論　文

コミュニティ型コンテンツにおける重要だが無視されているコメントの抽出手法の提案

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] 国立情報学研究所
東京都千代田区一ツ橋２－１－２, 101-8430
[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 2009 年 11 月 9 日）

概 要

Blog や SNS 等のコミュニティ型コンテンツにおいて，我々はユーザの気付いていない重要な情報をコンテンツホールと呼び，このコンテンツホールを検索する仕組みの提案を行ってきた．本論文では，コンテンツホール検索に関する研究の第一段階として，コミュニティ内のあるスレッドでは無視されているが実は重要な情報であるコメントの抽出手法の提案を行う．具体的には，コミュニティ型コンテンツの対話解析を行い，コメントグラフを作成し，関係度及び重要度から無視されているが重要なコメントをネグレクティッド・コンテンツ（Negrected Content）と呼び，このネグレクティッド・コンテンツを抽出する方法を提案する．

キーワード: コンテンツホール，ネグレクティッド・コンテンツ，コミュニティ型コンテンツ

1　はじめに

Blog や SNS 等のコミュニティ型コンテンツは Web2.0 を代表とするコンテンツである．コミュニティ型コンテンツはコミュニティのメンバーであるユーザがコミュニティ内の議論に集中するあまり視野が狭くなり，議論のテーマに対して重要な事柄に気づかず見落としたり，またテーマの全体像が見えなくなるといったような危険性が高い．我々は，このコミュニティ内で気づいていない重要な情報をコンテンツホールと呼ぶ．これまで我々はコンテンツホールをユーザに提示することによりコミュニティの視野が広がり議論がより活性化すると考え，コンテンツホール検索を提案してきた [1].

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，我々の提案するコンテンツホール検索はコミュニティ内で気づいていない重要な情報を探すことを目的としており，相違検索の一種である．そのため，コンテンツホール検索は類似検索と異なり様々な検索ターゲットが考えられる．そこで本論文ではコンテンツホール検索の 1 つとして，スレッド内において無視されているが重要な発言をネグレクティッド・コンテンツと呼び，そのコンテンツを抽出する方法を提案する．以下，コミュニティ型コンテンツの 1 つの発言をコメントと呼び，コメント群のある集合をスレッドと呼ぶ．スレッドにはテーマがあり，そのテーマはコミュニティーのテーマのサブテーマとなっている．例えば，阪神タイガースのコミュニティには，真弓監督のスレッドや本日の試合結果についてのスレッドなど，様々な種類のスレッドがある．

コミュニティ型コンテンツにおいて，あるコメントがそのスレッドのテーマと関係ないために無視されている場合は多数ある．本論文では，このようにスレッドのテーマに関係ないために無視されているコメントを対象とせず，スレッドのテーマと関係があり且つ重要な情報であるにもかかわらずコミュニティ内において無視されている情報をコンテンツホールとして抽出することを目的とする．さらに，コミュニティにおいて暗黙に既知の情報を無視している場合も考えられるが，このような情報を自動で取得する事は困難であるため，本論文では対象外とする．実際には，コミュニティ型コンテンツのスレッド内の対話解析を行うことにより，コメントグラフを生成し，そこから孤立しているコメントを抽出する．そして，その孤立しているコメントから，スレッドとの関係度が高く且つ重要なコメントを抽出しこれをコンテンツホールとする．スレッド内のコンテンツホールを提示することにより，ユーザが他のコメントにおいて重要なコメントを理解し，議論がより活性化されることを期待する．

以下，2 章では関連研究を，3 章ではネグレクティッド・コンテンツの抽出方法を，4 章ではプロトタイプシステムについて述べる．そして 5 章では実験と考察について述べ，6 章でまとめと今後の課題について述べる．

2 関連研究

コンテンツホール検索

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 [2] 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 [3] 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．

会話／談話分析

徳永ら [4] はチャットの発話の応答関係判別を提案している．徳永らは人手による辞書を用いて発話のタイプ（アクト）を決定し，それを素性の一部とする．それに対し我々は，タイプといった恣意的な区別を導入しないかわりに，呼応表現を学習するというアプローチをとっており人手を必要としない

点で新規性を持つ．さらに，他の多くの会話／談話の先行研究は，DAMSL [5] や discourse graph-bank [6] といった少量ではあるがフレーズ単位で 20〜40 種類の対話関係をアノテートしたコーパスにもとづいて研究されている．我々の扱うデータは，それらの先行研究で用いられたコーパスと比較して，より粗い単位（コメント単位）で構成され，さらに 1 種類の関係（対応しているかどうか）のみ扱っている．以上のような欠点はあるが，我々はかつてない大きな対話データを扱っており，これが統計的手法（PMI）の導入を可能としている．

Topic Detection and Tracking（TDT)
同じトピックを持つ文章を特定するタスクである Topic Detection and Tracking (TDT) のほとんどの手法は段落毎の単語の出現頻度を手掛かりにクラスタリング手法 [7], [8], [9] を用いてトピックを推定している．この手法は新聞記事など大量の文章においては有効であるが，本研究のようなコメント毎にトピックが変わる現象を扱うためには，単語の出現頻度が情報として過疎すぎ，効果を期待できない．

アウトライヤーの抽出
データの固まりから外れた極端なデータを抽出するアウトライヤーの研究は数多くある．成田ら [10] や He ら [11] はクラスタリングベースのアウトライヤーの抽出を提案している．Angiulli ら [12] は距離ベースのアウトライヤー抽出手法として HilOut を提案している．我々の研究はアウトライヤー抽出の一つであるといえる．先行研究に対し我々は対話解析を行いコメントグラフを生成し，そのコメントグラフから無視されている重要コメントを抽出する点が異なる．

3 ネグレクティッド・コンテンツの抽出手法

我々の提案する無視されている重要コメントであるネグレクティッド・コンテンツの抽出手順として，まず孤立しているコメントを対話解析に基づいて作成するコメントグラフから発見する．そして，その孤立しているコメントからスレッドのテーマについて発言しているコメントを無視されているコメントとして抽出する．最後に，他の Web ページでよく話されている情報を重要な情報とみなして無視されているコメントからその重要な情報について話しているコメントをネグレクティッド・コンテンツとして抽出する．

3.1 コメント間の関連性の生成

コミュニティ型コンテンツのコメント群は，不特定多数のユーザにより自由に記述されているため，発言が入り組んでしまうことが多い．表 1 に，ある BBS の書き込み（以降，コメント）の例を示す．表中の対話は (1)-(3)-(5) と (2)-(4) という二つの議論に分けることができ，それぞれ，別の議論のコメントが間に挟まりギャップが生じている．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．そこで，本論文では，2 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．我々は提案手法において対応するコメント間には内容的関連性と機能的関連性の 2 つの関連性があると仮定する．内容的関連性は，2 つのコメントが内容的に類似しているかどうかであり，機能的関連性は 2 つのコメントが応答関係になっているかどうかである．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による関連性を本論文では機能的関連性と呼ぶ．

表 1: コミュニティ型コンテンツのコメントの例

(1) 小さくて軽い MP3 プレイヤーを教えてください.
やっぱりシャッフルが一番なんでしょうか？
(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) iRiver の N12 はどうでしょうか．相当小さい上，カラーですよ.
(4) バッテリー表示は近似の値だからだと思います.
(5) <3 さん．N12 はもう生産中止ですよ.

以下の条件下でコミュニティ型コンテンツの対話解析を行う.

- 入力: あるスレッド内の 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 $False$）.

以下，表記の簡便のため本論文では，i 番目のコメントを P，j 番目のコメントを Q と表記する.

3.1.1 内容的関連性

内容的関連性は 2 つのコメントが内容的に関連している度合いを示す．そこで，2 つのコメントの類似度を求め，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 [13]．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 ($WEBPMI$) [14] を利用し，文同士の類似度である内容的関連度 ($Sim_r(P,Q)$) を求める.

$$Sim_r(P,Q) = \sum_{p \in W_P} \max_{q \in W_Q} WEBPMI(p,q) \tag{1}$$

ここで，W_P は文書 P に含まれる語の集合，W_Q は文書 Q に含まれる語の集合であり，$WEBPMI(p,q)$ は次の式によって定義される：

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

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．$H(p \cap q)$ はクエリ $p \cap q$ によって検索エンジンが返す文書数．L 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棄却し，先行研究 [14] にもとづいて c=5 とする．本研究では検索エンジンに “TSUBAKI [15]” を用いる.

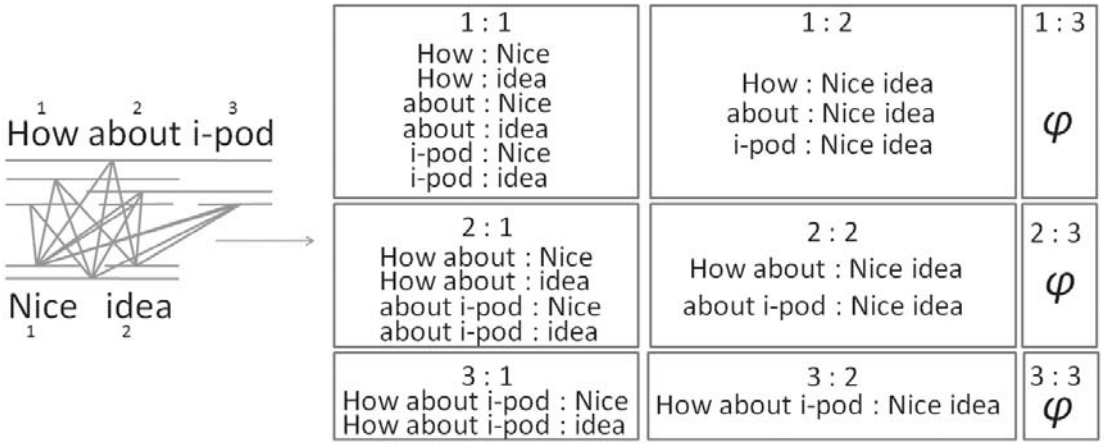

図 1: $n:m$-gram の例

3.1.2 機能的関連性

コメントには単語の重複が少ない場合も数多くある．たとえば,「ありがとう」には「どういたしまして」といった返答が自然であるが，これらの間に重複する単語はない．このようなコメントの関連性を本稿では機能的関連性と呼ぶ．機能的関連度 (Sim_d) を計算するために，我々は Corresponding-PMI (以降，CPMI) を提案する．これは WEBPMI と同様に相互情報量 MI を用いているが，以下の 2 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 (P,Q) 間での共起頻度を用いる．
2. WEBPMI は一語しか扱わないが，CPMI は語群（N グラム）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 つのコメントペアデータベースを構築する．

DB-A: コメントペア P と Q において P における N グラムのデータベース．

DB-B: コメントペア P と Q において Q における N グラムのデータベース．

DB-C: コメントペア P と Q において P，Q の $n:m$ グラムペアの組み合わせのデータベース．

ここで，n は P における N グラムの数であり，m は Q における N グラムの数である．実際には，$1 \leq n \leq 3, 1 \leq m \leq 3$ とする．例えば次のような 2 つのコメントがあると *P: How about i-pod?*，*Q: Nice idea* から図 1 に示すように $n:m$ グラムを取得する．

機能的関連度 $Sim_d(P,Q)$ は以下の通りである．

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

表 2: 応答先獲得のパターン

応答記号 (A)	レスポンス先表現 (B)	敬称表現 (C)
＜	【人名】	さん
＞	【コメント ID】	様
｜		氏
＞		ちゃん
＜		たん

* (A)+(B)+(C) または (B)+(C)+(A) のあらゆる組み合わせをパターンとする．
また，レスポンス先表現が【人名】である場合は敬称がない場合も応答関係であるとした．

ここで，N_P は，P に含まれる N グラムの集合であり DB-A より取得し，N_Q は Q に含まれる N グラムの集合であり DB-B より取得する．そして $CPMI$ 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 DB-A における N グラム p の出現頻度を，$H_b(q)$ は DB-B における N グラム q の出現頻度を示し，$H_c(p \cap q)$ は DB-C における N グラム対 $p:q$ の出現頻度を, M は検索エンジンが持つ文書数を示す．

3.2 コメントペアデータベースの生成

機能的関連性の統計量（CPMI）を計算するためには，3 つのデータベースが必要になるが，これらを構築するに当たり，大量のコメントペアが必要となる．実際には，そのようなデータは存在しないため，本研究では以下の手法でコメントペアを収集する．まず，ほとんどのコメントの対応関係が明示されていない中，表 1 の発言 5「<3 さん．N12 はもう生産中止ですよ.」のように，返答先の発言者 ID が示されている場合がある．このような対応が明示されているコメントは少ないが，大量のデータを扱うことで収集することでリカバーを行う．具体的には，まず SNS サイト「mixi」(http://mixi.jp/) のコミュニティの掲示板を中心に 130,000 の掲示板をクロールし，17,300,000 コメントを収集した．前述したように，これらのコメント間の対応関係は基本的には明示的でなくその中から応答先が明示されている物を抽出する．そこで表 2 に示されるような応答先を得るパターンを用いて，対応しているコメントを抽出した．このパターンで対応先を抽出できたのは 890,000 コメントペアつまりは 1,780,000 コメントであり，全体の 10.2%であった．

ここで，長い（文字数が多い）コメントは，複数のコメントへのレスポンスや，長い引用など，複雑な現象を含んでいる場合が多く，本研究の問題設定（コメントペアとして対応する/しないの二値

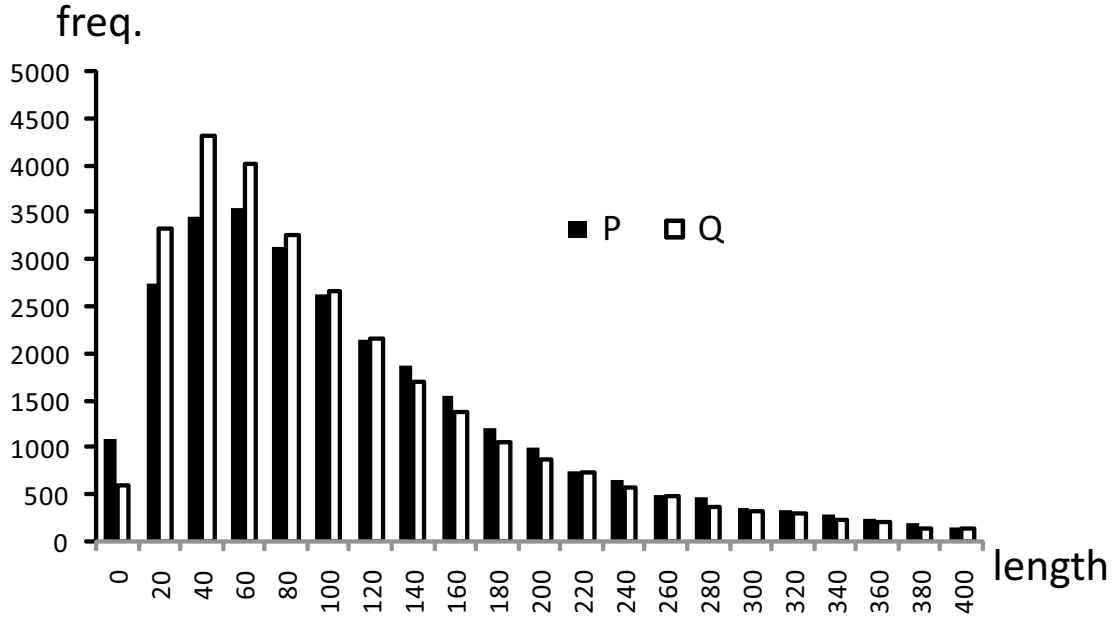

図 2: コメントペアの長さと頻度（発言 P，その応答 Q）

を出力）に沿わない場合がある．そこで長いコメントは取り除く事を行う．図 2 は収集したコメントペアの長さと頻度を示している．これを見ると 40 文字をピークとして，なだらかな曲線を描いているのがわかる．特に 20-100 文字で全体のほぼ 60%を占めていることがわかる．さらに図 3 では収集したコメント同士の長さの関係を示している．図 3 を見るとコメント 100 文字以内では，P と Q の長さの関係は一様ではなく相関関係がないことがわかる．つまりは，100 文字に近いコメントに対する応答が短いという特徴がないため，P と Q の長さの関係を考慮する必要がないことがわかる．これら 2 つのグラフより文字数 100 文字までが妥当であると考え，先に抽出した 890,000 コメントペアから，100 文字以上の長いコメントは棄却した．この結果，121,699 コメントペアを抽出した．

3.3 SVM による学習

上記の二つの指標 $Sim_r(P,Q)$ と $Sim_d(P,Q)$ がどれくらいの値以上であれば，コメントペアが応答関係にあるかどうかを決定するのは困難である．そこで SVM を用いてコメントペアが応答関係にあるかどうかを決定する．SVM の素性としては，上記の 2 つの値に加え，P と Q に含まれる語彙そのものを素性とする．SVM で学習を行うためには (1) 正例と (2) 負例が必要である．正例としては，次章で述べる $(P:Q)$ をそのまま用い，負例としては，コメントペア $(P:Q)$ の応答部分 Q を同一掲示板の他の応答 Q' と無作為に入れ替えこれを用いる．

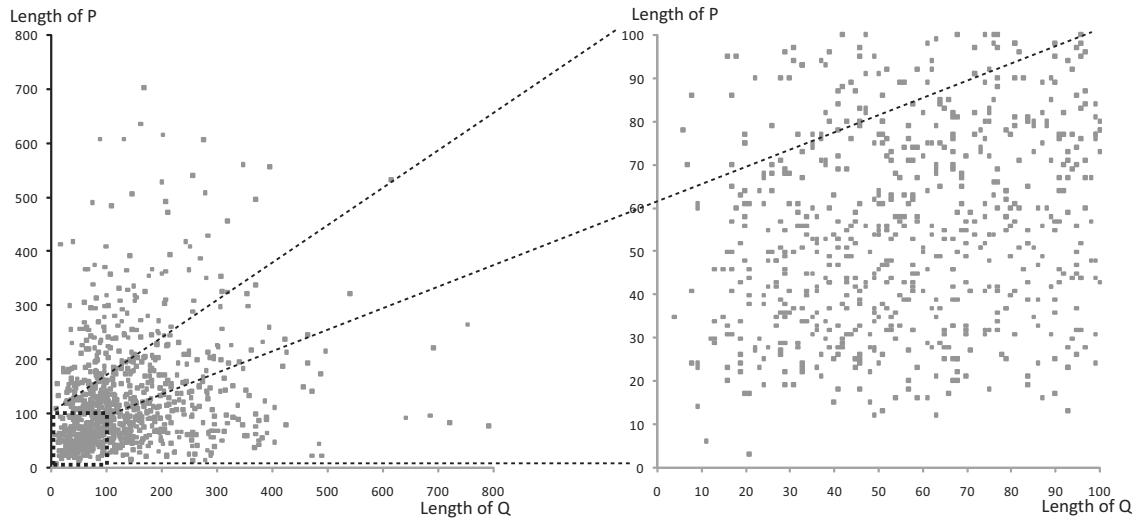

図 3: 収集されたコメントペアの長さ（発言 P，その応答 Q）

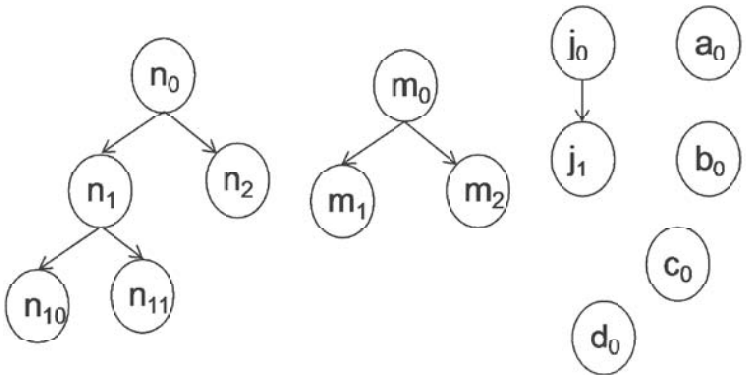

図 4: コメントグラフ

3.4 ネグレクティッド・コンテンツの抽出

我々はまずはじめに，無視されているコメントを抽出し，その中から重要なコメントを抽出しこれをそのコミュニティのネグレクティッド・コンテンツとする．

無視されているコメントの抽出

無視されているコメント候補である孤立しているコメントを抽出するために前節で求めたコメントペアの生成計算をすべてのコメントに対し行う．そして，その応答関係と決定されたコメントペアのうち，1 つのコメントが複数の異なるコメントと応答関係になっている場合それらを連結して図 4 に示すようなコメントグラフを生成する．図 4 の各節点はスレッド内の 1 コメントを示し，グラフの枝の方向は発言順番である．例えば，図 4 では m_0 と m_1，m_0 と m_2 のコメントペアが応答関係であると判定されたとき，m_0，m_1，m_2 を連結し図 4 に示すような部分グラフを生成する．

ここで，節点が 1 つもしくは 2 つで構成されている部分グラフは，他のコメントに対して孤立し

ていると考え，孤立コメントとする．つまりは，図 4 の場合，孤立コメントは $j_0, j_1, a_0, b_0, c_0, d_0$ である．

我々の提案する無視されている重要なコメントは，孤立しているだけでなく，そのスレッドのテーマと関係しているが無視されているコメントである．そこで，孤立コメントから関係度を求め，関係度がある閾値以上のコメントのみを抽出する．コメント P の関係度 $RO(P)$ は以下の様に内容的関連性を用いる．

$$RO(P) = \frac{\sum_{i=1}^{n-1} Sim_r(P, Q_i)}{n-1} \tag{5}$$

ここで，nn はスレッド内の総コメント数を示し，Q_i はコメント P 以外のスレッド内のコメントを示す．

ネグレクティッド・コンテンツの抽出

抽出した無視されているコメントから，一般には重要であるコメントを抽出する．これをネグレクティッド・コンテンツとする．一般に重要であるとは，その無視されているコメントに関する事が他の Web ページでも数多く掲載されていると考え，関連する Web ページの総数により，その重要性を計る．具体的には，共起辞書を用いて，無視されているコメントからそのスレッドのテーマ WT と共起度が最も高い単語をキーワード Key とし，クエリを $WT \cap Key$ としてサーチエンジンにて Web 検索を行う．その結果のページ数がある閾値以上のコメントは，一般的に重要とされているコメントであるとする．以上よりネグレクティッド・コンテンツを抽出する．

4 プロトタイプシステム

提案手法を用いてプロトタイプシステムを C#と JAVA を用いて開発した．プロトタイプシステムの画面図を図 5 に示す．プロトタイプシステムの手順は以下の通りである．

1. ユーザは自分が見たいコミュニティのキーワードを入力する．
2. システムはユーザが入力したキーワードを持つコミュニティを検索しウィンドウにリストを表示する．プロトタイプシステムでは，対象となるコミュニティを mixi とした．
3. ユーザは表示されたコミュニティから調べたいコミュニティ及びスレッドを選択する．そして，パラメータを入力する．
4. システムはユーザの入力に基づき，対象スレッドのコメントグラフを生成し，無視されているコメントを抽出する．
5. システムは無視されているコメントの重要度を求め，ネグレクティッド・コンテンツを決定する．この時 SVM には TinySVM を使用した．
6. システムはネグレクティッド・コンテンツであるコメントを赤枠で囲みユーザに提示する（図 5 参照）．

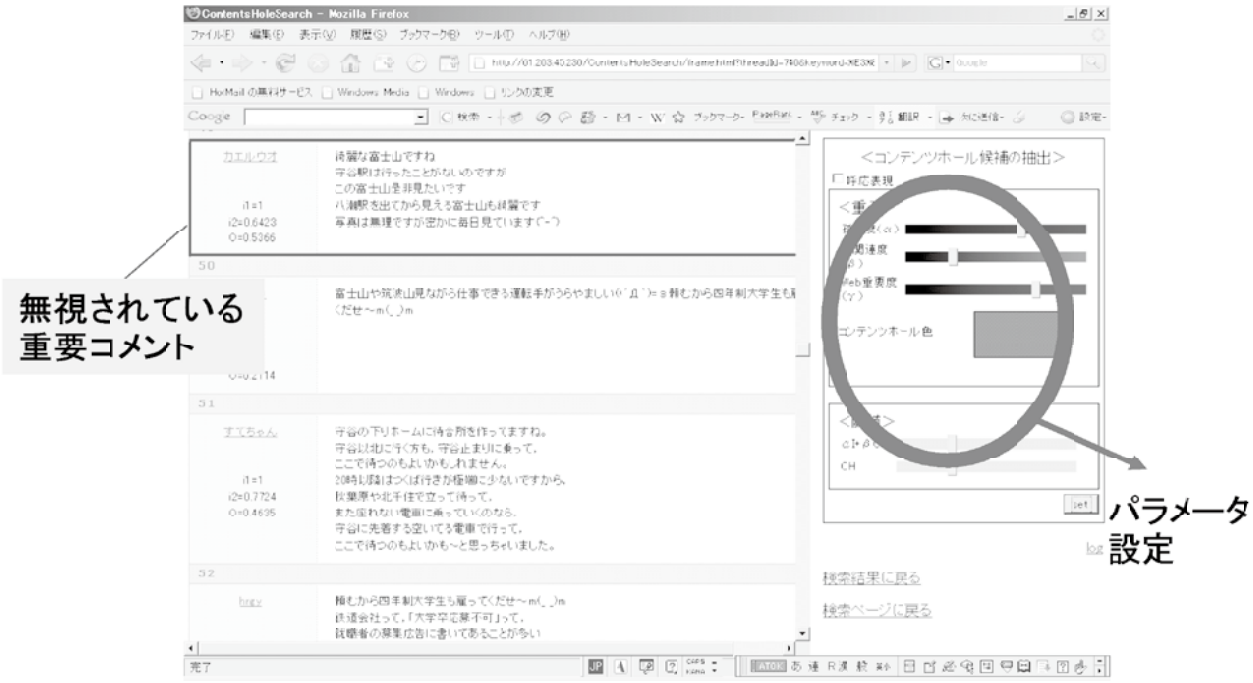

図 5: プロトタイプシステムの画面

5 実験

プロトタイプシステムを用いて，ネグレクティッド・コンテンツが実際に抽出されているかどうかを求める実験を行い，提案手法の有用性をはかる．提案手法では特に対象コミュニティのドメインを決定していないため，以下の種類のコミュニティのテーマを用いて実験を行うことにより，データのタイプ別における提案手法の有用性を計った．実際には，3 人の被験者を対象に以下のデータセットを用意して実験を行った．

- 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る．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して，提案手法の有用性を計る．
- 速報性の強い情報とそうでない情報との比較
  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の場合，呼応関係は定常的な話題に対してあまり多くなく，個々人が勝手にコメントしている場合が多い．このように呼応関係が明確ではないコンテンツに対して本提案手法が有用かどうかを計る．

組織名と個人名との比較に関する考察

表 3 より，組織名と個人名を比較した場合，個人名の方が組織名より適合率がよいことがわかる．ま

表 3: ネグレクティッド・コンテンツの抽出の適合率

コミュニティのテーマ	Human-D	Human-E	Human-F	平均
会社名と個人名				
トヨタ	0.43	0.55	0.38	0.45
JAL	0.35	0.63	0.52	0.50
柴崎コウ	0.51	0.65	0.62	0.59
山田優	0.53	0.62	0.67	0.61
時系列データと非時系列データ				
電車	0.73	0.63	0.68	0.68
たばこ	0.57	0.56	0.62	0.58
環境	0.63	0.58	0.65	0.62
プロ野球	0.48	0.46	0.42	0.45
政治	0.38	0.45	0.46	0.43

た，被験者からも同様に，個人名の話題の方がどのコメントが無視されているかを判断しやすかったとの意見があった．これは個人名の方が話が発散しておらず，対話ペアがとりやすかった為だと考えられる．特に，組織名の「トヨタ」は会社そのものを指すと言うよりも車の話題の方が多く且つ車種が多いため話が発散しており，同じ組織名の「JAL」と比較しても適合率が悪かった．

速報性の強い情報とそうでない情報との比較

表 3 より，速報性のある情報よりも，定常的な情報の方が適合率がよいことがわかる．実際に被験者からも，速報性のあるデータ特に野球に関する情報は実況中継のような物が多く，呼応関係を見つけるのが困難であったとの意見があった．

6 まとめと今後の課題

本論文では 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であるコンテンツホール検索の 1 つとして，あるコミュニティでは無視されているが，一般的には重要視されているコメントの抽出手法の提案を行った．具体的には，コミュニティ型コンテンツの各コメントの対話解析を内容的関連性と機能的関連性に基づいて行い，コメントグラフを作成することにより無視されている重要コメントの抽出方法を提案した．今後の課題は (1) 新語，コミュニティ語への対応, (2) 機能的関連性のデータベースの拡張, (3) ある特定分野への対応を行ってゆきたい．

謝辞

本研究の一部は，平成 21 年度科研費特定領研究域「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：21013044，代表：灘本明代) 及び甲南大学平生太郎基金科学研究奨励助成金によるものである．ここに記して謝意を表します．

参考文献

[1] 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.

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

[3] M. Qiang, A. Nadamoto and K. Tanaka, “Complementary information retrieval for cross-media news content,” *Elsevier ARTICLE Information Systems*, vol. 31, Issue 7, pp. 659-678, 2006.

[4] 徳永泰浩, 乾健太郎, 松本裕治, “チャット対話における発話間の継続関係と応答関係の同定,” 自然言語処理, vol. 12, no. 1, pp. 79-105, 2005.

[5] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Lingustics*, vol. 26, no. 3, pp. 340–373, 2000.

[6] F. Wolf and E. Gibson, “Representing discourse coherence: A corpus-based study,” *Computational Linguistics*, vol. 31, no. 2, pp. 249–287, 2005.

[7] F. Walls, H. Jin, S. Sista, and R. Schwartz, “Topic detection in broadcast news,” in *Proc. the DARPA Broadcast News Workshop*, pp. 193–198, 1999.

[8] K. Rajaraman and A. Tan, “Topic detection, tracking and trend analysis using self-organizing neural networks,” in *Proc. the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001)*, pp. 102–107, 2001.

[9] J. M. Schultz and M. Liberman, “Topic detection and tracking using idf-weighted cosine coefficient,” in *Proc. the DARPA Broadcast News Workshop*, pp. 189–192, 1999.

[10] K. Narita and H. Kitagawa, “Detecting outliers in categorical record databases based on attribute associations,” in *Proc. the 10th Asia-PacificWeb Conference International Conference (APWeb 2008)*, pp. 111–123, 2008.

[11] Z.He, Z.Zu and S.Deng, “Discovering cluster-based local outliers,” *Pattern Recognition Letters*, vol. 24, pp. 1641–1650, 2003.

[12] F.Angiulli and C.Pizzuti, “Outlier mining in large high-dimensional data sets,” *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, 2005.

[13] 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.

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

[15] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto and S. Kurohashi, “TSUBAKI: An open search engine infrastructure for developing new information access methodology,” in *Proc. the International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.