コンピュータによる文書群からの有用な情報の取り出し

大学院工学研究科　教授　村田　真樹

１　はじめに

コンピュータは、様々なことに役立つ。本講義では、コンピュータを利用して、文書群からの有用な情報の取り出し技術について紹介する。電子化テキストの増加にともない、電子テキストからの有用な情報の取り出しの重要性が高まっている。

テキストデータ中から有用な情報を取り出すことを、**テキストマイニング**と呼ぶ。これに対して、データ中から有用な情報を取り出すことを**データマイニング**と呼ぶ。テキストマイニングはデータマイニングの一種である。本講義で紹介する技術は、テキストマイニング技術 1)に相当する。

２　データマイニングとテキストマイニング

データマイニングとは、データ（data）を発掘（mining）して、宝物を見つける手法・プロセスのことである。データマイニングの例として、購入データの分析（傾向の分析）がある。例えば、商品の購入履歴のデータを収集し、そのデータを分析（データマイニング）することで、「商品 A を買った人は、商品 B も買う。」といった傾向を知ることができる。このような場合、商品 A と商品 B を近くに配置して販売すると商品の売り上げが上昇する可能性があり、そのような配置をとるということがなされる場合がある。

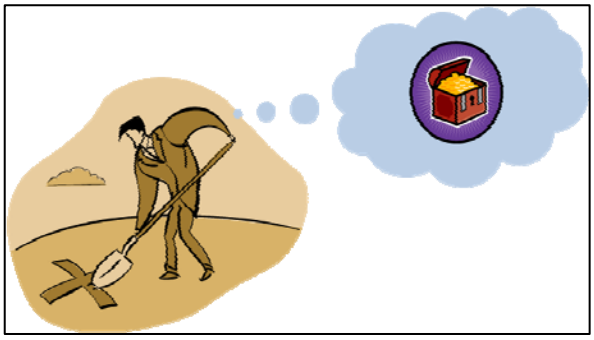

図１．データマイニングの様子

テキストマイニングは、データマイニングの一種であり、対象のデータが数値などだけでなく、文・テキストを対象とする。テキストマイニングの例として、新聞の社説のテキストデータを分析し、社会の動向を調べるというものがある。この社会動向調査については本講義内でも紹介する。

他のテキストマイニングの例としては、商品に関するアンケートのデータ（自由回答で、文で回答してもらったアンケートのデータ）を分析するというものがある。そのアンケートを分析することにより、商品 A を買った人は、商品 A にどういう印象を持っているかの情報を得て、その情報を商品 A の改良に役立てるということがある。

テキストマイニングについては、「事例で学ぶテキストマイニング、 上田太一郎監修, 村田真樹ほか著、 共立出版, 2008.1」という書籍を、筆者が主たる著者として執筆している。テキストマイニングについて、より詳細を知りたい場合はこの書籍も読まれることをお奨めする。この書籍では、種々のテキストマイニング事例を詳述している。また、なかなか目にすることの少ない高価なマイニングソフトの利用事例も紹介している。また、テキストマイニングの背景に存在する、テキストを処理する方法（自然言語処理技術）についても簡明に説明している。

本書籍では、テキスト処理に役立つツールやデータの紹介も行っている。（本講義の社会動向調査は本書籍の 4.1 節に対応する。）

図２．テキストマイニングの書籍の紹介（出典：事例で学ぶテキストマイニング、 上田太一郎監修、村田真樹ほか著、共立出版、 2008.1）

本講義ではテキストマイニングの例として、以下の三点について紹介する。

１．社会動向調査
２．数値、固有表現抽出に基づく情報の取り出し
３．研究者と研究分野の変遷情報の自動抽出

３　社会動向調査

ここでは、新聞の社説タイトルのデータをもとにした社会動向の調査について紹介する。１９９１年から２００５年までの合計１５年間の毎日新聞の社説のタイトルのデータ（１００５９個のデータ）を利用する。

社説のタイトルには「一国平和主義からの離脱へ」といったものがあり、それに対して形態素解析と呼ばれる技術を使うと、図３のように、文を単語に分割し、品詞を特定できる。

出現形	読み	見出し表記（基本形）	品詞
一国	イッコク	一国	名詞-副詞可能
平和	ヘイワ	平和	名詞-形容動詞語幹
主義	シュギ	主義	名詞-一般
から	カラ	から	助詞-格助詞-一般
の	ノ	の	助詞-連体化
離脱	リダツ	離脱	名詞-サ変接続
へ	へ	へ	助詞-格助詞-一般

図３．　形態素解析結果（出典：事例で学ぶテキストマイニング、上田太一郎監修、村田真樹ほか著、共立出版、2008.1、4.1 節。）

文から単語を取り出し、単語の頻度を使って傾向の分析をする。社説に出現した単語の頻度を調べると図４の結果を得る。

改革	662	2	247	責任	176	制度	140
化	429	者	237	国民	176	白書	135
政治	425	問題	236	時代	171	人	134
米	396	政策	224	委員	160	支援	133
0	345	経済	221	関係	157	世界	132
日	344	選	213	必要	154	力	131
年	275	首相	208	金融	154	論説	126
社会	274	事件	208	会談	151	情報	126
1	271	視点	201	3	151	憲法	124
法	262	5	195	外交	148	疑惑	121
国際	262	対策	190	核	145	大統領	118
日本	260	選挙	184	国	142		
的	260	国会	184	環境	142		

図４．社説における単語頻度分析（出典：事例で学ぶテキストマイニング、上田太一郎監修、村田真樹ほか著、共立出版、　2008.1、4.1 節。）

表の上位に、「改革」「政治」「社会」「国際」という単語がある。社説には「改革」「政治」「社会」「国際」に関連する内容が多く記述されていることがわかる。

このように単語の頻度をもとめるという簡単なことでもだいたいの傾向を知ることができる。

次に、社説のタイトルから、人名の単語を取り出し、その単語の頻度を時系列的に分析する事例について紹介する。人名の単語の頻度を時系列的にもとめてグラフ化したものを図５に示す。図において各人名の右につけている二つの数字は、一つ目は、その人名の合計の頻度であり、二つ目はその人名が平均してどこの年に出現していたかを示す。このグラフの一つの大きな特徴は、上下方向を、この二つ目の値（だいたい平均していつの年にその人名が出現していたか）によって並べかえているところにある。上の方が古い年で、下の方に新しい年を示している。これにより、古い時期に多く出現したものが上の方に集まり、最近出現したものが下の方に集まり、グラフを見て行う分析がしやすくなる。（自動的にこのように並べかえる方法は文献 1）にある。）

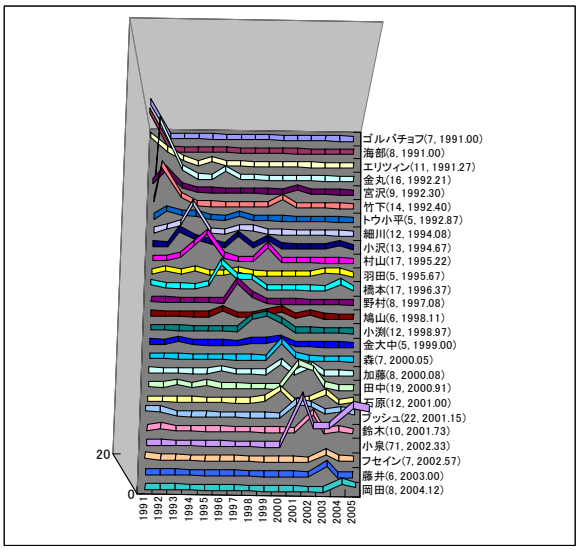

図５．人名の頻度の時系列的変化（出典：事例で学ぶテキストマイニング、上田太一郎監修、村田真樹ほか著、共立出版、2008.1、4.1 節。）

このグラフにより、「ゴルバチョフ」「海部」「エリツィン」は古い時期に多く出現し、「フセイン」「藤井」「岡田」は最近多く出現していることがわかる。

次に単語と人名を用いた分析について紹介する。人名と単語が同一の社説のタイトルに出現した回数を数えて集計した表を図６に示す。

同一のテキスト内に出現することを共起するという。図６のような表をクロス表と呼び、クロス表に基づく分析をクロス分析と呼ぶ。図６のようなデータに対して、双対尺度法と呼ばれる方法で分析した結果を図７に示す。双対尺度法により、よく似たデータを近くに、あまり似ていないデータを遠くに配置することができる。

図７より、「ブッシュ」「エリツィン」が「大統領」と近いこと、首相であった歴代の日本の政治家は「首相」と近いところに配置されていること、

「疑惑」と関連の深い「金丸」は「疑惑」の近くに配置されていることを確認できる。

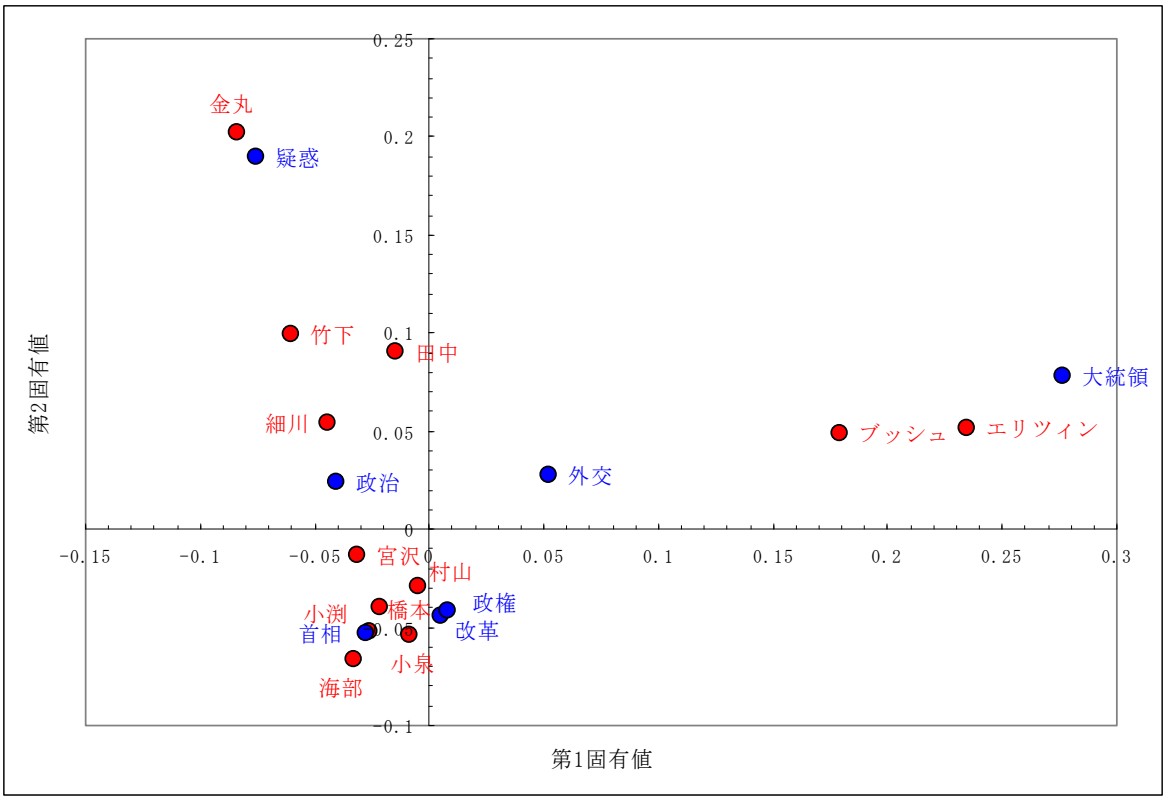

図７．双対尺度法に基づく分析結果（出典：事例で学ぶテキストマイニング、上田太一郎監修、村田真樹ほか著、共立出版、 2008.1、4.1 節。）

		人名に共起する単語						
		首相	政権	政治	改革	疑惑	大統領	外交
人名	小泉	22	14	8	22	0	0	4
	細川	2	3	5	2	4	0	0
	竹下	2	0	1	1	3	0	0
	村山	2	6	3	0	0	0	2
	橋本	3	5	4	1	0	0	0
	金丸	0	0	4	0	7	0	0
	ブッシュ	0	2	0	1	0	5	4
	田中	1	0	2	1	3	0	4
	宮沢	6	0	2	1	1	0	1
	小渕	4	3	2	0	0	0	0
	エリツィン	0	1	0	2	0	5	0
	海部	6	0	1	1	0	0	0

図６．人名と単語の共起頻度分析（出典：事例で学ぶテキストマイニング、上田太一郎監修、村田真樹ほか著、共立出版、2008.1、4.1 節。）

４　数値、固有表現抽出に基づく情報の取り出し

次に、テキストデータから、単位表現付きの数値情報と固有名詞を関連付けて抽出し、抽出したデータをグラフ化する技術を紹介する 2),3)。この研究は、筆者が（独）情報通信研究機構に在籍していたときに行った研究である。この研究は、（独）情報通信研究機構と（株）数理システムの共同研究により、 実用的なシステムも構築している。

そのシステムの動作例をここで紹介する。ガソリン価格に関する調査を例とする。実際の Web のニュース記事の一部を利用する。検索キーワード「ガソリン　リットル　円」を含むニュース記事を大量に収集する。記事の例を図８に示す。その記事から、数値情報や固有名詞を取り出す。

ガソリン価格が乱高下。＿＿新聞。 2008年7月24日。那覇近郊において、1リットル当たり百五十九円の小売価格。原油高騰による小売価格の上昇が止まらない。消費者も買い控えをする可能性あり。…

図８．一つの記事の例（概念図）

記事から取り出した数値情報や固有名詞を、図９のように表の形式で整理する。

テキスト	日	円	LOCATION
ガソリン価格が乱高下。＿新聞。 2008年7月24日。那覇	24	159	那覇近郊
記事2のテキスト	22	180	
記事3のテキスト	24	200	
・・・・	15	183	長野
・・・・	14	181	
・・・・	9	177	群馬
・・・・	2	180	石川
・・・・	14	181	中国地方
・・・・	21	120	中国
・・・・	7	116	米
・・・・	21	93	中国
・・・・	16	200	神奈川

図９．取り出した数値や固有名詞を整理した表

図９の表に対して、日付を横軸、円を縦軸にとってラベルに地名を設定してグラフを書くと図１０が得られる。

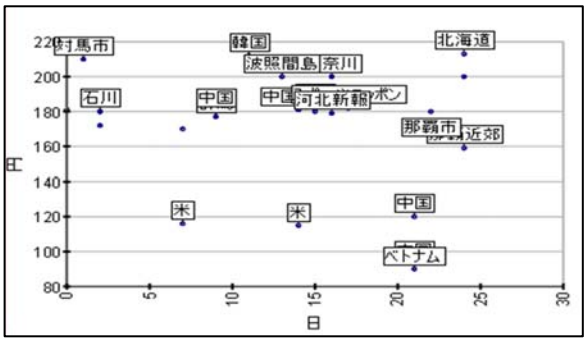

図１０．数値や固有名詞のグラフ化

本来的には数十におよぶ関連するニュース記事があったのだが、そこから主要な情報の円と日と地名を取り出して図１０ができている。数十におよぶ記事を読む代わりに、グラフを一つ見るだけでそれらの主要な情報を簡便につかむことができる。

図１０からは、その当時、ガソリン価格の移り変わりがわかる。また、海外のガソリン価格が安いこともわかる。

５　研究者と研究分野の変遷情報の自動抽出

次に、論文の書誌情報から、研究者と研究分野の変遷情報を自動で抽出する技術について紹介する。これは、鳥取大学の我々の研究グループが最近開発した技術 4)である。

この研究の概念図を図１１に示す。まず、図１１の左側にあるような論文の書誌データを準備する。論文のタイトルや著者名などの情報である。それを分析することで、図１１の右側にあるような研究分野や研究者の変遷情報を取り出す。

変遷情報の取り出し方は、あるものＸが出現した時に近い時期に同時に多く出現したものをそのＸのルーツ（先輩の研究者、またはルーツとなる研究分野）と考える。

例えば、図１２のようなデータがあったとする。各行は一つの論文を示し、それぞれが、その論文の発表年（出現年）、著者名を持つ。人名Ａが 1990 年の論文に初めて出現したとして、それ以外の人名は人名Ａより、前に出現していたとして、人名Ａの最も主たる先輩研究者（指導教官など）をこのデータから予測することを考える。

出現年	著者名データ	重み	a=0.5の場合
1990年	人名A, 人名B, 人名C	1	1
1991年	人名B, 人名A, 人名D	a	0.5
1992年	人名A, 人名C, 人名E, 人名F	a^2	0.25
…	…	…	…

図１２．ルーツの特定方法

我々の方法では、人名Ａが最初に出現した論文に近い時期に人名Ａとよく同時に出現するものを人名Ａのルーツと考える。具体的には、出現年ごとに重みを与えて、人名Ａが最初に出現した年の重みを１として、それ以降１年経つごとに重みが 0.5 ずつかけあわされていくとして、人名Ａと同時に出現した分だけその重みを加算し、その合計の重みが最も大きいものを人名Ａのルーツと考える。この計算を図１２のデータに対して行ったものを図１３に示す。

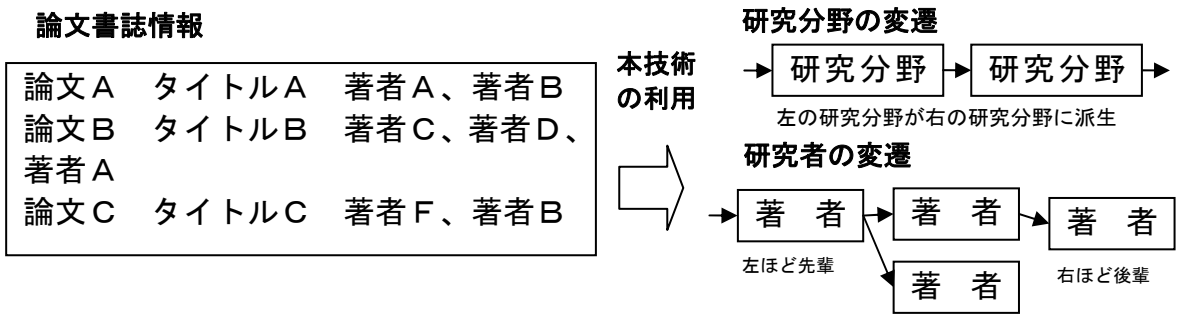

図１１．研究分野と研究者の変遷情報の取り出し

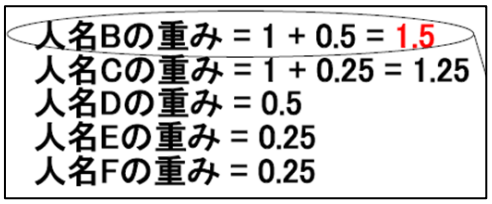

図１３．ルーツの特定のための計算例

図１３により、人名Ｂの重みが最も大きい。このため、人名Ａのルーツ（先輩研究者など）は人名Ｂであると推測する。

この方法により、言語処理学会の年次大会の論文1995年から2010年の書誌データから研究分野と研究者の変遷情報の取り出しを行ったところ、図１４のデータを得た。

人名Ａ（後輩）	人名Ｂ（先輩）
村上仁一	池原悟
馬青	井佐原均
宮尾祐介	辻井潤一
関根聡	井佐原均
丸山岳彦	柏岡秀紀
黒田航	井佐原均
難波英嗣	奥村学
松吉俊	佐藤理史
竹内孔一	影浦峡
橋本力	奥村学
…	…

分野名Ａ	分野名Ｂ（ルーツ）
自動評価	機械翻訳
統計的機械翻訳	統計
サンプリング	コーパス
タグ付きコーパス	コーパス
音声対話システム	音声対話
語義曖昧性解消	曖昧性解消
翻訳自動評価	機械翻訳
情報分析	分析
言語横断情報検索	情報検索
論文要約	情報抽出
…	…

図１４．抽出した変遷情報の例

６　おわりに

電子化テキストの増加にともない、電子テキストからの有用な情報の取り出しの重要性が高まっている。本講義では、電子テキストからの有用な情報の取り出しの例として、①社会動向調査、②数値、固有表現抽出に基づく情報の取り出し、③研究者と研究分野の変遷情報の自動抽出を紹介した。

参考文献など

1)　上田太一郎監修，村田真樹ほか著，事例で学ぶテキストマイニング，共立出版，2008.

2）Masaki Murata, Tamotsu Shirado, Kentaro Torisawa, Masakazu Iwatate, Koji Ichii, Qing Ma and Toshiyuki Kanamaru, Extraction and Visualization of Numerical and Named Entity Information from a Very Large Number of Documents Using Natural Language Processing, International Journal of Innovative Computing, Information and Control, Volume 6, Number 3(A), pp.1549-1568, March 2010.

3) 村田 真樹, 岩立 将和, 一井 康二, 馬 青, 白土 保, 金丸 敏幸，塚脇 幸代，井佐原 均，大規模記事群からの数値固有表現情報のテキストマイニング可視化システム，情報処理学会 第 184 回自然言語処理研究会，pp.25-32，2008.

4) 堀さな子，村田真樹，徳久雅人，馬青，研究者および研究分野の変遷の自動推定，言語処理学会第１７回年次大会，pp.236-239，2011.