イベントの属性抽出と系列化に基づくニュース記事閲覧支援システム

平田 紀史† 伊藤 太樹† 柿元 宏晃† 佐野 博之†
白松 俊† 大囿 忠親† 新谷 虎松†
名古屋工業大学大学院工学研究科情報工学専攻†

1 はじめに

本稿で提案するシステムは，閲覧中の記事に関係するイベントの系列を提示することで，閲覧中の記事の理解の支援を行う．複数のニュース記事として扱われるような大きな事件では，複数のイベントが関連している．ここでは，ニュース記事の閲覧支援として事件の全貌を表すためにイベントの関係を示すシステムを実現することを目指している．このとき，ニュース記事には様々な観点があるので，ユーザが望んでいる観点を考慮する必要がある．また，ユーザが望む観点はその状況によって変化することも考えられる．

この解決のために，システムとユーザのインタラクションによってイベント系列を取得していくシステムを提案する．品詞によって重み付けを変えることで，観点を変更する研究[1]もあるが，本稿ではより直接的にユーザに選択をする手法を用いる．これは，ユーザの興味の変化により素早く対応するためである．また，システムは過去の記事からイベントを抽出するが，本稿で述べるシステムはイベントを特徴付ける重要語を抽出する．これは，イベント抽出の対象とする記事を重要語でフィルタリングするためである．重要語の抽出において，単純にtf-idfなどを用いるだけでは，不十分な場合が確認できた[2]．この解決のために，記事からイベントの属性情報を抽出し，重要語の判定に利用する．これらにより，適切な重要語を抽出し，ユーザの観点の違いを考慮したイベント系列を求める手法を提案する．

イベントの定義には様々なものがあり，TDT(Topic Detection and Tracking)[3]においては“イベントとは特定の場所，時間で起こった出来事”と定義されている．本稿でのイベントもこの定義に従い，イベントは複数の記事で表現する．また，本稿ではイベント系列化をイベント同士を関連づけることとし，イベント系列をイベントをノードとした木構造として表す．

2 属性情報の抽出と系列化の手法

2.1 イベント系列化の概要

提案するイベント系列化手法は，1) 重要語抽出，2) 記事検索，3) イベント抽出，4) ユーザによるイベント選択という4ステップから構成される．4つ目のステップによって選択されたイベントから再度重要語を抽出し，処理を繰り返していく．提案手法は，この繰り返しによって，ユーザの観点を考慮したイベント系列を得る手法である．

図1にイベントを選択してイベント系列を生成していく様子を示す．横軸に時間軸を取るが，この例は過去にさかのぼってイベント系列を得ていく場合を考える．e_{i-j}と書かれてあるノードがイベントに当たり，イベント間はエッジで関連づけられている．始めに，閲覧中の記事を入力として与えると，それをイベントe_0と判断し，関連するイベント，e_{1-1}からe_{1-3}を提示する．提示されたイベントで興味のあるもの，例えばe_{1-1}を1回目の選択として選択すると，そのイベントe_{1-1}に関する別のイベントe_{2-1}からe_{2-4}が提示される．同様の動作を繰り返して，イベントをe_{2-3}，e_{3-2}と選択していった結果が図1である．この場合，イベント系列はe_0，e_{1-1}，e_{2-3}，e_{3-2}となる．

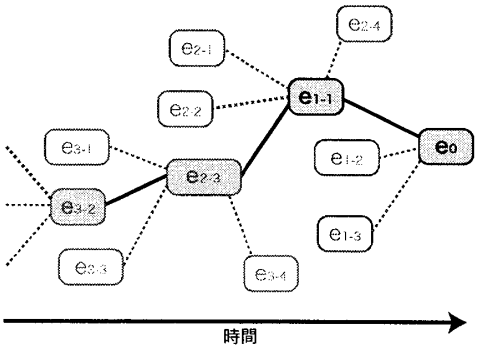

図1: イベントの提示とユーザの選択によるイベント系列の取得

2.2 属性情報による重要語の抽出

イベントを構成する記事から属性情報となる単語を重要語として抽出する．対象とする単語は名詞と動詞とする．各イベントの単語に対して評価値を計算し，高い値を示す単語を重要語とする．イベントeにおける単語wの評価値$eval(w,e)$を(1)式に示す．

$$eval(w,e) = weight(w,e)\sum_{i=0}^{N-1} tf\cdot idf(w,i) \quad (1)$$

評価値は，イベントeがN個の記事で構成される場合，それぞれの記事iにおける単語wのtf-idfの値$tf\cdot idf(w,i)$を足し合わせた値を基本とする．また，記事iに単語wが存在しない場合は，$tf\cdot idf(w,i)$を0として計算し，idfにおいて各単語wを含む文書数の範囲は，システムが扱う記事全てとする．そして，各イベントごとに$eval(w,e)$を求めて，評価値の高いwを重要語と判断する．

ここで，イベントの特徴を表す度合いが高いほど，高い評価値となるようにする．そこで，tf-idfだけでなくの値だけでなく，名詞に付属する助詞に着目する．その重みが$weight(w,e)$である．対象とする助詞は体言に付属する係助詞，格助詞である．強調される単語は文書中に多く出現するが，それに加え，強調される単語に高頻度で付属する助詞ほど，その単語を強調する役割があるという考えに基づいて計算する．

始めに，助詞cが付属する単語をw_cと表現する．そして，文書dにおいて，それぞれのw_cの出現割合を計算し，平均

A News Browsing Support System based on Event Attribute Detection and Event Arrangement
Norifumi HIRATA, Taiki ITO, Hiroaki KAKIMOTO, Hiroyuki SANO, Shun SHIRAMATSU, Tadachika OZONO and Toramatsu SHINTANI
† Dept. of Computer Science and Engineering, Graduate School of Engineering Nagoya Institute of Technology, 466-8555, Nagoya, Japan

を計算する．これを (2) 式に示すように $avg\_tf_c(d)$ とする．

$$avg\_tf_c(d) = \frac{1}{\#\{w_c \in d\}} \sum_{w_c} tf(w_c, d) \quad (2)$$

次に，これらを文書 d ごとに計算し，システムの扱う全ての文書 $d \in S$ に関して平均値を計算する．これを (3) 式に示すように $wc(c)$ とする．

$$wc(c) = \frac{1}{\#\{d \in S\}} \sum_{d} avg\_tf_c(d) \quad (3)$$

$wc(c)$ が助詞 c による単語の強調の程度を数値化したものである．

イベント e で出現する単語 w に付属する助詞の割合によって $wc(c)$ を加重平均した値を基に，重み $weight(w,e)$ を計算する．これを (4) 式に表す．

$$weight(w,e) = 1 + \alpha \sum_{c \in prt(w,e)} wc(c) \cdot p(w_c|e) \quad (4)$$

$$p(w_c|e) = \frac{e \text{ 中で } w_c \text{の出現回数}}{e \text{ 中の単語出現総数}}$$

ここで $prt(w,e)$ はイベント e において，単語 w に付属する助詞の集合であり，α は 0 以上の値である．

2.3 記事検索

イベント抽出の処理対象を限定するために，記事のタイトルと本文を対象に，重要語による検索を行う．入力された記事の配信時間に近い記事を優先して検索し，定数 L の記事を抽出する．これは，入力された記事と抽出するイベントが時間的に近いと，内容的にも類似する可能性が高いという考えに基づく．単純に検索を行うのではなく，ある時間範囲の記事から類似度計算によって対象を限定した方が，重要語による検索よりも再現率が上がることが期待できる．しかし，本システムではイベントの抽出をリアルタイムに行うため，処理速度に重点を置くため，検索によるフィルタリングを利用する．

また，イベントの発生時刻は，記事は対象とするイベントが発生してから，短い時間間隔で配信されるという速報性を考慮し，時間は記事の配信時間をイベントを表す時間と判断する．これは，記事中に記載されていない場合や，記載されていても表記方法が曖昧で正確に取得できない可能性が高いためである．

2.4 イベント抽出

まず，抽出された記事集合の中で，入力された記事 A に最も類似した記事 B を見つける．次に，記事 B を含むようなクラスタを見つけ，それをイベントとする．以下にその手法を示す．

step1 入力された記事 A と最も類似した記事 B を発見する．

step2 記事 B のみを含むクラスタ K を作成する．

step3 クラスタ K との類似度が閾値以上の記事をクラスタ K に追加する．

step4 クラスタ K をイベントとする．

本手法により，記事 B を含むクラスタとしてイベントを得ることができる．本手法の利点は，記事 B を含むクラスタだけを求めるため，単純な leader-follower 法よりも類似度計算の回数が少ないことである．また，本システムでの類似度計算にはコサイン類似度による群平均法に基づいた距離関数を用い，各単語のベクトルの大きさは tf-idf の値とする．

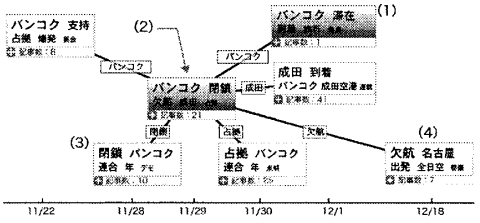

図 2: システムの実行例

3 システムの実行例

今回システムが扱う記事は，6 つのニュースサイト (asahi.com, The Japan Times ONLINE, 毎日 jp, NIKKEI NET, MSN 産経ニュース, YOMIURI ONLINE) が 2008 年 11 月 1 日から 2008 年 12 月 31 日に配信した 67,454 記事である．

システムに実際に記事を与えた場合の実行例を図 2 に示す．入力として与えた記事は，“タイからやっと帰れました 臨時第 1 便、中部空港到着” というタイトルで asahi.com から 2008 年 12 月 1 日に配信されたものである．イベントを表す矩形中の単語が抽出された重要語であり，横軸はイベントの発生日時を表す．図中 (1) で示すイベントが入力記事に該当し，そこからユーザが (2) のイベントを選択した状態が図 2 である．これ以降は，(2) のイベントから他のイベントを選択することで，イベント系列を求めていく．また，(3) のイベントを選択すれば，バンコクでのデモや情勢に関するイベント系列に，(4) のイベントを選択すれば，国内の航空関係のイベント系列に変化した．このように，イベントの選択によって求められるイベント系列が変化することが確認できた．

4 おわりに

本稿では，ニュース記事の理解を目的としたイベント系列を求める手法を提案した．そして，ユーザによって提示するイベントを変化させるため，システムのイベントの提示とユーザの選択の繰り返しによってイベント系列を求めるシステムを試作した．また，イベント系列を求める際に，記事本文から抽出した属性情報を利用した．そして，属性情報としてより精度の高い抽出が行えるように，tf-idf だけでなく，助詞の情報も利用する手法を提案した．

今回は助詞の情報も利用する手法する場合としない場合の差異などについて比較実験を行い，提案手法の効果を詳細に検討する必要がある．

参考文献

[1] 青島傳隼, 戸田智子, 福田直樹, 横山昌平, 石川博: 多様な視点からのブログ記事マイニングへの制約付きクラスタリングの適用, 情報処理学会研究報告, 2009-DBS, Vol.148, No.8, 2009.

[2] 平田紀史, 柿元宏晃, 白松俊, 大囿忠親, 新谷虎松: Web ニュース記事閲覧支援のための時間情報と重要語に基づくトピック系列化システム, 合同エージェントワークショップ&シンポジウム 2009, 2009.

[3] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron and Yiming Yang : Topic Detection and Tracking Pilot Study Final Repor, Proceedings of the DARPA broadcast news transcription and understanding workshop, pp.194–218, 1998.