時系列を考慮した階層的クラスタリングに基づくインタラクティブなニュース記事閲覧支援システム

An Interactive Browsing Support System for News Article Based on Hierarchical Clustering with Time-line

平田紀史[*1] 伊藤大樹[*1] 大囿忠親[*1] 新谷虎松[*1]
Norifumi Hirata Taiki Ito Tadachika Ozono Toramatsu Shintani

[*1]名古屋工業大学 大学院 工学研究科 情報工学専攻
Department of Computer Science and Engineering Graduate School of Engineering Nagoya Institute of Technology

It takes much computation time for analyzing topics in many news articles. In general, users focus on few topics for their interest. To speed up a topic analysis for news articles, the system needs to select articles that align with user's interest. In this paper, we propose a system to analyze topics based on hierarchical clustering. The system analyzes articles by keywords and interesting degrees. The system analyzes topics that are determinate by a few articles at a short time.

1. はじめに

Web 上のニュース記事は大量にかつ高頻度で配信されている．例えば，asahi.com[*1]の RSS を監視していると，1 日に約 150 の記事を配信していることが分かる．ニュース記事の分析のタイミングとして，記事の配信ごとに分析を行う，またはユーザの分析要求があるごとに分析を行うということが考えられる．前者の手法は前もって分析を行うため，ユーザへの分析結果の提示は高速になる．後者の手法は要求ごとに分析を行うため，結果の提示は遅くなる可能性がある．しかし，ユーザの要求に沿って結果を変化させることができるという利点がある．トピックの範囲の大きさはユーザによって異なるためである．例えば，“ 地震 ”全般に関して知りたいユーザもいれば，“ イタリア中部地震 ”に限定して知りたいユーザもいる．また，ユーザの要求が変化することも考えられる．不確かな要求からユーザの要求するトピックを確定させていくにはインタラクティブに分析を行うことが必要となる．以上のことから，本稿ではユーザの要求によって分析を行う形式をとり，分析時間の短縮と，要求を分析へ影響させることを目標とする．

具体的には，ユーザの入力するキーワードによって記事を限定し，分析する．そして，システムからのフィードバックをユーザが受け取り，分析を繰り返す．システムからのフィードバックはそのトピックに関連するキーワードである．これは，ユーザが本来求めるトピックを限定，拡張するための補助となる．

2. インタラクティブなニュース記事閲覧支援システム

2.1 トピックとサブトピックの関係

トピックの定義には様々なものがあり，TDT[Allan 98][Trieschinigg 04] での “互いに直接関連し合っているイベントおよび活動” の様に明示的に定義されている研究もある．この研究では，ニュース記事をコーパスとして整理し，記事とトピックとの対応関係が示しており，手法の比較や評価という面では非常に有用である．本稿では，

連絡先: 平田紀史，名古屋工業大学大学院 工学研究科 情報工学専攻，〒 466-8555 名古屋市昭和区御器所町，052-733-6550，nori@toralab.ics.nitech.ac.jp

∗1 http://www.asahi.com/

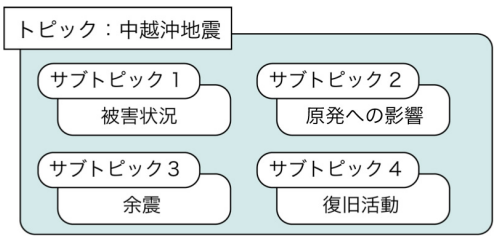

図 1: トピックとサブトピックの関係

ユーザが入力したキーワードに合致する記事の集合をトピックとする．これは，ユーザによって知りたいと考えるトピックの範囲が異なるという考えによる．

サブトピックは 1 トピックをさらに細かく内容ごとに分類したものである．したがって，トピックはサブトピックの集合からなる．例えば，中越沖地震に関するトピックは，その被害状況，原発への影響，余震，復旧活動などのサブトピックから構成させる．本システムでは，サブトピックを提示することで，内容の変化を把握できるようにする．

2.2 システム構成

図 2 にシステムの構成を示す．インタラクティブなトピック分析とは，提示された結果からキーワードや条件を修正し，システムに繰り返し問い合わせて，トピックを分析することである．一度のキーワード入力によりユーザが望むトピックの範囲を指定できるとは限らないからである．

インタラクティブなニュース記事閲覧支援システムでは，まず，ユーザには興味のあるキーワードを入力してもらう．すると，そのキーワードに関する記事の分析結果が提示される．分析対象とする記事はキーワードを含む記事であり，キーワードでフィルタリングを行っていることになる．分析結果が，ユーザが望むトピックでない場合は，繰り返しシステムに問い合わせることになる．分析結果にはトピックに含まれる評価値の高い単語が提示されるため，その単語をシステムからのフィードバックとして，次の分析の補助とする．

2.3 インターフェース

システムのインターフェースとして，始めに図 3 に示すような画面が表示される．ユーザは関心のある単語をキーワード

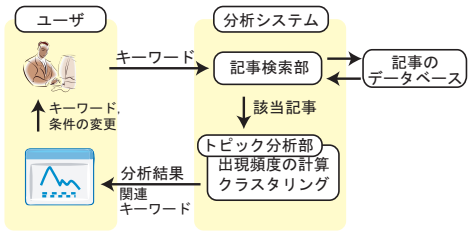

図 2: トピック分析システムの構成図

図 3: キーワードの入力ページ

として入力し，分析開始のボタンを押すことで，分析結果を得られる．複数の単語を入力キーワードとして与えたい場合は，スペースで区切りキーワードを入力する．キーワード間を明示的に “ or ” で区切らない場合は，AND 検索と同様の動作をする．また，プルダウンメニューで分析対象の期間を選択することができる．キーワードとして “北朝鮮” を入力した後，“核” を入力した場合の分析結果を図 4 に示す．

分析結果は図 4 の様に提示される．これは記事の出現頻度や単語の評価値の時間変化を提示することで，時系列的にどのような変化があったのかをユーザに理解させ，サブトピック間の関係を示すことで，トピック内の変遷を理解させることを目的とする．

図 4 中の (1) は横軸に時間，縦軸に配信された記事数をとったグラフであり，キーワードに対する記事の配信頻度を見ることができる．これにより，どの時期にキーワードに関するトピックが注目されていたかが分かる．

(2) は横軸に時間，縦軸に各単語の tf-idf の値の合計値をとったグラフである．ここで示される単語は期間内の tf-idf の合計値の多い上位 5 単語である．これにより，記事数より詳細なトピックの変化，各単語における注目度を知ることができる．

(3) はサブトピックを時系列に表示し，サブトピック間で類似度が高いものをエッジで関連づけた無向グラフであり，ノードがサブトピックに当たり，ノードの中の数字はそのサブトピックに属する記事数を表している．サブトピックに含まれる記事数が多いほど，大きなノードを示すようにしてある．また，図右の時間と文字列は，同じ高さにあるサブトピックを代表する記事のタイトルと配信時間である．サブトピックの代表はクラスタの重心に最も近い記事のタイトルとした．また，サブトピック数が多くなるとエッジが多くなるため，サブトピックに属する記事が 1 記事だけの場合は，表示しないようにしてある．

(4) は再分析する場合の補助のためのリストである．左から，現在の入力キーワード，分析結果より得られた関連のある単語，過去のキーワード入力の履歴である．各キーワードの右側

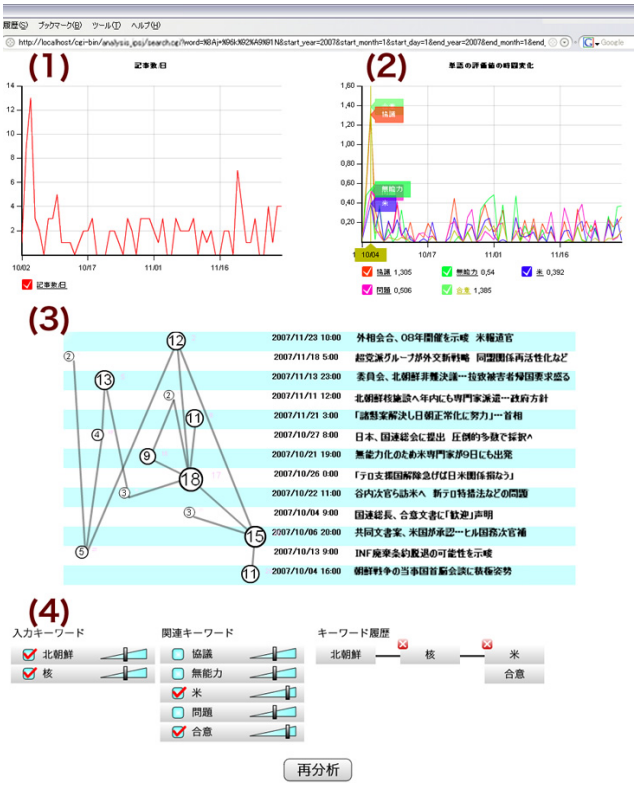

図 4: 分析結果の提示

のチェックボックスにチェックを入れて，再分析ボタンを押すと，そのキーワードにより分析が行われる．その際に，各キーワードの右側のスライドバーによって，キーワードに対する関心の度合いをシステムに与えることが可能である．より関心が高いキーワードはスライドバーを右に，関心の低いキーワードはスライドバーを左に移動することで，各キーワードに対する関心度を表すことができる．与えられた関心度は 3. 節で説明するクラスタリングの結果に影響する．

図 4(4) の例では，入力キーワードとして，“北朝鮮 核” を入力した結果が (1),(2),(3) の結果であるので，入力キーワードとして北朝鮮と核が表示されている．トピックに属する記事の中で，このトピックに関連キーワードには，(2) で選択された単語である “協議”,“無能力”,“米”,“問題”,“合意” が表示されている．“北朝鮮”,“核”,“米”,“合意” のキーワードの左にあるチェックボックスにチェックがしているため，再分析のボタンを押すと，これらのキーワードを含む記事をトピックとして分析を行う．分析結果については，“米”,“合意” に関するスライドバーは右に移動してあるため，これらの単語についてより詳細なサブトピックを示すことができる．これらによって，不確かなユーザの要求を確定させていく場合や，ユーザの要求の変化する場合に対応できるような，インタラクティブなトピック分析を支援することができる．

3. サブトピックのためのクラスタリング手法

3.1 分析時間減少のための階層的クラスタリング手法

サブトピックを得るために階層的クラスタリングを用いる．形態素解析には MeCab[*2] を用い，名詞のみを評価対象とした．しかし，凝縮型の階層的クラスタリングは記事数を n とする

*2 http://mecab.sourceforge.net/

と $O(n^2)$ となり，インタラクティブなトピック分析システムには適さない．しかし，クラスタ数の変更が容易であることや，k-means 法のように初期値に依存しないという利点がある．階層的クラスタリングの実行時間を減少させる手法として，k-means 法によりいくつかのクラスタを生成し，そのクラスタに対してクラスタリングする手法 [Zhao 02] がある．しかし，k-means 法は始めにクラスタ数を設定しておく必要があり，本稿で取り扱うトピックのように，記事数が一定でない場合や，適切なクラスタ数が曖昧な場合での設定は難しい．本稿では，この手法での k-means 法でのクラスタリングにおいて，記事数で区切り，k-menas 法を用いた分割型の階層的クラスタリングを行う手法 [平田 09] を用いる．以下にその手法を示す．

step1 時系列にソートした記事を記事数 L で区切り，初期のクラスタとする．

step2 クラスタを k-means 法で 2 つに分割を行い，自己類似度が閾値 S 以上となるまで繰り返し分割する．

step3 得られたクラスタに対して階層的クラスタリングを適用し，求まったクラスタをサブトピックとする．

step1 では，記事数で区切って初期クラスタとしている．これは，step2 で行う処理時間を減少させるためである．step2 では記事数によって区切られたクラスタを対象にそれぞれに対し分割を行う．分割に際して，終了条件をクラスタ数にするより，類似度などを終了条件とした方が，内容の類似したクラスタが得られると判断した．そこで，終了条件は分散の距離関数をコサイン類似度にした以下の式を用いた．

$$s_i = \frac{1}{N}\sum_{j=1}^{N} cos(i,j) \quad (1)$$

式 (1) においてクラスタ i に含まれる記事の平均ベクトルと，クラスタ i 内の記事 j のベクトルとのコサイン類似度を $cos(i,j)$ と表す．N はクラスタ内の記事数である．

クラスタとクラスタの距離を測るときに，クラスタに属する記事の規模によって，計算時間が増減する．そこで，記事数が増加した場合の計算時間を抑制するために，評価値が閾値 φ以下の単語は計算の対象外とした．また，各単語の評価値は tf-idf の値を取り，距離関数は Ward 法によるものとした．

3.2 関心度のクラスタリングへの影響

2 回目以降の分析ではユーザにキーワードごとの関心度を入力してもらい，関心度を考慮したクラスタリングを行う．関心度の低いキーワードを多く含むクラスタの場合は，クラスタの規模が大きくなるように，階層的クラスタリングの終了条件を変更する．関心度の低いキーワードを多く含むクラスタの規模が大きくなると，相対的に関心度の高いキーワードを多く含むクラスタの数が増加する．したがって，関心度の高いキーワードを含む記事ほど，詳細に分析を行うことができるようになる．

階層的クラスタリングを行う場合の閾値は，特に関心度の指定がない場合，ϕ とする．キーワード i に対する関心度 α_i を考慮する場合は，式 (2) で表す ϕ' を終了条件とする．また，この関心度は $0 \leq \alpha_i \leq 1$ の範囲で与えられる．しかし，すべての α_i が 0 となると，閾値 ϕ' が 0 となってしまうため，少なくとも 1 つの α_i は 0 より大きい必要がある．

$$\phi' = \phi \times \frac{\sum_{i=1}^{M}(\alpha_i \times t\_cnt_i)}{\sqrt{\sum_{i=1}^{M} t\_cnt_i^2}\sqrt{\sum_{i=1}^{M}\alpha_i^2}} \quad (2)$$

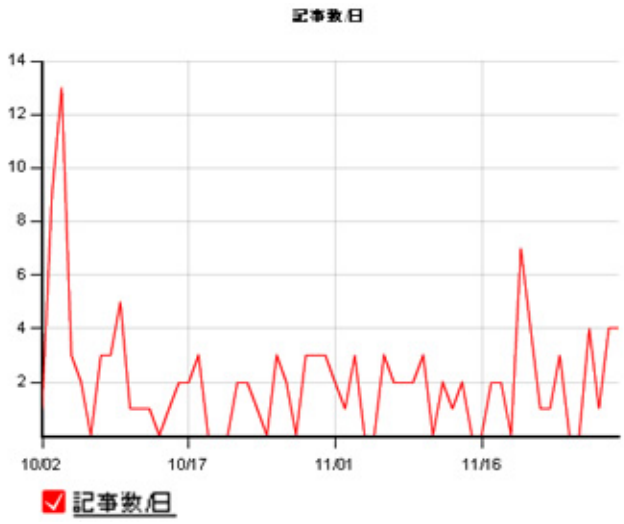

図 5: “北朝鮮 核” に関する一日当たりの記事数の時間変化

M はキーワードの数を表し，t_cnt_i はクラスタに属する記事でキーワード i が出現する回数の合計である．$\alpha_i \times df_i$ によって，キーワードに対する関心度が高く，かつクラスタ内に多くの記事が含まれる場合に，大きな値となる．

実際の処理では，まず，クラスタ c1 とクラスタ c2 の距離関数による類似度を得る．そして，c1 と c2 を結合した場合の ϕ' を求め，類似度が ϕ' 以下であれば，c1 と c2 の結合は行わないという処理になる．

4. 閲覧支援例

毎日 jp が 2007 年 10 月 1 日から 2007 年 11 月 30 日まで配信した 10,692 記事を対象にキーワードによりフィルタリングを行い，分析を行った．入力したキーワードは “北朝鮮 and 核” である．このキーワードに合致する記事は 125 記事あった．

4.1 記事数と単語の評価値の時間変化

図 5 に一日当たりに配信された記事数を示す．これにより，10 月 4 日，11 月 20 日に記事が多く配信されており，世間的な関心が高かったことと予測できる．

図 6 に一日ごとの各単語の評価値を示す．図 5 で示す記事数と同様に 10 月 4 日に評価値が高くなる単語が多い．しかし，11 月 20 日の各単語の評価値は低いままで，記事数の関連性がない．実際に 11 月 20 日には日中韓首脳会談があり，これに関する記事が配信されていた．北朝鮮の核に関する問題だけではなく，経済分野などで協力することなども記事となっていたため，“協議” や “合意” などの単語の評価値が記事数に比べ小さくなったと考えられる．

4.2 関心度による提示結果の違い

図 7 に関心度による影響のない状態でのサブトピックを，図 8 に影響のある状態でのサブトピックを示す．関心度による影響のない状態で分析を行った結果，表示されるサブトピック数は 13 となった．この関心度による影響のない状態とは，式 (2) で表す閾値の ϕ' を ϕ とした状態のことを指す．そして，関心度を北朝鮮が 0.2，核が 1.0 として分析を行った結果，表示されるサブトピック数は 11 となった．これにより，関心度を設定することで，サブトピック数が減少したことが確認できた．

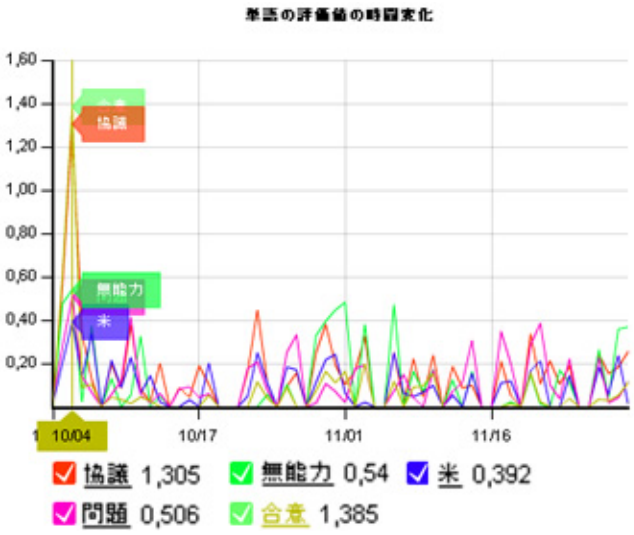

図 6: “北朝鮮 核” に関する評価値の時間変化

図 7: “北朝鮮 核” に関するサブトピック

4.3 課題と考察

図 6 の様に評価値の時間変化を提示することで，図 5 の様な配信頻度だけでは分からない，どのような単語が評価されていたのかを見ることができる．グラフに現れる単語は評価値の合計が高いものであり，システムが自動的に決めている．そのため，ユーザの興味のある単語についてもグラフ化できるように改良を行う必要がある．

また，ユーザに各キーワードへの関心度を設定してもらうことで，ユーザの意向をクラスタリングに影響させることが確認できた．しかし，どの程度の関心度を示せば，どの程度の影響があるのかという感覚は経験によるという課題がある．システムの方針としては，何度か問い合わせてユーザの望む結果を得ていくというものである．したがって，何度か試してもらうという課程で，関心度の影響を把握できるものと考える．

また，サブトピックの提示方法に課題がある．本システムでは，サブトピックを表す文字列は各クラスタの重心に最も近いタイトルとしている．クラスタに含まれる記事で評価値の高い単語を羅列するという手法も考えられたが，単語の羅列よりタイトルの方が内容を端的に表せられる考えたためである．しかし，実際にシステムとして作成してみると，記事数の少ないクラスタではタイトルがサブトピックの内容に対応するが，記事数が多い場合，サブトピックの内容が多岐に渡ることが多く，適切とは判断できなかった．したがって，記事数が少ない場合はタイトルをラベルとすることは良いが，その逆で，記事数が多い場合は単語の方が良いということが考えられる．

図 8: 関心度を変化させた場合の “北朝鮮 核” に関するサブトピック

5. おわりに

本稿では，ユーザがキーワードに関心度として重みを付けてトピックを分析するシステムを提案した．キーワードに重みを付けることで，ユーザは自分の知りたい事柄について，より詳しい結果を得ることができた．また，配信頻度や単語の評価値の時間変化を提示することで，トピックの理解の補助することができた．

しかし，サブトピックが多くなった場合など，サブトピックの表現方法について課題がある．ネットワークの可視化技術 [三末 09] や Web 検索におけるキーワードの可視化手法 [吉田 07] などの研究があるため，これらの分野の手法を参考に改良を行おうと考えている．サブトピックをより理解しやすい形式で表現できれば，トピックの分析によるニュース閲覧支援の効果がより大きくなる．

参考文献

[Allan 98] James Allan, Jaime. Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, “Topic Detection and Tracking Pilot Study Final Report”, Proc. of the DARPA broadcast news transcription and understanding workshop, 1998, pp.194-218.

[Trieschinigg 04] Dolf Trieschnigg, Wessel Kraaij, “TNO Hierarchical topic detection report at TDT 2004”, Topic Detection and Tracking 2004 Workshop, 2004.

[Zhao 02] Zhao Ying, George Karypis, “Evaluation of Hierarchical Clustering Algorithms for Document Databases”, Proc. of the 2002 ACM CIKM. 2002, pp.515-524.

[平田 09] 平田紀史, 浅見昌平, 大囿忠親, 新谷虎松, “ニュース記事のための対話的トピック分析システムとその高速化手法について”, 情報処理学会第 71 回全国大会, 2009.

[三末 09] 三末和男, “ネットワークの可視化技術 -大規模ネットワークと動的ネットワークへの挑戦-”, 電子情報通信学会誌, Vol.92, No. 2, 2009.

[吉田 07] 吉田大我, 小山聡, 中村聡史, 田中克己, “Web 検索結果におけるキーワード出現相関の可視化と対話的な質問変換”, 第 18 回データ工学ワークショップ, 2007, C7-2.