

テキスト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† Masaki Murata 一井 康二‡ Koji Ichii 馬 青†* Qing Ma 白土 保† Tamotsu Shirado 金丸 敏幸†# Toshiyuki Kanamaru 塚脇 幸代† Sachiyo Tsukawaki 井佐原 均† Hitoshi Isahara

1. はじめに

本研究では、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システムを構築した。このシステムは、まず主要な二つの単位表現、一つの項目表現を取り出し、次にこれらの表現が同時に出現している箇所から数値情報の二項組を抽出する。最後に抽出した二項組を整理し、二項組の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。

数量的な評価実験の結果、入力の記事群にグラフ化に有用な情報を含む場合の最良の精度では、一つ目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。入力の記事群に有用な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。ただし、評価 A は出力のグラフ作成に利用した数値情報対(グラフのプロットされた各点)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 は数値情報対の情報が全体の 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。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。

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。

本研究は、記事群から情報を取得する際に、数値情報に関するグラフを作成してそれをユーザに提示できる。グラフ化されたデータは視覚的に認知しやすく、記事群の情報を素早く理解することに役立つ。本研究は、このような記

事群からのユーザの情報収集に役立つ。

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. 主要表現抽出部

まず、ある話題に関連する記事群が与えられる。その記事群から、数値情報の二項組を抽出、整理する際に必要となる主要表現を抽出する。主要表現としては、単位表現、項目表現を抽出する。

1a. 主要単位表現抽出部

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、映画に関する記事だと、「興行収入の「5 億円」などの「円」や、「観客動員数の「30 万人」などの「人」を単位表現として抽出する。

1b. 主要項目表現抽出部

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。

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「観客動員数」をとり、縦軸に「興行収入」をとってグラフ化して表示する。

2.2 主要表現抽出部

数値情報の二項組を抽出、整理する際に必要となる主要表現を抽出する。主要表現としては以下のものを抽出する。

- ・単位表現
- ・項目表現

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

† 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.

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。

- 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こ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、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切れ目とし、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、記事中で最も最初に現れた情報対のみを取り出す。また、単位表現と接続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して判断して抽出する。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、S₁、S₂、S₃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、項目表現の一つを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有用なグラフと判断する。

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」、「ビール」と「希望小売価格」のそれぞれのAND検索を行い、三つの記事群を得た。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	ビールのデータ
単位表現		
円	号	円
人	メートル	ミリリットル
ドル	キロ	%
歳	ヘクトパスカル	ケース
本	ミリ	本
項目表現		
映画	台風	希望小売価格
興行収入	最大風速	発泡酒
作品	中心付近	ビール
千尋	気象庁	ビジネス情報
神隠し	時速	缶

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。また、ビールのデータだと、その記事群の主たる項目表現の「希望小売価格」「ビール」「発泡酒」がまた単位表現として「円」「ミリリットル」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

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、項目表現の抽出に、その表現の文字列長をかけあわせる方法(文字列長を利用する方法)とそうしない方法の二種類を用いたが、実験結果では文字列長を用いない方が若干よかった。しかし、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。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。

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ほしい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

我々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有用な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。入力の記事群に有用な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。

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, 言語パター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, Masao Utiyama, and Hitoshi Isahara, Use of multiple documents as evidence with decreased adding in a Japanese question-answering system, Journal of Natural Language Processing, Vol.12, No.2, (2005).