3B3-NFC-01a-3

ユーザの観点に基づく電子掲示板からのコミュニケーション抽出

Communication Extraction from Bulletin Board System Based on User's Point of View.

里中　晴日　　砂山　渡
Haruhi Satonaka　Wataru Sunayama

広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Recently, it is natural that we communicate with bulletin board systems (BBS) on the Internet. Though bulletin board systems are easily used for many purposes such as information collection, opinion exchange and discussion, those tend to include unnecessary comments. In this paper, a framework to identify comments related to the Key Words of a thread (a set of comments in a BBS) as communication is proposed so that users can find threads along with their interests.

1. はじめに

近年，インターネットの普及に伴い「2 ちゃんねる」や「Yahoo!掲示板」のような電子掲示板が一般的なコミュニケーションツールとなった．電子掲示板は匿名性が高く，PC 以外にもスマートフォンをはじめとする小型端末からも読み書き可能となった．また，上記 2 つに代表されるような大型掲示板だけでなく，レンタル式の電子掲示板も存在し，個人で立ち上げているブログやホームページ，または facebook を初めとするソーシャルウェアに組み込まれていることがある．このような背景から，多くの人々が電子掲示板で情報交換や議論などが盛んに行われるようになった．しかし，誰もが利用可能なため電子掲示板の数は膨大となり，1 つのスレッド（コメントとレスを一まとめにくくった時の名称）においても大量のレス（書き込み，返信）が多くなっている．そのため，閲覧しているスレッドにはどのように話が進んでいるかをすぐに把握することは難しく，ユーザの必要な情報を必要に応じて取り出す事が困難となっている．

本研究では，大量に書き込まれたスレッドのレスから瞬時にユーザの求める情報が取り出せるようにレスを抽出システムの構築を目指す．

以下，本論文では 2 章で関連研究として現在行われている電子掲示板に関する研究について述べる．3 章で提案システムの構成を述べる．4 章で本システムの有効性を確認するための評価実験とその実験結果，および考察について述べる．最後に，5 章で結論を述べる．

2. 関連研究

電子掲示板内のあるスレッドに書き込まれたコメント間のつながりを，複数のコメントで共通に使われる単語の情報をもとに，ツリー構造で表示する研究 [1] や，掲示板の要約を行う研究 [2] がある．これらの研究が対象とする掲示板においては，各コメントがどのコメントに向けて書かれたものかが明示されている．本研究においては，必ずしも返信先が明示されていない，時系列順に並べられたコメント集合の要約を目指す．

電子掲示板から評判情報を抽出してまとめる研究 [7] においては，商品やサービスを評価する書き込みを，肯定的なものと否定的なものとに分類して要約することを目指している．本研究におけるコミュニケーション抽出では，要点を抑えた指示的要約を作成するのではなく，一連の書き込みによるコミュニケーションを再現する報知的要約において必要と考えられるコメントを抽出する．

文章に対してその話の流れや単語の流れをもとに報知的要約を試みる研究 [5] はあるが，複数人が雑多な書き込みを行う電子掲示板において，そのコミュニケーションの主な流れを追うことを目的とした研究は見られない．そこで本研究では，電子掲示板の主題に沿ったユーザ間のコミュニケーションの流れの抽出を目指す．

連絡先: 里中晴日, 砂山渡, 広島市立大学大学院情報科学研究科 シス テム工学専攻, 広島市安佐南区大塚東三丁目 4 番 1 号, {haruhi,sunayama}@sys.info.hiroshima-cu.ac.jp

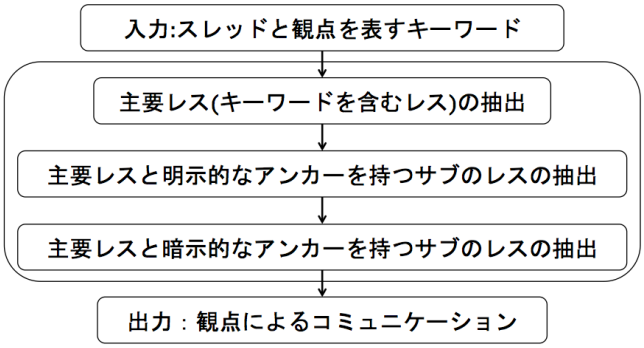

図 1: コミュニケーション抽出の枠組み

3. システム構成

本章では，提案するコミュニケーション抽出の枠組み (図 1) について述べる．

はじめに，電子掲示板のあるスレッドのテキスト，ユーザの観点に基づくキーワードを入力として，そのキーワードを含むレスを抽出する．次に，それらメインレスと関わりを持つレスを，明示的なアンカー情報および，暗示的なアンカー情報をもとに抽出する．その後，抽出されたレスを出力する．以下で，これらの詳細について述べる．

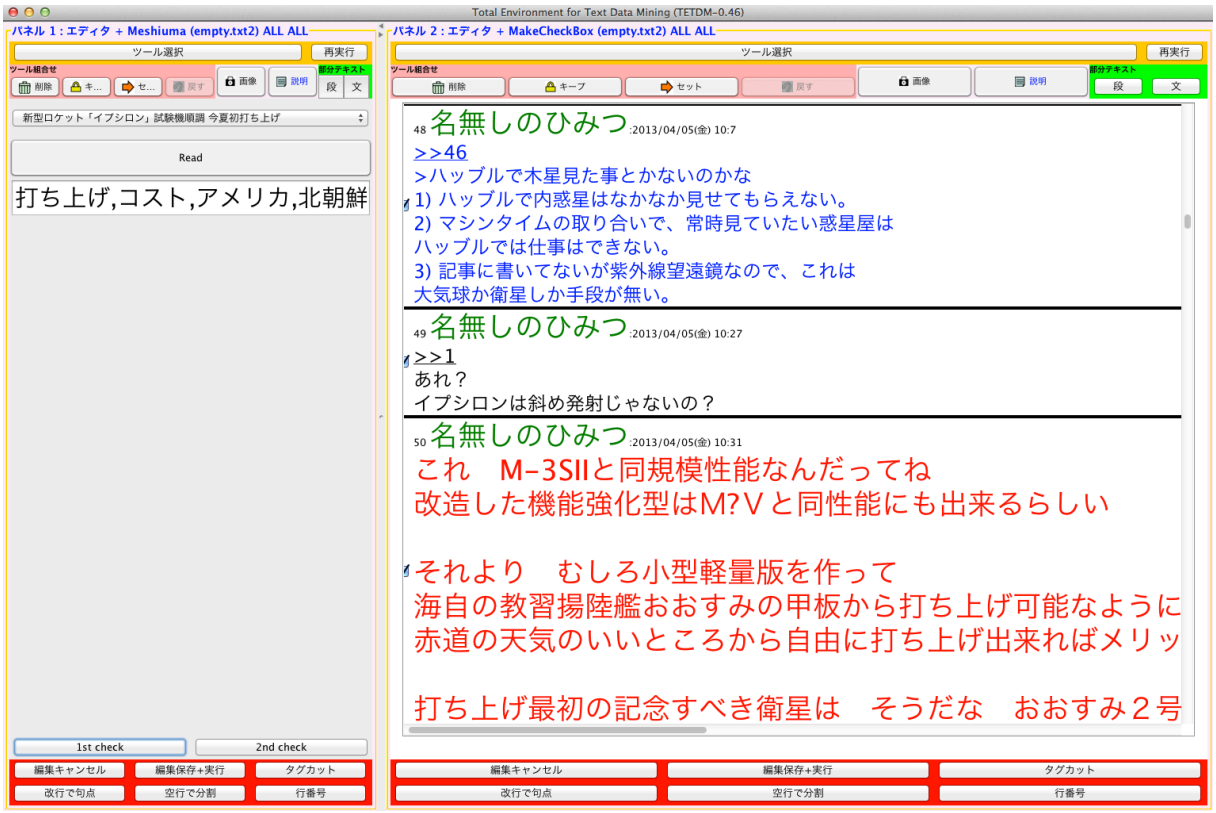

図 2: 提案システムによるスレッドタイトル「新型ロケットイプシロン試験機順調　今夏初打ち上げ」の抽出レス出力例

3.1 入力：スレッドの文章ファイルとキーワード

提案システムの入力はスレッドのテキストファイルとする．さらに，ユーザが求める情報 (ユーザの観点) に沿ったキーワードを入力する．このキーワードの入力は単数，複数どちらでも可能となっている．

3.2 メインレスの抽出

スレッドの主題となるメインレスを抽出する方法について述べる．1 節で入力されたテキストファイルからキーワードの単語を含むレスを，メインレスとして抽出する．

3.3 明示的アンカーによるサブレスの抽出

メインレスと強く関連するレスをサブレスとして抽出する．

関連の有無の判断には，掲示板上で用いられる他のレスへの明示的なアンカー情報を用いる．すなわち明示的なアンカーとは，他のレスの内容に対して返信を行う場合に，そのレスの番号を明示的に示すためのもので，例えば掲示板「2 ちゃんねる」[8] の場合，「>>」というアンカー記号を用いて「>>20」と書き込んだ場合，20 番のレスに対する返信レスであることを明示している．

そこで，この明示的アンカーを用いて，メインレスとのつながりをもつ次の 2 種類のレスをサブレスとして抽出する．

1. メインレスが含む明示的アンカーが指すレス
2. メインレスを指す明示的アンカーを含むレス

1. はメインレスが言及している内容を含み，主題に関するコミュニケーションが行われているレスとして抽出する．2. はスレッドの主題に関するコミュニケーションを意図したレスとして抽出する．

3.4 暗示的アンカーによるサブレスの抽出

メインレスと関連するレスが，必ずしも明示的なアンカーを用いているとは限らない．そこで，メインレスとの暗示的なつながりをもつレスをサブレスとして抽出する．すなわち，メインレス内の単語と同じ単語を用いてレスがなされ，かつそのレスがなされた時間と，つながりが想定されるメインレスが書き込まれた時間との差が一定時間内であれば，単語による暗示的なアンカーによるつながりを認定し，サブレスとして抽出する．

ただし，アンカーとして用いる単語は，一般的な単語では意味がないと考えられるため，日本語の場合，形態素解析器「茶筌」 [3] による名詞，または 3 文字以上のカタカナによる未知語であるものを用いる．

また時間差の条件として，前に書き込まれたレスを参照して，そのレスに対する返信であることが，明示的アンカーを用いなくても参照先が理解できる時間を設定する必要がある．この時間条件の設定の基準の一つとしては，通常のブラウザで一画面に収まると考えられる 10 レス程度の範囲であれば，参照先を理解しやすいと想定して，平均的に 10 レス程度が書き込まれる時間に設定する事が考えられる．またレス数ではなく時間を，暗示的アンカーの範囲条件に用いる理由は，書き込み頻度が多い掲示板では，レスの範囲が広くてもその内容を覚えている可能性が高く，より広い範囲で暗示的アンカーが発生する可能性が高いと考えたことと，逆に書き込み頻度が低い掲示板では，表示の上で近くのレスであっても，すでにその話題が終了している可能性が高いと考えたことによる．

表 1: 使用したスレッドのタイトル，レス数

No.	タイトル	レス数
1	古代ローマについて	889
2	メイドロボの作り方	558
3	豚汁のおいしい作り方	240
4	見ないと人生損するアニメ	204
5	新型ロケット「イプシロン」について	198
6	糖質オフダイエットについて	317

表 2: 使用したキーワードと抽出されたレス数の全レス数に対する割合

No.	キーワード	レスの割合
1	ローマ, キリスト教, ギリシャ, ヨーロッパ	29%
2	ところ, データ, みたい, メイド	32%
3	ジャガイモ, 作り方, カツオ, たくさん	8%
4	メディア, アニメ, コード, コイル	22%
5	打ち上げ, コスト, アメリカ, 北朝鮮	32%
6	カロリ, エネルギー, 炭水化物, アメリカ	48%

3.5 出力：抽出したレス集合を出力

抽出されたメインレスとサブレスの集合を，もとの掲示板の形式に基づいて出力する．この際，下品，侮辱的な言葉をまとめた NG ワードリストをあらかじめ作成しておき，その中の単語が 1 つでも含まれているレスを除去する．また出力を見やすくするために，メインレスのフォントサイズを大きくしたり，メインレスを赤色，明示アンカーを含む明示的サブレスを青色，暗示的アンカーを含む暗示的サブレスを黒とし，文字の色を変えることで視認性を高めた．図 2 にシステムの出力例を示す．使用したのは評価実験で使用したスレッドの No.5 であり，左のテキストスペースに高頻出の単語が自動で出力される．右に実際に抽出されたレスが表示される．

4. コミュニケーション抽出の評価実験

本章では，前章で述べたコミュニケーション抽出の枠組みに基づいて構築したシステムが，コミュニケーションの抽出に役立てられるかを確認した実験について述べる．

4.1 実験内容

電子掲示板「2 ちゃんねる」[8] から 6 つのスレッドを用意し，高頻出のキーワードを用いて提案システムにより抽出されるレス集合が，スレッド全体の要約となっているかを比較した．表 1 に使用したスレッドのタイトルとレス数を示し，表 2 に各スレッドで使用したキーワードと提案システムによって抽出されたレスの割合を示し，表 3 では各スレッドで抽出された各レスの割合を示す．

比較システムには，テキストの報知的要約を生成する展望台システム [6] を用いた．比較システムは，テキストの主題および副題となるキーワードをそれぞれ抽出した上で，テキスト中の各文に，テキスト全体における評価値と，文を含む段落内での評価値を与え，テキストのストーリーを作る重要文を抽出する．提案システムによるコミュニケーション抽出は，電子掲示板のスレッド内の話を理解するための要約生成とも考えられるため，このシステムを比較に用いた．なお比較システムは，文の区切りと段落の区切りを必要とするため，各レスの最後に句点「.」を挿入し，1 レスごとに段落の区切りを挿入した．

表 3: 提案システムが抽出したレスの種類ごとの割合

No.	メイン	明示的サブ	暗示的サブ
1	91%	8%	1%
2	85%	12%	3%
3	72%	11%	17%
4	64%	17%	19%
5	62%	10%	29%
6	64%	17%	19%

提案，比較システムでの実験に際しては，各システム情報科学を専攻する大学生・大学院生の男女 4 名 (計 8 名) を被験者とした．各被験者には，提案システムの出力と比較システムの出力を，それぞれ 3 つずつ合計 6 スレッドについて，以下の手順で評価を行ってもらった．本実験では，被験者の手間を抑え，集中力をもって実験に望んでもらうために，各システムが出力したレスにチェックをつけて提示を行い，被験者は，追加で必要なレスにチェックをつけ，不要なレスのチェックを外す作業を行った．

表 2 より，実験に用いた 6 つのスレッドのシステムにより抽出されたレス数の割合は，8%から 48%（平均 28.5%）となった．また，比較システムによる抽出レス数が全レス数の 30%程度となるように設定した．表 3 に，スレッドから抽出されたレスに対する各レスの割合を示した．

4.2 提案システムとメインレスの考察

被験者が必要だとチェックをしたレスを正解とし，各スレッドの提案システムと比較システムによって抽出されたレスの適合率と再現率を表 4 に，表 5 にメインレスの適合率と再現率を示す．表 4，表 5 のスレッド No.4 で適合率，再現率共に比較より高くなっている．このスレッドでは，他のスレッドに対して入力したキーワードに対して意見のやり取りが全体を通してコミュニケーションが行われている傾向があった．そのため，入力したキーワードはスレッド全体を表したものであり，比較システムに比べ適合率，再現率が高くなったと考えられる．そのため，他のスレッドで再現率や適合率で 7 割を下回る事があった理由は，キーワードがスレッドで行われたコミュニケーションがあまり行われておらず，入力したキーワードががスレッド全体を表しておらず、全体の要約には至らなかったと考えられる．また No.3 での，再現率が低くなったのは，スレッドのテーマ「豚汁のおいしい作り方」に対して、高頻度語のキーワード「ジャガイモ、作り方、カツオ、たくさん」が出現するレスが少なかったことが原因と考えられる．そのため，今後メインレスの抽出に用いるキーワード数を 4 つと固定するのではなく，抽出されるレス数に応じて設定することが考えられる． また，表 5 のメインレスの適合率，再現率が下がった理由も同様であると考えられる．

4.3 明示的なサブのレスの抽出の考察

明示的なサブのレスの適合率，再現率を表 6 に示す．この表より，No.1,3,4,5 の再現率が 5 割を下回っている．これは返信先を明示していても，ユーザに取ってはキーワードと関連が無く必要の無いレスだと判断されてしまったためと考えれられ

表 4: システムの抽出レスの適合率と再現率

	提案システム		比較システム	
No.	適合率	再現率	適合率	再現率
1	0.90	0.63	0.97	0.87
2	0.85	0.93	0.95	0.88
3	0.92	0.36	0.69	0.71
4	0.88	0.83	0.72	0.68
5	0.67	0.83	0.83	0.84
6	0.77	0.93	0.79	0.89

表 5: 提案システムのメインレスの再現率と適合率

No.	適合率	再現率
1	0.88	0.55
2	0.87	0.80
3	0.77	0.25
4	0.93	0.72
5	0.79	0.60
6	0.93	0.73

る．そのため，明示的サブレスの抽出の条件に，キーワードとの関連性の高い単語（メインレスの中で高い頻度で使用される単語）がある場合にサブのレスとする条件の追加が考えられる．

4.4 暗示的なサブのレスの抽出の考察

表 7 に抽出された暗示的なサブのレスの適合率, 再現率を示す．No.1〜No.3 での再現率が大きく下がった理由として，抽出されたレスの総数に対して暗示的なサブのレスの割合が少ないことが考えられる．それに対して No.4〜No.6 の再現率が上がった理由として，特定のキーワードに対するコミュニケーションが，提案システムで設定した 10 レスの間で行われてたことが考えられる．このことから、スレッドのレスの進行速度に応じて暗示的なサブのレスを抽出する時間の条件を変更する必要があると考えられる．

5. 結論

本論文では電子掲示板から，ユーザの観点に沿ったコミュニケーションを抽出するシステムを提案した．本システムは，入力されたスレッドからメインの特定とそのレスの抽出，メインのレスに関する明示的なサブのレスの抽出，暗示的なサブのレスを抽出し，抽出されたレスをまとめることによりコミュニケーションの抽出を行う評価実験を行い，抽出の際に必要だと思われる条件を確認した．

今後の課題として，レスの抽出率が低くならないうように抽出したい割合に応じてキーワードを設定することが考えられる．さらに，キーワードとの関連性の高いレスを抽出するために，キーワードと関連の高い単語を含む事を条件に追加する事が考えられる．最後に，スレッドのレスの進行速度に応じた抽出するレスの対象の時間を設定することが考えられる．

表 6: 提案システムの明示的サブレスの再現率と適合率

No.	適合率	再現率
1	0.77	0.36
2	0.77	0.71
3	1.00	0.06
4	0.75	0.24
5	0.67	0.31
6	0.58	0.76

表 7: 提案システムの暗示的サブのレスの再現率と適合率

No.	適合率	再現率
1	1.00	0.06
2	1.00	0.19
3	1.00	0.09
4	1.00	0.63
5	1.00	0.86
6	1.00	0.85

参考文献

[1] 松村真宏，大澤幸生，石塚満：テキストによるコミュニケーションにおける影響の普及モデル，人工知能学会論文誌，Vol.17, No.3, pp.259 – 267, (2002).

[2] 松尾豊，大澤幸生，石塚満：電子掲示板における会話からのトピック発見と要約，第 16 回人工知能学会全国大会, 3D1-07, (2002).

[3] 松本裕治，山下達雄，平野義隆，松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』，Ver.2.4.0，使用説明書，(2007).

[4] 岡村剛，角康之，西田豊明：電子掲示板からの興味のある会話の抽出支援，情報処理学会インタラクション 2005, A-109, (2005).

[5] 市丸夏樹，日高達：要約文の話題の流れの最大化による自動要約，自然言語処理, Vol.12, No.6, pp.45 – 61, (2005).

[6] 相良直樹，砂山渡，谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成，電子情報通信学会論文誌, Vol.J90-D, No.2, pp.427 – 440, (2007).

[7] 竹内啓祐，浦島智，畑田稔，安宅彰隆：電子掲示板からの評判情報抽出における P/N 判断，電子情報通信学会技術研究報告, KBSE, 知能ソフトウェア工学, Vol.106, No.473, pp.55 – 60. (2007).

[8] 2 ちゃんねる掲示板 (URL)http://www.2ch.net/