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

村田 真樹 *　西村 涼 *　金丸 敏幸 *　土井 晃一 **　松岡 雅裕 **　井佐原 均 *

* 独立行政法人 情報通信研究機構　({murata,nishimura,kanamaru,isahara} @nict.go.jp)

**(株)PSC　({doy,mage} @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．本研究では，手始めに新聞の構成情報を利用した，情報の重要度の研究を行った．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，被験者実験を行い，被験者の重要と考える記事を特定する研究を行った．

2　新聞記事を用いた機械学習に基づく実験

2006 年度の毎日新聞，読売新聞，日経新聞の三社の朝刊の新聞記事データを利用した．以下の三つの実験を行った．

- 実験 A: 1 面記事かそれ以外の面の記事かを特定する
- 実験 B: 1 面トップ記事かそれ以外の面の記事かを特定する
- 実験 C: 1 面トップ記事か 1 面内の他の記事かを特定する

1 面記事は他の面よりも重要度が高いと考えられる．また，1 面トップ記事はさらに重要度が高いと考えられる．このため，重要度に関する研究の手始めとしては上記実験を行った．

2006 年度のすべての日を使ったデータを作成した．実験 A は，1 年分の 1 面記事として，各社約 2000-3000 記事を利用し，それ以外の面の記事はそれ以外の面からランダムに 1 面記事と同数のものを取り出して，合計約 4000-6000 記事を利用した．実験 B,C は，1 年分の 1 面トップ記事として，各社約 350 記事を利用し，それ以外の面の記事または 1 面内の他の記事は，その場所からランダムに 1 面トップ記事と同数のものを取り出して，合計約 700 記事を利用した．

表 1: 素性

素性	説明
1	タイトルのみにあった名詞
2	タイトルのみにあった名詞の分類語彙表 [1] の番号の 1,2,3,4,5,7 桁 (ただし番号は論文 [2] 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
4	タイトルの直後にある 1 文のみにあった名詞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
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
8	タイトル，本文のいずれかにあった名詞の分類語彙表の番号の 1,2,3,4,5,7 桁

機械学習法には，サポートベクターマシン法 (SVM) [3] と最大エントロピー法 (ME)[4] を利用した．サポートベクターマシン法では，d=1,C=1 で実験した [5] (d=2 の実験も行ったが概ね d=1 の方がよいことを確認している)．素性としては，表 1に示すものを用いた．

まず，実験 A で，一つの記事を入力とし，それがどういう記事かを特定する実験を行った．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 2に示す．表の素性の列にある数字は，表 1のうちその行の実験で用いた素性を意味する．次に，実験 A で，二つの記事 (1 面記事とそれ以外の面の記事) を入力とし，どちらが 1 面記事かを特定する実験を行った．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 3に示す．

この実験結果から，一つの記事について 1 面かそれ以外かを特定するよりも，二つの記事のペアを与えて，どちらが 1 面かを特定する方が簡単であることがわかった．

次に，実験 B,C で，二つの記事 (1 面トップ記事とそうでない記事) を入力とし，どちらが 1 面トップ記事かを特定する実験を行った．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 4と表 5に示す．

実験 A,B,C の中では 1 面トップ記事かそれ以外の面の記事かを特定する実験 B の精度が比較的高いことが

表 2: 1 記事入力の場合の実験 A

素性	毎日新聞		読売新聞		日経新聞	
	SVM	ME	SVM	ME	SVM	ME
1,2	71.82%	73.36%	77.97%	79.35%	76.64%	73.73%
3,4	61.98%	51.13%	68.09%	52.90%	64.97%	55.69%
5,6	70.23%	57.51%	75.60%	53.47%	77.61%	49.67%
7,8	70.93%	81.54%	79.89%	84.60%	78.19%	86.03%
1,2,3,4	72.70%	73.73%	79.84%	80.82%	76.84%	77.59%
1,2,3,4,5,6	73.72%	82.64%	77.93%	85.54%	80.04%	85.18%
1,2,3,4,5,6,7,8	66.46%	82.98%	76.19%	85.56%	76.43%	85.63%
1	72.24%	73.68%	79.44%	80.99%	77.43%	75.20%
3	61.47%	51.10%	69.36%	54.06%	66.07%	55.34%
5	74.89%	57.35%	79.68%	54.01%	80.76%	49.88%
7	75.79%	81.64%	81.25%	85.11%	80.54%	86.26%
1,3	73.16%	74.29%	81.78%	82.64%	77.51%	78.84%
1,3,5	82.81%	83.92%	86.39%	87.21%	86.34%	86.63%
1,3,5,7	77.46%	83.86%	84.70%	87.07%	82.48%	86.05%

表 3: 記事ペア入力の場合の実験 A

素性	毎日新聞		読売新聞		日経新聞	
	SVM	ME	SVM	ME	SVM	ME
1,2	80.70%	81.34%	85.23%	84.82%	83.73%	84.16%
3,4	65.81%	63.10%	73.07%	74.26%	70.56%	71.01%
5,6	75.10%	84.35%	83.15%	86.09%	85.97%	88.67%
7,8	78.58%	88.85%	87.96%	91.55%	84.39%	91.50%
1,2,3,4	80.48%	81.02%	86.45%	86.41%	84.39%	83.89%
1,2,3,4,5,6	87.29%	89.80%	88.17%	90.00%	89.18%	90.65%
1,2,3,4,5,6,7,8	85.96%	90.11%	89.11%	91.23%	89.41%	90.73%
1	83.43%	83.02%	87.35%	88.00%	85.24%	84.70%
3	64.99%	62.69%	74.91%	75.36%	72.60%	72.56%
5	82.70%	83.62%	87.39%	88.21%	87.02%	88.21%
7	84.44%	88.21%	90.58%	92.21%	89.41%	91.65%
1,3	82.38%	82.38%	88.90%	88.78%	85.36%	84.66%
1,3,5	90.62%	90.59%	92.66%	92.37%	91.73%	90.80%
1,3,5,7	88.97%	90.91%	91.47%	92.49%	90.57%	90.84%

表 4: 記事ペア入力の場合の実験 B

素性	毎日新聞		読売新聞		日経新聞	
	SVM	ME	SVM	ME	SVM	ME
1,2	84.42%	84.42%	86.69%	85.84%	87.04%	86.48%
3,4	68.56%	71.39%	73.37%	74.22%	78.87%	77.46%
5,6	94.05%	92.92%	94.05%	96.03%	97.18%	98.59%
7,8	88.95%	92.35%	96.32%	96.03%	98.03%	98.31%
1,2,3,4	81.87%	82.15%	88.39%	87.54%	89.01%	89.86%
1,2,3,4,5,6	95.18%	93.48%	94.05%	96.03%	96.62%	99.15%
1,2,3,4,5,6,7,8	94.62%	93.48%	93.77%	96.88%	97.46%	99.44%
1	85.55%	84.14%	90.37%	87.25%	87.61%	87.89%
3	67.99%	67.99%	75.64%	73.65%	80.00%	80.85%
5	94.90%	94.62%	95.75%	95.47%	99.44%	100.00%
7	92.63%	94.33%	96.88%	97.17%	98.87%	98.59%
1,3	85.27%	84.70%	87.54%	83.57%	87.89%	87.89%
1,3,5	94.90%	94.62%	95.75%	96.03%	98.59%	99.72%
1,3,5,7	94.90%	94.33%	96.60%	96.03%	98.03%	100.00%

表 5: 記事ペア入力の場合の実験 C

素性	毎日新聞		読売新聞		日経新聞	
	SVM	ME	SVM	ME	SVM	ME
1,2	75.92%	76.49%	75.64%	76.20%	68.17%	68.73%
3,4	57.79%	56.66%	60.34%	60.62%	71.27%	72.68%
5,6	76.49%	77.34%	88.67%	89.52%	88.73%	89.86%
7,8	75.35%	76.20%	83.00%	88.67%	86.20%	89.01%
1,2,3,4	71.95%	72.24%	75.92%	77.05%	73.24%	74.65%
1,2,3,4,5,6	77.90%	78.47%	88.67%	90.65%	89.58%	90.14%
1,2,3,4,5,6,7,8	75.35%	77.90%	89.80%	90.08%	89.58%	90.42%
1	71.95%	71.67%	77.62%	79.04%	73.52%	74.37%
3	56.37%	56.94%	58.36%	62.04%	73.52%	71.83%
5	78.19%	77.34%	89.24%	89.24%	92.11%	91.83%
7	71.10%	74.79%	87.25%	89.24%	88.17%	90.42%
1,3	71.39%	71.10%	75.92%	76.77%	76.34%	75.21%
1,3,5	78.47%	78.47%	89.80%	89.24%	90.99%	91.27%
1,3,5,7	77.34%	76.77%	88.39%	88.95%	90.42%	90.70%

表 6: アンケートデータでの実験

素性	全データ		60% 以上		70% 以上		80% 以上	
	SVM	ME	SVM	ME	SVM	ME	SVM	ME
1,2	66.79%	67.14%	76.55%	75.52%	73.45%	72.57%	88.24%	88.24%
3,4	65.00%	66.79%	78.28%	77.93%	78.76%	78.76%	88.24%	88.24%
5,6	67.50%	69.29%	73.10%	74.48%	83.19%	81.42%	94.12%	88.24%
7,8	70.00%	71.96%	81.03%	78.97%	85.84%	80.53%	88.24%	88.24%
1,2,3,4	66.07%	70.00%	79.31%	81.38%	80.53%	77.88%	88.24%	88.24%
1,2,3,4,5,6	69.64%	70.71%	80.69%	79.66%	83.19%	79.65%	88.24%	88.24%
1,2,3,4,5,6,7,8	69.46%	71.25%	83.45%	81.03%	85.84%	80.53%	88.24%	88.24%
1	65.89%	66.25%	74.83%	75.86%	78.76%	73.45%	88.24%	88.24%
3	65.89%	67.14%	71.03%	73.10%	76.11%	73.45%	88.24%	88.24%
5	68.04%	68.93%	81.38%	79.66%	76.99%	76.99%	88.24%	94.12%
7	71.07%	70.71%	78.97%	78.28%	84.07%	78.76%	88.24%	94.12%
1,3	68.04%	69.46%	73.10%	74.48%	74.34%	74.34%	88.24%	88.24%
1,3,5	68.75%	69.82%	80.34%	78.62%	75.22%	73.45%	88.24%	94.12%
1,3,5,7	70.54%	70.89%	79.31%	78.28%	83.19%	76.99%	88.24%	94.12%

わかる.

3 アンケートデータを利用した機械学習に基づく実験

次にアンケートデータを利用した実験を行った．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，10 個のどちらが重要とされたかの情報を含む記事ペアを生成することで，56 個のデータから，計 560 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約 300 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約 300 文字のみを利用した.

ここで，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特定する実験を行った．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 6に示す．表 6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，全データや「60%」などの実験結果では性能は悪いが，「80%」の実験では高い精度を実現している.

次に，新聞記事を学習データとして，アンケートデータをテストデータとした実験を行った．その結果を表 7に示す．また，新聞記事とアンケートデータを学習データとして，アンケートデータをテストデータとした実験を行った．これはアンケートデータ部分については 10 分割のクロスバリデーションで実験した．その結果を表 8に示す．グラフ中の混合は実験 A,B,C のすべてのデータを利用したものを意味し，全新聞社は全新聞社のデータを利用したものを意味する．これらの実験は，重要記事と考えた被験者の割合が 80% 以上であったものだけで行った.

新聞社データだけを学習データとして用いる表 7では，毎日が 88% をあげ高精度であり，次は 82% の読売である．新聞社データから一般的な被験者が重要と思う記事を特定するには，毎日新聞，読売新聞の順に役立つことがわかる．実験環境がよいときには，一般的な被験者が重要と思う記事を特定するのに，新聞社データが利用でき，88% の精度で特定できることがわかった．また，そのときの新聞社は，毎日新聞で，実験の種類は実験 B であった．これは，2 節の実験でも実験 A,B,C の中で実験 B が比較的性能がよかったが，それと関係があると思われる．実験 B は，1 面トップ記事かそれ以外の面の記事かを特定するものであり，比較する 2 記事がかなりかけはなれたものであり，それが良い影響を与えたと思われる.

新聞データとアンケートデータを利用する方法では，最高精度 (94%) はアンケートデータだけを学習データに用いるものと同じであり，新聞データを学習データに追加で用いた効果は見ることはできなかった.

表 7: 新聞記事を学習データとしてアンケートデータをテストデータとした実験（80% 以上被験者一致）

種類	素性	毎日新聞		読売新聞		日経新聞		全新聞社	
		SVM	ME	SVM	ME	SVM	ME	SVM	ME
実験 A	1,3,5,7	70.59%	70.59%	64.71%	82.35%	58.82%	58.82%	58.82%	70.59%
実験 A	1,2,3,4,5,6,7,8	82.35%	76.47%	52.94%	58.82%	64.71%	70.59%	58.82%	52.94%
実験 B	1,3,5,7	88.24%	82.35%	64.71%	64.71%	64.71%	58.82%	76.47%	70.59%
実験 B	1,2,3,4,5,6,7,8	82.35%	70.59%	64.71%	64.71%	76.47%	64.71%	82.35%	76.47%
実験 C	1,3,5,7	70.59%	70.59%	70.59%	76.47%	41.18%	64.71%	76.47%	76.47%
実験 C	1,2,3,4,5,6,7,8	52.94%	58.82%	52.94%	64.71%	64.71%	70.59%	58.82%	58.82%
混合	1,3,5,7	82.35%	70.59%	82.35%	76.47%	58.82%	58.82%	70.59%	70.59%
混合	1,2,3,4,5,6,7,8	64.71%	76.47%	58.82%	52.94%	35.29%	52.94%	58.82%	58.82%

表 8: 新聞記事とアンケートデータを学習データとしてアンケートデータをテストデータとした実験（80% 以上一致）

種類	素性	毎日新聞		読売新聞		日経新聞		全新聞社	
		SVM	ME	SVM	ME	SVM	ME	SVM	ME
実験 A	1,3,5,7	82.35%	76.47%	70.59%	94.12%	82.35%	70.59%	64.71%	76.47%
実験 A	1,2,3,4,5,6,7,8	88.24%	76.47%	58.82%	70.59%	82.35%	88.24%	88.24%	64.71%
実験 B	1,3,5,7	94.12%	88.24%	88.24%	88.24%	70.59%	88.24%	94.12%	88.24%
実験 B	1,2,3,4,5,6,7,8	94.12%	82.35%	88.24%	88.24%	94.12%	82.35%	94.12%	88.24%
実験 C	1,3,5,7	76.47%	76.47%	88.24%	76.47%	82.35%	94.12%	88.24%	88.24%
実験 C	1,2,3,4,5,6,7,8	76.47%	82.35%	70.59%	76.47%	82.35%	88.24%	64.71%	70.59%
混合	1,3,5,7	88.24%	88.24%	76.47%	88.24%	70.59%	70.59%	76.47%	76.47%
混合	1,2,3,4,5,6,7,8	47.06%	82.35%	70.59%	58.82%	58.82%	82.35%	52.94%	64.71%

4 おわりに

本稿では，機械学習を利用した重要度に関する研究を行った．新聞記事での実験により，1 記事を与えて 1 面記事かどうかを特定するよりも，2 記事を与えてそのどちらが 1 面記事かどうかを特定する方が簡単であることがわかった．また，実験 A,B,C と行ったが，1 面トップ記事かそれ以外の面の記事かを特定する実験 B が最も高い精度をあげることがわかった．

被験者を利用した実験では，被験者の一致率が高い記事ペア (一致率 80% 以上) については，94% と高い精度で重要記事を特定できた．また，そのような記事ペアは新聞記事だけからでも，88% と高い精度で重要記事を特定できた．このことは，新聞データが，被験者データの代用としてもある程度利用できることを意味する．

今後はアンケートデータの分析 [6] も行いたいと考えている．例えば，テキストマイニングシステム Simpleminer[7] を用いると，「ライブドア」の重要度が低く，「年金」の重要度が高いという結果を得た．ここでは重要と被験者が判断した記事のタイトルに偏って多く出現したものを重要度が高いとしている．アンケートを 2007 年 11 月に実施したため，今はほとんどの人が「ライブドア」事件に興味がなく，年金問題に興味があることがわかった．また，「殺人，死亡，病院，保険，金融，与党，改革，天下り」の重要度も高いこと，「選挙，工事，談合，野球」の重要度が低いこともわかった．この種のアンケート分析についても今後機会があれば詳しく議論したいと考えている．

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] 国立国語研究所, 分類語彙表, (秀英出版, 1964).

[2] 村田真樹, 神崎享子, 内元清貴, 馬青, 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, Vol. 7, No. 1, (2000), pp. 51–66.

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

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

[5] 村田真樹, 馬青, 内元清貴, 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, 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).

[6] 上田太一郎, 村田真樹, 小木しのぶ, 高山泰博, 末吉正成, 今村誠, 渕上美喜, 事例で学ぶテキストマイニング, (共立出版, 2008).

[7] 村田真樹, 金丸敏幸, 一井康二, 白土保, 馬青, 井佐原均, テキストマイニングシステム simpleminer の開発, 言語処理学会第 14 回年次大会, (2008).