コミュニティ型コンテンツのコンテンツホール検索の提案

著者	灘本 明代, 荒牧 英治, 阿辺川 武, 村上 陽平
雑誌名	甲南大学紀要. 知能情報学編
巻	1
号	2
ページ	249-268
発行年	2008-12-20
URL	http://doi.org/10.14990/00001274

解説論文

コミュニティ型コンテンツのコンテンツホール検索の提案

灘本明代[a], 荒牧英治[b]
阿辺川武[c], 村上陽平[d]

[a] 甲南大学 知能情報学部 知能情報学科
神戸市東灘区岡本８－９－１, 658-8501
[b] 東京大学　知の構造化センター
東京都文京区本郷７－３－１, 113-8655
[c] 東京大学大学院 教育学研究科
東京都文京区本郷７－３－１, 113-8655
[d] 独立行政法人　情報通信研究機構
京都府相楽郡精華町光台３－５, 619-0237

（受理日 2008 年 11 月 3 日）

概 要

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．本論文では，コンテンツホールを定義し，コンテンツホール検索の第一歩となる (1) Web 空間における視点情報の抽出, (2) コミュニティ型コンテンツの視点抽出のための対話解析, (3) プロトタイプシステムの提案を行う．具体的には (1) では「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」であると定義し，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．(2) ではコミュニティ型コンテンツの二つのコメントの関係に注目し，コメント間の内容が関連している内容的関連性と，コメント間が応答関係になっている機能的関連性とを提案する．これにより，コミュニティ型コンテンツから複数の話題を抽出し，その話題毎の視点構造を抽出することが可能となる．(3) では Wikipedia との比較によるコンテンツホール抽出プロトタイプシステムを提案する．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: コンテンツホール，検索，コミュニティ型コンテンツ

1 はじめに

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，議論のテーマに対する全体像が見えなくなってしまう危険性がある．そこで本論文では，コミュニティ内で議論されていない重要な情報を検索し提示することを行う．従来の情報検索はユーザが求めている情報を探す類似検索が主流であるが，本論文ではコミュニティ型コンテンツにおいてコミュニティ内で気づいていない情報つまりは「ないものを探す」相違検索を目的とする．「コミュニティ内で気づいていない情報」をコンテンツホールと呼ぶ．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，以下の手順でコンテンツホール検索を行う．

(1) コミュニティ型コンテンツからのテーマの抽出

(2) Web 空間における視点情報の抽出

(3) コミュニティ型コンテンツの視点抽出

(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

(5) 抽出されたコンテンツホールの提示

現在の Web 検索はユーザの入力したキーワードを用いる情報検索が主流である．また，情報検索の研究分野でもキーワード検索に基づく研究が多い．近年の情報検索の研究では自然言語入力による検索手法やサンプル・コンテンツから Query-Free [1] による検索手法等の提案も行われているが，これらはすべてユーザがほしい情報を検索するのが目的である．このように現在の情報検索の技術ではユーザが気付いていない情報の検索が行えないのが現状である．また，ユーザが閲覧している情報に関連する詳細情報やより話題の広い情報を検索する情報補完に関する提案 [2] がされているが，これらはユーザが閲覧している情報に関連する情報を検索する研究であり，我々の提案する「気付いていない情報を探す」コンテンツホール検索とは異なる．本論文ではコンテンツ群の視点に注目し，コミュニティの視点と Web 空間の視点を比較することにより，コンテンツホールを抽出することを試みた．

本論文では 7 つのコミュニティ型コンテンツを提案すると共に，(2) を提案し，(3) の基盤技術となるコミュニティ型コンテンツの対話解析を提案した後，コンテンツホール検索システムのプロトタイプを紹介する．

2 コンテンツホールの種類

コンテンツホールはユーザの気づいていない情報であり，ユーザの発言の周辺の情報や反対の情報等様々な情報の種類が考えられる．図 2 に我々の考えるコンテンツホールのイメージ図を示す．ここで，コミュニティ型コンテンツはコミュニティが対象としているテーマ T とコミュニティ参加者であるユーザの発言 $C_i (i = 1, 2, \ldots, j)$ の集合からなる話題 $Sub_n (n = 1, 2, \ldots, m)$ で構成されているとする．話題 Sub_n とはテーマ内の同一話題を共有する発言の集合とし，例えば掲示板の場合は一つのス

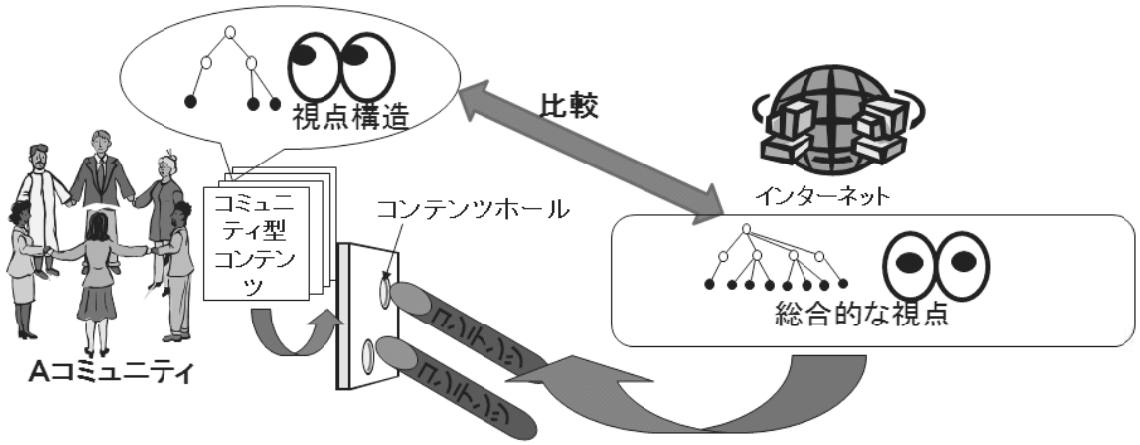

図 1: コンテンツホール検索のイメージ図

レッドを一つの話題とする．一つのテーマは複数の話題から構成される場合が多く，一つの話題は複数の発言から構成される場合が多いが，一つの発言のみであってもかまわないとする．図 2 ではコンテンツホールの種類をわかりやすく表現するために，一つの話題のみから構成されるコミュニティ型コンテンツの例を示している．図 2 に示すようにコンテンツホールは，周辺の話題や詳細な話題等コミュニティの話題に類似するものと，反対の印象の話題等の話題と相違するものに分けられると考える．尚，本論文ではコンテンツの時系列を考えないコンテンツホールの提案を行う．以下に各々のコンテンツホールの種類の定義をテーマ T をオリンピックの陸上競技とし，コミュニティ型コンテンツの話題 Sub_0 を陸上競技の選手であるウサイン・ボルト選手を話題とした例を用いて示す．

2.1 コミュニティの話題と類似するコンテンツホール

- 話題の内部
  コミュニティ型コンテンツ内の一つの話題に対して，抜け落ちている情報を話題の内部のコンテンツホールと呼ぶ．つまりは，Sub_n と Sub_0 は同一であるが，C_i と異なるコンテンツを示す．例えば，ボルトの好きな食べ物の話をしているコミュニティの場合，チキンナゲットが好きだという情報がなければ，そのチキンナゲットが好きだという情報が話題の内部のコンテンツホールとなる．

- 周辺の話題
  コミュニティ型コンテンツ内の一つの話題に対して，少し広い意味を持つ話題を周辺の話題のコンテンツホールと呼ぶ．つまりは Sub_n は Sub_0 と包含関係にあるコンテンツを示す．例えば，ボルトと彼の父親との関係を示す情報は周辺の話題のコンテンツホールとなる．

- 詳細な話題
  コミュニティ型コンテンツ内の一つの話題に対して，発言内容より詳しい情報を詳細な話題のコンテンツホールと呼ぶ．つまりは Sub_n と Sub_0 は同一であり，C_i より詳細なコンテンツ

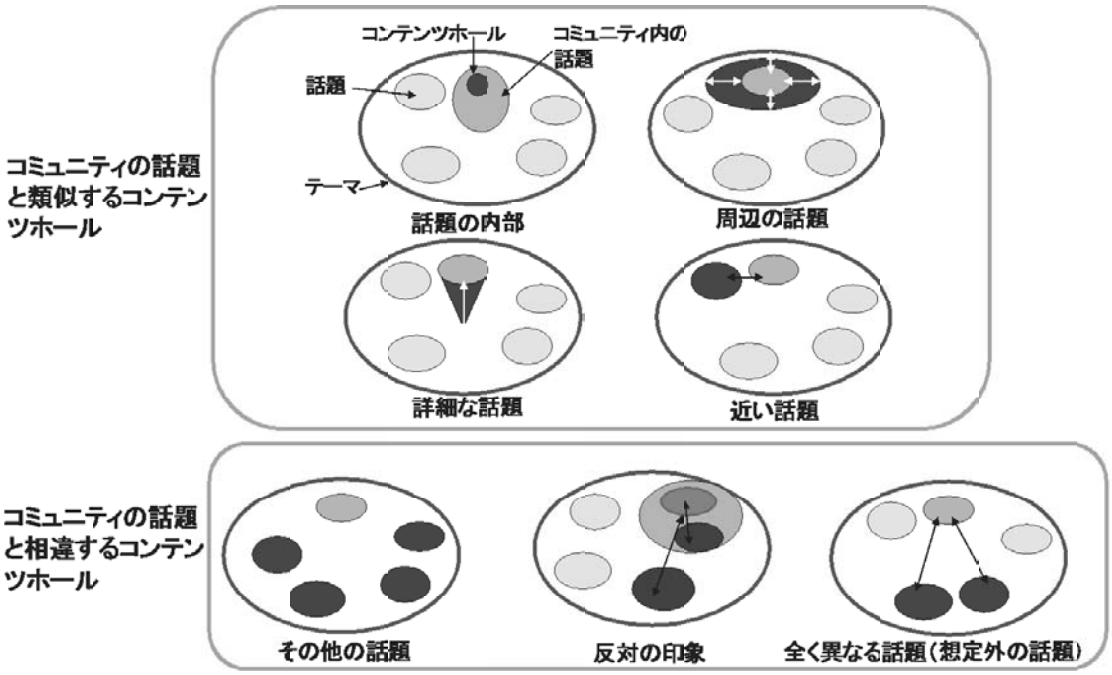

図 2: コンテンツホールの種類

を示す．ボルトが優勝した発言をしている場合，ボルトがどのような練習をして優勝したのかの情報は詳細な話題のコンテンツホールとなる．

- 近い話題
  コミュニティ型コンテンツ内の一つの話題と類似するが少し異なる話題を近い話題のコンテンツホールと呼ぶ．つまりは Sub_n と類似するが少し異なる話題を示す．例えば，他のジャマイカの選手の話題は近い話題のコンテンツホールとなる．

2.2 コミュニティの話題と相違するコンテンツホール

- その他の話題
  コミュニティ型コンテンツ内の話題と異なる話題すべてをその他の話題のコンテンツホールと呼ぶ．つまりは，T は同じであるが，Sub_n 以外のコンテンツすべてを示す．例えば，ボルト以外のすべての選手の話題はその他の話題のコンテンツホールとなる．例からもわかるように，その他の話題と近い話題のコンテンツホールは包含関係になっている．

- 反対の印象
  コミュニティ型コンテンツ内の話題がもつ印象と異なる印象を持つ話題，もしくは発言と反対の印象を持つ発言を反対の印象のコンテンツホールと呼ぶ．つまりは，Sub_0 の印象と異なる印象を持つ話題を指す場合と，Sub_n と Sub_0 は同じだが C_i と異なる印象を持つ場合とがある．例えば，ボルトの 100 m の最後の走りに対して最後まで全力で走ればと残念がる発言に対して，ふざけているといったような反対の印象を述べる情報は反対の印象のコンテンツホールとなる．

- 全く異なる話題（想定外の話題）
  コミュニティ型コンテンツ内の話題と全く異なる話題を想定外の話題のコンテンツホールと呼ぶ．その他の話題のコンテンツホールとは包含関係にあるが，その他の話題のコンテンツホールがある程度類似している話題も含まれるのに対し，想定外の話題はコンテンツ間の相違度が大きい話題を対象とする．例えば，200 m 決勝において 2 位 3 位の選手が実は失格になっていたというのは想定外の話題のコンテンツホールとなる．

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 [3] 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 [4] や，話題構造を抽出する手法 [5], [6] 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．本論文では，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 5,545 個に対してフィルタリング後の結果は 2,512 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 1,715，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．

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[1] を文節区切りには KNP[2] を利用した．文節の語形・主辞の定義は，文献 [7] にならう．また語の概念は，日本語語彙大系 [8] 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[3] 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．

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

表 1: 収集した属性名詞の例

名詞 A		名詞＋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

[1] JUMAN
http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[2] KNP
http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[3] SVM の実装には TinySVM
http://chasen.org/~taku/software/TinySVM/ を用いた

表 2: 素性を抽出する文節

	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

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

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

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．

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 5: 他の対象名詞における実験結果

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 %

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 コミュニティ型コンテンツの対話解析

SNS や Blog などのコミュニティ型コンテンツは通常の Web コンテンツと異なり，対話形式によるコンテンツが殆どである．そこで，コミュニティ型コンテンツの視点構造とはこれらコミュニティ型コンテンツの著者達の視点構造であり，彼らの対話を解析することにより，より実空間に近い視点構造が生成できると考え，コミュニティ型コンテンツの対話解析を行った．

実際には，表 7 に示す BBS の書き込み（以降，**コメント**）の例のようにコミュニティ型コンテンツでは，複数人が複数の話題について同時に議論するため，対応する 2 つのコメント同士の間にギャッ

表 6: Wii に対しての視点構造

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

表 7: BBS 書き込みの例

(1)	何を使ってる？
(2)	バッテリーがまだ残っているのに，ipod が止まってしまいます.
(3)	初代.
(4)	nano とシャッフル.
(5)	バッテリー表示は近似だからですよ.

プが存在している．このギャップはコミュニティ型コンテンツにおいて，頻繁に見られるもので，図 3 が示すように，ほぼ半数の対応するコメント間にギャップが存在している．そこで，本論文では，二つのコメント間が対応しているかどうかを識別する手法を提案する．これにより，コミュニティ型コンテンツの話題を抽出し，それに基づいてコミュニティ型コンテンツの視点構造を抽出する事が可能になる．

我々は，コミュニティ型コンテンツの任意の 2 つのコメントに関連性がある場合，以下の 2 種類の対話モデルがあると考える．一つは，2 つのコメントの内容的に類似している場合で，本論文ではこれを**内容的関連性**と呼ぶ．もう一つは，2 つのコメントが応答関係になっている機能的な対応である．例えば,「なぜ...」といったコメントに対する応答は「... だから」というコメントで応答することが考えられる．このようなコメント間で対応する表現を本論文では**対応ペア**と呼び，対応ペアによる手がかりを**機能的関連性**と呼ぶ．前者，**内容的関連性**は，文同士の類似度と近い概念である．我々は Web 上での単語の共起にもとづく類似度 [9] で，これを計算する．後者，**機能的関連性**を得るためには，**対応ペア**を得る必要がある．我々は，大量の対話コーパスを用いて，対応ペアを自動収集することを考える．大量の対話コーパスは，Web をクロールして，得た掲示板の中で確からしい部分だけを使って得た．

ここで，2 つの対話モデルを求めるために，まず，タスクを定式化する:

入力: ある BBS 内の 2 つのコメント（i 番目のコメントと j 番目のコメント ($j > i$)）.

出力: True または False（もし 2 つのコメントが対応しているならば True，そうでないなら False）.

本論文では，i 番目のコメントを P，j 番目のコメントを Q とする.

4.1 内容的関連性

2 つのコメントが内容的に関連している場合を内容的関連性と呼び，2 つのコメント（文）の類似度をもとめ，類似している文同士は内容的関連性が高いとする．これまで文同士の類似度または関連性を得る手法は数多く提案されている [10]．我々は，Web 上での単語の共起頻度にもとづいた単語類似度 (WEBPMI) を利用し，文同士の類似度 ($sim_r(P, Q)$) を求める.

$$sim_r(P, Q) = \sum_{p \in P} \max_{q \in Q} \mathrm{WEBPMI}(p, q), \quad (1)$$

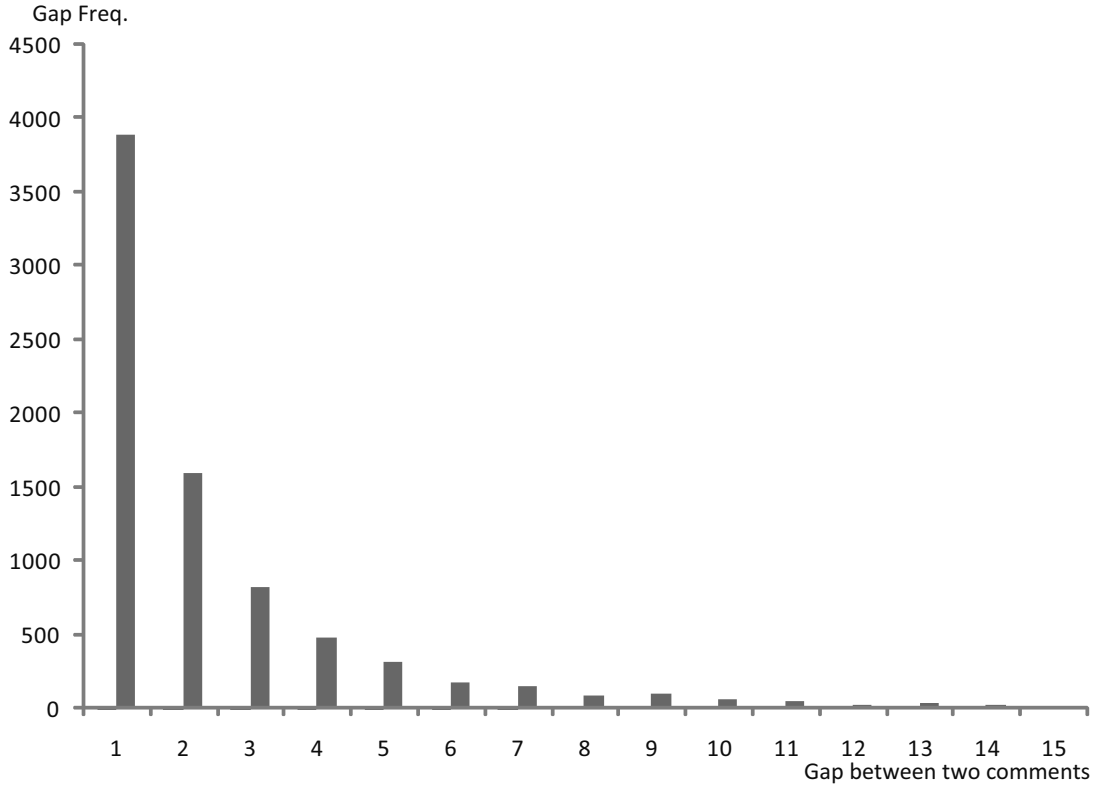

図 3: 対応するコメント間のギャップの長さとその頻度

ここで，p は P に含まれる語の集合，q は Q に含まれる語の集合であり，WEBPMI は次の式によって定義される：

$$\begin{aligned}&\text{WEBPMI}(p,q)\\&=\begin{cases}0, & \text{if } H(p\cap q)\le c,\\ \log\frac{\frac{H(p\cap q)}{N}}{\frac{H(p)}{N}\frac{H(q)}{N}}, & \text{otherwise,}\end{cases}\end{aligned}\tag{2}$$

ここで，$H(p)$ はクエリ「p」によって検索エンジンが返す文書数であり，$H(q)$ はクエリ「q」によって検索エンジンが返す文書数，$H(p\cap q)$ は「$p+q$」によって検索エンジンが返す文書数，N は検索エンジンが持つ文書数である．小さな値によるノイズを避けるため，閾値 c よりも小さいものはフィルターした[4]．

4.2 機能的関連性

2 つのコメントが応答関係になっている場合を機能的関連性と呼ぶ．機能的関連性を求めるために，Corresponding-PMI (CPMI) を定義する．これは WEBPMI と似ているが，以下の 2 つの点が異なる：

[4]先行研究 [9] にもとづいて $c=5$ とした．

(1) WEBPMI は Web での共起頻度を用いるが，CPMI は対応するコメント間での共起頻度を用いる．

(2) WEBPMI は一語しか扱わないが，CPMI は n-gram を扱う ($n = 1, 2, 3$).

CPMI を計算するために，まず，コメントペアである P と Q を用いて以下の 3 つのデータベースを構築する．

DB-A: P が生じる N-gram のデータベース．

DB-B: Q が生じる N-gram のデータベース．

DB-C: N-gram のペアの許容値のデータベース (possible n-grams in P: possible m-grams in Q) occurrences ($1 \leq n \leq 3, 1 \leq m \leq 3$).

例えば以下のような 2 つのコメントがあると

P: *How about a bus?*

Q: *Nice idea*

我々は以下の 3 つの 2 グラムのペアを取得することができる．(1) "*how about: nice idea*", (2) "*about a: nice idea*" and (3) "*a bus: nice idea*".

機能的関連性 $sim_d(P, Q)$ は以下の通りである．

$$sim_d(P, Q) = \sum_{p \in N_P} \max \sum_{q \in N_Q} CPMI(p, q), \tag{3}$$

ここで，N_P は，P に含まれる N-gram の集合，N_Q は Q に含まれる N-gram の集合，CPMI は次式によって定義される:

$$CPMI(p, q) = \begin{cases} 0, & \text{if } H_c(p \cap q) \leq c, \\ \log \frac{\frac{H_c(p \cap q)}{M}}{\frac{H_a(p)}{M} \frac{H_b(q)}{M}}, & \text{otherwise,} \end{cases} \tag{4}$$

ここで，$H_a(p)$ は N-gram p の頻度 P における出現数を，$H_b(q)$ は N-gram q の頻度 Q における出現数，$H_c(p \cap q)$ は N-gram 対（$p : q$) の出現頻度を，M は検索結果の文書数を示す．

4.3 対話解析の実験

4.3.1 テストセットの構築

テストの構築にあたっては，それぞれ [13] で述べたコメントペアコレクションから一定数のコメントペアを無作為抽出し，それらの応答部分 (Q) を，同 BBS 内の他の応答と入れ替えることによって構築した．実験では次の 2 つのテストセットを用いた．

表 8: Results in SMALL-SET

	Accuracy	Precision	Recall	$F_{\beta=1}$
human-A	79.28	83.33	75.34	79.13
human-B	75.71	78.26	73.97	76.05
human-C	70.71	71.62	72.6	72.10
$Overlap$	61.42	58.71	87.67	70.32
sim_r	61.42	72.09	42.46	53.44
sim_d	65.71	66.23	69.86	67.99

SMALL-SET: 人間も参加する小規模なデータ，140 コメントペアからなる．

LARGE-SET: コーパスサイズと機能的関連性の精度の関連性を調べるために用いる大規模データ．8,400 コメントペアからなる．

4.3.2 比較手法

次の手法を比較した．

$human$-A, B, and C: 人間（３人）による判定結果．

$Overlap$: 語の一致率による精度（ベースライン）．
語の一致率が閾値より高ければ TRUE を出力し; そうでなければ FALSE を出力する．

sim_r: sim_r の値が閾値より高ければ TRUE を出力し; そうでなければ FALSE を出力する．

sim_d: sim_d の値が閾値より高ければ TRUE を出力し; そうでなければ FALSE を出力する．

sim_r における WEBPMI の計算にあたっては正確なドキュメント数を得るために TSUBAKI [14] を用いた．

4.3.3 SMALL-SET の結果

表 8 に SMALL-SET での各手法の精度を示す．$Overlap$, sim_r と sim_d の精度は閾値に依存するため，様々な閾値で実験し，もっとも高い accuracy を示した点の精度を掲載する．表 9 に各手法による出力同士の一致率を示す．

4.3.4 人間の精度

人間の精度はたかだか 70–79%しかなく，本タスクの難しさを示している．これは多くのコメントに対する返答として成立する短い返答（「そう思います」や「ありがとうございます」など）による

表 9: 人間とシステム間の一致率と Kappa 値

	Human-A	Human-B	Human-C	$Overlap$	sim_r	sim_d
Human-A	-	0.78 (0.56)⊕	0.74 (0.49)⊕	0.52 (0.08)⊖	0.60 (0.20)	0.65 (0.28)
Human-B		-	0.73 (0.47)⊕	0.54 (0.09)⊖	0.60 (0.21)	0.62 (0.25)
Human-C			-	0.59 (0.15)⊖	0.52 (0.05)⊖	0.62 (0.25)
$Overlap$				-	0.63 (0.21)	0.45 (0.13)⊖
sim_r					-	0.56 (0.16)⊖
sim_d						-

*括弧内の数字は κ 値を示す．⊖ は κ 値の解釈が「slight」であることを示す．⊕ は κ 値の解釈が「moderate」であることを示す．

表 10: 高い $CPHI$ を持つ対応ペアの例

N-gram in P	N-gram in Q	$CPHI$
行き ます	お 待ち して	8.43
どこ に ある	あり ます	8.37
はじめ まして	はじめ まして	7.86
教えて ください	と 思い ますよ	7.62
いかがでしょう	早速	7.47
でき ます	やって み	7.38
と 思い ます	あり が とう	7.12
か な？	多分	6.93
あり が とう	いえいえ	6.80
私 は	私 も	6.73
か？	と 思い ます	6.72

False Positive，専門的すぎて評価者では判断できない返答による false negative が原因である．このような限界はあるものの人間同士の一致度は表 9 に示されるように高く（κ value = moderate），これらの限界は評価者間で一致しているものと考えられる．以上から，本タスクは難しいものの不合理ではないと言える．

4.3.5 二つの関連性の一致度

表 8 に示されるように, sim_d は sim_r や $Overlap$ よりも高い精度を示した．より重要なことは, $Overlap$ と sim_r はわずかに相関しているが（fair agreement; $0.2 < \kappa < 0.4$），これらの両方とも sim_r に対してわずかな相関しかみせていないことである (slight agreement; $\kappa < 0.2$)．これらの結果から，sim_r (or $Overlap$) と sim_r は互いに独立な手がかりを用いていることが推測される．表 10 に高い $CPHI$ を持つ対応ペアの例を示す．表にみられるように，これらの関連性を Web 文章での共起でとらえることは困難だと想像され，sim_d の有効性を示している．

4.3.6 LARGE-SET Results

図 4 に対応ペアを計算するために用いるコメントの数と sim_d の関係を示す．図 4 において精度はまだ飽和しておらず，今後，より大規模なデータを用いることで，さらに高い精度が期待される．

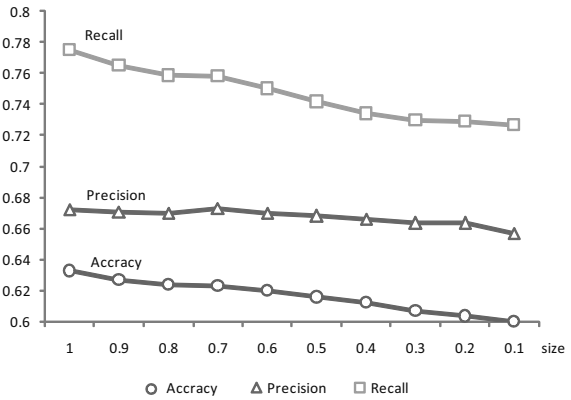

図 4: トレーニングセットのサイズと sim_d の精度

5 プロトタイプシステム

2 章で提案したコンテンツホールの種類の内，コミュニティの話題と相違するコンテンツホールの「その他の話題のコンテンツホール」を Wikipedia を用いて提示するプロトタイプシステムを開発した．Wikipedia に関連する研究は多数あるが, 中山 [15], Suchanek [16], Wu [17], Gabrilovich [18] らに代表されるように Wikipedia から知識を抽出し利用する研究が数多くある．これらの研究は Wikipedia のカテゴリ構造やリンク構造を用いて知識を抽出しているのに対し，本論文では Wikipedia から知識を抽出するのではなく，Wikipedia の記事の目次構造をその記事つまりはコミュニティ型コンテンツのテーマの属性情報である視点構造とみなす．また，川場ら citekawaba はあるトピックに有用なブログサイトを検索する応用例として Wikipedia を用い，Wikipedi の記事に対応したトピックのブログサイトを検索している．堀ら [20] はユーザのクエリからその意図に沿った拡張クエリを作成する際に Wikipedia を用いるシステムを提案している．これらの研究はクエリの拡張に対して Wikipedia を用いているが，本論文ではクエリを記事名とし，その記事の目次からコンテンツホールを求めることを行う．図 5 にプロトタイプシステムの検索画面を，また図 5 にコンテンツホールの表示画面を示す．コミュニティ型コンテンツの一つの話題を構成するコンテンツ群と Wikipedia の一つの記事を構成する目次を構成する最小の項目毎を比較することを行う．ここでは，各々の文書において形態素解析を行い，そこに含まれる名詞において TF/IDF 法により単語の重みを求め，それを用いてコサイン相関値により文書間の類似度を求める．ここでいう文書間とは，Wikipedia の目次の最小単位が指し示すコンテンツとコミュニティ型コンテンツ全体との文書間である．類似度がある閾値より小さいものをコンテンツホールの候補とする．

プロトタイプシステムのフローを以下に示す．

図 5: プロトタイプシステム初期画面図

1. ユーザは比較したいコミュニティのテーマをキーワードとして入力する．
2. ユーザの入力したキーワードからそのキーワードのコミュニティのサイトのリストと Wikipedia のページを検索し，コミュニティサイトを右画面に，Wikipedia を左画面に表示する．
3. ユーザは (b) で表示されたコミュニティサイトのリストからコンテンツホールを見つけたいサイトを選択する．
4. システムはユーザが指定したコミュニティのサイトの 1 テーマのコンテンツと (b) で検索した Wikipedia を Wikipedia の目次毎に比較し，類似していない目次のコンテンツをコンテンツホールとする．
5. Wikipedia の目次の階層構造を利用して，コンテンツホールを赤字で表示する（図 6 左画面参照）．

プロトタイプシステムの実験

種々のコミュニティのテーマを用いて提案手法の有用性を計り，提案手法の問題点を抽出する実験を行った．実験に用いたコミュニティのテーマは以下の方針で選んだ．

- 速報性の強い情報をテーマとしている場合
  スポーツ等のコミュニティの中では，その日にあった試合に対しての議論を行っている場合がある．このような速報性のある話題に対しては，Wikipedia の記事は速報性が弱いため有用ではないと予測されるが，その問題点は何かを抽出する．
- 固有名詞のうち，組織名と個人名との比較
  固有名詞をコミュニティ型コンテンツのテーマとしている場合，その固有名詞は有名な組織

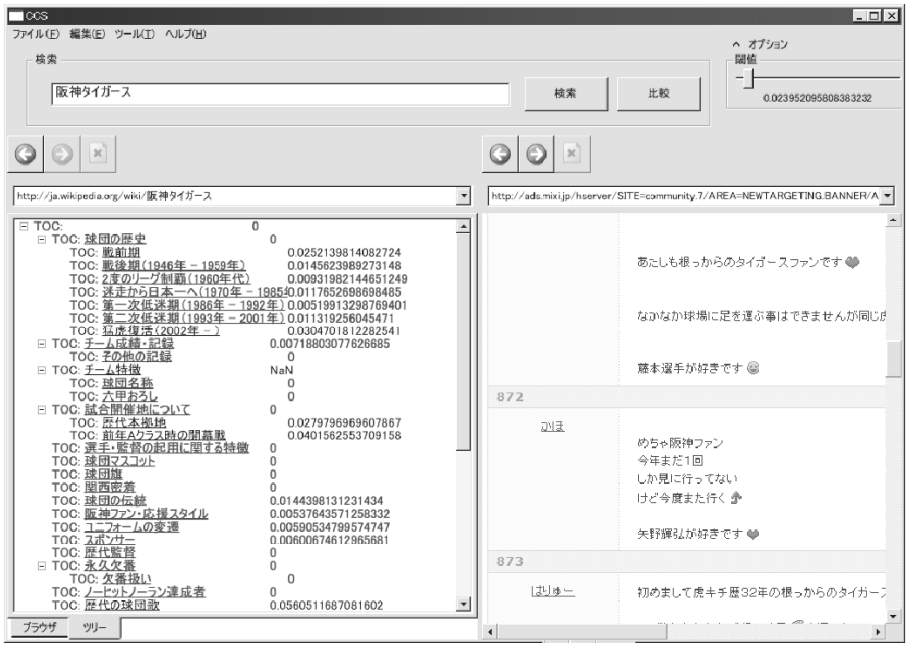

図 6: プロトタイプシステム解析結果画面図

や個人である場合がほとんどであり，Wikipedia に掲載されている可能性が高い．そこで，話題が広義な会社や団体等の組織名とそれと比較して話題が狭義な個人名とを比較する．組織名と個人名とで個人名の方が狭義の話題になり，Wikipedia を用いることが有効であると予測する．

- 上位概念，下位概念の比較
  テーマの構造が上位概念の場合とその下位概念の場合とを比較する．つまりは提案手法は Wikipedia を用いているため，よりインスタンスに近いテーマの方が有効であると予測するが，その比較実験を行う．例えば，JAL がテーマの場合，JAL をテーマとしているコミュニティとその下位概念であると考えられる JAL のサービスの一部であるマイレージ（JAL マイレージバンク）をテーマとしているコミュニティにおいて，どちらが Wikipedia を用いることにより有用なのかを比較する．

実際に実験に使用したテーマと実験結果の類似度が閾値以上とされた項目，つまりはコンテンツホールではないと判断された適合率を表 11 に示す．ここで求めた適合率は以下の通りである．

$$\text{適合率} = \frac{\text{類似度が閾値以上の目次項目の内正解の項目数}}{\text{類似度が閾値以上の目次項目数}} \times 100$$

6 おわりに

本論文では，これまで我々が提案してきた，コミュニティ型コンテンツにおいてユーザが気付いていない情報であるコンテンツホールの 7 つの種類を提案した．具体的には，7 種類のコンテンツホー

表 11: 評価実験に用いたテーマとその結果

対象テーマの説明	対象とするテーマ名（クエリ）	適合率 (%)
速報的な最新情報		
オリンピック	柔道	24
オリンピック	北京オリンピック野球日本代表	38
組織名と個人名		
組織名	JAL	36
組織名	阪神タイガース	28
個人名	柴崎コウ	35
個人名	金本知憲	61
上位クラスと下位クラス		
上位クラス	JAL	36
下位クラス	JAL マイレージバンク	81
上位クラス	ドコモ	38
下位クラス	ドコモダケ	61

ルをコミュニティ型コンテンツの話題に注目し，コミュニティの話題と類似するコンテンツホールとコミュニティの話題と相違するコンテンツホールに分類した．さらに，コンテンツホール検索の第一歩となる (1) Web 空間における視点情報の抽出, (2) コミュニティ型コンテンツの視点抽出のための対話解析, (3) プロトタイプシステムの提案を行った．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において独特な言葉や新語に対応する．
- 速報性のある話題への対応．
- コミュニティ型コンテンツの話題の下位概念に対応する単語（記事）の Wikipedia の項目を何処まで入れてコンテンツホールを抽出するかを検討．
- 「オリンピックの陸上 200m」といったようにコミュニティ型コンテンツでは Wikipedia において複数の記事に対応するテーマを対象としている場合が多いため，これらに対応．
- 本論文では提案した 7 つのコンテンツホールの内，その他の話題のコンテンツホールのプロトタイプシステムを開発したが，他の 6 つのコンテンツホールを抽出するシステムの提案．

謝辞

本論文の一部は，平成 20 年度科研費特定領研究域「Web2.0 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) と文部科学省 ORC 整備事業 (2004-2008) による．ここに記して謝意を表します．

参考文献

[1] Henzinger Monika, Chang Bay-Wei, Milch Brian and Brin Sergey, “Query-Free news search,” *World Wide Web Journal, Springer Science+Business Media B.V., ISSN: 1573-1413*, pp. 101-126, 2005.

[2] Ma Qiang, Akiyo Nadamoto and Katsumi Tanaka, “Complementary information retrieval for cross-media news content,” *Elsevier ARTICLE Information Systems,* vol. 31, iss. 7, pp. 659-678, 2006.

[3] 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, pp. 43-48, 2006.

[4] 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, pp. 08-01-08-08, 2003.

[5] M. Spitters and W. Kraaij, “A language modeling approach to tracking news rvents,” in *Proc. TDT 2002 Evaluation workshop*, pp. 101-106, 2002.

[6] 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*, pp. 347-367, 2005.

[7] 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, vol. 7, no. 5, pp. 3-17, 2000.

[8] 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, vol. 8, no. 1, pp. 143-173, 2001.

[9] Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka, “Measuring semantic similarity between words using web search engines,” in *Proc. 16th International World Wide Web Conference (WWW 2007)*, pp. 757-766, 2007.

[10] Marco De Boni and Suresh Manandhar, “An analysis of clarification dialogue for question answering,” in *Proc. the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003)*, pp. 48-55, 2003.

[11] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema , Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin and Marie Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Lingustics*, vol. 26, no. 3, pp. 340-373, 2000.

[12] Carlson, D. Marcu and M. E. Okurowski, “RST discourse treebank,” *Linguistic Data Consortium*, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07, 2002.

[13] 荒牧英治, 阿辺川武, 村上陽平, 灘本明代, “BBS 対話における発話間の応答関係の判定,” 言語処理学会 第 14 回年次大会, http://nlp2008.anlp.jp/program.htmlA1-5, 2008.

[14] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi, “TSUBAKI: an open search engine infrastructure for developing new information access methodology,” in *Proc. International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189-196, 2008.

[15] 中山浩太郎，原隆浩，西尾章治郎, “自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法,” 電子情報通信学会データ工学ワークショップ (DEWS’08) 論文集, http://www.ieice.org/ de/DEWS/DEWS2008/proceedings/, 2008.

[16] F.M. Suchanek, G. Kasneci and G. Weikum, “YAGO: a core of semantic knowledge unifying wordnet and wikipedia,” in *Proc. the 16th International World Wide Web Conference (WWW2007)*, pp. 697-706, 2007.

[17] Fei Wu and Daniel S. Weld, “Automatically refining the wikipedia infobox ontology,” in *Proc. the 17th International World Wide Web Conference (WWW2008)*, pp. 365-644, 2008.

[18] Evgeniy Gabrilovich and Shaul Markovitch, “Computing semantic relatedness using wikipedia-based explicit semantic analysis,” in *Proc. the International Joint Conference on Artificial Intelligence 2007 (IJCAI 2007)*, pp. 1606-1611, 2007.

[19] 川場真理子，中崎寛之，宇津呂武仁，福原知宏, “Wikipedia エントリとブログサイトの対応付けのための特定トピックのブログサイト検索,” 電子情報通信学会データ工学ワークショップ (DEWS’08) 論文集, http://www.ieice.org/ de/DEWS/DEWS2008/proceedings/, 2008.

[20] 堀憲太郎，大石哲也，長谷川隆三，藤田博，峯恒憲，越村三幸, “Wikipedia への関連単語抽出アルゴリズムの適用とその評価,” 情報処理学会研究報告, vol. 2008, no. 56, 2008-DBS-145, pp. 81-88, 2008.