3B2-3

テキストマイニングシステム Simpleminer の利用

Application of text-mining system "Simpleminer"

村田 真樹[*1] Masaki Murata
金丸 敏幸[*2] Toshiyuki Kanamaru
一井 康二[*3] Koji Ichii
馬 青[*4,*1] Qing Ma
白土 保[*1] Tamotsu Shirado
井佐原 均[*1] Hitoshi Isahara

[*1]独立行政法人 情報通信研究機構 National Institute of Information and Communications Technology
[*2]京都大学 Kyoto University
[*3]広島大学 Hiroshima University
[*4]龍谷大学 Ryukoku University

We have developed a simple text-mining system called "Simpleminer." This system works on Windows machines. It can be used, for example, to analyze questionnaires that include text and trends in journal titles. It can determine the usage frequencies of words and create multiway tables, which are the fundamental functions of text-mining systems. In addition, Simpleminer has two unique functions: it can generate information extraction tables and sort graphs. An information extraction table shows whether a sample of text data includes words with high usage frequencies. A sort graph shows the changes in the frequencies of word usage over a certain period of time shown in the figure. A useful feature of this graph is that words that appeared more frequently in the past are displayed higher on the list. This enables users to recognize such words more effectively. In this paper, we have used bibliographic information on conference papers of the Japanese Society for Artificial Intelligence (JSAI) as the sample data for text mining. We have carried out a trend survey of the conference.

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，昔多かった傾向，最近多くなった傾向を簡便につかむことができる機能である．

本稿では特に人工知能学会全国大会の書誌データを利用して Simpleminer の紹介を行う．このため，人工知能学会全国大会の傾向分析も行っている．

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あり，それをクリックすることで，図 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 を起動すると，図 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図 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，図 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．図 2 中の「表示」または「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，人工知能学会全国大会の 1992 年分から 2006 年分までの 15 年分の書誌情報を与えた．(関連研究としては言語処理学会の動向を調査したものがある [4, 5, 6]．) タイトルの部分を対象としてテキストマイニング処理を行った．

図 1: インストール画面

4. 単語集計

単語集計機能を使うと，図 4 の結果を得る．どの単語が何個の論文のタイトルに出現したかを示している．図 4 ではすべての品詞の単語を示した．これは，図 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，ひら仮名語削除の機能

連絡先: 村田 真樹，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，〒 619-0289 京都府相楽郡精華町光台 3-5，TEL: 0774-98-6833, FAX: 0774-98-6961, murata@nict.go.jp.

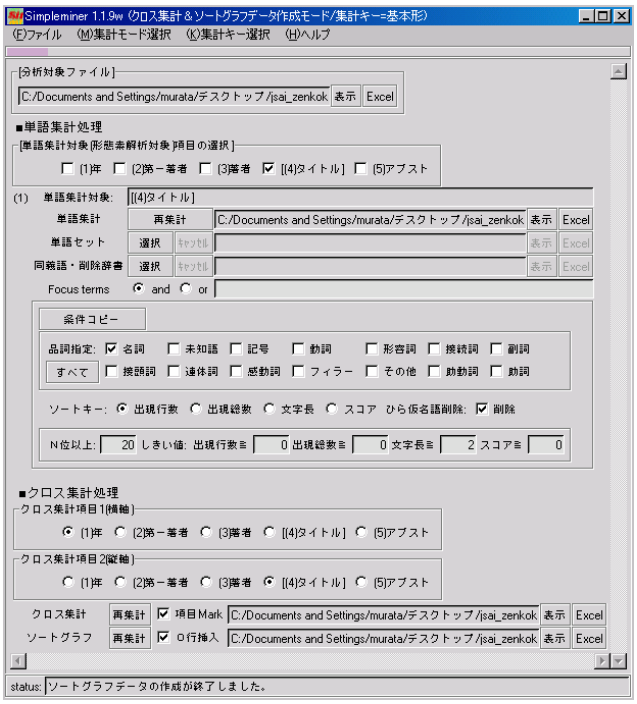

図 2: Simpleminer の画面

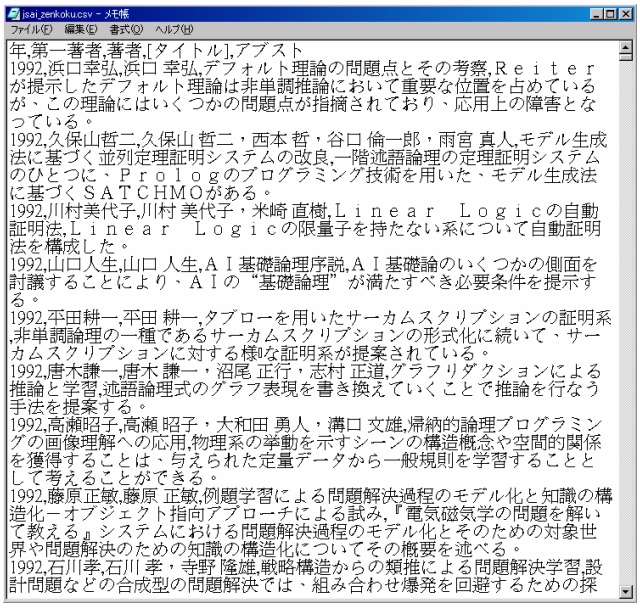

図 3: 入力の csv 形式のファイルの例

	A	B	C	D
1	出現形	基本形	品詞	出現行数
2	の	の	助詞	2601
3	を	を	助詞	1025
4	た	た	助動詞	866
5	と	と	助詞	751
6	システム	システム	名詞	648
7	に	に	助詞	583
8	における	における	助詞	550
9	による	による	助詞	513
10	し/さ/する/	する	動詞	495
11	用い/用い/	用いる	動詞	454
12	的	的	名詞	437
13	基づく/基づ	基づく	動詞	430
14	支援	支援	名詞	404
15	学習	学習	名詞	387
16	化	化	名詞	347
17	知識	知識	名詞	323
18	情報	情報	名詞	296
19	ため	ため	名詞	289
20	エージェント	エージェント	名詞	260
21	モデル	モデル	名詞	248

図 4: 単語集計の例

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．図 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 文字以上のもののみを対象とし，ひら仮名語も不要として対象外として単語集計をしてみる．これは，図 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，図 5 の結果を得る．図 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図 5 を見ることで，データの大雑把な傾向をつかめる．システムの研究が多いことがわかる．また，支援に関する研究が多いことがわかる．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 (図 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,「翻訳」という単語を含む論文だけを対象に実行した．図 2 の Focus terms の欄に「ロボット 移動」と入力すると「ロボット」と「移動」という単語を両方含む論文だけを使った処理を実行できる．さらに,「情報」「作成」「考慮」という単語をこの分析では不要であるのでそれ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」に「情報」を削除する単語として登録すると行える．このようにして，情報抽出表の処理を行った．その結果を図 6 に示す.「ロボット」と「移動」という単語を含む論文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．また各論文タイトルの右側の欄にはその列の単語をタイトルに含んでいればその単語を表示している．論文タイトルもソートしており，なるべく左側の単語を含むタイトルの順に表示している．

	A	B	C	D
1	出現形	基本形	品詞	出現行数
2	システム	システム	名詞	648
3	支援	支援	名詞	404
4	学習	学習	名詞	387
5	知識	知識	名詞	323
6	情報	情報	名詞	296
7	エージェント	エージェント	名詞	260
8	モデル	モデル	名詞	248
9	利用	利用	名詞	194
10	生成	生成	名詞	185
11	環境	環境	名詞	176
12	構築	構築	名詞	172
13	手法	手法	名詞	171
14	対話	対話	名詞	157
15	データ	データ	名詞	142
16	問題	問題	名詞	141
17	構造	構造	名詞	141
18	推論	推論	名詞	138
19	設計	設計	名詞	132
20	獲得	獲得	名詞	131
21	ロボット	ロボット	名詞	128

図 5: 名詞のみによる単語集計の例

[(4)タイトル]	移動	ロボット	行動	自律	環境	視覚	地図
	64	128	16	12	10	8	6
	32	32	8	6	5	4	3
	32	32	9	6	5	5	3
行為に基づく環境モデリングのための移	移動	ロボット	行動		環境		
画像運動情報に基づく単眼視覚移動ロ	移動	ロボット	行動			視覚	
ステレオ視覚の不確かさのモデリングと	移動	ロボット	行動			視覚	
GAによる小型移動ロボットの行動制御	移動	ロボット	行動				
多目的戦略を用いたGPのTreesizecor	移動	ロボット	行動				
状態・行動モデルに基づく移動ロボット	移動	ロボット	行動				
移動ロボットによるランドマーク間巡回	移動	ロボット	行動				
プラニングと行動の一貫性を考慮した移	移動	ロボット	行動				
知能ロボットの自律移動における経路	移動	ロボット		自律			地図
画像情報を用いた自律移動ロボットの	移動	ロボット		自律			
ベイジアンネットによる情報統合を用い	移動	ロボット		自律			
知能ロボットの自律移動のための実画	移動	ロボット		自律			
遺伝的プログラミングを用いた自律移動	移動	ロボット		自律			
遺伝的アルゴリズムによる自律移動ロ	移動	ロボット		自律			
視覚と触覚の統合に基づく移動ロボット	移動	ロボット			環境	視覚	
情報理論に基づく環境、身体性を考慮	移動	ロボット			環境		地図
部分観測環境における移動履歴情報	移動	ロボット			環境		地図
広域環境情報を利用する移動ロボット	移動	ロボット			環境		

図 6: 情報抽出表の例

この表では各論文タイトル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．図 6 から，移動ロボットの研究には，行動を対象とするもの，自立移動を特に扱っているもの，環境を対象とするものの三種類の研究アプローチが主にあることがわかる．その他，視覚や地図を扱う研究も存在していることがわかる．情報抽出表は，各データがどういった単語を含んでいるかを簡便な表の形で見るのに役立つ．

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．図 2 で,「タイトル」と「第一著者」でクロス分析を行う．クロス集計項目 1(横軸) に「タイトル」，クロス集計項目 2(縦軸) に「第一著者」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，図 7 に示す結果が得られる．この図は，どの著者がどういう単語を含む論文を何件発表したかを示す．図 7 の結果に対して，双対尺度法 [7] を実

	A	B	C	D	E	F	G	H	I	J	K
1		■システム	■支援	■学習	■知識	■情報	■エージェント	■モデル	■利用	■生成	■環境
2	溝口文雄	8	1	5	0	1	4	1	0	0	0
3	砂山渡	9	7	1	0	3	0	1	1	1	0
4	西山裕之	6	0	0	0	0	6	0	0	0	1
5	吉岡真治	3	2	0	5	1	0	1	1	1	1
6	鷹岡亮	3	3	2	1	0	2	2	1	0	2
7	武田英明	1	4	0	5	2	1	1	0	0	2
8	藤本和則	5	4	0	3	6	0	0	0	0	0
9	宮原哲浩	2	0	1	2	2	0	0	0	1	0
10	笹島宗彦	3	0	0	2	0	0	4	2	1	1
11	角康之	3	6	0	1	1	2	1	1	1	0

図 7: クロス分析の例

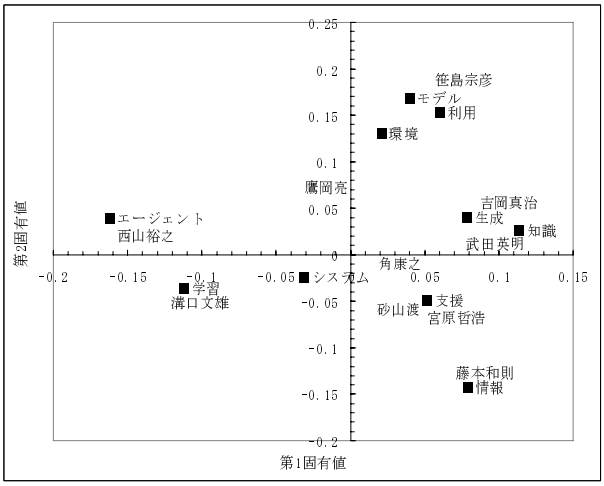

図 8: 双対尺度法の結果

行すると，図 8 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．図 8 から，吉岡真治先生，武田英明先生は知識に関する研究が多く，溝口文雄先生が学習の研究，西山裕之先生がエージェントの研究が多いことがわかる．

最近は 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落すことができれば，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 [7, 8, 9]．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」，クロス集計項目 2(縦軸) に「タイトル」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図 9 に示す．図で等高線の高さが論文の数を意味する．ソートグラフの横軸は発表年で，右側の単語は，タイトル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，その単語を含む論文の合計で，二つ目の数字はその単語が多く出現している発表年の平均を示す．二つ目の数字は厳密には，発表年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

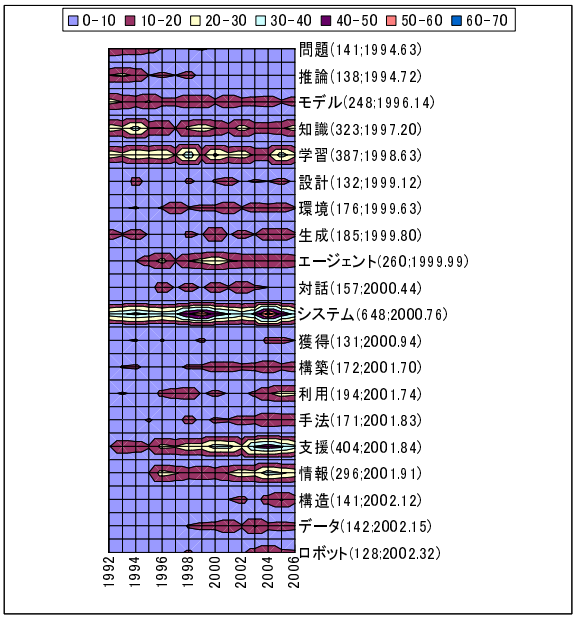

図 9: ソートグラフの例 (等高線の高さは件数を示す．図中の各単語に付与した二つの数字は左が合計件数を右が発表件数の巻数の平均を意味する (厳密な定義は本文を参照のこと))

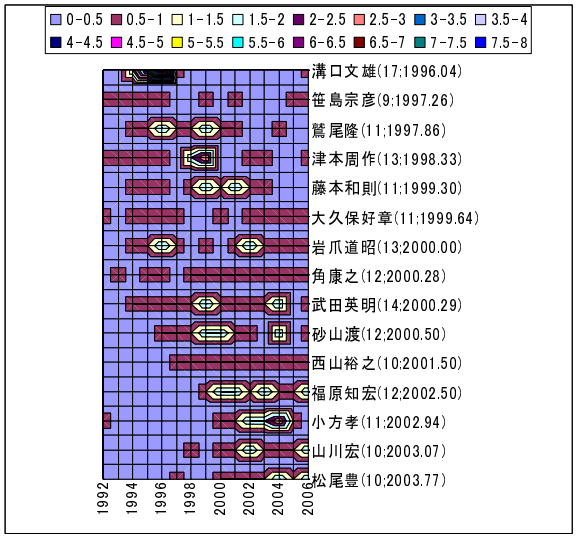

図 10: 第一著者のソートグラフ

線の高いところを見ることで，どの巻でどの単語を含む論文が多かったかがわかる．昔は「推論」「知識」の研究が多かったが，最近は「支援」「データ」「ロボット」という研究が多いことがわかる．ソートグラフでは文献 [10] を参考にして等高線表示を利用している．

入力データを発表件数上位 15 位までの人の書いた論文の書誌情報にして，クロス集計項目 1(横軸) に「年」，クロス集計項目 2(縦軸) に「第一著者」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した結果を図 10 に示す．図から，どの人がいつどのくらい発表しているかがすぐわかる．また，昔発表が多かった人や最近多い人もわかる．

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] 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．

具体例として人工知能学会の全国大会の書誌情報を対象にテキストマイニングを行った結果を示したが，本システムを利用することでもっと多くの学会動向を簡便に調べることができる．また，本システムは，動向調査のみならず，自由記述のアンケートデータの分析にも利用できる．

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).

[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, 渕上美喜, 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.html).

[7] 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).

[8] 上田太一郎, 高橋玲子, 村田 真樹美喜, 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).

[9] 上田太一郎, 近藤宏, 渕上美喜, 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).

[10] 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).