すものは情報の重要度に大きな寄与をしているものと思われる。この知見は今後の重要度推定システムの構築に役立つものである。情報の重要度の推定処理の応用および発展を目指して特許文書中で特に重要な箇所である請求項とその実施例の比較と対応付けの研究も行った。

(3)先行研究では、情報の重要度を扱う研究は、リンク情報に基づいてどの情報が重要かを調べているものであり、文章の内容を利用しているものではない。それに比べて本研究は、文章の内容のみから情報の重要性を推定するものであり、応用範囲が大きい。また、本研究では、新聞社や一般の多くの人が重要と考える情報の重要度を約 90%程度で推測することができ、性能も高い手法である。また、具体的に文章の内容を利用して分析したため、どういう内容のものが重要かの分析も行うことができた。

(4)今後の展望としては、本研究で構築した情報の重要度を推定する技術を種々の応用先に用いることが考えられる。例えば、Web の文書を具体的に本技術で重要な順に並べ替えることが考えられる。また、風評被害の原因になりそうな Web 文書で本技術を利用してみることも考えられる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 村田真樹, 白土保, 金丸敏幸, 井佐原均, System Displaying Differences between Claims and Matching of Claims with Corresponding Parts in Embodiments, Journal of Information, 査読有, 11 巻, 2008, p.407-425.
② 村田真樹, 金丸敏幸, 白土保, 井佐原均, 入力文の格助詞ごとに学習データを分割した機械学習による受身文の能動文への変換における格助詞の変換, システム制御情報学会論文誌, 査読有, 21 巻, 2008, p.165-175.

〔学会発表〕（計３件）

① 村田真樹, ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析, 言語処理学会第１５回年次大会, 2009 年 3 月 4 日, 鳥取大学.
② 村田真樹, Analysis of the Degree of Importance of Information Using Newspapers and Questionnaires, 国際会議 IEEE NLPKE-08, 2008 年 10 月 20 日, 北京・首都師範大学.
③ 村田真樹, 情報の重要度を決める要因の抽出・分析と重要度の自動推定, 言語処理学会第１４回年次大会, 2008 年 3 月 20 日, 東京大学駒場キャンパス.

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計１件）

名称：情報の重要度推定システム及び方法及びプログラム
発明者：村田真樹
権利者：独立行政法人情報通信研究機構
種類：特許権
番号：特願 2008-134888
出願年月日：2008 年 5 月 23 日
国内外の別：国内

○取得状況（計０件）

〔その他〕
特になし。

６．研究組織
(1)研究代表者
村田　真樹（MURATA MASAKI）
独立行政法人情報通信研究機構・知識創成コミュニケーション研究センター言語基盤グループ・主任研究員
研究者番号：５０３５８８８４