コミュニティ型コンテンツのコンテンツホール抽出手法の提案

Extracting for Content Hole of the Community Type Content

灘本 明代♥ 阿辺川 武♦
荒牧 英治♣ 村上 陽平♠
Akiyo NADAMOTO Takeshi ABEKAWA
Eiji ARAMAKI Youhei MURAKAMI

SNSやブログのようなコミュニティ型コンテンツの場合，コミュニティ内での議論に集中するあまり視点が狭くなり，議論におけるテーマを多面的に捉えられなくなる危険性がある．我々は，このような見落とされた視点をコンテンツホールと呼び，SNS やブログにおけるコミュニティ内の議論の履歴からコンテンツホールを抽出しユーザに提示することを試みている．本研究では，そのための第一歩となる(1)Web 空間のあるテーマに対する視点抽出(2)視点間の比較によるコンテンツホールの抽出を行う．

In a case of the community type content such as SNS and Blog, sometimes users don't understand the theme of the content from multi-viewpoint and lost much information. Because they concentrate the discussion in the community, they become narrow viewpoint. We call the user's unawareness information "the content-hole". We try to extract and represent the content-hole from the history of their discussion on the SNS and Blog. In this paper, we propose the first technique of (1)Extract the viewpoint of theme on the Web, (2)Extract the content-hole based on comparing the viewpoints.

1. はじめに

Web2.0を代表するコンテンツのひとつであるブログやSNSのようなコミュニティ型コンテンツがインターネット上に多数存在する．これらコミュニティ型コンテンツはこれまでのWebコンテンツと異なり，そのコミュニティに参加しているユーザが気楽にコンテンツを記述し，コミュニティ内において議論や情報交換を容易に行えるといった特徴がある．それ故にコミュニティ内での議論に集中するあまり視点が狭くなり，テーマを多面的に捉えられなくなる危険性がある．例えば，京都のラーメン店をテーマとしたコミュニティの場合，ある有名なラーメン店の話題に集中しているばかりに，他に京都の美味しい穴場のラーメン店がたくさんあることに気づかないことがある．このような場合，このコミュニティのメンバーが京都へ行ったときに本当に美味しいラーメンを食べ損なう可能性がある．このようにコミュニティ内ではそのテーマの中の一部の話題に集中してしまい，本来そのテーマの全容を見ることなく，「井の中の蛙，大海を知らず」という事態に陥っている場合が多い．特に医療等においてこのような状態が起こることは大変深刻な問題である．そこで，本研究ではこのようにコミュニティが本来テーマとしている事柄に対し，他ではどのような議論が行われているかを抽出し，そのコミュニティ内で議論されている事柄と比較し，その差分情報をコミュニティに提示することを行う．この差分情報はコミュニティ内においてそのユーザの気づいていない情報であり，本研究ではこれをコンテンツホールと呼ぶ．そしてこのコンテンツホールを探し出すことつまりは「ないものを探す」ことをコンテンツホール検索と呼ぶ．

実際には，そのコミュニティがテーマとしているコミュニティ内の視点情報を抽出する．次にWeb空間よりそのテーマの視点情報を抽出する．これらを比較し，その差分情報を取得する．この差分情報がそのコミュニティにとってのコンテンツホールとなる．そしてそのコンテンツホールをコミュニティのユーザに提示する．本論文ではその第一歩となる(1)Webページからの視点抽出(2)視点間の比較手法の提案を行う．

2. コンテンツホール検索の基本コンセプト

コンテンツホール検索の処理方法は以下の通りである．

(1) コミュニティ型コンテンツからのテーマの抽出
SNSやブログなどからそのテーマとなる名詞を抽出する．テーマは時系列的に変化する場合や複数のテーマを取り扱っている場合が想定されるが,本研究ではテーマは一つとし，時系列に変化しない物とする．

(2) コミュニティ型コンテンツにおけるコミュニティ内の視点情報の抽出
コミュニティ型コンテンツにおいて，「名詞A+が+形容詞+名詞B」という構文に注目し，自然言語処理によりそのコミュニティの視点構造を抽出する.この視点構造は階層構造を持ちXMLで記述する．

(3) Web空間における抽出したテーマの視点情報の抽出
大規模Web空間において，Webページ群から「名詞A+が+形容詞+名詞B」という構文に注目し，自然言語処理によりWeb空間全体の視点構造を抽出する．この視点構造は階層構造を持ちXMLで記述する．

(4) Web空間における視点構造とコミュニティ内の視点構造を比較しその差分情報であるコンテンツホールを抽出
Web空間における視点情報とコミュニティにおける視点情報各々から視点構造グラフを作成する.そしてこれらの視点構造グラフを比較することにより差分情報を取得する.グラフ作成時に概念構造を考慮することにより，同一テーマにおける概念の異なる情報を抽出することを行う．

(5) 抽出されたコンテンツホールをユーザに提示
抽出したコンテンツホールをコミュニティ内のユーザに提示するユーザインタフェースを開発する．

本論文では上記手順の内，(2)(3)(4)の手法を提案する．

コンテンツホール検索の特徴は以下の通りである．

・ ユーザはコミュニティ内の議論において不足してい

♥ 正会員 独立行政法人 情報通信研究機構
nadamoto@nict.go.jp
♦東京大学大学院 教育学研究科
abekawa@p.u-tokyo.ac.jp
♣東京大学 医学部付属病院
aramaki@hcc.h.u-tokyo.ac.jp
♠独立行政法人 情報通信研究機構
yohei@nict.go.jp

る内容を把握でき，これにより，より公平性のある議論を行うことが出来る．
- ユーザはコミュニティ内のテーマの一般的な視点が把握できる．

3. Web ページからの視点抽出

大規模Web空間における各Webページの視点構造を抽出する．情報検索において膨大な検索結果から必要な情報を選択あるいは統合するために，ある対象に対して関連する属性を抽出する研究が試みられており，Webの表形式から属性表現を収集する手法[1]や，ルールを用いて属性表現を収集し，オントロジーを構築する手法[2]や，話題構造を抽出する手法[3][4]などが提案されている．本論文ではテーマを示す単語に係る形容詞に注目しWebページの視点抽出を行う．具体的には「オムライスが美味しいレストラン」のような「名詞A+が+形容詞+名詞B」という構文に注目し，この「名詞Aが」が形容詞に係るとき，その名詞Aと名詞Bが属性とテーマ（対象）との関係になっていることが多いという前提のもとに，Web上からこのような表現を収集し，あるテーマに対してその属性集合を収集する．

ここであるテーマに対する視点情報とは属性である「名詞A＋形容詞」と定義する．「名詞A＋が＋形容詞＋名詞B」という構文に着目した理由は，形容詞はその多くが必須格を1つしか持っていないからである．「名詞Aが」が形容詞に係り，形容詞の各スロットを埋めているとき，名詞Bは形容詞に連体修飾されていても形容詞との格関係を持つことが出来ず，名詞Aと関係を持つことになるからである．また，表形式やリスト構造を利用した既存手法では，あるテーマについて情報発信者により明示的に選択され記述された属性を抽出しているのに対し，本手法で用いる「名詞A+が+形容詞+名詞B」というフレーズは文章中に自然に出現するため，情報発信者は無意識のうちに属性を記述していると捉えることができる．そのため既存手法では取得できない属性が得られる可能性がある．また扱うデータがコミュニティ内での議論という特性上，一般に文章により構成されていることが多いことから，本手法の利点が活かされると考えられる．

本手法は以下の手順で行う．

1. 学習モデルの訓練データとして，あるテーマについてWeb上から「名詞A+が+形容詞+名詞B」を収集する．
2. 人手により「名詞Aが」が形容詞に係る事例のラベル付けを行う．
3. ラベル付けされた事例から名詞Aを抽出し，実際に属性名詞となってるかを確認する．
4. 名詞Aが形容詞に係るか否かを認識する学習モデルを構築する．
5. 4で構築した学習モデルを用いて他の対象名詞について属性抽出を行う．
6. 「抽出した属性（名詞A）＋形容詞」をそのテーマの視点とする．

本章では対象名詞「レストラン」を例にWebからその属性名詞集合を収集する手法を説明する．

3.1 「名詞 A+が+形容詞+名詞 B」の収集

ある対象名詞に対する上記構文を検索エンジンを用いて収集する．本研究では，対象名詞が上記構文で使用されている事例を出来るだけ多く収集したいため，新聞コーパスで頻出するイ形容詞，ナ形容詞をそれぞれ500個ずつ用意し，それぞれの形容詞に対して「が＋形容詞＋レストラン」といったフレーズ検索を行い，そのフレーズを含むsnippetを収集する．

一般にフレーズ検索では，間に記号が挿入されていてもそれらは無視される．例えば「が美味しいレストラン」でフレーズ検索を行った場合，「～が美味しい．レストランでは～」のように間に句点が入ったページも検索される．このような文を削除する．さらに「名詞A＋が＋形容詞＋名詞Bの名詞C」の場合では，形容詞の係り先に曖昧性が生じてしまうので，名詞Bに助詞「の」が後接する場合も削除する．その後，句点などの記号を文区切りとみなし，snippetからフレーズが含まれる文を抽出する．

検索エンジンは内容が全く同じページでもホストが異なれば別々の検索結果として表示されることがあるため，文字列が完全に一致する文は1文にまとめる．実際にYahoo!Japanを使用し「が美味しいレストラン」のフレーズ検索を行った結果が5545個に対してフィルタリング後の結果は2512文となった．

3.2 ラベリング

得られた文集合を名詞Aの係り先と対象・属性の観点から以下の4つに分類する．

ラベル1：「名詞Aが」が形容詞に係り，名詞Aが名詞Bの属性である．
「予約が困難なレストラン」

ラベル2：名詞Aが名詞B以降の文節に係る．
「ここが高級なレストランである」

ラベル３：「名詞Aが」が形容詞に係るが，名詞Aが名詞Bの属性ではない．
「私が好きなレストラン」

ラベル４：文区切りの失敗のため得られた文字列が文をなしていない．対象名詞が文節の主辞となっていない．
「読者コメントがおもしろいレストラン情報」

上記分類を人手により行った結果，分類内訳はラベル1が1715，ラベル2が325，ラベル3が433，ラベル4が39となった．ラベル１に分類した事例から名詞A及び名詞A+形容詞のペアを抜き出し，頻度順位に列挙した結果の上位10位を表1に示す．名詞Aのリストをみると殆どの名詞は対象名詞(テーマ)「レストラン」の属性名詞であると考えられる．これにより「名詞A＋が＋形容詞＋名詞B」の構文を用いることでWeb上から対象名詞と属性名詞のペアが収集できることが分かった．

表１　収集した属性名詞の例
Table 1　Example of Attribute Proper Noun Collection

名詞A	名詞A＋形容詞
122　夜景	72　夜景　綺麗だ
89　料理	43　料理　美味しい
81　雰囲気	36　予約　必要だ
65　予約	29　パン　美味しい
63　眺め	27　人気　高い
46　景色	23　夜景　美しい
38　パン	22　予約　困難だ
28　眺望	22　雰囲気　良い
28　人気	21　雰囲気　いい
25　インテリア	19　眺め　いい

3.3 学習モデル

ラベル付けして得られた事例を学習データとみなして，機械学習により「名詞Aが」が形容詞に係るか否かを識別する学習モデルを構築し，ある対象名詞に対しその属性名詞集合を自動的に収集することを行う．

(1)素性

例文としてラベル１である「小娘には敷居が高いレストランだが，高級すぎるほどではない」を用いて機械学習アルゴリズムで使用する素性について説明する．はじめに文を形態素解析した後，表2に示すように構文を含む文節とその1つ前の文節を取り出す．そして，それぞれの文節について表3にある素性を抽出する．尚，形態素解析にはJUMANを文節区切りにはKNPを利用した．文節の語形・主辞の定義は，文献[5]にならう．また語の概念は，日本語語彙大系[6]を参照し，ルートからの一定の深さにある概念番号を使用する．

(2)予備実験1．一般名詞

3.2節でタグ付けを行ったデータに対して，分類実験を行なった．実験ではラベル4を除き，ラベル1,2,3の3値分類とする．機械学習アルゴリズムにはSVMとして実装にはTinySVMを使用し，カーネルは線型カーネル，多値分類にはpair wise法を用いた．評価は5分割交差検定で行った．実験の結果は，正解率が0.945，ラベル1の精度が0.955，ラベル１の再現率が0.979だった．正解率とは，分類器が正しくラベルを判別したときの正解率，ラベル1の精度とは，分類器がラベル1と判別した内，正解した事例の割合，ラベル1の再現率とは，全てのラベル1の事例に対して，分類器がラベル1と判別した割合である．

表２　素性を抽出する文節
Table 2　Paragraph of Extracted Features

	文節１	文節２ 名詞A	文節3 形容詞	文節4 名詞B
例	小娘には	敷居が	高い	ﾚｽﾄﾗﾝだが

表３　使用した素性
Table 3　Used Features

素性	例
1.文節1の語形の見出し	は
2.文節1の語形の品詞	助詞
3.文節1の語形の品詞細分類	副助詞
4.文節1の読点の有無	無
5.文節2の主辞の見出し (接尾辞がある場合は接尾辞も含める)	敷居
6.文節2の主辞の品詞	名詞
7.文節2の主辞の品詞細分類	普通名詞
8.文節2の主辞の活用	無
9.文節2の主辞の活用形	無
10.文節2の主辞の見出しの概念 (固有名詞は深さ2，一般名詞は深さ3)	0533
11.文節3の主辞の見出し	高い
12.文節2と文節4の概念が同一か	異なる
13.文節4が文中か文末か	文中

得られた学習モデルを利用して他の名詞について，その名詞を対象名詞として属性名詞集合を収集する実験を行った．対象名詞「レストラン」において学習したモデルなので，語の見出しを素性とするものは「レストラン」に特化した素性のままである．実験には，対象名詞として「ホテル」「会社」「デジカメ」を選択した．それぞれの名詞について，3.1節の手法でWebから文を収集し，3.3節で用いた学習モデルを使用して構文を分類した．収集された文数，ラベル1の文数，そしてランダムに選択した200個の事例に対して人手で評価した分類正解率を表4に示す．3つの名詞についてランダムに200個を抽出し評価した分類正解率では「ホテル」「デジカメ」は「レストラン」よりも良い正解率を示している．その理由として「ホテル」は「レストラン」とドメインが似通っているため，また「デジカメ」は名詞Aに「私が」「夫が」のように人称名詞が来る事例が多く，名詞Aに関する素性が有効に働いたためであると考えられる．

表４　他の対象名詞における実験結果
Table 2　Experiment Result of Other Target Nouns

名詞	ホテル	会社	デジカメ
取得snippet数	12,411	26,075	4,629
フィルタリング後	5,090	10,348	2,203
ラベル1に分類	3,609	7,120	955
正解率(任意200個)	95.5%	81.0%	95.0%

(3)予備実験2.固有名詞

ブログやSNSの場合，ある歌手に対するコミュニティやあるゲーム機に対するコミュニティ等一般名詞だけでなく固有名詞をテーマとするコミュニティが多く存在する．そこで，本提案手法が固有名詞にも有用であるかどうかを図るための予備実験を行った．実験手法は予備実験1と同一であり，テーマはゲーム機の「Wii」とした．検索エンジンの総Hit数は10,802,取得snippet数は2,598,フィルタリング後 1,105,重複除去後565，SVM分類後287であった．表5に収集した視点構造（名詞A+形容詞）の全体75件の内，上位10件を示す．表1の結果と比較して，頻度が低く全体的に幅広い話題となっている．しかしながら，上位10件を見ても，Wiiの特徴が分かるように，固有名詞においても十分にその話題における視点が抽出出来ていることがわかる．

表５　Wii に対しての視点構造
Table 5　Example of Viewpoint Structure for "Wii"

名詞A	名詞A＋形容詞
63　ソフト	56　ソフト　少ない
24　範囲	24　範囲　狭い
12　入手	12　ないほう　いい
12　ないほう	11　攻略　素っ気無い
11　攻略	9　コントロール　楽しい
9　コントロール	8　熟練度　凄い
9　ゲーム	6　入手　難しい
8　熟練度	6　環境　無い
6　販売度	5　入手　困難だ
6　環境	5　ゲーム　多い

4. 視点間の比較

3章で述べた手法で求めた「名詞A＋形容詞」がテーマ名詞Bの視点となる．ここで，Web空間全体から視点を抽出するとその結果は名詞Bに対するWeb空間全体の視点構造となり，あるコミュニティ型コンテンツ内において同様の方法で視点

を抽出するとその結果は，名詞Bに対するそのコミュニティ内の視点構造となる．これらWeb空間全体の視点構造とコミュニティ内の視点構造を比較しその差分情報をコンテンツホールとする．ここでは，あるテーマ名詞Bに対し求められた視点構造を構成する名詞Aの概念的関係を考慮する必要があるが，本論文ではコンテンツホールを求める手法の第一歩となる最も単純な手法である概念関係を考慮しない手法で視点間の比較を行う．その上で将来，概念関係を考慮した手法を提案する．視点を抽出する手順は以下の通りである．

(1) Web空間全体の視点構造からテーマ（名詞B）をルートとし，名詞Aを経由して形容詞を葉節点とする視点構造グラフを作成する．この時，全ての名詞Aと形容詞のペアを対象にすると膨大なグラフとなる可能性がある．ある閾値α以上の頻度を持つ名詞A，または，ある閾値β以上の形容詞を持つ名詞Aと形容詞のペアを対象とする．閾値β以上の形容詞を持つ名詞Aを対象とした理由は，多数の形容詞との係り受け関係を持つ名詞Aは多様な視点から構成されている可能性が大きいと考えた為である．

(2) 作成した視点構造グラフからシソーラスを用いて類似する単語は同一の意味を持つと考え，類似する単語を示すノードをマージする．

(3) コミュニティの視点構造から(1)(2)と同様に視点構造グラフを作成する．

(4) (1)と(2)のグラフを比較し，その差分情報の候補を取得する．

図1にαを80，βを3とした表3のWeb空間全体の視点構造グラフとあるコミュニティの視点構造グラフを示す．ここでは，Web空間全体の視点構造のグラフにて丸で囲まれている部分があるコミュニティにおけるコンテンツホールとなる．このグラフから分かるように，ここでコンテンツホールとなっている「夜景」「景色」「眺め」は同一概念であるといえる．ここで，「景色」「眺め」は同一概念であり，また，「夜景」は「景色・眺め」の一種（下位概念）である．これらを正確に扱うために，概念構造を用いてコンテンツホールを抽出する必要があり，今後の課題である．

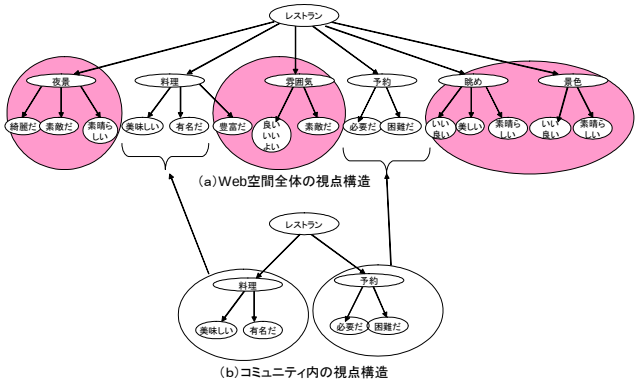

図1　視点構造グラフの比較例

Fig.1　Example of Viewpoint Structure Graph

5. まとめ

本論文ではコミュニティ型コンテンツのコンテンツホール検索を行う手法の提案の第一歩として(1)Web空間のあるテーマに対する視点抽出(2)視点間の比較によるコンテンツホールの抽出の提案を行った．このように見落とされた視点情報であるコンテンツホールを提示することにより，ユーザはこれまで気づかなかった情報について知ることができ，より公平性のある議論をすることができるようになると考えられる．

[謝辞]

本研究の一部は，平成 19 年度科研費特定領研究域「Web 2.0 時代のコミュニティ型コンテンツのコンテンツホール検索に関する研究」(課題番号：19024072，代表：灘本明代)による．ここに記して謝意を表します．

[文献]

[1]大前 信弘, 黄瀬 浩一, "Web の表を対象とした属性の自動識別", 情報処理学会研究報告 171-NL-8, 2006

[2]松平 正樹, 上田 俊夫, 大沼 宏行, 森田 幸伯, "Web コンテンツの分析に基づくオントロジー構築および情報整理の試み", 人工知能学会セマンティックウェブとオントロジー研究会, SIG-SWO-A302-08, 2004

[3]M.Spitters and W.Kraaij, "A Language Modeling Approach to Tracking News Events," TDT 2002 Evaluation workshop, Gaithersburg, MD, USA, 2002.

[4]Akiyo Nadamoto, Ma Qiang, and Katsumi Tanaka “B-CWB: Bilingual Comparative Web Browser Based on Content-Synchronization and Viewpoint Retrieval”, World Wide Web Journal, Springer Science+Business Media B.V., ISSN: 1573-1413 (Online)

[5]内元 清貴, 村田 真樹, 関根 聡, 井佐原 均, "後方文脈を考慮した係り受けモデル", 自然言語処理, Vol7, Number5, PP.3-17, 2000

[6]池原 悟, 中井 慎司, 村上 仁一, "多義解消のための構造規則の生成方法と日本語名詞句への適用", 自然言語処理, Vol8, Number1, pp.143-173, 2001

灘本 明代　Akiyo NADAMOTO

独立行政法人情報通信研究機構主任研究員，2002 年神戸大学大学院自然科学研究科博士後期課程修了，博士（工学）．Web コンテンツ，Web サービスの検索，閲覧，配信に関する研究に従事．情報処理学会，日本データベース学会会員．

阿辺川 武　Takeshi ABEKAWA

東京大学大学院教育学研究科学術研究支援員，2006 年東京工業大学知能システム工学研究科博士後期過程修了．博士(工学)．自然言語処理・翻訳支援の研究に従事．

荒牧 英治　Eiji ARAMAKI

東京大学医学部附属病院，特任助教．2005 年東京大学大学院情報学研究科博士後期課程修了，博士（情報理工学）．自然言語処理，医療情報の研究に従事．情報処理学会，言語処理学会，医療情報学会，ACL 会員．

村上 陽平　Youhei MURAKAMI

独立行政法人情報通信研究機構研究員，2003 年京都大学大学院社会情報学専攻修士課程修了．2006 年，同大学院社会情報学専攻博士課程修了．博士(情報学)．現在，言語グリッドプロジェクトを推進．人工知能学会会員．