様式 C-19

科学研究費補助金研究成果報告書

平成　21　年　3　月　31　日現在

研究種目：特定領域研究
研究期間：2007～2008
課題番号：１９０２４０７２
研究課題名（和文）　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
研究課題名（英文）
Content Hole Search of Community-type Content in the Web 2.0 era

研究代表者
灘本　明代（NADAMOTO AKIYO）
甲南大学・知能情報学部・准教授
研究者番号：３０３５９１０３

研究成果の概要：
SNS やブログのようなコミュニティ型コンテンツを対象とし，ユーザの気づいていない情報である「コンテンツホール」を検索するシステムの提案を行った．
コンテンツホール検索の基盤技術として，コミュニティ型コンテンツの対話解析に取り組み，内容的関連性，機能的関連性の抽出を行いスレッドにおける話題の抽出の成果を上げた．
7 つのコンテンツホールを定義し，そのうちの 2 つのコンテンツホール検索（話題の内部のコンテンツホール，その他の話題のコンテンツホール）に取り組み，コンテンツホールとなっている話題の抽出の成果を上げた．その 2 つのコンテンツホール検索の研究は以下の通りである．話題の内部のコンテンツホール検索においては，本年度コミュニティ内の対話解析を行い，コミュニティ内では無視されているけれども実は重要な話題である，Neglected Content の検索に取り組んだ．その他の話題では，Wikipedia と比較することにより，現在ユーザが閲覧しているコミュニティ型コンテンツに記載されていない情報を提示するシステムの研究に取り組んだ．これらは，すべてプロトタイプシステムを作成し，それらを用いた実験を行い，その有用性を示した．今年度の研究成果は，論文誌 2 本，国際会議 2 本，紀要 1 本，研究会 6 本，受賞 1 本である．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,900,000	0	2,900,000
2008 年度	2,900,000	0	2,900,000
年度			
年度			
年度			
総　計	5,800,000	0	5,800,000

研究分野：総合領域
科研費の分科・細目：情報学，　メディア情報学・データベース
キーワード：コミュニティ，Web2.0，　コンテンツホール検索，　Wikipedia

１．研究開始当初の背景

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，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．そこで本研究では，コミュニティ内で議論されているテーマにおいて議論されて

いない情報を検索し提示する事が必要であると考える．Google に代表されるように，現在の Web 検索はユーザの入力したキーワードに関連す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 QueryFree 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「ここにはない情報が欲しい」といったような検索が行えないのが現状である．そこで，我々はこの「ここにはない情報」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行う．

２．研究の目的

本研究では，コミュニティ型コンテンツにおいてそのコミュニティ内で議論されていない重要な情報であるコンテンツホール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，本研究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．

３．研究の方法

データベース，自然言語，人工知能の３領域の若手研究者が連携することにより，様々な知見からコンテンツホール検索の研究開発に取り組む．具体的には以下の３つの研究を行う．

・コミュニティ型コンテンツの対話解析
・コンテンツホールの定義
・コンテンツホール検索システムの提案

４．研究成果

4.1 コミュニティ型コンテンツの対話解析

コミュニティ型コンテンツの対話解析の研究として，SNS の対話解析に取り組み，内容的関連性，機能的関連性の抽出を行いスレッドにおける話題の抽出の成果を上げた．本研究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においての基盤技術となる．具体的には内容的関連性，機能的関連性の抽出方法を考案し，コミュニティ型コンテンツの対話解析を試みた．その結果，内容的関連性及び機能的関連性を組み合わせることにより，コミュニティ型コンテンツから話題の遷移を抽出することができ，対話解析ができることが判明した．以下に提案手法の概要を示す．

4.1.1 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が，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($REL_c(P,Q)$)を求めた．

$$\mathrm{REL}_{c(P,Q)} = \sum_{p \in Wp} \max_{q \in Wq} \mathrm{WEBPMI}(p,q)$$

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るり，WEBPMI は次の式によって定義される：

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

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 ∪ 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．

4.1.2 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：

(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..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した．

- DB-A: P が生じる N グラムのデータベース

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

¥begin{equation}

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 の集合である．

4.2 コンテンツホールの定義

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．そこで，コンテンツホールの定義として、本年度は７つのコンテンツホールを定義した．図１に提案したコンテンツホールの種類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,...，j)の集合からなる話題 Sub_n (n=1,...，m)で構成されているとする．

4.2.1 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．

● 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．

● 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 はより詳細なコンテンツを示す．

● 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．

4.2.2 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．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．

● 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．

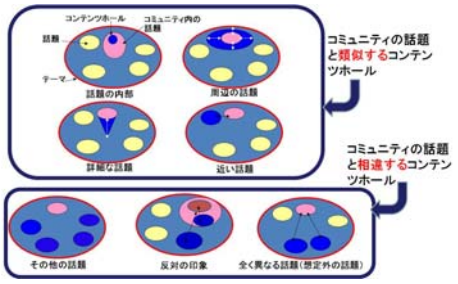

図１．コンテンツホールの種類

4.3 ２種類（話題の内部，その他の話題）のコンテンツホール検索

上記で提案した７つのコンテンツホールの内，話題の内部，周辺の話題，その他の話題の２つの課題に取組、それぞれコンテンツホール検索手法を提案すると共に，プロトタイプシステムを作成し，実験を行うことによりその有用性を示した．以下に各々のコンテンツホールの概要を示す。

4.3.1 話題の内部のコンテンツホール検索

今年度の研究結果である「コミュニティ型コンテンツの対話解析」における内容的関連性と機能的関連性を用いて，話題の内部のコンテンツホール検索を行った．

具体的には，無関係度と孤立度を求め，そこから重要だけど無視されている話題を抽出することにより話題の内部のコンテンツホール検索を行った．

無関係度と孤立度は以下の通りである．

● 無関係度(NO)

無関係度とは，一つのコメントとそのコメントの同一スレッド内の他のコメントとの相違度を示す．無関係度はその一つのコメントと他のコメントとの単語の重なり度合いで求めた．

● 孤立度(IS)

孤立度とはスレッド内のコメント同士との関係を示す度合いである．

我々は活性化モデルを用いてスレッド内の対話を木構造に変換し，IS(n_{jk})が葉節点の場合及びは接点でない場合に分類して孤立土(IS)を求めた．

無関係度と孤立度から，無視度(NDn = α ISn + β NOn) を求めた．そして重要だけど無視されている度合い(INDn)は以下のように求めた．

$$INDn = NDn + \gamma IWn$$

ここで IWn はそのテーマにおけるコミュニティ型コンテンツの総数を正規化したものである．この重要だけど無視されている度合い

(INDn)が高いものが話題の内部のコンテンツホールとした.
話題の内部のコンテンツホールのプロトタイプシステムの図を図２に示す.

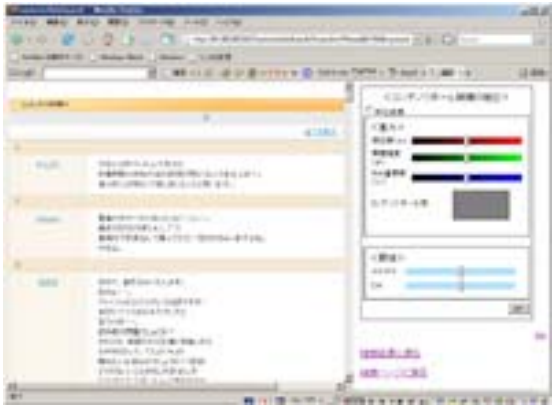

図２．プロトタイプシステム１の画面図

4.4 その他の話題のコンテンツホール検索
コミュニティ型コンテンツに全くない話題を提示する「その他の話題のコンテンツホール検索」として，Wikipedia を用いたコンテンツホール検索を提案した.
ここでは，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た．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．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総合的視点抽出に Wikipedia を用いた.
実際には，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った.
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した.
図３にプロトタイプシステムの検索画面を示す.

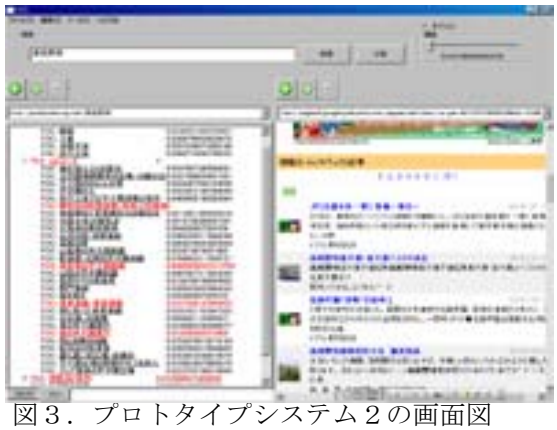

図３．プロトタイプシステム２の画面図

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３参照）

今後の展望は下記のとおりである。
① 7 種類のコンテンツホール検索の提案を行いそのうち 2 種類（話題の内部，その他の話題）のコンテンツホール検索を提案したが，残る 5 種類のコンテンツホール検索の提案を行う.
② コミュニティ型コンテンツには新語だけでなく，コミュニティ独自の単語であるコミュニティ語が多数使われている．このコミュニティ語の自動抽出を行う.
③ コンテンツホール検索をコミュニティ内だけでなく他のコンテンツにも適応させる

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）
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, Vol. 6, No. 2, pp. 29-32, 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 灘本明代, 荒牧英治, 阿辺川武, 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検索の提案,”甲南大学紀要知能情報学編, Vol. 1, No. 2, pp.249-268, 2008

〔学会発表〕(計 8 件)

① 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

② Akiyo Nadamoto, Eiji Aramaki, Takeshi Abekawa, Yohei Murakami, ″Searching for Important but Neglected Content from Community-type-content″, The Fourth International Conference On Signal-Image Technology &Internet-based Systems (SITIS’2008), pp.161-168, 2008

③ 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, Vol.6, No.2, pp.29-32, 2007 年 10 月

④ 荒牧英治, 灘本明代, 阿辺川武, 村上陽平, “コンテンツホール検索の為のコミュニティ型コンテンツの対話解析,” 電子情報通信学会 データ工学ワークショップ(DEWS2008), 2008

⑤ 荒牧英治, 灘本明代, 阿辺川武, 村上陽平, “コンテンツホール検索のための掲示板対話の解析,” 情報処理学会研究報告, Vol.2008, No.56 2008-DBS-145, pp.123-123 2008 年 6 月

⑥ 灘本明代, 荒牧英治, 阿辺川武, 村上陽平, “Wikipedia を用いたコンテンツホール検索の提案,” 情報処理学会研究報告, Vol.2008, No.88 2008-DBS-146, pp.259-264 2008

⑦ 渡邉康平, 灘本明代, “概念構造に基づく周辺の話題のコンテンツホール検索,” 第 1 回データ工学と情報マネジメントに関するフォーラム(DEIM2009), i1-21, 2009 年 3 月

⑧ 内村圭佑, 灘本明代, “Wikipedia からの概念抽出に基づくなぞかけ文の自動生成,” 第 1 回データ工学と情報マネジメントに関するフォーラム(DEIM2009), i2-40, 2009 年 3 月

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 0 件)

○取得状況 (0 件)

〔その他〕

受賞：

荒牧英治, ”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 / 電子情報通信学会データ工学研究会 / 情報処理学会データベースシステム研究会, 優秀若手研究者賞, 2008

６．研究組織

(1)研究代表者

灘本 明代 (NADAMOTO AKIYO)

甲南大学・知能情報学部・准教授

研究者番号：30359103

(2)研究分担者

村上 陽平 (MURAKAMI YOHEI)

独立行政法人情報通信研究機構・第二研究部門知識創成コミュニケーション研究センター言語基盤グループ・研究員

研究者番号：00435786

(3)研究分担者

荒牧 英治 (ARAMAKI EIJI)

東京大学・知の構造化センター・特任講師

研究者番号：70401073

(3)連携研究者

阿辺川 武 (ABEKAWA TAKESHI)

東京大学・教育学研究科・学術研究支援員

研究者番号：00431776