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

Dialog analysis of the Community Type Content for Content Hole Search

荒牧 英治♠ 阿辺川 武♦
村上 陽平♣ 灘本 明代♥

Eiji ARAMAKI Takeshi ABEKAWA
Yohei MURAKAMI Akiyo NADAMOTO

これまで我々は，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見落とされた話題を検索するコンテンツホール検索の手法の提案をしてきた．そして，まず第一段階として，Web 空間における視点抽出を行った．本論文では，改めて，コミュニティ型コンテンツのコンテンツホールを定義すると共に，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について提案する．具体的には，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．

We have proposed the method of extracting for content hole of the community-type content. The community type content such as SNS and Blog are representative technique of Web2.0. In this paper, we formalize the content hole, and propose the dialogue analysis of the community type content for extracting of the viewpoint of the community type content. We focus on the two types of the relation between two comments in the community type content. One is "Relevance consistency" which is relation of the content between two comments, the other is "Discourse consistency" which relates question-answer between them. By using the proposed method, we can extract multiple subjects and also extract viewpoint of each subject.

1. はじめに

現在，SNS上の掲示板やブログ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，コミュニティ型コンテンツにおいて，コミュニティ内でユーザが気付いていない情報を探し提示する事を目的としコミュニティ型コンテンツのコンテンツホール検索を提案する．我々は，この「ユーザが気付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．

これまで我々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ために，以下の手順を提案してきた．

(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)を提案してきた[1]．本論文では，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に焦点を当てる．ここで問題となるのは，SNS上の掲示板やブログ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，対話形式で構成される点である．さらに，その対話は通常の発話者同士が対面する音声対話と異なり，不特定多数人で複数のトピックについて同時に議論が行われうる．このようなテキストを扱うためには，どのコメントがどのコメントと対応しているのか，その対応関係を抽出する処理が必須となる．

そこで，本論文では，コメント間に存在する二種類の関連性を定式化し，その指標を用いることで，コメント間の関係を判定する手法を提案する．これにより，対話を時系列に辿ることが可能となり，コミュニティ型コンテンツの視点構造の構築につながる．

以下，2章ではコンテンツホール検索システムの定義を，3章ではコミュニティ型コンテンツの対話解析の手法を，4章では対話解析の実験，5章では関連研究，6章にて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

現在のWeb検索はGoogleをはじめとして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においてもTREC等キーワード検索を想定したタスク設定での研究が主流である．最近ではPowersetなどの自然言語入力による検索やサンプル・コンテンツからQueryFree[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．本研究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．

我々の提案するコンテンツホール検索の手順は以下の通りである．

♠ 正会員　東京大学 知の構造化センター　eiji.aramaki@gmail.com
♦ 正会員　東京大学大学院 教育学研究科　abekawa@p.u-tokyo.ac.jp
♣ 非会員　独立行政法人　情報通信研究機構　yohei@nict.go.jp
♥ 正会員　甲南大学知能情報学部　nadamoto@konan-u.ac.jp

- **コミュニティ型コンテンツからのテーマの抽出**

SNS上の掲示板やブログなどからそのテーマとなる名詞句を抽出する．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ものとする．

- **Web空間における視点情報の抽出**

Webページ群から「名詞A+が+形容詞+名詞B」という手がかり表現を利用し名詞Bについての視点を抽出する．この視点からシソーラスを用いて視点構造グラフを生成する．

- **コミュニティ型コンテンツの視点抽出**

SNS上の掲示板やブログ等のコミュニティ型コンテンツは対話形式のコンテンツになっていることが多い．そこで，対話解析を行うことによりそのテーマの視点を抽出する．そして，その視点から視点構造グラフを生成する．

- **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．

- **抽出されたコンテンツホールの提示**

抽出したコンテンツホールをコミュニティ内のユーザに提示するユーザインタフェースを開発する．

以上のように本研究では，あるテーマのコミュニティの話題には範囲があると仮定し，その範囲内の視点（情報）から抜け落ちている視点（情報）をコンテンツホールと呼んでいる．しかし，あるコミュニティにおいて視点の偏りがあるのは当然のことであり，すべてを網羅した視点で議論が行われることの方がむしろ稀であると考えられる．そこで我々は現在の視点と足りない視点との関係から図1のようなさまざまなコンテンツホールが考えられると予想している．これらのいずれがユーザにとって提示する価値のある情報かは，今後の研究の中で実証的に調査する予定である．

3. コミュニティ型コンテンツの対話解析

SNS上の掲示板やブログ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多数を占める．また，これは通常の音声対話と異なり不特定多数の人間により複数のトピックが議論されるため，どのようなトピックについて誰がどれくらい発言しているかを知ること容易ではない．

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，頻繁に見られるもので，図2が示すように，ほぼ半数の対応するコメント間に距離2～5程度のギャップが存在していることがわかる．

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タスクに挑戦する．このタスクにおいて我々は対応するコメント間には次の2つの関連性があると仮定している．まず，2つのコメントが内容的に類似しているかどうかで本論文ではこれを**内容的関連性**と呼ぶ．次に，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による関連性を本論文では**機能的関連性**と呼ぶ．

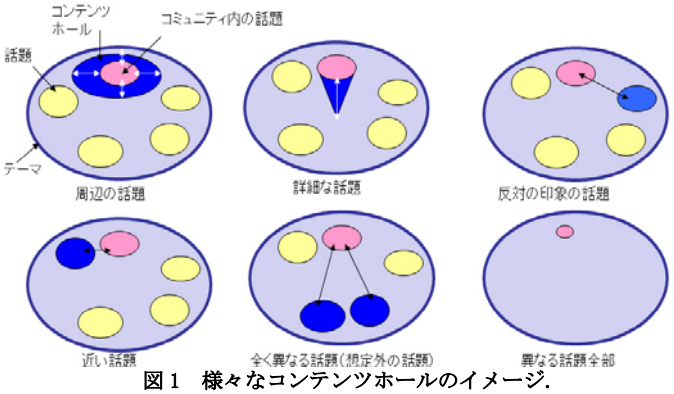

図1 様々なコンテンツホールのイメージ．
Fig.1 Image of Various Contents Holes.

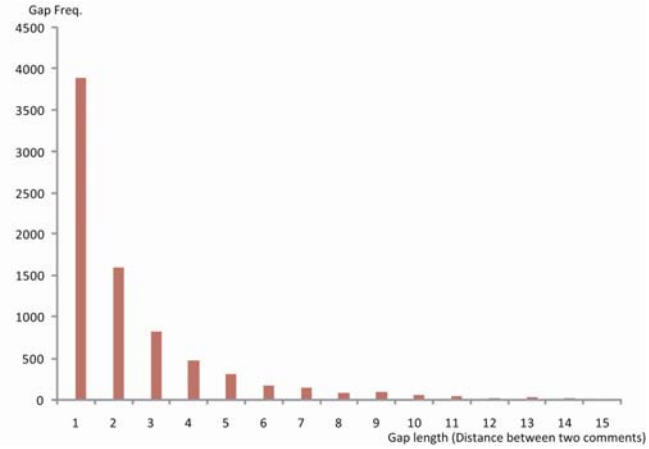

図2 対応するコメント間の距離とその頻度．
Fig.2 Distance between a Comment and its Response (X) and its Occurrence (Y).
* 統計（頻度）は3.4節の手法で自動抽出した応答関係を用いて集計した．

内容的関連性は，文同士の類似度と近い概念であると考え，Web上での単語の共起に基づく類似度[4]を用いて求める．そして，**機能的関連性**は，**対応する表現**（以降，**対応ペア**）を得る必要があると考え，本論文では，大量の対話コーパスを用から対応ペアを自動収集する手法を同時に提案する．

3.1 問題設定

表 1 BBS 書き込みの例
Table 1 BBS Examples

ID	コメント
(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．

まず，本研究の問題設定を定式化する：

- **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）．

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．

3.2 内容的関連性

２つのコメントが内容的に関連している度合い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[5]．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(*SIMr(P,Q)*)を求める．

$$sim_r(P,Q) = \sum_{p \in W_p} \max_{q \in W_q} WEBPMI(p,q),$$

ここで，*Wp*は*P*に含まれる語の集合，*Wq*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：

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

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．*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)* はクエリ「*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棄却した（先行研究[4]にもとづいて*c=5* とする）．

3.3 機能的関連性

機能的関連性(*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-gram）を扱う(*n=1..3*)．

表 2 応答先獲得のパターン
Table 2 Response Target Extraction Patterns

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
く		ちゃん
＜		たん

＊「＞【人名】さん」「 / 【人名】様」など，(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して用いた．
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（【コメントID】の場合は，敬称がない場合，箇条書きの番号などが誤って抽出されてしまう恐れがある）．

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

ここで，*NP*は，*P*に含まれるn-gram の集合，*NQ*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：

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

ここで，*Ha(p)* はn-gram *p*のP における出現数．*Hb(q)* はn-gram *q*の*Q* における出現数．*Hc(p∩q)*はn-gram 対(*p:q*) の共起頻度数である．

3.4 Web からのコメントペアの自動抽出

前節の統計量（*CPMI*）を計算するためには，大量の対応するコメントデータが必要となる．そこで，以下の手法で，データを自動抽出した．

まず，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BBS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い．しかし，表１最下部の発言「*>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は対応関係を得ることができる．そこで「＞」のような応答先を示す記号(応答記号)，発言者IDや発言者名，さらに「さん」といった敬称を手掛かりに「＞【人名】さん」といったパターンを作成し，対応するコメントのペアを抽出した．この際に用いた応答記号，敬称のリストを表２に示す．予備実験の結果，このようなパターンで対応先を抽出できる割合は5%と低い精度であるが，我々は大量のクロールにより母数を増やすことでこの問題をクリアできると考えている．

また，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（対応する/しないの二値を出力）に沿わない場合がある．そこで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（全コメント量の1.4%）を得ることができた．表3に抽出したコメントペアの例を示す．

表 3 自動抽出したコメントペアの例.
Table 3 Several Examples of Extracted Comment Pairs.

P	Q
アドマイヤムーンは秋天で引退ですか？	■＞正確にはドバイへ移籍です
年金問題が解決する前に、リア・ディゾンが日本語を完璧にマスターする方が早いと思うよ。	■＞アグネス・チャンの方は一体いつになったら日本語が流暢になるのでしょうか？
わはー♪新入りです。よろしくおねがいします	■ お､100レスおめでとうございますヽ(´▽`)ノわはー
ラストベガはベガの子供ではありませんよ！父がアドマイヤベガなのでその名前なのだと…	＞■すみません＾＾；勘違いしてました ┌〇″ﾍﾟｺｯ

4. 実験

実験により次の３点を調査した：（1）まず，提案する２つの関連性の応答関係判別の精度，（2）次に，人間との精度とシステムの精度の差，（3）最後に，データの数（CPMIを計算するコメントペアの数）と精度の関係，である

4.1 テストセット

テストセットの構築にあたっては，それぞれ3.4節にて獲得したコメントペア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.

- **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係を調べるために用いる大規模データ.8400コメントペアからなる.

4.2 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ヒット数を得るために検索エンジン基盤TSUBAKI[6]を用いた.

4.3 SMALL-SET の結果

表4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した. 表5に各手法による出力同士の一致率を示す.

手法の中でもっとも高い精度を示したのは機能的関連性であり，対話の応答関係を推定する上で機能的関連性が重要な手がかりであることがいえる.

表 4 SMALL-SET での結果.
Table 4 Result 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	72.60	72.10
Overlap	61.42	58.71	87.67	70.32
Simr	61.42	72.09	42.46	53.44
Simd	65.71	66.23	69.86	67.99

表 5 人間とシステム間の一致率と Kappa 値.
Table 5 Kappa Value Matrix between Methods.

	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) ▼

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

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. これは一部の短い返答（「*そう思います*」や「*ありがとうございます*」など）がどのような発言の返答としても考えられるためしばしばFalse Positiveとなってしまうのが原因である. このような誤りはあるものの人間同士の一致度は表5に示されるように高く(kappa value = moderate)，これらの限界は評価者間で一致している. 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.

2 つの関連性の独立性

前述したように *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，表5が示すように，*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$_d$* 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$_r$は互いに独立であり，別々のtest exampleを正解していることがわかり，関連性を2つの指標に分解する妥当性を示している.

表6に高いCPHIを持つ呼応表現の例を示す. 表にみられる

表 6 高い CPHI を持つ対応ペアの例.
Table 6 Examples of phrase pairs with high CPHI values.

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

ように，これらの関連性を内容的関連性でとらえることは困難だと想像される．例えば，以下のような例では，まったく単語が重なっておらず，内容的関連性では応答関係をとらえることは困難であるが，機能的関連性を用いれば高い確度でこれをとらえることができる：

P「*次のオフ会にはさっそく行ってみたいと思います*」
Q「*来月の２２日お待ちしています*」

4.4 LARGE-SET の結果: コーパスサイズと精度

図3の対応ペアを計算に用いたコメントの数とsim_dの精度（Accuracy）関係を示す．表に示されるように最大サイズにおいても精度はまだ飽和しておらず，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．このため，今後は，今回のように手がかり表現を人手でデザインするのではなく，自動抽出するなどして被覆率を上げる技術が望まれる．

5. 関連研究

本研究のように応答かどうかを判別するという問題設定は会話研究では珍しく，同様の研究は，我々の知るかぎりでは，徳永ら[7] によるチャットの発話の応答関係判別のみである．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していた．一方，本研究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．

5.1 会話／談話分析

他の多くの会話／談話の先行研究は，DAMSL[8] や RST-DT[9] や discourse graph-bank[10] 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きた．本研究で扱うデータは，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）しか扱っていない．

以上のような欠点があるものの，本研究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．本研究により，今後の大規模なデータを用いた会話研究が活性することを期待している．

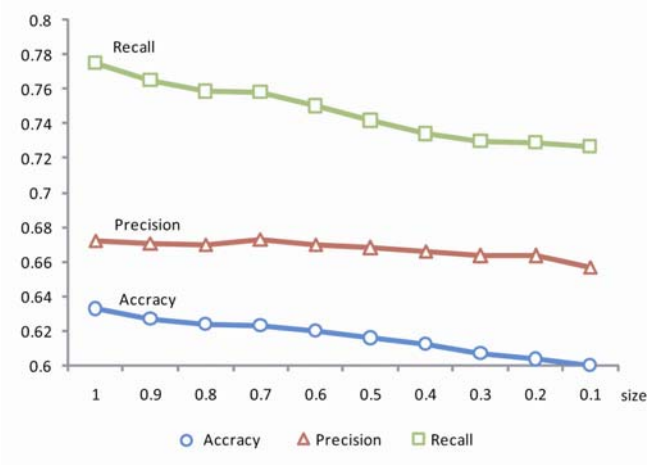

図 3 データのサイズと sim_d の精度.
Fig. 3 Data size (# of comment-pairs) and sim_d Performance..

5.2 Topic Detection and Tracking

もう一つの関連分野は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) [11] である．多くの TDT の手法は段落枚の単語の出現頻度を手掛かりにクラスタリング手法[12, 13, 14] 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語が含まれるテキスト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．本実験結果が示すように，コメントのような短いテキストにおいては，機能的関連性が手がかりになる．

5.3 言語学: 語用論

言語学では Grice から近年の neo-Gricean にいたるまで様々な会話分析の理論が提案されてきた．Grice は 4 つの conversational maxim (maxim of quantity,quality, relevance, and manner) [15]を提案した．Sperber & Wilson は関連性理論(relevance theory)[16]にて 4 つの maxim を関連性というひとつの考え方で説明した．Levinson は Generalized Conversational Implicature (GCI) [17] にて Grice の Maxim を拡張し， 省略など多くの言語を取り込んだ．現象の説明を試みた．以上のような語用論の研究はさまざま言語現象に説明を与えてきたが，数学的に定式化されておらず，実装できるような性質の理論ではない．本研究は，会話における関連性の定式化を試みており，言語学の分野にも影響を与えることを期待している．

6. まとめと今後の展望

本論文では、(1)コンテンツホールの定義，(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を行った．

今後の展望は下記のとおりである。

- **コミュニティ型コンテンツからの視点情報の抽出**： 今後，本論文にて提案したコミュニティ型コンテンツの対話解析の結果を用いて，コミュニティ型コンテンツからの視点情報の抽出及び，視点構造グラフの生成方法．
- **コンテンツホールの抽出**：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する手法の提案を行う．

- **コンテンツホールの提示**：抽出したコンテンツホールをコミュニティ内のユーザに提示するユーザインタフェースを開発する．

[謝辞]
本研究の一部は，平成19年度科研費特定領研究域「Web 2.0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, vol.6 No.2, pp29-32, 2007.
[2] Henzinger Monika, Chang Bay-Wei, Milch Brian, Brin Sergey, "Query-Free News Search", World Wide Web Journal, Springer Science+Business Media B.V., ISSN: 1573-1413 pp.101-126, 2005.
[3] Ma Qiang, Akiyo Nadamoto, Katsumi Tanaka, "Complementary information retrieval for cross-media news content", Elsevier ARTICLE Information Systems, Volume 31, Issue 7, pp. 659-678, 2006.
[4] Danushka Bollegala and Yutaka Matsuo and Mitsuru Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines", Proceedings of 16th International World Wide Web Conference (WWW2007), pp.757-766, 2007.
[5] Marco De Boni and Suresh Manandhar, "An Analysis of Clarification Dialogue for Question Answering", Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003), pp.48-55, 2003.
[6] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto and Sadao Kurohashi, "TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology", Proceedings of International Joint Conference on Natural Language Processing (IJCNLP2008), 2008.
[7] 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.
[8] 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, Volume 26, number 3, pp.340-373, 2000.
[9] Carlson, D. Marcu and M. E. Okurowski, "RST Discourse Treebank", Linguistic Data Consortium, 2002
[10] Florian Wolf and Edward Gibson, “Representing discourse coherence: A corpus-based study”, Computational Linguistics, Vol.31, No.2, pp.249-287, 2005.
[11] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, “Topic detection and tracking pilot study: Final report”, In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, pp. 194-218, 1998.
[12] K. Rajaraman and A. Tan, “Topic detection, tracking and trend analysis using self-organizing neural networks”, In Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001), pp. 102-107, 2001.
[13] J. M. Schultz and M. Liberman, “Topic detection and tracking using idf-weighted cosine coefficient”, In Proceedings of DARPA Broadcast News Workshop, pp. 189-192, 1999.
[14] F. Walls, H. Jin, S. Sista, and R. Schwartz, ”Topic detection in broadcast news”, In Proceedings of DARPA Broadcast News Workshop, pp. 193-198, 1999.
[15] H. P. Grice, “Logic and conversation”, In Cole, P. and Morgan, J. (eds.) Syntax and semantics, vol 3. New York: Academic Press, 1975.
[16] Dan Sperber and Deirdre Wilson, "Relevance: Communication and Cognition", Cambridge: Harverd University Press, 1986.
[17] Stephen C. Levinson, “Presumptive meanings: The theory of generalized conversational implicature”, MIT Press, 2000.

荒牧 英治 Eiji ARAMAKI
東京大学知の構造化センター特任講師．2005．東京大学情報理工学系研究科博士課程修了，情報理工学博士．自然言語処理の研究に従事．

阿辺川 武 Takeshi ABEKAWA
東京大学大学院教育学研究科，特任研究員．2006．東京工業大学総合理工学研究
科博士課程卒業．工学博士．自然言語処理の研究に従事．

村上 陽平 Yohei MURAKAMI
(独)情報通信研究機構研究員．2006．京都大学大学院社会情報学専攻博士課程修了．博士（情報学）．現在，言語グリッドプロジェクトを推進．

灘本 明代 Akiyo NADAMOTO
甲南大学知能情報学部知能情報学科 准教授．2002．神戸大学大学院自然科学研究科後期博士課程修了，博士（工学）．Web コンピューティング，データ工学の研究に従事．