3G1-1

ユーザの選好に基づくトピック分析システムの試作

An Implementation of a Topic Analyzing System Based on Users' Preferences

平田紀史*1　大囿忠親*1　新谷虎松*1
Norifumi Hirata　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 topic analysis of many articles. But actually the topics which users need is few. To improve the computation time, we need to focus on users' preferences. In this paper, we propose a system to analyze topics based on users' preference. The system select articles by users' preferences. And the system analyzes topics which are consisted by a few articles at a short time.

1. はじめに

ニュース記事を読んでいるとき，記事の内容が理解できないことがある．これには，背景知識の不足や各単語の意味の理解不足など，様々な原因が考えられる．本論文では特に，記事の背景となるトピックの変遷の理解不足が原因の場合を考える．そして，トピックの変遷の把握を支援するシステムについて提案する．

毎日 jp*1から配信される記事のタイトルには，その記事のトピック，話題を表す情報がある．これを仮にトピックと見なして，1ヶ月間のトピックの数とトピックに属する記事数の関係を調べると図 1 のようになる．これは，トピックを記事数によってソートした結果である．全体のトピック数は 954 記事であるのに対し，トピック分析が可能となる数の記事を含むトピックは，全体に対して少ない．例えば，5 記事以上を含むトピックは 54 個である．したがって，すべての記事に対して分析を行う必要はないと考える．

ニュース記事を対象として自動的にトピックを分析する研究は，TDT(Topic Detection and Tracking)[Allan 98] をはじめ，多く存在する．このような研究ではトピックを定義し，各記事とトピックの対応関係を持つデータを用いて，分析を行っている．文献 [Allan 98] によれば，トピックは “互いに直接関連し合っているイベントおよび活動”と定義される．イベントは “特定の時間および場所で起こった出来事”を表し，活動は “共通の関心や目的を持った行動の連鎖”を表す．しかし，各ユーザが考える記事のまとまりと，上記のように定義されるトピックとは必ずしも一致しない．例えば，大リーグのイチローの記事を読んでいたユーザを想定する．ユーザが知りたいことは，大リーグで活躍する日本人選手全般に関してなのか，イチロー個人に関してなのかは，ユーザにより異なると考えられる．そこで，本論文ではユーザの選好により選択された記事集合をトピックとする．

提案するシステムは，ユーザが入力したキーワードから，記事を選択し，結果をトピックと見なす．すなわち，トピックはユーザによって定義される．そして，システムはトピックを分析し，ユーザに提示することでトピックの理解を支援する．

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 愛知県名古屋市御器所町，nori@toralab.ics.nitech.ac.jp

*1 http://mainichi.jp/

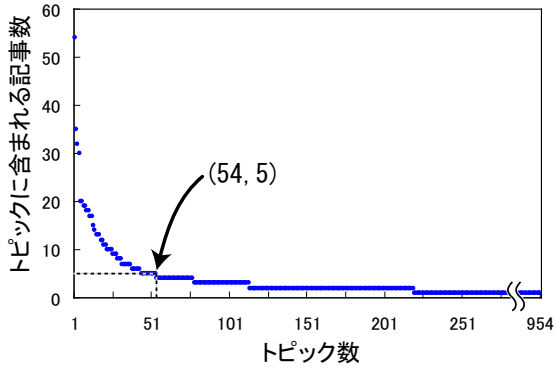

図 1: トピックに含まれる記事数の分布

2. トピック分析手法

トピック分析は記事の配信頻度の提示，単語の評価値の変化の提示，およびトピック追跡を行う．配信頻度を提示することで，時間経過によるトピック全体の注目度の起伏が理解できる．また，評価値の変化を提示する

ことで，トピックを特徴付ける単語の変化が理解できる．そして，トピック追跡を行うことで，トピックの変遷を把握することができる．

2.1 単語の評価値の時間変化

トピック分析の結果の一つとして単語の評価値の時間変化を提示する．このためには，記事の配信された時点での単語の評価値を計算する必要がある．

文書中の単語の特徴は $tf{\cdot}idf$ によって表すことができる．

$tf{\cdot}idf$ は文書中の単語の出現頻度である tf と，全文書中に単語が出現しない文書数の割合である idf の乗算によって計算できる．特に，idf は式 (1) によって示される．

$$idf(w) = log\left(\frac{N}{n(w)}\right) + 1 \quad (1)$$

w は対象とする単語，N は全文書数，$n(w)$ は単語 w の出現する文書数である．idf の $n(w)$ は，トピック内で w が出現する文書数に置き換えた方が，トピック内での特徴を評価するには良い．しかし，本来のトピックに関係のない文書が含まれて

いた場合，無関係な単語が高評価になる場合がある．したがって，idf は全文書からの文書数の割合とする．単語の評価値の大きさは tf によって決定され，その変化は idf によって決定されることになる．

また，$tf{\cdot}idf$ を計算するためには，文書を単語に分解する必要がある．これには，形態素解析器である MeCab[*2]を用い，計算対象とする品詞は名詞のみとする．

2.2 サブトピックの抽出手法

本論文では，トピックは内容毎に分割されたサブトピックから構成されているとする．そして，トピック追跡とは，サブトピックを時系列に並べることで，トピックの変遷を把握することである．

トピックの変遷を把握するという意味ではトピックのスレッド構造を検出する研究 [井手 03] もある．この研究では一つ一つの記事を見ているが，本論文では，記事のまとまりごとにトピックを追跡する．これには，ユーザへの提示する情報の量を減らせられる効果がある．

トピック追跡を行うためには，トピックをサブトピックに分割する必要がある．そのために，まず，文書をベクトル化する．具体的にはベクトルの各次元に単語を割り当て，各成分には単語の評価値を割り当てる．各単語に対する評価値は $tf \cdot idf$ の値を用いる．そして，階層的クラスタリングの Ward 法を用いてサブトピックを得る．Ward 法の距離関数は式 (2) で表される．

$$Ward(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \qquad (2)$$

ここで，C_1，C_2 はクラスタを，$E(C)$ はクラスタ C の平方和を表す．

トピック内で，時間的に距離のある記事は同じサブトピックを表している可能性は低い．そこで，記事間の配信時間が閾値 L 以上の場合は距離関数の結果を無限大として，計算の対象から外す．距離関数は式 (3) で表す値を用いる．

$$\begin{aligned} D(C_1, C_2) &= \begin{cases} if\ \ t_C(C_1, C_2) > L,\ \ \infty \\ else,\ \ Ward(C_1, C_2) \end{cases} \qquad (3) \\ t_C(C_1, C_2) &= \max_{i,j} \left(t_d(d_{1i}),\ t_d(d_{2j}) \right) \end{aligned}$$

$t_C(C_1, C_2)$ はクラスタ C_1，C_2 の時間的距離を，$t_d(d)$ は文書 d の配信時間を示す．また，クラスタ C_1 は文書 d_{1i} から，クラスタ C_2 は文書 d_{2i} から構成されるとする．以後，L を制約時間と呼ぶ．

時間情報を用いるため，サブトピックを得るためのクラスタリングの精度向上が期待できる．また，階層的クラスタリングはユークリッド距離を用いる場合，計算量は $O(N^2)$ である．しかし，対象を制限することため，クラスタリングの高速化も期待できる．

文書間の時間的な距離を類似度に反映させる手法 [Zhang 06] も存在する．それに対し，本手法は，閾値を越えた場合は距離計算を行わないことが特徴である．

また，クラスタリングする範囲を制限する手法としてスライディングウィンドウ方式 [Brants 03][平田 07] がある．スライディングウィンドウ方式では，ウィンドウと呼ばれる範囲の記事のみを対象に処理を行う．そして，そのウィンドウの範囲 L を step S ごとに移動し，それぞれに処理を行う手法である．しかし，その結果は L と S に依存する．本論文で用いる手法も対象を制限するという考え方は同じであるが，S を設定しなくて良いという利点がある．

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

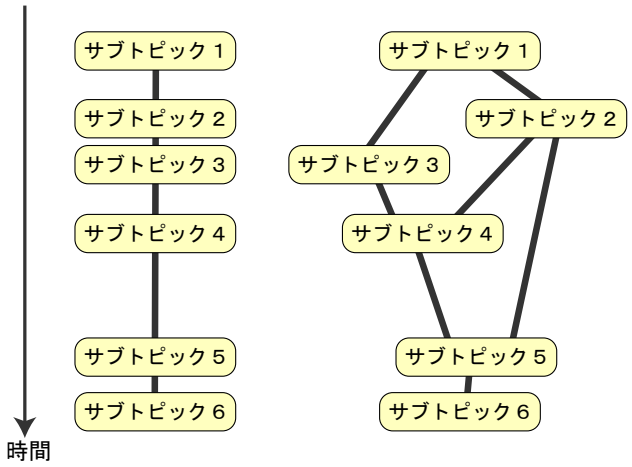

図 2: サブトピック間の関係の表現

2.3 トピックとサブトピックの表現

トピックとサブトピックを表現するものとしてラベルを定義する必要がある．トピック追跡を行う際には，クラスタのラベルによってトピックの変化を把握するため，ラベル付けは重要である．ラベルの付け方は，クラスタ内で評価値の高い単語を複数提示するという手法があるが，単語の羅列だけでは内容の把握が困難である場合がある．

そこで，クラスタの平均ベクトルと式 (2) で表す距離が最も近い文書のタイトルもラベルとする．タイトルは文書本文を短く表す文字列であることが多い．クラスタの平均との距離の近い文書のタイトルは，クラスタの特徴を表すと考えられる．したがって，ラベルとして文書のタイトルも用いる．

2.4 サブトピック間の関係

トピック追跡はサブトピックを時系列に整列することで行われる．図 2 の右と左のグラフは同一トピックを追跡した例である．各サブトピックをエッジで関連付けて，時系列に並べている．サブトピック間の関係は右図の方が詳細に提示することができる．サブトピックを，図 2 の左のように，一次元的に提示すると，本来のトピックの変遷の把握が困難になる場合がある．したがって，図 2 の右に示すように，サブトピックをノードとしたグラフ構造を持つことを考える．

具体的には，式 (2) で表されるサブトピック間の距離が閾値以上なら，エッジを結ぶ．また，エッジが一つもないノードは，閾値以下であっても最も距離の近いノードへとエッジを結ぶ．

3. システム構成

本論文で提案するシステムの構成図を図 3 に示す．図 3①で，ユーザは興味のあるトピックに関するキーワードを入力し，システムに与える．図 3②で，キーワードに合致する記事集合をトピックと見なし，トピック分析部に与える．そして，図 3③で，記事集合を分析した結果をユーザに提示する．ユーザはキーワードを変化させることで，トピックの規模や条件を変化させることができる．

3.1 記事収集

トピックの検索や分析をする前に，記事を収集する必要がある．新聞社等のサイトでは過去の記事は削除される傾向があり，参照できない記事が存在する．そのため，記事を収集して保存しておく必要がある．

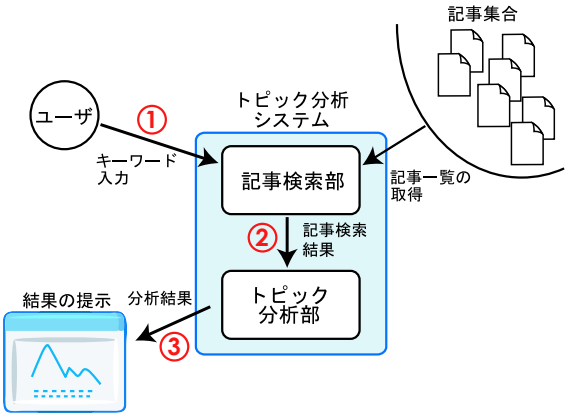

図 3: システムの構成図

図 4: 記事検索のためのインターフェース

記事は RSS を用いて新聞社のサイトから各記事を収集する．また，RSS に記述されている <dc:date> の時間を記事の配信時間とする．

3.2 記事検索

キーワードから記事を選択する過程は検索に置き換えられる．図 4 にキーワード入力に関するインターフェースを示す．システムは，入力されたキーワードと収集しておいた記事集合とを比較し，条件に合致する記事だけを選択する．キーワードが複数の場合は and 検索と or 検索を，それぞれ “and” と “or” で分けることが可能である．また，図 4 ① において，記事の配信時間の範囲も設定可能である．ユーザはこれらの機能を組み合わせることで，興味のあるトピックに関する記事を選択する．

4. トピック分析実験

4.1 実験環境

毎日 jp が 2007 年 10 月 1 日から 2007 年 11 月 30 日まで配信した 10692 記事を収集し，分析を行った．入力したキーワードは “ＮＯＶＡ or 英会話” である．実験環境は OS が Windows XP Home Edition，CPU が Pentium M 1.20GHz，メモリが 1GB DDR2 SDRAM，Java の実行環境が JRE1.5.0 である．

トピック分析の結果は図 5 に示すようにユーザに提示される．結果としては，70 の記事が一つのトピックとして得られた．図 5 ① にトピックに属する記事の配信頻度，図 5 ② にトピック内での単語の評価値の変化，図 5 ③ にトピック追跡の結果が提示される．

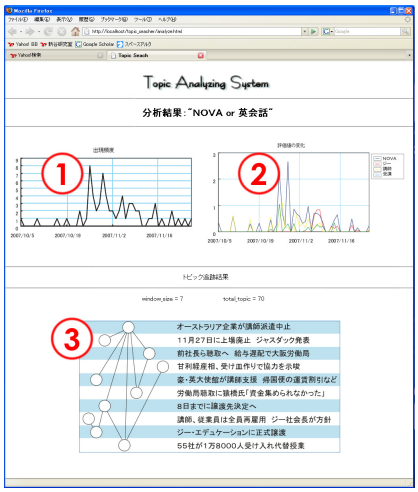

図 5: 分析結果一覧

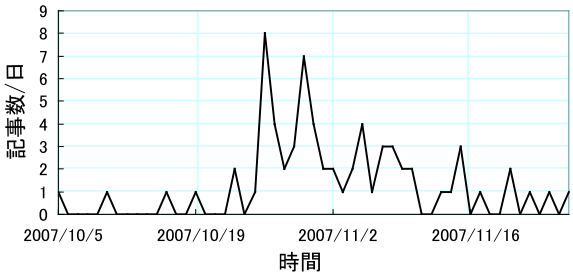

図 6: 記事の配信頻度の時間変化

4.2 記事の配信頻度

図 6 に配信頻度の時間変化の図を示す．横軸は時間を，縦軸は一日ごとの記事数を示す．この図を見ると，10 月下旬頃に記事数が増加しているため，トピックに何らかの変化があったことが予想できる．本例では，10 月 26 日に経営破綻となり，記事数と注目度が連動していることが分かる．

4.3 単語の評価値の変化

図 7 に単語の評価値の時間変化を示す．表示してある単語は出現数が単語の評価値の高かった上位の 4 単語である．期間の前半では，“受講” や “講師” といった単語の評価が高かったが，後半になると “ジー” という単語の評価が高くなっている．本例では，ジー・コミュニケーションという会社がＮＯＶＡを買収するといった変化があった．前半には，講師や受講といった部分が注目されていたのに対し，後半になると具体的な買収先に注意が移っていることが分かる．また，“ＮＯＶＡ” という単語は図 6 に示した記事の配信頻度と変化の仕方が類似している．したがって，トピック内の変化を把握する目的には適切な単語でないと考えられる．

4.4 トピック追跡

図 8 にトピック追跡を行った結果を示す．本例では式 (3) に示す時間的な制約時間は 7 日間とした．図 8 の左にはサブトピックを現すグラフを，右にはグラフのノードを表すサブトピックのタイトルを示す．左にある丸がサブトピックを表し，距離の近いサブトピック同士をエッジで関連付けてある．そして，これによって，サブトピック間の関係を提示する．単純に時系列に並べた場合は，図の右の文字列だけとなり，トピック

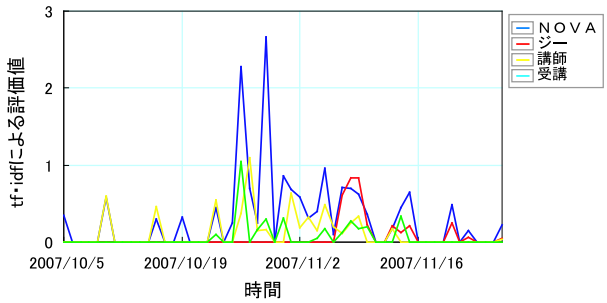

図 7: 単語の評価値の時間変化

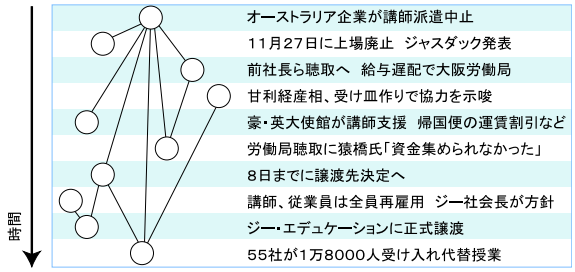

図 8: 制約時間 L が 7 日でのトピック追跡の結果

の変化の把握には向かない．

サブトピックに属する記事数が一つだけの場合は，表示していない．一つの記事で構成されるサブトピックは，他のサブトピックと関連性の低い記事である可能性が高いためである．また，サブトピックを多く表示すると出力されるグラフ構造が複雑になる．

制約時間を 3 日に変化させてトピック追跡を行った結果を図 9 に示す．制約時間が 7 日である図 8 に比べるとサブトピックの数が 10 から 12 に増加し，詳細にトピックを追跡できる．しかし，グラフの構造が複雑になっている．

4.5 実行時間

記事の検索と記事の読み込みを除く，トピック追跡のためのクラスタリングに費やした時間を計測した．その結果，制約時間が 7 日の場合は約 0.92 秒，制約時間が 3 日の場合は約 0.53 秒であった．制約時間が短い方が，トピック追跡を行うための時間は短くなることが確認された．

また，記事の検索からトピック追跡までの実行時間は，7 日の場合，約 25.03 秒であり，3 日の場合は 23.89 秒であった．クラスタリングには 1 秒程度しか時間を必要としないため，記事の検索に時間を費やしていることが考えられる．これは，転置ファイルなどを用いずに全文探索を行っているためである．

事前にトピックの抽出と分析を行い，分析結果を読み込むシステムも考えられる．この場合，ユーザがトピックを検索し，選択したトピックの分析結果を提示することになる．結果の提示は本システムより高速になる．しかし，ユーザはトピックを検索する必要があるため，ユーザの作業量としては変わらない．そして，ユーザによるトピックの決定が不可能となる．本実験と同じ条件で，すべての記事を対象にクラスタリングを行い，トピック抽出をした処理時間は約 4590 秒であった．対象とする記事が増加した場合に，計算時間が $O(N^2)$ に従って増加することが考えられる．処理時間が増加すると，最新の記事に対して処理を行うことが難しくなる．

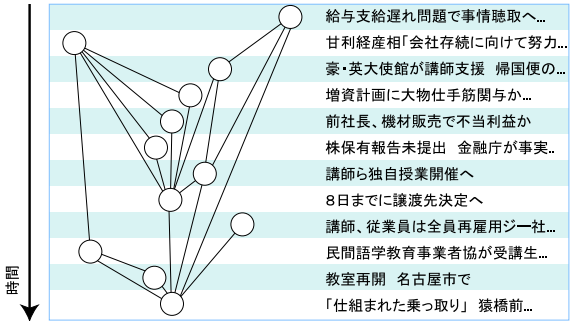

図 9: 制約時間 L が 3 日でのトピック追跡の結果

5. おわりに

本論文では，トピックはユーザによって決められる文書集合と定義した．トピック分析を行うために，ユーザにキーワードを入力させ，対象を限定することで，分析時間を短縮する手法を用いた．実際に，記事数を制限することにより，制限しなかった場合と比べ短時間での処理が可能となった．

トピック追跡を行うためのクラスタリングでは，制約時間を用いる手法を提案した．この手法により，時間の短縮と，時間情報の結果への反映が可能となった．また，実験において，トピック分析での主な処理であるクラスタリングの時間が 1 秒未満となり，短時間で処理が可能であることが確認できた．ユーザは制約時間を変更することで，トピック追跡の粒度の変更を短時間で行うことができる．

課題として，評価値の時間変化に関して，実験結果の “NOVA” という単語のような，トピック全体を表す単語を選択しない手法が必要である．トピック全体に均一に出現する単語を提示しても，トピック内の変化の把握は困難である．また，記事検索の高速化も課題の一つであり，検索技術を導入する必要がある．

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, pp.194-218, 1998.

[Zhang 06] 張 一萌，何 書勉，小山 聡，田島敬史，田中克己，“時系列データに意味的に関連するニュースト ピックの発見”，日本データベース学会 Letters Vol.5，No.1，pp.133-136，2006.

[Brants 03] Thorsten Brants and Francine Chen, “A System for New Event Detection”, Proc. ACM SIGIR, pp.330-337, 2003.

[平田 07] 平田紀史，大囿忠親，新谷虎松，“エージェントによる滑走窓方式を用いた複数情報源からのトピック追跡”，合同エージェントワークショップ＆シンポジウム 2007，CD-ROM，2007.

[井手 03] 井手一郎，孟洋，片山紀生，佐藤真一，“大規模ニュース映像コーパスの意味構造解析”，電子情報通信学会技術研究報告，pp.13-18，2003.