ディスカッションマイニング：構造化されたコミュニケーションによるトピックの検索と視覚化

村上 明子　　長尾 確

日本アイ・ビー・エム（株） 東京基礎研究所
{murakami,nagao}@trl.ibm.co.jp

1 はじめに

現在大量の文章を分析して知識を発見するテキストマイニングが注目を集めている。また、最近ではインターネットの普及に伴い、対話に近い形で行われているメーリングリストや掲示板などネット上のディスカッション履歴をとることが可能になり、テキストマイニングのなどの技術を用いて知識として再利用することが考えられている。しかし、非同期で行われる上記のようなディスカッションの履歴は、一つのドキュメントの中に話題が多数存在したり、一つのトピックが複数のドキュメントにわたったりしており、そのままではテキストマイニングの対象としては扱えない。会議などの現実世界でのディスカッションは報告書などの形で保存され、マイニングの対象となっているが、ディスカッション履歴から人手によってトピックごとに文書をまとめるにはコストがかかりすぎてしまう。

我々は今回、このような非同期ディスカッションの履歴から、引用情報を用いて履歴文書の構造化・視覚化の例を示し、新たなテキストマイニングの対象となる文書の自動生成を行い、それを対象とするマイニングについて検討する。

2 文書における知識管理

大規模な記録媒体の登場で情報を大量に保存しておくことが可能となった。そのような大量の情報から必要なものを知識として管理し、検索する技術が要求されている。それらの技術を総称して知識管理（ナレッジマネジメント）と呼んでいる。

知識管理という点から見ると、再利用される文書ほど価値が高いと考えられる。そこで、文書を知識の再利用性という観点で見る。

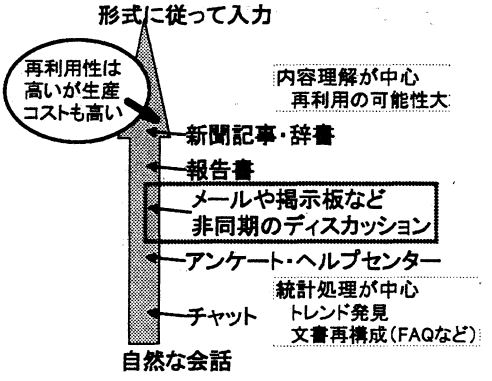

図 1: 知識管理から見たテキスト

図1の下に位置する会話に近い形のもので、コールセンターやアンケートの文書がある。この分野の文書に対して、単語レベルに加えて係り受けの情報やモーダルなどの情報を用いて時間・場所依存性を示し，問題発見や全体のトレンドを見るといった知識発見型の研究が行われている [1]。これはデータマイニングの手法を用いて、全体的な傾向を統計処理して求め知識として得ているものであり、文書そのものを再利用しているわけではない。また、言語処理から得られた単語間の係り受け関係から文書の中の頻出トピックを求め、得られた知識を FAQ(Frequently Asked Question) などを生成するのに役立てるといた研究も行われている [2]。

図1で上に位置する新聞や辞書などはそのままの形で知識として利用できるものであり、検索や要約などによって再利用される。これまでに新聞記事を同じ話題の時間依存性・場所依存性を考慮して視覚化する研

究 [3] や、共起する単語により相互関連性をもとめて、スレッドにするという研究が行われた [4]。しかし、このような文書は再利用性が高いが、生成するためにはコストが高いといった問題がある。

今回、新しい対象として上の二つの中間に位置すると思われる非同期ディスカッションの履歴文書を取り上げる。ディスカッション中の一文書は話題の一部であったり、複数の話題を含んでいたりするものであり、そのままでは再利用困難である。したがって、これらの文書を再利用が容易な形に変え、知識を得ることを目標とする。

3 議論の構造化の重要性

人のコミュニケーションの結果として得られるディスカッションの履歴は、知識・トレンド発見型のテキストマイニングを用いては有効な結果は得ることができない [5]。また、知識獲得という意味でテキストマイニングを考えていく場合、内容そのものを今までに行われたディスカッションの当事者はもちろん、第三者が見て知識として再利用できる形に再編成する必要がある。

3.1 ネット上でのコミュニケーション

現在インターネット、イントラネットを用いるディスカッションとしては、

- チャット
- ネットニュース
- 電子メールを用いたメーリングリスト
- Web 掲示板

などがあげられる。これらは作成者や時間といったごく限られた情報でのみ識別され、内容まで言及されていないことが多い。そこで、グループウエアや一部の掲示板などにおいては発話者が発言する内容に基づいて構造化し表示する方法が取られている。ネットニュースでは、これまでにダイジェストやFAQの自動生成をする研究 [6][7] が行われている。また、ネットニュースをキーワード情報に基づいてクラスタリング・構造化する研究も行われている [8]。

一方で電子メールを用いたディスカッションの形態であるメーリングリストはメール間に依存関係があるにもかかわらず、それが明示化されていないことが多い。そのため、話題間の関係に基づいた履歴の表示が提案されてきた [9][5]。しかし、構造化を行っただけでは、発話（メール）一つが文書の一単位であることには変わりなく、一つの文書が複数の話題を含んでいたり複数の文書に一つの話題が広がっていたりする場合に文書の単位として適当でないといえる。

3.2 ディスカッションの構造化

そこで、我々はメールに対して内容によって構造化を行った上、それを新たな知識として蓄積することを提案する。

3.2.1 引用によるメール相関関係の抽出

以前、メールの構造化において我々は引用部分に着目した [5]。メール中において他のメールを引用している部分を引用部とする。メール中で引用部に続く文章は内容的に連続している。この引用部に対して内容が連続している部分をコメント部とする。この引用部とコメント部の情報を利用することにより、複数の話題に対して引用・コメントしているメールなど、1つのメールから複数の議論に派生するものに対応できると考える。

3.2.2 XML によるメールの構造化

メール内容を統一したフォーマットで処理するため、XML を用いる。メールに付随している情報としてはタイトル、筆者、日付、などが考えられる。その他、引用とその引用に対するコメントについてもタグで明示化する。メールでは慣用的に他のメールを引用するときには「>」などの引用符をつける。そこで、引用符を用いてメールの表記のみによって引用部を判断し、引用を示すタグ <QUOTE> で囲む。そこにはリンク情報としてメールのID ではなく、メールのどの部分を引用しているかの情報を文単位あるいは単語単位で付加しておく。また、そのメールに対するコメントの部分についても、同じようにタグ <COMMENT> で囲むようにする。引用の直前にコメントがくることも考えられるが、ほとんどの場合引用部の後に表れ、ひとつの段落で形成されることが多い。まず引用部直前の段落のなかに「次の」「下の」といった引用部をさす指

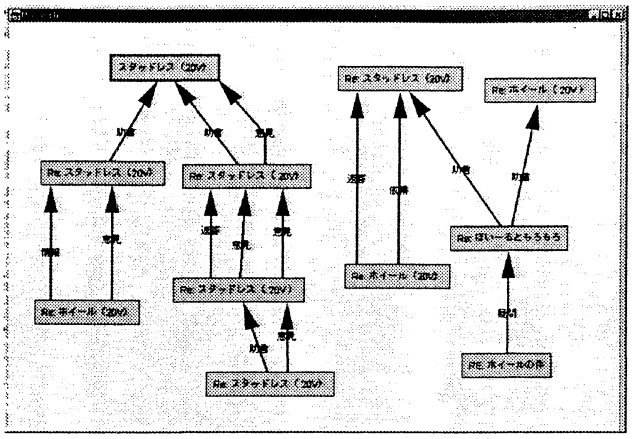

図 2: 視覚化されたメールのグラフ構造

示語があるかをみて、あれば直前の段落を、ない場合は引用部の直後の段落をコメント部と判断する。

3.2.3 コメント部の意味抽出

引用に対するコメントの部分から、メールとメールの相関関係を取り出す。今回のデータに対しては「助言」「情報」「依頼」「疑問」「返答」「意見」という意図に分けた。これについては今回疑問などの基本的なものについてのみ行いあとは人手で行った。自動化するにはコメント部について形態素解析したのちに、それぞれの意図に対して動詞のパターンを記憶することによって可能であると考えられる。

3.2.4 スレッドサマリーの生成

議論の関連性の判断は、コメント部を引用しているかどうかの引用情報を元にして考えた。話の関連性を複数のメールにわたって見たとき、その流れは一つのトピックについての文書として扱えるであろう。これをスレッドサマリーと呼ぶ。

この生成方法は以下のとおりである。1．引用→コメントの組み合わせを見つける2．引用がほかの引用のコメントである場合、その組み合わせを前に追加する3．コメントをさらに引用しているメールがある場合、その組み合わせを後ろに追加す4．以上を繰り返し、それらの組み合わせの集合を引用されている順に並べて一つの文書を生成する　このスレッドサマリーをひとつの文書として扱い、要約をして表示することを考えている。メーリングリストのデータのようなディスカッションの履歴は、グラフ構造している。これらを要約しようとしても、グラフ構造になることは避けられない。したがって、スレッドサマリーを生成することによって、話の流れをひとつの文書としてまとめ、要約が可能になる。

3.2.5 視覚化

メールの表示では引用された部分とそれに対するコメントが色分けされて表示される（引用中に含まれるほかのメールの引用についても区別される）。　XMLで書かれたメールのデータから、引用、コメント、リンク情報、意図情報を取り出して新たなXMLファイルを作る。メールの相互関係情報を書いたXMLファイルをXMaiLと呼ぶ。このXMaiLファイルを元にして、メールのグラフ構造を視覚化する。図2にグラフ化されたメール構造を示す。

ノード（節）のひとつである引用先のメールから引用元のメールにアーク（弧）が引かれている。このアークはひとつの引用について一つ作られるので、2つのノード間に対して複数本生成されることがある。このアークに引用に対するコメントの内容である「意見」「疑問」といった意図が示されている。　ここで、一つのメールは一つのノードとして扱われるが、アークの端点がメール全体ではないことに注意したい。一つのアークはあくまでもメールの一部分である「引用」と「コメ

ント」のペアを表している。スレッドサマリーはグラフ上で複数のノードをアークで結んだ集合として考えられる。引用・コメントのペアをつないだスレッドサマリーはグラフ表現上で図３のように表される。

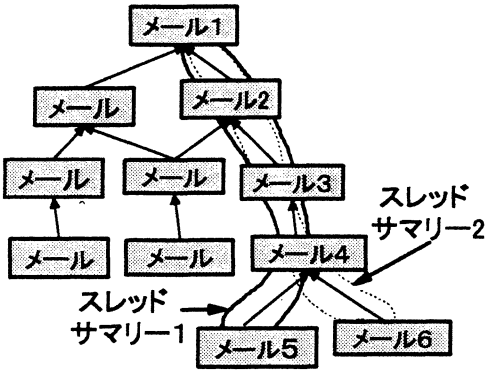

図 3: スレッドサマリー

したがって、一つ一つのスレッドサマリーを独立した文書として扱うといっても、一部内容の重なりが出来てしまう。たとえば上の例を見る場合、スレッドサマリー１とスレッドサマリー２はメール４の同じところを引用して意見をそれぞれメール５，６の中で述べており、スレッドサマリーはメール４（の一部分）まではまったく同じ文書である。

4 関連研究

メールにおける引用構造が重要なことは上で十分に述べた。そのために、正しい引用構造を生成する（そしてリアルタイムで文にタグをつける）システムを提案する。こうすることにより、メールを書くこと＝再利用可能な情報を生成することにつながり、一つの知識としてメールを捉えることが出来るようになる。これからはメールをクライアントで持つ時代は終わり、サーバー側で管理するようになるだろう。以上のシステムをJAVA アプリケーションのようなクライアントを選ばないもので構築することにより、クライアントはそれを表示するブラウザの役割を果たすことになる。知識管理においては既存の情報を再利用するだけではなく最初から情報を再利用する形で保存することも考えなくてはいけないだろう。

5 まとめ

今回はメーリングリストに代表されるメールの引用構造を用いたディスカッションの履歴を構造化し、表示する方法を提案した。これは今まで蓄積した情報であるメールの情報を再利用しやすい形に直したものである。また、議論の関連性によって分割されたスレッドサマリーをひとつの文書として扱うことにより、新たな知識発見の対象としてディスカッションの履歴を再利用することが可能になる。今後は、こうして作られたスレッドサマリーの文書を検索・視覚化するだけではなく、複数のメーリングリストや他の文書と組み合わせて考え、全体のトレンドなどを見ることができるようになるのではないかと考えている。

参考文献

[1] 那須川哲哉, 諸橋正幸, 長野徹. テキストマイニング ー膨大な文書データの自動分析による知識発見ー. 情報処理, Vol.40 No.4, pp.358-364 (1999)

[2] 松澤裕史、福田剛志　大規模データベースからの構造化相関パターンの抽出、アドバンスト・データベース・シンポジウム'99, p151-160

[3] Masayuki Morohashi, Koichi Takeda, Hiroshi Nomiyama, Hiroshi Maruyama: Information Outlining -Filling the Gap between Visualization and Navigation in Digital Libraries, Workshop on Digital Library, [1995]

[4] N. Uramoto and K. Takeda: A Method for Relating Multiple Newspaper Articles by Using Graphs, and Its Application to Webcasting, COLING-ACL'98,[1998].

[5] 村上明子、長尾確：引用に基づくオンラインディスカッションの構造化、自然言語処理シンポジウム1999、http://www.pluto.ai.kyutech.ac.jp/plt/inui-lab/pub/NLP_Sympo99/murakami/murakami.html

[6] 佐藤円、佐藤理史、篠田陽一：電子ニュースのダイジェスト自動生成、情報処理学会誌、Vol.36,No.10, pp.2371-2379

[7] 佐藤円、佐藤理史：ネットニュース記事群の自動パッケージ化、情報処理学会誌、Vol.38,No.6, pp.1225-1234

[8] 内元清貴、小作浩美、井佐原均：キーワードによるネットワークニュース記事群の構造化、言語処理学会第４回年次大会発表論文集、p544-547

[9] 松浦文崇、高田眞吾、中小路久美代：計算機上におけるコミュニケーションの履歴表示に関する研究、情報処理学会研究報告、99-GW-31-8,P43-48