1V－9

分類された文書集合における特徴的なキーワードパターンの抽出

小中 裕喜
三菱電機株式会社 先端技術総合研究所

1 はじめに

検索対象に対する知識や検索要求が曖昧で適切な検索式が構成できない場合、文書集合を分類/クラスタリングするアプローチが有望である。同一の検索要求に対して関連度の高い文書は互いに類似しているという仮説に基づき、類似した文書をまとめることにより、再現率は低いかもしれないが適合率の高い検索が期待されるクラスタが構成される。ここで各クラスタの特徴が何らかの形で提示されていれば、利用者にとって興味深い文書が集まったクラスタを見い出すことが容易となる。そのような特徴づけを自動的に行う1つの方法は、各クラスタに属する文書のキーワードに頻度などに応じて重みを与え、クラスタごとに最も重みを得たキーワードをいくつか提示することである。しかしながら、そのような方法ではキーワードの共起情報が考慮されず、また他のクラスタにも頻出するキーワードが選択されることもあるため、各クラスタの特徴を十分表しているとは言えない。

本稿では、分類/クラスタリングされた文書集合から、各クラスタを他のクラスタと識別するキーワードの組合せを抽出するアルゴリズムを提案する。抽出結果は各クラスタの概要を把握する手がかりとして、検索対象を絞り込むための候補検索式として、また後続の文書をインクリメンタルに分類するための知識として用いられる。

2 問題の記述

あらかじめ分類/クラスタリングされた文書集合において、各文書にはキーワード集合が付与されているものとする。$\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$ を全キーワード集合、$\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ を全文書集合、$d_i \subseteq \mathcal{W}$を各文書に対応するキーワード集合とする。また$\mathcal{C} = \{c_1, c_2, \ldots, c_l\}$ を全クラスタの集合とする。ただし $c_i \subseteq \mathcal{D}$, $\bigcup_i c_i = \mathcal{D}$である。パターン $p \subseteq \mathcal{W}$をキーワードの集合とし、$|p| = k$ のとき k-パターンと呼ぶ。パターン p を含む文書集合を $S_p = \{d_i | p \subseteq d_i\}$ とする。

パターン p のクラスタ c_iにおける支持率 sup と確信度 cnfを $sup = |S_p \cap c_i|/|c_i|$, $cnf = |S_p \cap c_i|/|S_p|$ とする。これらはパターン p をAND検索式、クラスタ c_iを適合文書集合とみなした時の再現率、適合率に相当する。

各クラスタにおいて与えられた最小支持率 $minsup$以上の支持率をもつパターンをラージパターンと呼ぶ。ラージパターンのうち、与えられた最小確信度 $mincnf$及びその部分パターンの確信度の最大値以上の確信度を持つものを特徴パターンと呼ぶ。また最大確信度 $maxcnf$が与えられた場合、それ以上の確信度をもつ特徴パターンは十分特徴的であるとみなされ、それを含むパターンは特徴パターンから除外される。

3 アルゴリズム

上記の問題を解くアルゴリズムは入力フェーズとマイニングフェーズからなる。

入力フェーズではクラスタごとに各文書のキーワード情報が入力される。このとき、どのクラスタにおいても $minsup$ %以上の文書に含まれていないキーワードは、決して特徴パターンに現れることがなく、結果に影響を与えることもないため、全文書情報から削除される。

マイニングフェーズは図1に示すように各クラスタ c_i ごとに、パターンの構成キーワード数 kを k_{max}まで増やしながら、クラスタ内処理とクラスタ間処理を交互に行っていく。クラスタ内処理は基本的にAprioriアルゴリズム[2]を用いてラージパターン集合 L_kを生成するが、ラージ2-パターンの候補集合 C_2の生成を効率化するために、ラージ1-パターン集合 L_1を収集した後にDHPアルゴリズム[3]のハッシング法を適用している。filter()とfilter_and_count_other_support()は、ラージ1-パターン集合 L_1に現れないキーワードの情報を一時的に削除する。各文書の残りのキーワード情報及び各パターンのキーワード情報はビットベクタで表現し、その後の処理を効率化している。count_support()はパターン pのクラスタ c_iからの支持 $|S_p \cap c_i|$ を計算し、count_other_support()とfilter_and_count_other_support()は各ラージパターンの他のクラスタからの支持 $|S_p \cap \overline{c_i}|$ を計算する。ただし前回の反復でラージパターンを支持する文書がなかったクラスタは計算を省略する。apriori_gen()とgen_candidate()は[2]及び[3]と同様であるが、最大確信度に関する変更がなされている。最大確信度が与えられた場合 $maxcnf$以上の確信度をもつラージパターンが次の反復の候補パターンとして用いられることはない。

Extraction of discriminative keyword patterns from categorized documents
Hiroki Konaka
Advanced Technology R&D Center, Mitsubishi Electric Corp.
8-1-1 Tsukaguchi-Honmachi, Amagasaki, Hyogo 661-8661, Japan

```
forall cluster c_i ∈ C do begin
    /* find discriminative 1-patterns */
    L_1 := { large 1-patterns };
    set all the buckets of H_2 to zero;
    D_i := ∅;
    forall document d ∈ c_i do begin
        filter(d, L_1, d̂);                    /* d̂ ⊆ d */
        D_i := D_i ∪ {d̂};
        forall 2-subsets x of d̂ do begin
            H_2[h_2(x)] + +;
        end
    end
    D̄_i := ∅;
    forall cluster c_j ∈ C, c_j ≠ c_i do begin
        forall document d ∈ c_j do begin
            filter_and_count_other_support(d, L_1, d̂);
            D̄_i := D̄_i ∪ {d̂};                 /* d̂ ⊆ d */
        end
    end
    forall pattern p ∈ L_1 do begin
        p.cnf := |S_p ∩ c_i|/|S_p|;
    end
    G_1 := {p ∈ L_1 | p.cnf ≥ mincnf};
    L_1 := {p ∈ L_1 | p.cnf < maxcnf};

    /* find discriminative 2-patterns */
    gen_candidate(L_1, H_2, C_2);
    count_support(D_i, C_2);
    forall pattern p ∈ C_2 do begin
        p.sup := |S_p ∩ c_i|/|c_i|;
    end
    L_2 := {p ∈ C_2 | p.sup ≥ minsup};
    count_other_support(D̄_i, L_2);
    forall pattern p ∈ L_2 do begin
        p.cnf := |S_p ∩ c_i|/|S_p|;
    end
    G_2 := {p ∈ L_2 | p.cnf ≥ mincnf,
                      p.cnf > p.maxAncestralCnf};
    L_2 := {p ∈ L_2 | p.cnf < maxcnf};

    /* find discriminative k-patterns (k ≥ 3) */
    k := 3;
    while (|L_{k-1}| ≥ 2 & k ≤ k_max){
        C_k := apriori_gen(L_{k-1});
        count_support(D_i, C_k);
        forall pattern p ∈ C_k do begin
            p.sup := |S_p ∩ c_i|/|c_i|;
        end
        L_k := {p ∈ C_k | p.sup ≥ minsup};
        count_other_support(D̄_i, L_k);
        forall pattern p ∈ L_k do begin
            p.cnf := |S_p ∩ c_i|/|S_p|;
        end
        G_k := {p ∈ L_k | p.cnf ≥ mincnf,
                          p.cnf > p.maxAncestralCnf};
        L_k := {p ∈ L_k | p.cnf < maxcnf};
        k + +;
    end
    Answer_i := ⋃_k G_k;
end
```

図 1: マイニングフェーズ

4 予備評価

CD-毎日新聞94年版[1]を対象とした予備評価を行った。各記事にはキーワード集合に加え、RWCテキストデータベース[4]においてUDCコードが付与されている。そこで、UDC主標数の上位2けたが「65」の2822記事(経営関係)を、下位2けたまでで分類し、31のクラスタを得た。総キーワード数は71100である。これに対して上記のアルゴリズムをSun Ultra 1の上で適用したところ、$minsup = 10\%$, $mincnf = 60\%$, $maxcnf = 80\%$, $k_{max} = 3$として、4629の特徴パターンが約3.1秒で得られた(I/O時間は除く)。具体的な特徴パターンとしては、例えばUDC658.15(私有財務管理.企業の財務管理)に対しては「決算&不良債券」(同19%、78%)、UDC658.16(財務整理)に対しては「救済&信用組合」(支持率17%、確信度100%)、「合併&信用組合」(同17%、85%)などが得られた。このように「不良債券」「信用組合」など単独のキーワードでは十分特徴づけできないクラスタに対しても、キーワードの組合せによって特徴づけられる場合があり、本アルゴリズムの有用性が示されている。一方、1000を超える特徴パターンが生成されるクラスタもあれば、幅広い記事が集まって特徴パターンが見い出せない場合も見られ、適応的なパラメータチューニングやシソーラス援用の必要性が示唆された。

5 おわりに

本稿では、あらかじめ分類/クラスタリングされた文書集合から、各クラスタを特徴づけるキーワードの組合せを抽出するアルゴリズムを提案した。現在、本アルゴリズムを応用した検索/分類システムを開発しているが、詳細は別稿にゆずる。今後の課題としては、適応的制御の導入、シソーラスや係受け関係などの考慮、階層的分類への適用などがある。

参考文献

[1] http://cactus.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html.

[2] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. of VLDB'94*, pp. 487–499, 1994.

[3] J. Park, M.-S. Chen, and P. Yu. An Effective Hash Based Algorithm for Mining Association Rules. In *Proc. of SIGMOD'95*, pp. 175–186. ACM, 1995.

[4] Real World Computing Partnership. RWC Text Database: RWCP-DB-TEXT-95-3, 1995. http://www.rwcp.or.jp/wswg/rwcdb/text/index.html.