1F1-01

コミュニティウェブにおけるアクティブ情報検索ためのトピック抽出

Topic Distillation for Active Information Retrieval in Community Webs

余 東明　　石川 孝
Dongming Yu　　Takashi Ishikawa

日本工業大学
Nippon Institute of Technology

In this paper, we describe the design and implementation of topic distillation from communication logs (e.g., messages written in the BBS or email, etc.) in community webs, which not only distills the topics after clustering messages, but also supplies keywords for searching the related information (e.g., news, links, data, etc.) from World Wide Web. Topic distillation is implemented over three phases. First, the messages are converted into the feature vectors comprised of the weight of nouns, verbs, adjectives and adverbs. Then the messages are classified by clustering based on the feature vectors. Finally, the topics of each cluster and the keywords for active information retrieval are distilled. Here we also show the results of a topic distillation experiment made on a real community site in yahoo! egroups, and extract the problems found by the experiment.

1. はじめに

今日のインターネットにおいて電子メールや電子掲示板などの通信手段はすでに不可欠な存在である. 電子メールや電子掲示板などの通信手段は, 時間と空間の制約を越え, 人々のコミュニケーションに多大な利便性をもたらす. 最近では, インターネット人口の急増にしたがって, メーリングリストや電子掲示板などを中心サービスとして, 同じ興味や目標を持っている人たちにコミュニケーションの場を提供するコミュニティウェブも増え続けている. その中にはインターネットプロバイダが運営するコミュニティサイト[egroups]や実社会の共同体が運営するウェブサイト[学会 NET]などが多数存在しているが, 一般的にはメッセージ(メーリングリストへのメールや電子掲示板への投稿)の追加機能と表示機能しか提供していない. コミュニティウェブにおいては, 毎日, 数十件, 数百件のメッセージが増えていて, ユーザはその内容に対する把握が難しくなっている. 一方, コミュニティウェブの管理者も, 一般的にはウェブの運営者ではないので, コミュニティウェブに有用な機能を追加したり, ウェブコンテンツを適切に更新したりすることはほとんど不可能である. そこで, コミュニティウェブを支援するソフトウェアが必要になる.

本研究の目的はコミュニティウェブの利用と管理を支援するソフトウェブの開発[石川 2003]である. 具体的には, コミュニティウェブにおいて, まず, 膨大なコミュニティログから自動的にトピックを抽出してユーザに提供する. 次に, 抽出されたトピックを使って, WWW から関連情報(ニュース, リンク, 資料)をアクティブに検索してユーザに表示することである. このソフトウェアの仕組みを図 1 に示す.

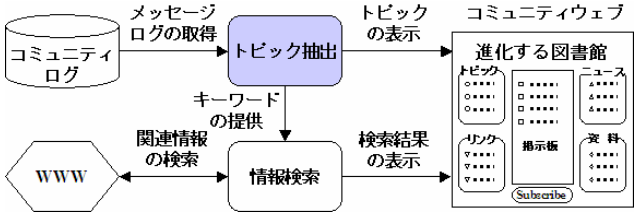

図 1:本研究で開発するソフトウェアの仕組み

連絡先: 余 東明, 日本工業大学大学院工学研究科情報工学専攻, 〒 345-8501 埼玉県南埼玉郡宮代町学園台 4-1, Tel: 0480-34-4111(代), E-mail: c2025010@cstu.nit.ac.jp

トピック抽出は, コミュニティ支援のために単独でも有用な基本機能であり, アクティブ情報検索に対する前処理の機能でもある. 従って, 抽出するトピックの適切さはコミュニティ支援の効用およびコミュニティウェブのユーザビリティに大きな影響をもたらす. 本論文では, コミュニティウェブのメッセージから情報検索に有効なトピックを抽出する手法について述べる.

以下, 2 節では関連研究と本研究の位置付けについて述べる. 3 節ではトピック抽出のアルゴリズム, 4 節ではトピック抽出の実験について述べる. 5 節では, 今後の課題を記す.

2. 関連研究と本研究の位置付け

トピック抽出は, 特定の文書集合(コーパス)から, 重要な, 話題性を持つ概念を自動的に抽出する技術である. 従来, 本研究とは異なる目的として, 索引語としての専門用語の抽出 [中川 2003]や, 要約ためのキーワードの抽出[松尾 2002]などの研究が行われている. 従来の研究と比べると, 本研究は以下の特徴をもつ.

(1) 対象となる文書集合はメッセージの集合である.

コミュニティウェブでは, トピック抽出の対象は, メーリングリストや電子掲示板への投稿である. 本研究ではこれらをまとめて「メッセージ」と呼ぶ. メッセージは, 以下の性質を持っている.

- メッセージの文体は一般に口語体で, その本文には構造化のため改行などが含まれる.
- メッセージには, サブジェクト, 投稿者, 投稿時間などの情報が含まれている. さらに, 他のメッセージとの応答関係が含まれる場合もある.
- 多くのメッセージには, 投稿者の肩書きとシグネチャーが含まれている.

(2) トピック抽出の目的は, メッセージの分類と WWW からの関連情報の検索である.

本研究では, 話題性を持つ概念が一般的に名詞句によって表現されることに着目して, 名詞句を構成する名詞であるキーワードのリストによってトピックを表現する. この表現は, キーワードによる関連情報検索に対しても有用である. 名詞句の構造は, BNF 記法により, 表1に定義する. 品詞の判定には辞書を用いる.

表 1:名詞句の定義

〈名詞句〉	::=	〈名詞〉|〈名詞相当句〉	(1)
〈名詞相当句〉	::=	〈複合名詞〉|〈連体修飾構造〉	(2)
〈複合名詞〉	::=	〈名詞〉〈名詞〉{〈名詞〉}	(3)
〈連体修飾構造〉	::=	〈名詞〉の〈名詞〉{の〈名詞〉}	(4.1)
	|	[〈名詞〉(を|が)]〈動詞〉〈名詞〉	(4.2)
	|	[〈名詞〉(を|が)]〈動詞タ接続〉〈助動詞〉〈名詞〉	(4.3)
	|	〈副詞〉の〈名詞〉	(4.4)
	|	〈形容詞〉〈名詞〉	(4.5)
	|	〈形容動詞〉な〈名詞〉	(4.6)
	|	〈連体詞〉〈名詞〉	(4.7)
	|	〈名詞〉〈格助詞〉の〈名詞〉	(4.8)
	|	〈名詞〉〈連語〉〈名詞〉	(4.9)

3. トピック抽出のアルゴリズム

トピック抽出処理は，対象とする各メッセージに対し，まず構造上のノイズを除いて，形態素解析[松本]を行う．形態素解析の結果に出現する語句を次元とし，その重要度により各メッセージの特徴ベクトルを構成する．次に，その特徴ベクトルを使ってメッセージのクラスタリングを行い，話題性が類似したメッセージのクラスタを形成する．最後に，各クラスタにおいて共通するトピック候補の組を選択してトピックを構成する．以上のトピック抽出のプロセスを図 2 に示す．

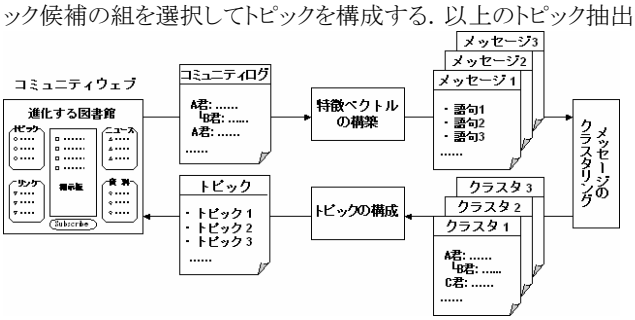

図 2:トピック抽出のプロセス

以下，メッセージの特徴ベクトルの構築，メッセージのクラスタリング，トピックの構成について説明する．

3.1 メッセージの特徴ベクトルの構築

メッセージの内容を数量的に表すために，その内容を表す特徴ベクトルを構築する．内容を表す具体的な概念は主に名詞，動詞，形容詞，副詞により表現されるので，本研究では，その四つの品詞に属する単語の出現と重要度によりメッセージの特徴ベクトルを構築する．

(1) 語句の重要度

あるメッセージに m における語句 t の重要度 $w(t, m)$は，以下の式によって定義する．

$$w(t,m) = td * idf = \frac{tf(t,m)}{M} * \left(\log_{10}\left(\frac{N}{df(t)}\right) + 1 \right) \quad (1)$$

ここで，語句 t の単メッセージ m 内の出現頻度（回数）を $tf(t, m)$，メッセージ m の中に全語句数を M，全メッセージ数を N，語句 t を含むメッセージ数を $df(t)$とする．式(1)は，単メッセージに密度が高くかつ特定のメッセージにしか出現しない語句の重要度を高くする．語句の密度は，ある語句 t のあるメッセージ m 内で出現頻度 $tf(t, m)$を m の全語句数 M で割った値 td を使った．あるメッセージの全語句数 M が一定の場合，ある語句の出現頻度 tf が高くなれば，その密度は高くなる．特定のメッセージにしか現れないことは，全メッセージ数 N を，語句 t を含むメッセージ数 $df(t)$で割った数の対数 idf で表した．メッセージ集合において，ある語句が多くのメッセージ中に現れる語句である場合に idf は小さくなり，逆に，特定のメッセージにしか現れない場合に idf は大きくなる．

一般的に，語句の重要度として $tf*idf$ という値[長尾 2001]がよく用いられているが，本論文では tf を td に置き換えている．この理由は，一般にメッセージ中の語句数が一定でないために，頻度より密度の方が比較に適すると考えるためである．たとえば，同じ語句が二つのメッセージに 30 回ずつ出現するとき，ひとつのメッセージの語句数は 2 千，もうひとつのは 1 万とすれば，その語句はそれぞれのメッセージにおける重要度が違うはずである．

(2) 特徴ベクトルの構築

メッセージ m の特徴ベクトル V_m は，その第 i 成分 $w(t_i, m)$を単語 t_i が m に含まれるとき $td*idf$，含まれないとき 0 として式(2)で定義する．

$$V_m = (w(t_1,m), w(t_2,m), ..., w(t_n,m)) \quad (2)$$

ここでは，メッセージの内容を表現するために使う単語の数を n，単語を $t_1, t_2, ..., t_n$ としている．

3.2 メッセージのクラスタリング

メッセージを内容で分類するために，各メッセージに含まれるパターンのベクトルとした類似度によって，階層的クラスタリングを行う．階層的クラスタリングにより形成された各クラスタ内のメッセージは，同じトピックに関わると考えられる．具体的には，次のプロセスに従ってメッセージのクラスタリングを行う．

(1) 二つのクラスタ間の類似度

二つのクラスタ c_j, c_k の類似度 $sim(c_j, c_k)$は式(4)で求める．

$$sim(c_j,c_k) = \frac{\sum_{i=1}^{n} w(t_i,c_j)w(t_i,c_k)}{\sqrt{\sum_{i=1}^{n}(w(t_i,c_j))^2 \sum_{i=1}^{n}(w(t_i,c_k))^2}} \quad (3)$$

$w(t_i, c_j)$と $w(t_i, c_k)$ はそれぞれ単語 t_i のクラスタ c_j と c_k における重要度である．

(2) クラスタリングのアルゴリズム

メッセージ内容の主題性と差異性を表すために，本研究ではクラスタリングに階層的方法を用いる．メッセージのクラスタリング処理は表 2 の手順で行う．

表 2:クラスタリングのアルゴリズム

① すべてのメッセージを，ひとつのメッセージからなるクラスタとする．
② すべてのクラスタ間の類似度を計算し，類似度が一番高い二つのクラスタを統合する．
③ ②の処理を，すべてのクラスタがひとつのクラスタになるまで繰り返し行う．

クラスタリングには重心法を用い，p 個のメッセージ V_1, V_2,..., V_p が形成するクラスタ c の重心ベクトル CV を式(3)で求める．

$$CV = \left(w(t_1,c), w(t_2,c), ..., w(t_n,c)\right) = \frac{1}{p}\left(\sum_{h=1}^{p} w(t_1,m_h), \sum_{h=1}^{p} w(t_2,m_h), ..., \sum_{h=1}^{p} w(t_n,m_h)\right) \quad (4)$$

3.3 トピックの構成

3.2 節のクラスタリング処理によって，二分木構造のクラスタ木が形成される．木のルートにあるクラスタ（ルートクラスタ）はすべてのメッセージを含むクラスタで，木の葉にあるクラスタは単メッセージからなるクラスタである．葉にあるクラスタを除いて，木の各ノードにあるクラスタはひとつ下の階層にある二つのクラスタからなる．

トピックの構成方法は，まず，表1の名詞句の定義を満たすパターンを抽出し，トピック候補とする．トピック候補の重要度は 3.1 の式(1)を使う．ただし，対象となる語句を名詞句とする．次に，クラスタ木の最上位レベルから，最下位レベルまで，各クラスタにおける重要度が最上位の共通トピック候補をそのクラスタのトピックとし，上位クラスタのトピックを合わせて，そのクラスタのトピックとする．共通トピック候補とは，各クラスタのすべてのメッセージに含まれるトピック候補である．このようにトピックを構成する理由は，同じクラスタに属するメッセージは同じトピックに関わっており，あるクラスタに属する二つのクラスタはトピックが内容の細部では異なるためである．ただし，ルートクラスタの共通トピック候補はすべてのメッセージに含まれるので，トピックにならない．

トピック構成のアルゴリズムは表 3 のとおりである．アルゴリズム中，「共通トピック候補」，「トピック不可」，「トピック」というリストは，名前通りに，「共通トピック候補」リストには各クラスタにあるすべてのメッセージが含まれたトピック候補を格納し，「トピック不可」リストにはトピックになれない語句を格納し，「トピック」リストにはメッセージ集合のトピックを格納している．

表 3:トピック構成のアルゴリズム

① ルートクラスタの「共通トピック候補」リストにあるすべてのトピック候補を「トピック不可」リストに追加する．
② 最上位レベルを除いて，その一つ下位のレベルから，最下位レベルまで，各レベルにあるすべてのクラスタに対し，次の処理を繰り返す．
  (a) クラスタの最上位にあるトピック候補は「トピック不可」と「トピック」リストに含まれない場合に「トピック」リストに追加する．
  (b) さもなければ，当該トピック候補を「共通トピック候補」リストから削除し，(a)に戻る．

このアルゴリズムによって構成したトピックは，単語のリストであるため，リスト中の単語をキーワードとして AND 検索に利用できる．トピックを見出しとして表示する場合は，リスト中の単語を最も多く含む名詞句を抽出して使用する．

4. トピック抽出の実験

4.1 実験の対象と方法

本論文で提案したトピック抽出の手法を実際のコミュニティウェブのメッセージデータによって評価した．実験の対象は「Yahoo! e グループ」の「進化する図書館」[sinka]というグループ（会員制）で，そのグループでは毎月平均 80,000 字相当のメッセージが増えている．

トピックの抽出は奈良先端科学技術大学院大学で開発された形態素解析システム「茶筌」[松本]を用い，実験のプログラムは Perl によって作成した．実験方法は，まず，毎日，前の一週間分のメッセージに対し，トピック抽出処理を行う．トピックは指定した数の葉クラスタについて「話題」として表示する．次に，抽出されたトピックに基づいて，関連するニュース情報を検索する．抽出するトピックと情報検索の結果は JSP による作成した実験サイトを通じてグループのメンバーに提供する．実験サイトのトップページの一例を図 3 に示す．

4.2 実験の結果と評価

実験の結果，図 3 では「知識の時代に求められる図書館員の能力」，「フリーライブライアン」，「社会企業家への支援」などの独立な話題を表すトピックが抽出されている．ただし，「進化する図書館」グループのメンバーからコメントによると，「高山市図書館」などのメッセージの肩書きとシグネチャーに含まれた名詞句や，重要度がそんなに高くない名詞句もトピックとして抽出された場合がある．抽出結果の適合率は 80%前後と推定される．

5. 今後の課題

実験の結果明らかになった以下の課題を解決する必要がある．

(1) 共通トピック候補において重要度が低い語句が含まれる．

本研究では，クラスタごとに共通するトピック候補からトピックを構成するので，クラスタ木のより上位のレベルにあるクラスタでは，クラスタに共通するトピック候補が少なくなることにより，重要度が低い語句がトピックになる．この問題の解決にはトピックとして選択する語句の重要度に下限を設けることが考えられる．

(2) 肩書きやシグネチャーが含まれてしまう．

メッセージに含まれる肩書きとシグネチャーにおける住所などの名詞句は，*idf* が大きく，長さも大きいので，高い重要度が付き，トピックになる可能性が高くなる．この問題の解決には，肩書きとシグネチャーなどの文書構造を解析して，本文から取り除く処理が必要である．

(3) クラスタリングに時間がかかる．

現在の実験では，一週間分（30 件前後）のメッセージをトピック抽出の対象とするので，クラスタリングにはまだ問題がない．しかし，メッセージの件数がふえるに伴い，クラスタリングの時間も長くなる．その問題を解決するために，メッセージ間の応答関係などを利用して，クラスタリングの計算量を低減する必要がある．

6. おわりに

本論文はコミュニティウェブにおけるアクティブ情報検索のためのトピック抽出の手法と実験結果について述べた．本論文で提案したトピック抽出の手法は，メッセージをクラスタリングによって分類し，各クラスタに共通する語句のリストをトピックとして抽出する．抽出したトピックは情報検索のキーワードとしてアクティブに利用できる．現在，名詞句の重要度の計算方法やクラスタリングのアルゴリズムなどを改良して，実コミュニティサイトにおいて，評価実験を続けている．

図 3:実験サイトの一例 (「進化する図書館」コミュニティウェブ)

参考文献

[石川 2003] 石川 孝: コミュニティウェブソフトウェアにおけるアクティブマイニング, アクティブマイニング合同研究会, 2003.

[余 2002] 余 東明, 石川 孝: コミュニティウェブにおける掲示板からのトピック抽出, 第 1 回情報科学技術フォーラム, 一般講演論文集, 第 2 分冊, E-17, pp.115-116, 2002.

[中川 2003] 中川裕志, 森辰則, 湯本紘彰: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[松尾 2002] 松尾 豊, 大澤幸生, 石塚 満: 電子掲示板における会話からのハイライト部分の抽出, SIG-FAI 研究会, 2002.

[長尾 2001] 長尾 真: 自然言語処理, 岩波書店, pp.421, 2001.

[松本] 奈良先端科学技術大学院大学情報科学研究科松本研究室. 茶筌:http://chasen.aist-nara.ac.jp/index.html.ja/

[egroups] Yahoo! Japan e グループ:http://www.egroups.co.jp/

[学会 NET] http://www.skysoft.co.jp/gakkai/

[sinka] 進化する図書館: http://www.egroups.co.jp/group/sinka/