1M-3

Wikipediaの時系列アクセス数に着目した関連度算出

瀧口 裕一† 倉門 浩二†† 大石 哲也††† 越村 三幸†††† 藤田博†††† 長谷川 隆三††††
†九州大学工学部電気情報工学科 ††九州大学大学院システム情報科学府
†††九州大学情報基盤研究開発センター ††††九州大学大学院システム情報科学研究院

1 概要

近年，情報化社会が進むにつれて，新しい言葉が多く生み出されるようになった．その新語の意味や，それに関する事象を調べるとき Wikipedia を利用する人が多い．

Wikipedia は，大規模な Web 上の百科事典というべきもので，誰でも記事内容を変更できることが大きな特徴である．また幅広い分野の記事（概念）を網羅しており，既存の概念から新しい概念までの記事が存在する．その記事数は日本で 80 万弱 (2010 年時点) に上り，今でもその記事数は増加している．加えて，Wikipedia の全記事の記事名や本文を ダンプデータとして提供しているため，容易にデータの解析が可能である．

上記のような特徴のため Wikipedia を利用した研究が多く，その中でも概念間の関連度算出の研究が盛んに行われている．

一方で，ダンプデータに含まれていないが，全記事の 1 時間ごとのアクセス数も公開されている．その時系列データを用いた研究はあまりされていない [2].

関連度算出に関する既存研究においても，文書の数と文書間の中に含まれる単語の頻度から推定するものが多く，時系列データを用いたものはなかった．そこで本研究では，新しい手法として Wikipedia に存在する時系列データを用いて，記事同士の関連度を算出する新たな方法を提案する．

本研究では大きく以下の 3 つのアプローチから関連度の算出を試みる．

1. 記事同士のバースト期間の関連を比較する．
2. リンクを利用する．
3. ニュース記事を利用する．

2 節では，上記の 1，2 について述べ，3 については 5 節で述べる．

2 バースト共起

2.1 システム

時系列データ（アクセス数）を取得し，解析を行う．Wikipedia の全ての記事に対してアクセス数が極端に上がった（以下バーストしたと言う）時間を算出する．この期間を比較することにより，同じ時期にバーストした二つの記事の関連度を算出する．2 つのバーストの関連度の計算法として曽根らは, バーストの共起を算出する手法を提案している [3]．我々はそのバースト共起を参考にし，以下のシステムを考案した．

1. Wikipedia の各記事に対する 1 時間毎のアクセス数を取得する．
2. 取得したアクセス数から全ての記事に対してバースト期間 p_t を算出する．
3. 各記事の p_t を比較し，バースト期間がどれほど重なっているかを表す数値 $Overlap$ を出す．
4. $Overlap$ の高い記事を比較する．

2.2 実験

使用した時系列アクセス数のデータは，2010 年 4 月 1 日から同年 9 月 30 日までの 6ヶ月間に 1 時間毎で取られたもので，Wikipedia の約 78 万の記事を対象としている

2 節で説明した方法で実験を行い，関連度の高い記事の対を挙げ，人手でその関連性を確認した．

バースト検出を用いた方法では以下の 2 通りのバースト期間の比較方法を試した．

1. 記事 A とバースト期間が算出できた全記事とを比較.
2. 記事 A と，A 内のバースト期間が算出できたリンク先の記事のみとを比較.

またバースト期間は 1 日単位で算出した．

2.3 結果

2.3.1 全ての記事との比較

ほとんどの記事でこのような，関係性がおよそ無いと考えられる結果が得られ，関係性のあるものが上がることは無かった．例えば,「サッカー」と関連度の高いものに「数学」が提示されてしまい，明らかに関係性のないものが上位に挙がってしまった．

2.3.2 リンク先の記事のみとの比較

記事とその内部にあるリンクは関連性が高いと考えられる．実際，関連のより高いリンク先が上位に上がってきており，全ての記事との比較による場合と比べて良い結果を得た．

2.4 考察

以上のように，全般的には良好な結果を得ることはできなかった．この理由としていくつか考えられる．

1. 平均アクセス数が少数の場合，アクセス数が数十件増えただけでバーストと判定されてしまう．そのため結果的にノイズとの判別がつかなかった．
2. 偶然バースト期間が同じだった．
3. 記事同士の関連度と時系列アクセス数の相関は本質的に極めて小さい．

実験結果から，1 と 2 が主な原因とは言い難い．よって 3 が主な要因と考えられる．時系列アクセス数の比較から記事同士の関連度を算出することは困難であると考える．

3 ニュース記事の利用

関連語というものはその時代の情勢によって変わってくる．今まで関連性は無いと思われた単語同士に関連性が生じるということが大いに有り得る．たとえば多くの人にとって，相撲と野球賭博は今まで関連性がなかったが，最近では大きく関連しているといったようなことである．

そこで，既存の関連度に時事性を加味した，新たな関連度を算出することを試みる．

時事性という観点から今回，インターネット上のニュース記事を利用した．

3.1 手法

予備実験において，大きなニュースが報じられたその日に，ニュースに関連する Wikipedia の記事のアクセス数が急激に増加することを確認した．従って，ニュースに関連する Wikipedia の記事各々のアクセス数がバーストした時を算出し，比較することで関連度に時事性を反映させることができると考えた．手順を以下に示す．

1. ニュース記事の見出しからキーワード (Wikipedia に記事が存在するもののみ) を決定する．
2. キーワードと関連度の高い単語（Wikipedia に記事が存在するもののみ）を抽出する [1].
3. ニュース記事の日付と 2 で抽出した単語とで *Overlap*(2.1 節参照) を算出し比較する．

3.2 実験

まず Yahoo!ニュース Web API を用いてニュースの見出しを取得した．次に形態素解析を行い，Wikipedia に含まれる単語 (キーワード) を抽出し，上記の手順で実験を行った．今回は「琴光喜啓司」をキーワードとした．

結果は「佐渡ヶ嶽，時津海正博，豊ノ島大樹，…」といったように野球賭博に関わったとされる力士が上位に選出された.「佐渡ヶ嶽」はもともと関連度は 50 位ほどだったことを考慮すると，上手く時事性を反映させることができたと考えられる．

3.3 関連度算出

実験で得た *Overlap* 値を最大値 1 で正規化した値と，3.1 節の 2 で抽出した単語の関連度を足し合わせることで，新しい関連度を算出した．以下に結果の一部を示す．実験前に比べ，実験後は関連度がかなり大きくしているが，これは時事性を重視したためである．

表 1: 関連度算出結果

名前	佐渡ヶ嶽	時津海正博	豊ノ島大樹
(1) 正規化値	0.833	0.833	0.666
(2) 時事性反映前	0.433	0.427	0.868
(1)+(2)	1.266	1.260	1.534

4 まとめ

時系列データから記事同士の関連度を算出する手法について検討した．単純にバースト期間を比較するということでは，関連性のあるものを抽出すること自体が困難であった．

またニュース記事利用については，従来，関連語として認識されていた単語対の関連度を加味して，適切に修正することに成功した．今後は以前は全く関係のなかった単語に，新しい関連性が生じた対を発見する手法について研究を進めたい．

謝辞 本研究は科研費 (21500102) の助成を受けたものである．

参考文献

[1] K. Nakayama, T. Hara, and S. Nishio. Wikipedia mining for an association web thesaurus construction. *Web Information Systems Engineering–WISE 2007*, pp. 322–334, 2007.

[2] 曽根広哲, 山名早人. ウィキペディア記事閲覧回数の特徴分析.

[3] 平野真太郎, 成凱, 岩井原瑞穂. 階層型カテゴリを用いたウェブサイトのアクセス履歴の時系列相関性解析.