ブログ分析のための制約付きクラスタリングシステムの実装

青島　傳隼†　　福田　直樹†　　横山　昌平†　　石川　博†
†静岡大学情報学部情報科学科

1　はじめに

Web上のサービスとしてのブログは，ここ数年で急速に普及した．従来一般的であったgooブログなどのブログサービスに加え，Twitterを代表とするマイクロブログサービス，写真や画像を主体に更新するフォトログサービスなど，ブログの多様性が重要な研究対象となりつつある．また，ブログの分析を行う目的も多様になってきている．例えば，あるキーワードについてブログの記事群を用いて分析を行いたい場合においても，そのキーワードの評判情報を知りたいのか，どのキーワードと相関性が高いのかなど，ブログマイニングの目的は一様ではない．本論文では，多様なブログサービスの普及と，それぞれのユーザが持つ分析目的・意図に対応するために，制約付きクラスタリング手法を採用したシステムの実装について述べる．

2　制約付きクラスタリング

制約付きクラスタリング[1]は，クラスタリングの対象となる要素に対して，同一クラスタへの所属の可否や所属クラスタの指定などの制約を加え，その制約を満たすようにクラスタリングを行う手法である．

本研究では，Chengらの提案した制約を用いた非階層型クラスタリング手法であるCLWC(Constrained Locally Weighted Clustering)法[2]を採用し，制約の種類は，データ間の関係を指すmust-linkとcannot-linkの2種類の制約を扱う．must-linkで結ばれた二つのデータは同じクラスタに，cannot-linkで結ばれた二つのデータは異なるクラスタにそれぞれ属するようにクラスタリングされる．

3　システムの実装

本システムでは，複数のブログサービスの記事を扱い，制約付きクラスタリングを適用可能としている．制約の付与方法としては，ユーザによる手動での付与方法，ユーザの設定した条件に従い制約を付与する半自動的な付与方法の2種類が考えられる．本研究ではこの二つの制約の付与方法に注目してシステム「Mi-ke」の実装をJavaを用いて行った．制約付きクラスタリングは，制約を付与し，結果を確認した後に，再び制約の付与・削除を行うことが一般的である．そのため，本システムでも再クラスタリングが可能な機能を持たせた．図1に，Mi-keにおける処理の流れを示す．

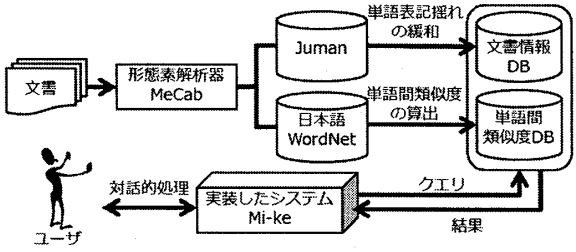

図1: 実装したシステム「Mi-ke」の処理の流れ

3.1　前処理

本システムの対象となる記事群は一般的なブログ記事を想定している．ブログ記事は人の書いた記事であることが一般的であるため，表記揺れの問題が発生する．また，対象がマイクロブログ記事の場合，記事が短いため，単語の頻度に基づく$TF \cdot IDF$の特徴量のみでは，各文書の特徴を効率的に表すことが難しい．これらの問題を解決するために，Jumanを用いた表記揺れの緩和と，日本語WordNetを用いた概念に基づく単語間類似度の算出を行っている．詳しくは文献[3]を参考されたい．

3.2　手動での制約の付与

ユーザの背景知識や目的に応じて制約をデータ間に付与する方法の一つに，ユーザ自身がデータを確認しながら手動で制約を付与する方法がある．Mi-keでは，データ間に“線”を引くような操作で制約の付与を行えるようにし，視覚的に各クラスタに属するデータや制約の効果を確認しながら制約を付与していくためのユーザインタフェースを実装している．

3.3　半自動的な制約の付与

大規模な記事群を分析対象としたとき，全ての記事を確認しながら，一つ一つ手動で文書間に制約の付与を行うことはユーザにとって負担となる．Mi-keでは，複数記事に対する制約付与機能を用意した．複数記事に対する制約付与機能とは，ユーザが複数の記事に対して制約の付与を行いたい場合，ユーザの負担を軽減する機能である．例えば，大規模な記事群の中から一部だけ手動で似た記事をまとめた後で，それらを制約

On Implementing Constrained Clustering System for Blog Analysis
Tsugutoshi Aoshima†, Naoki FUKUTA†, Shohei YOKOYAMA†and Hiroshi ISHIKAWA†
†Department of Computer Science, Faculty of Informatics, Shizuoka University

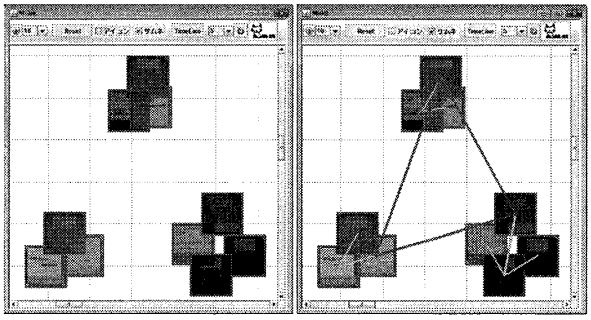

図2: 複数記事に対する制約付与，左：複数の記事を選択した状態（赤い太枠で囲まれたもの），右：選択中の記事に対して半自動的な制約の付与を行った状態（緑線：must-link，赤線：cannot-link）

付きクラスタリング結果に反映させたい場合，まとめられた記事群が同じクラスタに，別々にされた記事群は異なるクラスタに属させる必要があると考えられる．

この機能を使用することで，図2のように，選択中のまとまった記事間にmust-linkを付与し，別々にされた記事群の間にcannot-linkを付与することが可能である．

3.4 各クラスタの詳細情報の提示

分析を行う場合には，各クラスタはどのようなデータの集合となっているのか知る必要がある．そのため，各クラスタの詳細情報としてクラスタに属するデータ群の持つ単語の重み（*weight*）と，頻度（*frequency*）を表示する．クラスタ C_i の持つ記事群 E の中に含まれる単語 t の重み $weight(t)$ は，式（1）のように求める．$|C_i|$ はクラスタ C_i に属するデータ数，E_j はクラスタに属する一つのデータ，$w^t_{E_j}$ はデータ E_j の持つ単語 t の重みを指す．

$$weight(t) = \frac{1}{|C_i|} \sum_{E_j \in C_i} w^t_{E_j} \tag{1}$$

頻度はクラスタに属する各データ内で登場した回数の和とする．

また，クラスタ内で重要と思われる語句をクラスタを表す特徴語として抽出しユーザに提示する．本論文では，簡易的に式（2）のように単語 t の *feature* を求め，上位5つの単語をそのクラスタの特徴語とする．

$$feature(t) = \frac{weight\_rank(t) + local\_tf\_rank(t)}{2} \tag{2}$$

ただし，頻度がクラスタに属する文書の5分の1以下の単語は特徴語とはしない．これは，文書量の短い記事があった場合，その記事に属する単語の TF の値が大きくなり，登場回数が低いにも関わらず *weight_rank* で高い順位になることがあるためである．

weight_rank(t) と *local_tf_rank*(t) は，各単語の *weight* と *local_tf* を降順に並べたときの順位である．*local_tf*(t) は，単語 t のクラスタ内での TF を指し，式（3）のように求める．$n^t_{C_i}$ はクラスタ C_i 中での単語 t の登場回数，分母はクラスタ C_i の持つ総単語数を指す．

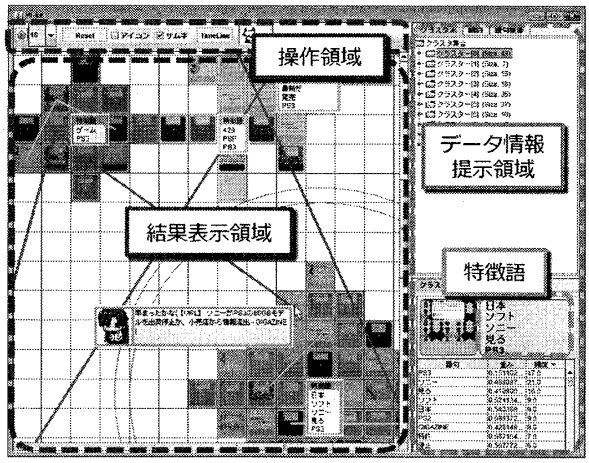

図3: 実装したシステム「Mi-ke」の実行例

$$local\_tf(t) = \frac{n^t_{C_i}}{\sum_{E_j \in C_i} \sum_{k \in E_j} n_k} \tag{3}$$

3.5 実行例

本論文で試作したシステム，Mi-keの実行例を図3に示す．本例は，一度クラスタリングを実行した後にいくつかのデータを確認し，制約を手動で付与した後に，再びクラスタリングを実行したときの結果である．左側にはクラスタリング結果を主に表示し，右側ではクラスタ情報や制約の情報などを主に表示する．

4 おわりに

制約付きクラスタリングシステム「Mi-ke」の実装を行った．制約付きクラスタリングを多様な記事群に対して適用し，対象記事群に応じて柔軟に分析可能な環境を整えることができたと考えられる．

今後の課題としては，具体的に大規模な記事群を分析する状況を想定した場合での，本システムの有効性の評価がある．

謝辞 本研究の一部は科学研究費補助金基盤研究（B）（課題番号 19300026）の助成による．

参考文献

[1] Sugato Basu, Ian Davidson, and K. Wagstaff, editors. *Constrained Clustering –Advances in Algorithms, Theory, and Applications–*. CRC Press, 2009.

[2] Hao Cheng, Kien A. Hua, and Khanh Vu. Constrained locally weighted clustering. *Proc. VLDB Endow.*, 1(1):90–101, 2008.

[3] 青島傳隼, 福田直樹, 横山昌平, 石川博. マイクロブログを対象とした制約付きクラスタリングの実現. 第1回データ工学と情報マネジメントに関するフォーラム DEIM2010, 2010.