ニュース記事の主題に着目したニュース収集・理解支援に関する研究

上村 紘輝†　　東 基衛†
早稲田大学大学院 創造理工学研究科 経営システム工学専攻†

1. はじめに

ニュースサイトでは、日々ニュース記事の配信が行われ、ニュースの閲覧が容易である。しかし、ユーザがニュースを理解する上で、「1 つの記事の情報が少ない」という問題が生じている。この問題に対し、記事の情報を収集する手段として関連記事が一般的である。

そこで本研究では記事の情報を、関連記事を用いて補うことでニュースの収集・理解支援を図る。

2. 現状分析と問題点

従来の関連記事では、記事の特徴語を抽出し、特徴語と関連の深い記事を提示している。その際の問題点を以下にまとめた。

記事に含まれる複数の主題を抽出できていない:

ニュースサイトの記事は、記事本文の前半に重要な語が出現する場合が多い[1]。そのため、重要な語が均一に分布していることを前提とした従来の特徴値算出法では、重要でない語が重要な語として抽出されてしまうことが考えられる。例えば TF(Term Frequency)法を用いた場合、記事の第 1 文(記事本文の最初の文)に存在する重要な語と出現回数が同数の重要でない語の特徴値が等しくなり、重要な語を抽出できなくなってしまう。

ユーザの興味を考慮していない:

提示されている関連記事は、ユーザに関わらず一意である。ユーザが記事に対して持つ興味を取得し、興味に関連する記事を提示する必要がある。

3. 研究目的と研究アプローチ

本研究では、記事に含まれる重要な語(主題候補語)を抽出し、それらの中からユーザが興味を持つ主題語を定め、主題語に関連する記事を提示する。

本研究のアプローチは以下の2つとする。

記事に含まれる主題候補語を抽出する:

記事において重要な語が存在する箇所を考慮し、特徴値を求める。事前調査として、人手により記事に含まれる主題候補語を抽出した結果、タイトルと第 1 文に多く出現することが分かった。そのため本研究では主題候補語を抽出する際にタイトル・第 1 文・その他(記事本文の第 1 文以外)に分けて特徴値を算出する。以下主題候補語の特徴値算出式を(1)に示す。

記事 k_j における語 t_i の特徴値 $feat(t_i,k_j)$

$$= tf(t_i,k_{j,title}) + \alpha \cdot tf(t_i,k_{j,sent}) + \beta \cdot tf(t_i,k_{j,other}) \cdots (1)$$

$$tf(t_i,d) = \frac{freq(t_i,d)}{M(d)} = \frac{\text{記事}d\text{に含まれる語 } t_i\text{の数}}{\text{記事}d\text{に含まれる総形態素数}}$$

$k_{j,(出現箇所)}$: 記事 k_j の(タイトル or 第 1 文 or その他)

α : 記事の第 1 文の重み　　β: 記事のその他の重み

feat 値の上位 N 件を記事の主題候補語とする。なお、TF 法を用いるため、高頻出の不要語は予め除去する。

ユーザが興味を持つ主題語を定め、主題語に関連する記事を提示する:

主題候補語のうち、ユーザが興味を持つ主題語を定め、主題語に関連する記事を提示する。橋本はWebコンテンツに対する興味には長期的興味・一時的興味・潜在的興味の 3 つが存在することを定義している[2]。**長期的興味**とは以前から頻繁に目にしている語への興味である。**一時的興味**とは普段あまり目にしないが、現在閲覧している Web コンテンツに含まれている語への興味である。**潜在的興味**とは一時的興味と関わりの深い語への興味である。ニュース記事を閲覧しているユーザも同様にこれらの興味を持つと考えられるため、興味を持った語に関連する記事を収集する必要がある。

以上より本研究では、主題候補語から各興味として長期的興味語、一時的興味語、潜在的興味語を抽出し、主題語とする。具体的には閲覧履歴からユーザプロファイルを生成し、ユーザプロファイルと主題候補語から各興味語を抽出し、これらに関連する記事を提示する。

4. 提案システム

4.1 システム概要

提案システムの概要図を図1に示す。

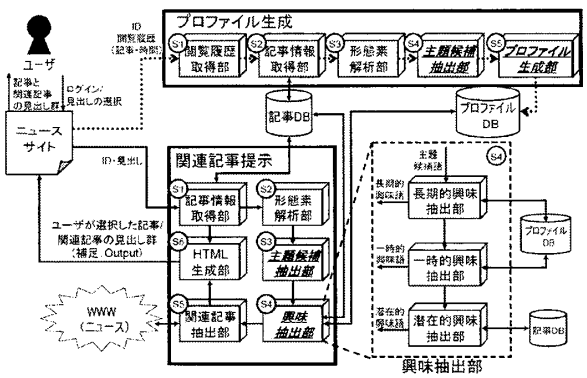

図 1.システム概要図

提案システムにおいて、ユーザはブラウザを通して本システムにログインする。これはユーザプロファイルを生

Support for Collecting and Understanding News focused on Subject of the News Article

Kouki Uemura, Motoei Azuma, Dept of IMSE, Graduate School of Creative Science and Engineering, Waseda University

成・利用するためである。プロファイル DB(データベース)にはユーザの閲覧履歴から、主題候補語 t と feat 値を算出し、総計を $Score_t$ として格納する。

4.2 関連記事提示

処理手順:

ユーザは記事を閲覧する際に見出しを選択する。

S1:ユーザが選択した記事の情報(タイトル・本文)を記事 DB より取得し、タイトル・第 1 文・その他に分割する。

S2:形態素解析を行い、名詞のみを抽出する。

S3:feat 値を算出し、上位 N 件を主題候補語とする。

S4:主題候補語とプロファイル DB からユーザの記事に対する興味語(長期的・一時的・潜在的)を抽出する。

S5:検索エージェントを用いてニュース検索を行い、関連記事の見出し群を抽出する。

S6:関連記事の見出し群と、ユーザが選択した記事の情報を用いて HTML を生成し、ユーザに提示する。

主題候補抽出部(S3):

ユーザが選択した記事に含まれている名詞について 3 節より得られた特徴値算出式(1)を用いて、記事に出現する主題候補語を抽出する。

長期的興味抽出部(S4):

ユーザが選択した記事の主題候補語のうち、プロファイル DB に格納されている語について、長期的興味値 *LongScore* を数式(3)により算出する。この値は普段閲覧している語の数値が高くなる。上位 n'件の語 t_m を長期的興味語とする。これにより、ユーザが普段から目にしている語を抽出できる。

$$LongScore_{t_m} = \frac{Score_{t_m}}{\sum_{m=1}^{M} Score_{t_m}} \cdots (3)$$

t_m :プロファイル DB と主題候補語に含まれる語

M :プロファイル DB と主題候補語に含まれる語の総数

一時的興味抽出部(S4):

ユーザが選択した記事の主題候補語に対して、一時的興味値 *ShortScore* を数式(4)により算出する。この値は、普段閲覧している語の値を下げ、ユーザが普段あまり目にしないが、記事の中で重要な語の値を上げる。上位 n'件の語 t_n を一時的興味語とする。これにより、ユーザが一時的に興味を持つ語を抽出できる。

$$ShortScore_{t_n} = \frac{1}{1+Score_{t_n}} \times \frac{feat(t_n, k_{origin})}{\sum_{n=1}^{N} feat(t_n, k_{origin})} \cdots (4)$$

k_{origin} :ユーザが選択した記事

t_n :ユーザが選択した記事の主題候補語

N :ユーザが選択した記事の主題候補語の総数

潜在的興味抽出部(S4):

記事 DB に格納されている記事の主題候補語と feat 値から、一時的興味語との共起頻度を算出し、潜在的興味値 *PotentialScore* とする。その算出法を数式(5)に示す。この値は、一時的興味語と共起している語の値を上げる。上位 n'件の語 t_w を潜在的興味語とする。これにより、一時的興味語と関わりが深く、ユーザが潜在的に興味を持つ語を抽出できる。

$$PotentialScore_{t_w} = \sum_{l=1}^{L} feat(t_w, k_l) \cdot feat(t_n, k_l) \cdots (5)$$

t_n :一時的興味語(n=1,2...n')

L :記事 DB に格納されている記事の総数

k_l :記事 DB に格納されている任意の記事

t_w :記事 k_l に含まれる任意の主題候補語

関連記事抽出部(S5):

各興味語(長期的・一時的・潜在的)に関連する記事を取得する。本研究では、関連記事の収集元として WWW(World Wide Web)上に存在するニュース記事を利用する。これは、既に閲覧済みの記事がユーザに提示されることを防ぐためである。

ニュースの検索エージェントを利用し、各興味語をクエリとしてニュース検索を行う。検索結果を関連記事の見出し群としてユーザに提示する。

5. 実験と考察

本研究の有効性を検証するために、プロトタイプを利用し、特徴値算出式(1)における α 、β の推定と提案の有用性の検討を行った。α 、β の推定では、プロトタイプで抽出された主題候補語と事前調査により人手で抽出した語との適合率を測定したところ、α =1.1、β =0.5 の場合に最も高い適合率であったため、プロトタイプではこの値を用いた。また、本研究の有用性の評価について、定量的評価としてシステムにより抽出した各興味語と従来手法の TF 法で抽出した語を比較したところ、適合率が向上した。さらに定性的評価として実施したアンケートでも過半数以上から肯定的な回答が得られた。

6. 結論と今後の課題

本研究では、ニュース記事の主題に着目し、ユーザの興味を持つ主題に関連する記事を提示することで、ニュースの収集・理解支援を達成した。今後更なる改善としてアルゴリズムの改善を考える必要がある。

参考文献

[1]佐藤 吉秀他:時系列ニュース記事における最新話題語抽出方法,情報処理学会研究報告,No.73(2005)

[2]橋本 雅幸他:興味派生を考慮した Web コンテンツ推薦システム,情報処理学会全国大会,2005