C210

情報抽出表とソートグラフを利用した テキストマイニングシステム Simpleminer

村田 真樹[○] (情報通信研究機構), 金丸 敏幸 (京都大学), 一井 康二 (広島大学),

馬 青 (龍谷大学), 白土 保, 井佐原 均 (情報通信研究機構)

Text-mining System Simpleminer: Use of IE Table and Sort Graph

Masaki MURATA, Toshiyuki KANAMARU, Koji ICHII,
Qing MA, Tamotsu SHIRADO and Hitoshi ISAHARA

ABSTRACT

We have developed a simple text-mining system called "Simpleminer." The system has been designed to function on machines equipped with the Windows operating system. It can be used, for example, to analyze questionnaires that include text and trends in journal titles. The system performs fundamental text-mining functions such as the determination of the usage frequencies of the words and the creation of multiway tables. In addition, Simpleminer has two unique functions: it can generate information extraction (IE) tables and sort graphs. An information extraction table shows whether a sample of text data includes words with high usage frequencies. A sort graph can be used for data with two attributes x_1 and x_2 ; x_1 is a numerical value, while x_2 is text. A useful feature of this graph is that the words in x_2 that have a small corresponding value of x_1 appear more frequently and are displayed higher on the list. This enables users to recognize such words more effectively.

Keywords: Text-mining, Simpleminer, Information extraction table, Sort graph

1. はじめに

本稿では、われわれが開発した簡易テキストマイニングシステム Simpleminer について紹介する^{1),2)}。このシステムは、Windows上で簡便に動作する。自由記述のアンケートデータの分析や、論文書誌情報・論文タイトル情報からの動向分析に用いることができる。一般的なテキストマイニングシステム³⁾が持つ、単語の頻度分析、クロス分析が可能である。そのうえ、情報抽出表とソートグラフと呼ぶ、他のシステムにない新規な技術を利用した分析も可能である。情報抽出表は、データを分かりやすい表の形で整理して表示する機能である。ソートグラフは、片方が数値である場合のクロス表のデータの分析に利用できるものであり、若い年齢層の傾向や年配の方の傾向を簡便につかむことができる機能である。

2. インストール

Simpleminer は、Windows に簡便にインストールができる。インストーラーが用意されており、それをクリックすることで、Fig. 1 に示すようなインストール画面が立ち上がる。ここで、「次へ」をクリックしていくことで簡単にインストールできる。Simpleminer を起動する

Fig. 1 Installation

と、Fig. 2 に示すような画面が立ち上がる。

3. 入力

入出力ファイルは csv 形式(カンマ区切りのデータ形式)である。入力ファイルの例を Fig. 3 に示す。図のような入力ファイルを準備し、そのファイルを Simpleminer にセットして、Fig. 2 の各種ボタンを押すことで様々な処理ができるようになっている。Fig. 2 中の「表示」ま

Fig. 2 Simpleminer

Fig. 3 Input File in the CSV Form

たは「Excel」ボタンを押すと、それに該当するファイルを、Note Pad または、MS Word で開いて表示できる。

入力データとしては、Simpleminerを利用した先行研究^{1),2),4),5),6)}では、言語処理学会の論文書誌情報や人工知能学会全国大会の論文書誌情報を利用していたが、本論文では、アンケートデータを利用した。筆者らはどのような情報が重要かを明らかにする研究⁷⁾を進めておりその際に収集したアンケートデータを利用した。2007 年の秋に 309 人の被験者に対して、どのような記事や情報を重

	A	B	C	D
1	出現形	基本形	品詞	出現行数
2	事件	事件	名詞	113
3	問題	問題	名詞	80
4	行方	行方	名詞	67
5	不明	不明	名詞	67
6	香川	香川	名詞	65
7	祖母	祖母	名詞	56
8	人	人	名詞	49
9	孫	孫	名詞	42
10	原油	原油	名詞	37
11	県	県	名詞	36
12	殺人	殺人	名詞	35
13	、	、	記号	33
14	化	化	名詞	31
15	高	高	名詞	30
16	プライム	プライム	名詞	30
17	温暖	温暖	名詞	30
18	高騰	高騰	名詞	30
19	防衛	防衛	名詞	29
20	地球	地球	名詞	29
21	食品	食品	名詞	29

Fig. 4 Word Count

要と考えるかのアンケートを取った。実際のアンケートでは 60 問の設問があったが、ここでの分析では、その一部分のみを利用した。ここで用いたアンケートデータには、被験者の性別 (1: 男性, 2: 女性), 年齢 (歳代), 地域 (8 地域), 結婚 (1: 未婚, 2: 既婚), 子供 (1: 無, 2: 有), 趣味・興味 (自由回答文), 職 (自由回答文), 被験者の興味のある事件や話題 (自由回答文), 作業の基準 (自由回答文), 重要な記事を提示するシステムに関する要望 (自由回答文), 今回のアンケートの感想 (自由回答文) の情報がある。ここには含めていないが、5 つの記事を見せて、自分の考える重要な記事の順に並べ替える作業もしてもらっており、その際に自分がどういう基準でその並べ替えを行ったかの作業の基準についても答えてもらっている。

4. 単語集計

単語集計機能を使うと、自由回答文においてどの単語が何個出現したかを調べることができる。例えば、被験者の興味のある事件や話題の質問項目「事件」の自由回答文に対して単語集計機能を使うと、Fig. 4 の結果を得る。Fig. 4 ではすべての品詞の単語を示した。これは、Fig. 2 で、すべての品詞をチェックし、文字長の設定値も 0 とし、ひらがな語削除の機能のチェックを外して単語集計の「再集計」ボタンを押して実行したものである。Fig. 4 では分析に役立たない単語が多く出現している。Simpleminer では分析に用いる単語を簡便に指定することができる。ここでは、名詞のみを対象とし、文字長も 2 文字以上のものでのみを対象とし、ひらがな語も不要として対象外として単語集計を試みる。これは、Fig. 2 のように設定させて、単語集計の「再集計」ボタンを押して単語集計を実行する。そうすると、Fig. 5 の結果を得る。Fig. 5 は不要な単語が削除され見やすくなる。単語集計機能により得られた Fig. 5 を見ることで、データ

	A	B	C	D
1	出現形	基本形	品詞	出現行数
2	事件	事件	名詞	113
3	問題	問題	名詞	80
4	行方	行方	名詞	67
5	不明	不明	名詞	67
6	香川	香川	名詞	65
7	祖母	祖母	名詞	56
8	原油	原油	名詞	37
9	殺人	殺人	名詞	35
10	プライム	プライム	名詞	30
11	温暖	温暖	名詞	30
12	高騰	高騰	名詞	30
13	防衛	防衛	名詞	29
14	地球	地球	名詞	29
15	食品	食品	名詞	29
16	サブ	サブ	名詞	28
17	年金	年金	名詞	27
18	ガソリン	ガソリン	名詞	25
19	値上げ	値上げ	名詞	22
20	偽装	偽装	名詞	22
21	失踪	失踪	名詞	20

Fig. 5 Word Count for Nouns

の大雑把な傾向をつかめる。これは、被験者の興味のあった事件や事柄について自由に記述してもらったデータを分析しているものだが、結果から、なにかよくわからないが「香川の事件」に興味を持っている被験者が多いことがわかる。また、「行方不明事件」「殺人事件」「サブプライム」「温暖化」「食品偽装」「ガソリン値上げ」についても多くの被験者が興味を持っていることが予想される。ここでは名詞のみを取り出して分析したが Simpleminer では対象とする品詞を変更することもできる。また、同義語辞書により、異なる単語を同じ単語として扱ったり、削除辞書により集計対象から単語を強制的に排除することもできる。単語への分割と品詞の推定には ChaSen を利用している。

5. 情報抽出表

次に情報抽出表の機能を示す(Fig. 2 の画面にはないが、集計モードを切り替えると情報抽出表の処理画面が表示される)。ここでは、「香川」という単語を、アンケートの「事件」の項目に含むデータだけを対象に実行した。Fig. 2 の Focus terms の欄に「香川」と入力すると「香川」という単語を含む論文だけを使った処理を実行できる。その結果を Fig. 6 に示す。「香川」という単語を含むデータの中で出現が大きかった単語の順に左から右に表示している。また各データの右側の欄にはその列の単語をそのデータが含んでいればその単語を表示している。データも上下方向にソートしており、なるべく左側の単語を含むデータの順に表示している。この表では各データがどのような単語を含んでいるかを簡便に把握することができる。Fig. 6 から、前節で多くの被験者が興味を持っていた「香川の事件」が実は、「香川の祖母孫の行方不明事件」のことであることがわかる。また、「香川」という単語を記述した被験者はほとんど、この「香川の祖母孫の行方不明事件」のことを書いているこ

	J	K	L	M	N
1	18) 事件	原油	高騰	事件	問題
2	19) 事件	原油	高騰	事件	問題
3	20) 事件	原油	高騰	事件	問題
4	21) 事件	原油	高騰	事件	問題
5	22) 事件	原油	高騰	事件	問題
6	23) 事件	原油	高騰	事件	問題
7	24) 事件	原油	高騰	事件	問題
8	25) 事件	原油	高騰	事件	問題
9	26) 事件	原油	高騰	事件	問題
10	27) 事件	原油	高騰	事件	問題
11	28) 事件	原油	高騰	事件	問題
12	29) 事件	原油	高騰	事件	問題
13	30) 事件	原油	高騰	事件	問題
14	31) 事件	原油	高騰	事件	問題
15	32) 事件	原油	高騰	事件	問題
16	33) 事件	原油	高騰	事件	問題
17	34) 事件	原油	高騰	事件	問題
18	35) 事件	原油	高騰	事件	問題
19	36) 事件	原油	高騰	事件	問題
20	37) 事件	原油	高騰	事件	問題
21	38) 事件	原油	高騰	事件	問題
22	39) 事件	原油	高騰	事件	問題
23	40) 事件	原油	高騰	事件	問題
24	41) 事件	原油	高騰	事件	問題
25	42) 事件	原油	高騰	事件	問題
26	43) 事件	原油	高騰	事件	問題
27	44) 事件	原油	高騰	事件	問題
28	45) 事件	原油	高騰	事件	問題
29	46) 事件	原油	高騰	事件	問題
30	47) 事件	原油	高騰	事件	問題
31	48) 事件	原油	高騰	事件	問題
32	49) 事件	原油	高騰	事件	問題
33	50) 事件	原油	高騰	事件	問題
34	51) 事件	原油	高騰	事件	問題
35	52) 事件	原油	高騰	事件	問題
36	53) 事件	原油	高騰	事件	問題
37	54) 事件	原油	高騰	事件	問題
38	55) 事件	原油	高騰	事件	問題
39	56) 事件	原油	高騰	事件	問題
40	57) 事件	原油	高騰	事件	問題
41	58) 事件	原油	高騰	事件	問題
42	59) 事件	原油	高騰	事件	問題
43	60) 事件	原油	高騰	事件	問題
44	61) 事件	原油	高騰	事件	問題
45	62) 事件	原油	高騰	事件	問題
46	63) 事件	原油	高騰	事件	問題
47	64) 事件	原油	高騰	事件	問題
48	65) 事件	原油	高騰	事件	問題
49	66) 事件	原油	高騰	事件	問題
50	67) 事件	原油	高騰	事件	問題
51	68) 事件	原油	高騰	事件	問題
52	69) 事件	原油	高騰	事件	問題
53	70) 事件	原油	高騰	事件	問題
54	71) 事件	原油	高騰	事件	問題
55	72) 事件	原油	高騰	事件	問題
56	73) 事件	原油	高騰	事件	問題
57	74) 事件	原油	高騰	事件	問題
58	75) 事件	原油	高騰	事件	問題
59	76) 事件	原油	高騰	事件	問題
60	77) 事件	原油	高騰	事件	問題
61	78) 事件	原油	高騰	事件	問題
62	79) 事件	原油	高騰	事件	問題
63	80) 事件	原油	高騰	事件	問題
64	81) 事件	原油	高騰	事件	問題
65	82) 事件	原油	高騰	事件	問題
66	83) 事件	原油	高騰	事件	問題
67	84) 事件	原油	高騰	事件	問題
68	85) 事件	原油	高騰	事件	問題
69	86) 事件	原油	高騰	事件	問題
70	87) 事件	原油	高騰	事件	問題
71	88) 事件	原油	高騰	事件	問題
72	89) 事件	原油	高騰	事件	問題
73	90) 事件	原油	高騰	事件	問題
74	91) 事件	原油	高騰	事件	問題
75	92) 事件	原油	高騰	事件	問題
76	93) 事件	原油	高騰	事件	問題
77	94) 事件	原油	高騰	事件	問題
78	95) 事件	原油	高騰	事件	問題
79	96) 事件	原油	高騰	事件	問題
80	97) 事件	原油	高騰	事件	問題
81	98) 事件	原油	高騰	事件	問題
82	99) 事件	原油	高騰	事件	問題
83	100) 事件	原油	高騰	事件	問題

Fig. 6 Information Extraction Table

	J	K	L	M	N
1	18) 事件	原油	高騰	事件	問題
2	19) 事件	原油	高騰	事件	問題
3	20) 事件	原油	高騰	事件	問題
4	21) 事件	原油	高騰	事件	問題
5	22) 事件	原油	高騰	事件	問題
6	23) 事件	原油	高騰	事件	問題
7	24) 事件	原油	高騰	事件	問題
8	25) 事件	原油	高騰	事件	問題
9	26) 事件	原油	高騰	事件	問題
10	27) 事件	原油	高騰	事件	問題
11	28) 事件	原油	高騰	事件	問題
12	29) 事件	原油	高騰	事件	問題
13	30) 事件	原油	高騰	事件	問題
14	31) 事件	原油	高騰	事件	問題
15	32) 事件	原油	高騰	事件	問題
16	33) 事件	原油	高騰	事件	問題
17	34) 事件	原油	高騰	事件	問題
18	35) 事件	原油	高騰	事件	問題
19	36) 事件	原油	高騰	事件	問題
20	37) 事件	原油	高騰	事件	問題
21	38) 事件	原油	高騰	事件	問題
22	39) 事件	原油	高騰	事件	問題
23	40) 事件	原油	高騰	事件	問題
24	41) 事件	原油	高騰	事件	問題
25	42) 事件	原油	高騰	事件	問題
26	43) 事件	原油	高騰	事件	問題
27	44) 事件	原油	高騰	事件	問題
28	45) 事件	原油	高騰	事件	問題
29	46) 事件	原油	高騰	事件	問題
30	47) 事件	原油	高騰	事件	問題
31	48) 事件	原油	高騰	事件	問題
32	49) 事件	原油	高騰	事件	問題
33	50) 事件	原油	高騰	事件	問題
34	51) 事件	原油	高騰	事件	問題
35	52) 事件	原油	高騰	事件	問題
36	53) 事件	原油	高騰	事件	問題
37	54) 事件	原油	高騰	事件	問題
38	55) 事件	原油	高騰	事件	問題
39	56) 事件	原油	高騰	事件	問題
40	57) 事件	原油	高騰	事件	問題
41	58) 事件	原油	高騰	事件	問題
42	59) 事件	原油	高騰	事件	問題
43	60) 事件	原油	高騰	事件	問題
44	61) 事件	原油	高騰	事件	問題
45	62) 事件	原油	高騰	事件	問題
46	63) 事件	原油	高騰	事件	問題
47	64) 事件	原油	高騰	事件	問題
48	65) 事件	原油	高騰	事件	問題
49	66) 事件	原油	高騰	事件	問題
50	67) 事件	原油	高騰	事件	問題
51	68) 事件	原油	高騰	事件	問題
52	69) 事件	原油	高騰	事件	問題
53	70) 事件	原油	高騰	事件	問題
54	71) 事件	原油	高騰	事件	問題
55	72) 事件	原油	高騰	事件	問題
56	73) 事件	原油	高騰	事件	問題
57	74) 事件	原油	高騰	事件	問題
58	75) 事件	原油	高騰	事件	問題
59	76) 事件	原油	高騰	事件	問題
60	77) 事件	原油	高騰	事件	問題
61	78) 事件	原油	高騰	事件	問題
62	79) 事件	原油	高騰	事件	問題
63	80) 事件	原油	高騰	事件	問題
64	81) 事件	原油	高騰	事件	問題
65	82) 事件	原油	高騰	事件	問題
66	83) 事件	原油	高騰	事件	問題
67	84) 事件	原油	高騰	事件	問題
68	85) 事件	原油	高騰	事件	問題
69	86) 事件	原油	高騰	事件	問題
70	87) 事件	原油	高騰	事件	問題
71	88) 事件	原油	高騰	事件	問題
72	89) 事件	原油	高騰	事件	問題
73	90) 事件	原油	高騰	事件	問題
74	91) 事件	原油	高騰	事件	問題
75	92) 事件	原油	高騰	事件	問題
76	93) 事件	原油	高騰	事件	問題
77	94) 事件	原油	高騰	事件	問題
78	95) 事件	原油	高騰	事件	問題
79	96) 事件	原油	高騰	事件	問題
80	97) 事件	原油	高騰	事件	問題
81	98) 事件	原油	高騰	事件	問題
82	99) 事件	原油	高騰	事件	問題
83	100) 事件	原油	高騰	事件	問題

Fig. 7 Information Extraction Table (Crude Oil)

とがわかる。情報抽出表は、単語集計結果で出現した単語が実際にどういう文中に出現していたのか、どういう単語とともにその単語が多く出現していたかを確認するのに役立つ。Fig. 2 の Focus terms の欄に「原油」を入力すると「事件」「問題」という一般用語を除いて考えると、「原油」「価格」「高騰」が並んで頻度高く出現していることがわかる。このことから、原油は、価格の高騰の文脈で被験者により関心を持たれている事柄であることがわかる。

次に、アンケートの「作業の基準」の項目のデータを対象に情報抽出表を作成した。Focus term は指定せずに実行した。「作業の基準」の項目のデータを対象に実行するには、Fig. 2 の画面で、上部の単語集計対象(形態素解析対象)項目の選択で、「事件」の項目のチェックを外して、「作業の基準」の項目にチェックをしてから「情報

	K	L	M	N	O	P	Q
1. 作業の基準	自分	興味	生活	記事	重要	身近	
2. コピー	248	126	22	26	26	26	20
3. 出現回数	138	58	38	26	38	38	30
4. 出現回数	138	58	41	38	36	43	36
5. 自分の興味・関心のある記事と順番に並べ	自分	興味	生活	記事	重要	身近	
6. 自分の記事の中で自分に興味があると思う順番に並べ	自分	興味	生活	記事	重要	身近	
7. 最初から、記事の内容に目を通し、自分の興味がある順番に並べ	自分	興味	生活	記事	重要	身近	
8. 大抵的な重要度と自分の興味のある順番、未決の予定なのか	自分	興味	生活	記事	重要	身近	
9. 関心や興味があった順番を自分優先で決めた	自分	興味	生活	記事	重要	身近	
10. 自分に直接関係があるかどうか、又は興味がある順番	自分	興味	生活	記事	重要	身近	
11. 自分の興味がある順番、および古い記事に注目した	自分	興味	生活	記事	重要	身近	
12. 自分の興味のある順番で選んだ	自分	興味	生活	記事	重要	身近	
13. 自分にとって興味がある順番	自分	興味	生活	記事	重要	身近	
14. 自分に興味がある順番	自分	興味	生活	記事	重要	身近	
15. 自分の興味のあること、自分の身の周りのこと、直接関係あること	自分	興味	生活	記事	重要	身近	
16. 自分の興味のある話題を感覚的な順番で並べた	自分	興味	生活	記事	重要	身近	
17. 自分に興味のある順番に並べた	自分	興味	生活	記事	重要	身近	
18. 自分にかかわること、興味があること、順番をきめたように思い	自分	興味	生活	記事	重要	身近	
19. 自分の生活に関わってくる記事、興味のある記事と基準に並べ	自分	興味	生活	記事	重要	身近	
20. 自分の興味のある内容、自分の生活とかわりがありそうな内容	自分	興味	生活	記事	重要	身近	
21. 自分の生活に直接関係する内容は、順位が上にはいりやすかった	自分	興味	生活	記事	重要	身近	
22. 自分の興味のある記事や社会的に重要な記事の順番を自分	自分	興味	生活	記事	重要	身近	
23. 自分の興味関心の深いもの、近頃に関心があるもの	自分	興味	生活	記事	重要	身近	
24. 自分が関心を持っているかどうか、興味があるかどうかという事、自分	自分	興味	生活	記事	重要	身近	
25. 経済的豊乏(1)経済的豊乏(2)自分が興味ある内容に関	自分	興味	生活	記事	重要	身近	
26. 自分の興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	
27. 自分の興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	
28. 自分が興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	

Fig. 8 Information Extraction Table (Criteria of Selection of Important Documents for Each Subject)

抽出表」の「再集計」のボタンを押して実行する。実行結果を Fig. 8 に示す。アンケートの「作業の基準」の項目では、記事を自分が思う重要と思う順番に並べ替える作業をしてもらったがその作業をする際にどういう基準でその作業をしたかを自由回答文で書いてもらった箇所である。Fig. 8 から、記事の並べ替えの基準は、主に、自分の興味の順番に並べ替えているということがわかる。また、生活に関わりの深いもの、身近なものを重要と考えて並べ替えをしたことがわかる。情報抽出表は、自由回答文の分析において、アンケート回答者がどう回答しているかを手軽に大雑把につかむことに役立つ。

また、重要な記事を提示するシステムで要望する機能に関するアンケート項目に対して情報抽出表を作成した。「記事」「機能」「ニュース」「システム」「興味」「事件」「情報」「選択」「表示」という単語はこの分析では不要であるのでそれらは事前に分析から取り除く設定を行った。この設定は、単語集計の「単語セット」に分析に利用する単語だけをセットするか、「同義語・削除辞書」で削除する単語を登録すると行える。このようにして、情報抽出表の処理を行った。その結果を Fig. 9 に示す。Fig. 9 から、アンケート回答者が、「検索」の機能、「重要」な事柄を扱う機能、「ジャンル」「キーワード」を利用した機能、「自分」を主体に考慮した機能、「クリック」操作でできる機能、「個人」に応じた機能、「ソート」に関する機能を要望していることがわかる。Simpleminer では、こういった事柄を具体的に要望しているかを、実際の回答文と照らし合わせながら見ていくこともできる。

最後に、アンケートの感想の項目に対して情報抽出表を作成した。ここでは形容詞のみを利用して分析を行った。その結果を Fig. 10 に示す。Fig. 10 から、アンケートの項目が多くて長かったこと、また、難しいと回答者が感じていたことがわかる。また、回答者が良いと思っている事柄の情報もすぐに入手できる。2006 年度の新聞記事データを題材にしたアンケートも行っていたため、データが「古い」という回答もあったが、そういう回答

	K	L	M	N	O	P	Q
1. 作業の基準	自分	興味	生活	記事	重要	身近	
2. コピー	248	126	22	26	26	26	20
3. 出現回数	138	58	38	26	38	38	30
4. 出現回数	138	58	41	38	36	43	36
5. 自分の興味・関心のある記事と順番に並べ	自分	興味	生活	記事	重要	身近	
6. 自分の記事の中で自分に興味があると思う順番に並べ	自分	興味	生活	記事	重要	身近	
7. 最初から、記事の内容に目を通し、自分の興味がある順番に並べ	自分	興味	生活	記事	重要	身近	
8. 大抵的な重要度と自分の興味のある順番、未決の予定なのか	自分	興味	生活	記事	重要	身近	
9. 関心や興味があった順番を自分優先で決めた	自分	興味	生活	記事	重要	身近	
10. 自分に直接関係があるかどうか、又は興味がある順番	自分	興味	生活	記事	重要	身近	
11. 自分の興味がある順番、および古い記事に注目した	自分	興味	生活	記事	重要	身近	
12. 自分の興味のある順番で選んだ	自分	興味	生活	記事	重要	身近	
13. 自分にとって興味がある順番	自分	興味	生活	記事	重要	身近	
14. 自分に興味がある順番	自分	興味	生活	記事	重要	身近	
15. 自分の興味のあること、自分の身の周りのこと、直接関係あること	自分	興味	生活	記事	重要	身近	
16. 自分の興味のある話題を感覚的な順番で並べた	自分	興味	生活	記事	重要	身近	
17. 自分に興味のある順番に並べた	自分	興味	生活	記事	重要	身近	
18. 自分にかかわること、興味があること、順番をきめたように思い	自分	興味	生活	記事	重要	身近	
19. 自分の生活に関わってくる記事、興味のある記事と基準に並べ	自分	興味	生活	記事	重要	身近	
20. 自分の興味のある内容、自分の生活とかわりがありそうな内容	自分	興味	生活	記事	重要	身近	
21. 自分の生活に直接関係する内容は、順位が上にはいりやすかった	自分	興味	生活	記事	重要	身近	
22. 自分の興味のある記事や社会的に重要な記事の順番を自分	自分	興味	生活	記事	重要	身近	
23. 自分の興味関心の深いもの、近頃に関心があるもの	自分	興味	生活	記事	重要	身近	
24. 自分が関心を持っているかどうか、興味があるかどうかという事、自分	自分	興味	生活	記事	重要	身近	
25. 経済的豊乏(1)経済的豊乏(2)自分が興味ある内容に関	自分	興味	生活	記事	重要	身近	
26. 自分の興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	
27. 自分の興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	
28. 自分が興味のあること、または身近な関心がある内容	自分	興味	生活	記事	重要	身近	

Fig. 9 Information Extraction Table (Request Submitted by Each Subject to the Systems Handling Important Documents)

	L	M	N	O	P	Q	R
1. (1)感想	多い	多い	多い	多い	多い	多い	多い
2. スコア	116	108	38	20	18	12	
3. 出現回数	58	54	12	10	8	6	
4. 出現回数	58	54	12	11	6	6	
5. 質問が多くて良いので、最後はどうでもよくなってしまいました。先に要望、その	多い	多い	多い	多い	多い	多い	多い
6. 記事の文量が多すぎ、読めなくて、途中でやめて、後でそこから拾うことが多い	多い	多い	多い	多い	多い	多い	多い
7. 文量が多すぎる。多量なものはわかるが、必要がないものも多かった	多い	多い	多い	多い	多い	多い	多い
8. あまりに良いので、疲れました。似たような記事が多く、大事でないもの同士	多い	多い	多い	多い	多い	多い	多い
9. あまりに長すぎて、また同じような内容が多くて、途中で寝てしまった	多い	多い	多い	多い	多い	多い	多い
10. 長い、多すぎ	多い	多い	多い	多い	多い	多い	多い
11. 問題が多くて多すぎ、もっとどこに何回かに分けたほうが良いと思う	多い	多い	多い	多い	多い	多い	多い
12. 興味のない記事も多い中、重要度で並びかえるシステムは良いと思うけど、	多い	多い	多い	多い	多い	多い	多い
13. 古いニュースが多かったので、思い出すのが大変だった	多い	多い	多い	多い	多い	多い	多い
14. 60分は長すぎる。意識が低くなりやすいので、20分程度にして、理由等を記せる	多い	多い	多い	多い	多い	多い	多い
15. 今回のアンケートは非常に長く手間が掛かりました。それなりのポイント 付与が欲しい	多い	多い	多い	多い	多い	多い	多い
16. 長すぎる。今どれ位アンケートが進んでいるのかわかるようにして欲しい	多い	多い	多い	多い	多い	多い	多い
17. 少し長いです。小さい子を抱えながら答えているので中断機能がほしいものを何回	多い	多い	多い	多い	多い	多い	多い
18. システム自体は有意義なものだともうがニュース記事の重要度は個々の主題に大	多い	多い	多い	多い	多い	多い	多い
19. 何が重要かは、人の価値観で 統一して欲しい、難しいと思いますが、私達は、いろいろ	多い	多い	多い	多い	多い	多い	多い
20. 今回のアンケートは大変多くて、なかなか難しかったので疲れました...	多い	多い	多い	多い	多い	多い	多い
21. 非常に良いと思う。ただし、関連するものをまとめるのと見やすくして欲しい(システムについて)	多い	多い	多い	多い	多い	多い	多い
22. 良いシステムだと思います。推定、でもアンケートはちょっとしんどかったです。バーで選	多い	多い	多い	多い	多い	多い	多い
23. 国内外の出来事が分かれていたのも、こういうことがあったと思う出すことが出来	多い	多い	多い	多い	多い	多い	多い
24. 古い記事に対してはどうしても重要性が低くなってしま	多い	多い	多い	多い	多い	多い	多い
25. 古いニュースもまだ見ても興味がない、興味がなくな	多い	多い	多い	多い	多い	多い	多い

Fig. 10 Information Extraction Table (Responses of the Subjects to the Questionnaire)

があることも、情報抽出表を利用すると容易に見出すことができる。

6. クロス分析

次にクロス分析を実行した。クロス分析では二つの事柄を指定して分析を行う。Fig. 2 で、「地域」と「事件」でクロス分析を行う。クロス集計項目 1(横軸)に「地域」、クロス集計項目 2(縦軸)に「事件」を選択して、クロス集計の「再集計」ボタンを押す。そうすると、Fig. 11 に示す結果が得られる。この図は、何人のどの地域の人がどういう単語を含む事件について興味を持っているかを示す。Fig. 11 の結果に対して、双対尺度法[7]を実行すると、Fig. 12 の結果が得られる。Simpleminer には、双対尺度法の機能は付いていない。双対尺度法の実行は、上田データマイニング塾(<http://www.datamining.jp/>)から別途購入したツールを利用した。Fig. 12 から、香川の事件を言及した四国地方の人はむしろ少なく、近畿地方

	A	B	C	D	E	F	G	H	I
1	■関東地方	■中部地方	■近畿地方	■九州地方	■東北地方	■北海道	■中国地方	■四国地方	
2	事件	44	24	18	13	4	6	3	1
3	問題	31	17	16	7	4	1	4	0
4	行方	26	9	11	9	3	4	2	1
5	不明	30	11	10	8	2	4	1	1
6	香川	24	12	13	7	1	4	3	1
7	祖母	23	12	5	7	3	3	2	1
8	原油	12	7	3	5	6	3	0	1
9	殺人	14	4	6	5	3	1	2	0
10	高騰	8	6	4	4	3	2	1	2
11	温暖	21	4	0	1	2	0	1	1
12	プライム	16	3	7	1	1	0	2	0
13	食品	13	5	6	3	2	0	0	0
14	地球	19	5	1	1	1	0	1	1
15	防衛	11	4	6	3	1	1	1	2
16	サブ	15	3	6	1	1	0	2	0
17	年金	12	8	3	2	0	0	2	0
18	ガソリン	8	9	2	1	2	0	2	1
19	偽装	9	4	4	4	0	0	1	0
20	値上げ	6	6	2	1	2	3	1	1
21	失踪	5	7	1	3	1	1	1	1

Fig. 11 Cross Analysis

Fig. 12 Result of the Dual Scaling Method

や九州地方の人が多く香川の事件に言及していることがわかる。また、東北地方の人は特に原油高を問題に考えているようであり、関東地方や中国地方の人は年金問題に興味があるようである。

最近Excelや簡単なツールを用いることで種々の統計分析ができるようになっており、テキストデータから数値データに変換することができれば、上述の双対尺度法など、種々の数値解析手法が利用できる^{8),9),10)}。Simpleminerは、テキストデータから数値データを生成するところで役立っていることになる。

7. ソートグラフ

最後にソートグラフの機能を示す。これは片方が数値である場合のクロス表のデータの分析に利用できる。図2で、クロス集計項目1(横軸)に「年齢(歳代)」、クロス集計項目2(縦軸)に「事件」を選択して、ソートグラフの「再集計」ボタンを押して、ソートグラフを作成する。作成したソートグラフの例をFig. 13に示す。図で等高線の高さがその年齢でその単語を含む事件を言及した人の数を意味する。ソートグラフの横軸は年齢で、右側の単語は、「事件」の自由回答文に出現した単語である。この単語につけている一つ目の数字は、その単語を使った

Fig. 13 Sort Graph (The Numbers in the Parentheses Indicate the Total Number of Subjects and the Average Value of Their Ages.)

人の合計で、二つ目の数字はその単語が多く出現している年齢の平均を示す。二つ目の数字は厳密には、年齢の平均値と最頻値と中央値の平均である。この値の小さいものから順に上から表示している。システムは等高線グラフを描きやすいcsv形式のファイルを出力する。ユーザはそのファイルからExcelを使って簡単に等高線グラフを描くことができる。等高線の高いところを見ることで、どの年齢の人がどの事件に多く興味を持っているかがわかる。「地球温暖化」「サブプライム」「防衛」の問題については、比較的少しだけ歳のいった平均35歳から40歳程度の人が興味を持っており、それ以外の問題は、30歳前後の人が興味を持っていることがわかる。ソートグラフでは文献¹¹⁾を参考にして等高線表示を利用している。

次に性別による違いを調べてみる。Fig. 2で、クロス集計項目1(横軸)に「性別」、クロス集計項目2(縦軸)に「事件」を選択して、ソートグラフの「再集計」ボタンを押して、ソートグラフを作成する。その結果、Fig. 14の結果を得る。図の横軸では男性を1、女性を2としている。図の上の方にある、「防衛」「サブプライム」「原油」は男性の方で多く言及されている事柄であり、図の下の方にある、「香川」の事件、「年金」「食品偽装」「値上げ」は、女性の方で多く言及されている事柄である。男性、女性という二分類の事柄は、それぞれを0と1として数値化することでソートグラフで分析できるデータになる。性別に基づく分析もソートグラフで実行することができる。

Fig. 14 Sort Graph on the Basis of Gender

Fig. 15 Sort Graph on the Basis of Whether a Subject Has a Child

子供の有無による分析を行うと、Fig. 15 の結果を得る。横軸では子供のいない人を 1、子供のいる人を 2 としている。子供のいない人は、「サブプライム」「殺人」の事柄に興味を持ち、子供のいる人は、「年金」「地球温暖化」「防衛」の事柄に興味を持っていることがわかる。

8. おわりに

本稿では、われわれが開発した簡易テキストマイニングシステム Simpleminer について紹介した。一般的なテキストマイニングシステム³⁾を持つ、単語の頻度分析、クロス分析が可能である。そのうえ、情報抽出表とソートグラフと呼ぶ、他のシステムにない新規な技術を利用した分析も可能である。

具体例として、アンケートデータを対象にテキストマイニングを行った結果を示したが、本システムは、論文書誌情報を対象にすることで、学会動向を調べることができる^{1),2),4),5),6)}。

参考文献

- 1) 村田真樹, 金丸敏幸, 一井康二, 馬青, 白土保, 井佐原均, 簡易テキストマイニングシステム simpleminer, 第 15 回インタラクティブシステムとソフトウェアに関するワークショップ, (2007), pp. 103-104.
- 2) 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), p. 1093-1096.
- 3) 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, 瀬上美喜, 事例で学ぶテキストマイニング, (共立出版, 2008).
- 4) 村田真樹, 一井康二, 馬青, 白土保, 井佐原均, 過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査, 言語処理学会第 11 回年次大会, (2005).
- 5) Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru, and Hitoshi Isahara, Trend survey on Japanese natural language processing studies over the last decade, The Second International Joint Conference on Natural Language Processing, Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts, (2005).
- 6) 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 井佐原均, 過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査, (2007), 言語処理学会ホームページ (<http://www.nak.ics.keio.ac.jp/NLP/trend-survey.htm>).
- 7) 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 松岡雅裕, 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, 言語処理学会第 14 回年次大会, (2008), p. 907-910.
- 8) 上田太一郎, 刈田正雄, 本田和恵, 実践ワークショップ Excel 徹底活用多変量解析, (秀和システム, 2003).
- 9) 上田太一郎, 高橋玲子, 村田真樹, 瀬上美喜, 藤川貴司, 近藤宏, 上田和明, Excel で学ぶ時系列分析と予測, (オーム社, 2006).
- 10) 上田太一郎, 近藤宏, 瀬上美喜, 末吉正成, 村田真樹, Excel でかんたん統計分析 --- [分析ツール]を使いこなそう! ---, (オーム社, 2007).
- 11) 谷口敏夫, 『人工知能と人間/長尾真』のテキスト可視化 ---KT システムによるテキスト分析---, (<http://www.koka.ac.jp/taniguti96M/0/30/2000/Note2Nagao/Note20000409.htm>, 2000).