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

村田 真樹 * 西村 涼 * 金丸 敏幸 ** 土井 晃一 *** 鳥澤 健太郎 *
* 独立行政法人 情報通信研究機構 ({murata,nishimura,torisawa} @ nict.go.jp)
** 京都大学 (kanamaru@hi.h.kyoto-u.ac.jp) ***(株)PSC (doy @ pharmasecurity.jp)

1 はじめに

本研究の大目標は，情報の重要度を決める要因を明らかにし，その知見に基づき情報の重要度を自動推定するシステムを構築することである．情報の重要度を推定する技術は，記事のランキング [1, 2, 3, 4] や，重要な情報の自動収集など，種々の場面で役立つ重要なものである．

われわれは，過去に新聞を利用した情報の重要度の研究を行った [5, 6]．例えば，新聞の 1 面は他の面よりも情報の重要度が高いと考えられるので，記事ペアのうち，どちらが 1 面であるかを特定する研究を行った．さらに，文献 [5, 6] では，新聞データに頼った研究だけでなく，被験者実験を行い，どのような情報が重要であるかを調べるアンケート調査を行った．これは，新聞データのみによる研究では，新聞社の意図や主義主張が偏り，一般の人が重要と考える情報の分析にならない可能性があるためである．一般の人が重要と考える情報の分析につながるようにアンケート調査に基づく情報の重要度の研究を行った．

本研究では，上記研究をさらに発展させたものである．文献 [5, 6] では，一般の人が重要と考える情報はどういうものであるかを分析するものであったが，どのような情報を重要と考えるかは個々人によって異なるものである．本研究では，ユーザ個人ごとに異なる情報の重要度について特に焦点をあてて分析した研究である．ユーザごとの興味をアンケートにより抽出しその結果を利用してユーザごとに異なる情報の重要度について調査を行った．

2 一般的な情報の重要度の推定

まず，被験者実験に基づくアンケートデータを利用して，一般的な情報の重要度について分析を行った．アンケートは 2007 年 11 月に実施し，309 人の被験者を対象に，56 個の 5 組の新聞記事を与えてその 5 組を自分にとって重要な順に並べかえてもらった．56 個の新聞記事の内訳は，異なる 5 個の日の新聞 1 面トップ記事 (毎日新聞 15 個，読売新聞 15 個，日経新聞 8 個) が計 38 個，1 面トップ記事を含む同じ日の 1 面内の 5 記事 (各社 2 個ずつ) が計 6 個，同じ日の 1 面トップ記事と 4 個のランダムに取り出した 1 面以外の記事 (各社 2 個ずつ) が計 6 個，同じ日の毎日新聞の 1 面トップ記事，次の記事，読売新聞の 1 面トップ記事，次の記事，日経新聞の 1 面トップ記事 (この 5 記事の記事内容が重複しない日を選択) が計 6 個である．5 組の並べ替えのデータから，20 個のどちらが重要とされたかの情報を含む記事ペア (5 組から 2 個の組み合わせを取ることで 10 種類でき，さらにペアの二つの入力順の違いにより 2 倍のデータができるため，5 組のデータからは 20 個のペアが作成される) を生成することで，56 個のデータから，計 1,120 個の記事ペアを生成した．この記事ペアを実験に用いた．アンケートでは字数の制限のため記事の最初の 350 文字のみを利用した．また，これにあわせて本節の実験では，すべての記事について最初の 350 文字のみを利用した．

本節では，一般的な情報の重要度を特定することを目的としている．このため，全体データで被験者で多数決をとり，重要と答えられた数の多い方の記事を重要記事と考え，記事ペアを入力としてその重要記事を機械学習法を利用して特定する実験を行った．

機械学習法には，サポートベクターマシン法 (SVM) [7] と最大エントロピー法 (ME)[8] を利用した．サポートベクターマシン法では，線形カーネルを用い C=1 のパラメータを利用して実験した [9]．素性としては，表 3 に示すものを用いた．

実験は 10 分割クロスバリデーションで行った．その結果を表 1 に示す．表 1 では，さらに，重要記事と考えた被験者の割合が 60%，70%，80%以上であったものだけで行った実験 (それぞれの場合の実験で用いられた事例数は，290 個，113 個，17 個である) も記載している．

被験者の意見もわかれる，全データの実験結果では 7 割程度と性能は悪いが，60%以上の実験では 80%を超える高い精度を実現している．60%以上の被験者が重要と考える一般的な情報の重要度は，80%以上の精度でもとめることができることがわかった．

次に，どういう情報が重要かを調べる実験を行った．その調査結果を表 2 に示す．この実験では，分析しやすいようにタイトルにあった名詞のみを素性として用いた．機械学習法には最大エントロピー法を用いた．最大エントロピー法でもとまる α 値を正規化した値 (ここでは正規化 α 値と呼ぶ．2 個の記事のうちどちらの方が重要かを示す二分類においてこの α 値の和が 1 になるように正規化している．) をもとめた．この値が大きいほどその素性が重要であり，小さいほど重要でないことを示す．

表において「年度」「政府」「事故」「殺人」などは一般的に重要と思われれる.「年金」「北朝鮮」は最近特に重要と思われる.「安倍 (元首相)」「ライブドア」は，問題をおこして現在は人気が下がっており重要でない事柄と思われる．

表 1: 実験結果

素性	全データ		60%以上		70%以上		80%以上	
	SVM	ME	SVM	ME	SVM	ME	SVM	ME
1,3,5,7	73,48%	73.39%	84.48%	83.62%	89.38%	91.15%	94.12%	88.24%
1,2,3,4,5,6,7,8	72.68%	72.50%	85.86%	85.17%	92.04%	85.84%	100.00%	94.12%

表 2: 機械学習により取り出した特徴的な素性

素性の単語	重要度	素性の単語	重要度
年度	0.74	...	...
政府	0.72	談合	0.32
事故	0.72	総裁	0.32
トヨタ	0.70	共同	0.32
電話	0.69	国連	0.31
方向	0.68	選挙	0.30
殺人	0.67	安倍	0.30
年金	0.66	過半数	0.29
最高	0.66	協議	0.28
北朝鮮	0.66	ライブ	0.23
...	...	ドア	0.23

表 3: 素性

素性	説明
1	タイトルのみにあった名詞，形容詞，副詞，動詞，未知語の単語
2	タイトルのみにあった名詞，形容詞，副詞，動詞，未知語の単語の分類語彙表 [10] の番号の 1,2,3,4,5,7 桁 (ただし番号は論文 [11] のように変更)
3	タイトルの直後にある 1 文のみにあった名詞，形容詞，副詞，動詞，未知語の単語
4	タイトルの直後にある 1 文のみにあった名詞，形容詞，副詞，動詞，未知語の単語の分類語彙表の番号の 1,2,3,4,5,7 桁
5	タイトルの直後にある 1 文を除いた本文にあった名詞，形容詞，副詞，動詞，未知語の単語
6	タイトルの直後にある 1 文を除いた本文にあった名詞，形容詞，副詞，動詞，未知語の単語の分類語彙表の番号の 1,2,3,4,5,7 桁
7	タイトル，本文のいずれかにあった名詞，形容詞，副詞，動詞，未知語の単語
8	タイトル，本文のいずれかにあった名詞，形容詞，副詞，動詞，未知語の単語の分類語彙表の番号の 1,2,3,4,5,7 桁

3 個々のユーザごとの情報の重要度の推定

被験者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被験者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人によって異なることがわかる．本節では，このユーザごとに異なる情報の重要度を推定することを試みる．

前節のデータを用いて，個々の被験者ごとに二つの記事のうちどちらが重要とされたかを自動推定する実験を行った．前節と同様の機械学習手法と素性を用い，10 分割クロスバリデーションで評価した．その結果のヒストグラムを図 1 に示す．また，被験者全体での精度の平均

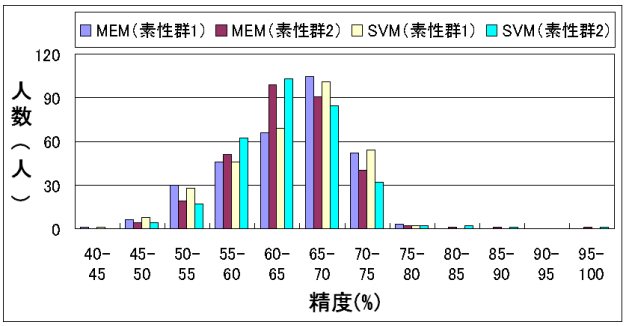

図 1: 個々人の情報の重要度の推定結果

表 4: 個々人の情報の重要度の推定結果の平均

素性	SVM	ME
1,3,5,7	63.96%	64.02%
1,2,3,4,5,6,7,8	63.80%	64.03%

を表 4 に示す．図の素性群 1 は素性 1,3,5,7 を，素性群 2 は全素性を用いたことを意味する．65%前後の精度で個人の情報の重要度を推定できることがわかった．

4 属性ごとの有効な素性

被験者の性別 (男性・女性)，居住地 (関東地方・近畿地方など) といった属性ごとに，どういう情報が重要かを調べる実験を行った．その調査結果を表 5,6 に示す．

本節の実験では，分析しやすいようにタイトルにあった名詞のみを素性として用いた．機械学習法には最大エントロピー法を用いた．記事ペアに対して各属性ごとに被験者の多数決をとり重要とされた記事を特定し，そのデータを学習データとして最大エントロピー法で正規化 α 値を求めた．表の重要度は，この属性ごとにもとめた正規化 α 値を，その属性と同じカテゴリのすべての属性 (男性の場合は，男性と女性，関東地方の場合は，8 地方) の正規化 α 値の和で割ったものである．表の重要度は，同じカテゴリの他の属性に対して相対的に重要であるかどうかを示したものになっている．表 5 では，男性と女性の属性について，それぞれ重要度の上位 30 個を示している．表 6 では，8 地方の属性について，それぞれ重要度の上位 15 個を示している．

表 5 から男性が重要としている素性と女性が重要としている素性に違いがあることがわかる．例えば，男性では「野球」や「トヨタ」などを重要と考えているのに対

表 5: 男性と女性の素性

男性 (148 人)		女性 (161 人)	
素性の単語	重要度	素性の単語	重要度
東証	0.636	訴訟	0.658
中間	0.634	採択	0.627
代表	0.630	消費	0.626
民主	0.621	不明	0.622
海域	0.619	時代	0.609
官民	0.618	対象	0.606
落札	0.614	義務	0.606
投資	0.613	解明	0.604
決算	0.612	さん	0.601
工事	0.604	**めぐみ**	0.600
トヨタ	0.603	最終	0.600
協議	0.601	皇室	0.599
主導	0.598	建物	0.599
工場	0.597	世論	0.599
商業	0.594	吸水	0.595
反発	0.592	支援	0.594
利益	0.592	大阪	0.594
委員	0.589	提供	0.592
ぶり	0.588	高校	0.587
野球	0.588	9月	0.585
国連	0.588	売却	0.584
把握	0.587	大学	0.580
議員	0.587	最高裁	0.580
ジャパン	0.586	幕開け	0.580
監視	0.585	和歌山	0.580
金利	0.585	予定	0.579
株式	0.582	**出産**	0.579
会議	0.580	紀子	0.579
廃止	0.580	**懐妊**	0.579
法案	0.577	さま	0.579

して，女性では「めぐみさん」や「出産」,「懐妊」などを重要と考えていることがわかった．

表 6 からは，北海道地方では「北海道」などの単語が得られている．東北地方では，福島県の談合に関わる単語が得られている．福島県の県知事の「佐藤」の名前や「発注」,「工事」などはまさに地域的なニュースである．中部地方では，トヨタ，ホンダなどの地域性を反映した単語が重要とされている．近畿地方は,「株式」「資金」などの金銭的なものが重要であり，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．

5 ユーザ個人の興味が重要な記事の判断に与える影響の分析

ユーザ個人の趣味・興味，仕事，気になっているニュースをそれぞれ 5 個以下ずつアンケート時に答えてもらっていた．これらの情報をここでは興味情報と呼ぶ．この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かを分析した．

前節と同様に，分析しやすいようにタイトルにあった名詞のみを素性として用いた．機械学習法には最大エントロピー法を用いた．ユーザ個人ごとに機械学習し，1,469 個の種類の単語が素性として得られた．ユーザ個人ごとに正規化 α 値の上位 500 個の単語と下位 500 個の単語を取り出した．ユーザ個人の興味情報として書いたものとこれらの単語の重なりを調べた．上位 500 個の単語の方が下位 500 個の単語よりも興味情報として書いたものと重なりが多かった被験者は 304 人中 141 人であった．逆に下位 500 個の単語の方が重なりが多かった被験者は 304 人中 91 人であった．141 人と 91 人では有意水準 5%の片側符号検定で有意差がある．このため，最大エントロピー法で重要と判断された単語と被験者の書いた興味情報は，重なる方が多く，相関があることがわかる．

また，有意水準 5%の片側符号検定で，上位 500 個の単語との重なりの個数と下位 500 個の単語との重なりの個数で有意差があった被験者だけで調べると，上位 500 個の単語の方が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 人でも有意水準 5%の片側符号検定で有意差がある．このため，最大エントロピー法で重要と判断された単語と被験者の書いた興味情報が有意に重なる方が多く，最大エントロピー法で重要と判断された単語と被験者の書いた興味情報は，相関があることがわかる．

これらの結果により，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る．

具体例を示す．ある被験者の趣味・興味は,「旅行」「海外投資」「サッカー」「中国語」「経済」であり，仕事は大学非常勤講師・中等教育学校教諭であり，気になっているニュースは「サブプライムローンの今後の行方」「サッカー（男子）日本代表チーム監督人選」「株価下落」であった．これらの興味情報のうち，この被験者で学習した場合の正規化 α の上位 500 個にあった単語は,「日本」「大学」「教育」「サッカー」「海外」「代表」「投資」であり，下位 500 個にあった単語は「行方」のみであった．またこの被験者の上位 10 個の素性は,「拡大」「日本」「大学」「教育」「研究」「トヨタ」「一致」「資金」「ニッポン放送」「家電」であり，この被験者の職業の大学・教育，趣味の「投資」とも一致する結果である．

6 おわりに

本稿では，機械学習を利用した情報の重要度に関する研究を行った．実験の結果，60%以上の被験者が重要と考える一般的な情報の重要度は，80%以上の精度でもとめることができることがわかった．また，素性を分析し，一般的に重要な事柄，重要でないとされた事柄を示した．また，ユーザ個人の考える情報の重要度は約 65%で推定できた．

男性・女性，関東地方・近畿地方といった属性ごとに，どういう情報が重要かを調べる実験を行った．素性を分

表 6: 日本の 8 地方の素性

北海道地方 (15 人)		東北地方 (19 人)		関東地方 (122 人)		中部地方 (59 人)	
素性の単語	重要度	素性の単語	重要度	素性の単語	重要度	素性の単語	重要度
主導	0.180	**発注**	0.182	**戦後**	0.168	工場	0.176
高校	0.177	さん	0.172	原点	0.164	対応	0.169
日本	0.175	要求	0.169	認識	0.162	交渉	0.166
談合	0.174	自治体	0.165	**教育**	0.161	設備	0.165
直後	0.174	追及	0.165	返還	0.160	賛成	0.162
担当	0.174	**工事**	0.164	拡大	0.158	住宅	0.159
北海道	0.171	逮捕	0.163	**政府**	0.153	政治	0.159
野球	0.171	社長	0.161	対策	0.152	**トヨタ**	0.157
支配	0.171	**佐藤**	0.159	安保理	0.148	取得	0.154
認識	0.168	弁護士	0.157	利益	0.148	首相	0.153
海域	0.166	大手	0.155	行方	0.148	共同	0.153
ドイツ	0.166	整備	0.155	不明	0.147	政府	0.153
サッカー	0.166	格差	0.155	方向	0.147	過半数	0.152
W杯	0.166	企業	0.155	入学	0.147	工事	0.152
松本	0.165	管理	0.155	申告	0.147	**ホンダ**	0.152
近畿地方 (46 人)		中国地方 (12 人)		四国地方 (5 人)		九州地方 (31 人)	
素性の単語	重要度	素性の単語	重要度	素性の単語	重要度	素性の単語	重要度
選挙	0.189	工場	0.213	内部	0.200	**安倍**	0.194
中国	0.188	粉飾	0.188	捜査	0.188	地検	0.178
強化	0.180	会議	0.186	大統領	0.187	拡大	0.175
東証	0.179	違反	0.182	民主	0.187	輸出	0.174
友好	0.178	村上	0.182	機関	0.183	決算	0.174
株式	0.168	ファンド	0.179	北朝鮮	0.180	提携	0.171
障害	0.167	松下	0.177	負担	0.180	幕開け	0.167
新株	0.166	ライブ	0.176	解決	0.178	貨物	0.166
資金	0.163	ドア	0.176	中間	0.178	ホテル	0.165
投資	0.158	作成	0.176	成立	0.175	一部	0.163
訪中	0.155	防止	0.172	提言	0.175	大学	0.161
財界	0.155	経済	0.171	防衛施設庁	0.174	緊急	0.161
被告	0.155	全国	0.169	ブッシュ	0.172	国連	0.160
訴訟	0.154	自治体	0.169	議長	0.171	断念	0.160
発行	0.153	取引	0.169	大差	0.170	検査	0.160

析し，それぞれの属性ごとに異なる重要な事柄を示した．

アンケートにおいて答えたもらったユーザ個人の興味情報と，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．

謝辞: 本研究は科研費 (19700154) の助成を受けたものである．

参考文献

[1] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Newsinessence: a system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the first international conference on Human language technology research*, pages 1–4, 2001.

[2] Gianna M. Del Corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. In *WWW 2005*, pages 97–106, 2005.

[3] Jinyi Yao, Jue Wang, Zhiwei Li, Mingjing Li, and Wei-Ying Ma. Ranking web news via homepage visual layout and cross-site voting. In *Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, pages 131–142, 2006.

[4] Yang Hu, Mingjing Li, Zhiwei Li, and Wei-Ying Ma. Discovering authoritative news sources and top news stories. In *AIRS 2006*, pages 230–243, 2006.

[5] 村田 真樹, 西村 涼, 金丸 敏幸, 土井 晃一, 松岡 雅裕, and 井佐原 均. 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. In **言語処理学会第** *14* **回年次大会**, pages 907–910, 2008.

[6] Masaki Murata, Ryo Nishimura, Kouichi Doi, Toshiyuki Kanamaru, and Kentaro Torisawa. Analysis of the degree of importance of information using newspapers and questionnaires. *2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2008)*, pages 137–144, 2008.

[7] Taku Kudoh. TinySVM: Support Vector Machines. http://cl.aist-nara.ac.jp/ taku-ku// software/TinySVM/ index.html, 2000.

[8] Masao Utiyama. Maximum Entropy Modeling Package. http://www.nict.go.jp/x/x161/ members/mutiyama/software.html#maxent, 2006.

[9] 村田 真樹, 馬 青, 内元 清貴, and 井佐原 均. サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳. In **電子情報通信学会 言語理解とコミュニケーション研究会** *NLC2000-78*, 2001.

[10] 国立国語研究所. **分類語彙表**. 秀英出版, 1964.

[11] 村田 真樹, 神崎 享子, 内元 清貴, 馬 青, and 井佐原 均. 意味ソート msort — 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例 —. **言語処理学会誌**, 7(1):51–66, 2000.