Vol.2009-SLP-77 No.2
2009/7/17

Wikipedia ページへの tfidf 法の適用

宮崎将隆[†]　　川端豪[†]

本報告では tfidf 法に基づく話題キーワード選択法の改良を行う．ブログなどの限定された少数ページから tfidf を計算しようとすると，その基となる tf 及び idf の値が精度良く求められない．まず，idf については Web ページ全体から算出した idf で Wikipedia から算出した idf を近似できることが分かった．次に，tf については単語共起に基づくクラスタリング手法を導入し，キーワードのグループを構成した．少数ページから tf の計数を行う際に，グループに含まれるすべての単語の計数値の総和で代用する．実験によって，このようにして求めたグループ tf が真の tf と強い相関を持つことを確認した．

A Study of “tfidf” Measure using Wikipedia Statistics

Masataka Miyazaki[†] and Takeshi Kawabata[†]

This paper describes an improvement of the keyword selection criteria based on the “tfidf” measure. It is very difficult to estimate “tf (term frequency)” and “idf (inverse document frequency)” values from small amount of weblog pages. First, we investigate an approximation of the world wide idf value as the Wikipedia idf value. Experiments show that this idf approximation is promising. Secondly, we apply the clustering method to word co-occurrence and make several word groups. The tf value of a keyword is extrapolated as the sum of its group word frequency. Experiments show that the group-word based tf values counted in small amount of pages are strongly correlated to the true tf values.

1. はじめに

ネットワークユーザの一個人が，ブログなどから情報を簡単に発信できるようになり，Web のコンテンツは増大する一途である．ページの執筆において，アクセスのキーとなるタグを付けることが必要であるが，個人によって基準も様々であり，これをコンピュータが自動的に付与する技術は有用である．

ページの話題判定には，予め設定したキーワードの出現数がよく用いられるが，ブログに代表される限定ページでは，文書の絶対量が少ないためいろいろな問題が生じる．話題判定キーワードの選択には tfidf [1]と呼ばれる尺度がよく用いられている．この量は，話題ページ中のキーワード出現数 tf と全体ページ中のそのキーワード出現の偏りを表す idf の積で計算される．

本報告では，ある Web ページ集合に関する idf をより一般的な Web ページ集合に対する idf で近似することを検討する．また，クラスタ化された単語について，少数のページに関する tf 計数を行うことによって，全体のページの tf の近似を検討する．

2. ブログページと話題キーワード

Web2.0 の時代になり，エンドユーザの一人一人が例えばブログページを介して情報を作成，配信する機会が増加している．ブログページ執筆の際に，そのページに興味を持つ相手がアクセスしやすいように「タグ」を付けることがよく行われるが，これをコンピュータが判断して自動的に行えれば便利である．また，共通的な「話題」に基づいてタグが付けられるようにすれば，タグの一貫性が向上し，検索の観点からも有用である．

Web 上のあるページがどのような話題に属するか判定するために話題ごとに予め設定したキーワードがそのページ中にどれだけ現れるかを調べる手法が用いられる．しかし，ブログに代表される限定ページについて，このようなキーワードを設定しようとすると，文書の絶対量が少ないためいろいろな問題が生じる．

2.1 キーワード抽出のための tfidf 法

ここでは情報検索の分野でキーワードの重要性を計算する手法である tfidf 法について述べる．この量は，ある文書中に出現するキーワードの頻度 tf と，キーワードの全文書中における頻度の偏りを表す idf の積で計算される．

「tf」とは Term Frequency の略語で，ある単語が文書中に出現する頻度を表す．頻度を求めるには文書を最初から最後まで検索して，単語が何回でてきたかを数える．繰り返し出現する単語は重要であるという概念の基に成り立っているので，頻度の多い単語に重みを大きくする．ある話題に属するページ集合 d に関する単語 w の出現頻度を tf（w,d）と記す．

「idf」とは Inverse Document Frequency の略語で，ある単語が他の文書中のどれくらいの文書に出現するかという尺度で表す．次のような計算式で求めることができ，他

[†] 関西学院大学　理工学研究科
School of Science and Technology, Kwansei Gakuin University

Vol.2009-SLP-77 No.2
2009/7/17

の文書にあまり出現していない単語の重みが大きくなる．単語 w に対する idf を下式により計算する．

$$idf(w) = \log \frac{N}{df(w)} \cdots(1)$$

w ：キーワード候補の単語

N ：検索対象となる全ページ数

$df(w)$：w が出現するページ数

「tfidf」は上で説明した tf と idf の積で定義される尺度によってキーワードを抽出する手法である．繰り返し出てくる単語は重要であるという概念を背景に持つ tf と，他の文書にあまり出てこない単語はこの文書を特徴付けるという概念を背景に持つ idf を組み合わせることで，文書を代表するキーワードを見つけている．tf だけではどの文書にも出てくるような文書を書く上でよく用いられる単語(助詞など)もキーワードとして誤検出してしまうので，idf を組み合わせることで精度を向上させている．

tfidf の計算は具体的に次のような手順で行う．図 1 に示すように，キーワード候補のある単語に注目し，全ページ数 N の中でその単語の出現するページ数 df を数え，(1)式で idf を算出する．また，ある話題に属するページ中からその単語が出現する数 tf を数える．両者の積が tfidf であり，その単語がその話題を抽出するキーワードとしての有効性に対応する．

$$tfidf = tf(w, d) \times idf(w) \cdots(2)$$

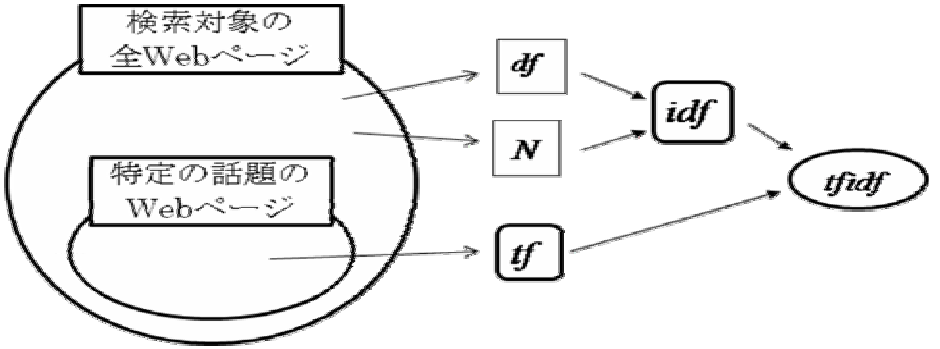

図 1：tfidf の計算手順

2.2 idf 算出の問題点

idf の算出には検索対象となる全ページ中での単語 w の出現頻度を数えなければならないが，ブログのような限定されたページにのみアクセスできる条件では，その「検索対象となる全ページ」を特定することが困難である．しかし，idf の意味が，ある単語がさまざまな話題に関わりなく広く使われるか，あるいは特定の話題に対し，偏って出現するかを表す量であることを考えると，これはその語の持つ「意味」そのものに左右されると思われ，より一般的なページ集合に対する統計から求められる可能性がある．

本報告では，ある Web ページ集合に関する idf をより一般的な Web ページ集合に対する idf で近似することを検討する．

2.3 tf 算出の問題点

tf の算出には特定の話題の Web ページ中での単語 w の出現頻度を数えなければならないが，ブログのようなアクセスできるページが限定される条件では，「特定の話題のページ」を特定することが困難である．idf の場合ではより一般的な Web ページ集合に関する統計をとることが考えられるが，tf の場合はある特定の話題に属するページであるかどうかを検索ヒットのみから判定することはできないので，さらに工夫が必要である．

松尾ら[2]は単語間の類似度を測る指標として，単語の共起の偏りに着目している．このような考えを用いれば，類似した単語をクラスタ化することができ，限定された

ページ中における単語の計数における標本の少なさを同じクラスタ内の単語と合わせて計数することによって，補えないだろうか．

本報告ではクラスタ化された単語についてある少数のページに関する tf 計数を行うことによって，全体のページの tf の近似を検討する．

3. idf の検討

3.1 基本的な考え方

情報検索における idf は，基本的には検索対象となる全ページを対象とする統計によって求められる．しかし，複雑化する Web 利用シーンの中には，検索対象となるページ集合を明確に決められない状況も生じる．本節では idf の算出を検索対象ページそのものではなく，それを含むより一般的なページ集合に対して算出し，両者の相関を観察する．一般的なページ集合での統計を行う手段として，検索エンジンのヒット数を利用する．idf に対してこのようなアプローチが成立するかどうかを検証するために，明確にページ範囲を特定できる Wikipedia に注目する．Wikipedia の「サッカー」という項目中に含まれる頻出単語 100 をキーワード候補と考え，後に述べる手法により idf を求める．この値を Wikipedia 中のページに関して算出した idf と比較し，両者の一致を検討する．

3.2 実験条件

ある単語に対して，Wikipedia 内検索を用いて求めた idf と Yahoo 検索を用いて求めた idf を比較して，2 つの値が相関を持っているかを調べる．idf の計算は(1)式を使って Yahoo と Wikipedia それぞれ次のように求まる．

【Yahoo】 N ： Yahoo の総インデックス数(約 192 億, ’08)

$df(w)$ ： Yahoo 検索でのヒット数

【Wikipedia】 N ： Wikipedia の総記事数(約 53 万, ’08)

$df(w)$ ： Wikipedia で全文検索をした際の記事のヒット数

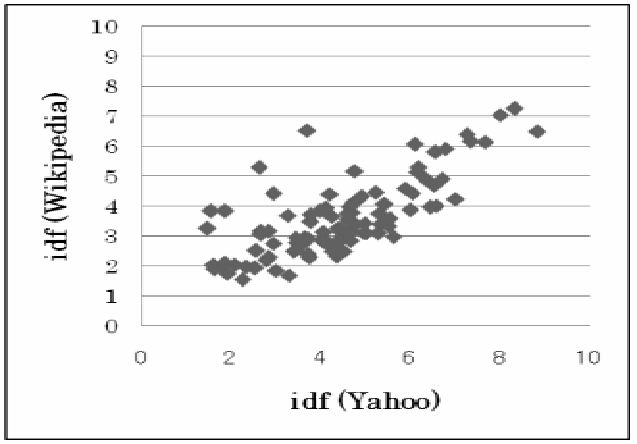

図 2：Yahoo から求めた idf と Wikipedia から求めた idf の散布図

3.3 結果と考察

Wikipedia の「サッカー」のページの頻出単語上位 100 個について，Yahoo 検索を利用して求めた idf と Wikipedia から求めた idf の散布図を図 2 に示す．

右上がりの直線に沿って点が集中している．これは Wikipedia の idf で大きい数値を示した単語は Yahoo の idf でも大きくなるということで，2 つの idf に強い正の相関があることを示している．

今回の実験の結果，全体の idf でその部分集合の idf を近似できることが分かったので，Wikipedia のような検索対象が明確に定義できる集合以外の部分集合に対しても，全体の idf が代用できる．

4. tf の検討

4.1 基本的な考え方

今ある Web ページの集合があって，限定された少数のページから構成されている状況を考える．例えばブログのサイトがこれにあたる．この中に複数の話題（テーマ）が設定されており，それらの話題を判定するためのキーワードを含まれる少数のページ中の単語から選ぶという問題を考える．しかし，従来からある単語の出現回数を数えるという方法で得られる tf では少数ページから計数するために単語に統計的に意味

のある頻度が得られないという問題が生じる. idf の場合はより大きなページ集合に対する統計によって近似がうまくいったが，tf の場合はその話題に対して関係の深いページ中で頻度が大きくなることがキーワードとして選択されるために重要であるから，ページ集合を大きくすることはできない.

そこで，単語の中から類似度の高い単語を見つけてそれらをグループ化することで tf を補完しようと考えた．そして，ある単語の tf を計数する際に，類似単語の tf も加算する.

本節では，少数ページからの tf の算出を対象単語そのものに加え，同じクラスタに属する単語の tf を加算する手法の有効性を検討する．明確にページ集合を決定できる話題を設定する必要があるが，Web ページそのものにはそのような情報は含まれていないので,「話題」集合を用いた検討は難しい．そこで，本報告では，明確にページ集合を決定できる基準として「Wikipedia」を利用することを考えた.

Wikipedia を構成する全ページをひとつの集合と考える. これは話題集合とは概念的に異なるが，明確にページ集合が規定されるという点では一致しており，そこを利用する．Wikipedia 中から特定の 1 ページのみを取り出し，これを限定ページと考える．限定ページに含まれる高頻出単語について，クラスタを考慮した tf を算出し，この値を Wikipedia 全体から算出されるその単語に対する真の tf と比較する．両者はスケール的に大きく異なっているが相対的に意味があると思われるので，散布図上の位置関係でその妥当性を評価する.

4.2 実験条件

4.2.1 単語のクラスタリング

単語のクラスタリングを行う際に，２つの単語の類似度を測る指標が必要になる．さまざまな指標があるが[3]，一般的によく用いられるのは文書における２つの単語の一文中での共起頻度を調べることである[4][5][6].

閲覧者が見やすいように箇条書きにしている部分があったり，句点のないものがあったりするので，多くの Web ページにおいて，一文を定義するというのは簡単ではない．さらに，文単位の共起を調べるにはある程度の文数が必要になってくるが，Web ページでは論文のように 1 つの文書に情報を集約するといった必要がなく，１ページが短く必要な文数を得ることができない場合も多い．そこで，本研究では共起の範囲を一文から一文書に拡張して考えることにした.

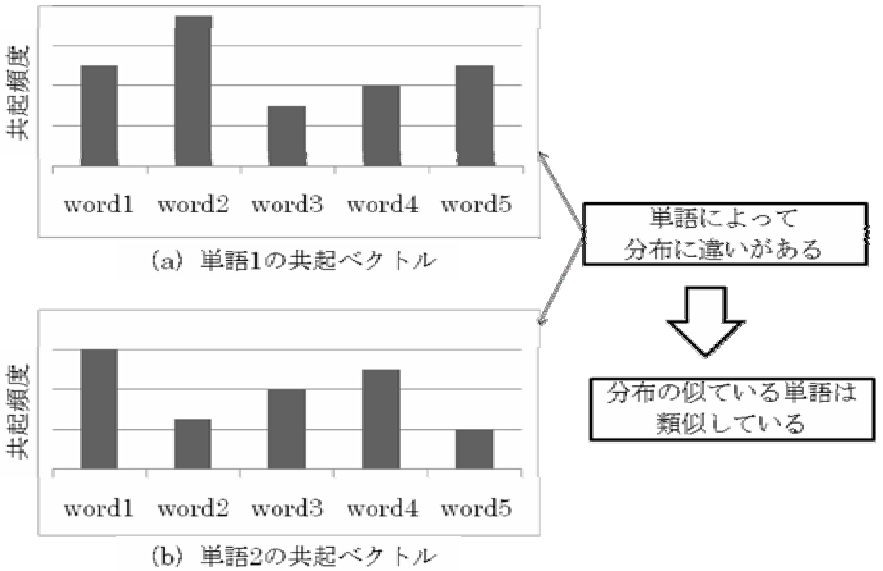

図 3：類似度計算のイメージ

本報告ではある単語が他の単語と Web 全ページ中で共起する頻度をベクトルで表し，この共起ベクトル間の内積によって単語同士をクラスタリングする.

例として，単語 1 と単語 2 の類似度計算のイメージを図 3 に示す．単語 1 と頻出単語 5 種に対し，各々Yahoo の「and 検索」を行い，図上段の共起ベクトルを求める．次に，単語 2 について同じく頻出単語 5 種との共起ベクトル（下段）を求め，両共起ベクトルの内積を計算する．分布の似ている単語はこの内積が大きくなる．この尺度に対して，後述する閾値を設定し，単語のクラスタリングを行う.

Wikipedia の『サッカー』というページを使って，クラスタリングを行う際の閾値の判定を行う．このページの単語の種類は 823 種あるが，Yahoo 検索 API のリクエストの上限回数による規制を受けて，全単語については検証不可能なので，頻出単語の上位 100 個までをデータとして用いる．使用する単語数と頻出単語として設定する数の間にある有効な比率と，クラスタリングを行うための閾値を調べる.

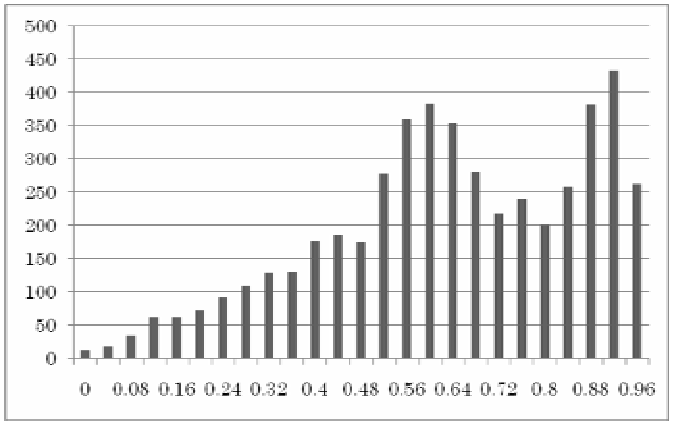

図 4：共起ベクトル間の内積のヒストグラム＜単語数:100 個，共起単語数:30 個＞

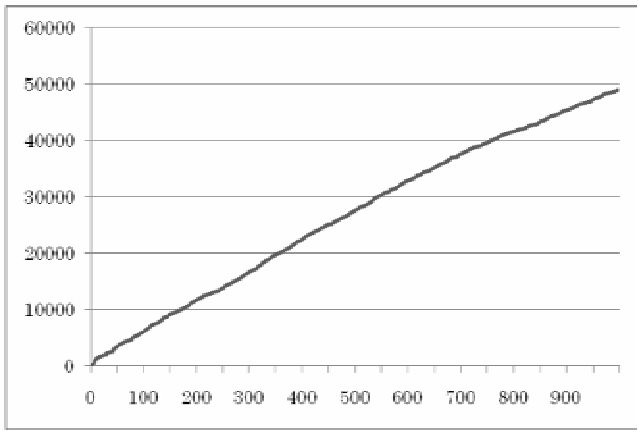

図 5：検索でヒットした URL1000 件に対する「サッカー」の tf の累積

本報告では単語数を 100 個，共起単語数を 30 個にした．100 個のベクトルすべての組み合わせについて内積を計算し，内積値(0～1)を 0.04 ごとの範囲に区切り，その範囲内に存在する組み合わせの数を数えてヒストグラムを作成した. 結果を図 4 に示す．

全体で見ると 0.8 付近で谷ができて，0.92 付近の類似度の高いかたまりが一つのクラスタになりそうに見える．このことから，単語数と頻出単語の割合を 10：3 と設定し，クラスタリングを行う閾値を 0.8 とする．

4.2.2 Wikipedia 全体に対する tf の推定

Wikipedia の全ページにおける単語の出現回数を測定することで，Wikipedia という集合における tf の真値を求めることができる．これは先ほどの手法でクラスタリングを行った tf の値を評価する際に用いる．

Wikipedia は idf を計算するときのようにヒット数ならば参照可能なのだが，tf の場合はヒットしたすべてのページの URL を集めて，その URL の全ページで出現回数を数えなければいけない．しかし，一般の人が検索エンジンを使う時にはヒットしたうちのせいぜい 50 件程度しか見ない上に, 検索結果が多いとサーバーへの負担も大きくなるので，1000 件までしか URL を取得できない仕様になっている．

そこで，「サッカー」という単語を使って，取得できる 1000 件のデータから残りの部分を類推し，tf を外挿することを検討した．1000 件の URL を横軸にとり，各ページでの tf を累積させていった時のグラフを図 5 に示す．グラフを見ると累積値が直線的に増加している．このことから，1000 件を超えるヒット数になっていても，1000 件の tf を数えて，そのまま 1000 とヒット数の比率に従って tf を外挿する．

4.3 結果と考察

クラスタリングによるグループベース tf の値が相対的な意味で，Wikipedia 全体から求める tf とうまく対応するかどうか検討する．本来 tf の目的は出現頻度を指標にして多く出現した単語に重みを大きくするというものなので，単語の大小関係が重要である．そこで，①限定ページ単語 tf，②限定ページグループ tf，③Wikipedia 全体の単語 tf の 3 つの値を使って，相関を調べることで評価する．

10 個の単語について，①と③，②と③の 2 つの組み合わせを散布図で表示するとそれぞれ図 6，図 7 のようになった．図 6 では左下に点が集まっていて単語間の大小関係がはっきりしていないが，図 7 では右へ行く（tf が大きくなる）ほど点が上に向かっている（真値の tf も大きくなっている）のがわかる．

また，いずれの図においても，「はずれ値」が存在していたので，これは取り除いておく．残りの 9 個の単語について，同じように 2 種類の組み合わせで相関係数を調べると結果は表 1 のようになった．

表 1：相関係数の比較

	Wikipedia 全体の単語 tf
単語 tf	0.5765
グループ tf	0.8905

数値上でも，相関のなかった単語群が本手法を適用することによって強い相関が見られるようになっているのが分かる．このことから，単語の出現頻度が少ないページ集合に対して，今回用いたクラスタリングによる，類似単語の tf を加算するというグループベースの tf は有効である．

5. おわりに

今回はブログに代表されるような限定されたページ中から話題判定に有効なキーワードを選択する手法を探るために tfidf をベースに，idf の近似とグループ tf をそれぞれ個別に検討した．

idf は，Wikipedia と検索エンジン，2 種類の idf を比較すると，相関を持つということが分かったので，限定されたページ集合の idf はより一般的なページ集合の idf で近似できることが分かった．

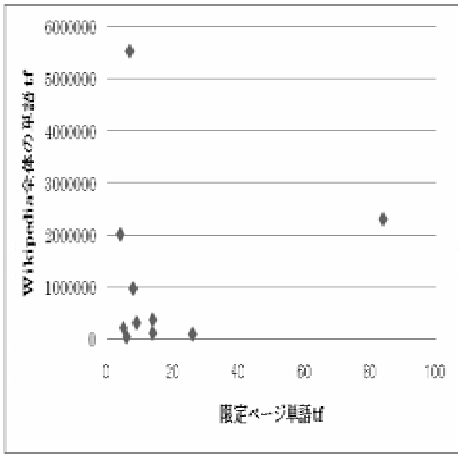

図 6：限定ページ単語 tf と
Wikipedia 全体の単語 tf の散布図

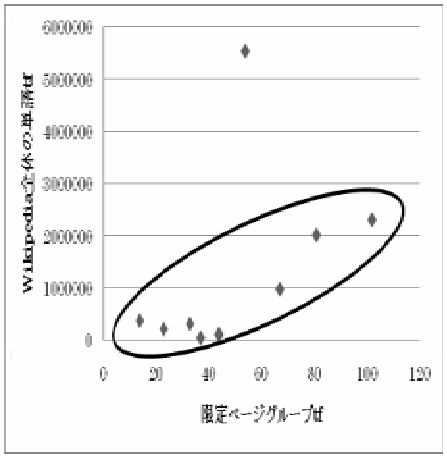

図 7：限定ページグループ tf と
Wikipedia 全体の単語 tf の散布図

一方，tf は類似の単語をクラスタリングするグループベースの tf に拡張することを検討した．本研究ではクラスタリングを行う際に，「検索エンジンでの 2 単語の and 検索のヒット数」という新しい指標によって単語の類似度を測り，共起ベクトルの内積を計算することで類似単語を検出した．そして，クラスタリングによって得られるグループ tf の有効性を示した．

今回 tf を検証する際に，「話題」集合として，明確に範囲を定義できる Wikipedia という集合を設定したが，本来の方針に沿うように，話題集合をうまく設定した時のクラスタリングによる tf の補完の有効性も検証していきたい．その後，tf と idf を組み合わせる tfidf の有効性を検討していく．

参考文献

1) 徳永健伸：情報検索と言語処理，東京大学出版会
2) 松尾豊，石塚満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人口知能学会論文誌，17 巻 3 号 a，2002 年
3) 相澤彰子：共起に基づく類似尺度，日本オペレーションズ・リサーチ学会　2007 年 11 月号
4) 大澤幸生，ネルス E.ベンソン，谷内田正彦：語の共起グラフの分割・統合によるキーワード抽出，電子情報通信学会音声論文誌，Vol.J82-D-I(2)，pp.391-400，19990225
5) 松尾豊，大澤幸生，石塚満：Small World 構造に基づく文書からのキーワード抽出，情報処理学会論文誌，Vol.43 No.6，2002/06
6) 松村真宏，大澤幸生，石塚満：語の活性度に基づくキーワード抽出法，人工知能学会論文誌，17 巻 4 号 F(2002)