タグの共起関係を利用した類似ソーシャルブックマーク抽出システムの開発

池田 善昭† 木村 昌臣†
芝浦工業大学情報工学科†

1. 研究背景，目的

近年，インターネット上では情報量が増加し続けており，求める情報を見つけ出すまでに時間がかかることが多くなっている．従来の検索手順はまず検索単語を考えて入力し，検索する．次に検索結果一覧から選択し，内容吟味をする．最後に再検索を行うもしくは終了という流れになっている．この手順の中で時間がかかるのは再検索に関る部分であると考えられる．この要因を解決するために閲覧中のページに関連ページ情報を自動的に推薦することによりユーザーの再検索の必要回数を削減させる仕組みが必要である．既存の研究 [1] ではオンライン上で公開，共有されたブックマークの集合であるソーシャルブックマーク（以下 SBM）に着目して TF-IDF スコアからページ間類似度を算出する仕組みを提案している．しかし，コンテンツ以外のサイト共通記述部によって類似度スコアが高くなる，単語の表記揺れに対応するために辞書が必要であり新語や略語に対応しづらい等の問題がある．SBM においてユーザーはブックマークする際にページの内容を特徴付けるタグを付ける．本研究ではタグに着目することで上記の問題を解決したシステムの作成を目的とする．

2. 提案手法

本システムは代表的な SBM であるはてなブックマーク [2] を対象として開発をする．まず SBM からブックマークされたページごとに URL，タイトル，タグ(図１)を取得する．例えば図１においてタグの ruby と Programing は共起しているといい、各ページに同時に出現しているタグの共起関係に着目する．さらに他のページも合わせて共起したタグすべての組み合わせについて共起した回数を求める．共起しているタグは関連性があり，関連性の高いタグは共起数が大きくなると予想できることから，共起数の逆数をタグ間の距離としてネットワークを構成し，最短経路アルゴリズムのダイクストラ法を用いて各タグ間における最短距離を求める．

Related contents extraction system using tags collocational relationship
†Ikeda Yoshiaki, Masaomi Kimura
†Shibaura Institute of Technology

図１．ブックマークされたページのタグ

図２は実際にタグ「ruby」で検索した結果 50 件からタグの起関係を求め，距離が近い単語間ほど太いエッジが発生するように Kamada-Kawai 法を用いて Pajek [3] で可視化したものである．図から ruby と rails が非常に関連度の高い単語であると判断でき，さらに rails の同義語である ruby on rails も関連語として判断できる．これによりあらかじめ単語の類義語辞書を用意しなくても共起関係を見ることによって関連単語の自動判別が期待できる．

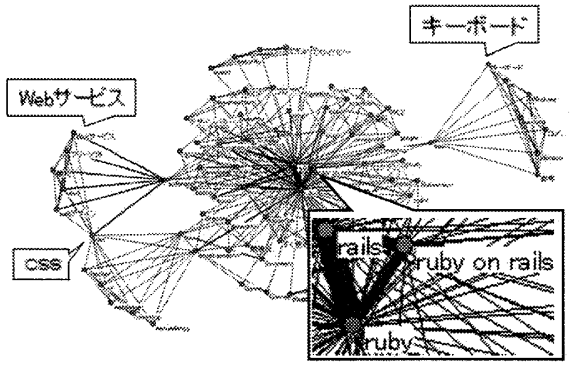

図２．共起関係を可視化した図

次に，求めたタグ間距離を元にページごとに付いている各タグの距離の平均を求めてページ間距離とする．しかし，単に平均を取ると距離が非常に大きくなるタグの組み合わせが１つでも含まれると計算結果に与える影響が大きくなる．これは記事に対して関係が乏しいタグが１つでも付いていた場合に他のタグが複数一致していても関連度が低くなってしまうことを指す．ユーザーがタグを付加することから精度の良いタグだけが付けられるとは限らないため，関係のないタグが計算に与える影響を小さくする必要がある．関連性の低いタグがページ間距離に与える影響を減らすために調和平均を利用する．その際 n は比較するタグのパターン数，x_i はタグ間距離とする．ダイクストラ法において同一ノードまでの距離は 0 であり辿れないノードま

での距離は∞とされているが，調和平均の式に当てはめた場合に数学的に計算できない値となることがあることから同一ノードまでの距離を微小な値とし，辿れないノードを非常に大きな値とした．各ページ間に調和平均の式の適用を繰り返し，収集したすべてのページ間距離を求め，収集した各ページにおいて関連度の高い上位５つのページを算出する．

しかし実際のデータで計算をした結果，多くのページにおいてタグ数が少ないページほど関連度が高くなる傾向があったことから，比較するページのタグで多い方の値を t として割ることによって，多くのタグが関連するほど距離が短くなるようにした結果，ページ間距離Mを求める式は図３のようになった．

$$M = \frac{n}{t\sum_{i=1}^{n} x_i^{-1}}$$

図３．ページ間距離算出式

図３の式で計算した結果，タグ数が同じものであると関連度が高くなる傾向が現れた．そこでタグ数の差があった場合は，タグ数が少ないページに対してどのタグからも距離が１となる仮想ノードを配置し計算することとした．

最後に出力部分を用意した．閲覧中のページ URL を取得してデータベースから検索し，該当するページが存在した場合に関連度の高い上位５ページを閲覧中のページ上部に表示する処理を行うようにした結果が図４である．

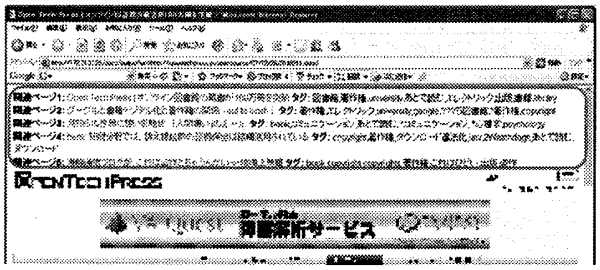

図４．類似 SBM 抽出システム

3. 実験

本研究では関連性を調べるページとして，はてなブックマークから無作為に選んだタグで検索してヒットした結果を用いる．被験者に収集したページからサンプリングした各ページに対する５つの推薦ページのタグや内容を見てもらいページの関連度をそれぞれ評価してもらい，関連性があるページが推薦される精度を調べる．また，アンケートには関連ページではないと判断した際に何を見て判断したかを書く記述欄を設ける．

表１ 分析対象データ

対象 SBM	はてなブックマーク
検索タグ	ruby, tool, web サービス， java, 社会, 猫, programming, 著作権
ブックマーク	各タグ 125 件ずつ計 1000 ページ
評価ページ	サンプル４ページ 推薦ページ合計２０ページ
評価人数	１２人

4. 結果・考察

推薦されたページごとに関連性が高いと感じたページ，関連性があると感じたページ，関連性がないと感じたページの３つから選択する形式のアンケートを集計した結果，関連性があるページが推薦される割合は 82.08%（関連度が高いページが推薦された割合は 60.83%，関連性があるとされたページは 21.25%）という結果が出た．関連性がないと回答されたページに対しての記述の多くから，web サービスであるという観点からは関連性があったが実際は違う web サービスについて書かれたページであったことから関連性が低いと判断した等，関連性の抽象度が高いことに端を発する問題が発生していることが分かった．

5. 今後の課題

タグの中でキーワードの重要度を判別して重み付けをし，抽象度が高いっ場合の問題を解決する方法を確立することや，前処理としてページを収集した後のタグの変更や，新しいページの追加などの更新に対する処理を実装する必要がある．

6. まとめ

本研究ではソーシャルブックマークにおけるタグの共起関係から類似，関連ページを推薦する仕組みと，実際にブラウジングできるシステムの開発を行った．これによりインターネットユーザーが関連ページを探す手間を助力し，関連単語の想起をさせることが可能になると考えられる．

参考文献

1) 矢島健太郎, 井上潮：ソーシャルブックマークにおける文書解析を利用した類似文書および類似ユーザの推薦方法の提案, 電子情報通信学会第18回データ工学ワークショップ論文集, C9-Web 情報検索2, (2007)

2) はてなブックマーク, http://b.hatena.ne.jp

3) Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek/