文書検索のための大規模文書クラスタリング

岩山 真
(株) 日立製作所基礎研究所
iwayama@harl.hitachi.co.jp

徳永 健伸　桜井 直之
東京工業大学工学部
{take,naoyuki}@cs.titech.ac.jp

概 要

大規模な文書集合をクラスタリングするための近似法を幾つか提案する．また，近似クラスタリングにより構成したクラスタ木上で文書の二分木検索を行い，本近似法を評価する．近似クラスタリングは，厳密 (非近似) クラスタリングをサブルーチンとして使いクラスタ木の層を構成していくため，厳密クラスタリングを文書データベース全体に適用する場合に比べ，時間／空間的資源を大きく軽減できる．精度の点では，自己検索（自分自身を一位で検索できる割合），トピック割り付けにおいて，特にトップダウンにクラスタ木の層を構成する近似法の優位性が確認できた．

1 はじめに

文書検索において，検索対象の文書集合が大きくなると高速/高精度な検索が困難になる．例えば，検索要求との比較を全文書に対して行う単純な網羅検索では，文書数を N とすると $O(N)$ の時間を要する．この問題を解決するために本論文では，**クラスタ検索** [4] と呼ばれる検索法を用いる．クラスタ検索では．文書集合から自動的に二分木を構成し(このステップを**クラスタリング**と呼ぶ)，検索要求を各クラスタ (ノード) と比較することによって，検索要求と類似する文書を指定された数だけとりだす (このステップを**検索**と呼ぶ)．最も単純な検索法は**二分木検索**であり，クラスタ木の根からトップダウンに木をたどり，指定された数の文書を含むクラスタを探す．二分木検索は，平均 $O(\log_2 N)$ の検索時間しか必要としないので，網羅検索 ($O(N)$) に比べ高速な検索が可能である．ところがクラスタ検索は，現実のシステムにおいてほとんど用いられていない．大きさ N の文書集合をクラスタリングするのに $O(N^2)$ の時間/空間的計算資源を要してしまうからである．本論文では，クラスタリングのスケーラビリティを高めることを目標として，大規模文書集合のための近似クラスタリング法を幾つか提案し評価する．

2 厳密 (非近似) クラスタリング

クラスタ検索におけるクラスタリングの目的は，検索を行った際に高い精度を与えるようなクラスタ木を構築することである．まず，クラスタリングで使う尺度として**自己再現率 (self recall)** を定義する．あるクラスタ C に関する自己再現率 $SR(C)$ を以下のように定義する．

$$SR(C) = \prod_{d \in C} P(C|d). \quad (1)$$

ここで，$P(C|d)$ は，文書 d から文書集合 C への方向性のある類似性と解釈できるため，自己再現率は，クラスタに含まれる各文書が自分自身を含むクラスタを検索結果として見つけることができる確率，と解釈できる．$P(C|d)$ の推定法については Iwayama 等の推定法 [1] を用いる．

文書集合 $\mathcal{D}$ がクラスタの集合 $\{C_1.C_2....\}$ に分割されているとき，その文書集合 $\mathcal{D}$ に対する自己再現率を以下のように定義する．

$$SR(\mathcal{D}) = \prod_{C \in \mathcal{D}} SR(C) = \prod_{C \in \mathcal{D}} \prod_{d \in C} P(C|d). \quad (2)$$

これは，文書集合全体に関する自己検索の精度に関連する．自己再現率の定義を用いるとクラスタ検索のためのクラスタリングの目的は「文書集合 $\mathcal{D}$ が与えられた時，$SR(\mathcal{D})$ が最大となる分割を見付けること」であると定式化できる．

文書集合 $\mathcal{D}$ に対して階層的な二分クラスタ木を構築するには，以下に示す凝集型アルゴリズムを適用すればよい．

1. 初期クラスタ集合を，$\mathcal{D}$ 内の各文書それ自体のみからなるクラスタの集合とする．

2. マージにより $SR(\mathcal{D})$ の増分が最大になるようなクラスタのペアを見つけ実際にマージする.
3. 残りのクラスタの数が 1 でなければステップ 2 に戻る.

アルゴリズムの詳細については, [3. 2] を参照されたい.

3 近似クラスタリング

厳密クラスタリングの問題点は, 必要とする時間/空間的計算量が $O(N^2)$ である点である. これは, ほとんど全ての階層的クラスタリングにあてはまるため, クラスタ検索は大規模な文書集合に適用されなかった. 本節では, 厳密クラスタリングの計算量を大幅に軽減する近似クラスタリングを幾つか提案する.

最初の方法は, **分割-収集法**である (図 1 参照). 分割-収集法ではまず, 文書集合 $\mathcal{D}$ を L 個の部分集合に分割する. 本論文では乱数分割を行った. それぞれの部分集合に含まれる文書数は計算量的に扱える数にする. 次に, 各部分集合に対して厳密クラスタリングを適用しクラスタ木を構成する. 構成したクラスタ木を水平にスライスすることで, 各クラスタ木から M 個ずつクラスタを収集する. これにより $L \times M$ 個のクラスタができる. もし, $L \times M$ が計算量的に扱える数なら, このクラスタ集合に対して厳密クラスタリングを適用する. $L \times M$ が依然として大きな数ならば, さらに一連の分割/クラスタリング/収集手続きを $L \times M$ が計算量的に扱える数になるまで繰り返す.

二番目の方法は, **サンプリング-分類法**である (図 1 参照). サンプリング-分類法ではまず, 文書集合 $\mathcal{D}$ から S 個の文書をサンプリングする. 本論文では乱数サンプリングを行った. そして, S 個の文書に対して厳密クラスタリングを適用し, 種となるクラスタ木を構成する. 次に, 残った $|\mathcal{D}| - S$ 個の文書を, 種クラスタ木の葉ノードのいずれかに分類する. 最適な葉ノードを見つける際には, 種クラスタ木上で二分検索を行う. $|\mathcal{D}| - S$ 個の文書を分類した後には, 種クラスタ木の各葉ノードは幾つかの文書を持つことになる. 言いかえると, 文書集合 $\mathcal{D}$ が S 個の部分集合に分割され, それぞれの部分集合が種クラスタ木を介して検索可能になっている. 最後に, もし各部分集合の大きさが計算量的に扱える数なら, その部分集合に対して厳密クラスタリングを適用し各々のクラスタ木を構成する. もしある部分集合の大きさが計算量的に扱えなければ, その部分集合に対して更にサンプリング/分類を適用する.

以上の二つの方法は, クラスタ木の層構造を構成していることに相当する. 分割-収集法では層をボトムアップに構成し, サンプリング-分類法ではトップダウンに構成する. よって, 以後, 前者を**ボトムアップ近似クラスタリング**, 後者を**トップダウン近似クラスタリング**と呼ぶことにする. また. トップダウン法では, 種クラスタ木の各葉ノードに分類された文書を, 種クラスタ木に反映するか否かという二つの選択がある. 反映する方法を, **更新型トップダウン近似クラスタリング**と呼ぶ.

4 実験

4.1 データと設定

評価実験では, RWC 文書データベース (RWC-DB-TEXT-95-3) [5] の一部分を用いた. このデータベースは CD-毎日新聞 94 年版の 30.207 記事に対し, UDC(国際十進分類法) コードを付与したものである. 本実験では, 計算機環境の制限により, 以下の基準で記事を選択した. まず, 相対的に頻度の高い 300 番台の主標数 (「社会科学, 法律, 行政」に相当) のうち 100 記事以上に付与されているものを 35 個を選択し, これらの主標数が付与されている 14.904 記事を対象データとして選択した. これらの記事に付与されている選択した 35 以外の主標数は削除した. 各記事に付与されている UDC コードの平均は 1.46 個である. また, 同じ文書集合に対して形態素情報を提供する別の RWC 文書データベース (RWC-DB-TEXT-95-1) を用いて, 各文書の単語ベクトルを作成した. ベクトルの要素となる単語には,「名詞」と「未知語」を用いた. 以上によって作成した文書集合を Mainichi300 と呼ぶ.

実験では, 以下の 5 つの文書検索法を比較する.

- 網羅検索 ("exhaustive" で表記)
- 厳密クラスタリングを用いたクラスタ検索 ("precise" で表記)
- ボトムアップ近似クラスタリングを用いたクラスタ検索 ("bottom up" で表記)
- トップダウン近似クラスタリングを用いたクラスタ検索 ("top down" で表記)
- 更新トップダウン近似クラスタリングを用いたクラスタ検索 ("updated top down" で表記)

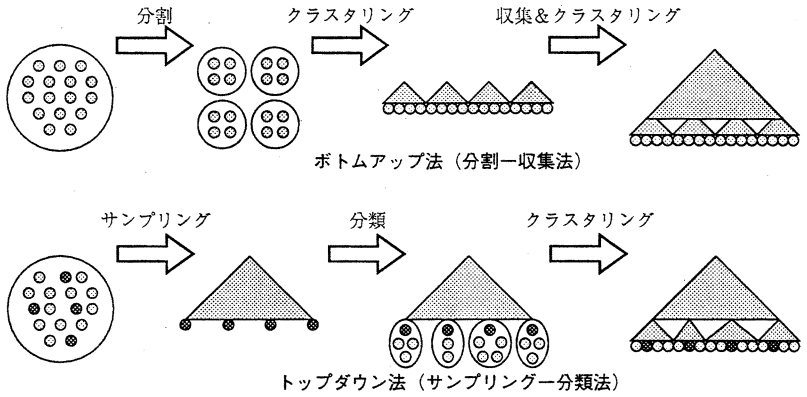

図 1: 近似クラスタリング

exhaustive	0.993
precise	0.654
bottom up	0.555
top down	0.991
updated top down	0.871

表 1: 自己検索の精度 (正解率)

4.2 自己検索 (closed set)

連想検索の精度を調べる最も単純な実験は，自己検索である．この実験では，与えられた文書集合内の全ての文書を検索入力として使う．実際に検索を行い，検索入力と同一の文書が 1 位の順位で検索できた場合，その検索結果を成功とみなす．

表 1 に実験結果を示す．実験では，近似クラスタリングにおけるパラメータ L、M、S はそれぞれ 10. 160. 1.000 に設定した．

表 1 から当然ではあるが網羅検索の優位性がわかる．注目して欲しいのは，トップダウン近似クラスタリングを用いたクラスタ検索も，網羅検索と同等の精度 (99%を上回る正解率) を出している点である．これは，網羅検索の結果が 14.903 回の比較 (理論的) によって得られているのに対し，クラスタ検索の結果は平均 28 回の比較しか必要としないことを考えると注目すべき結果である．しかし，全てのクラスタ検索の結果が高い精度であるとは限らない．例えば，ボトムアップ近似クラスタリングを用いた場合，56% の正解率しか得られていない．厳密クラスタリングでさえも 65% にとどまっている．これらの結果から，ボトムアップにクラスタ木を構築していく手法よりも，トップダウンに構築していく手法のほうが優位であると考えられる [1]．考えられる理由の一つは，トップダウンな方法では，クラスタ木を構築していく際，つまり，種クラスタ木の葉ノードに文書を分類するプロセスにおいて，検索そのものを行っている点である．これにより，ある文書 d は，クラスタ木の構築時と自己検索の評価時に全く同じ種クラスタ木をたどることになる．ここで注目して欲しいのは，更新トップダウン法では，分類された文書情報が種クラスタ木に反映されるため，クラスタリング終了時の種クラスタ木は与えられた文書集合全体を特徴化したものになっていることである．よって，文書 d は，クラスタリング時と検索時で全く同じ種クラスタ木をたどるわけではないため，自己検索の精度もトップダプン法に比べると低くなっている．しかし，更新トップダウンが持つ汎化能力は，未知データに対して有効であることが期待できる．次節では，この点についての実験を行う．

4.3 トピック割り付け (open set)

本節では，未知文書 (クラスタリングに使わなかった文書) に対するクラスタ検索の有効性をトピック割り付けにより調べる．トピック割り付けとは，あらかじめ定義されたトピックの中から任意個のトピックを文書に割り付けるタスクである．自動的なトピック割り付け法としては，k-NN 法 (k-Nearest

[1] 厳密クラスタリングもボトムアップにクラスタ木を構築していく．

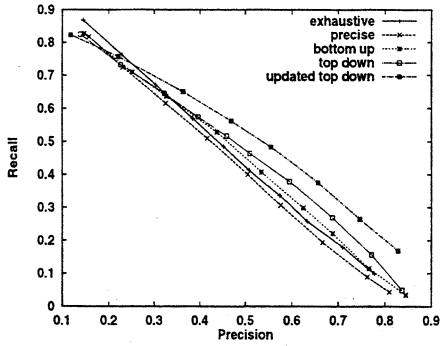

図 2: 再現率/適合率 (トピック割り付け)

Neighbor classifiers) [6] を用いた. k-NN 法ではまず, あらかじめ専門家によりトピックが割り付けられている文書集合の中から 割り付け対象文書 d に近いものを k 個検索し, 検索された k 個の文書に割り付けられているトピックを用いて d にトピックを割り付ける. 実験では, トピックとして UDC コードを用い, 4-fold のクロスバリデーションを行った.

図 2 に, トピック割り付けにおける再現率/適合率を示す. 近似クラスタリングにおける L, M, S は, それぞれ 10, 130, 800 に設定した. 図から, 差はそれほど大きくはないものの, 更新トップダウン法の優位性がわかる. 単純なトップダウン法も, 残りの 3 つと比べると優れている. この実験でも, トップダウンにクラスタ木を構築していく方法の優位性, つまりクラスタリング時と検索時で同じ原理を用いることの優位性を確かめることができた. 更新トップダウン法が単純なトップダウン法よりすぐれているのは種クラスタの違いが主な要因である. 前節でも述べたように, 更新トップダウン法における種クラスタ木は, クラスタリングの対象文書全体を特徴化しているため, 汎化能力も高いことが予想できる. 本実験結果は, この予想の妥当性を支持している.

4.4 クラスタリングの時間的効率

クラスタリングに要する時間に関しては, 本来なら理論的な計算量を議論することも必要であるが, 近似クラスタリングの計算量はそのパラメータに依存するため理論的解析が困難である. よって, 今回は HP K460 (主記憶 2 GB) 上で, 実際に Mainichi300 をクラスタリングした時の実行時間 (real CPU time) を幾つかのパラメータ設定のもとで測定した. 詳細は省くが, 近似クラスタリングを用いることで, 厳密クラスタリングに比べ計算時間を平均 100 倍以上高速化することができた.

5 おわりに

本論文では, クラスタリングにおける規模の問題を克服するために, 近似クラスタリングの方法を幾つか提案し, CD-毎日新聞 94 年版を用いた実験でその有効性を確認した.

参考文献

[1] M. Iwayama and T. Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of 4th Conference on Applied Natural Language Processing*. pp. 162–167. 1994.

[2] M. Iwayama and T. Tokunaga. Cluster-based text categorization: A comparison of category search strategies. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 273–280. 1995.

[3] M. Iwayama and T. Tokunaga. Hierarchical bayesian clustering for automatic text classification. In *Proceedings of the International Joint Conference on Artificial Intelligence*. pp. 1322–1327. 1995.

[4] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Publishing Company. 1983.

[5] The text group of RWC Database Workshop Group. `http://www.rwcp.or.jp/WSWG/rwcdb/text/index.html`. 1995.

[6] S. M. Weiss and C. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann. 1990.