B1-3 分類カテゴリの因子分解

星合 忠 (hoshiai@flab.fujitsu.co.jp)
(株)富士通研究所

1. はじめに

従来の情報検索や情報配信などの情報利用技術は単語・文字列レベルの分析であり、対象とする文書群をそのまま集合として記述することはできない。実際の情報検索の場面では、検索式は単語(文字列)"A"を含む文書集合を表すのにもかかわらず、ユーザは対象分野<A>に属する文書集合という異なる解釈に引きずられて検索を行いがちである。

情報検索で漠然とした検索式を入力した時や、文書の分類において網羅的でない分類カテゴリに分割した場合などに、対象となる文書集合の内部に、かなり異質の文書群が混在して、情報構造が分かりにくくなる場合がある。例えば、検索式として"インターネット"を入力した場合の検索結果の文書群には、ユーザの当初の意図とは異なり、〈インターネット技術〉、〈インターネット上の商売〉、〈インターネット・コミュニティ〉、などの〈インターネット〉分野の下位分類に属する文書だけでなくて、〈「名称に"インターネット"を含むニュースシステム」の全ての記事〉等の異種の情報内容が混在する。また、文書分類においては、例えば、新聞記事を〈政治〉、〈経済〉、〈社会〉の3つのカテゴリに分類した場合、〈社会〉記事の中には、〈事件〉、〈スポーツ〉、〈文芸・科学〉などの異質なジャンルの文書群が混在している。

このような場合、対象とする文書群の混沌とした情報構造を明確にするため、分類カテゴリをより細かい単位に分解して分析することを考える。

本研究では、この分類カテゴリの細分類を「カテゴリ因子」と呼ぶ。人手により定められた分類カテゴリに対して、カテゴリ因子は計算機により自動生成する。カテゴリ因子を媒介として、文書、分類カテゴリ、検索要求等を分析したり、マルチリソースの語彙の違いを吸収するため語彙グループを対応付けたり、直接的にカテゴリ因子のレベルで集合演算を行うことを意図している。

カテゴリ因子への分解では、例えば、〈経済〉や〈スポーツ〉などのカテゴリの記事データがあった場合、経済記事からは〈景気〉、〈企業活動〉、〈金融市場動向〉などのようなカテゴリ因子に分解することを想定している。

また、単語レベルの情報利用技術に関しては、扱えるデータ範囲が世の中の電子化されたテキストデータ全体と広いが、演繹、帰納などの推論は行えない。一方、知識処理技術では、扱えるデータ範囲が狭い(専用の知識ベース記述形式、推論ルール記述形式など)のが問題である。本研究では、その中間レベルとして分類カテゴリやカテゴリ因子のレベルでの情報利用技術を対象として、扱えるデータ範囲は広く保ちつつ演算可能な範囲を広げていきたい。

2. カテゴリ因子分解の準備

以下に、分類カテゴリの因子分解の準備としてのデータ処理について説明する。

2.1. 形態素解析

形態素解析プログラム Breakfast [1]を用いて、単語への分かち書きを行っている。以降の統計処理では、基本的にはこの分かち書きに基づいて単語がカウントされる。

2.2. 同義語処理

同義語を別々のエントリーとしてカウントすると、その同義語グループ全体の統計的特徴度が反映されないので、同義語処理を行う。処理内容としては、(1)同義語辞書の参照時に代表語以外の同義語が見つかった場合は、すべて代表語が出現したものとしてカウントする、また、(2)英数字のみからなる語は,まず、半角→全角変換と、小文字→大文字変換を行った後に、同義語辞書の参照を行う。

2.3. 特徴語抽出

本研究においては、分類カテゴリごとの特徴語を抽出する。各カテゴリにおける単語の出現確率分布の統計的特性として、Kullback-Leibler の情報量[2]を構成する式

$$J[w_i, c_j] \equiv p(w_i \mid c_j) \log \{p(w_i \mid c_j) / p(w_i)\}$$

w_i:単語, c_j:分類カテゴリ, $p(w_i \mid c_j)$: c_jにおけるw_iの出現確率, $p(w_i)$: 全文書におけるw_iの出現確率

を用いて、カテゴリ c_jにおける特徴語 w_iの尺度を設定した。ここで、$\Sigma\ J[w_i, c_j]$が、Kullback-Leibler の情報量であり、確率分布 $p(w_i)$と確率分布 $p(w_i \mid c_j)$とのずれの程度を表す。この$J[w_i, c_j]$の値の上位の単語を特徴語とする。なお、キーワードの選択にこの式を用いた例としては、 渡部[3]の研究等がある。

また、特徴語の対象となる品詞は、基本的には機能語を除き内容語を対象とした。

2.4. 主成分分析

特徴語の相関に関する固有値問題 $A \cdot x = \lambda \cdot x$ (A:相関行列, x:固有ベクトル, λ:固有値)を解けば、結果の固有ベクトルがが主成分に対応する。

主成分分析[4]を行うと、対象とするデータ分布の特徴をよく表す順に座標軸(主成分)を得ることができ、元の座標空間よりも少ない次元数で同程度の情報量を持つ座標空間を形成することができ、計算量を軽減することができる。また、上位の主成分軸には、固有値に比例して識別の情報量が集中し、統計的に重要な因子が対応すると考えられる。

3. カテゴリ因子分解の手法

3.1. 主成分レベルのカテゴリ因子分解

分類カテゴリと正の相関が高い主成分を、当該分類カテゴリに対応するカテゴリ因子の候補とする。

対象文書と主成分との相関は、主成分得点(主成分座標値)に現われるので、座標値の大きい主成分ほど当該文書との正の相関が強い。従って、分析空間上で当該分類カテゴリ中の文書の分布の重心の各主成分得点を比較して、上位の主成分をカテゴリ因子の第1近似とする。

上記の結果は、カテゴリ名と相関の強い主成分座標軸のラベル(主成分番号)との対応付けであって、その主成分の持つ意味までは自動的に求められない。そこで、主成分の表す意味の推定をするために、それぞれの主成分に対して相関の強い特徴語群を求めた。主成分と特徴語との相関の強さを表す数値は、因子負荷量として、以下のように求めて比較できる。

$$\text{因子負荷量} = \sqrt{\text{固有値}} \times \text{対応する固有ベクトル要素}$$

この2つの結果を統合するとカテゴリとカテゴリ因子との相関、および、カテゴリ因子と特徴語との相関が求まり、分類カテゴリがカテゴリ因子へ分解(細分類)される。例えば、後述の図2. のように、カテゴリG08G(交通制御システム)と相関の強いカテゴリ因子は、それぞれ、〈走行制御システム〉と〈カーナビゲーションシステム〉と類推される。但し、カテゴリ因子の名前は、対応する文書を読むか、特徴語リストを参照するかして、人間が類推する必要がある。

3.2. 語の相関レベルのカテゴリ因子分解

同一主成分に関して、異なる種類の特徴語群が存在する場合は、特徴語同士の相関を利用してさらにカテゴリ因子を細分割する。

特徴語同士の相関の有意性を調べるために、相関係数の簡易検定法[5]を応用した。本実験では定数パラメータを入れて、単語 w_iと単語 w_jの出現確率に関する相関係数 r_{ij}が、$r_{ij} > k \cdot 2 / \sqrt{n+2}$ の条件(n:データ数、k>0, t 検定の t 値=2 の場合に相当する)により、単語間の正の相関の有意性を判定した。

カテゴリ因子に対応する特徴語の全ての組合せについて、正の相関の有意性の簡易検定を行い、相関の有無による弱連結グラフを形成し、これにより因子の細分解を行う。

3.3.カテゴリ因子の冗長性除去

3.2 の結果は、同一カテゴリ中の複数の主成分の特徴語リストの中で特徴語が重複して現われるといった冗長性を残したものであるので、カテゴリ因子に対応する特徴語リスト同士の重複する部分を見つけて除去したり、カテゴリ因子の主成分得点寄与分の値の小さいものを除去する。

カテゴリ	特 徴 語 リ ス ト
G06G アナログ計算機	光,トランジスタ,カーペット,入力,MOS,乗算,神経,端子,関数,電気,層,回路,出力,電圧,薄膜,拡散,信号,乗算器,抵抗,マスク,抵抗器,光学,使用者,回路網,発生器,反転,推測,アレイ,磁束密度,快適,変調,学習,試料,数式,素子,極性,模式,表面,ベキ,級数,接続,可変,空間,特性,温度,ドレイン,シリコン,図,正弦波,入射,アルミ,波形,非線形,シフト,電流,不純物,照射,次数,エミッタ,熱流束,マトリックス,アナログ,蛍光体,ジェネレータ,周波数,発光素子,バリア,掛算,電極,メタル,ニューロン,オフセット,利得,差動,等価,暖房,電源電圧,BSO,リセット,レベル,ニューロネットワーク,ネットワーク,比較,ソース,赤外線,FET,Rbd,導電層,他方,室内,シナプス,コンタクト,コレクタ,ゲート,形,接地,条件
G07F コイン解放装置または類似装置	商品,カップ,自動販売機,販売,容器,飲料,投入,搬出,氷,扉,展示,貯蔵庫,冷却,開口,前面,庫,玉,台,見本,選択,製氷,棚,パチンコ,シュート,シュータ,コラム,リフタ,返却,払い出し,排出,落下,ボタン,商品取,コーヒー,冷飲料,原料,貯氷室,硬貨,コイン,記憶,スイッチ,本体,形成,価格,自販機,機体,吐出,メダル,プリペイドカード,表示,補充,合計,機,ラック,出口,収容,蛍光灯,空き缶,収納,予備,押ボタン,検知,ケース,業者,カップディスペンサ,通路,円板,台本,計数,川崎,磁気カード,機構,コールド,張出,取出口,上面,冷凍,移動,保管,ハンド,一杯,ホット,G07F,動作,補給,看板,駆動,数,ヒーター, 供給,内扉,操作,請求,停止,突出,上下,カロリー,売上,富士電機,排水
G08G 交通制御システム	車両,道路,位置,地図,表示,情報,走行,距離,ステップ,交差点,現在,経路,データ,交通,車,目的地,画像,受信,渋滞,制動,手段,判定,装置,自車,ナビゲーション,誘導,前方,交通量,運転,事故,進行,判断,ビーコン,場合,速度,地点,接近,GPS,障害物,計測,配車,地域,検索,自動車,駐車,先行,ルート,方位,車線,間,バス,位相,否,検出,信号機,方向,測位,上記,標識,抽出,移動,相対,左折,確定,存在,自動,地図帳,PS,出発,探査,衝突,認識,状況,ブレーキ,照明,動態,加速度,指示,自車両,物体,算出,衛星,地,旋回,座標,警告,車載,右折,NO,現在地,地区,ポーリング,タクシー,表示画面,エリア,画面,角,2値,送信,YES

図1. 特徴語抽出結果の例: 特許文書データ(一部)

4. 実験結果

4.1. 実験環境

対象データは、特許明細書文書の 808 文書、360 万語である。特許文書に付与されている IPC (国際特許分類)コード G05B-G08G を 16 個の分類カテゴリとした。なお、各カテゴリの特徴を切り分けて分析をより正確に進めるため、特許文書に G05B～G08G の内の2個以上の IPC コードが付与されている文書を分析対象から外した。

また、PCは、CPU: Pentium Pro 200MHz, 実メモリ 96MB, スワップ 210MB のものを用いた。

4.2. 因子分解前の特徴語

16 個の各カテゴリについて、上位 100 位までの特徴語を抽出した。図1. にその一部を挙げる。重複を除いた総特徴語数は、1248 語である。この特徴語出現確率に関する相関行列を求めて、主成分分析への入力として用いた。

主成分数の決定は、固有値 ≧ 1 の条件で定め、その結果 392 本の主成分を、寄与率約 88%で得て、主成分分析の分析空間とした。

4.3. 因子分解の結果

全カテゴリの重心点の主成分値を求め、各カテゴリに関して正の相関の強い上位 5 位の主成分を求めた。次に、それぞれの主成分に対して相関の強い特徴語を上位 100 位まで求めた。

上記の結果に対して、特徴語同士の相関を考慮しながら特徴語リストの細分解を行った。なお、判定条件では n=808, k = 1.0 とし、危険率約 5%での有意性の検定とした。

最後に、複数回出現する特徴語の除去を行った。この部分は一部プログラム未完のため手作業で行った。

最終的に、各カテゴリは、2～5個のカテゴリ因子に分解され、それぞれの主成分に対応する特徴語リストには、相関の強いものが残り、主成分間の特徴語の重複も減少した。

. カテゴリの因子分解結果の一部を図2. に挙げる。これによると、例えば特許文書のカテゴリ G08G(交通制御システム)は2つのカテゴリ因子〈走行制御システム〉と〈カーナビゲーションシステム〉に分解されると考えられる。

5. おわりに

文字列・単語レベルと知識レベルの中間として、カテゴリ因子を用いた情報利用技術を開発中である。

カテゴリ因子の応用について、以下が考えられる。

分類体系の分析や構築の支援:

カテゴリ因子を基にした文書の動的自動分類。

既存の分類カテゴリをカテゴリ因子で記述して、カテゴリ間の関係(階層、並置、排他など)をカテゴリ因子の構造から分析する。

文書特徴分析:

文書ベース中の各文書がどんなジャンルの要素(カテゴリ因子)を含んでいるか一覧する。

オントロジー:

複数の情報源における語彙の違いを吸収するため、カテゴリ因子のレベルで語彙の対応付けを行う。

文脈/状況識別子:

自然言語処理における文脈情報や、対話処理ソフトウェア等の対話状況における話題の変化に応じて、カテゴリ因子を対応させて記述し、文脈の同定や、文脈の類似性、連続性を分析に利用する。

検索の絞り込み支援:

カテゴリ因子の特徴語リストを検索の絞り込み時に提示する。

総合情報利用環境:

カテゴリ因子は、検索、自動分類、プッシュ配信、文書ベース管理等の特徴分析手段として利用可能であるので、カテゴリ因子による共通の統計的特徴量を基盤とする統合的情報利用環境が考えられる。

カテゴリ因子の課題としては、以下の項目が挙げられる。

・検索要求を任意の文書ベースの適切なカテゴリ因子へ変換する仕組み
・カテゴリ因子の命名
・実験の大規模化

[参考文献]

(1) 颯々野学, 難波功: 利用者による調節が可能な高速日本語形態素解析, 情報処理学会第 52 回全国大会講演論文集, 分冊3, 58-4, pp.75-76, (1996).
(2) 坂元慶行、石黒真木夫、北川源四郎: 情報量統計学、共立出版、(1983).
(3) 渡部勇: 緩い協調:協調フィルタリングシステム、情報処理学会ヒューマンインタフェース研究報告、91-HI-35, 35-24, (1991).
(4) 大隅昇、L. ルバール、A. モリノウ、K. M. ワーウィック、馬場康雄: 記述的多変量解析法、日科技連、(1994).
(5) 上田太一郎: 相関があるかを見つける簡便法, オペレーションズ・リサーチ, 1997 年 7 月号, pp.493-496, (1997).

カテゴリ	カテゴリ因子識別番号(仮名)	特徴語リスト
G06G アナログ計算機	8 (ニューラルネット)	神経,回路網,学習;ニューロン,ネットワーク,模式,カーペット,シナプス,快適,推測,使用者,暖房,電気,非線形,表面,室内,条件,温度,関数,シリコン,正弦波,磁束密度
	2 (電気回路)	回路,電圧,抵抗,トランジスタ,接続,出力,エミッタ,コレクタ,接地,ドレイン,差動,特性,電源電圧,MOS,抵抗器,素子,ソース,波形,他方,ゲート,電極,反転,等価
	6 (光学系)	光,アレイ,使用者,非線形,入射,周波数,信号,光学,照射
	21 (半導体基板)	層,シリコン,アルミ,メタル,薄膜,バリア,BSO,蛍光体,導電層,不純物,拡散,表面,コンタクト,試料
G07F コイン解放装置または類似装置	16 (自販機商品系)	自動販売機,冷飲料,販売,商品,一杯,飲料,富士電機,氷,製氷,貯氷室,吐出,冷却,容器,コールド,ホット,機体,ラック,開口,川崎,売上,原料,ケース,搬出,前面,商品取,扉,落下,合計,棚,シュータ,選択,台本
	5 (自販機入金系)	取出口,台,上下,収納,機体,形成,上面,出口,排出,硬貨,庫,見本,通路,展示,計数,台本,本体,シュート,請求,リフタ
G08G 交通制御システム	3 (走行制御システム)	車,障害物,前方,ブレーキ,距離,車両,相対,走行,位置,速度,角,道路, 検出,自車,判断,先行,存在,画像,ステップ,制動,運転
	17 (カーナビゲーション)	GPS,ナビゲーション,衛星,方位,算出,地図,経路,目的地,現在地,出発,現在,道路,交差点,進行,方向,測位,手段,YES,NO,否,車載,座標

図2. 分類カテゴリの因子分解の例: 特許文書データ(一部)