E-017

時間情報を考慮したニュース記事のトピック分類とトピックの関連付け

Web News Clustering based on Time and Detection of Topic Relevance

坪井 幸雄† 山田 剛一† 絹川 博之†
Yukio Tsuboi Koichi Yamada Hiroshi Kinukawa

１．はじめに

近年、個人から企業までの様々な人々や組織が情報提供やコミュニケーションの場として、Web 上にサイトを開設している。一方、提供される情報も膨大となったため、利用者にとって必要な情報の抽出が難しくなっている。この解決方法として、様々な検索エンジンが提供され、広く利用されている。例えば、Google では PageRank と呼ばれる Web ページ間のリンクを考慮した Web ページの順序付けを行うことで利用者の直感に近い順位で検索結果を提供している。

Web 検索エンジンによって、必要な情報が掲載されている個々の Web ページは容易に抽出が可能になった。しかし、互いに関係し合う Web ページ群を順序立てて調べたいときに、現在の Web 検索エンジンは十分でない。例えば、2003 年の鳥インフルエンザ問題の経緯を調べたいとする。Web 検索エンジンを用いる場合なら「鳥インフルエンザ 2003」などのように検索質問を入力することになる。検索結果は、適当に順位付けされているか、さらに適当に分類されているかで提示される。しかし、このような検索結果の提示では、鳥インフルエンザの時間的な遷移を十分に考慮しているとは言えない。そのため、例えば上から順番に閲覧すると、利用者は時系列を行き来しながら情報を取得することになり、利用者に情報整理能力を要求することになる。また、どの情報が最新であるかを判断することは難しい。[1]

本研究では、Web 上で配信されるニュース記事を対象に、事件などのトピック（話題）を抽出し、関連性を提示することで上記の問題の解決を図る。ここで言う関連性とは、以下の処理によって得られるトピックとトピックの関連性である。

(1) 時間情報を踏まえ、ニュース記事をトピック別に分類する。（２章）
(2) 分類されたトピック同士を関連付ける。（３章）

† 東京電機大学　大学院
工学研究科 情報メディア学専攻

２．ニュース記事のトピック分類

ニュース記事をトピック別に分類するために、クラスタリングを用いる。クラスタリングとは、内容の類似しているもの同士をクラスタというグループに分ける処理のことである。ニュース記事は時々刻々と配信される無限データストリームであるため、ここではデータの走査数が１回である１パス処理のクラスタリングを扱う。なお、本稿では時間的な要素を考慮するため、類似度計算に忘却関数[2] [3]を導入する。

2.1 ニュース記事の表現

ニュース記事をクラスタリングするために、まずニュース記事を特徴量で表現する必要がある。ここでは、ニュース記事の特徴量に、文書を単語の集合と考えるベクトル空間モデルを利用し表現する。ベクトルを構成する単語は、形態素解析器によって得られる名詞と未知語のみに限定する。語の重みベクトルの生成には、次の tfidf 式を利用する。

$$tfidf(T) = f(T)\log\frac{N_0}{N(T)}$$

語 T の tfidf は、語 T の全文献における頻度を f(T)、語 T を含む文献の数を N(T)、全文献数を N_0 とする。

2.2 単一パスクラスタリング

今得た特徴量を用いて、ニュース記事を単一パス法によりクラスタリングする。単一パス法は非常に単純な方法で、文書とクラスタの類似度が閾値以上ならばクラスタに追加し、超えない場合は文書を新しいクラスタとする方法である。それゆえ、少ない計算時間で、かつ発行順序を考慮したクラスタリングを行うことができる。

この単一パスクラスタリングは、以下の手順で実行される。

① 閾値を設定する。
② 最初は、１つ目の文書自身をクラスタとする。
③ 次の文書を読み込み、既存の全クラスタのそれぞれと類似度を計算する。
④ 最も類似したクラスタとの類似度が、閾値より大きいなら、文書をクラスタに追加し、クラスタの重心を再計算する。もし、類似度が閾値より小さいなら

ば、文書を新しいクラスタとする。

⑤　③から④の処理をデータがなくなるまで繰り返す。

2.3 時間情報

本稿で取り扱う時間情報は、ニュース記事中で述べられる内容時間ではなく、ニュース記事が配信される時間（タイムスタンプ）とする。

2.4 忘却関数

本稿では、文書とクラスタ間の時間距離を考慮した類似度の計算を行うために、忘却関数[2][3]を導入する。忘却関数とは、「文書は古くなれば古くなるほど分類への重要度が減少（忘却）する」という概念に基づいた関数である。すなわち、時間的に近いものほど重要と考えたクラスタリングを行うことができる。

クラスタリングに用いる文書とクラスタ間の類似度はCosine尺度と忘却関数を組み合わせた次式とする。

$$sim(\vec{D},\vec{C}) = \omega_\lambda(|Time_D - Time_C|) \times \frac{\vec{D}\cdot\vec{V_c}}{|\vec{D}|\cdot|\vec{V_c}|}$$

ここでDは文書、V_CはクラスタCの重心を表す。$Time_D$と$Time_C$はそれぞれ文書Dのタイムスタンプ、クラスタのタイムスタンプを示す。

$$\omega_\lambda(t) = \lambda^t (0 \le \lambda \le 1.0)$$

tは時間距離を示し、その単位は日数である。λはターゲットとなる文書データに応じてチューニングされる定数である。

上記で述べた忘却関数は、定数λに基づいた類似度の減衰が行われるため、本来同一のクラスタに含まれるべきが、広がり（話題規模）と伸び（時間的距離）のあるトピックに対して、クラスタを分割してしまう可能性がある。

そこで、忘却関数に対して広がり（話題規模）を考慮し、減衰速度を任意に決定することで解決を図る必要がある。今後この処理について具体的な方法を検討していきたい。

３．トピックの関連付け

トピックが主張する概念を共有する割合が大きいほど、トピックとトピックの意味的な関連が強いという仮定をもとに、KeyGraph[4]によりトピックの主張語を取り出し、トピックの主張語同士の比較を行い強い関連性が見られるものに対して関連付けを行う。

3.1 KeyGraphによる主張語の取り出し

KeyGraphとは、文書中に出現する単語の出現頻度と共起関係からグラフを作成し、そのグラフより文書の主張点を把握し、キーワードを抽出する方法である。

本稿では、分類されたトピックに含まれるニュース記事から主張語を順次取り出していき、抽出されたすべての主張語をトピックの主張語とする。

3.2 主張語の重複による関連付け

トピックの主張語と他のトピックの主張語を次に述べる指標で比較を行い、トピックとトピックの関連付けを行う。

トピックとトピックの関連性の強さを表す指標を以下に定義する。

$$link(T_a, T_b) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

linkの値はトピックT_aとトピックT_bとの間で共有する概念の割合を表し、linkの値が大きいほどトピックT_aとT_bの関連性は強く、linkの値が小さいほど関連性は弱くなる。AはトピックT_aに含まれる主張語の集合であり、BはトピックT_bに含まれる主張語の集合である。

ここではトピックの関連性が強いトピック同士に対して関連付けを行う。

４．おわりに

本研究では、Web上で配信されるニュース記事を対象に、事件などのトピック（話題）を抽出し、関連性を提示するために必要な処理の提案を行った。

今後は、忘却関数の改良点の検討、関連付けられたトピック同士の順序付け方法の検討を進めた上で、実装および実験による有効性の検証を行う予定である。

参考文献

[1] 森幹彦, 山田誠二, “Web における話題の時間変化の提示”, JSAI2006, 2006.

[2] Yang.Y, Pierce.T, Carbonell.J, “A Study on Retrospective and On-Line Event Detection”, SIGIR98, 1998.

[3] Ishikawa.Y, and Kitagawa.H, “An Improved Approach to the Clustering Method Based on Forgetting Factors”, ECDL01, 2001.

[4] 大沢幸生, “KeyGraph －語の共起グラフの分割統合によるキーワード検出”, 電子情報通信学会論文誌 D-I, 82-D-I2,pp.391-400,1999

[5] 上嶋 宏, 三浦 考夫, 塩谷 勇, “時系列ニュース記事集合に基づくニュース記事の順序付け”,DEW2004,2004.

[6] 森 正輝, 三浦考夫, 塩谷 勇, “時制クラスタのトピック追跡”, DEWS2006, 2006.