3H9-1

大規模記事群からの数値情報に関わるテキストマイニング・可視化

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

村田 真樹[*1] Masaki Murata　一井 康二[*2] Koji Ichii　馬 青[*3,*1] Qing Ma　白土 保[*1] Tamotsu Shirado　金丸 敏幸[*1] Toshiyuki Kanamaru　塚脇 幸代[*1] Sachiyo Tsukawaki

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]広島大学 Hiroshima University　[*3]龍谷大学 Ryukoku University

We have constructed a system for semi-automatically extracting numerical sets from a large number of documents and making various kinds of graphs including one where the vertical axis indicates the central atmospheric pressure of a typhoon and the horizontal axis indicates its wind speed. Our system can also extract three or more numerical values from documents and make a graph indicating these values. In experiments, our system semi-automatically created about one hundred kinds of graphs from two years' worth of newspaper articles with human labor of about one hour.

1. はじめに

テキスト文書は，台風の中心気圧や風速など多くの数値情報を含んでいる．そのような情報を取り出しグラフ化することは，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情報を取り出し種々の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 [1]．例えば，生成したグラフの一つは，台風の中心気圧を横軸，風速を縦軸にとったものである．実験の結果，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．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 [2]．松下らはデータベースデータベースに格納された情報からグラフを作成した [3]．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 [4, 5]．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 [6]．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．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

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自然言語グループ，〒 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.

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，数値情報の抽出や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，単位表現と項目表現の二つに分類される．

1a. 主要単位表現抽出部

システムは数値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，「980 ペクトパスカル」などの「ヘクトパスカル」や「40 メートル」などの「メートル」を単位表現として抽出する．

1b. 主要項目表現抽出部

システムは数値情報のセットの抽出やグラフ化に役立つ項目表現を抽出する．例えば，「中心気圧」や「風速」などの数値情報の項目となる表現を，項目表現として取り出す．

システムは上記の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述したリストをユーザに出力する．例えば，「項目表現：中心気圧」「単位表現１：メートル」「単位表現２：ヘクトパスカル」が同一文に出現する文が多数あるため，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されたリストから，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

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：中心気圧」「単位表現１：メートル」「単位表現２：ヘクトパスカル」が主要表現として与えられる場合，システムは「気象庁によると、中心気圧は 960 ヘクトパスカル、中心付近の最大風速は 35 メート

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
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

ルで、」といった文から，「項目表現：中心気圧」「数値表現１：35 メートル」「数値表現２：960 ヘクトパスカル」のセットを抽出する．

システムは抽出された数値情報のセットを集めて，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，中心気圧を横軸，風速を縦軸に示したグラフを作成する．

2.2 主要表現セットのリスト作成部

まず大規模な記事群がシステムに入力される．システムは，数値情報の抽出や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，単位表現と項目表現の二つに分類される．

われわれはそれらの表現の抽出には形態素解析システム ChaSen[7] により得られる品詞情報を利用した．

数字に隣接する名詞連続を単位表現として取り出した．時間表現の単位表現は取り除いた．「年」「月」「時」「秒」などの時間に関する表現を含むものを除いた．項目表現としては名詞連続を抽出した．

システムは同じ文に高頻度で出現する主要表現のセットを作り，同じ文に出現する共起頻度の大きい順にそのセットを並べたリストを作成し，それをユーザに見せる．

ユーザは主要表現の構成要素の種類を指定できる．例えば，二つの単位表現と一つの項目表現を主要表現のセットに指定できる．また，三つの単位表現と一つの項目表現を主要表現のセットに指定できる．

2.3 主要表現セットのユーザによる選択部

ユーザはリストから主要表現のセットを選択する．リストには，グラフの作成には適さない主要表現のセットが含まれている．ユーザはそれらのセットを人手で削除する．システムはユーザの選択を受け取る．

2.4 選択された主要表現セットのグラフ作成部

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．

システムは抽出された数値情報のセットを集めて，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．

3. 実験と考察

3.1 実験

われわれのシステムを用いて大規模文書群から様々なグラフを作成する実験を行った．この実験では，1998 年と 1999 年の２年分の毎日新聞 [8] 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: 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	no

表 3: 3 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
円	ドル	円	銭	153	no
…					
出火	階	階建て	平方メートル	57	yes
…					
前年同月比	円	社	店	56	no
福井県敦賀市	キロワット	号機	次	56	yes
通算	アンダー	ボギー	位	56	no
各組	チーム	位	組	24	no
…					
中心	キロ	ヘクトパスカル	メートル	24	yes
…					
調査	%	歳	人	24	no
スタート時	%	メートル	度	24	yes
賞金	回	着	頭	23	no

表 2，表 3，表 4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

表 4: 4 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
通算	アンダー	バーディー	ボギー	位	54	no
…						
売上高	%減	円	社	店	35	no
…						
芝	メートル	円	回	頭	28	no
…						
末端価格	キロ	トン	円	人	8	yes
…						
モーリタニア	キロ	ステージ	位	回	8	no
地域	各国	競技	種目	人	8	yes
リーグ	カ国	位	次	組	8	no
…						
走行中	号	号車	人	両	7	yes
…						
賞金	競争	歳	着	頭	7	yes

表 5: 評価結果

単位表現の数	評価 A	評価 B
2	0.47 (28/60)	0.72 (43/60)
3	0.37 (13/35)	0.71 (25/35)
4	0.70 (14/20)	0.85 (17/20)

で，一つのトピックについての一貫した良い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 5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 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 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 6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である．2 個の単位表現を使うグラフが最も多くのプロット数を持ち，単位表現の数が増えるほどプロットの数が小さくなっている．

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．

システムが誤った例を考察した．多くの事柄について同じ単位表現が用いられる場合に誤りが多いことがわかった．例えば，「円」が単位表現として取り出された．しかし，「円」は「売上高」「純利益」「月額」「年額」のような様々なものに用いられる．システムはこれらの混ざった数値情報を取り出しそれをグラフ化して，一貫性のない良くないグラフを生成した．

表 6: プロット数の平均

単位表現の数	グラフの数	プロット数の平均
2	60	36
3	35	14
4	20	4

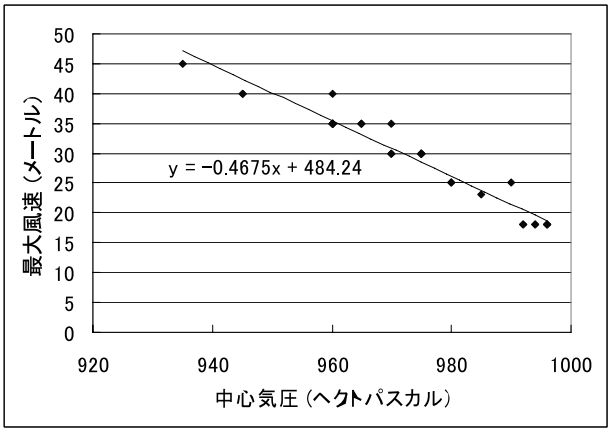

図 1: 2 個の単位表現を利用した台風に関するグラフ

3.2 システムにより作成したグラフ

本節では，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に 2 個の単位表現を使って生成したグラフを示す．図 1は「中心気圧」を項目表現として，「ヘクトパスカル」「メートル」を単位表現として利用して作成したものである．グラフは台風に関するものである．グラフでは横軸は「中心気圧」を縦軸は「最大風速」を示している．グラフから，気圧が低いと風速が大きくなることがわかる．また同じ気圧の時でも異なる風速になることがわかる．図のプロットに対して単回帰線とその式を計算した．その式を用いることで中心気圧から最大風速を予測できる．

図 2は 3 個の単位表現を利用して作成したグラフである．図 2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．グラフはマラソンに関係するものと判断できる．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 3，図 4は 4 個の単位表現を利用して作成した．図 3と図 4は「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 3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．三つの棒グラフは，国の数，競技数，種目数を意味する．折れ線グラフは参加人数を意味する．図 4では顔グラフを利用した [9]．顔グラフでは耳の高さは国の数を，顔の幅は競技数を，顔の長さは種目数を，顔上半分の楕円の離心率は人間の数を意味する．顔グラフは多くの数値データを一つのグラフで示す時に役立つ．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

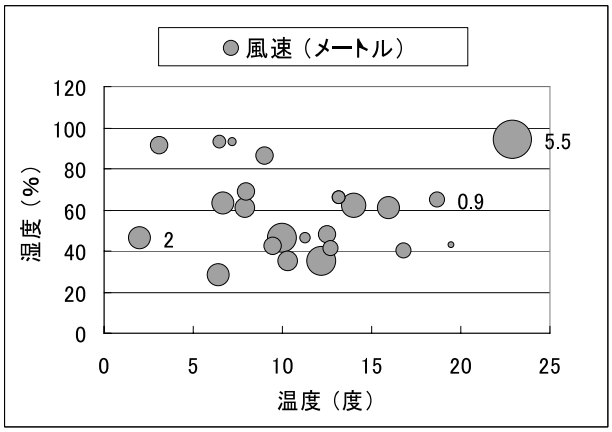

図 2: 3 個の単位表現を利用したマラソンに関するグラフ

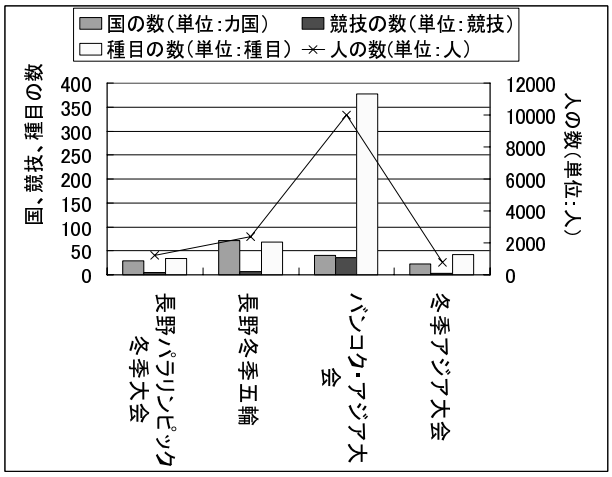

図 3: 4 個の単位表現を利用したスポーツ大会に関するグラフ

バンコクアジア大会が最も大規模であることがわかる．

われわれのシステムはこのような興味深い数値情報の抽出とそのグラフ化を半自動で行うことができた．

4. 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 個の有用はグラフを作成できた．

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情報を取り出してしまうことであった．混ざった数値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．

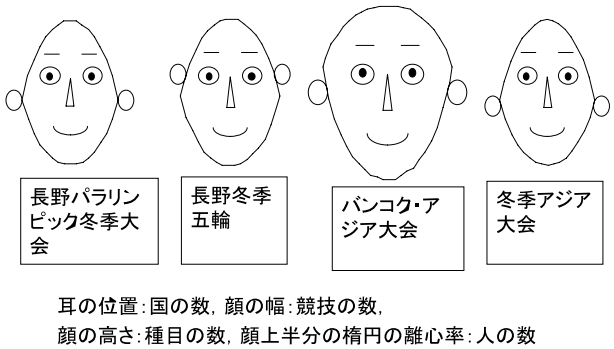

図 4: スポーツ大会の顔グラフ

参考文献

[1] Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru, Sachiyo Tsukawaki, and Hitoshi Isahara. Text mining and visualization system for numerical information from a large number of documents. *The International Workshop on Data-Mining and Statistical Science (DMSS 2006)*, pp. 46–53, 2006.

[2] 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.

[3] 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, pp. 87–100, 2002.

[4] 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, pp. 67–74, 2005.

[5] 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.

[6] 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, pp. 73–76, 2006.

[7] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. 1999.

[8] Mainichi Publishing. Mainichi Newspaper 1998-1999, 1999.

[9] 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.