2008－NL－184 (5)
2008／3／27

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

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[†]
独立行政法人 情報通信研究機構[†]
〒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
奈良先端科学技術大学院大学[‡]，広島大学[§]，龍谷大学[*]，京都大学[+]

あらまし

本研究では，大規模な記事群から数値固有表現情報を取り出し，様々な重要な情報を含むグラフや表を半自動で作成するシステムを構築した．われわれのシステムは約2時間の人的労力の半自動的処理で，2年分の新聞記事からおよそ300個の有用なグラフを生成した．本システムは大規模なデータからの情報抽出に役立つ．

キーワード テキストマイニング，可視化，数値情報，固有表現

Text mining and visualization system for numerical and named entity information from a large number of documents

Masaki Murata[†] Masakazu Iwatate[†,‡] Koji Ichii[§] Qing Ma[*,†]
Tamotsu Shirado[†] Toshiyuki Kanamaru[†,+] Sachiyo Tsukawaki[†] Hitoshi Isahara[†]
National Institute of Information and Communications Technology[†]
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
TEL:+81-774-98-6833 FAX:+81-774-98-6961 murata@nict.go.jp
NAIST[‡], Hiroshima University[§], Ryukoku University[*], Kyoto University[+]

Abstract

We constructed a system for semi-automatically extracting numerical and named entity (NE) sets from a large number of documents and creating various kinds of tables and graphs containing the important numerical and NE information. Our system semi-automatically created about three hundred kinds of graphs and tables using the contents of two years' worth of accumulated newspaper articles. The total amount of work performed by a human subject to compete this task was only about two hours. Our system will be useful for information extraction.

key words Text mining, Visualization, Numerical information, Named entity

1 はじめに

近年，テキストマイニング[1]は，アンケート分析やマーケティングリサーチに利用されるなど，社会的にも重要視されている．本研究はテキストマイニングの高度化を目的としている．われわれは，半自動で，大規模記事群から数値・固有表現情報を取り出し，種々の表やグラフを生成するテキストマイニング・可視化システムを構築した[2]．例えば，生成したグラフの一つは，マラソンのスタート時の気温，湿度，風速の三つの値を示すものである．数値情報に関わる実験では，われわれのシステムを利用することで約1時間の人的労力の半自動的処理で，2年分の新聞記事からおよそ100個の有用なグラフを生成した．また，固有表現に関わる実験では，150個から250個ほどの有用な表，グラフを作成した．

いくつかの関連研究がある．藤畑らは記事群から数値情報と関連する項目を抽出した[3]．松下らはデータベースに格納された情報からグラフを作成した[4]．難波らや村田らは記事群から動向情報を取り出し時間情報を横軸に数値情報を縦軸に示したグラフを作成した[5, 6]．村田らはある事柄に関連する記事群から数値ペアを取り出し，それを二次元の散布図で表示した[7]．しかし，上述の研究は，数値固有表現を利用して，様々な情報を含む，大規模なテキスト文書から，様々な種類のグラフを半自動で作成するものではない．

われわれのシステムを利用することで様々な種類の情報を含む大規模なテキスト文書から様々な種類のグラフを半自動で作成することができる．このシステムはユーザに対して，大規模なテキスト文書に含まれる様々な種類の数値・固有表現情報をグラフ化して見せることでそれら情報を容易に理解させることが可能である．さらにこのシステムは，3次元以上の数値情報を取り出しグラフ化することも可能である．表・グラフは文書中の情報を人間が容易に理解することに役立つ．

2 システム

2.1 システムの構成

われわれのシステムは，以下の構成要素からなる．

1. 主要表現セットのリスト作成部

まず，システムに大規模なテキスト文書群が入力される．システムは，数値・固有表現情報の抽出や表・グラフ化に役立つ主要表現のセットを記述したリストを出力する．主要表現は，単位表現と固有表現の種類と項目表現の三つに分類される．固有表現はIREX[8]の固有表現に従い，人名，地名，組織名，固有物名，時間表現，日付表現，金額表現，割合表現の8種類を利用した．固有表現抽出には中野らの方法[9]を利用した．単位表現と項目表現は以下の方法で抽出する．

1a. 主要単位表現抽出部

システムは数値・固有表現情報のセットの抽出やグラフ化に役立つ単位表現を抽出する．例えば，「１８度」などの「度」や「６５％」などの「%」を単位表現として抽出する．表現の取り出しには形態素解析を利用して数値に連接する名詞連続を取り出す．

1b. 主要項目表現抽出部

システムは数値・固有表現情報のセットの意味を限定するのに役立つ項目表現を抽出する．例えば，「マラソン」や「スタート時」などの対象データを限定する表現を，項目表現として取り出す．表現の取り出しには形態素解析を利用して名詞連続を取り出す．

システムは上記二つの抽出部において単位表現と項目表現を抽出する．同じ文に同時に出現する単位表現と固有表現の種類と項目表現のセットをシステムは特定し，そのセットの出現頻度を調べ出現頻度の多いセットを抽出する．システムはそのセットを記したリストをユーザに出力する．

例えば，「項目表現：スタート時」「単位表現１：度」「単位表現２：%」「単位表現３：メートル」「固有表現の種類：地名」「固有表現の種類：組織名」が同一文に出現する記事があるとそれを1個と数えて，このような記事が多数あるとこれらの表現のセットを抽出する．

2. 主要表現セットのユーザによる選択部

ユーザは上記で作成したリストから，主要表現のセットを選択する．システムはユーザの選択結果を受け取る．

3. 選択された主要表現セットのグラフ作成部

表 1: 主要表現のセットの数

単位表現の数	2	3	4	5	6	7
合計	572828	122640	46123	31427	54857	34025
5以上	36263	8977	4029	1600	490	84
削除後	511343	80345	23071	19210	50125	32647
5以上，削除後	28648	4174	1287	372	91	11
チェック数	3000	1411	1287	372	91	11
選択数	60	35	20	0	0	0

選択されたそれぞれのセットに対して，システムは主要表現同士が近くに出現する箇所を特定する．単位表現に隣接する数値表現を，その単位表現の数値情報として取り出す．例えば，「項目表現：スタート時」「単位表現１：度」「単位表現２：%」「単位表現３：メートル」「固有表現の種類：地名」が主要表現として与えられる場合，システムは「スタート時の京都の気象条件＝曇り，気温１４度，湿度６２%，北北東の風２メートル」といった文から，「項目表現：スタート時」「数値表現１：１４度」「数値表現２：６２%」「数値表現表現１：１４度」「数値表現２：６２%」「数値表現３：２メートル」「固有表現の種類：地名：京都」のセットを抽出する．システムは抽出された数値・固有表現情報のセットを集めて，表・グラフを作成する．例えば，上記の主要表現のセットの場合，システムは，横軸で気温を，縦軸で湿度を，プロットの大きさで風速を，プロットのラベルで地名を示したバブルチャートグラフを作成する．三つ以上の単位表現からなる主要表現のセットの場合，システムは，三次元散布図や顔グラフやバブルチャートなどの三次元以上の数値を表現できるグラフを用いる．現在のシステムでは，システムの出力は表データを csv ファイルで出力し，グラフの作成は Excel を利用して人手で行っている．各グラフや表の軸や項目の名称は人手で与えた．

3 数値情報のみに関する実験

まず，主要表現として単位表現と項目表現のみを用いた実験を行った．ここでは主要表現として固有表現の種類は用いない．1998 年と 1999 年の２年分の毎日新聞の記事群 (220,078 記事) を利用した．われわれは１個の項目表現と２個から７個の単位表現を主要表現セットとして利用した．実験結果を表 1に示す．

表 2: ２個の単位表現を用いた場合のリストでのユーザによる主要表現セットの選択

項目表現	単位表現		頻度	選択
昨年	歳	人	189	no
価格	円	平方メートル	189	yes
午前	階建て	平方メートル	187	no
原電	キロワット	号機	62	no
台風	キロ	号	62	yes
…				
利益	ドル	円	32	no
パナマ船籍	トン	人	32	yes
縦	センチ	枚	32	no
…				
中心気圧	メートル	ヘクトパスカル	30	no

表の１行目の「単位表現の数」は主要表現セットとして利用した単位表現の数を意味する．項目表現は常に一つを利用した．「合計」は取り出した主要表現セットの数である．「5 以上」は主要表現セットが出現した記事数が 5 以上であったものの数を意味している．主要表現セットはしばしば「局」「歩」「勝」「負」のような将棋や野球に関係する単位表現を含んだ．新聞ではこれらの表現は多数出現し主要表現セットの上位をこれらの表現が占めた．人手により主要表現セットを選択する際，他の単位表現のセットを見落とす恐れがあるため，これらを含む主要表現セットをすべて取り除くことにした．「削除後」はそのような主要表現セットを取り除いた後の主要表現セットの数を意味する．「5 以上，削除後」はそのような主要表現セットを削除しなおかつ 5 個以上の記事に出現した主要表現セットの数を意味する．「チェック数」は人手によりチェックした主要表現セットの数である．人手によりリストの頭から「チェック数」の数の主要表現セットをチェックした．このチェックではそれぞれの主要表現セットがグラフの作成に役立つかどうかを判断した．「選択数」は人手のチェックにより役立つと判断された主要表現セットの数である．

表 2に主要表現セットの例を示す．「頻度」は主要表現が同時に一文に現れた記事の数を意味する．「選択」の「yes」は人手により選ばれたことを，「no」は選ばれなかったことを意味する．例えば，表 2では，１行目の主要表現セットは「昨年」「歳」「人」である．そのセットはそれほど意味を限定するもので

表 3: 評価結果

単位表現の数	評価 A	評価 B	平均プロット数
2	0.47 (28/60)	0.72 (43/60)	36
3	0.37 (13/35)	0.71 (25/35)	14
4	0.70 (14/20)	0.85 (17/20)	4
合計	0.48 (55/115)	0.74 (85/115)	4

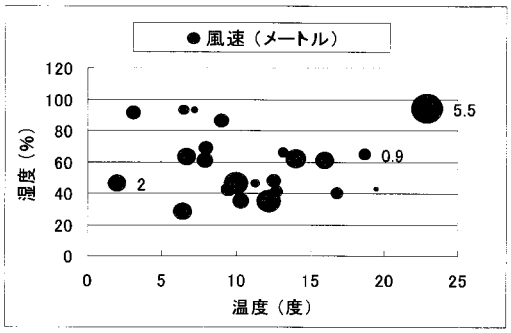

図 1: 3 個の単位表現を利用したマラソンに関するグラフ

はなく，種々のトピックを含むもので，一つのトピックについての一貫したグラフを作成するには役立たないと判断した．そのため，それを選ばなかった．2 行目の主要表現セットは「価格」「円」「平方メートル」である．主要表現セットは限定されたもので土地の価格に関する一貫した良いグラフを作成すると判断した．そこでそのセットを選択した．すべてのチェックに 1 時間を要した．

次に，人手によって選択された主要表現を使ってグラフを作成した．作成したグラフを評価した．結果を表 3に示す．一つ目の欄の「単位表現の数」は主要表現セットに用いた単位表現の数を意味する．「評価 A」は，グラフのプロットのうち 75% がある一つのトピックについて正しい情報を示す場合にそのグラフを正しいと判断し，その正しいとされたグラフの割合を意味する．「評価 B」は，グラフのプロットのうち 50% がある一つのトピックについて正しい情報を示す場合にそのグラフを正しいと判断し，その正しいとされたグラフの割合を意味する．評価 B では正解率は 0.7 から 0.8 くらいであった．評価 A を満足するグラフを 55 個作成できた．評価 B を満足するグラフを 85 個作成できた．表 3にはわれわれのシステムが作成したグラフのプロット数の平均 (平均プロット数) も示し

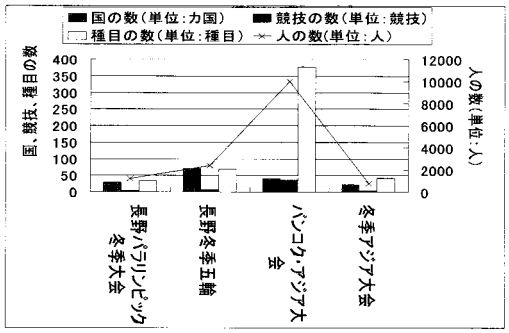

図 2: 4 個の単位表現を利用したスポーツ大会に関するグラフ

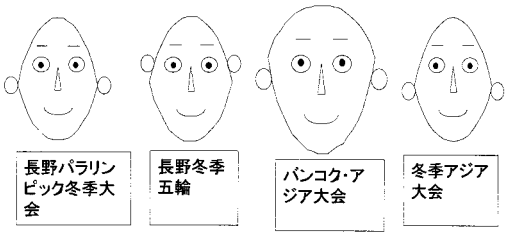

図 3: スポーツ大会の顔グラフ

ている．

実験では主要表現セットのチェックに 1 人が 1 時間を要した．つまり 1 時間の人的資源で 2 年分の新聞記事から約 100 個の有用なグラフを半自動で作成できたことになる．2 年分の新聞記事は膨大な量であり，短時間で人が読んだりチェックしたりできないものである．この観点から，われわれのシステムは便利で有用と考えることができる．

われわれのシステムで作成したいくつかのグラフを示す．図 1は 3 個の単位表現を利用して作成したグラフである．図 1は「スタート時」を項目表現として「度」「%」「メートル」を単位表現として用いて作成された．グラフの作成に用いられた記事はマラソンについて記述しているものだった．グラフで横軸はマラソンのスタート時の温度，縦軸は湿度，それぞれの円の直径は風速を示す．グラフからそれぞれのマラソンのコンディションがわかる．例えば，右上隅のプロットのデータは高温 (23 度)，多湿 (94%)，強風 (5.5 メートル) とわかる．　図 2,3 は 4 個の単位表現を利用

表 4: 主要表現セットの数

	固有表現のみ			数値と固有表現を利用		
NE 数	抽出数	5 以上	選択数	抽出数	5 以上	選択数
1	1000156	107007	3	422672	5961	20
2	972428	82780	2	494159	5182	7
3	606420	44183	3	434067	5562	25
4	271353	18749	6	255301	4615	44
5	66561	3466	69	87232	1928	41
6	8929	288	75	18727	104	53
7	325	1	1	748	1	0
8	41	0	0	123	0	0

表 5: 1 個の固有表現と 2 個の単位表現を用いた場合のリストの例

項目表現	単位表現		固有表現の種類	固有表現の例	頻度
名人戦	局	期	人名	谷川浩司, 佐藤康光, 羽生善治, 中原誠, 森内俊之, 森下卓, 丸山忠久, 加藤一二三, 井上慶太, 島朗	514

表 6: 固有表現を用いた場合の評価結果

固有表現数	評価 A	評価 B	記事数の平均
1	0.67 (2/ 3)	0.67 (2/ 3)	1165
2	0.50 (1/ 2)	1.00 (2/ 2)	2165
3	0.00 (0/ 2)	0.50 (1/ 2)	363
4	0.40 (2/ 5)	0.60 (3/ 5)	254
5	0.21 (12/58)	0.69 (40/58)	64
6	0.19 (13/69)	0.83 (57/69)	14
7	0.00 (0/ 1)	1.00 (1/ 1)	6
合計	0.21 (30/140)	0.76 (106/140)	103

表 7: 数値と固有表現を用いた場合の評価結果

固有表現数	評価 A	評価 B	記事数の平均
1	0.75 (15/20)	0.85 (17/20)	249
2	0.14 (1/ 7)	0.29 (2/ 7)	76
3	0.56 (14/25)	0.76 (19/25)	99
4	0.70 (31/44)	0.93 (41/44)	105
5	0.73 (30/41)	0.83 (34/41)	25
6	0.62 (33/53)	0.79 (42/53)	8
合計	0.44 (124/190)	0.56 (155/190)	74

して作成した．「地域」を項目表現として，「カ国」「競技」「種目」「人」を単位表現として利用して作成した．グラフの作成に利用された記事はオリンピックとアジア大会のものだった．図 2のグラフ化には，折れ線グラフと棒グラフの複合グラフを利用した．図 3では顔グラフを利用した．これらのグラフからそれぞれのオリンピックとアジア大会の規模を容易に知ることができる．これらの中では，夏に開かれたバンコクアジア大会が最も大規模であることがわかる．

上述以外にも，台風の中心気圧と風速の関係，電車の窓のひび割れ事故における故障車番号・編成車両数・電車号数・乗客数の関係等を示す多様なグラフを得た．

4 固有表現情報も含めた実験

次に主要表現として固有表現も含めた実験を行った．1998 年と 1999 年の 2 年分の毎日新聞の記事群 (220,078 記事) を利用した．われわれは 1 個の項目表現と 1 個から 8 個の固有表現を主要表現セットとして利用する固有表現のみを利用する実験と，1 個の項目表現と 1 個から 8 個の固有表現と 2 個の単位表現を主要表現セットとして利用する数値と固有表現を利用する実験の二種類を行った．実験結果を表 4に示す．表の「抽出数」はシステムで取り出した主要表現セットの数を，「5 以上」は主要表現セットが出現した記事数が 5 以上であったものの数を意味する．この 5 以上であったもののうち上位 100 件を人手によりチェックした．このチェックではそれぞれの主要表現セットがデータ抽出に役立つかどうかを判断した．「選択数」は人手のチェックにより役立つと判断された主要表現セットの数である．固有表現も含む実験での人手によるチェックでは，主要表現のセットにおいて，固有表現の種類だけを見て判断するのが難しいために，各固有表現について実際に出現している頻度の上位 10 位までの固有表現も見てチェックした．チェックに用いたリストの一例を表 5に示す．

次に選択した主要表現セットにより得られるデータが役立つかどうか調べた．その結果を表 6と表 7に示す．

「評価 A」は，抽出記事数個取り出した数値・固有表現の情報対のうち 75% がある一つのトピックについて正しい情報を示す場合にそのデータを正しいと判断し，その正しいとされたデータの割合を意味する．「評価 B」は，抽出記事数個取り出した数値・固有表現の情報対のうち 50% がある一つのトピックについて正しい情報を示す場合にそのデータを正しいと判断し，その正しいとされたデータの割合を意味する．た

表 8: 固有表現のみを利用して抽出したデータの例

(a) スライダーを投げる選手とそのチーム名

選手名	チーム名
井上	日本航空
ブロス	ヤクルト
矢野	高鍋
クロフォード	西武
酒井	近鉄
古井	メッツ
森本	市船橋
…	…

(b) ミサイルとその国名

ミサイル名	国名
シャヒーン	パキスタン
ノドン 1 号	北朝鮮
テポドン 2	北朝鮮
アロー	イスラエル
テポドン	北朝鮮

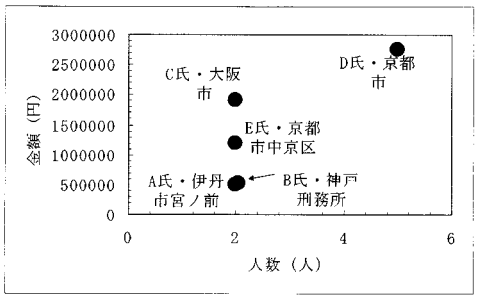

図 4: 数値固有表現情報を利用した収賄罪に関するグラフ

だし，同一文に複数の同種の固有表現が出現した場合はそのどれかが正解として解釈できるものであれば正解とした．「記事数の平均」は，数値情報のみの実験の表 3の「平均プロット数」に相当するもので，情報の取り出しに用いた記事数の平均である．

固有表現のみを用いた場合は評価 B で 76% という高い精度を得た．数値と固有表現を用いた場合は評価 A,B ともに 50% 程度の精度を得た．固有表現のみを用いた場合と数値と固有表現を用いた場合をあわせると評価 A を満足するデータを約 150 個，評価 B を満足するデータを約 250 個得た．人手による選択の時間は固有表現のみを用いた場合と数値と固有表現を用いた場合をあわせて 1 時間であった．すなわち，約 1 時間の人的資源で 2 年分の新聞記事から約 150 個から 250 個の有用なデータを抽出できた．

本システムにより抽出したデータを表 8，図 4に示す．表 8には固有表現と項目表現を主要表現とした場合に得られたデータである．表 8(a) は項目表現「スライダー」，人名と組織名の固有表現の種類を主要表現セットとした場合のものである．表から当時スライダーを投げていた選手とそのチーム名がわかる．表 8(b) は項目表現「弾道ミサイル」，固有物名と地名の固有表現の種類を主要表現セットとした場合のものである．表から当時の弾道ミサイルに関係するミサイル名とそのミサイルの保有国がわかる．その他，囲碁将棋などの毎日新聞社主催行事の開催時期・場所・主催団体・棋士名のデータ，家宅捜索を受けた組織・日付・場所・人・金額・関連する法律のデータなど多様なデータが得られた．

表 9: 分類語彙表の分類番号の変更

意味素性	分類語彙表の分類番号	変換後の分類番号
ANI(動物)	[1-3]56	511
HUM(人間)	12[0-4]	52[0-4]
ORG(組織・機関)	[1-3]2[5-8]	53[5-8]
PRO(生産物・道具)	[1-3]4[0-9]	61[0-9]
PAR(動物の部分)	[1-3]57	621
PLA(植物)	[1-3]55	631
NAT(自然物)	[1-3]52	641
LOC(空間・方角)	[1-3]17	657
QUA(数量)	[1-3]19	711
TIM(時間)	[1-3]16	811
PHE(現象名詞)	[1-3]5[01]	91[12]
ABS(抽象関係)	[1-3]1[0-58]	aa[0-58]
ACT(人間活動)	[1-3]58,[1-3]3[0-8]	ab[0-9]

図 4は固有表現と項目表現を主要表現とした場合に得られたデータである．項目表現「収賄罪」，単位表現「人」「円」，人名と地名の固有表現の種類を主要表現セットとした場合のものである．図の横軸は収賄罪に関係した人数，縦軸は収賄罪の金額を示す．各プロットには人名と関連する場所を記載した．ただし人名はシステムではとれているがここでは匿名で表示している．その他，何階建ての何階で火事が起きたかとその住民の氏名と時間，スポーツ競技の順位とその競技のメートル数・選手・組織・場所などを示す多様なグラフを得た．

5 分類語彙表の利用

本研究では分類語彙表を利用した抽出も試みた．意味ソート[10] の文献において構築した表 9の分類を利用した．まず分類語彙表の最初の 3 桁を表のように変換する．変換した番号において最初の 2 桁の分類を利

表 10: 分類語彙表を利用して抽出した電車の遅延・運休の影響のデータ

原因	電車の本数 (本)	人数 (人)
事故	64	34000
事故	16	15000
事故	29	25000
事故	18	58000
事故	11	11000
事故	16	10000
事故	4	18000
事故	4	8000
トラブル	50	700
トラブル	11	7000
除去作業	16	10000
確認	4	3000
スト	39000	900000
…	…	…

用した．単語を表 9の 2 桁の分類にわけて利用した．合計 13 種類の分類を利用した．

システムでは，分類語彙表[11] は固有表現と同様に利用した．分類語彙表の 13 分類を固有表現の 8 分類のように利用した．

数量的な評価はまだ行っていないが，分類語彙表の利用により得られたいくつかの事例を紹介する．

1 個の分類語彙表の分類の種類と単位表現 2 個と項目表現を主要表現をセットにした実験において，電車の遅延・運休に関わるデータを得た．その例を表 10に示す．表 10は，表 9の「活動」の分類を分類語彙表の分類として利用し，「本」「人」を単位表現，「影響」を項目表現として利用した場合の結果である．このデータは，どういう原因で何本の電車の運行に影響を与えて何人の人の足に影響を与えたかを示す．電車の遅延・運休の原因は，固有表現では表現されず一般名詞で表現される．そのような表現を本システムでも抽出しようとすると一般名詞の分類も扱ったような情報の利用が必要となる．本節では一般名詞の分類を扱うために分類語彙表を利用したものである．

その他，死亡日時，死亡した人の名前，その人の役職や職業，死亡した原因を示すデータも分類語彙表を利用する方法で抽出できた．人の役職や職業，死亡した原因も固有表現ではなく一般名詞で表現されるため，この情報の取り出しにも分類語彙表の一般名詞の分類が役立った．

図 5: デモシステム

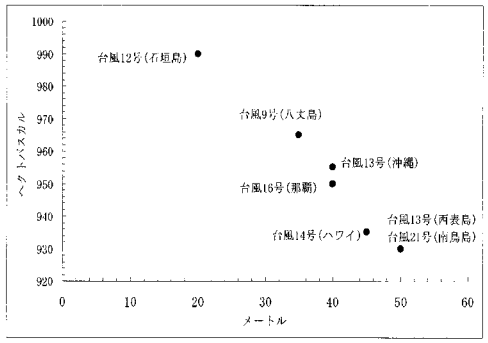

図 6: 台風のグラフ

6 デモシステム

本システムを， Web 上のアプリケーションとして構築した．その様子を図 5に示す．

図 5では，新聞記事を用いずに Web 文書からの情報の取り出しを行った．このシステムでは右下のフレームで web ページを閲覧できるようになっている．そのフレームで，まず，ユーザは「台風 中心気圧 最大風速 メートル ヘクトパスカル 号」を google で検索する．その検索結果のページを表示させている状態で，オプション設定をして demo というコマンドを入力して実行させる．オプション設定では，単位表現とし

て「メートル」「ヘクトパスカル」「号」を固有表現として地名を項目表現として「台風」をセットしておく．そうすると，googleの各検索結果から，「メートル」「ヘクトパスカル」「号」の前についている数字を抽出し，地名を抽出しその結果を表形式で整理する．その表をcsvで生成する．そのcsvファイルは左下の出力ファイル一覧に表示される．図5では，そのcsvをユーザがExcelで開いている状態を表している．図5のようなcsv形式のデータが得られるとそこから，Excelの機能などを用いることで容易に様々なグラフを作成できる．例えばそのcsvから図6のグラフを作成した．このグラフから，何号の台風がどの地域に行ったか，またそのときの台風の規模(最大風速と中心気圧)を知ることができる．

現在このWebアプリケーションはまだ一般には公開していないが，近いうちに公開して一般の人に利用できる形にしたいと思っている．

このWebアプリケーションは汎用的な構成にしている．本研究では数値固有表現の情報抽出システムとして利用しているが，他のテキストマイニングシステムのデモシステムとしても利用できる構成になっている．Webアプリケーションの裏で走らせている内部プログラムを交換することで様々なデモシステムを構築できる．例えば，内部プログラムを変更することで，入力ファイルに与えたデータの頻度分析をしたものを右下のフレームに表示するようなシステムも構築できる．今後はこのシステムを利用してweb上で一般の人が利用できる，種々のテキストマイニングシステムを構築したいと考えている．

7 おわりに

本研究では，大規模な記事群から数値固有表現情報を取り出し，様々なグラフや表を半自動で作成するシステムを構築した．例えば，台風の最大風速を横軸に中心気圧を縦軸に示したグラフを作成した．われわれのシステムは記事群から3次元以上の数値情報を取り出しそれのグラフを作成することもできる．数値情報に関わる実験では，われわれのシステムは約1時間の人的労力の半自動的処理で，2年分の新聞記事からおよそ100個の有用なグラフを生成した．また，固有表現に関わる実験では，約1時間の人的労力で150個から250個ほどの有用な表，グラフを作成した．さらに固有表現と同様に分類語彙表のデータを利用した情報の取り出しも行った．また，Web上でも実装した本システムの紹介も行った．

われわれのシステムが誤りを起こした主な原因は2個以上のトピックに関係する混ざった数値固有表現情報を取り出してしまうことであった．混ざった数値固有表現情報を一つのトピックに関する数値情報に分割するためにクラスタリングの技術を利用してみたいと思っている．また，クラスタリング以外の技術も利用して不要なデータを取り除き精度の高いデータを抽出する技術も開発したいと考えている．

参考文献

(1) 上田太一郎, 村田真樹, 小木しのぶ, 髙山泰博, 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).

(2) 村田真樹, 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に関わるテキストマイニング可視化, 第6回情報科学技術フォーラム(FIT), (2007), pp. 157–160.

(3) 藤畑勝之, 志賀正裕, 森辰則, 係り受けの制約と優先規則に基づく数量表現抽出, 情報処理学会 自然言語処理研究会 2001-NL-145, (2001).

(4) 松下光範, 米澤勇人, 加藤恒昭, 表題に基づく統計データの自動可視化手法, 情報処理学会論文誌, Vol. 43, No. 1, (2002), pp. 87–100.

(5) 難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学, 文書横断文間関係を考慮した動向情報の抽出と可視化, 情報処理学会自然言語処理研究会 2005-NL-168, (2005), pp. 67–74.

(6) Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru, Sachiyo Tsukawaki, and Hitoshi Isahara, Development of an automatic trend exploration system using the must data collection, *Proceedings of the ACL 2006 Workshop on Information Extraction Beyond The Document*, (2006).

(7) 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 塚脇幸代, 井佐原均, テキストからの主要数値ペア群の抽出とそのグラフ化, 情報科学技術レターズ, Vol. 5, (2006), pp. 73–76.

(8) Satoshi Sekine and Hitoshi Isahara, IREX project overview, *Proceedings of the IREX Workshop*, (1999), pp. 7–12.

(9) 中野桂吾, 平井有三, 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, Vol. 45, No. 3, (2004).

(10) 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.

(11) 国立国語研究所, 分類語彙表, (秀英出版, 1964).