UserGeneratedContent におけるコメントの重要度計算手法の提案

内村　圭佑†　　灘本　明代††

† 甲南大学大学院自然科学研究科　〒 658-8501 兵庫県神戸市東灘区岡本 8 丁目 9 番地 1 号
†† 甲南大学知能情報学部　〒 658-8501 兵庫県神戸市東灘区岡本 8 丁目 9 番地 1 号
E-mail: †mn924002@center.konan-u.ac.jp, ††nadamoto@konan-u.ac.jp

あらまし　近年インターネットにおいて，SNS やブログといった一般の人々によって作成されるコンテンツが多く存在している．このようなコンテンツにおいて，しばしばユーザは何らかのテーマに沿ったコミュニティを形成し，テーマに対する情報の交換を行っている．テーマに対してある程度の知識を持っているユーザは，コミュニティ内でやり取りされる情報の中から重要な情報を判断し取得する事が比較的容易であると考えられるが，一方でテーマに対する知識をあまり持っていないユーザにとっては，それらの情報を重要であると判断する事が困難である．そこで我々はこれまで，このようにコミュニティに特有で且つ，コミュニティのテーマに対して重要な情報を潜在情報と呼び，コミュニティから潜在情報を抽出する方法を提案してきた．しかし，実験において提案手法の有用性を測ったところ，あまり良い結果が得られなかった．そこで本論文では，潜在情報抽出手法におけるコメントの重要度を求める手法である重要度計算手法を再構築すると共に，対象ドメインを「医療」として評価実験を行い，提案手法の有用性を示す．
キーワード　コミュニティ，潜在情報，重要度計算手法

Calculating Imortant Degree of Comments in User Gerated Content.

Keisuke UCHIMURA† and Akiyo NADAMOTO††

† Konan University, Okamoto 8–9–1, Higashinada-ku, Kobe-shi, Hyogo, 658–8501 Japan
†† Konan University, Okamoto 8–9–1, Higashinada-ku, Kobe-shi, Hyogo, 658–8501 Japan
E-mail: †mn924002@center.konan-u.ac.jp, ††nadamoto@konan-u.ac.jp

Abstract　Nowadays, there are many contents on the Internet, such as social net work services (SNSs) or blogs are created by general users. In such contents, users create communities based on a theme, and they exchange information about the theme. It is easy for users who know detail of the information of the theme to extract important information from the content. It is difficult, however, for users who do not know about the information of the theme to extract only important information from the content. We call specific and important information for the community as “hidden information”. We had proposed the method to extract it from the content. Our proposed method consists of extraction of difference method and importance degree method. Our proposed importance degree method, however, had some problem and we could not get good results from our experiment. In this paper, we improve the importance degree method. Then we had experiment by using medical data.
Key words　Community, Hidden Information, Importance Method

1. はじめに

近年インターネットにおいて，SNS やブログといった一般の人々によって作成されるコンテンツが多く存在し，それらは UserGeneratedContents（UGC）と呼ばれている．

UGC 上では，あるテーマに対して興味のあるユーザ同士がコミュニティを形成し，情報を共有しあっている．コミュニティに属するユーザはコミュニティ内で扱われるテーマに対しある程度の知識を持っていると予想される．よってユーザ同士でやり取りされる情報の中にはテーマに対して重要でかつ，一般の Web には載っていないような情報が多く存在すると考えられる．これらの情報は本来，そのテーマに対する知識を得たいユーザにとっても有益なものである．しかしながら，コミュニティに属していないユーザにとって，コミュニティ内でやり取りされる情報の重要性を判断する事は難しく，時間のかかる作業である．そこで我々は，コミュニティに特有で且つ重要な情報を潜在情報と呼び，UGC から潜在情報を抽出する手法を提案する．これまで我々は，一般の Web から得られる情報を基準情報と

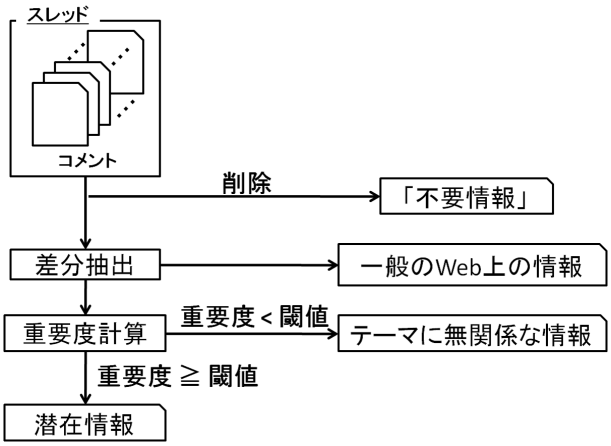

図 1　潜在情報抽出の手順

し，コミュニティ内の情報と基準情報との差分を抽出することでコミュニティに特有の情報を取得する差分抽出手法と，熟知度，貢献度，客観度という三つの尺度を用いて情報のテーマに対する重要度を計算する重要度計算手法を提案し，それら二つの提案手法を用いて潜在情報の抽出を提案してきた [1]．しかしながら，実験において提案手法の有用性を測ったところ，重要度計算手法に問題があり，あまり良い結果が得られなかった．そこで本論文では，重要度計算手法を再構築する．具体的にはコメントのタイプごとの熟知性を計算する「集散度」，コメントの理解し易さを表す「理解容易性」，そのコメントがどの程度客観的に書かれているかという「客観度」から，そのコメントの重要度計算を行う．

以下に潜在情報抽出の手順を示す（図 1 参照）．

（1） スレッドとスレッドのテーマを入力する．

（2） スレッドから不要情報を削除する．

（3） 差分抽出手法を用いて SNS のあるスレッドからコミュニティ特有のコンテンツを抽出する．

（4） 重要度計算手法を用いて，上記の 3 で抽出したコミュニティ特有のコンテンツから重要なコンテンツを抽出する．

（5） 上記の 4 で抽出したコンテンツを潜在情報とし，ユーザに提示する．

以下，2 章で関連研究について述べる．3 章では不要情報の削除について，4 章で差分抽出手法について述べ，5 章では重要度計算手法について述べる．そして 6 章で評価実験を行い，7 章でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 UGC からの情報抽出

ブログ等の一般のユーザが発信した情報から有益な情報を抽出する研究は数多く存在する．瀬藤ら [2] や立石ら [3] は，ブログやレビューサイトに書かれている何らかのテーマや商品に対するユーザの評価や評判の抽出を行っている．また佐々木ら [4] は，あるテーマに対する複数の賛成意見，反対意見の中から，構文パターンを用いて論点となる単語を抽出し，論点の可視化を測っている．これらの研究は，ユーザが発信した何らかのテーマに対する情報から評判や論点を得ることで，テーマの持つ本質の抽出を目的としている．

また，ユーザの発信した情報からそのユーザの経験を抽出し知識化する研究として倉島ら [5] の研究と乾ら [6] の研究が挙げられる．倉島らはブログ記事に記載されているユーザの経験情報を構造化し，それらの知識化を行っている．また乾らはブログ記事に記載されているユーザの経験情報から経験データベースを構築する研究を行っている．これらの研究は，ユーザの発信する経験情報から経験知の集積を目的としている．

さらに，ブログの情報からブロガーの熟知度を測る手法を提案する研究として中島ら [7] の研究と竹原ら [8] の研究が存在する．ブログの過去エントリを解析することでブロガーの熟知度を測っており，さらに中島らは求めた熟知度に基づいたブログのリランキング手法を提案している．

また一般のユーザが情報を発信する場として掲示板があるが，掲示板から情報を抽出する研究として岡村ら [9] の研究がある．岡村らはユーザの発信した情報を解析し，発言の重要度や会話の意味構造からユーザの興味のある会話を抽出している．

これらは全てユーザの発信した情報から有益な情報を抽出するという部分で我々の研究と関連している．我々の研究はコミュニティサイトからそのコミュニティが扱うテーマに対して重要な情報を抽出することで，そのテーマに対して興味があるユーザへの支援を目的としている．

2.2 情報難易度の推定

ある情報がユーザにとって理解容易な情報であるかを測る研究として，近藤ら [10] の研究と中谷ら [11] の研究が挙げられる．近藤らは独自に構築した言語モデルを用いて文書の難易度を測る研究を行っている．また，中谷らは可読性と専門性の二つの尺度を提案し，文章の理解容易性を判定している．

我々の研究では，コミュニティから重要な情報の抽出を行う際，その情報がユーザにとって理解容易な情報であるかを考慮している．しかし我々は文書に出現する専門用語の有無は重視せず，提案する専門性判断語を用いて，その情報の理解容易性を求める．

3. 不要情報の削除

SNS の中にはコミュニティのテーマと関係ないコメントや議論から無視されているコメント等コメントの重要度を測る上で，明らかに不要なコメントが存在している．我々はこの不要なコメントを不要情報と呼ぶ．そこで，潜在情報を抽出する前に，この不要情報を抽出し削除することにより計算時間の効率化を図る．我々は以下の 2 種類の不要情報を抽出し，削除する．

- 意味のないコメント

我々は，内容的に何を述べたいのか意味的に不明なコメントを「意味のないコメント」とするのではなく，コメントそれ自体が単語レベルで意味をなさないコメントを意味のないコメントとする．そこで，コメントが一つの単語のみで構成されているコメント及び複数の単語からなるが一つの品詞のみで構成されているコメントを意味のないコメントとし，それらのコメントを不要情報として削除する．ここで削除されるコメントとして

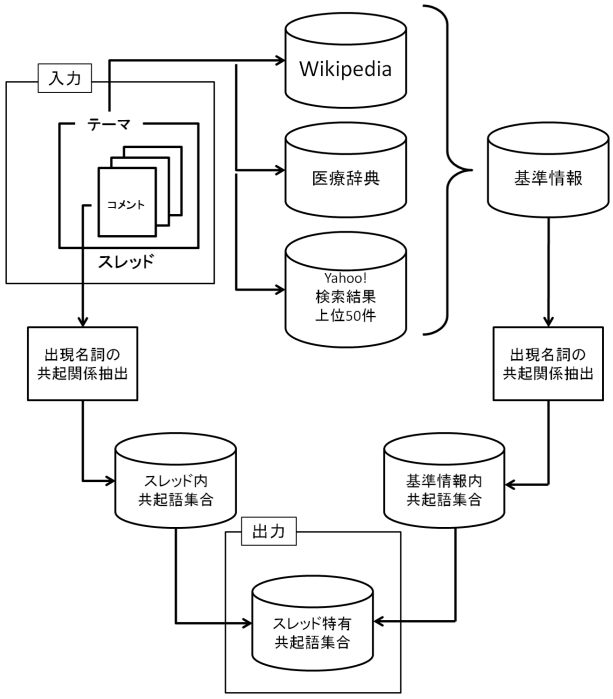

図 2　差分抽出の流れ
Fig. 2　Flow of Difference extraction

は，挨拶のみのコメント等が挙げられる．

- 孤立した質問コメント

一般に質問，応答はそのコミュニティにとって重要なコメントであると言える．しかしながら，応答の得られていない質問はそのコミュニティにとって重要であるとは言えないと考え，そのような質問のコメントを不要情報として削除する．質問コメントの抽出は，コメントの最後 5 単語内に”?”記号が含まれているコメントを質問コメントとし，コメント間の呼応関係抽出には，対話解析は用いず，コメント内の呼応に関する明示的な記述を抽出し，コメント群の呼応関係を抽出する．具体的にはコメント内にコメント番号やユーザ名を記述している，もしくはされているコメントを呼応関係のあるコメントとする．

4. 差分抽出手法

我々の提案する差分抽出手法とは，基準情報とコミュニティ内の情報とを比較し，コミュニティに特有の情報を抽出する手法である．基準情報は SNS のコミュニティが持つテーマに基づき選定する．本研究では対象ドメインを医療とし，基準情報は Wikipedia の記事，医学辞典の記事及び Yahoo!検索結果上位 50 件のサマリとする．ここで医学辞典は医学書院「医学大辞典」[12] と南山堂「医学大辞典」[13] を用いた．以下に差分抽出手法の手順を示す (図 2 参照)．

（1） 基準情報の一文の中のすべての名詞を抽出し，その名詞の組み合わせすべてを共起語対とする．基準情報内の全ての文から抽出した共起語対の集合を基準情報内共起語集合と呼ぶ．

（2） スレッド内のコメント中のすべての名詞を抽出し，その名詞の組み合わせをすべて共起語対とする．そしてスレッド内すべてのコメントから抽出した共起語対の集合をスレッド内共起語集合と呼ぶ．ここでは，一つ一つのコメントは情報量が少量である為，スレッドにおける共起語対は一つの文内で共起している名詞対を抽出するのではなく，一つのコメント内で共起している全ての名詞の共起語対を取得する．

（3） (1) で得られた基準情報内共起語集合と (2) で得られたスレッド内共起語集合から，下記の式を用いてスレッド特有共起語集合を抽出する．

$$\text{スレッド特有共起語集合} = \left(\overline{\text{基準情報内共起語集合}} \cap \text{スレッド内共起語集合}\right)$$

上記の式より得られた，スレッド特有共起語集合内の共起語対を含むコメントを「コミュニティに特有な情報」とする．

5. 重要度計算手法

我々の提案する重要度計算手法とは，コミュニティ内のあるコメントが，そのコミュニティのテーマに対してどの程度重要であるかを求める手法である．なお，本研究における潜在情報とは「コミュニティに特有でかつ重要な情報」である為，ここでは差分抽出手法により得られた「コミュニティに特有な情報」であるコメントのみを対象として重要度計算を行う．本提案手法では「集散度」「理解容易性」「客観度」の三つの尺度を提案し，それらの尺度からコメントの重要度を計算する．計算手順は以下の通りである．

（1） Wikipedia の記事を目次構造に従いセグメントに分割する．

（2） 得られたそれぞれのセグメントに対するコメントの網羅値を計算する．

（3） それぞれのセグメントに対するコメントの網羅値から分散値を計算する．

（4） 上記の 3 で得られた分散値より，コメントを「集中型コメント」と「分散型コメント」に分類する．

（5） 上記の 4 で分類したコメントのタイプごとに異なる方法で集散度を計算する．

（6） 上記の 5 で求めた集散度，コメントの理解容易性，客観度からコメントの重要度を計算する．

5.1 集 散 度

我々の提案する集散度とはそのコメントがどの程度コミュニティのテーマに対して詳しい情報であるかを求める尺度である．ここであるコメントがテーマに対して詳しい情報である時，そのコメントについて二つのタイプが考えられる．一つはテーマの中のある一部分について詳しく述べられているコメントである．このようにテーマの一部分に対して集中して書かれたコメントを「集中型コメント」と呼ぶ．もう一つはテーマ全体について詳しく述べられているコメントというのも考えられる．このようにテーマ全体について詳しく述べられているコメントを「分散型コメント」と呼ぶ．これらのコメントはタイプは違うがどちらもテーマに対して詳しいコメントであると言える．しかしタイプが異なる為，同じ手法によりそのコメントの詳しさを測る事が難しい．そこで我々はコメントのタイプごとに「テーマに対する詳しさ」として集散度を求める．具体的にはあるコ

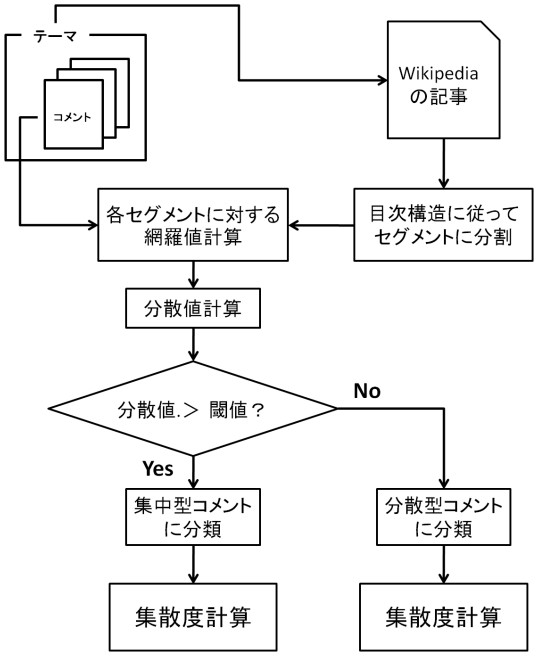

図 3　集散度計算の流れ

メントがテーマのどの部分について詳しく述べられているのかを，Wikipedia の記事とそのコメントを比較する事で計算する網羅値により求め，網羅値の分散によりコメントを「集中型コメント」と「分散型コメント」に分類する．集中型コメントと分散型コメントはそれぞれ異なる方法で集散度の計算を行う (図 3 参照)．

5.1.1　網　羅　値

網羅値とは，あるコメントがそのコミュニティのテーマが持つ情報をどの程度網羅しているかという値と定義する．そこで，あるコメントが Wikipedia の記事をどの程度網羅しているのかを Wikipedia の目次構造に着目して求める．網羅値は以下の手順により求める．

（1） Wikipedia から SNS のコミュニティのテーマに関する記事を取得する．

（2） 上記の 1 で取得した記事の目次構造に従って記事内の情報を分割し，分割された各部分をセグメントと呼ぶ (図 4 参照)．それぞれのセグメントに出現する名詞を全て取得する．

（3） あるコメントが，上記の 2 で取得したそれぞれのセグメントに出現する名詞を網羅している割合を網羅値とし，セグメント N_i におけるコメント j の網羅値 $cov(N_i)_j$ を求める．

ここで Wikipedia から取得したセグメントは，目次を表す木構造のうち葉節点である場合と子節点を持つ場合が考えられる (図 5 参照)．コメントの網羅値を求める際，子節点を持つ節点はその子節点の概要も含んでいる場合が多い為，その子節点の内容も加味した方がよいと考え，対象となるセグメントが葉節点である場合と子節点を持つ場合で違う計算方法を用いる．以下にそれぞれの求め方を述べる．

- 葉節点のセグメントに対する網羅値

あるセグメントが図 5 の節点 C のように葉節点である場合は，以下の式で網羅値 $cov(N_i)_j$ を計算する．

図 4　Wikipedia の記事をセグメントに分割

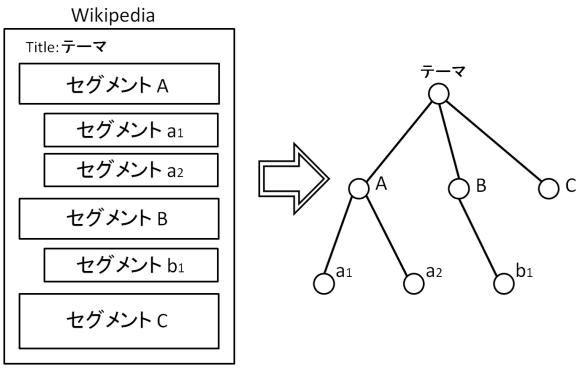

図 5　Wikipedia の記事を木構造化

$$cov(N_i)_j = \frac{num(N_i \cap j)}{num(N_i)} \tag{1}$$

ここで，$num(N_i \cap j)$ はコメント j がセグメント N_i に対して網羅している名詞の総数を示し，$num(N_i)$ は，葉節点となっているセグメント N_i 内の名詞の総数とする．

- 子節点を持つセグメントの網羅値

Wikipedia の記事の目次構造を木構造と捉えた時，親節点の内容はその項目に対する大まかな情報であり，子節点の内容は親節点の項目内容の一部分について細かい情報である場合が多い．ここで，親節点と各子節点の両方に存在する単語は，その子節点の項目内容に対し重要な単語であると考えられる．また子節点が複数存在する場合，それぞれの子節点と親節点の全てに存在する単語は，親節点の項目内容に対して重要な単語であると考えられる．これらのことから，網羅値は以下の式により計算する．

$$\begin{aligned} cov(N_i)_j &= \frac{num(N_i \cap j)}{num(N_i \cup n_1 \cup \cdots \cup n_m)} \\ &+ \sum_{k=1}^{m} \frac{num\{(N_i \cap n_k - N_i \cap n_1 \cap \cdots \cap n_m) \cap j\}}{num(N_i \cap n_k - N_i \cap n_1 \cap \cdots \cap n_m)} \\ &+ \frac{num(N_i \cap n_1 \cap \cdots \cap n_m \cap j)}{num(N_i \cap n_1 \cap \cdots \cap n_m)} \end{aligned} \tag{2}$$

ここで n_m は節点 N_i の子節点であり，m は節点 N_i が持つ子節点の総数を示す．

5.1.2　コメントの分類

コメント j の分散値 $var(j)$ は，5.1.1 項で述べた Wikipedia

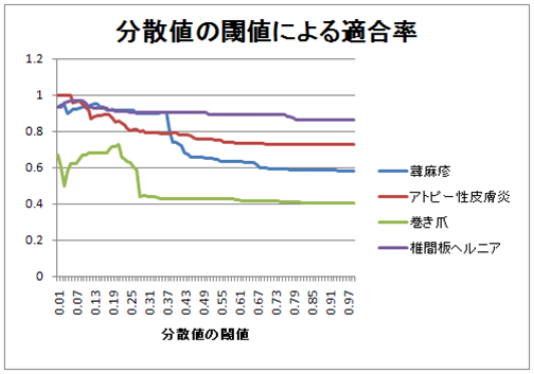

図 6　閾値毎の適合率

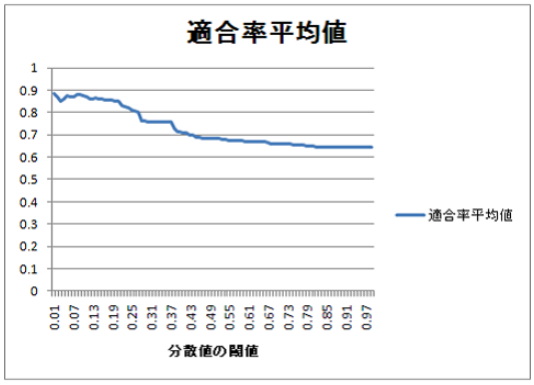

図 7　適合率の平均値

の記事に対するコメント j のセグメント N_i に対する網羅値 $cov(N_i)_j$ から網羅値の平均値 $ave(cov(N_i)_j)$ を計算し，それらを用いて以下の式により求める．なお式中の n は，Wikipedia 記事内のセグメントの総数を表す．

$$ave(cov(N_i)_j) = \frac{1}{n}\sum_{i=1}^{n} cov(N_i)_j \tag{3}$$

$$var(j) = \frac{1}{n}\sum_{i=1}^{n}\{ave(cov(N_i)_j) - cov(N_i)_j\}^2 \tag{4}$$

得られた分散値からそのコメントが「集中型コメント」か「分散型コメント」かを決定するために，本研究では分散値に閾値を設ける．これにより分散値が閾値以下のコメントを「分散型コメント」に分類し，分散値が閾値を上回るコメントを「集中型コメント」と分類する．なお今回，分散値の閾値は筆者一人により予備実験を行い決定した．以下分散値の閾値を決定するための予備実験について述べる．

分散値の閾値

分散値の閾値は以下の手順により求める．

（1）　スレッドから適当なコメントを 100 件抽出する．

（2）　内容的に分散していると思うコメントに対してチェックを入れる．

（3）　上記の 2 で作成した正解データと，閾値により分散型に分類されるコメントとの適合率を求め，適合率が 0.8 以上となる閾値の最大値を求める．

本研究では上記の手順を医療をテーマとする四つのスレッドに対して行った．各スレッドにおける閾値毎の適合率と，実験に用いた 4 スレッドの適合率の平均値をそれぞれ図 6，7 に示す．

得られた結果より，本研究では適合率の平均値が 0.8 以上となる閾値の最大値 0.26 を分散値の閾値として利用する．これにより分散値が 0.26 以下のコメントを分散型コメントとし，分散値が 0.26 を上回るコメントを集中型コメントに分類する．次に分散型コメント及び集中型コメントのそれぞれ集散度計算方法について述べる．

分散型コメントの集散度計算

分散型コメントとは，コミュニティの持つテーマ全体について書かれた内容のコメントである．よって集散度はそのコメントがテーマ全体に対してどの程度詳しいのかを測る必要があるため，分散型コメント j の集散度 CD_j は以下の式により計算する．ここで式中の n は Wikipedia の記事から得られるセグメントの総数であり，また $cov(N_i)_j$ はセグメント N_i に対するコメント j の網羅値を表す．

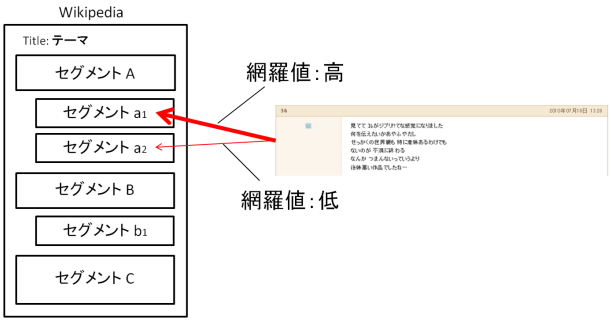

図 8　集中型コメントの例

$$CD_j = \sum_{i=1}^{n} cov(N_i)_j \tag{5}$$

集中型コメントの集散度計算

集中型コメントとは，コミュニティの持つテーマのある一部分について書かれた内容のコメントである．我々は集中型コメントの集散度を計算する際，テーマ全体に対しての詳しさを測るのではなく，そのコメントが対象としているテーマの一部分についてどの程度詳しく書かれた内容のコメントであるかを計算する．これは先にも述べたように，テーマの一部分について詳しく書かれた内容のコメントも重要なコメントであると言え，よってテーマの全体についてどの程度詳しい内容であるのかを測るのではなく，そのコメントが対象とするテーマの一部分についてどの程度詳しい内容であるかを測るのが妥当であると考えた為である．

しかし集中型コメントの集散度を計算する際は注意が必要である．それは例えば図 8 に示すように，あるコメントの網羅値が Wikipedia の記事中のある一つのセグメント a_1 に対してのみ非常に高い網羅値であり，他のセグメントに対しては低い網羅値である場合，網羅値の平均が高くなり，これにより分散値も高くなる．このような場合，そのコメントは集中型コメントに分類されるが，セグメント a_1 以外のセグメントに対しては網羅値も低く，対象としているセグメントは a_1 のみであると

言え，よって僅かにしか述べられていないその他のセグメントはそのコメントの対象になっているとは考え難い．以上の事から，分散値が閾値を上回る集中型コメントに対しては更に網羅値の閾値を用いることにより，そのコメントが対象とするセグメントを判断する．網羅値の閾値を決定するために，予備実験を行った．尚，本予備実験は筆者一人によるものである．

網羅値の閾値は以下の手順により求める．

（1） スレッドから分散値が 0.3 以上のコメントを全て抽出する．

（2） 各コメントと Wikipedia の記事を見比べ，上記の 1 で抽出された全てのコメントに対し，そのコメントの対象となっていない Wikipedia の項目にチェックを入れる．

（3） 上記の 2 でチェックを入れた項目に対する網羅値の平均値を求める．

上記実験により，網羅値の平均値 0.08 が得られた．得られた値は，コメントが対象としていない Wikipedia の項目に対する網羅値の平均であるため，本研究では集中型コメントが対象とするセグメントは，網羅値が 0.08 以上のセグメントとする．

集中型コメントの集散度計算

我々が提案する集中型コメント j の集散度 CD_j は，Wikipedia の目次項目 N_i に対する網羅度 $cov(N_i)_j$ を用いて計算する．ここで式中の n_j は Wikipedia の記事中のセグメントの内，コメント j の網羅値が 0.08 以上となるセグメントの数を示す．

$$CD_j = \sum_{i=1}^{n_j} cov(N_i)_j \tag{6}$$

5.1.3 理解容易性

理解容易性とは，コミュニティ内のあるコメントが，そのコミュニティに属していない外部ユーザにとって理解し易いものであるかを測る尺度である．一般に理解容易性を判断するには専門用語に注目して求める場合が多いが，我々は専門用語が羅列されているだけの情報が熟知性の高い情報とは限らないと考え，専門用語には注目しない．その代わり，我々の提案する理解容易性判断語を用いる．理解容易性判断語は Q&A サイトにおいて，ベストアンサーに選ばれている回答はその分野において専門性が高く且つ分かり易いコンテンツであると考え，このコンテンツを解析することにより抽出する．本論文では Q&A サイトに Yahoo!知恵袋(注1)を使用する．以下に理解容易性判断語抽出の流れを示す．

（1） Q&A サイトの質問の中から「〜について教えて下さい」等の，熟知性の高い回答を求める質問を人手により収集する．

（2） 上記の 1 で収集した質問のベストアンサーから「名詞，形容詞，形容動詞，動詞，助動詞，記号」以外の出現単語を取得する．

（3） 上記の 2 で得られた単語からストップワードを除き，出現頻度が閾値以上の単語を理解容易性判断語とする．

今回，理解容易性判断語を得る為に Yahoo!知恵袋から収集した質問の数は 42 件である．これらのベストアンサーから上記の手順により，115 種類の単語を抽出した．単語が理解容易性判断語として適当か否かは全て人手により判断した．これにより得られた 115 種類の単語の内，42 件中 5 件以上のベストアンサーに出現する 27 語を理解容易性判断語としている．得られた理解容易性判断語を表 1 に示す．また表 1 の理解容易性判断語の種類は，得られた理解容易性判断語を分類する際に人手により付けたラベルである．

表 1　理解容易性判断語

種類	理解容易性判断
比較	より・よりも，に対し・に対して
制限	の時，なら，だけ，まで，ただ
付加	更に，例えば，以外，など，また，ちなみに，他に
因果	つまり，よって，による，から，ので
対象	について，そういう，として，同じ，という
推測	かも
打消	あまり
その他	でも

表 1 の理解容易性判断語は「〜について教えて下さい」といった，熟知性の高い回答を求める質問のベストアンサーに選ばれたコメントに多く含まれている事から，これらの単語を多く含む情報は，テーマに対してあまり知識のないユーザにとっても理解し易いと考えられる．よって定義した理解容易性判断語は重要度を測る尺度として適当であると考えられる．

5.2 客　観　度

客観度とは，あるコメントが客観的な立場で書かれているかを示す尺度であり，コメント内の形容詞の出現頻度から計算する．形容詞の出現頻度が高いコメントは客観性を欠くコメントであり，万人に受け入れられる情報ではないと考えられる．本研究ではコメント j の客観度 OD_j はコメント内の形容詞の出現頻度とする．

5.3 重要度計算

重要度計算は集散度 CD_j と，理解容易性判断語の出現頻度 $judg(j)$，客観度 OD_j を用いて計算する．コメント j の重要度 ID_j は以下の式により求める．

$$ID_j = CD_j * judg(j) * \frac{1}{OD_j} \tag{7}$$

上記の計算式により得られる重要度が閾値以下のコメントをテーマに無関係な情報とし，閾値を超えるコメントを潜在情報としてユーザに提示する．また本研究では分散型コメントと集中型コメントにそれぞれ別の閾値を用いることで，分散型コメントにおける潜在情報と集中型コメントにおける潜在情報をそれぞれ抽出する．

6. 実　　験

本論文では二つの実験を通して提案手法の有用性を測った．まずノイズとなるコメントが正しく抽出できているかを調査する為の実験を行い、次に潜在情報の抽出に関する実験を行った．

6.1 実験１：ノイズコメントの抽出

本実験では，掲示板サイト内の映画に関するスレッドを用い

(注1)：http://chiebukuro.yahoo.co.jp/

表 2　ノイズ抽出における再現率・適合率

タイトル	コメント数	再現率	適合率
ダ・ヴィンチ・コード	38	0.613	0.87
アイ・アム・レジェンド	48	0.647	0.73
アバター	10	1.0	1.0
ゲド戦記	89	0.656	.0857
20 世紀少年	115	0.695	0.852

表 3　F 値の各テーマ別最高値

テーマ	分散型コメントの閾値	集中型コメントの閾値	適合率	再現率	F 値
アトピー性皮膚炎	0.1	0.003	0.675	0.355	0.466
片頭痛	0.001	0.013	0.621	0.878	0.728
椎間板ヘルニア	0.006	0.007	0.409	0.4	0.404
蕁麻疹	0.04	0.007	0.54	0.662	0.595
痔	0	0.013	0.259	1.0	0.411
巻き爪	0.003	0.021	0.278	0.82	0.415

た．筆者が人手によりコメントを不要情報とそれ以外に分類し，システムの抽出した不要情報と比較することで，適合率，再現率を測っている．なおノイズコメントの抽出実験においては「学生コミュニティポータルサイト キャスフィ！(注2)」を用いた．実験の結果は表 2 に示す通りである．平均再現率が 72.2%，平均適合率は 86.2%という結果が得られており，これにより不要情報の削除においては効果的にノイズコメントを抽出できていると言える．

6.2　実験 2：潜在情報の抽出

データセット

提案する潜在情報抽出手法の有用性を測るため，mixi 内の「医療」をテーマとするスレッドをデータセットとし 7 名の被験者に対し実験を行った．任意のスレッドであるアトピー性皮膚炎，片頭痛，椎間板ヘルニア，蕁麻疹，痔，巻き爪をテーマとする 6 つのスレッドを用いて実験を行った．また本実験では基準情報として，各スレッドのテーマに関連する Wikipedia の記事，医学書院「医学大辞典」[12]，南山堂「医学大辞典」[13]，Yahoo! 検索結果上位 50 件のサマリを用いた．実験の手順は以下のように行った．

（1） mixi の医療をテーマとするコミュニティから任意のスレッドを被験者に提示する．

（2） 被験者は提示されたスレッド内の重要であると思うコメントを選択する．

（3） 上記の 2 の結果のうち，被験者の過半数が選択したコメントを正解データとする．

（4） プロトタイプシステムを用いて，潜在情報を抽出する．

（5） システムの抽出した潜在情報を上記の 3 で得られた正解データと比較し，適合率，再現率，F 値をそれぞれ求める．

上記の手順により実験を行った結果を表 3 に示す．表 3 はそれぞれの F 値が最も高い時の結果である．

(注2)：http://www.casphy.com

6.3　考　　察

実験の結果，表 3 から片頭痛，蕁麻疹に関しては良い結果が得られたと言える．しかしその他のテーマに関しては結果が悪く，提案手法の有用性は確認できなかった．良い結果の得られなかった 4 つのスレッドに関しては以下の特徴が挙げられる．

- 民間療法に関するコメントが多い．
- 業者からの書き込みが多数存在する．

民間療法に関するコメントはユーザの実経験に伴う重要な情報であると言えるが，あくまで民間療法の域を出ない治療法である為，基準情報からは民間療法に関する単語の抽出が行えず，それらはユーザの実経験に伴う場合が多いため，表記揺れにもつながり，その結果重要度をうまく計算できなかった．またアトピー性皮膚炎等の有名でかつ大きな病気に関しては業者からのコメントが複数見受けられた．業者からのコメントは，基準情報から得られるテーマに関連する単語を多く含み，システムは重要なコメントであると判断しがちである．しかし業者のコメントはそのほとんどが，病気に悩むユーザを勧誘する為のコメントであり，特に重要な治療法を記す内容のコメントではない為，被験者からは重要でないと判断され，その結果システムの抽出した結果の適合率の低下を招いたと考えられる．

7.　まとめと今後の課題

本論文では UGC 上に存在するコミュニティから潜在情報を抽出する手法を提案した．基準情報とコミュニティ内の情報を比較し差分を取ることで「コミュニティに特有な情報」を抽出する差分抽出手法と，コミュニティのテーマに対するコメントの重要度を計算する重要度計算手法について述べ，差分抽出手法と重要度計算手法を用いた潜在情報抽出を行い，実験により提案手法の有用性を確認した．なかでも重要度計算手法ではコメントの重要度を計算する為の新たな尺度とし，集散度を提案した．これはあるコメントがテーマの持つ情報に対して分散した内容のコメントであるか，また集中した内容のコメントであるかをコメントの分散値計算により求め，それぞれのタイプ別に，テーマに対する詳しさを測る為の尺度である．そして実験を行った結果，表記揺れ等の問題があり全体としてあまり良い結果が得られなかった．

今後は表記揺れに対応するための方法を考え，また基準情報として選定するデータに関しても議論の必要があると考えられる．現段階では Wikipedia やテーマに関する辞書的なデータを主にテーマの基準情報としているが，今後は民間的な知識や経験則のようなデータも基準情報とすることでより高い精度の潜在情報抽出が可能になると考えられる．そして今回，理解容易性判断語は一つの Q&A サイトから抽出したが，今後は別の Q&A サイトでも抽出を行い，より精度の高い理解容易性判断語を抽出する必要がある．

謝　　辞

本研究の一部は，平成 22 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) によるものである．ここ

に記して謝意を表します.

文　　献

[1] 内村圭佑，灘本明代，”User-Generated Content における潜在情報抽出手法の提案 ”第 151 回データベースシステム研究会，4C-42，Nov. 2010
[2] 瀬藤亮，佐藤哲司，”商品説明ページを用いた評価視点別評判情報提示システム ”，電子情報通信学会データ工学研究専門委員会他共催，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)，C6-5，Mar. 2009
[3] 立石健二，石黒義英，福島俊一，”インターネットからの評判情報検索 ”，人工知能学会誌，Vol.19, No 3, 2004.
[4] 佐々木千晴，藤井敦，石川徹也，”意思決定支援のための主観情報マイニング ”，言語処理学会第 12 回年次大会発表論文集，pp.77-80, Mar. 2006
[5] 倉島健，藤村考，奥田英範，”大規模テキストからの経験マイニング ”，電子情報通信学会論文誌 D Vol.J92-D No.3 pp.301-310, 2009.03
[6] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 314-321, Dec. 2008.
[7] 中島伸介，稲垣陽一，草野奉章，”ブロガーの熟知度に基づいたブログランキング方式の提案 ”，電子情報通信学会第 19 回データ工学ワークショップ，第 6 回日本データベース学会年次大会 (DEWS2008) 論文集, 2008.
[8] 竹原幹人，中島伸介，角谷和俊，田中克己，”Web 情報検索のための Blog 情報に基づくトラスト値の算出方式 ”，日本データベース学会論文誌 (DBSJ Letters) ,Vol.3, No.1, pp.101-104, 2004 年 6 月
[9] 岡村剛，角康之，西田豊明，”電子掲示板からの興味ある会話の抽出支援 ”，インタラクション 2005，情報処理学会主催，2005
[10] 近藤陽介，松吉俊，佐藤理史，”教科書コーパスを用いた日本語テキストの難易度推定 ”，言語処理学会 第 14 回年次大会 (NLP2008), D5-5. 2008 年 3 月
[11] 中谷誠，アダム・ヤフト，田中克己，”理解容易性を考慮した用語説明のランキング手法 ”，Web とデータベースに関するフォーラム (WebDB Forum)2009, 3A-3, 2009 年 11 月
[12] 伊藤正男，井村裕夫，高久史麿，“ 医学書院 医学大辞典 ”, 医学書院, 2003.
[13] “ 南山堂 医学大辞典 ”, 南山堂, 2006.