5K－4

テキストマイニング：膨大な文書データからの知識獲得 —概要—

那須川哲哉　諸橋正幸　長野徹
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

近年，Knowledge Management という観点から，知識の共有化と再利用を意図して，様々な情報を電子化し蓄積する動きが急速に進んでいる．その際，計算機環境の向上により入力コストが低下していることから，特に形式を定めずに種々雑多な情報をとにかく入力してしまうケースが多い．そのためデータ中には，数値のように集計可能な定型データの部分と，文章のように単純には集計の出来ない非定型データの部分とが混在している．ところが，従来のデータマイニングでは定型データの部分のみが対象となるため，非定型データの部分はそのままでは利用できず，蓄積されたデータが充分に活用されていないことが多い．

本稿では，自然言語処理を適用することで非定型データである文章から概念情報を抽出し，定型データの部分と組み合わせた分析を行なう手法について，その概要を示すと共に，実際に弊社お客様相談センターで収集している 40 万件以上の問い合わせレポートを処理した結果を報告する．

2. 定型データと非定型データからの情報抽出

顧客からの問い合わせや営業活動報告，技術報告などの様々なデータが，計算機上に蓄積されるようになってきたが，その量が増大すると，単に特定のデータを再利用するだけではなく，全体的な傾向や特定のデータ群の特徴を分析することにより新たな知識を獲得できる可能性がでてくる．

例えば，データマイニングの技術を用いることにより，大量の顧客情報から顧客層を識別・分析し，効果的な営業活動に結びつけるといった応用が可能になる．

ところが従来のデータマイニング手法で分析対象になるのは，選択肢から得られた値や記述者名などのように限定された値や数値などの集計可能な定型データのみである．

大抵のデータに付随しているフリーコメントなどの文章情報は，基本的に非定型であるため，数値を中心とした定型データと違って集計が困難であり，従来のデータマイニングの分析対象にはならない．従って，従来のデータマイニング技術で非定型の文書データを扱うには，書かれた内容を何らかの基準で分類・評価し，記号化しなければならないが，膨大な量のデータに対し，これを人手で網羅的に行なうのは非常に困難である．そのため，蓄積された膨大の量のデータにおいて，非定型データの部分は充分活用されていないのが現状である．

例えば，アンケート調査結果の分析を考えてみると，データには，

- 与えられた選択肢から選んだ値
- 年齢などの数値

のような定型情報と，フリーコメントのような非定型情報が混在しているが，フリーコメントの部分は定型情報と異なり集計が困難なため，分析結果に参考情報としてそのまま添付されることが多い．そのため，量が膨大になると，フリーコメントの大半が活用されないことになる．しかし，基本的に，定型情報の部分には，アンケートを作成する側が前もって予想した回答の候補が列挙されており，予想できなかった回答は，非定型のフリーコメントの中に存在する．従って，非定型データの部分が活用できない損失は非常に大きい．

そこで我々は，自然言語処理により，従来扱えなかった非定型データの部分から，その内容を表現する概念情報を抽出し，定型データに変換することで，非定型データの内容と，定型データの内容を統合的に扱い，原データの全体に対しデータマイニングの手法を適用して分析する仕組みを開発した．

3. 処理の流れ

3.1. 概念の抽出

テキストに記述された内容を定型データと統合的に処理するためには，テキストから，その内容を表現する概念情報を抽出し，定型データに変換する必要がある．その際，単なる文字列ではなく概念として扱えるよう，同義表現を吸収すると共に，カテゴリを付加する．具体的には，まず形態素解析を適用し，自立語を抽出すると同時に，辞書を参照して同義語は特定の語に変換し，品詞とカテゴリ辞書を参照してカテゴリ付きの概念を抽出する．カテゴリは，対象データ，及び，分析目的により異なるが，例えば，「人名」「地名」「組織名」といったカテゴリが存在すれば，「日本ＩＢＭの長野氏が…」という文章からは，

Text Mining: Knowledge Acquisition from Voluminous Data with Texts —Framework—
Tetsuya NASUKAWA, Masayuki MOROHASHI, Tohru NAGANO
IBM Research, Tokyo Research Laboratory

- 「人名」の「長野」
- 「組織名」の「日本アイ・ビー・エム」

といった概念が抽出される.

ここで, テキストに記述された内容をいかにうまく表現できる概念を抽出できるかがマイニングの精度向上のポイントであるため, 単語レベルの概念を抽出するのではなく, 「○が□する」といった組合せの概念も抽出する [2].

3.2. 定型データと非定型データとの統合及びマイニング

非定型データから抽出された概念は, カテゴリ別に定型データと統合される. 原データが何らかのレポートであるとすれば, 各レポートの非定型データ (テキスト部分) から抽出されたカテゴリ別概念と各定型データとがレポート毎にまとめて蓄積され, 分析対象となる. マイニングを行なう仕組みとしては, Information Outlining [3] を利用することにより,

- カテゴリ別の各特徴量を含むデータの件数
- 期間毎のデータの件数
- 特定の時期に集中的に出現する概念の推移

などの情報を用いてユーザがインタラクティブに内容分析を行なう機能が提供できる.

4. お客様相談センターのデータを用いたマイニング実験

弊社お客様相談センターでは, PC 製品に関する問い合わせをフリーダイヤルで受けており, そこにかかってくる毎月 4 万件程度の問い合わせ内容を全て記録し, データベースに蓄積している. 記録内容には,

- 問い合わせ種別 (「技術的なQA」「購入相談」「要望」等)
- 問題種別 (「導入」「操作」「カタログ表記」等)
- 機種名

などの定型情報と, 実際の応対内容を文章で入力した非定型情報が含まれている. 今回は, 97 年 7 月から 98 年 4 月までの 10ヶ月分のデータ約 43 万件を用いて実験を行なった.

まず日本語形態素解析システム JMA [1] を用いて各月のデータの文章部分を形態素解析し, 副詞以外の自立語, 及び, その連鎖からなる複合語をキーワードとして抽出した. 例えば, 97 年 8 月の問い合わせレポート約 4 万件においては, レポート一件あたりの平均的な文章量は 150 文字程度であり, そこから抽出されるキーワード数は 36 語程度であった. この月のデータの総キーワード数は約 147 万語であったが, 異なり語は 10 万語程度であり, 2 回以上出現している語が全体の 95.6%, 10 回以上出現している語が全体の 89.1%, 1000 回以上出現している語が全体の 45.0%と, 同じ語が何度も繰り返し出現しているという結果が得られたため, 高頻度 (具体的には 10 回以上出現) の語を中心にカテゴリ辞書を構築した. カテゴリ項目としては, 体言に対しては,

「ハードウェア」「ソフトウェア」「専門用語 (ハードウェア, ソフトウェア以外の電算システム関連用語)」「コマンド」「組織名」「地名」「人名」

用言に対しては,

「好評」「問題」「質問」「要望」「感覚」「形容」「動作変化」

を設定し, 10ヶ月分のデータに対するカテゴリ辞書のエントリ数は約 1 万 6 千語となった. また, 表現のゆれなどを吸収するための同義語辞書のエントリ数は, 約 2 百語と, 比較的少なかった.

こうして構築した辞書を適用して, 抽出されたキーワードにカテゴリを付加し, 各レポートの属性として, 他の定型情報と共にマイニングを行なった. マイニング結果としては,

- 特定の商品に対して多い問い合わせの具体的内容
- 特定のトピック (例えばインターネット) に関する問い合わせの時期的傾向 (いつ頃どの機種に関連したものが多いか)
- 特定の機種に関して特に問い合わせが多いソフトウェア

などの情報を容易に引き出せるようになった.

5. まとめ

テキストを含む膨大なデータから知識を獲得する方法として, 自然言語処理により非定型のテキスト情報を定型情報に変換し, 原データの定型情報と統合的に分析を行なう手法を示し, 実際に, 40 万件を越す PC 製品に関する問い合わせを処理した結果を示した.

従来自然言語処理で対象としていた新聞記事やマニュアルのように文章作成の専門家が記述した文章と異なり, 今回対象とした問い合わせレポートは, 誤字脱字や書き手による個人差などが大きいと予想されたが, 本手法では基本的に文中の自立語部分しか抽出せず, また高頻度の語を中心に利用したことから, 定型情報に変換する上の質的な問題は特に目立たなかった.

また, 対象が PC 製品に関する問い合わせという比較的狭い領域であるため, 同じ語が繰り返し出現する割合が高く, 同義語辞書などの知識を構築する上では, ごく一部の高頻度語の情報を整備するだけでも, 高い効果が得られた.

さらに, 月毎にデータをまとめて処理する上では, 新たな月のデータにおいて新しく出現する異なり語の数が比較的少なく, 最初の月のデータに対する辞書データを整備した後, 継続してデータを追加していく際の労力は少なくなることが確認できた.

参考文献

[1] 丸山宏, 荻野紫穂: 正規文法に基づく形態素解析, 情報処理, Vol.35, No.7, pp.1293-1299 (1994).

[2] 諸橋, 那須川, 長野: "テキストマイニング：膨大な文書データからの知識獲得 —意図の認識—," 情報処理学会第 57 回全国大会, 5K-03 (1998).

[3] M.Morohashi, K. Takeda, H. Nomiyama, and H. Maruyama: Information Outlining - Filling the Gap between Visualization and Navigation in Digital Libraries, Intl. Symp. on Digital Libraries, pp.151-158 (1995).