

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

灘本 明代[†] 阿辺川 武^{††} 荒牧 英治^{†††} 村上 陽平[†]

[†] 独立行政法人 情報通信研究機構 〒 619-0289 京都府相楽郡精華町光台 3-5

^{††} 東京大学大学院 教育学研究科 〒 113-0033 東京都文京区本郷 7-3-1

^{†††} 東京大学 医学部付属病院 〒 113-8655 東京都文京区本郷 7-3-1

E-mail: [†]{nadamoto,yohei}@nict.go.jp, ^{††}abekawa@p.u-tokyo.ac.jp, ^{†††}aramaki@hcc.h.u-tokyo.ac.jp

あらし Web2.0 を代表する技術のひとつである 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野が狭くなり、テーマを多面的に捉えられなくなる危険性がある。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。具体的には「名詞 A + が + 形容詞 + 名詞 B」の構造に注目し、あるテーマ名詞 B に対しその視点構造を「名詞 A + 形容詞」として定義し、Web 空間とコミュニティ内との 2 つの視点構造を抽出する。それら視点構造を比較しその差分情報を取得することによりコンテンツホールを抽出する。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。

キーワード コミュニティ型コンテンツ, コンテンツホール, 視点抽出

Extracting for Content Hole of the Community Type Content

Akiyo NADAMOTO[†], Takeshi ABEKAWA^{††}, Eiji ARAMAKI^{†††}, and Youhei MURAKAMI[†]

[†] National Institute of Information and Communications Technology Hikaridai 3-5, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan

^{††} The University of Tokyo Graduate School of Education Hongou 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan

^{†††} The University of Tokyo Hospital Hongou 7-3-1, Bunkyo-ku, Tokyo, 113-8655 Japan

E-mail: [†]{nadamoto,yohei}@nict.go.jp, ^{††}abekawa@p.u-tokyo.ac.jp, ^{†††}aramaki@hcc.h.u-tokyo.ac.jp

Abstract The community type content such as SNS and Blog are representative technique of Web2.0. Sometimes users who use the community type content 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. We focus on the structure of a sentence that is "NounA+Ga(one of a Japanese conjunction)+Adjective+NounB", we define that the viewpoint of theme(NounB) is "NounA+adjective". We extract the viewpoint of web content and viewpoint of community. Then, we compare two viewpoints and extract different points. the different points becomes the content-hole. In this way, by representing the content-hole, user can get the new information and they can do fair discussion.

Key words Community type content, Content-hole, Viewpoint detection

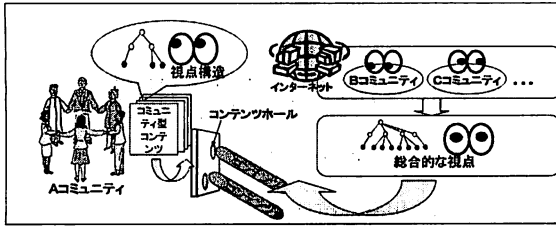


図1 コンテンツホール検索のイメージ図
Fig.1 Image of Search for Content-Hole

1. はじめに

近年 Web2.0 [1] が騒がれ、その代表であるブログや 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ー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野が狭くなり、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ー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。図1にコンテンツホール検索のイメージ図を示す。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。これらと比較し、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ーに提示する。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。

以下、2章ではコンテンツホール検索システムの基本コンセプトを、3章では Web ページからの視点抽出手法を、4章では視点間の比較手法について述べ、5章でまとめと今後の課題について述べる。

2. コンテンツホール検索の基本コンセプト

本研究では、SNS やブログ等のコミュニティ型コンテンツにおいて、見落とされている情報であるコンテンツホールをそのコンテンツの視点情報に基づき抽出し提示するコンテンツホール検索を提案する。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ーに提示
抽出したコンテンツホールをコミュニティ内のユーザーに提示するユーザーインター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。

- ユーザーはコミュニティ内の議論において不足している内容を把握でき、これにより、より公平性のある議論を行うことが出来る。
- ユーザー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 [2] 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 [3] や、話題構造を抽出する手法 [4] [5] 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「鼻の

長い象」や「オムライスが美味しいレストラン」などのような「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いるかを確認する。

(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

フレーズ検索

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。

フィルタリング

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^(注1)を使用した。にて「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されていることが分かった。

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。

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4にある素性を抽出する。尚、形態素解析には JUMAN^(注2)を文節区切りには KNP^(注3)を利用した。

(注1): Yahoo!検索 Web サービス

<http://developer.yahoo.co.jp/search/>

(注2): JUMAN

<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>

(注3): KNP

<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html>

文節の語形・主辞の定義は、文献[6]にならう。また語の概念は、日本語語彙大系[7]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。

予備実験 1. 一般名詞

まず、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。実験に用いたデータは、ラベル4を除いたもので、ラベル1,2,3の3値分類となる。機械学習アルゴリズムにはSVM^(注4)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。

実験の結果を表4に示す。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
20 内装	18 夜景 素敵だ
17 居心地	14 雰囲気 素敵だ
16 評判	14 景色 いい
16 ビール	13 眺め 良い
15 種類	11 敷居 高い
14 値段	11 眺望 素晴らしい
13 窓	11 値段 高い
13 気持ち	11 種類 豊富だ
13 シーフード	10 料理 有名だ
12 敷居	10 料理 豊富だ
12 ピザ	10 雰囲気 よい
12 サービス	10 眺め 美しい
11 展望	10 眺め 素晴らしい
11 天井	10 インテリア 印象的だ
11 建物	9 景色 良い
11 外観	8 評判 良い
10 ワイン	8 ピザ 美味しい
9 料金	7 夜景 素晴らしい
9 見晴らし	7 天井 高い
9 魚介類	7 景色 素晴らしい
9 飲茶	7 居心地 良い
8 夕日	7 気持ち いい
8 評価	7 管理 必要だ

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

表4 SVMによる「レストラン」の分類精度

Table 4 Accuracy of Classification of "Restaurant" by Using SVM

正解率	0.945
ラベル1の精度	0.955
ラベル1の再現率	0.979

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。もちろん対象名詞「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5に掲載する。また、それぞれの対象名詞について得られた属性名詞集合のリストを表A.1に掲載する。

表5 他の対象名詞における実験結果

Table 5 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%

(注4): SVMの実装には TinySVM

<http://chasen.org/~taku/software/TinySVM/> を用いた

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。

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 6 に収集した視点構造（名詞 A + 形容詞）の全体 75 件の内、上位 20 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 20 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。

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」がテーマである名詞 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ためである。

(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 2 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。

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。(1) では「名詞 A + 形容詞 + 名詞 B」の構造に注目し、これらの構造からテーマである名詞 B の視点構造「名詞 A + 形容詞」のペアを抽出した。(2) ではこの抽出されたペアを用いてグラフ構造を作成し、同一テーマの Web 空間全体での視点構造グラフとコミュニティでの視点構造グラフを比較しその差分情報であるコンテンツホールを抽出する方法を提案した。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と考えられる。今後の課題は以下の通りである。

- コミュニティ型コンテンツからテーマの自動抽出手法の提案

表 6 Wii に対しての視点構造

Table 6 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
5 発売	3 評価 高い
3 本体	3 発売 間近だ
3 評価	3 謎 多い
3 謎	3 値段 安い
3 値段	3 サード 協力
3 価格	3 イメージ 強い
3 印象	2 歴史 長い
3 タイトル	2 裏技 素っ気無い
3 サード	2 様子 無い
3 イメージ	2 品切れ 多い

