テキスト文書群からの主要数値ペア群の抽出とそのグラフ化

Extraction of important numerical pairs from text documents and visualization of them

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

われわれは MuST の 2 年目の試みとして，あるトピックに関連する電子テキスト群から自動で数値情報の二項組を抽出し，グラフ化するためのデータを出力するシステムを構築した[8]．このシステムは，まずトピックに関連する二つの単位表現と一つの項目表現とを取り出し，次にこれらの表現が同時に出現している文から数値情報の二項組を抽出する．抽出した数値の一方を横軸に，一方を縦軸にプロットしたグラフを作成して表示する．MuST でなされている研究では，主に横軸に時間軸，縦軸に数値を示す二次元のグラフを自動的に出力する研究が多い．これに対して本研究は横軸，縦軸ともに数値情報としたところが新しいと考えている．

関連研究としては，テキストから数値と項目の組と取り出す研究[1]，テキストから属性・属性値の組を取り出す研究[2]，データベースからグラフを表示する研究[3]，時間情報に関わる数値情報をテキストから取り出し時間情報を横軸に数値情報を縦軸にグラフを表示する研究[4,5]はあるが，本研究のように，時間情報に関わらない二つの数値情報の組を自動抽出しそれらを横軸，縦軸にグラフ表示するものはない．

本研究において作成したシステムは，テキスト群から取得した数値情報をグラフで表示してユーザに提示することを目的としている．グラフ化することにより，テキスト文書群に含まれる数値情報の把握を容易にする．本研究は，MuST で提供されているデータを利用していることから，新聞記事程度の規模のテキストの集合を対象とする．

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. 主要表現抽出部

あるトピックに関連するテキスト群から主要表現を抽出する．本システムにおける主要表現は，単位表現と項目表現の 2 種類からなる．

1a.単位表現抽出部

数値情報の二項組を抽出，整理する際に必要となる単位表現を抽出する．

例えば，映画に関する記事群から，興行収入を示す「5 億円」の「円」や，観客動員数を示す「30 万人」の「人」を単位表現として抽出する．

1b.項目表現抽出部

数値情報の二項組を抽出，整理する際に必要となる項目表現を抽出する．

例えば，映画に関する記事だと，「興行収入」や「観客動員数」などを項目表現として抽出する．

2. 数値情報対抽出部

対象のテキスト群において，主要表現抽出部において取り出した表現が同時に出現している箇所を特定し，その部分に記載されている主要表現対を数値情報対として抽出する．

単位表現についてはそれに関連する数値も同時に抽出し，数値と単位表現をあわせて数値表現として抽出する．

例えば，映画の記事だと，「項目表現：興行収入」「数値表現：5 億円」「数値表現：30 万人」の情報対を数値情報対として抽出する．

3. 主要グラフ抽出及び表示部

対象の記事群において，数値情報対抽出部において抽出した数値情報対を整理して，グラフ化して表示する．

例えば，映画の記事だと，数値情報対抽出部で取り出した，「興行収入」「観客動員数」に関する数値情報対をグラフ化して表示する．横軸に「人」（「観客動員数」を示す）をとり，縦軸に「円」（「興行収入」を示す）をとってグラフ化して表示する．

２．２　主要表現抽出部

与えられたテキスト群から単位表現および項目表現を抽出する．

各表現の抽出には，ChaSen を利用した．ChaSen の出力において品詞の情報を利用して，各表現の抽出を行った．

単位表現は，数値の前方または後方に接続する名詞連続を取り出した．本稿の数値情報には時間表現を含めないことにした．このため，単位表現として得られた表現のうち，時間に関する表現(例：「年」「月」「日」)を含む表現を取り除いた．項目表現は，名詞連続を取り出した．

次に今扱っている分野の記事群で主たる役割を果たす主要な単位表現，項目表現を取り出した．

主要な表現の抽出方法としては，以下の 3 種類の方法を作成した．システムではこれらの方法の一つを利用して実行する．以下の式の値が大きいものほど主要な表現であると判断する．

† 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.
#京都大学.
kanamaru@hi.h.kyoto-u.ac.jp
Kyoto University.

- 式１：Okapi の TF 項の式[6]

$$Score = \sum_{i \in Docs} \frac{TF_i}{TF_i + \frac{l_i}{\Delta}}$$

- 式２：総頻度

$$Score = \sum_{i \in Docs} TF_i$$

- 式３：総出現記事数

$$Score = \sum_{i \in Docs} 1$$

ただし，i は記事の番号，Docsは記事の番号の集合，TF_iは記事iでの表現の出現回数，l_iは記事iの長さ，Δは記事群Docsにおける記事の平均の長さを意味する．OkapiのTF項の式は，複数の記事に万遍なく出現しなおかつ頻度が大きい表現のスコアを大きくする効果がある．

項目表現については，長い文字列を優先して取ってくることができるように，TF_iを記事iでの表現の出現回数とせずに，記事iでの表現の出現回数とその表現の文字列長の積とする方法も利用した．

２．３　数値情報対抽出部

数値情報対抽出部では，対象の記事群において，主要表現抽出部において取り出した表現が同時に出現している箇所を特定し，その部分に記載されている主要表現対を数値情報対として抽出する．

現時点のシステムでは，句点，改行，文書の切れ目を示す特殊記号を文区切りとし，これらをはさまずに同時に二つの単位表現と一つの項目表現が出現した箇所を，同時に出現した箇所とした．また，一記事につき，数値情報対は一つとし，記事中で最初に現れた情報対のみを取り出す．また，単位表現と連接した数値と単位表現を組み合わせたものを数値表現として取り出す．

２．４　主要グラフ抽出及び表示部

主要グラフ抽出及び表示部では，対象の記事群において，数値情報対抽出部において抽出した数値情報対を整理して，グラフ化して表示する．数値情報対の一方を横軸に，もう一方を縦軸にしたグラフを作成する．

本稿の手法では，主要表現抽出部において，複数の単位表現，項目表現を取り出し，それら複数の表現のすべての組み合わせ分のデータにおいて，数値情報対抽出部を用いて複数種類のグラフを作成し，それら複数種類のグラフにおいて，以下の評価式を計算し以下の評価式の値が大きいものほど有用なグラフと仮定する．本稿では以下の４種類の評価式を利用した．システムではこれらのうちの一つを利用して実行する．

方法 1 --- 数値情報対の頻度と主要表現のスコアを使う．

$M = Freq \times S_1 \times S_2 \times S_3$

方法 2 --- 数値情報対の頻度と主要表現のスコアを使う．

$M = Freq \times (S_1 \times S_2 \times S_3)^{\frac{1}{3}}$

方法 3 --- 数値情報対の頻度を使う．

$M = Freq$

方法 4 --- 主要表現のスコアを使う．

$M = S_1 \times S_2 \times S_3$

ただし，Freq は２．３節の方法によって取り出した数値情報対の数，S_1, S_2, S_3は，２．２節の方法によって算出した三つの主要表現のScoreの値である．

また，主要表現抽出部において，単位表現，項目表現についてはそれぞれ上位 5 つずつ取り出し，この中から単位表現を二つ，項目表現を一つ選び出しそのすべての組み合わせ 50 個(=5×4×5÷2)に対して上記の計算をしてその値が大きいものほど優先的に出力すべきデータと仮定する．

３．実験と考察

本節ではまず３．１節，３．２節において本システムの実行例を示す．３．３節で本システムの性能を調べた数量的な評価について述べる．

３．１　主要表現抽出

主要表現抽出の実験を行った．毎日新聞の 2000 年と 2001 年の記事より「映画」と「興行収入」，「台風」と「最大風速」のそれぞれの AND 検索を行い，二つの記事群を得た．これを実験に用いた．実験結果を表１に示す．

表１　主要表現抽出の例

映画のデータ	台風のデータ
単位表現	
円	号
人	メートル
ドル	キロ
歳	ヘクトパスカル
本	ミリ
項目表現	
映画	台風
興行収入	最大風速
作品	中心付近
千尋	気象庁
神隠し	時速

OkapiのTF項の式を利用し，項目表現では，TF_iを表現の出現回数とその表現の文字列長の積とする方法を利用した．表１にはScoreの上位５つを示している．

表を見ると，それぞれその記事群の主要な表現がうまく取り出せている．例えば，映画のデータだとその主たる項目表現の「興行収入」がまた単位表現として「円」「人」などが取れている．台風のデータだとその記事群の主たる項目表現の「最大風速」がまた単位表現として「号」「メートル」「キロ」「ヘクトパスカル」など台風に関連する単位表現が取れている．

３．２　二項組の数値情報のグラフ化

次に，二項組の数値情報のグラフ化の実験を行った．前節と同じ三つの記事群を用いた．主要グラフ抽出及び表示部の方法３を利用して，有用なグラフの抽出を行った．方法３の M の値が最も大きいグラフを作成した．そうすると，映画のデータでは，「円」「人」が単位表現で「興行収入」が項目表現の場合のグラフをシステムは作成した．台風のデータでは，「メートル」「ヘクトパスカル」が単

位表現で，「最大風速」が項目表現の場合のグラフを作成した．それぞれのグラフを図１から図２に示す．実際のグラフ化には Excel を用いた．

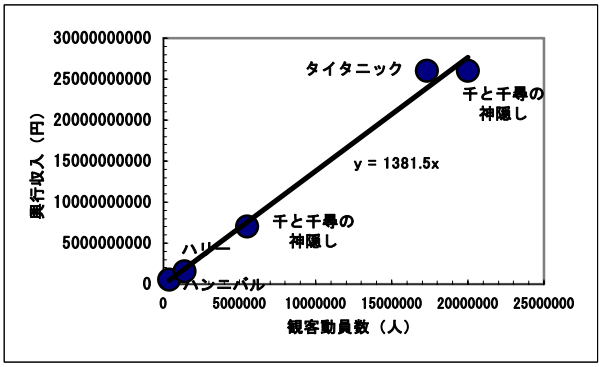

図１　映画のデータの数値二項組のグラフ

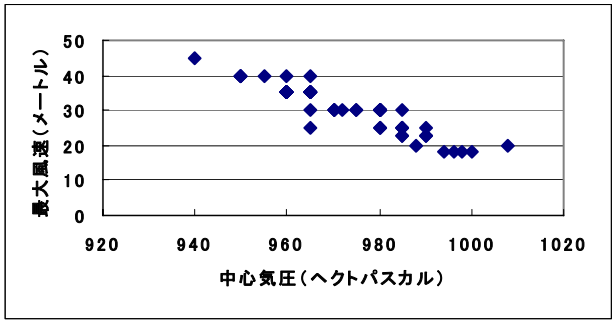

図２　台風のデータの数値二項組のグラフ

映画のグラフについては，数値情報対を取り出した文またはその記事のタイトル，先頭の文などから鍵括弧内の表現を取り出すことで，各プロットに対するラベルに相当する表現を容易に自動で取り出すことができたため，それもグラフに表示している．映画のデータで『千と千尋の神隠し』が複数あるのは異なる日時におけるそのデータが記事から得られたためである．映画のデータではプロットされた点に対して原点を通る単回帰直線も求めた．この直線の式からだいたい一人 1400 円を払って入場していることがわかる．また，タイタニックは直線の左側に，『千と千尋の神隠し』が直線の右側にあるため，大人向けと思われるタイタニックは平均よりも一人当たり高い料金を払っており，子供向けと思われる『千と千尋の神隠し』は平均よりも一人当たり安い料金を払っていることもわかる．台風のデータでは，気圧が低いと風速も大きくなる様子が観察できる．システムの出力にノイズが含まれる可能性があるため，確認用に数値情報の抽出元を文単位で出力している．グラフに表示した各プロットにマウスを持っていくとそのプロットのデータを抽出した際に利用した文や記事を表示させ，データが正しいかをユーザが容易に確認できるようにする工夫を設けると便利と思われる．

３．３　数量的評価

本システムの数量的評価を行った．数量的評価の対象となるデータは，毎日新聞の 2000 年と 2001 年の記事よりキーワード検索によって抽出した記事データ群で構成した．キーワード検索によって抽出された記事群を 1 分野とし，全部で 24 分野の記事データ群を用意した．キーワード検索によって抽出された記事群から人手で記事を取り除くなどの操作は一切していない．また，このデータはシステムの構築には利用していない．キーワードとしては，「内閣支持率」「市町村合併」「最低気温」「労働力人口」などを与えた．特に複数の数値情報の対がすぐに連想されるキーワードを選ぶということはしていない．

まず，本システムの主要グラフ抽出及び表示部の出力の上位 30 個のグラフ作成に利用した主要表現の三つ組を参考にして記事群にグラフを作成するための情報が含まれているかを判断した．そのような情報が含まれている記事群は 8 分野であった．以下の評価ではこれらのデータを利用した．

評価では，主要グラフ抽出及び表示部の出力において，どれだけ正しい二項組の数値情報を示すグラフを上位で抽出することができたかを調べた．その結果を表２に示す．

精度の計算には以下の三種類の手法を用いて，それぞれ精度を求めた．表の TP1 は上位 1 個での適合率の平均，TP5 は上位 5 個での適合率の平均，RP は r-precision の平均である．(ここでの平均は分野ごとに精度を求め，その精度の和を分野の数で割ったものである．) r-precision とは正解事例の数だけシステムが出力した際の適合率のことである．r-precision は適合率と再現率が一致する時のそれらの精度の値を示すものになっている．適合率はシステムの出力のうち正解したものの割合で，再現率は正解事例のうちシステムが正解を出力できたものの割合である．

表２中の方法 1 から方法 4 は，２．４節で述べた有用なグラフを抽出する方法を意味する．式 1 から式 3 は２．２節で述べた主要表現を抽出するスコアの算出方法を意味する．文字列長の利用は，主要項目表現のスコア算出において表現の出現回数に文字列長をかけあわせることを意味し，文字列長の利用なしは，主要項目表現のスコア算出において表現の出現回数に文字列長をかけあわせないことを意味する．

r-precision の算出には，正解の集合が必要となるが，今回の問題設定ではこれを定めるのが困難なため，上記 24(=4×6)個の手法の上位 30 個の出力を人手で確認しこの中にあった有用なグラフのみを正解としている．

評価 A は，出力のグラフ作成に利用した数値情報対の情報がそのグラフにおいて全体の 75%以上正しく抽出されている場合に正解としたものであり，評価 B は，出力のグラフ作成に利用した数値情報対の情報がそのグラフにおいて全体の 50%以上正しく抽出されている場合に正解としたものである．

表２　数量的評価

	評価A			評価B		
	TP1	TP5	RP	TP1	TP5	RP
	式１と文字列長の利用					
方法1	0.250	0.625	0.188	0.500	0.750	0.395
方法2	0.375	0.625	0.251	0.625	0.750	0.458
方法3	0.500	0.625	0.257	0.750	0.750	0.464
方法4	0.250	0.375	0.102	0.375	0.500	0.220
	式２と文字列長の利用					
方法1	0.375	0.500	0.126	0.500	0.625	0.280
方法2	0.250	0.625	0.189	0.500	0.750	0.342
方法3	0.500	0.625	0.189	0.750	0.750	0.342
方法4	0.125	0.375	0.091	0.250	0.500	0.227
	式３と文字列長の利用					
方法1	0.250	0.625	0.138	0.500	0.750	0.333
方法2	0.250	0.625	0.201	0.500	0.750	0.396
方法3	0.500	0.625	0.201	0.750	0.750	0.396
方法4	0.250	0.375	0.102	0.375	0.500	0.220
	式１の利用，文字列長の利用なし					
方法1	0.250	0.625	0.200	0.500	0.750	0.407
方法2	0.250	0.625	0.263	0.500	0.750	0.470
方法3	0.500	0.625	0.269	0.750	0.750	0.476
方法4	0.125	0.375	0.120	0.125	0.500	0.191
	式２の利用，文字列長の利用なし					
方法1	0.375	0.500	0.150	0.500	0.625	0.316
方法2	0.375	0.500	0.150	0.625	0.625	0.316
方法3	0.500	0.625	0.213	0.750	0.750	0.378
方法4	0.125	0.250	0.070	0.125	0.375	0.124
	式３の利用，文字列長の利用なし					
方法1	0.250	0.625	0.138	0.500	0.750	0.333
方法2	0.250	0.625	0.201	0.500	0.750	0.396
方法3	0.500	0.625	0.201	0.750	0.750	0.396
方法4	0.250	0.375	0.102	0.375	0.500	0.220

実験結果から，主要表現のスコアの算出(Score の算出)には Okapi の TF 項の式を利用する方法(式 1)を用いるのが最もよいことがわかった．

また，実験結果から，有用な数値情報対を選択する方法としては，数値情報対の頻度のみを用いる方法 3 が最もよいことがわかった．

また，項目表現を抽出する際，スコアに文字列長をかけあわせる方法とそうしない方法の二種類を用いたが，実験では文字列長を考慮しない方が若干よい結果となった．しかし，両者の差は極めて小さい．

最良の精度では，一つ目に出力したものがあっているかどうかの精度（TP1）では評価 A で 0.500，評価 B で 0.750 である．有用な情報が記事群にあるかないかを調べずに入力として記事群を与えた場合の精度では，評価データの元の個数 24 個で計算し直すと(8/24 を掛ける)，最良の精度では，一つ目に出力したものがあっているかどうかの精度では評価 A で 0.167，評価 B で 0.250 である．表２から，評価 A での最良の精度では，RP は，0.269 であった．再現率と適合率が同じ値になる場合の精度を示す，RP で 0.269 であるため，再現率と適合率が同じ値になる場合で正解のうちのだいたい 1/4 程度のグラフを抽出できていることがわかる．

４．おわりに

われわれはある話題に関連する電子テキスト群から自動で数値情報の二項組を抽出し，それをグラフ化して表示する我々のシステムを開発した．数量的な評価実験の結果，入力の記事群にグラフ化に適した情報を含む場合の最良の精度では，一つ目に出力したものがあっているかどうかの精度では評価 A で 0.500，評価 B で 0.750 である．入力の記事群に抽出すべき数値情報があるかないかを調べずに記事群を与えた場合，評価 A で 0.167，評価 B で 0.250 である．入力として記事群を与えただけでこれだけの精度で二項組の数値情報を示すグラフを出力できる本システムは，便利で有効と思われる．

われわれは次の段階として，数値二組のデータを取り出すのではなく，三つ以上の数値の組の情報を取り出し，三次元以上を表現できるグラフなどで可視化する研究[7]を進めている．また，関連記事でなく，数年分の新聞記事などの大規模テキスト群から半自動で数値情報の組の抽出とその可視化をする研究[7]も進めている．

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] 高橋哲朗, 乾健太郎, 松本裕治, 言語パタンと統計的共起尺度による属性関係抽出, 言語処理学会第 11 回年次大会, (2005).

[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, (2002), pp. 87-100.

[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, (2005), pp. 67-74.

[5] 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, MuST データを利用した自動動向調査システムの開発, 電子情報通信学会言語理解とコミュニケーション研究会 NLC2005-119, (2006).

[6] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford, Okapi at TREC-3, TREC-3, (1994).

[7] 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) , 2006, p.46-53.

[8] 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，5 巻．p.73-76, 2006.