特開 2008-21052 号

情報抽出装置、情報抽出方法及び情報抽出プログラム

発明者
村田（むらた）　真樹（まさき）

テキストマイニング結果例
（特許関連のキーワード抽出イメージ）

技術の概要

本発明は、電子化して記録されている情報群から、あるトピックに関連する情報の対を自動で抽出し、グラフ化することを目的としています。この装置の構成は図 1 の通り、関連記事 DB から主要表現を抽出する主要表現抽出部と、この手段によって抽出された主要表現に基づいて、関連記事 DB を構成する記事から複数の情報の対を情報対として抽出する情報対抽出部とを備えています。抽出する情報は、複数の項目表現と、これに対する単位表現の対になります（図 2）。項目表現は、例えば日経平均株価や最高気温等であり、これに対する単位表現は、9100 円の円や 35 度の度になります。情報対抽出部は、主要表現抽出部によって抽出された主要表現に基づいて、記事群を構成する記事から複数の情報の対を情報対として抽出します。情報対抽出部は、例えば、関連記事 DB に格納された記事群において、主要表現抽出部によって抽出された主要表現が同時に出現している箇所を特定し、その箇所に記載されている数値情報の対を抽出し、抽出した数値情報の対と上記主要表現のうちの項目表現との対を情報対とします。この主要表現のうちの単位表現については、情報対抽出部は、その単位表現に関連する数値（例えば、単位表現に隣接して記事中に出現している数値）も同時に抽出し、数値と単位表現とをあわせて数値表現として抽出します。

表示部は、情報対抽出部によって抽出された数値情報対を整理して表示します。例えば、映画の記事の場合、情報対抽出部が抽出した、「興行収入」、「観客動員数」に関する数値情報対を、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します。表示部は、主要表現抽出部が抽出した主要表現が複数の場合に、情報対抽出部が各主要表現に基づいて抽出した複数種類の情報対から、各主要表現について所定の評価値算出式に基づいて算出される評価値に基づいて主要な情報対を選択した上で、選択

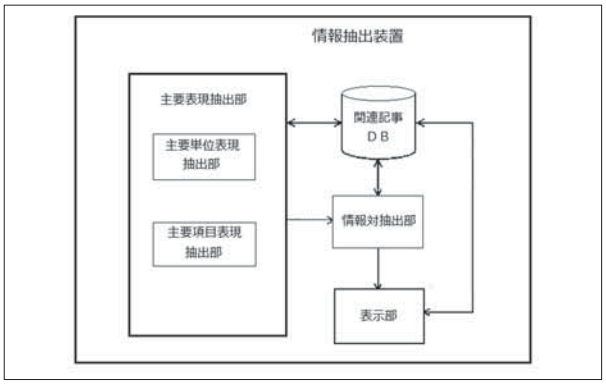

図 1　システム構成例

映画のデータ	台風のデータ	ビールのデータ
項目表現		
映画 興行収入 作品 千尋 神隠し	台風 最大風速 中心付近 気象庁 時速	希望小売価格 発泡酒 ビール ビジネス情報 缶
単位表現		
円 人 ドル 歳 本	号 メートル キロ ヘクトパスカル ミリ	円 ミリリットル % ケース 本

図 2　主要表現の例

した主要な情報対をグラフ化します。また、情報対抽出部が、ユーザの指定入力に従って、上記複数種類の情報対から主要な情報対を選択することや、表示する円の大きさが数値表現の数値の大きさを示すバブルチャートの形式で画面表示することもできます。

応用

企業においては、製品やサービスのアンケート結果、お客様相談センターに寄せられた苦情等の内容等の多くが電子データ化されています。これらの蓄積された膨大なテキストデータのデータベースについて、その内容と傾向の変化を把握し、今後の企業の商品販売やサービス提供の戦略へ反映し、さらには売り上げを増加させるための施策を行うことが必要とされています。しかし、顧客のアンケート結果全てを１枚１枚読んでいたのでは、時間がかかりすぎます。テキストマイニングは、膨大なテキスト情報の中から、必要な情報のみを素早く切り出してくることができます。回答が選択式のアンケートであれば、顧客の傾向や満足度等は、容易に把握可能かもしれませんが、アンケートなどの最後などにある自由記述文には対応はできません。これを機械で行うには文章を理解する知識が不可欠であり、キーワードを検索できる程度のツールでは実現不可能です。そこで、関連記事 DB の中の文章における品詞の情報を利用し、例えば時間表現であれば数値の後方に連続する名詞であって、「時」、「分」等を含むものを抽出します。このようにして得られた結果をグラフにしたものを図３に示します。縦軸に興業収入、横軸に観客動員数をとって、映画がどれだけヒットしたのかを見ることができます。この場合、上映開始直後のデータでは、興業収入、観客動員数も少ないのは当然ですので、当該映画の上映終了後のデータで比較する必要があります。入場者の単価は、大雑把に見れば一定ですので、興業収入と観客動員数は比例関係になります。また、細かく見れば入場者の単価は、年齢層により階段状に変化していますので、興業収入が同額程度の「タイタニック」と「千と千尋の神隠し」では、「タイタニック」の方が観客動員数が少ないので、客単価は逆に高い、つまり「タイタニック」では、大人の観客の比率が「千と千尋の神隠し」より高いと推測することができます。

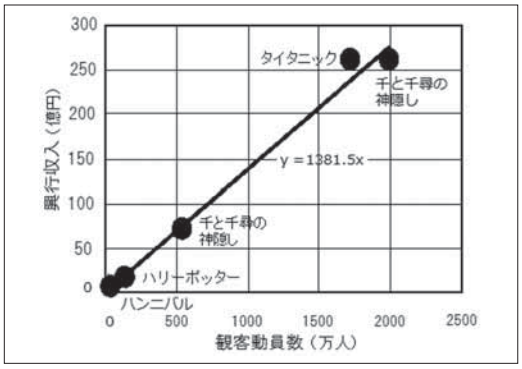

図３　映画の興行収入と観客動員数

もうひとつのサンプルとして、台風に関する情報をグラフにしたものを図４に示します。縦軸に「最大風速」、横軸に「中心気圧」をとると、反比例の関係が見えてきます。つまり、中心気圧が低いと最大風速が大きい（大型の台風）ことがわかります。

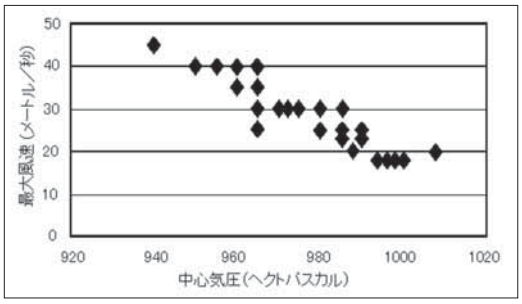

図４　台風の最大風速と中心気圧

おわりに

テキストマイニングは、今後ますます電子化され増大する情報の中から、自分の欲しい情報を効率的に抽出する手段として重要になってくると思われます。もちろん現在のテキストマイニングの技術が、十分検索ノイズ無く抽出できている訳ではありませんが、今後さらに進歩し人間が読んで抽出するのと同じくらいの正確さで、しかも瞬時に抽出できる日がやってくるのは、そう遠くない気がします。

（文責：研究推進部門　知財推進グループ　主幹　澤田史武）