1A3-1

NewsMLのための特徴語の自動抽出

Automatic Extraction of Feature Terms for NewsML

大川原雄也[*1] 大圃忠親[*1] 伊藤孝行[*1] 新谷虎松[*1]
Yuya Okawara Tadachika ozono Takayuki Ito Toramatsu Shintani

[*1]名古屋工業大学大学院 工学研究科 情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

This paper presents a feature term extraction algorithm. Extracted terms can be applied to NewsML. We presents an algorithm based on a class model used for the natural language processing. Our algorithm is based on set of occurrence information between each term of a class and terms of other classes. By using a class model, the classification of articles is also possible at the same time as extracting terms. The feature terms are also representive terms of classificated groups. We can retrieve articles relative to an aritcle, by addition of feature terms to articles in the NewsML format.

1. はじめに

NewsML は，インターネットを通じ，テキスト，画像，動画などの異なる性質を持つニュース素材を配信する方法として発案されたものである [1]．NewsML は，ニュースを配信する際に，ニュースの発信者，作成日およびニュースの内容などのメタデータを付加して，報道機関同士で情報を扱いやすくしたものである．

NewsML には，多くのメタデータを記述することが可能である．例えば，記事の種類，作成者，更新時刻，重要度，関連記事，提供者，記事の内容，場所などを記述できる．メタデータを利用することで，コンテンツの比較，再利用，検証などが格段に容易になる．また，多くの記事の中からユーザが興味のある記事を見つけることも容易になる．これらのメタデータを自動的に抽出することができれば，文書の内容や関係を把握することが容易になる．また，読むべき文書を選択することも容易になる．一方，メタデータの付加には大きなコストがかかる．メタデータを有効に利用するためには，メタデータを自動的に抽出および追加する仕組みが不可欠である．

本稿では，NewsML のメタデータとして利用可能な，記事の特徴語の抽出および記事のクラスタリングを行うことを目的とする．特徴語は他の語との共起の度合いによって求まる．特徴語は，記事の内容に関係し，クラスタに属する他の記事の内容にも関係する．本稿では，特徴語は，記事をクラスタリングすることにより得られる．

提案するアルゴリズムにより抽出された特徴語を，NewsML のメタデータとして記事に追加することで，記事の分類および関連した記事の検索が可能である．

2. クラスモデルによるクラスタリング

2.1 クラスモデル

本稿では，クラスモデルに基づいた手法によってクラスタリングを行う．クラスモデル [3] は，学習用データから得られた情報をもとに単語を自動的に分類し，その分類結果を用いた言語モデルである [2]．クラスモデルでは，学習データにおけるクラスモデルの対数尤度を最大化させるクラス分類を最適な分類とする．この分類結果を，推定だけでなく他の用途に利用することも可能である．本稿では，クラスモデルによる記事の分類結果をクラスタリングに利用した．単語を分類し，言語モデルとして利用する利点は，N グラムモデルでは十分に学習できない，学習パラメータが少ない場合でもクラスモデルは推定すべきパラメータ数が少ないため，有効な言語モデルとして用いることができる点が挙げられる．単語 w_n の属するクラスを c_n とするとき，クラスモデルは以下のように表される．

$$P(w_n|w_{n-1}) = P(w_n|c_n)P(c_n|c_{n-1}) \tag{1}$$

確率 $P(w_n|c_n)$ は，単語 w_n がクラス c_n から生起する確率であり，次式により推定できる．

$$P(w_n|c_n) = \frac{N(w_n)}{N(c_n)} \tag{2}$$

ここで，$N(w_n)$ は，学習データ中で単語 w_n が出現した回数であり，$N(c_n)$ は，クラス c_n の単語が出現した回数である．

クラスモデルでは，学習データの対数尤度を最大化するクラス分類をより最適な分類と考える [3][4]．クラスモデルにおける学習データの対数尤度 $L(\pi)$ は以下のように計算できる．

$$L(\pi) = \sum_{c_1,c_2} C(c_1,c_2)\log\frac{C(c_1,c_2)}{C(c_1)C(c_2)} + \sum_{w} C(w)\log C(w) \tag{3}$$

右辺の第２項は，クラスに依存しないため，第１項を最大化することで最大の $L(\pi)$ が求まる．

2.2 クラス分類を求めるアルゴリズム

クラスモデルでは，各単語に１つのクラスを割当て，$L(\pi)$ を最大化するようにクラスを次々に併合していく方法でクラス分類を求める．以下にクラス分類を求めるアルゴリズムを示す．

1. すべての単語に対して１つのクラスを割り当てる．
2. 各クラスの単語を，他のクラスに移動したときの $L(\pi)$ を計算する．
3. $L(\pi)$ を最大化させるクラス C_m を求める．
4. 単語をクラス C_m に移動する．
5. ステップ２〜4 を $L(\pi)$ が収束するまで，または決められた回数繰り返す．

連絡先: 大川原雄也，名古屋工業大学大学院 情報工学専攻 新谷研究室 〒 466-8555 名古屋市昭和区御器所町 名古屋工業大学，TEL: (052)733-6550，FAX: (052) 735-5477，yuya@ics.nitech.ac.jp

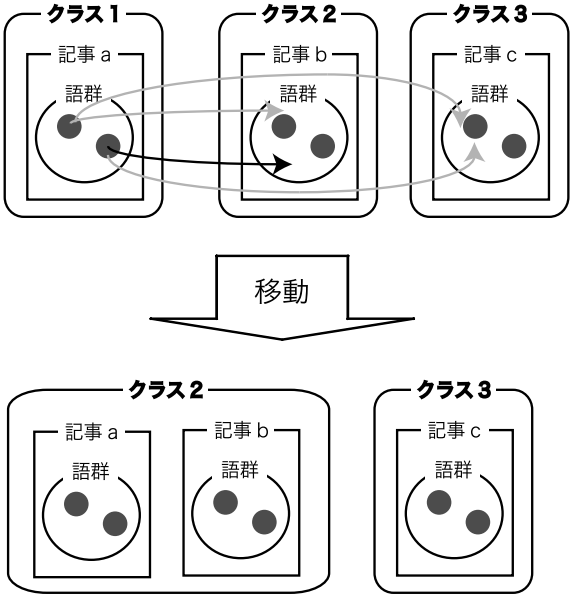

図 1: クラスタリング処理の例

3. 特徴語抽出のためのクラスタリング

3.1 記事のクラスタリング

本稿では，２節のクラスモデルのクラス分類に基づいたクラスタリングを行う．提案する手法では，記事のクラスタリングの結果から特徴語を抽出する．

通常，クラスモデルは，データ中のすべての単語を対象に処理を行う．本稿では，記事を対象としている．記事のタイトルは，人手によって，記事内容を最大限に要約したメタデータである．本稿では，その特性を利用し，タイトル中のすべての名詞および本文中の重要度が高い名詞を対象に処理を行う．名詞を抽出するための前処理として，形態素解析を行う．形態素解析には茶筌[*1]を用いた．重要度は，TF-IDF によって計算する．重要度が高い語は，記事の本文での重要度が上位の語のことである．

また，通常のクラスモデルでは，クラスのメンバは単語であるが，記事をクラスタリングするために，クラスモデルにおけるクラスに属するメンバの単位を記事とする．クラスの移動は，前述の語群を用いて，対数尤度 $L(\pi)$ の評価することで行われる．$L(\pi)$ を最大化させる移動先が求まったとき，記事をその移動先に移動する．

記事 A_i のタイトル中のすべての名詞および本文中の重要度が高い名詞の集合を W_i とする．対数尤度 $L(\pi)$ の計算は，他のクラス C_k に対して，W_i に含まれる語と，$W_j(A_j \in C_k)$ に含まれる語のすべての組み合わせに対して行われる．W_i の語数を m，$W_j(A_j \in C_k)$ の語数を n_k とするとこの組み合わせは mn 個ある．他のすべてのクラスでこの組み合わせに関して対数尤度の評価を行う．通常のクラスモデルと同様に，対数尤度を最大化させるクラスを求めて，クラスのメンバである記事をそのクラスへ移動する．

以下にクラスタリングのアルゴリズムを示す．図 1 にクラスタリング処理の例を示す．

1. 各記事にそれぞれ 1 つのクラスを割り当てる．
2. 各記事のタイトル中の名詞および本文中の重要度が上位の語を抽出する．記事 Ai より抽出された，これらの語の集合を W_i とする．
3. W_i 中の語を，他のクラスに移動したときの $L(\pi)$ を計算する．
4. $L(\pi)$ を最大化させるクラス C_m を求める．
5. 記事 A_i および W_i をクラス C_m に移動する．
6. i を変更し，適当なクラス数になるまでステップ 3〜5 を繰り返す．

3.2 クラスタリング結果からの特徴語抽出

記事 A より抽出された W 中に含まれる，語 t のクラス移動に伴う対数尤度 $L(\pi)$ の変量を ΔL とする．閾値 α に関して，$\alpha < \Delta L$ を満たす語 t を，クラスタに関連する語および語 t が属する記事 A の特徴語とする．

4. NewsML への利用

本提案手法による特徴語の抽出および記事のクラスタリングの結果は即座に NewsML 化することが可能である．クラスタリング結果を NewsML 化するには，クラスタに ID などを割り当てて，その ID をメタデータとして付加すればよい．記事から抽出された特徴語をメタデータとして記事に付加することで，記事の内容を考慮した記事検索が可能である．また，記事の主題をつかむ手がかりとなることも考えられる．

クラスタリング結果をメタデータとして付加することで，関係する記事の検索が容易になる．記事の分類は，記事の管理に直接影響する要素である．政治やスポーツなどの大きなカテゴリを小分類できれば，管理が容易になると考えられる．提案手法における入力は記事のみであるため，人手による作業のコストが軽減できる．

5. おわりに

本稿では，クラスモデルにおけるクラス分類の考え方をもとに，対数尤度を変化させる語に注目した．そして，クラスモデルに基づいた手法により記事をクラスタリングした．クラスタリングの結果として得られた，クラスタに関係のある語を記事の特徴語として抽出を行った．特徴語は，NewsML に付加可能なメタデータである．特徴語をメタデータとして付加することで，記事の分類および関連記事の検索が可能である．

参考文献

[1] 井上明, 猪狩淳一, 金田重郎, ”ニュース配信のための国際データフォーマット NewsML：その概要と現状について”, 情報処理学会研究報告, 2002.

[2] 北研二, ”確率的言語モデル”, 東京大学出版会, 1999.

[3] S. Martin, J. Liermann and H. Ney,”Algorithms for Bigram and Trigram Word Clustering”, Proc. EUROSPEECH-95, Madrid, pp.1253-1256,1995.

[4] MacMahon, J. G. and Smith, F. J., ”Improving statical language model performance with automatically generated word hierarchies”, Computational Linguistics, 22(2), pp. 217-247, 1996.

*1 http://chasen.naist.jp/hiki/ChaSen/