形態素を用いた内容比較によるニュースポータルサイトにおける同一トピック類似記事群の統合方式

加藤　孝祐　　加藤　裕樹　　竹内　誠　　濱川　礼
中京大学　情報理工学部　情報システム工学科

1. はじめに

本論文では，ニュースポータルサイトにおける同一トピック類似記事群の統合を行う方式とその実装，情報取得の効率化について述べる．

現在，ニュースポータルサイトにおいて数多くの新聞社から配信される記事を閲覧することが出来るが，同一時事（以下，同一トピック）を扱った記事でも内容に差がある場合が見られる．そこで，我々は形態素を用いた内容比較によって同一トピック類似記事群を1つに統合することで効率的に記事群の内容を把握出来ると考え，情報取得の効率化を目的に類似記事群統合方式の提案と実装を行った．

2. 特徴

本方式では，ニュースポータルサイトから複数の新聞社の同一トピック類似記事を取得し，記事の内容を形態素を用いて比較し，差異となる文を抽出，ベースとなる記事の適所への挿入を自動で行い記事の統合を実現している．この手法を用いることで，記事の文脈を保ったまま記事群全ての情報を統合記事に保持することが出来る．統合記事はデータベースに格納されており，HTML ファイルに出力することが出来るため，ユーザはニュースサイトを閲覧するように統合記事を読むことが出来る．

3. 手法の実現

第 2 章で述べた方式を元に記事の取得･記事の統合･統合記事の公開を自動で行う方式を提案･実装した．

また，取得する記事や画像の著作権は各新聞社に帰属する．そのため，各新聞社に記事や画像の利用許諾の申請を行い，許諾を得られた新聞社の記事に対して，実験的に使用した．

本方式では，ニュースポータルサイトからの記事取得，取得した記事群の統合，統合記事の出力という流れで処理を行う．

4. 記事取得方法

ニュースポータルサイトから記事統合の対象である同一トピック類似記事群を取得する．

まず，ニュースポータルサイト上に掲載されている記事ページの HTML ファイルを取得する．次に，その HTML ファイルから DOM 解析モジュール Zend_Dom[1]を用いて HTML の構造に沿って同一トピック類似記事群を取得する．その後，取得した同一トピック類似記事群を最も配信日時の古い記事とそれ以外の記事に分け，前者をベース記事，後者を挿入記事にしている．また，単位時間（30 分）ごとにニュースポータルサイト上の新たに掲載された記事を取得する．

なお，今回利用許諾の得られた提供社の記事・画像のみを取得するため，記事・画像利用許諾フィルタリングを行っている．

5. 記事統合方法

記事取得によって得られた記事データを形態素の比較を用いて統合する．具体的には，ベース記事と挿入記事の各文に含まれる形態素の比較を行い，各文の類似度を算出して挿入文と挿入位置の決定を行っている．

本システムでは形態素を取得する際，形態素解析エンジン MeCab[2]を使用している．その際，名詞・動詞・形容詞のみを抽出し，形態素の基本形を用いて比較を行う．

MeCab を用いて抽出した形態素を用いて，以下の二つの値を算出する．

・内容一致率(*Agr*)
・形態素評価値(*Evl*)

内容一致率とは，記事 A と記事 B それぞれにある 2 つの文の類似度を指し，後述の形態素評価値を用いて算出する．また，文間で共通しな

The integration method for identical topics' similar article groups by substance comparison adopts morpheme on the news portal site.
Kosuke Kato, Hiroki Kato, Makoto Takeuchi and Rei Hamakawa
Chukyo University Department of information engineering

い形態素は評価に用いないため，計算式で分母がゼロになることはない．

形態素評価値とは，記事 A と記事 B それぞれにある 2 つの文に共通して出現する形態素に与えられる評価値を示す．ある形態素が記事 A 記事 B に 1 回ずつ出現する場合，それぞれの文で形態素が示す内容は重複する可能性が高いが，各記事中に同じ形態素が複数回出現する場合，形態素の示す内容がそれぞれ別の事柄を指している場合もあることから，内容が重複する可能性は低くなる．そこで，本研究では形態素の出現頻度によって，何度も出現する形態素には低い形態素評価値が与えられる．

以下に各値の算出方法を示す．

・記事 A と記事 B に共通して出現する形態素に番号をつけ，x 番目の形態素に対し以下の式を用いる．

$$Evl(A,B,x)=\frac{100}{apr(A,x)\times apr(B,x)}$$

*$apr(T,x)$ ⇒ 記事 T における x 番目の形態素の出現回数

・記事 A 内の文 a と記事 B の文 b にある 2 つの内容一致率の計算式は以下を用いる．

$$Agr\ (a,b)=\frac{\sum_{k=1}^{n} Evl\ (A,B,k)}{mrp\ (a,b)}$$

*$mrp(x,y)$⇒文 x, y 中の出現形態素数

*n⇒文 a，文 b に共通して出現する形態素数

なお，上記 2 つの式は[3]を参考にしている．

この式をベース記事と挿入記事の 2 つの文の組み合わせ全てに当てはめる．このとき，どのベース記事の文とも閾値を超えなかった挿入記事の文はベース記事から独立した内容だと見なすことが出来，ベース記事への挿入対象となる．

例として，[図 1]の記事の全ての文の内容一致率を求めると，[表 1]のようになる．

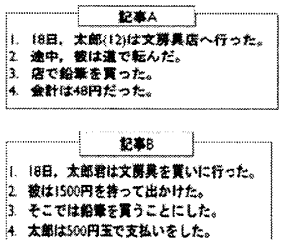
挿入位置の決定には，独立文の前後と内容一致率の高いベース記事の文を探す．例として，独立文 b の 1 つ前の文 a がベース記事の文 c と内容一致率が高いとき，表す内容は文 a≒文 c ある．ここで，文脈が文 c→文 b となるようにベース記事に挿入することで，文脈を損なうことなく記事の統合を行うことが出来る．

また，本章で解説した処理を記事群全ての記事に対して行う．

表 1 内容一致率表の例

	A1	A2	A3	A4
B1	68	0	11	0
B2	0	12	0	0
B3	0	0	50	0
B4	4	11	0	14

図 1 ベース記事・挿入記事の例



6. 出力

[図 2]に実際の出力例である表示画面を示す．

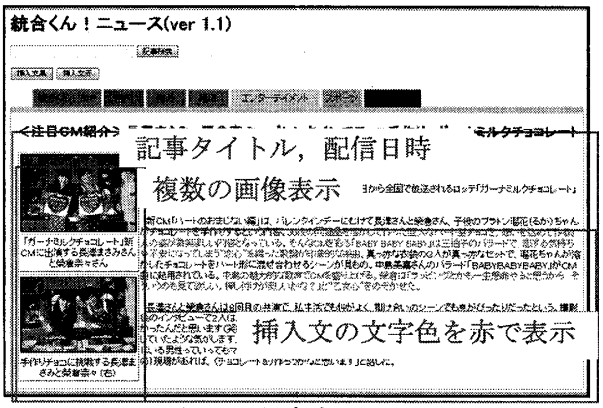


図 2　記事表示画面

7. まとめと今後の展望

21 名の被験者に本システムを使用した評価を 4 段階評価で行った．「統合された記事は，元になる記事全ての情報を含んでいるか」という項目では，平均 3.8 点と高評価を得たことから，内容を網羅した記事の生成ができたと言える．

今後の改善点として，記事統合に用いるベース記事の自動選定機能の実装が上げられる．統合記事に占める挿入された文の割合が少ないほど，より文脈を損なわない記事の統合が可能であると考えられる．さらに，挿入する文量の自動調節機能や，ベース記事に適した記事を自動選別する機能を追加することで，読みやすい統合記事を生成できると考えられる．

参考文献

[1]Zend_Dom
http://www.zend.com/en/
[2]MeCab
http://mecab.sourceforge.net/
[3] 柴田 昇吾，上田 隆也，池田 裕治：複数文章の融合，電子情報通信学会技術研究報告