A-018

2 部グラフによる電子掲示板のスレッド分類手法
Topic Categorization with Bipartite Graph in Bulletin Board System

田中 航介† 鈴木 育男‡ 山本 雅人† 古川 正志††
Kousuke TANAKA Ikuo SUZUKI Masahito YAMAMOTO Masashi Furukawa

1. はじめに

電子掲示板は Web 上でユーザ同士が情報交換を行う場として，インターネットの普及と共に急速な発展を遂げ，日本では話題とそれに対する投稿の纏まりであるスレッドを単位とする，スレッドフロート型掲示板が主流となった．ユーザはスレッド一覧から目的の情報を発見するために，キーワード一致による検索や，電子掲示板やブラウザ側の提供するシステム等を利用しスレッドの探索を行うが，これらの手法の多くはスレッドのフィルタリングを主軸に置くものであり[1]，スレッド一覧そのものを俯瞰する手法は少ない．

本研究ではスレッド一覧の俯瞰によるユーザのスレッド探索の補助を目的とし，そのためにスレッド，及びスレッドより抽出されるキーワードをノードとする 2 部グラフを構築、スレッドの分類を行う手法を提案する．

2. スレッドからのキーワード抽出

2 部グラフの構築にあたり，各々のスレッドからキーワードの抽出を行う．抽出するキーワードとして，大半のユーザがはじめに目にする情報であるスレッド名，またはスレッド自身の特徴を最も表すスレッド内の投稿本文を用い，それぞれに対し MeCab[2]により形態素解析を行う．その後，得られた単語に対してそれぞれ重要度を算出，キーワードノードとして抽出する単語を決定する．形態素解析において，今回は名詞である単語のみを抽出の対象とする．

2.1 単語の重要度の算出

形態素解析により得られた単語について，スレッド名とスレッド本文でそれぞれ重要度を算出する．スレッド名は基本的にそのスレッドの方向性を端的に特徴付けるものであり，文字数が限られている．そのため使用される単語数が少なく，単語の重要度を算出するにあたり，そのスレッドにおける単語の出現回数である単語頻度(tf 値)を用いる手法では重要度の差がはっきりと現れない．したがって，スレッド一覧において出現頻度の高い単語を重要度の高い単語とする．具体的には文書頻度(df 値)を用い，全 N 個のスレッド名より抽出された i 個の単語 t_i に対し，その単語の出現するスレッドの数 $df(t_i)$から，式(1)をスレッド名における単語の重要度 $W(t_i)$と定義する．

$$W(t_i) = \frac{df(t_i)}{N} \quad (1)$$

スレッド本文はユーザの求める情報そのものであり，文字数もスレッド名と比較して最大で何千倍もの差が存在する．故に使用される単語数が非常に多く，頻出語よりもそのスレッドを特徴付ける単語の重要度が高く算出

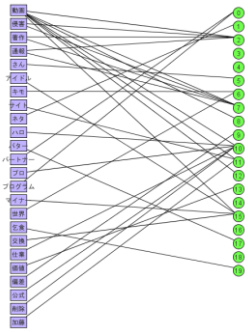

図1　構築される 2 部グラフの例(左:キーワードノード，右:スレッドノード)

されるべきである．したがって今回は tf-idf 法[3]を用い，全 N 個のスレッドの本文より抽出された i 個の単語 t_i の出現頻度 $tf(t_i)$，及び $df(t_i)$の逆数の対数を取った逆文書頻度 $idf(t_i)$を用いて，それぞれのスレッド，及びスレッド全体における各単語の $tfidf(t_i)$を式(2)のように求め，これをスレッド本文における重要度 $W(t_i)$と定義する．

$$tfidf(t_i) = tf(t_i) \times idf(t_i) = tf(t_i) \log \frac{N}{df(t_i)} \quad (2)$$

2.2 抽出語の絞り込み

2.1 節でそれぞれ抽出された単語及びその重要度について，そのまま全てを 2 部グラフにキーワードノードとして適用を行うとその数は膨大な量となる．故に単語の出現回数，及び重要度に閾値を設け，その個数を絞り込む．スレッド名の場合，全体における単語の出現回数がそのまま重要度に結び付くため，出現回数による閾値は設けず，重要度の上位 X%の単語のみを抽出語の対象とする．スレッド本文の場合，出現回数の少な過ぎる，明らかに重要でない単語や逆に多過ぎる，言わばありふれた単語が重要度に影響を与えるため，これらを出現回数，及び重要度上位による閾値によって絞り込む．

3. 2 部グラフの構築

スレッドノードの集合を V_t，抽出されたキーワードノードの集合を V_k，両集合間に結ばれる重み付きリンクの集合を E とするとき，図 1 のような 2 部グラフ G は G={ V_t, V_k, E}により表される．リンクの生成にあたり，V_t，V_k は同集合内でリンクを決して結ばず，またリンクが一切結ばれないノードを V_t ではそのまま残し，V_k では削除する．リンクの重みは V_t と V_k の関連性の強さを表し，V_k をスレッド名から抽出した場合は V_t に V_t の単語が含まれているとき，V_t の重要度をそのまま重み付きリンクとして生成する．スレッド本文から抽出した場合は V_t と V_t の重要度の積を重み付きリンクとして生成する．

† 北海道大学　大学院情報科学研究科
‡ 北見工業大学　情報システム工学科
†† 北海道情報大学

4. 2 部グラフの構築実験

実際のスレッドを用いて，提案手法の評価をスレッド名を用いた場合と本文を用いた場合でそれぞれ行う．

4.1 実験の設定

実験に使用するスレッドとして，2 ちゃんねる[4]の youtube 板より表 1 に示すスレッドを取得する．取得するスレッドの条件としてスレッド名に”youtube”，”ニコニコ動画”，”ニコニコ生放送”のいずれかが含まれるものを選択した．これらのグループはいずれもキーワードノードを介して同グループ内全てのスレッドノードの組み合わせでパスが存在する状態を理想とする．また，閾値として，スレッド名を用いる場合は重要度の上位 10%をキーワードノードとして採用，スレッド本文を用いる場合は出現回数が 5 回未満 1000 回以上の単語を排除，重要度は得られる単語数の差を考慮し，上位 0.1%とする．

表 1　実験に用いるスレッド

スレッド名に含まれる語	スレッド数
youtube	33
ニコニコ動画	38
ニコニコ生放送	27

4.2 実験結果・考察

前節の条件で 2 部グラフの構築を行い，それぞれのグループにおけるパスの構築率を算出した結果，表 2,3 及び図 2,3 のようになった．図は枠に囲まれたものがキーワードノード，番号を囲む円がスレッドノードを表し，番号による色分けとグループが対応している．結果に対する考察は以下のようになる．

- スレッド名の場合では単純に出現頻度の高いキーワードが重要度の上位となる関係で，各グループに含まれる語を中心とするクラスタが形成されるのが確認できた．一方で今回の結果では英文に見られる大文字小文字等の表記揺れが含まれており，その判別，修正が必要であると判明した．
- スレッド本文の場合では”ニコニコ生放送”のグループにおいて，スレッド名のときよりも高いパス構築率が確認された．この理由としては，グループ内で頻出かつ独自性・専門性の高い単語が多く，それによりとりわけ高い重要度を持つ単語を複数抽出できたためと考えられる．

表 2　実験結果(スレッド名，キーワード数 19)

	youtube	ニコニコ動画	ニコニコ生放送
孤立スレッド数	2	7	9
リンク数	88	73	51
パス構築率	88.1%	66.0%	51.0%

表 3　実験結果(スレッド本文，キーワード数 42)

	youtube	ニコニコ動画	ニコニコ生放送
孤立スレッド数	10	14	5
リンク数	101	89	99
パス構築率	47.9%	39.3%	77.0%

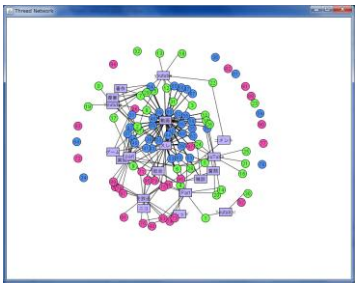

図 2　2 部グラフ構築結果(スレッド名)

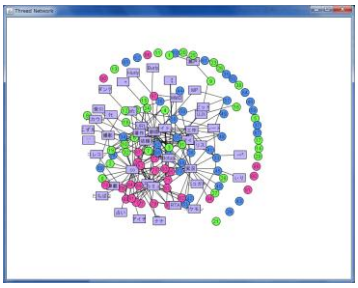

図 3　2 部グラフ構築結果(スレッド本文)

5. おわりに

本研究ではスレッドフロート型掲示板のスレッドを，スレッド名及び本文から重要語を抽出，スレッドとキーワードによる 2 部グラフの構築手法を提案した．

実験では理想解を設定した上で，スレッドに対し提案手法を適用し前述の結果が得られたが，実際のスレッドは実験で使用したものよりも遥かに多く，名称・内容共にその傾向も千差万別である．また，2 ちゃんねるのような掲示板は独自の表現やスラング，表記揺れ等が多く，今回使用した辞書ではそれをカバー出来ていない．したがって実験結果について，その精度は保証が難しい．

今後の課題としては，形態素解析周りの前述する問題点への対処や 2 部グラフの表示の改善，及び構築した 2 部グラフに対するコミュニティ抽出によるスレッドの解析などが挙げられる．

参考文献

[1]深谷 雅志, 倉本 到, 渋谷 雄, 辻野 嘉宏, “電子掲示板における行動履歴を用いたユーザ にとって興味あるスレッドの推薦手法”, 電子情報通信学会技術研究報告, Vol.106, No.410, pp.149-154(2006).
[2]MeCab (和布蕪), http://mecab.sourceforge.net/
[3]Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval* ., McGraw-Hill Book Company (1984).
[4]2 ちゃんねる，http://www.2ch.net/