

掲示板からの深層情報抽出手法の提案

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 SNS やブログ等に代表される、ユーザから発信された情報により構成されるコンテンツ (UGC) には、今までの Web には無かったような重要な情報が多く存在している。しかしそういった重要な情報は、同じ趣味・趣向を持つユーザ同士が形成しているコミュニティ内に埋もれがちであり、コミュニティ外のユーザにとってそれらを取得するのは多くの時間を要する作業となる。そこで本研究では、このようなコミュニティに特有でかつ重要な情報を深層情報と呼び、この深層情報を抽出する手法を提案する。本論文ではその第一歩とし「第三者同士がある一つのテーマに対して議論する」という性質を持つ「掲示板」を対象とし、二つの提案手法を用いて「掲示板に特有でかつ重要なコメント」を深層情報として抽出する。これにより掲示板上的コメントをすべて読むことなく、一般の Web には書いていない重要な情報を自動的に取得する事が可能となる。さらにどんなコメントが既に書き込まれているかを解析することで、その他の情報の書き込みを促すことができ、掲示板の活性化が期待できる。

キーワード UGC, コミュニティ, 深層情報, 掲示板

Extracting Deep Content from the BBS

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 Today, we can exchange information easily on BBS and SNS, when a lot of people post a small sentence based on a same theme. Therefore, unexpected content on UGC is more than on the general Web content. We call the content which is difference between UGC and general Web content, as deep content. In this paper, we propose extracting the deep content from the BBS. In particular, we extract familiarity-content and experiment-content based on data structure in the BBS. We regard the familiarity-content and experiment-content as deep content. By using our system, user can get the information, even if they don't read the content well.

Key words BBS, Experiment-content, Familiarity-content, Deep Content

1. はじめに

インターネット上の掲示板や SNS はある 1 つのテーマについて、参加者同士がコメントを書くことにより情報交換を行い、ネット上のコミュニティを形成している。これらのコンテンツは一般のユーザによって作成されており、UserGeneratedContent (UGC) と呼ばれている。特に、掲示板や SNS は通常の Web コンテンツと比較して、経験的に書かれた情報やテーマに対して精通した情報など、様々な観点からの情報が投稿されている場合が多い。これらの情報は一般の Web にはほとんど記載されておらず、掲示板や SNS 特有の情報であるといえる。しかしながら、これらは重要な情報であるにも関わらず、掲示板や

SNS におけるコメント間の関係の複雑さや読みにくさから、容易に内容を把握することが困難なため、ユーザに認識されずにコミュニティ内に埋もれている場合が多い。例えば、何らかのテーマに対する情報を探しているユーザが、掲示板に、探しているテーマについて書かれたスレッドを発見しても、そのスレッドの中から重要な情報だけを取得するのは容易ではなく、スレッド上の数百に上るコメントの中から自分の欲している情報を探すのにはかなりの時間を要する。さらに、既に多くのコメントが存在するスレッドに対して新たにコメントを投稿する場合、自分の投稿しようとしている情報と重複する内容のコメントが既に存在していないかを調べるにも多くの時間を要する。このため、重複する内容のコメントが多数存在するという事態

が生じている。つまり掲示板や SNS は重要な情報を探すにしても、情報を投稿するにしても多くの時間がかかってしまうという問題があり、その結果重要な情報は掲示板や SNS を構成するコミュニティ内に埋もれてしまいがちであるという現状がある。

本研究ではこのような、一般の Web にはほとんど記載されていない、コミュニティに特有でかつ重要な情報を深層情報と呼び、深層情報を抽出することを目的とする。

本論文では深層情報抽出の第一歩として、「映画」をテーマとする掲示板を対象とし、掲示板から得られる情報と一般の Web で得られる情報との差分を取ることで掲示板特有の情報を抽出する差分抽出手法と、あるコメントのテーマに対する重要度を計算する重要度計算手法の 2 つの手法を提案し「ゴミ情報、孤立質問、一般の Web 上の情報、テーマに無関係な情報、深層情報」の 5 つに種類分けされるスレッド内のコメントの中から、深層情報の抽出を行う。深層情報を抽出することにより、多量なコメントの中から多くの時間を要することなく重要な情報を取得する事が可能となる。さらに得られた深層情報から、スレッド内で既にどのような情報が存在しているのかが判断できるため、ユーザが新たにコメントの投稿を行う際に、既に存在する重要な情報以外の情報の投稿を促すことでスレッド内の情報をより充実させることが可能になると考えられる。

以下、2 章で関連研究について述べ、3 章で深層情報を抽出するための二つの提案手法について述べる。4 章で評価実験を行い、5 章で考察を述べる。6 章ではまとめと今後の課題について述べる。

2. 関連研究

2.1 評判検索

ブログやレビューサイトで書かれている、あるテーマや商品に対するユーザの評判や評価を抽出する研究として、瀬藤らの研究 [1]、立石らの研究 [2] 等が挙げられる。瀬藤らはクチコミサイトにおける評判情報抽出の際、商品説明ページから抽出した評価視点ごとにクチコミサイトユーザの意見を分類し、提示することでクチコミサイトの閲覧支援システムの提案を行っている。また、立石らは対象物に対する評価表現とその極性の組を登録した辞書（評価表現辞書）を人手により構築し、辞書と単語のパターンマッチングルールにより、Web ページから抽出した対象物の意見を肯定意見と否定意見に分け、提示するシステムの提案を行っている。評価表現辞書に関しては、Web 上の文書を解析し、独自に提案した単語の共起パターンや語彙統合パターンから自動的に辞書の規模拡大を行っている研究なども存在する ([3], [4])。

一方佐々木ら [5] 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、固有度と重要度の二つの尺度に基づいたスコアリングを行うことで、論点の可視化を図っている。またその論点がコミュニティ内で賛成寄りなのか反対寄りなのかを提示することで、ユーザの意思決定支援となるシステムの提案を行っている。

これらの研究では第三者の意見や論点を抽出する際、構文パ

ターンを用いて対象物に対する特有の単語とその評価となる単語の抽出を行っている。しかし本研究では「第三者同士があるテーマに沿って意見を投稿する」といった特徴を持つ、掲示板を対象としていることから、全ての情報が第三者の意見として捉えられるため、構文パターンを用いた意見抽出は行っていない。また、これらの研究では、意見や論点を評価に沿って分類する際、第三者が下した評価に基づいて分類を行っているが、本研究は対象物の評判を抽出する事が目的ではなく、あくまであるテーマの深層情報、すなわち一般の Web 上でほとんど記載されていない、テーマに対して重要な情報を抽出する事が目的である。よって第三者の主観的な評価の解析までは踏み込まず、客観的な評価尺度を用いた第三者の意見分類を行う。これにより、万人に対して重要な情報の抽出を試みる。

2.2 経験情報マイニング

倉島ら [6] は、経験情報を状況（時間、空間）、行動（動作、対象）、主観（評価、感情）から成る情報と定義し、評価を除いた 5 要素（時間、空間、動作、対象、感情）を抽出することで、非構造データであるブログ記事に記載されている経験情報を構造化している。さらに、感情を 8 つのカテゴリに分類し、経験情報が動作主にとって成功だったか失敗だったかを導き出すことで、構造化された経験情報の知識化を行っている。また乾ら [7] は、ブログ上の経験に基づく情報から、表現やイベントの種類等を抽出し、自動的に意味的な分類を行うことで経験データベースを構築する研究を行っている。これらの研究では、ユーザの発信した情報を解析し、UGC 特有の情報の抽出を行い、得られた知識を体系化している。

比べて本研究では抽出する情報をユーザの経験に基づく情報に限らず、一般の Web 上には存在しないテーマに対して重要な情報を深層情報としており、またコメントの分類を行う際も、ユーザの感情等は考慮せず分類を行っている点でこれらの研究とは異なる。

2.3 掲示板からの情報抽出

岡村ら [8] は、ツリー構造を持つ掲示板の発言の重要度や会話の意味構造を可視化し、そこからユーザの興味のある会話を抽出する研究を行っている。この研究では、掲示板内で返信を行っている発言が、返信先の発言からどの程度の影響を受けているかを出現単語から測り、その影響量を発言の重要度として掲示板の解析を行っている。しかし、この影響量は発言同士のレスポンス関係が分かりやすいツリー構造型の掲示板においては計算が可能であるが、例えば「2ちゃんねる」等の非構造型掲示板では発言のレスポンス関係を抽出できない場合が多く、同じ手法を用いた計算は困難であると考えられる。以上のことから本研究では、コメントの重要度に関して、コメント間のレスポンス関係も加味した独自の計算手法を提案し、コメントの重要度計算を行う。

3. 深層情報抽出手法

深層情報を抽出する手法として (1) 掲示板特有の情報を抽出する差分抽出手法と、(2) あるコメントのテーマに対する重要度をスコアリングする重要度計算手法を提案する。

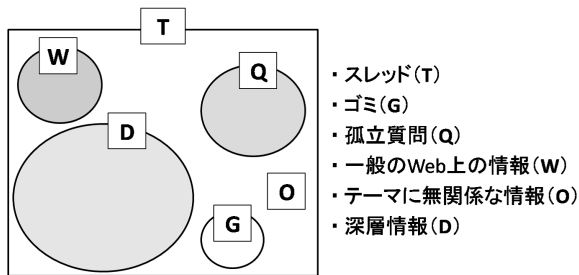


図 1 コメントの分類

Fig.1 Classification of Comment on BBS

この2つの手法は、一般の Web 上の情報を基準としてスレッドの情報解析を行う。この一般の Web 上の情報を基準情報と呼ぶ。また、本論文では「映画」をテーマとするスレッドを対象とするため、基準情報は映画の情報を調べる際に閲覧する Web ページから取得する。そして映画の情報を調べる際に閲覧する基準情報を、Wikipedia、レビューサイト (allcinema^(注1)、キネマ旬報^(注2))、Yahoo!検索における検索結果上位 50 件のサマリとする。我々は掲示板を「ゴミ情報、孤立質問、一般の Web 上の情報、テーマに無関係な情報、深層情報」の 5 つの種類のコメントから成ると定義する (図 1 参照)。そこで深層情報を抽出する流れとして、まず「ゴミ情報」、「孤立質問」に分類されるコメントの削除を行う。次に残ったコメントを差分抽出手法により、「一般の Web 上の情報」と「掲示板に特有の情報」に分類する。最後に「掲示板に特有の情報」に分類されたコメントを重要度計算手法により「テーマに無関係な情報」と「深層情報」に分類する。これにより得られた「掲示板に特有」でかつ「テーマに対して重要」なコメントを深層情報とし、ユーザに提示する。以下、深層情報抽出の流れに沿って詳しく説明する。

3.1 「ゴミ情報、孤立質問」の削除

本研究では「ゴミ情報」「孤立質問」を以下のように定義する。

- ゴミ情報：出現する名詞が一つ以下のコメント (ユーザ同士の挨拶のみ等)。
- 孤立質問：文末に疑問符があるコメントの内、回答が得られていないコメント。

また質問に対して回答が存在しているか否かは、掲示板に特有の呼応表現に注目することで簡易的に判断しており、自然言語処理を用いたコメントの内容的呼応関係からは判断していない。

本研究において、これらに分類されるコメントは必要ではない為、処理速度を上げるために初めに削除しておく。この処理により、残ったスレッド内のコメントは「一般の Web 上の情報、テーマに無関係な情報、深層情報」の 3 種類となる。

3.2 差分抽出手法

我々の提案する差分抽出手法とは、スレッド内の情報と基準情報の差分を取ることで、そのスレッドにおいて「特有」な情報を抽出する手法である。

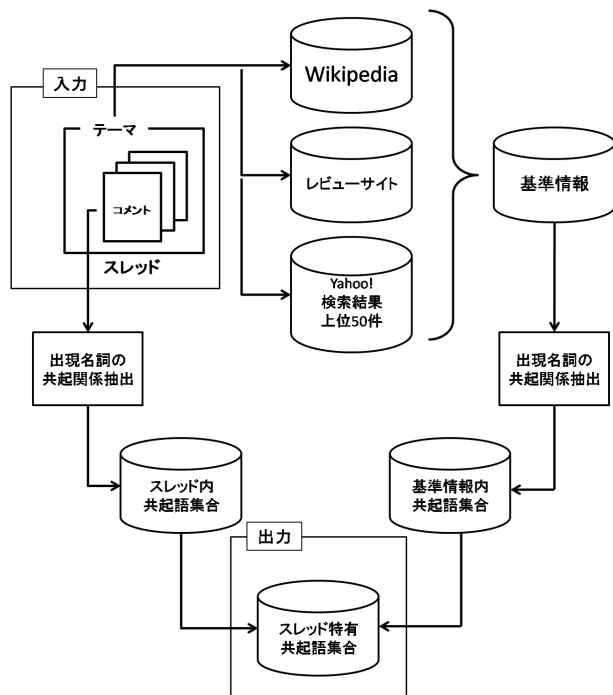


図 2 差分抽出の流れ

Fig.2 Flow of Difference extraction

抽出する際の差分抽出手法の処理を図 2 に示し、以下に手順を示す。尚、入力はいくつかのコメントから成るスレッドとそのテーマとする。

(1) テーマとなる映画のタイトルから Wikipedia の記事、レビューサイトの記事、Yahoo!検索における検索結果上位 50 件のサマリを取得し、これらを基準情報とする。

(2) 形容詞は主観情報が多く含まれると考え、本研究では客観情報を取得したい為、名詞のみを対象とし、全ての基準情報内の名詞の共起関係を抽出する。共起関係にある名詞の対を一つの情報とし、全文から抽出した名詞の対を総合して基準情報内共起語集合と呼ぶ。

(3) スレッドに対しても名詞の共起関係を求め、スレッド内共起語集合を生成する。一つ一つのコメントは情報量が少量であることから、スレッドにおける共起語集合は、一つのコメント内での名詞の共起関係とする。

(4) 得られた基準情報内共起語集合とスレッド内共起語集合から、以下の式よりスレッド特有共起語集合を出力する。

スレッド特有共起語集合

$$= (\text{基準情報内共起語集合} \cap \text{スレッド内共起語集合})$$

上記より得られた結果の一例を表 1 に示す。ここで表中の「重なり」とは、基準情報及びスレッドから得られる共起関係の内、両方に属する共起関係の事を指す。これにより得られたスレッド特有共起語集合内の共起語を含むコメントを「スレッドに特有な情報」とし、得られた共起語を含まないコメントを「一般の Web 上の情報」に分類する。また、「スレッドに特有な情報」に分類されるコメントの中には「テーマに無関係な情報」と「深層情報」がある。

(注1): <http://www.allcinema.net/prog/index2.php>

(注2): <http://www.kinejun.jp/>

表 1 差分抽出手法結果例

Table 1 Example : Results of Difference extraction

映画タイトル	コメント数	スレッド内名詞共起数	基準情報内名詞共起数	重なりの数	スレッド内特有情報数
「20 世紀少年」	115	8074	42084	128	7946
「レッドクリフ」	8	424	14472	26	398
「アイ・アム・レジェンド」	48	458	5602	12	446
「ゲド戦記」	89	4172	13726	174	3998
「ターミネーター 4」	41	676	5464	26	650

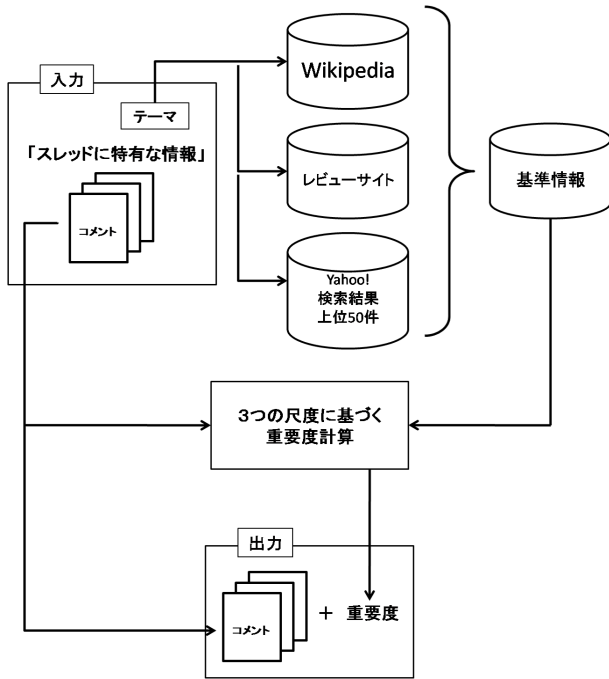


図 3 重要度計算の流れ

Fig.3 Flow of Scoring

3.3 重要度計算手法

我々の提案する重要度計算手法とは、3つの尺度「テーマに対する熟知度、議論に対する貢献度、客観度」を用いてスレッドにおけるコメントの重要度を測る手法である。入力、差分抽出手法により得られた「スレッドに特有な情報」に分類されるいくつかのコメントとそのテーマであり、テーマから基準情報を抽出するまでの流れは3.1の差分抽出手法と同じである(図3参照)。以下それぞれ3つの尺度について説明する。

(1) テーマに対する熟知度

あるコメントがテーマに対し、どの程度の知識を持って書き込まれたコメントであるかを測る尺度を熟知度とする。実際には、基準情報内に用いられている名詞を多く用いているコメントはテーマに対してより深い知識を持っているとし、コメントと基準情報との類似度をテーマに対する熟知度とする。計算は以下の式により行う。

(a) tf 値

$$tf^i = \frac{\log_2(freq(i)+1)}{\log_2(\text{文章中の形態素数})}$$

ここで、 $freq(i)$ はキーワード $i(i=1, 2, \dots, n)$ の出現頻度である。

(b) コメント (j) における特徴ベクトル

$$V_j^c = (tf^1, \dots, tf^n)$$

(c) 基準情報における特徴ベクトル

$$V^b = (tf^1, \dots, tf^n)$$

(d) コメント (j) の熟知度 $DK(j)$

(コメント (j) と基準情報の類似度)

$$DK(j) = \text{sim}(V_j^c, V^b) = \frac{V_j^c \cdot V^b}{\|V_j^c\| \cdot \|V^b\|}$$

(2) 議論に対する貢献度

あるコメントがどの程度スレッド内の議論に対して貢献しているかを測る尺度をその議論に対する貢献度とする。貢献度はスレッド内でのコメントの呼応関係に着目し、そのコメントがレスポンスをしている、もしくはレスポンスされている場合、そのコメントは議論に対して貢献していると判断する。ここでの呼応関係とは、コメントの内容から判断するのではなく、他のコメントを参照する際に用いられる「>>(参照するコメントの番号)」等の掲示板独特の表現から判断する。また、テーマに沿っていないコメントの呼応関係は議論に対して貢献していないと考えられるため、呼応関係にあるコメントであっても熟知性に欠けるコメントの貢献度は0とする。以下にコメント j の貢献度 $CD(j)$ を示す。ここで、 $DK(j)$ とはコメント j に対する類似度(熟知度)である。

$$CD(j) = \begin{cases} DK(j) * 0.1 & \text{呼応関係がある場合} \\ 0 & \text{呼応関係がない場合} \end{cases}$$

(3) 客観度

コメントが客観的な立場で書かれているかを測る尺度であり、コメント内の形容詞の出現頻度に注目する。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。よって形容詞の出現頻度がある閾値を超えるコメントの重要度には減点を行う。以下にコメント j に対する客観度 $OD(j)$ を示す。ここで、 $AT(j)$ はコメント j 内の形容詞の出現頻度を示し、 α はその閾値を示す。

$$OD(j) = \begin{cases} DK(j) * (-0.1) & \text{if } AT(j) > \alpha \\ 0 & \text{otherwise} \end{cases}$$

以上3つの尺度を用いてコメントの重要度計算を行う。コメント j の重要度 $ID(j)$ は以下の式により求める。

$$ID(j) = DK(j) + CD(j) + OD(j)$$

得られたコメントの重要度をもとに、差分抽出手法により得ら

れた「スレッドに特有な情報」を「テーマに無関係な情報」と「深層情報」に分類する．具体的にはコメントの重要度が閾値 β 未満のコメントを「テーマに無関係な情報」とし、閾値 β 以上のコメントを「深層情報」とする．

これにより得られた「深層情報」をユーザに提示する．

4. 実 験

提案手法では、掲示板のスレッドと一般の Web ページを比較し、その差分情報から深層情報を抽出しているため、基準情報となる一般の Web ページの選定は非常に重要であると考えられる．そこで本章では、まず前章で述べた Wikipedia の記事、レビューサイトの記事、Yahoo!検索結果上位 50 件のサマリの 3 種類の Web サイトそれぞれ 1 つずつを基準情報とした時の再現率・適合率を算出した．次に我々の提案するこれら 3 種類の Web サイトを合算して基準情報とした時のシステムの評価実験を行い、個々の実験の結果と比べる事で提案手法の基準情報の妥当性を調べ、更にシステムの有用性を測った．

具体的には本実験では以下の評価を行った．(1) システムのコメント分類が人手により事前に分類したコメント分類をどの程度再現しているか、(2) システムのコメント分類が適切であるか、の 2 項目を評価する．尚、今回評価するスレッドは「学生コミュニティポータルサイト キャスフィ!」^(注3)の映画掲示板から取得した．

システムのコメント分類の正誤を判別するために、事前に人手でスレッド内のコメントを「ゴミ情報、孤立質問、一般の Web 上の情報、テーマに無関係な情報、深層情報」の 5 つに分類する．今回この分類作業は筆者 1 人により行った．以下に人手による分類作業の要領を示す．

- ・ 挨拶のみ等のテーマに全く関係のないコメントは「ゴミ情報」に分類する．
- ・ 質問をしているコメントの内、回答のないコメントは「孤立質問」に分類する．
- ・ Wikipedia やレビューサイトから得られる情報で構成されるコメントは「一般の Web 上の情報」に分類する．
- ・ テーマに関係のないコメントは「テーマに無関係な情報」に分類する．
- ・ テーマに対してかなり重要なコメントは「深層情報」に分類する．

4.1 3 種類の基準情報各々の実験の結果

本実験では「映画」に関するスレッドを解析する為、Wikipedia の記事、レビューサイトの記事、Yahoo!の検索結果上位 50 件のサマ리를基準情報とした．そこでこれら 3 種類の Web サイトそれぞれを基準情報とした際の、システムのコメント分類に対する再現率と適合率を調べた．結果をそれぞれ表に示す．(表 2、表 3、表 4、参照)

結果から、例えば「20 世紀少年」では Yahoo!検索の結果を基準情報とした時、再現率・適合率は最も良い値となるが、「アイ・アム・レジェンド」では Wikipedia を基準情報とした時

の結果が再現率・適合率ともに最も良い値となる等、最適な基準情報はテーマによってバラつきがある事が分かる．また、Wikipedia のみを基準情報とした時のシステムのコメント分類結果は、ほかの 2 つと比べて全体的に良い結果が得られた．

4.2 提案手法を用いた実験の結果

本手法の有用性を測る為に、システムのコメント分類に対して再現率と適合率を算出した．また 3.3 節で述べた重要度計算を行う際の、客観性を判断する為に用いる形容詞の出現頻度に対する閾値 α と「一般の Web 上の情報」と「深層情報」を分ける為の重要度の閾値 β は、経験的に $\alpha = 2$ 、 $\beta = 4.5$ と設定した．算出した再現率と適合率を表 5 に示す．

4.3 考 察

Wikipedia の記事、レビューサイトの記事、Yahoo!の検索結果上位 50 件のサマリのそれぞれを基準情報としてコメント分類を行った事前実験の結果と、それらすべてを基準情報としてコメント分類を行った評価実験の結果を比べてみると、適合率・再現率ともに評価実験の方が安定した結果が得られている．これは事前実験において全体的に良い結果が得られた Wikipedia の記事が、多くの場合、レビューサイトの記事や Yahoo!の検索結果と比べて文章量が多く、その結果、それらすべてを基準情報としても、大半が Wikipedia の情報となったのが、原因であると推測される．

また表記揺れの問題があり、再現率や適合率の低下を招く原因となっている部分が見られた．

5. まとめと今後の課題

本論文では「映画」をテーマとする掲示板から「一般の Web 上にほとんど記載されていない、テーマに対して重要な情報」すなわち深層情報を取得するという目的で、差分抽出手法と重要度計算手法の二つの手法を提案し、実験を行うことで提案手法の有用性を測った．

まず、差分抽出手法により、スレッド内におけるスレッド特有の情報を取得し、次に 3 つの尺度「テーマに対する熟知度、議論に対する貢献度、客観度」を提案し、重要度計算手法によるコメントの重要度計算を行った．これらを用いてコメントを「ゴミ情報、孤立質問、一般の Web 上の情報、テーマに無関係な情報、深層情報」の 5 つの種類に分類し、深層情報の抽出を図った．また、実験においては基準情報の妥当性を調べるために事前実験と評価実験との比較を行い、その結果、評価実験の結果の方が事前実験の結果と比べて安定した結果が得られた．しかし、事前実験の結果をすべてにおいて上回る結果とはならず、今後さらに基準情報を選定し、また重要度計算手法に関しても再考が必要であると考えられる．

深層情報の自動的な取得が可能となることで、多くの時間を費やしスレッド内の全てのコメントを読むことなく、重要でかつ一般の Web 上にあまり記載されていない、掲示板特有の情報を取得する事が可能となる．また深層情報を提示することで、そのスレッド内で既にどのような重要な議論がなされているのかを知ることが出来、新たにコメントの投稿を行うユーザに対し、既出の情報と同じ内容の投稿を防ぐことが可能になると考

(注3): <http://www.casphy.com/>

表 2 Wikipedia を基準情報とした時のシステムの再現率・適合率

Table 2 Recall Ratio & Precision Ratio Based on Wikipedia

映画タイトル	コメント数	再現率	適合率
「20 世紀少年」	115	64.3%	76.5%
「レッドクリフ」	8	75.0%	75.0%
「アイ・アム・レジェンド」	48	64.6%	77.1%
「ゲド戦記」	89	58.4%	76.4%
「ターミネーター 4」	41	48.8%	63.4%

表 3 レビューサイトを基準情報とした時のシステムの再現率・適合率

Table 3 Recall Ratio & Precision Ratio Based on Review site

映画タイトル	コメント数	再現率	適合率
「20 世紀少年」	115	56.5%	72.2%
「レッドクリフ」	8	75.0%	75.0%
「アイ・アム・レジェンド」	48	60.4%	70.8%
「ゲド戦記」	89	53.9%	74.2%
「ターミネーター 4」	41	39.0%	56.1%

表 4 Yahoo!検索結果を基準情報とした時のシステムの再現率・適合率

Table 4 Recall Ratio & Precision Ratio Based on Yahoo! results

映画タイトル	コメント数	再現率	適合率
「20 世紀少年」	115	67.0%	80.0%
「レッドクリフ」	8	50.0%	50.0%
「アイ・アム・レジェンド」	48	52.1%	70.8%
「ゲド戦記」	89	56.2%	78.7%
「ターミネーター 4」	41	46.3%	61.0%

表 5 システムの再現率・適合率

Table 5 Recall Ratio & Precision Ratio

映画タイトル	コメント数	再現率	適合率
「20 世紀少年」	115	67.8%	77.4%
「レッドクリフ」	8	75.0%	75.0%
「アイ・アム・レジェンド」	48	58.3%	75.0%
「ゲド戦記」	89	55.1%	74.2%
「ターミネーター 4」	41	53.7%	65.9%

えられる。

今後の研究課題として以下のものが考えられる。

- 映画以外の領域における提案手法の有効性確認
- 表記揺れへの対応
- 基準情報の選定
- コメントの内容からの呼応関係の抽出
- コメントの信頼性・信憑性の考慮

謝 辞

本研究の一部は、平成 21 年度科研費特定領域研究「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: 瀬本明代) 及び甲南大学平生太郎基科学研究奨励助成金による。ここに記して謝意を表します。

文 献

- [1] 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
- [2] 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.
- [3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, “テキストマイニングによる評価表現の収集”, 情報処理学会研究報告 NL154-12, pp. 77-84, 2003.03.
- [4] 鍛冶伸裕, 喜連川優, “HTML 文書からの評価表現辞書の自動構築”, 言語処理学会第 13 回年次大会, 420-423, 2007.03
- [5] 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
- [6] 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
- [7] 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.
- [8] 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 ”, インタクション 2005, 情報処理学会主催, 2005