ソーシャルブックマークを使った関連 Web ページの検索の研究

青野雅樹 (第４工学系, セマンティックアーカイブコア)

1　はじめに

われわれは、MSRC の中のセマンティックアーカイブコアの属し、主としてメディア（テキスト、画像、三次元モデル形状など）の概念的な類似度、H18 年度は主として三次元モデルの形状類似検索を行ったが、H19 年度では、Web 上のテキストデータに関して、集団が「タグ」として与えたソーシャルブックマークを使った関連 Web ページ検索を行ったのでそれを報告する。

ある Web ページに興味を持ったとき，その Web ページと関連する Web ページを探したいという要求がある．このような場合，一般的な検索エンジンでは，ユーザは，目的とする Web ページに含まれていると思われる単語やフレーズを推測し，検索キーワードを入力しなければならない．これはユーザにとって負担となるため，Web ページをそのままクエリとして利用し，それに関連したページを検索する手法が求められている．本稿では，近年注目を集めているソーシャルブックマーク（SBM）を利用し，関連ページを検索する手法を提案する．提案手法では，同じユーザは同じタグを，関連するページ群に付与する傾向があることを利用する．この手法を実装し，評価実験により有効性を検証したので，これを報告する．

2　ユーザのタグ付けの傾向分析

2.1　ユーザ間で異なるタグ付与行動

ソーシャルブックマーク（SBM）では，ユーザは Web ページに対して自由にタグ付けできる．そのため，ユーザの嗜好によって，タグの付け方が異なる．例として，ある「Web デザイン」に関する Web ページに対して付与されたタグの上位 10 種類を表 1 に示す．表 1 を見ると，ユーザによって“ web デザイン ”や“ webdesign ”といった，同じ意味でも表記の異なるタグを付与していることがわかる．さらには“ *web デザイン ”や“ *webdesign ”といった，先頭にアスタリスクのような記号を付けたタグもある．このような先頭に記号や数字を付けたタグは，ユーザ自身が良く使うタグを探しやすくするために，しばしば使われる．また，言葉の多義性やユーザの嗜好により，同じタグでもユーザによって意味合いが異なることがある [9]．例えばタグ“ library ”は，あるユーザにとっては「プログラミングにおけるライブラリ」であるが，別のユーザにとっては「図書館」を意味する．

Table 1　「Web デザイン」に関するページに付けられたタグ上位 10 種類

タグ名	付与ユーザ数
web デザイン	40
css	39
まとめ	37
webdesign	32
デザイン	24
design	19
*web デザイン	17
web	12
リンク集	12
*webdesign	10

また，タグ“ word ”は，ユーザによって「名言」を意味したり,「Microsoft Word」のことを指したりする．

2.2　同じユーザの同じタグのブックマーク内容

表 1 の「Web デザイン」に関するページには,“ web ”や“ まとめ ”といった，広い意味を持つタグが付与されている．これらのタグは，このページ以外にも，他の大勢のユーザが様々なページに付与している．単純にこれらのタグが付与されたページを探すと,「Web デザイン」に関係のないページも数多く見つかると考える．ここで，表 1 の「Web デザイン」に関するページに対しタグ“ web ”を付与したユーザ 12 人が，他にタグ“ web ”を付与したページを調べたところ，12 名中「Web デザイン」に関するページのみのユーザ数が 6 名であり、この 6 名だけが、タグ“ web ”を付与していた．比較のため，タグ“ web ”をよく使用している別のユーザ 10 人について調べたところ，16 名中、「Web デザイン」に関するページにタグ“ web ”で管理，分類しているユーザはいなかった．このように，タグ“ web ”のような抽象度の高い曖昧なタグであっても，同じユーザが同じタグを付与したページ群を探すことにより，関連ページを見つけられると考える．

3 ユーザとタグのペアの共起を用いた関連ページ検索手法

まず，ユーザ U にとってのページ P1 と P2 の関連度を以下の式で定義する．

$$J(U,P_1,P_2)=\frac{|Tags(U,P_1)\cap Tags(U,P_2)|}{|Tags(U,P_1)\cup Tags(U,P_2)|}$$

ここで，$Tags(U,P)$ は，ユーザ U がページ P に付与したタグの集合を表す．上式は，ユーザ U がページ P_1 および P_2 に付与したタグの集合間の Jaccard 係数である．一方，ページ P1 と P2 の関連度を表すスコア $R(P_1,P_2)$ を以下のように与える．

$$R(P_1,P_2)=\frac{\sum_{con1}J(U_i,P_1,P_2)}{|TaggedUsers(P_1)\cup TaggedUsers(P_2)|}$$

$$con1=U_i\in\{TaggedUsers(P_1)\cup TaggedUsers(P_2)\}$$

ここで，$TaggedUsers(P)$ は，ページ P にタグを付与したユーザの集合を表す．この式の値としては，0 以上 1 以下の値をとる．この値が高いほど，ページ P_1 との関連度が高いページと解釈する．P_1 にタグを付与しているユーザ群が，P_2 にも全く同じようにタグを付与している場合，1 となる．いま，大勢のユーザがページ P_1 と P_2 をブックマークしている場合を考える．このとき，付与されたタグが大きく異なっていたとすると，ページ間の関連度は低いと考えられる．そこで，ページ P_1 と P_2 の両方にタグを付与したユーザ群のタグの共起率の平均値を以下の式により求め，

$$M(P_1,P_2)=\frac{\sum_{con2}J(U_i,P_1,P_2)}{|TaggedUsers(P_1)\cap TaggedUsers(P_2)|}$$

$$con2=U_i\in\{TaggedUsers(P_1)\cap TaggedUsers(P_2)\}$$

この値が閾値 M_{theta} より低い場合は，$R=0$ とし，関連ページ群から除外することとした．ただし，M_{theta} の値を高く設定すると，関連が強いページも除去される．事前に M_{theta} を変化させた実験を行ったところ，0.2〜0.4 近辺で概ね良い結果が得られた．

4 評価実験

4.1 データ収集

我々は，はてなブックマークからデータを収集した．はてなブックマークは 2007 年 12 月時点で，日本で最大規模の SBM サービスであり，10 万人以上のユーザが利用している．我々は，約 5 万人のユーザのデータと，それらユーザによってブックマークされたページ約 150 万ページを収集した．さらに，収集したユーザによってそれらのページに付与されたタグ約 20 万種類を収集した．

4.2 クエリページ

本研究では，クエリページのブックマークユーザの人数が，検索結果に影響を与えると考えられる．そのため，クエリページを選ぶ際に，ブックマークユーザの人数に応じて，以下のクラス分けを行った．

A ブックマーク人数が 30 人未満

B ブックマーク人数が 30 人以上 100 人未満

C ブックマーク人数が 100 人以上 500 人未満

D ブックマーク人数が 500 人以上

そして，それぞれのクラスからクエリページを 10 ページずつ設定した．

4.3 比較手法

閾値 $M_{theta}=1/3$ と設定した提案手法と，$M_{theta}=0$ と設定した提案手法（すなわち，式の閾値による除去を行わない場合）を用意した．また，比較手法として以下の 2 つの手法を用意し，実験を行った．

4.3.1 タグのベクトルの類似度に基づく手法 (タグ手法)

この手法では，ページに付与されたタグを使用し，ページに対してタグの特徴ベクトルを計算し，ページ間の関連度を式 (4) のコサイン類似度で求める．

$$R_{Tags}(P_1,P_2)=\frac{V_{Tags}(P_1)\cdot V_{Tags}(P_2)}{|V_{Tags}(P_1)|\,|V_{Tags}(P_2)|}$$

ただし，

$$V_{Tags}(P)=\{rel(P,T_1),rel(P,T_2),\cdots,rel(P,T_n)\}$$

$$rel(P,T)=TF(P,T)\times IDF(T)$$

$$TF(P,T)=\frac{w(P,T)}{\sum_{T_i\in TAGS}w(P,T_i)}$$

$$IDF(T)=\log\frac{\sum_{P_j\in PAGES}\sum_{T_i\in TAGS}w(P_j,T_i)}{\sum_{P_j\in PAGES}w(P_j,T)}$$

である．ここで，$w(P,T)$ はページ P に付与されたタグ T の数，$TAGS$ は全てのタグの集合，$PAGES$ は全てのページの集合を表す．$rel(P,T)$ の式は，文献 [1] を参考にした．

4.3.2 ユーザの共起を用いた手法 (ユーザ手法)

この手法では，ページをブックマークしたユーザを用いて，ページ間の関連度を式 (5) の Jaccard 係数で計算する．

$$R_{Users}(P_1,P_2)=\frac{|Users(P_1)\cap Users(P_2)|}{|Users(P_1)\cup Users(P_2)|}$$

ここで，$Users(P)$ は，ページ P をブックマークしたユーザの集合を表す．

4.4 評価尺度

情報検索における評価尺度として，精度と再現率がよく用いられる．このうち再現率は，検索対象のデータ中に適合ページがいくつあるのかが分かっていなければならない．本研究における検索対象は，SBM から収集した約 150 万ページであるため，再現率を求めるためには，150 万ページ中に適合ページがいくつあるかを確認する必要があり，これは困難である．また通常，ユーザは検索結果の上位から見ていくため，適合ページがより上位に検索されることが望まれる．そのため評価尺度としては，上位ページほど重みが大きくなる DCG (Discounted Cumulative Gain)[2] を用いた．

$$DCG_i = \begin{cases} G_1 & if\ i = 1 \\ DCG_{i-1} + \frac{G_i}{\log_2 i} & otherwise \end{cases}$$

ここで G の値は，各順位の検索結果の適合度の高さによって，多値を取ることができる．今回は，検索結果の Web ページ上位 20 件に対し，適合，不適合の判定を，3 名の評価者に行ってもらった．判定基準は以下の通りである．

- 適合　…クエリページと関連がある
- 不適合…クエリページと関連が見られない

または，ページにアクセスできないそして，適合と判定した評価者の人数を G の値とした．

4.5 考察

提案手法は，タグ手法，ユーザ手法と比べ，比較的，ブックマークユーザの人数に影響を受けず，安定して関連ページを検索することができていることがわかった．クエリページのブックマーク人数が少ない場合では，提案手法はタグ手法と比べ，良い結果となった．また，クエリページのブックマーク人数が多い場合では，提案手法はユーザ手法と比べ，良い結果となった．タグ手法にはやや劣っているが，M の閾値 $M_{theta} = 1/3$ と設定することにより，結果を改善できることがわかった．ブックマーク人数が多いページをクエリにした場合，閾値を設定しない提案手法（$M_theta = 0$）は，複数の評価者から関連していないと判定されたページが上位に来た．しかし，$M_{theta} = 1/3$ と設定することにより，関連していないページを除去することができた．ブックマーク人数が少ないページをクエリにした場合は，提案手法はユーザ手法と比べ，検索結果上位の精度が低くなっている．これは，関連のあるページ群をブックマークしているユーザが，それらのページ群にタグを付与していないことがあったため，提案手法では，関連が弱いと解釈されたことが原因である．ブックマーク人数が多いページをクエリにした場合は，提案手法はタグ手法に比べ，精度が劣っている．これは，クエリページのブックマーク人数が多い場合には，それぞれのユーザが複数の関連ページに対し，同じようにタグを付与していることが少ないためと考える．SBM では，同じ意味で表記の異なるタグが複数ユーザ間で付与されることがある [3]．しかし提案手法では，[3] と同様に，同じユーザの同じタグが付与されたページの集合を利用するため，ユーザ間のタグ名の揺らぎは問題にならない．

5 まとめ

SBM のユーザとタグの共起情報を利用し，関連ページを検索する手法を提案した．実験の結果，提案手法はタグのベクトルやユーザの共起を用いた手法と比べ，検索精度へのブックマークユーザの人数の影響が少ないことがわかり，有効性が確認できた．

参考文献

[1] 丹羽智史, 土肥拓生, 本位田真一, “Folksonomy マイニングに基づく Web ページ推薦システム,” 情報処理学会論文誌, Vol.47, No.5, pp.1382-1392, 2006.

[2] K. Jarvelin and J. Kekalainen, “IR Evaluation Methods for Retrieving Highly Relevant Documents,” Proc. of the 23rd Annual International ACM SIGIR Conference (SIGIR 2000), pp.41-48, July 2000.

[3] 佐々木祥, 宮田高道, 稲積泰宏, 小林亜樹, 酒井善則, “ Social Bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム, ” 情報処理学会論文誌:データベース, Vol.48, No.SIG20(TOD36), pp.14-27, 2007.

発表論文

[1] 杉山典之, 関洋平, 青野雅樹 “ユーザのタグ付けの傾向を利用したソーシャルブックマーク内の関連ページ検索手法”, Journal of the DBSJ, Vol.7, N0.1, pp.239-244, 2008