E-012

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

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†
Masaki Murata Koji Ichii Qing Ma Tamotsu Shirado Toshiyuki Kanamaru Sachiyo Tsukawaki Hitoshi Isahara

1. はじめに

テキスト文書は，マラソン開催時の気温，湿度，風速，開催場所，主催団体など多くの数値情報や固有表現の情報を含んでいる．そのような情報を取り出し，表やグラフの形で整理することは，テキスト文書からの情報抽出に役立つ．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．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，20 個の有用な表，グラフを作成した．また，今後 2 万近くの有用な表，グラフを作成できる見込みを得た．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[1]．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[2]．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[3]．しかし，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先行研究はない．

われわれのシステムは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抽出することができる．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．テキスト文書から３次元以上の数値情報を含むグラフを作成する先行研究はない．また，分野を限定せずに，テキストから数値情報だけでなく固有表現も同時に抽出してグラフを作成する先行研究もない．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

† 独立行政法人 情報通信研究機構
{murata,qma,shirado,kanamaru,tsuka,isahara}@nict.go.jp
National Institute of Information and Communications Technology.

‡ 広島大学
ichiikoji@hiroshima-u.ac.jp
Hiroshima University

* 龍谷大学
qma@math.ryukoku.ac.jp
Ryukoku University.

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[4]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[5]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情報抽出およびグラフ作成部

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 1　主要表現セットの数

単位表現の数	2	3	4	5	6	7
合計	572828	122640	46123	31427	54857	34025
頻度 5 以上	36263	8977	4029	1600	490	84
削除後	511343	80345	23071	19210	50125	32647
頻度 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

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
ノルディックスキー	メートル	位	30	no
...				
中心気圧	メートル	ヘクトパスカル	30	yes

表 3　評価結果とプロット数の平均

単位表現の数	評価 A	評価 B	プロット数の平均
2	0.47(28/60)	0.72(43/60)	36
3	0.37(13/35)	0.71(25/35)	14
4	0.7(14/20)	0.85(17/20)	4

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．グラフの作成には Excel を利用した．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．頻度 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．「頻度 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」は被験者によりチェックされた主要表現セットの数である．被験者はリストの頭から「チェック数」の数の主要表現セットをチェックした．このチェックでは被験者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」は被験者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」は被験者によって選ばれたことを，「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．被験者はそのセットはそれほど意味を限定するものではなく，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，被験者は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．被験者は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．そこで被験者はそのセットを選択した．被験者はすべてのチェックに 1 時間を要した．

次に，被験者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(=28+13+14)作成できた．評価 B を満足するグラフを 85 個(=43+25+17)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も示している．

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 個の有用は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．

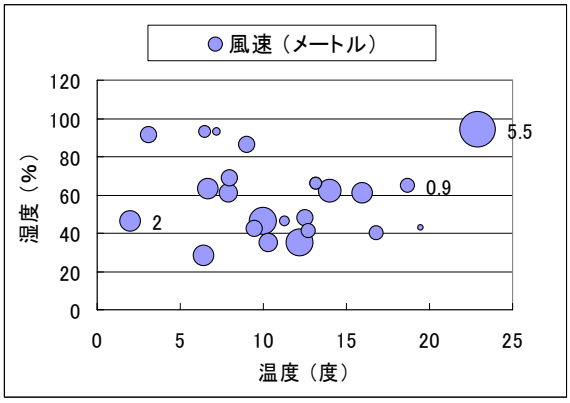

図 1　3 個の単位表現を利用したマラソンに関するグラフ

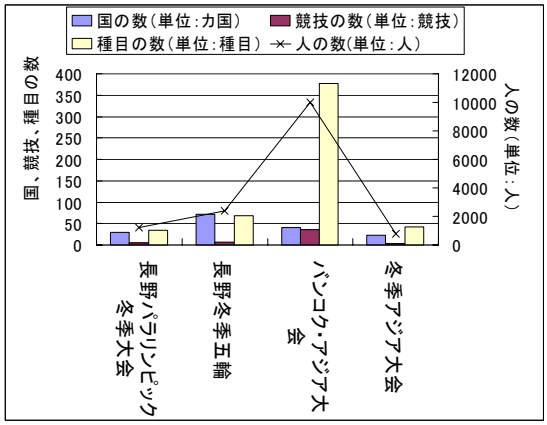

図 2　4 個の単位表現を利用したスポーツ大会に関するグラフ

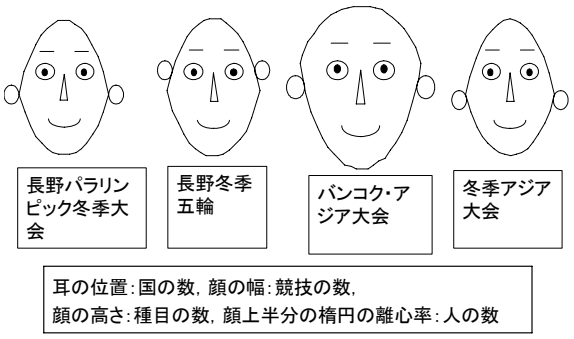

図 3　スポーツ大会に顔グラフ

図 2,3 は 4 個の単位表現を利用して作成した．「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[6]．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

表 4　　固有表現情報のみの場合の精度（評価 B）

記事数	10	30	50	70	90	全体
NE1	0.10	0.10	0.20	0.20	0.30	0.180
NE2	0.10	0.10	0.20	0.40	0.40	0.240
NE3	0.20	0.60	0.20	0.30	0.10	0.280
NE4	0.10	0.50	0.70	0.30	0.56	0.429
NE5	0.20	0.40	0.00	0.00	1.00	0.259
NE6	0.40	0.00	0.00	---	---	0.333
全体	0.183	0.333	0.283	0.293	0.350	0.282

表 5　固有表現情報のみの場合の抽出総数

記事数	5～20	21～40	41～60	61～80	81 以上	合計
NE1	270765	43377	15356	8000	24031	361529
NE2	249475	36653	12489	6140	15708	320465
NE3	128497	16492	5134	2501	5337	157961
NE4	35796	3758	1124	478	952	42108
NE5	4774	427	117	47	65	5430
NE6	275	21	10	1	0	307
合計	689582	100728	34230	17167	46093	887800

表 6　数値・固有表現情報の場合の精度（評価 B）

記事数	10	30	50	70	90	全体
NE1	0.40	0.30	0.60	0.10	0.10	0.300
NE2	0.30	0.40	0.20	0.40	0.10	0.280
NE3	0.30	0.10	0.00	0.10	0.10	0.120
NE4	0.60	0.20	0.20	0.00	0.29	0.255
NE5	0.60	0.20	0.67	1.00	---	0.458
NE6	0.50	---	---	---	---	0.500
全体	0.442	0.240	0.279	0.171	0.135	0.265

表 7　数値・固有表現情報の場合の抽出総数

記事数	5～20	21～40	41～60	61～80	81 以上	合計
NE1	91406	6920	1791	817	1166	102100
NE2	91959	6952	1931	907	1062	102811
NE3	52181	3976	1203	561	506	58427
NE4	16325	1246	404	161	115	18251
NE5	2424	180	49	14	7	2674
NE6	95	3	0	0	0	98
合計	254390	19277	5378	2460	2856	284361

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．この実験では主要表現セットのリストを眺めてもそのセットから有用な情報が得られるかどうかの判断が難しかったため，ユーザによる選択は行わず，抽出された主要表現セット全体を実験対象とし，評価はそのうちいくつかを選んで人手で評価した．評価結果を表 4 から表 7 に示す．表 4,表 5 は 1 から 6 個の固有表現と 1 個の項目表現を主要表現のセットとして用いた場合で，表 6,7 は 1 から 6 個の固有表現と 2 個の数値表現と 1 個の項目表現を主要表現のセットとして用いた場合である．表の NEx は x 個の固有表現を用いる場合を意味する．評価は抽出記事数(主要表現が同時に出現した 1 文を持つ記事の数)がちょうど 10, 30, 50, 70, 90 であったデータからそれぞれ 10 個を上限としてランダムに取り出し，それが正解かどうかを人

手で調べた．「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．表では評価 B の結果のみ載せた．評価 A では全体データで固有表現情報のみの場合 0.084，数値・固有表現情報の場合 0.018 であった．表 5 と表 7 にはデータの抽出総数を示す．

評価 A で取り出せたデータの個数は，固有表現のみを用いた場合，数値・固有表現を用いた場合の両方を合わせて 24 個であった．また，評価 B の全体データでの精度は固有表現のみを用いた場合 0.28 で数値・固有表現の情報を用いた場合 0.26 であった．本節の実験は前節と異なり人手で有用な主要表現セットを選択するなどをしておらず，その条件でこれだけの精度を得ることのできた本システムは有用と考える．また，抽出総数と精度をかけあわせて合計どのくらい有用なデータを抽出できるかを見積もった．これは例えば 21～40 の記事数の NE2 の抽出総数と記事数 30 の精度の積を 21～40 の記事数の NE2 の場合の抽出できる有用データとする手順で求めた．この見積もりでは抽出可能な評価 A のデータは固有表現のみを用いた場合，数値・固有表現を用いた場合の両方を合わせて約 2 万個であった．

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．

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. おわりに

本研究では，大規模な記事群から数値情報を取り出し，様々なグラフを半自動で作成するシステムを構築した．

例えば，台風の中心気圧を横軸に最大風速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．実験では約 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作成できた．また，固有表現を含めた実験では，20 個の有用な表，グラフを作成した．また，今後 12 万近くの有用な表，グラフを作成できる見込みを得た．

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
吉井	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

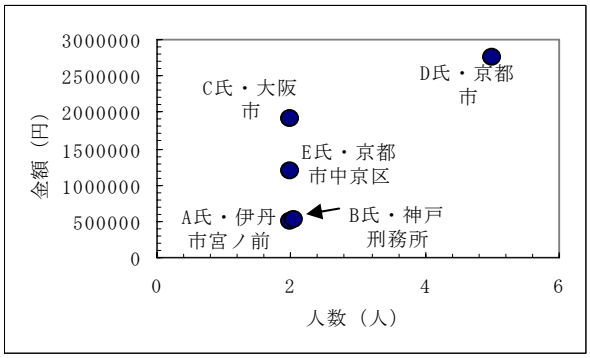

図 4　数値固有表現情報を利用した収賄罪に関するグラフ

将来的には，より大きな新聞記事を利用してみたいと考えている．また，Web のテキスト文書も利用してそれら文書からグラフを作成したいと考えている．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．混ざった情報を一つのトピックに関する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．

参考文献

[1] 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).

[2] 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.

[3] 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 塚脇幸代, 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.

[4] http://nlp.cs.nyu.edu/irex/

[5] 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), pp. 934-941.

[6] 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).