ニュース記事分類と検索のための自動タグ付け手法の提案

塩津 和貴
東京工科大学大学院
バイオ情報メディア研究科
g211002968@st.teu.ac.jp

岩下 志乃
東京工科大学
コンピュータサイエンス学部
iwashita@cs.teu.ac.jp

1　はじめに

Yahoo!ニュース [1] などのニュースサイトに対する評価として, ニュースサイトは「情報の種類が豊富」, 「情報更新が早い」といった利点があり, 実際, 1 日に 1 回以上サイトを通してニュースを閲覧するユーザは 6 割を超えているという調査結果がある [2]. 一方で, 記事の量が多く「読みたい記事が探しにくい」といった不満点も挙げられている. そこで大量の記事をどのように分類し提示するかが問題となっている.

我々はこれまで, TF-IDF によってキーワードを抽出しタグ付けを行うという手法を用いて, 記事の分類および記事検索システム [3] を提案してきた. 1 つの記事において TF-IDF により推薦されたキーワード群の中から上位 5 つのキーワードをタグとして付与させてきた. しかしこの手法において, ある記事において関連性が高いがそれぞれ異なるタグが付与されたため, 記事抽出ができないという問題が生じた.

そこでより快適に記事検索を行うために, 記事同士の関連性に注目してタグ付けする手法を提案する. 収集した記事についてコサイン類似度を用いて記事同士の類似性を測定し, 記事間の距離を求める. また, この計算結果を用いて複数のクラスターを作成することで, 関連の高い記事だけを収集する. これら収集した記事において高い頻度で出現するキーワードをそのグループの特徴単語とし, 該当する全ての記事に付与する.

最終的には記事 1 つにつき, TF-IDF により推薦されたタグが 5 つ, 関連記事の中から抽出したキーワードタグが 5 つ付与されることになる.

2　コサイン類似度による記事間類似度

Yahoo!ニュースで扱われる記事はそれぞれサッカーやゴルフ, 政治, 経済などのカテゴリ情報を持つ. そこで本研究ではカテゴリごとにコサイン類似度を用いて記事間の類似度を測定する.

コサイン類似度はベクトル間の角度を用いた類似性測度であり, 情報検索の分野におけるベクトル空間型モデルにおいてよく利用されている [4]. 2 つの記事の類似度は, 各記事の単語ベクトル x, y を用いると, 式 (1) で表される.

$$cos\ \theta = \frac{< x, y >}{|x||y|} \tag{1}$$

記事の特徴は記事に出現する単語とその出現数により, その記事の特徴として捉えることができる. コサイン類似度を用いることで 2 つの対象についてそれらの特徴から類似性を計算する.

3　関連タグの付与

3.1　処理の流れ

本研究では, 収集した関連記事群の中から抽出した特徴的なキーワードのことを関連タグと呼ぶ. 関連タグを付与するまでの処理の流れを図 1 に示す.

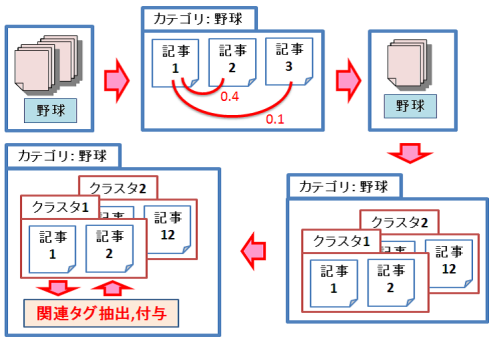

図 1: 関連タグ付与までの様子

各処理の詳細は次の通りである.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

処理 1 カテゴリ別に収集した記事群においてコサイン類似度を用いて記事の類似度を測定する.

処理 2 類似度が高く関連性が高い記事を収集する.

処理 3 記事群の中でクラスタを生成する.

処理 4 クラスタ内において頻出単語 (関連タグ) を抽出する.

処理 5 頻出単語 (関連タグ) をクラスタを構成する記事に対して付与する.

3.2 コサイン類似度による記事間類似度の計算

カテゴリごとに収集した記事群において記事同士の類似度を測定する. 記事の特徴ベクトルとなる要素としてタイトル部分と本文の第 1 段落部分に含まれる名詞を用いる. 記事の主題を表すキーワードが多く含まれる可能性が高いためである.

収集した記事には識別用の ID を割り振る. 記事の組み合わせ類似度の結果をデータベースに保存する. 得られた記事は類似度を基準に降順にソートを行う. また事前調査で類似度 0.25 を比較する記事が類似するかどうかの閾値として定めた.

3.3 類似度に基づくクラスタリング

類似度の計算結果より類似度が高いものから 0.25 までの記事をクラスタの生成対象として用いる. ただしクラスタ生成対象として抽出した記事群には様々なトピックの記事が含まれる. 例えばサッカー関連記事で抽出した記事群には, J リーグ, なでしこジャパン, W 杯など複数のトピックが存在する. そこでカテゴリごとに抽出した類似度が高い記事群に対して最適な数のクラスタを生成する.

本研究では, ある期間に配信された Yahoo!ニュース記事に対して, カテゴリごとに生成されるクラスター数に関する調査を行う.

3.4 関連タグの抽出

生成されるクラスタは, あるトピックに関連する記事が集まった集合体である. そこでクラスタごとに収集された記事を用いてキーワードの出現頻度を求める. 頻出するキーワードはこのクラスタを特徴づけるキーワードとして記事の分類に有効であると考えられる. 本研究ではこのキーワードを関連タグとして, クラスタごとに収集された記事全てに対してタグ付けを行う. また, 出現頻度の高いキーワードの中から上位 5 つが関連タグとして付与される.

4 記事の分類と関連タグの生成

記事データ (カテゴリ:野球) に対してクラスタ生成を行った. 生成されたクラスタの一部を図 2 に示す.

楽天・岩隈、マリナーズ入りで笑顔「僕を必要としてくれた」(産経新聞)
---(類似度：0.54)岩隈、今月末にマリナーズ本拠地でお披露目（サンケイス
---(類似度：0.538)岩隈、メジャー仕様！もうブルペン５０球（サンケイス
---(類似度：0.519)岩隈、来週にも渡米…ファンフェスタ参加へ（サンケイ
---(類似度：0.468)「皆さんに笑顔を」ＭＬＢ日本開幕戦で両監督意気込み
---(類似度：0.324)ＦＡ岩隈、移籍先にマリナーズ再浮上！（サンケイスポ
---(類似度：0.296)ポスティングは「なくすべきだ」レンジャーズ上原、報
---(類似度：0.276)ダルビッシュ＆マー君が合同自主トレ！（サンケイスポ

図 2: 生成されたクラスタ (一部)

楽天の岩隈選手の大リーグ移籍に関連する記事が収集されていることが分かる. またこのクラスタの関連タグとしては投手, 岩隈, 楽天, マリナーズ, 大リーグなどのキーワードが抽出された. これらのキーワードはこのクラスタを特徴付けるキーワードであり, 関連タグとして付与されるキーワードとしては適切であると考えられる.

5 おわりに

本研究は, ニュースサイトで配信される記事に対して記事の分類と簡単な検索を実現するための有効なタグ付け手法を提案した. コサイン類似度を用いて記事をクラスタリングすることで, クラスタに関連するタグを付与することができた.

今後の課題としては, 対象の記事データに対して適切な数のクラスタを生成できる手法の導入とクラスタごとに収集された記事に対して付与する関連タグの精度を高めることが挙げられる.

参考文献

[1] Yahoo!ニュース, http://headlines.yahoo.co.jp/hl
[2] MyVoice 第 2 回ニュースサイトに関する調査, 2008, http://www.myvoice.co.jp/biz/surveys/11708/index.html
[3] 塩津 和貴, 岩下志乃 :“ 自動タグ付けによるニュース記事の分類と快適な検索システムの提案, ”, ファジイシステムシンポジウム講演論文集, pp.99-101, 2010
[4] 岸田和明, 情報検索の理論と技術, 勁草書房, 1998

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved