社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (44)
2008／9／23

Wikipediaを用いたコンテンツホール検索の提案

灘本 明代† 荒牧 英治†† 阿辺川 武††† 村上 陽平††††

† 甲南大学 〒658–8501 兵庫県神戸市東灘区岡本8–9–1
†† 東京大学 知の構造化センター 〒113–8655 東京都文京区本郷7–3–1
††† 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
†††† 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
E-mail: †nadamoto@konan-u.ac.jp, ††aramaki@gmail.com, †††abekawa@p.u-tokyo.ac.jp, ††††yohei@nict.go.jp

あらまし Blog やSNS等のコミュニティ型コンテンツにおいて，ユーザが気付かずに抜け落ちている情報を我々はコンテンツホールと呼び，このユーザが気づいていない情報を探すことをコンテンツホール検索と呼ぶ．コンテンツホール検索は様々な種類のコンテンツホールが考えられる．そこで本論文では，コンテンツホールの種類を提案しそれらを定義する．また，これまで我々はコミュニティ型コンテンツの視点情報とWeb空間との視点情報を比較することによりコンテンツホール検索を行うことを提案してきた．本論文では，Web空間の視点構造としてWikipediaの記事の目次構造を用い，コミュニティ型コンテンツとWikipediaの情報と比較しコンテンツホールを検索するシステムの提案を行う．
キーワード コンテンツホール，検索，Wikipedia

Content Hole Search System by Using Table of Contents of Wilipedia

Akiyo NADAMOTO†, Eiji ARAMAKI††, Takeshi ABEKAWA†††, and Yohei MURAKAMI††††

† Konan University Okamoto 8–9–1, Higashinada-ku, Kobe city, Hyogo, 658–8501 Japan
†† The University of Tokyo Center for Knowledge Structuring, Hongou 7–3–1, Bunkyo-ku, Tokyo, 113–8655 Japan
††† The University of Tokyo Graduate School of Education Hongou 7–3–1, Bunkyo-ku, Tokyo, 113–8655 Japan
†††† National Institute of Information and Communications Technology Hikaridai 3–5, Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan
E-mail: †nadamoto@konan-u.ac.jp, ††aramaki@gmail.com, †††abekawa@p.u-tokyo.ac.jp, ††††yohei@nict.go.jp

Abstract In the community type contents such as Blog and SNS, We call the user's unawareness information "the content hole" and search for it "the content hole search". The content hole search is different from a current similarity searching and it has a variety type. In this paper, we propose the type of the contents hole and define them. We have proposed the contents hole search by comparing a viewpoint of community-type-content with a viewpoint of the whole web space. In the paper, we also propose the content hole system by using Wikipedia
Key words Content hole, Search, Wikipedia

1. はじめに

ブログやSNS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発言者の視野が狭くなり，議論のテーマに対する全体像が見えなくなってしまうおそれがある．このような場合，そのテーマに対して発言者であるユーザの気づいていない情報を提示することにより，より広い視野に基づいた議論ができ，コミュニティの活性化につながると考える．そこで我々は，この「ユーザが気付いていない情報」をコンテンツホールと呼び，コミュニティ型コンテンツにおけるコンテンツホール検索を提案してきた．これまで我々の提案してきたコンテンツホール検索の手順は以下のおとりである．

（1） コミュニティ型コンテンツからのテーマの抽出
（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

これまで，上記手順の (2) を提案する [1] と共に，上記手順 (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 [2] [3] を提案してきた．ここでいう視点構造とは，そのキーワードを構成する属性情報を示す. [1] では，「名詞 A ＋が＋形容詞＋名詞 B」の構造に注目し，視点構造を抽出した．しかしながら，提案手法は計算時間がかかるため，あらかじめあるキーワードに対する Web 空間の視点構造を求めておかなければならず，あらゆるキーワードに対応することは不可能であった．そこで本論文では，Web 空間におけるキーワードの視点情報として，Wikipedia の記事の目次 (TOC) 構造をそのキーワードの属性情報つまりは視点情報と見なして，コンテンツホール検索を行う．具体的には，コミュニティ型コンテンツ群と Wikipedia の目次構造を構成するセクションとを比較し，コミュニティ型コンテンツと相違しているセクションをコンテンツホールとして提示することを行う．Wikipedia は 2008 年 8 月現在日本語で約 51 万本余りの記事があり，英語に至っては約 250 万本の記事がある [4]．これは，ブリタニカの百科事典 (Encyclopedia Britannica) が約 65000 の項目数 [5] であることから，Wikipedia は一般的な単語を十分網羅していると考えられる．さらに Wikipedia は不特定多数の人により編集されており，あるコミュニティに依存するコンテンツである可能性が低く，世間一般的な属性情報が取得可能であると考え，Web 空間における視点情報抽出に Wikipedia を用いる．

一方，コンテンツホールはこれまでの類似検索のような絞り込み検索と異なり，ない情報つまりは相違検索の一部であるため，様々な種類が考えられる．そこで本論文ではコンテンツホールを分類し，改めてコンテンツホールの定義を行う．

以下，2 章では関連研究を，3 章ではコンテンツホールの種類を，そして 4 章では Wikipedia を用いたコンテンツホール検索システムについて述べ，5 章でまとめと今後の課題について述べる．

2. 関 連 研 究

馬ら [6] は TV コンテンツから話題構造を取得し，話題の広がりや話題の詳細を求める研究を行っている．これらは TV コンテンツにおける周辺の話題及び詳細な話題のコンテンツホールを求めているともいえる．また，鳥澤 [7] らは鳥式を提案し想定外の話題の抽出の研究を行っている．これらは想定外の話題のコンテンツホールを求めているともいえる．

一方 Wikipedia に関連する研究は多数あるが中山 [8], Suchanek [9], Wu [10], Gabrilovich [11] 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している点が異なる．また，川場ら [12] 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 [13] 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為，これらの研究とは異なる．

3. 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 1 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 $C_i(i = 1 \ldots j)$ の集合からなる話題 $Sub_n(n = 1 \ldots m)$ 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レッドを一つの話題とする．一つ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 1 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 1 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．

3.1 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．

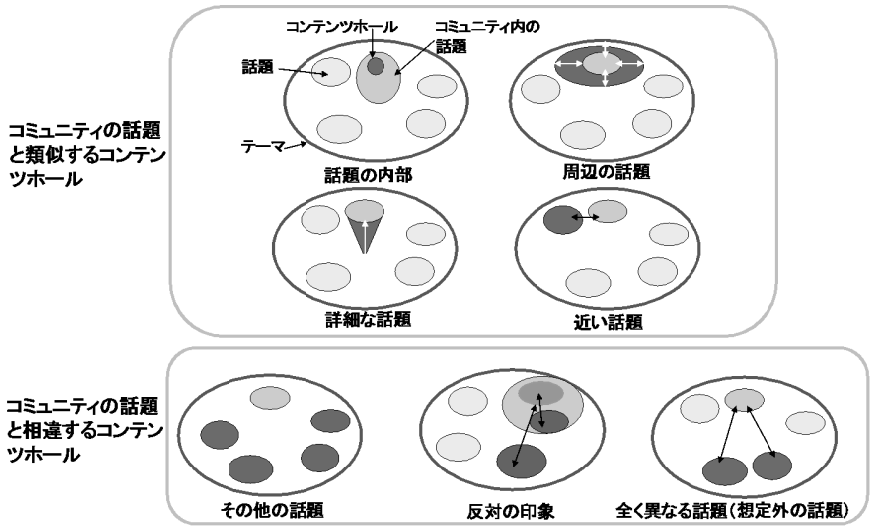

図 1　コンテンツホールの種類
Fig. 1　Image of Content-Hole type

- 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．

3.2　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 100m 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，200m 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．

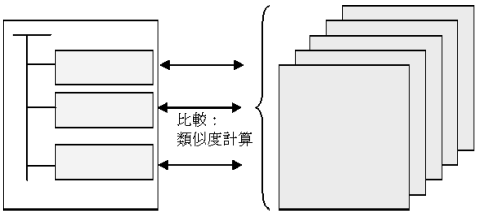

図 2　比較イメージ図
Fig. 2　Image of Comparison

4.　Wikipedia を用いたコンテンツホール検索システム

本論文ではコンテンツホール検索のリアルタイム性を重視し，Web 空間の視点構造を抽出するために Wikipedia の記事の目次構造を用いる．Wikipedia は不特定多数の人々により記述されているコンテンツであり，ある意味様々な視点でかかれたコンテンツであると考えられる．Wikipedia の記事の目次はユーザが特に指定しない限り，見出しが 4 つ以上あるページには，基本的にセクション見出しから自動生成される [14]．従って，その記事の属性情報を顕著に表したものであると考えられる．

4.1　Wikipedia とコミュニティ型コンテンツとの比較

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．(図 2 参照) 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．ここで3章で述べた各コンテンツホールのうち，以下のコンテンツホールを求める．

- 周辺の話題

コンテンツホールの候補となった目次の項目と同じ親項目を持つ目次の項目を周辺の話題と呼ぶ．つまりは，目次構造を木構造と考え，目次の項目を木構造の各ノードとした場合，コンテンツホールの候補となったノードと親ノードを共にする兄弟ノードは，話題Sub_nの拡張であると考え，周辺の話題と定義する．

- 詳細な話題

コンテンツホールの候補となった目次の項目のコンテンツにおけるリンク先の記事を詳細な話題と呼ぶ．つまりは，Sub_nは同一であり，リンク先の記事はC_iを解説するためのより詳細なコンテンツと考え，詳細な話題と定義する．

- その他の話題

類似度がある閾値以下の目次の項目全てをその他の話題のコンテンツホールとする．

4.2 プロトタイプシステム

3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」を提示するプロトタイプシステムを開発した．図3にプロトタイプシステムの検索画面を，また図3にコンテンツホールの表示画面を示す．プロトタイプシステムのフローを以下に示す．

（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(2)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(3)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4左画面参照)

4.3 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1に示す．ここで求めた適合率は以下の通りである．

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

表1　評価実験に用いたテーマとその結果
Table 1　Thema used for evaluation experiment and results

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

4.4 考　　察

- 速報性の強い情報をテーマとしている場合

予測通り，速報性の高い情報に対してはWikipediaを用いたのではあまり良い結果が得られなかった．しかしながら，北京オリンピック野球日本代表のように速報性が高く且つ注目度の高い話題に対しては，すぐにWikipediaでも反映されており，ある程度は有効である事がわかった．

- 固有名詞のうち，組織名と個人名との比較

組織名と個人名とで個人名の方が狭義の話題になり，Wikipedia

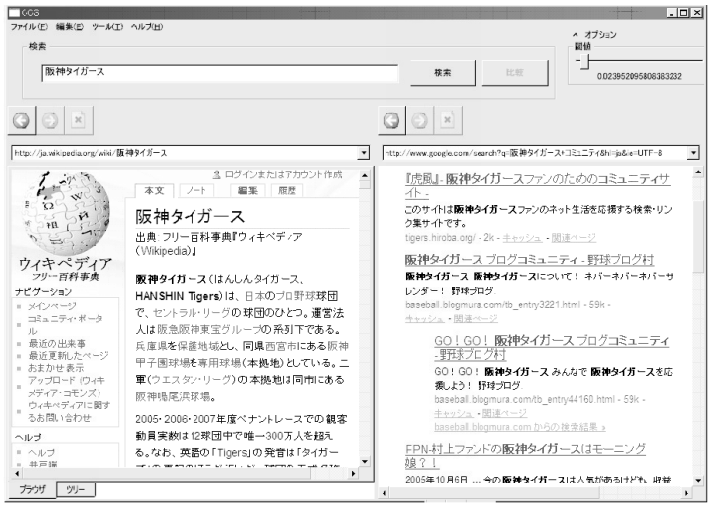

図 3　プロトタイプシステム初期画面図
Fig.3　Initial Display of the Prototype System

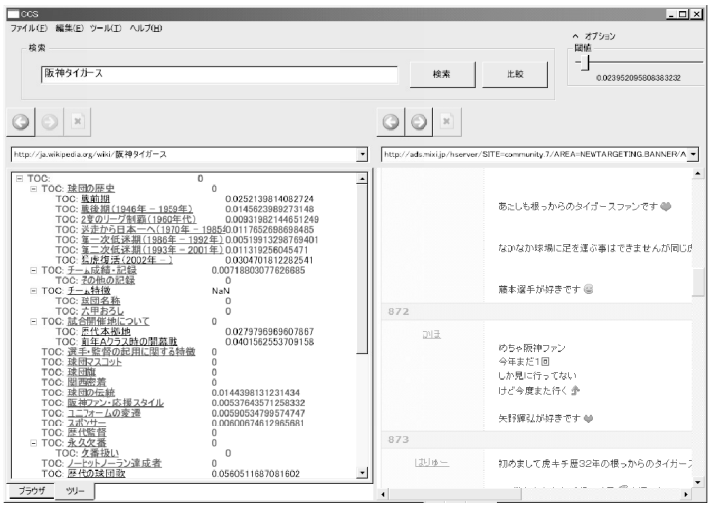

図 4　プロトタイプシステム解析結果画面図
Fig.4　Result Display of the Prototype System

を用いることが有効であると予測したが，実際には組織名，個人名に有用性の因果関係があるとは言えない結果となった．例えば，個人名における「柴咲コウ」は歌手と女優の 2 つの大きな項目を Wikipedia ではあるのに対し，柴咲コウのコミュニティに取っては，柴崎コウが好きであり，歌手と女優の顔を区別するような発言がないため類似の適合率が低かったと考えられる．

- 上位概念，下位概念の比較

ここでは，上位概念は話題が広義であるため，下位概念の方がより有効であると予測したが，予測した通りの結果となった．結果より，上位概念をテーマとするコミュニティに対してもその下位概念の項目の目次も有効であるということがわかり，どの範囲での下位概念を考慮するかが今後の課題となった．

- その他

コミュニティ型コンテンツにはコミュニティ特有の言葉や略語等が多く非常に解析がしにくい．そこで，これらコミュニティの依存度の高い言葉を一般用語に変更することが必要である．

5.　まとめと今後の課題

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，コミュニティ型コンテンツと Wikipedia の目次の項目を比較することによりコンテンツホールを求めるシステムの提案を行い，そのプロトタイプシステムを開発すると共に，そのプロトタイプを用いて実験を行い，提案手法の問題点を洗い出した．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

本研究の一部は，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) による．ここに記して謝意を表します．

文　　献

[1] 灘本明代，阿辺川武，荒牧英治，村上陽平, 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, 日本データベース学会 Letters, vol6 No2, pp29-32, 2007
[2] 荒牧英治, 阿辺川武, 村上陽平, 灘本明代: コンテンツホール検索のためのコミュニティ型コンテンツの対話解析 日本データベース学会論文誌 (DBSJ), Vol.7, No.1, pp.109-114, 2008.
[3] Eiji Aramaki, Takeshi Abekawa, Yohei Murakami, Akiyo Nadamoto: Discriminative Dialog Analysis Using a Massive Collection of BBS comments International World Wide Web Conference (WWW2008) Workshop on NLP Challenges in the Information Explosion Era (NLPIX2008), 2008.
[4] Wikipedia:http://wikipedia.org/
[5] ブリタニカ：http://www.britannica.co.jp/
[6] Qiang Ma, Katsumi Tanaka: Topic-Structure-Based Complementary Information Retrieval and Its Application, ACM Transactions on Asia Language Information Processing, Vol. 4, No.4, pp.475-503, 2005
[7] 鳥澤健太郎, 隅田飛鳥, 野口大輔, 風間淳一: 自動生成された検索ディレクトリ「鳥式」の現状, 言語処理学会 第 14 回年次大会, pp.729-732, 2008 年 3 月
[8] 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, 2008 年 3 月.
[9] F.M.Suchanek, G.Kasneci, G.Weikum: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, Proceedings of the 16th International World Wide Web Conference(WWW2007), pp.697-706, Banff, Canada, May 2007.
[10] Fei Wu, Daniel S. Weld: Automatically Refining the Wikipedia Infobox Ontology, Proceedings of the 17th International World Wide Web Conference(WWW2008), pp.365-644, Beijing, China, April 2008.
[11] Evgeniy Gabrilovich, Shaul Markovitch: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, Proceedings of the International Joint Conference on Artificial Intelligence 2007(IJCAI 2007), pp.1606-1611, Hyderabad, India, Jabuary 2007.
[12] 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, 2008 年 3 月.
[13] 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 年 6 月.
[14] Wikipedia ヘルプ：http://ja.wikipedia.org/wiki/Help: