多量の文書データからの特徴抽出支援システムの開発*

5M−07

渡邉 雄一　　榑松 理樹†
岩手県立大学 ソフトウェア情報学部‡

1. はじめに

近年、情報化の進展とともにコンピュータ上に蓄えられる情報が急増している。その結果、情報を利用する側は、人手では処理が困難な大規模のデータの中から目的にあった情報を検索しなければならないという、Information Overload（情報過多）と呼ばれる現象が生じている。その負荷を軽減するために、多くの手法が提案され、実現されてきている。しかし、利用者の要求を十分に満たす情報アクセス手法はまだ整っていないのが現状である。

特に、記述の柔軟性や容易性、可読性などの観点から、蓄えられる情報量の増加が顕著である自然言語で記述されたデータ（文書データ）は、そのまま計算機で処理することは困難であり、また、内容が膨大かつ多岐にわたるため、多くの場合、人手で処理しなければならないのが現状である。[1]

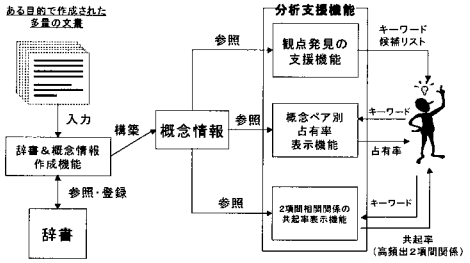
以上のような背景に基づき、本稿では、分析者の負担を軽減することを目的に、多量の文書データを効率良く処理するシステムを提案する。本システムは、始めに、ある目的で作成された多量の文書データから、頻度情報に基づき概念(単語とカテゴリーのペア)と、名詞概念と述語概念のペアとを抽出する。次にそれらの文書データにおける占有率、概念ペアまたは概念間の相関関係の共起率をデータ群の特徴として抽出することで、文書データ処理の支援を試みる。

2. システムの概要

図1に本システムの構成を示す。本システムは、辞書･概念情報構築機能と分析支援機能から構成される。以下、それぞれの機能について説明する。

2.1 辞書･概念情報構築機能

本機能では、入力文書群の各文を、既存の自然言語解析手法を用いて、概念情報と呼ぶ形式に変換する。ここで概念情報とは、①頻度情報に基づき抽出した単語とラベルのペア（以後、概念と表記）、②名詞概念と述語概念から構成される概念ペア、③文書ID、④原文、から構成される情報である。また概念情報に変換する際に、分析者が、システムが参照する概念辞書の構築を行う。



※キーワード＝単語もしくは、カテゴリー

図1：システム構成

2.2 分析支援機能

分析支援機能は、以下の3つの機能から構成される。

1）観点発見の支援機能：概念情報群を整理し、分析の観点となる単語を抽出・提示する機能。

2）概念ペア別占有率表示機能：分析者が与えた観点に基づいて、文書データの傾向の一つとして名詞概念と述語概念のペアの占有率を表示する機能。

3）2項間相関関係の共起率表示機能：分析者によって任意に指定された2つのキーの相関関係として、キー間の共起率を算出し、分析者に提示する機能。

3. システムの処理手順

2章で述べた各機能の処理手順について説明する。

3.1 辞書･概念情報構築機能

本機能の処理手順を図2に示す。本機能は、文章群変換フェーズと概念情報構築フェーズの二つから構成される。

文書群変換フェーズでは、入力文書群中の各文章に

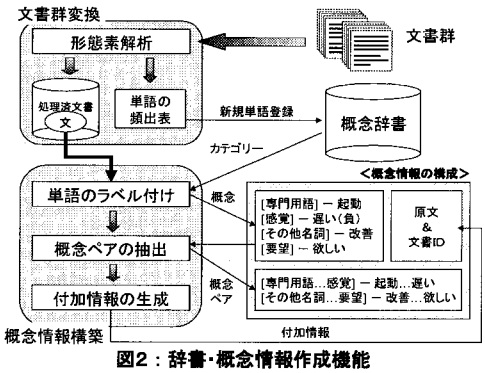


図2：辞書･概念情報作成機能

---

* Research on a Support System for Feature Extraction in Documents

† Yuichi Watanabe and Masaki Kurematsu

‡ Faculty of Software and Information Science, Iwate Prefectural University*

対し、形態素解析を行う。この結果を以後、処理済文書と呼ぶ。また同時に、全文書中に現れる単語の出現頻度を示す、単語の頻出表を作成する。概念辞書に無い頻出表中の単語については、手動で概念辞書に登録する。この際、概念辞書に登録する内容は、①単語名、②カテゴリー名（その単語の属する分類名）、③属性名（良い意味か悪い意味かを示す）、④同義語、⑤概念種類（名詞概念か、述語概念か）、の５つの情報であり、これらが概念辞書の持つ情報となる。

概念情報構築フェーズにおいては、処理済文書中の各文を次の手順で概念情報に変換する。（1）文中に出現する単語に対し、概念辞書を用い、ラベルづけを行う。ここでラベルには概念辞書中に登録されているカテゴリー名を利用する。（2）ラベル付けを行った単語のうち、同一文中に出現する名詞概念の単語と述語概念の単語のペアを、部分的な係り受け関係に注目し、抽出する**[2]**。（3）（2）までで作成した情報に、文書ID、原文を付加し、概念情報を構築する。

3. 2 分析支援機能

3. 2. 1 観点発見の支援機能

本機能では、ユーザに対し、分析観点の候補として、文書群中に現れる単語を、出現頻度順にラベルを付加した形式で提示する。分析支援機能における残りの二つの機能では、本機能で提示される単語またはカテゴリー名（ラベル名）を利用する。単語を検索する場合、同義語は同一の単語としてみなす。これは、以下の機能でも同様である。

3. 2. 2 概念ペア別占有率表示機能

本機能では、利用者が任意に指定した名詞概念の単語またはカテゴリーに対し、文書群中において、それらを含む概念ペアを、出現頻度の高い順に、その頻度とともに提示する。単語で指定した場合は単語で、カテゴリーで指定した場合はカテゴリーで概念ペアを検索する。この機能により、例えば、頻出する名詞概念と対になっている述語概念の種類を提示することができ、それを文書データの傾向を表す特徴、観点として利用すべきなのかを判断する指標として利用することが期待できる。

3. 2. 3 2項間相関関係の共起率表示機能

本機能では、利用者がキーとして、カテゴリー単独か、概念ペアの一方の概念を与え、それらに含まれる概念、概念ペア間の共起率を求めることにより、相関関係を提示する。ここで共起率は次式により算出する。

$$\text{共起率} = \frac{\text{両方のキーが含まれる文書数}}{\text{どちらか一方のキーが含まれる文書数}}$$

共起率が事前に決めた閾値よりも高い組合せをその２つのキー間の「特徴」とみなすことで、ある観点からみた時の文書群の「傾向」をユーザに提示することが期待できる。

4. 評価実験

2、3章で示した基本設計の有用性を評価するために、以下の要領で評価実験を行う計画である。なお、実験結果については、発表時に報告する予定である。

１．システムの実装

前述の基本設計に基づき、Perl/Tk を用いて、システムのプロトタイプを実装する。なお、辞書・概念情報構築機能の実装において、形態素解析部分には、奈良先端科学技術大学院大学の松本らが開発した茶筌**[3]**の利用を検討している。本プロトタイプを用いて処理を行うことにより、提案手法の評価を行う。

２．評価実験

入力データとしては、インターネット上の映画「タイタニック」に関するレビュー文書35件（1件あたり、平均400字～800字）を用いる予定である。それらの文書群から、それらを特徴づける概念や概念ペアの抽出を、分析者（被験者）単独の場合、およびシステムを利用した場合の双方で行う。そして、抽出した特徴の比較、抽出に要した作業量、時間を比較することにより、本システムの有用性を評価する。また、実験においては作業に対する錬度も影響することが考えられることから、錬度も評価の際に考慮する予定である。

5. おわりに

本稿では、多量の文書データに対し、頻度情報に基づき、単語やカテゴリーを抽出し、それらの出現頻度や相関関係を捉えることで、文書データからの特徴抽出支援を行うシステムについて、その枠組みについて述べた。

今後の課題としては、本枠組みに基づいたシステムを実装し、評価実験を行うことで、本システムの有効性を示すこと、実験結果に基づいて、問題点を洗い出し、それらを改善することによって、よりユーザを効率よく支援するシステムへと向上させることがあげられる。

参考文献

[1] 那須川哲哉、他：「テキストマイニング－膨大な文書データの自動分析による知識発見」,情報処理, 40巻4号, 1999年4月

[2] 諸橋正幸、他：「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3－75

[3] 松本裕治、他。http://chasen.aist-nara.ac.jp