電子メールの傾向分析への知識獲得手法の適用

上田宏高[†] 柳沢豊[††]
塚本昌彦[†] 西尾章治郎[†]

本稿では，大量に蓄積された電子メールアーカイブから，知識獲得手法を用いることによって，有用な知識を得ることを目的とした電子メールからの知識獲得（KDM: Knowledge Discovery in electronic Mail）の手法について論じる．KDMでは，特にメーリングリストのアーカイブに注目し，各メールのヘッダ情報や，本文中に出現する単語頻度情報を抽出する．得られた情報を関係データベースに格納し，知識獲得アルゴリズムを用いて相関ルールを導出する．その際，メール中に含まれる単語の重要度をtf*idf法により推定し，導出ルール数を抑えることができる．さらに本稿では，実際にKDMをあるメーリングリストのアーカイブに適用し，KDMによる知識の導出例を示す．KDMにより，そのメーリングリストにおける全体的な傾向や，話題となった事柄，あるいは，メーリングリストのメンバに関する知識の手がかりを得る一助となる．また，精度が向上すれば，得られた相関ルールを用いることで，メールに対する反応を推し量ったり，検索語が明示的に現れないメールを検索することが可能になると考えている．

Applying Knowledge Discovery Techniques to Trend Analysis of Electronic Mail

Hirotaka Ueda,[†] Yutaka Yanagisawa,[††] Masahiko Tsukamoto[†] and Shojiro Nishio[†]

In this paper, we present the Knowledge Discovery from electronic Mail (KDM) system, a technique for discovering knowledge from large electronic mail archives. In KDM, the system extracts several kinds of data from mail archives, including the header information of each mail and the frequency of words in a mail. Afterwards, the data obtained is stored in a relational database. Furthermore, the association rules are extracted from the database by using the Apriori algorithm. Here, we can restrict the number of rules by estimating the importance of each word by using the tf*idf method. In this paper, we show some examples of rules derived after applying KDM to a mailing list archive. In general, applying KDM to mailing list archives allow us to gather clues of knowledge about each member of the mailing list, including their general behavior and topics in conversation. Finally, we believe that it will be possible to predict the response for any incoming mail and to search for mail with ambiguously specified keywords by increasing precision.

1. はじめに

近年の情報化により電子メールの利用が急速に進んでいる．電子メールの普及率は1997年の9月末には6.6%だったのが2000年の3月末には16.3%にまで増加しており，その増加傾向は今後も持続していくことは間違いない[21]．電子メールの利用は当初，ビジネスなどフォーマルなものが主であったが，電子メールの普及にともない，その利用形態も多様化している．今後，個人が行うコミュニケーションにおける電子メールの占める割合はますます増大していくものと見られる．

電子メールの量が増大するにともない，その処理が大きな負担となっており，膨大なメールから有用なメールだけを選択しようとする情報フィルタリングの研究がさかんに行われている[15]．一方，データベースに蓄積された膨大な生データから，価値ある情報を発掘することを目的とした，データベースからの知識獲得（KDD: Knowledge Discovery in Databases）に関する研究や，関連するシステムの研究がさかんに

† 大阪大学大学院工学研究科情報システム工学専攻
Department of Information Systems Engineering, Graduate School of Engineering, Osaka University
†† NTTコミュニケーション科学基礎研究所社会情報研究部
Social Communication Laboratory, Communication Science Laboratories, NTT

なっている[3),10),12),20)].

本研究の目的はKDDの戦略を電子メールアーカイブに適用して，その中から有用な情報を抽出することにある．本稿ではその際に必要になる，電子メールの特徴抽出およびその表現手法，知識獲得手法について論じる．筆者らはこの一連の処理を電子メールからの知識獲得（KDM: Knowledge Discovery in electronic Mail）と呼んでいる．KDMではメーリングリストのアーカイブなどから，各メールのヘッダ情報や，本文や引用文に含まれる単語頻度情報を抽出する．さらに得られたデータを関係データベースに格納し，Aprioriアルゴリズム[1)]を用いて相関ルールを導出する．その際，各単語の重要度を推定し，その結果を知識獲得プロセスに反映させることで有用なルールを効率良く導出することが可能である．以下，2章ではKDMの概要を述べ，3章ではKDMのアルゴリズムについて論じる．4章では実際にメーリングリストアーカイブにKDMを適用させた結果およびその評価について考察する．5章で関連研究との比較を行い，6章で検索へのKDMの応用およびKDMの今後の展望について述べる．最後に7章で本稿のまとめを行う．

2. KDMの概要

日々流通する電子メールは膨大な量にのぼる．広告など不特定多数に送信されるものを除けば，次のようなものがある．

- 個人間で送信されるメール．
- メーリングリストを通してグループ内で送信されるメール．
- 公的窓口や企業のユーザサポートなどへ不特定多数から送信されるメール．

これらのメールは，こまめに整理しないとどんどん蓄積されていく一方である．その整理を支援するために新着メールの自動分類や自動応答などの研究がなされ，メールを溜めない努力がなされてきた．一方，溜められたメールは，検索という能動的なアクションを行わない限り，利用することができないため，有効に利用されていないのが現状である．だが，蓄積されたメールには重大な情報が眠っている可能性がある．たとえば，次のような情報が考えられる．

- 個人間で送信されるメールからは，コミュニケーションの対象が比較的限られるため，各個人の癖や性格などが分かるかもしれない．ある人が機嫌が悪いときにある特定の単語を利用する可能性が高いということが分かれば，その単語が出てくるメールを受け取ったときに，なるべく相手を刺激しないような応答をすることができる．
- メーリングリストのメールからは，そのメーリングリストでどのような事柄が話題になったのか，メーリングリストの参加者にはどのような人がいるのかといったことを知ることができるだろう．これはそのメーリングリストに参加するかどうかの判断の助けになる．また，あるトピックスに対するメーリングリストメンバの認識を知ることができれば，メールを出す前にそのメールに対する反応を推測することができる．
- 公的窓口などに不特定多数から送られるメールは往々にして膨大な数にのぼり処理しきれないことがある．しかしそれらのメールからは，そうしたメールを送ってくる人々に共通する傾向や，どのような点に不満を持っているのかという点についての手がかりを得ることができるだろう．企業などの顧客サービス係には多くのメールが送信されるが，それらの膨大なメールの中からユーザのニーズを見いだすことができる．

蓄積された膨大な生データから，価値ある情報を獲得することを目的としてKDDに関する研究が数多く行われているが，KDDの戦略をメールアーカイブに適用することにより，前述のようなメールにひそむ情報の抽出が期待できる．すなわちKDMは，メールから抽出した特徴を関係データベースに格納し，知識獲得アルゴリズムを用いて，上にあげたような有用な知識を発掘しようとする試みである．

3. KDMアルゴリズムの概要

KDDを電子メールアーカイブに適用するためにはメールの持つ特徴を抽出しデータベースに格納するという作業が必要になる．その後，知識を獲得することができるようになるが，どのような知識が獲得できるかは，この特徴抽出の精度に影響される部分が大きい．本章ではまず電子メールの特徴抽出手順について述べ，次に知識獲得手順について述べる．以下ではある1つのメーリングリストのアーカイブがデータベースに収納されており，そのメーリングリスト内で成立する知識を導出することを例として取り上げながら，話を進める．

3.1 電子メールの特徴抽出

電子メールのヘッダにはその電子メールの送信者，サブジェクト，日付などの情報が付加されており，一般的なメーラなどのメール仕分け機能はこれらの情報を基にメールの仕分けなどを行っている．ヘッダ情報は非常に有益な情報であるが，メールから何らかの有

To: space@mailinglist.co.jp
From: かに <kani@rokojudo.kcn.ne.jp>
Date: Sun, 24 May 1998 19:12:30 +0900
Subject: [space-ML 766] Re: ガチャピン宇宙へ

かに@KCNです.

Subject: "[space-ML 765] ガチャピン宇宙へ"
satoh@hoge.nasda.go.jp Taro Satoh wrote:
佐藤> 佐藤@NASDAです。
佐藤>
佐藤> http://www.kp-777.co.jp/NEW/event/cosmo.htm を見てください。
佐藤> なんとあのガチャピンがロシアの宇宙船で宇宙に行くのです。

某庁では，「宇宙ハンドブックの日本人宇宙飛行士にガチャピンを載せなくてはならないのか」と大騒ぎをしているようです．
仮に（というか間違いなく）ロシア人宇宙飛行士が扮装するだけだとしても，掲載してほしいですね．なにせ，世界のヒーロー初の快挙なのですから．

****
* かに(Yutaka Yanagisawa) *
* Kani Community Network *

図 1 電子メールの例
Fig. 1 An example of e-mail.

メール ID: 766
日付： 1998/05/24 19:12
送信者： kani@rokojudo.kcn.ne.jp
サブジェクト：ガチャピン宇宙へ
単語： KCN, かに, ガチャピン, ハンドブック, ヒーロー, ロシア, 宇宙, 宇宙船, 宇宙飛行士, 快挙, 掲載, 佐藤, 世界, 大騒ぎ, 庁, 日本人, 扮装, 某, http://www.kp-777.co.jp/NEW/event/cosmo.htm

図 2 抽出情報例
Fig. 2 An example of information extraction.

用な知識を得るためにはこれだけでは不十分であり，本文から特徴を抽出することが必要になる．ただしメール本文は非常に自由度が高い自然言語で構成されており，その内容は実に多様性に富む．その意味解析はきわめて困難であり，現状では実用的ではないうえに，計算量が大きく本研究で扱うような大きなサイズのデータに適用しにくい．そこで本研究ではメールの特徴として，形態素解析を行うことで本文中に含まれる単語の抽出を行い，その種類のみに注目する．抽出した情報は，すべて関係データベースに格納する．たとえば，**図 1** に示したメールからは**図 2** に示すような情報が抽出される☆．この例に基づき以下に細部の補足を行う．

メール ID はデータベース内のメールに一意に割り当てられるもので，現在の実装ではメーリングリストのサブジェクトに自動的に付加されるシーケンスナンバをそのまま用いている．サブジェクトについては，メーリングリストサーバで自動的に付加されるメーリングリストラベルや，リプライメールなどに付加される 'Re:' などの修飾語を削り，同一メールに対するリプライメールにはすべて元メールと同じサブジェクトに統一する．

また，電子メール本文からシグネチャや，'佐藤>' などの引用符，引用文の前に置かれる 'satoh@hoge.nasda.go.jp Taro Satoh wrote:' といったリファレンス情報を削除する．たとえば，図 1 に示したメールのうち，網掛けした部分が抽出の対象となる．次に，メール本文について形態素解析を行い，普通名詞，固有名詞，サ変名詞を抜き出す．

3.2 単語重要度の算出

メールの内容は多肢にわたるため，データベース内のメールすべてを対象とすると，一般的なルールしか導出されない．たとえば,「思う ⇒ する」「できる ⇒ する」などといったよく使われる動詞の組からなるルールが導出されるが，これらに意味はない．有用なルールを導出するには，送信者やサブジェクトなどを基に互いに関連のあるメールを選択し，そのメール集合に知識獲得手法を適用する必要がある．ただメール集合の要素数が少なくなりすぎると，導出されたルールの信頼性に疑問が出てくるので，最小支持メール数と呼ぶ，ルールにあてはまるメールの最小数をユーザが指定することとする．最小支持メール数を 10 とすると少なくとも 10 のメールにあてはまる知識が導出される．

メール集合の選択にヘッダ情報の送信者やサブジェクトを用いれば，あるサブジェクトの一連のメールで行われた議論における関連ルールや，ある個人の性格を表すような関連ルールが導出される可能性がある．送信日時を用いれば，時系列に沿った傾向の移り変わりが導出されるかもしれない．

また，あるトピックスを表すキーワードを含む一連のメールを選択することが有効であると考えられる．しかし，膨大に存在する単語のうちどの単語に注目すべきか決定する必要がある．最小支持メール数以上のメールに存在する単語をキーワードにして総当たりで調べるというのは，本研究で対象としているような膨大なアーカイブには適用しにくい．

メール本文のキーワードはサブジェクトに出現することが期待されるが，現状では，日本語だと正常に表示できないメーラがまだ存在するため，サブジェクトは英語あるいはローマ字で書くことが推奨されている．そのため，サブジェクトからキーワードを得るには英語から日本語への翻訳，もしくはローマ字から漢字かな混じり文への変換という作業が必要になり，適切な

☆ ガチャピンはフジテレビジョン株式会社の登録商標である．

キーワードを得るのは困難である．

キーワードの抽出については様々な研究があるが，本研究では計算量が少なく，最も一般的に用いられるtf*idf法を用いて，本文中に含まれる各名詞の重要度を算出し，重要度が上位に位置する名詞をキーワードとする．各単語 t の重要度 $I(t)$ の算出は次の式で行う．ここでデータベース内のメール総数を m とする．

$$I(t) = \sum_{i=1}^{m} tf(t, d_i) \cdot idf(t) \cdot W_q(t)$$

$tf(t,d)$（normalized within-document frequency）は，ある語 t があるメール d 中に現れる頻度をメール d 内の形態素数 $M(d)$ で割った値であり，メール長を考慮した正規化を行っている．$idf(t)$（inverse document frequency）は，データベース内の全メールにおいてある語 t が現れるメールの頻度に基づく値であり，次式で定義される．

$$idf(t) = \log \frac{\text{データベース内のメール総数 } m}{\text{語 } t \text{ が現れるメール数}} + 1$$

$idf(t)$ はある語 t が一部のメールに集中している度合いを表しているので，$tf \cdot idf(d,t)$ はある語 t がある文書 d を弁別する能力を表している．たとえば，未知語 t が出現した場合には，$idf(t)$ が大きくなるため，語 t の重要度 $I(t)$ は大きくなるが，その単語が1度しか出現しないような単語であると $tf(t,d)$ が小さくなり，重要度 $I(t)$ も小さくなる．

電子メールを用いたコミュニケーションでは相手のメールの引用が頻繁に行われる．また冗長な引用は嫌われ，必要な部分のみを引用することが推奨されている．なかには全文を引用し末尾に添付するスタイルも存在するが，それは全体から見れば少数派である．すなわち，メールにおける引用部分は元メールのエッセンスを表していると見なすことができ，引用部分に含まれる単語ほど，重要度が高いと推測できる．逆にまったく引用されない単語は，重要でないと考えられる．$W_q(t)$ はデータベース内の全メールにおける単語 t の引用率に基づく重みであり，より多く引用されている単語ほど大きい値になる．

3.3 知識獲得

知識獲得アルゴリズムとしては相関ルールを導出するAprioriアルゴリズム[1]や分類階層（taxonomy）によりアイテム集合の一般化を行う属性指向アルゴリズム[6),7)]などが存在する．属性指向アルゴリズムは非常に有効なアルゴリズムではあるが，属性指向アルゴリズムを用いるためにはシソーラス辞書を用いるなどの手段により概念木を構築し，一般化を行う必要があり，概念木の構築管理に多大な手間がかかる．そこで本研究では，Aprioriアルゴリズムを用いて知識獲得を行う．まず，前節で述べた手順により抽出されたデータを次の2つのテーブルに保持する．

- HEADER（メールID，日付，送信者，サブジェクト）
- WORD（メールID，単語）

この結果生成されるテーブルの例を**表1**，**表2**にあげる．表2のWORDテーブルでは，スペースの都合上，同一メールIDの単語を1行にまとめているが，実際はWORD（メールID，単語）のバイナリリレーションである．SQLで記述することにより関係データベース上でルール抽出が可能である[9),12)]．

膨大なメールアーカイブからの知識獲得を考えた場合，その組合せの数は爆発的に増大し，実用的な時間内で解析が終了しないことが懸念される．そこで本研究では，知識獲得の対象となるメール集合をサブジェクトや送信者などで制限し，その中で成立するルールの導出を目指す．すなわち，HEADERテーブルを参照して知識獲得の対象となるメール集合を選択し，それに含まれるメールについてWORDテーブルから「単語A ⇒ 単語B」といった単語間の相関ルールを導出する．ここで相関ルール「$B \Rightarrow H$」は，B が成立するときに H が成立する事例が多いことを示すものであり，B を本体（Body），H を頭部（Head）と呼ぶ．最終的に導出されるメールは，サブジェクトや送信者による条件が付いた条件付き相関ルールとなる．

表1 HEADERテーブル例
Table 1 An example of HEADER table.

ID	日付	送信者	サブジェクト
765	1998/05/24 15:03	satoh@hoge...	ガチャピン宇宙へ
766	1998/05/24 19:12	kani@rokoj...	ガチャピン宇宙へ
767	1998/05/24 19:20	akiyama@is...	のぞみ打上げ
768	1998/05/25 22:02	naga@cre.n...	ガチャピン宇宙へ
769	1998/05/25 22:08	naga@cre.n...	PLANET-B

表2 WORDテーブル例
Table 2 An example of WORD table.

ID	単語
765	ガチャピン，佐藤，宇宙，ロシア，宇宙船，ニュース，···
766	扮装，ハンドブック，某，ヒーロー，宇宙，掲載，ガチャピン，···
767	発射，全段，飛行，理由，内之浦，プラネット，···
768	ガチャピン，中継，ソユーズ，宇宙，長崎，明日，NEC，出社，···
769	NY，留学，みなさま，ロマン，お願い，自己紹介，公募，···

このような手順を踏むことで，数万通のメールからなるようなデータベースに対しても実用的な範囲内で解析が行えると考えられる．

Apriori アルゴリズムでは，ラージアイテム集合の導出がその第 1 段階となるが，先に述べた最小支持メール数を 10，最小支持度を 0.2 とすると，最小支持メール数と最小支持度の逆数の積（最小候補メール集合要素数）である 50 以上のメールからなるメール集合のみを知識獲得の対象とすればよい．

次に，メール集合を選択する手順について例をあげて述べる．

(1) メールアーカイブ全体に対する知識獲得

データベース内のメール全体に対し，最小確信度，最小支持度を満足するルールを導出する．データベースサイズが大きくなるに従い，最小確信度を満たすルールは存在しにくくなる．データベース全体から得られる知識はそのメーリングリストの性格を示すと考えられる．

(2) ある「サブジェクト」または「送信者」を持つメール集合からの知識獲得

ある共通のサブジェクトあるいは送信者を持つメール集合のうち，その要素数が最小候補メール集合要素数以上であるメール集合において，Apriori アルゴリズムを適用し，相関ルールを導出する．

(3) ある「キーワード」に関連するメール集合からの知識獲得

あるキーワードに関するメール集合を選択するには，そのキーワードを含むかどうかが重要な指針となる．さらにそのキーワードを明示的に含んでいなくても，そのキーワードから派生した話題に関するメールは同じメール集合の中に含めたい．そこで本研究では，メールの相関性を用いて，あるキーワードに関連のあるメール集合を導出する．

3.2 節で求めたキーワードリスト（重要度降順にソート済み）からその出現回数が最小候補メール集合要素数以上である 1 つの単語に注目し，その単語を含むメールのサブジェクト（複数）を求める．それらのサブジェクトを持つメール集合の要素数が最小候補メール集合要素数以上である場合，そのメール集合に対し，Apriori アルゴリズムを適用し，ルールを導出する．

4. 実装および評価

今回，システムは Perl5 による実装を行った．メールの特徴抽出段階における形態素解析には奈良先端科学技術大学院大学で実装された形態素解析ツール茶筅[14]を用いた．ただし，茶筅に付属の基本辞書に加えて，フリーの IME 用辞書を用いて，基本辞書にない 6 万余語の名詞の辞書増強を行っている．また，メール特徴を格納する関係データベースには米 Oracle 社の Oracle7☆を用いた．データベースインタフェースとして DBI for Perl を，データベースドライバとして Oraperl を用いて Perl から Oracle を操作している．

実験用メーリングリストとして宇宙関連のメーリングリストのアーカイブ 760 通を対象とした．先にあげたメール例（図 1），テーブル例（表 1，表 2）および次にあげる導出ルール例は，すべてこのメーリングリスト由来のものであるが，個人名やメールの内容など，評価の正当性を損なわない範囲で改変を行っている．760 通のメールから抽出した単語は 7464 種類．それらに対して tf*idf 法を用いて重要度を算出した結果を **表 3** に重要度上位順に 20 あげる．個人名 A，B，C は当メーリングリストに最もよく投稿する常連の名前である．

表 3 単語重要度
Table 3 Word significance.

順位	単語	出現回数	重要度
1	宇宙	1092	29.73
2	開発	454	18.68
3	メール	214	17.67
4	NASDA	168	15.25
5	日本	294	14.85
6	お願い	174	14.42
7	個人名 A	140	13.71
8	衛星	359	13.45
9	ML	156	13.32
10	ロケット	223	13.25
11	個人名 B	133	12.13
12	メーリングリスト	64	11.97
13	何	186	11.61
14	人	190	10.60
15	話	168	10.48
16	技術	286	10.45
17	個人名 C	42	10.38
18	ガチャピン	38	9.59
19	興味	116	9.45
20	軌道	187	9.21

関係データベースに表 1，表 2 で示した形式で格納したタプル数は計 10 万程度となった．次に最小支持度 0.2，最小確信度 0.3，最小支持メール数 10 として 3.3 節に示した方法でルール抽出を行った．最小候補メール集合要素数は $1/0.2 * 10 = 50$ となる．これら

☆ Oracle および Oracle7 は，米 Oracle 社の登録商標である．

表 4 送信者 A の送信したメールにおける相関ルール
Table 4 An example of association rules in the mail sent by A.

No.	相関ルール	確信度
1	NASDA 衛星 ⇒ 個人名 A	1.00
2	NASDA 開発 ⇒ 個人名 A	1.00
3	開発 ⇒ 個人名 A	1.00
4	衛星 ⇒ 個人名 A	1.00
5	宇宙 ⇒ 個人名 A	1.00
6	NASDA 宇宙 ⇒ 個人名 A	1.00
7	NASDA ⇒ 個人名 A	1.00
8	宇宙 開発 ⇒ 個人名 A	1.00
9	衛星 個人名 A ⇒ NASDA	0.95
10	衛星 ⇒ NASDA	0.95
11	宇宙 個人名 A ⇒ NASDA	0.89
12	宇宙 ⇒ NASDA	0.89
13	個人名 A ⇒ NASDA	0.86
14	開発 ⇒ NASDA	0.83
15	開発 ⇒ 宇宙	0.83
16	開発 個人名 A ⇒ NASDA	0.83
17	開発 個人名 A ⇒ 宇宙	0.83
18	宇宙 ⇒ 開発	0.56
19	宇宙 個人名 A ⇒ 開発	0.56
20	NASDA ⇒ 宇宙	0.38

表 5 「管理者」に関するメールにおける相関ルール
Table 5 An example of association rule regarding the mailing list administrator.

No.	相関ルール	確信度
1	D 大学 ⇒ 個人名 B	1.00
2	開発 ⇒ 宇宙	1.00
3	メール 管理者 ⇒ 個人名 B	0.94
4	現在 ⇒ 宇宙	0.94
5	space-ML ⇒ 個人名 B	0.94
6	お願い ⇒ 宇宙	0.89
7	管理者 ⇒ 個人名 B	0.88
8	NASDA ⇒ 宇宙	0.80
9	人 ⇒ 宇宙	0.80
10	メール ⇒ 個人名 B	0.79
11	宇宙 ⇒ 開発	0.78
12	メール ⇒ 管理者	0.75
13	何 ⇒ 宇宙	0.75
14	個人名 B ⇒ 管理者	0.74
15	衛星 ⇒ 宇宙	0.73
16	メール 個人名 B ⇒ 管理者	0.71
17	人 ⇒ 開発	0.70
18	管理者 ⇒ メール	0.69
19	管理者 個人名 B ⇒ メール	0.65
20	管理者 ⇒ 人	0.62

の閾値はルール数が爆発的に増加しない範囲の最小の値を選んだ．

(1) メールアーカイブ全体に対する知識獲得

ルールとして「開発 ⇒ 宇宙」,「宇宙 ⇒ 開発」などが得られた．宇宙関連のメーリングリストでは妥当な結果である．

(2) ある「サブジェクト」または「送信者」を持つメール集合からの知識獲得

最小候補メール集合要素数 50 を満たすサブジェクトは存在しなかった．最小候補メール集合要素数を満たす送信者は 1 人（個人名 A）のみであった．彼の送信したメール 74 通から導出された全 31 ルールのうち確信度上位 20 ルールを**表 4** にあげる．「NASDA 開発 ⇒ 個人名 A」「NASDA ⇒ 個人名 A」などのルールが確信度 1.0 で導出され，彼が NASDA と宇宙開発に何らかの関係を持つ人物であることが分かる．実際彼は NASDA で宇宙開発に携わっている．

(3) ある「キーワード」に関連するメール集合からの知識獲得

最小候補メール数を満たす単語は全部で 98 種類．そのうち 69 種類が重要度順の上位 100 に入っていた．得られた相関ルール集合のうち，最も分かりやすい知識を導出した例の 1 つとして，重要度 63 位の「管理者」関連の 78 通のメールをあげる．このメール集合からは，計 67 の相関ルールが得られた．**表 5** にそのうちの確信度上位 20 ルールをあげる．「メール 管理者 ⇒ 個人名 B」「space-ML ⇒ 個人名 B」「D 大学 ⇒ 個人名 B」といったルールが確信度 0.9 近くで発見された．個人名 B は space-ML の管理者で D 大学の学生である．

今回の実験では，用意したメーリングリストのアーカイブ総数が 760 通と比較的少量だったこともあり，最小候補メール集合要素数を満たすメール集合が少なく，同一サブジェクトあるいは同一送信者によるメール集合からは，特に有用なルールは発見されなかった．しかし，より多くのメールアーカイブを対象にすれば，たとえば個人の癖や傾向，さらに性格が推測できると期待できる．一方，トピックスによる条件下での知識獲得においては，話題やユーザの傾向を推測する手がかりが発見できた．今回の実験では取り上げなかったが，メールが送信された日時ごとに KDM を行えば，時系列に沿った傾向もつかむことができる．

また，KDM の根幹をなす Apriori アルゴリズムは比較的単純で高速なアルゴリズムであるため，メール数が数万通のオーダーになっても実用的な時間内で処理することが可能である[1)]．また，KDM ではメールのヘッダ情報などを基にメールを複数の集合に分割したうえで知識獲得を行うので，メールアーカイブ全体が大きくなっても知識獲得の対象となる集合は，多くの場合それほど大きくならないと考えられる．

5. 関連研究

本研究が目指しているように大量の文書データに記述されている多種多様な内容を対象として，その相関関係や出現傾向などから新たな知識を発見することを目的とする技術をテキストマイニング（text mining）と呼ぶ[16]．テキストマイニングの分野では，いかにして文書中の概念を抽出するか，いかにしてその概念から有用な知識を得るか，いかにして得られた知識を分析するかについて様々な研究がなされている．対象とする文書類は多岐にわたっているため，すべてを同様の手法で扱うのは無理がある．有用な知識を得るためには，知識獲得の対象となる文書類の特性を考慮し，その特性を活かすように知識獲得手法を工夫する必要がある．本研究では特にこれから先利用が拡大すると見込まれる電子メールを対象に知識獲得を試みた．テキストマイニングでなされている各種の手法と電子メールの特性を組み合わせることで，より精度の高い知識を獲得できる可能性がある．本稿では取り上げなかった他の手法との連携は今後の課題である．

電子メールからの知識獲得として，シグネチャから送信者の住所録情報を抽出し，それを用いた住所録管理システムを構築する研究[2]が報告されているが，これはシグネチャに特化したものであり，本研究のように電子メールの本文を扱うものではない．会議の告知や論文募集に関する電子ニュースからセンタリングなどのレイアウトを考慮したパターンマッチングを行うことにより，重要語を検出し，サマリーを自動生成する研究[18]も報告されているが，その適用文書は意図的に整形したものや，目的のはっきりしたものに限られる．また，電子メールから，日時・場所などのスケジューリング情報をその表現パターンとのパターンマッチングによって抽出する研究[8]が報告されているが，やはりスケジューリング情報の抽出に特化したものであり，本研究とは方向性が異なる．ただ，重要語の検出手順などは，本稿で提案した手法の精度を向上させるのにも有効であると考えられ，それらを検討するのは今後の課題である．

一方フィルタリングの分野では，コンテンツベースの情報フィルタリング技術が数多く報告されている．しかし，現在までに報告されているフィルタリング技術は定型的な文書を対象にしたものが多く，メールのように自由度の高いものについては，ヘッダ情報によるフィルタリングが主流である．中には，SPAM メールを対象に意味解析を行い，SPAM メールを自動的にはねる商用ソフトウェア[13]も発売されているが，その目的は SPAM メールかそうでないかを判別することに限られる．

獅々堀らは，新規メールの重要度算出に各個人が優先度付けした既存のメール文書に基づいたコーパスベースの手法を提案しているが[19]，本研究のように知識の導出を目的としたものではない．本研究と比較的アプローチが類似していると考えられる商用アプリケーションとして，シャープの Datahunter[4]があげられる☆．Datahunter ではテーマに沿って検索した複数のメールの中から，特徴的に含まれるキーワードを取り出したり，ある語の使用頻度を時間や日付に沿ってグラフ化したりできるので，情報の分析に役立てることができる．しかし，それらの知識発見プロセスはすべて人間の手に委ねられており，本研究のように KDD を用いたものではない．

北川らは，メール間の構造情報および内容情報を用いたメール群の組織化機構の提案と，それを用いたメールダイジェストの生成手法について論じている[11]．メール間の類似度を手がかりにして，話題ごとにメール群を分割することは有効であると考えられるが，本手法が対象としているような大量のデータを扱う場合には適用しにくい．

6. KDM の応用例および今後の展望

本手法で導出される相関ルールの利用例について議論する．サーバに蓄えられたメールからある事象に関するメールを検索する場合，従来全文検索を行い tf*idf により順位付けして出力することが行われてきた．しかし検索を行うユーザが適切な検索キーワードを指定することができない場合には，期待するような検索が行われない．そのため検索キーワードに適当な語を追加する問合せ拡張（query expansion）と呼ばれる手法が提案されている．これには様々な手法が存在するが，代表的なものとして，シソーラス辞書を用いて検索キーワードの同義語を追加するシソーラス法，一度検索を行い得られた結果（みなし正解）中の重要語を抽出し，検索キーワードとして追加する関連フィードバックなどがある．

相関ルールを用いることで従来の手法では検索できなかった文章を検索することが可能である．まず，データベースから検索キーワードを含むメールを抽出し，ユーザの決める閾値以上のメールに共通するサブジェクトを導出する．それらのサブジェクトごとにメールを分割し，それぞれのメール集合において，検索キー

☆ Datahunter はシャープ株式会社の登録商標である．

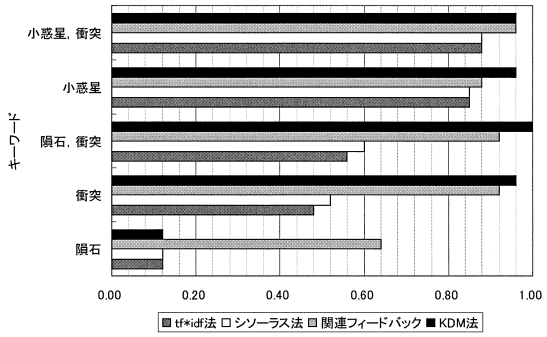

図 3 キーワードに対する再現率
Fig. 3 Recall to keyword.

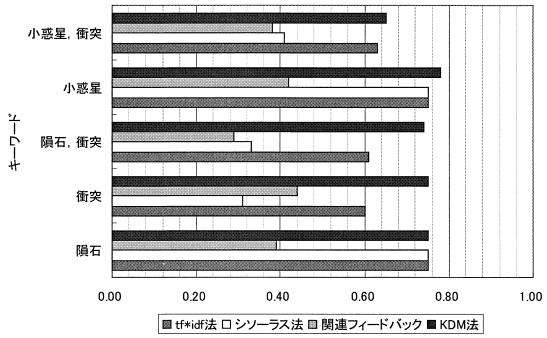

図 4 キーワードに対する適合率
Fig. 4 Precision to keyword.

ワードを本体とする相関ルールを導出し，その頭部を検索キーワードとして追加して再検索を行う．最後に各メール集合で導出された結果をまとめて出力する．こうすることで，不必要に検索結果を増大させることなく，目的のメールを得ることが可能になる．

一例として，近年話題になっている地球への隕石衝突についてのメールを検索したいとする．先の実験で用いた電子メールアーカイブを対象に，単純な tf*idf 法，シソーラス法，関連フィードバック，それに KDM による相関ルールを用いた KDM 法のそれぞれについてキーワードを変化させて再現率と適合率を算出した結果を**図 3**，**図 4** にあげる．再現率はデータベース内の正解メール中実際に導出できた割合であり，適合率は導出したメール中の正解メールの割合である．

図 3 の tf*idf 法の結果から，上側に行くに従いメール本文中に実際に出現する良いキーワードであるといえる．「小惑星，衝突」と検索キーワードを指定した場合は，すべての手法で高い再現率が達成されている．しかし，結果に著しい差が生じるのは，そのほかの不完全な検索キーワードを与えたときの結果であり，こうした不完全なキーワードでも，適切な結果を返すことができる検索手法が有効であるといえる．

直感的には「隕石」というキーワードが良い結果を返すと期待していたが，関連フィードバックを除く手法の再現率が 0.12 と非常に低くなっている．これはこのメーリングリスト内では隕石ではなく小惑星という言葉が使われていたことによる．シソーラス法で隕石に対して追加されたキーワードは「いん石」「メテオ」「流星」の 3 つでありキーワード追加による効果はなかった．関連フィードバックでは「爆発」などが追加され，再現率に若干の改善が見られている．KDM 法では，1 回目の検索で最小支持度を満たすだけのメールが得られなかったため tf*idf 法にくらべて改善は見られない．

しかしそれ以外のキーワードでは，KDM 法は高い再現率を出している．再現率においては関連フィードバックも良い結果を出しているが，適合率が著しく低くなっている．それに対して，KDM 法では適合率も十分に高い．このように，KDM 法はメーリングリストのアーカイブなどの大量のメールから情報を検索する際に，有効であると考えられる．

このようにキーワードの共起性に基づき関連検索を行う商用システムとしてはジャストシステム社の ConceptBase [5),17)]☆があげられる．ConceptBase では，高度な自然言語処理技術を駆使して，対象に合わせて最適な検索条件を生成し，大量のテキスト文書に関して，自然文による類似情報抽出を行える．ConceptBase では，あらゆるテキストを同様に扱うことが可能だが，本研究で用いたようにその対象となる文章の特性を考慮することでさらに精度の高い検索を行える可能性がある．

有効な相関ルールが数多く発見されれば，メーリングリストにメールを出す際に，自分のメールに対する反応をある程度予測するという利用方法も考えられる．たとえば，ある人はこのメールの話題に乗ってくるだろうとか，ある人はこの話題には否定的だろう，などということが分かる可能性がある．さらに，メールに含まれる単語間の相関ルールから，そのメールに含まれた意味を推測するという利用方法も考えられる．

7. ま と め

本稿では，大量に溜まった電子メールアーカイブから有用な情報を得るために，ヘッダ情報や形態素解析を用いて抽出した単語をデータベースに格納し，それに Apriori アルゴリズムを適用して知識獲得を行う手法について述べた．さらに実際にメーリングリスト

☆ ConceptBase は株式会社ジャストシステムの登録商標である．

アーカイブに本手法を適用した結果，メーリングリストでの話題やユーザの傾向を推測する手がかりを得ることができることを示した．知識の発見は，メール集合の選択の仕方に非常に大きく左右されると考えられ，どのようにメール集合を選択するかの考察は今後の最重要課題である．また，膨大に導出される知識から有用な知識を選択し提示する工夫が必要である．

今後，知識獲得精度を上げていくことにより，たとえば，企業などがユーザから受け取った大量のメールに対して，KDM を行うことによりユーザのニーズをつかむことができると考えている．

謝辞 本稿を進めるにあたり，貴重なご助言をいただいた春本要講師，原隆浩助手をはじめとする当研究室諸氏に謝意を表す．なお，本稿は，日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」（プロジェクト番号：JSPS-RFTF97P00501）によっている．ここに記して謝意を表す．

参 考 文 献

1) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. 20th International Conference on Very Large DataBases*, pp.487–499 (1994).
2) 浅野久子，加藤恒昭，高木伸一郎：Signatue の局所的パターンマッチングによる電子メールからの送信元住所録情報の抽出とそれを用いた住所録管理システム，情報処理学会論文誌，Vol.39, No.7, pp.2196–2206 (1998).
3) Chen, M.-S., Han, J. and Yu, P.S.: Data Mining: An Overview from a Database Perspective, *IEEE Trans. Knowledge and Data Engineering*, Vol.8, No.6, pp.866–883 (1996).
4) Datahunter: http://www.sharp.co.jp/datahunter/.
5) 藤田澄男：自然言語処理を利用した情報の検索・分類へのアプローチ，情報処理，Vol.40, No.4, pp.352–357 (1999).
6) Han, J., Cai, Y. and Cercone, N.: Knowledge Discovery in Databases: An Attribute-Oriented Approach, *Proc. 18th International Conference on Very Large Data Bases*, pp.547–559 (1992).
7) Han, J. and Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases, *Proc. 21st International Conference on Very Large Data Bases*, pp.420–431 (1995).
8) 長谷川隆明，高木伸一郎：電子メールコミュニケーションにおけるスケジュール情報抽出，情報処理学会自然言語処理研究会資料，123-10, pp.73–80 (1998).
9) Houtsma, M.A.W. and Swami, A.N.: Set-Oriented Mining for Association Rules in Relational Databases, *Proc. 11th International Conference on Data Engineering*, pp.25–33 (1995).
10) 河野浩之：データベースからの知識発見の現状と動向，人工知能学会誌，Vol.12, No.4, pp.497–505 (1997).
11) 北川雅嗣，水内祥晃，田島敬史，田中克己：メイリングリスト情報の組織化のためのクラスタリングとカット検出，電子情報通信学会データ工学ワークショップ（DEWS'97）論文集，pp.221–226 (1997).
12) 喜連川優：データマイニングにおける相関ルール抽出技法，人工知能学会誌，Vol.12, No.4, pp.513–520 (1997).
13) MailGoGoGo!: http://www.makie.com/mailgogogo.html.
14) 松本裕治，北内 啓，山下達雄，今一 修，今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書，技術報告 NAIST-IS-TR97007, NAIST (1997).
15) 森田昌宏，速水治夫：情報フィルタリングシステム—情報洪水への処方箋，情報処理，Vol.37, No.8, pp.751–758 (1996).
16) 那須川哲哉，諸橋正幸，長野 徹：テキストマイニング—膨大な文書データの自動分類による知識発見，情報処理，Vol.40, No.4, pp.358–364 (1999).
17) 野村直之：ConceptBase の言語処理と新しいソリューション，情報処理学会自然言語処理研究会資料，129-1, pp.1–8 (1999).
18) 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，情報処理学会論文誌，Vol.36, No.10, pp.2371–2379 (1995).
19) 獅々堀正幹，藤井 誠，安藤一秋，青江順一：各個人のプロファイルを用いたメイル文書のフィルタリング手法，電子情報通信学会技術研究報告，DE98-31, pp.9–16 (1998).
20) 寺野隆雄：KDD ツールの動向と課題，人工知能学会誌，Vol.12, No.4, pp.521–527 (1997).
21) 日経マーケット・アクセス：インターネット普及率調査（2000 年春），http://www.nikkeibp.co.jp/MA/.

(平成 12 年 5 月 5 日受付)
(平成 12 年 10 月 6 日採録)

上田　宏高（学生会員）

1997年大阪大学工学部情報システム工学科卒業．1998年同大学大学院工学研究科情報システム工学専攻博士前期課程修了．現在，同後期課程在籍．現在の研究テーマは，ウェアラブルコンピューティング，拡張現実感等であり，宇宙旅行が現実になることを夢見ている．

柳沢　　豊（正会員）

1990年大阪大学工学部情報システム工学科卒業．1996年同大学大学院博士課程修了．同年，日本電信電話（株）に入社，現在同社コミュニケーション科学基礎研究所に所属．博士（工学）．知識処理，データベース，空間情報処理の研究に従事．人工知能学会ほか4学会の各会員．

塚本　昌彦（正会員）

1987年京都大学工学部数理工学科卒業．1989年同大学大学院工学研究科修士課程修了．同年，シャープ（株）入社．1995年大阪大学大学院工学研究科情報システム工学専攻講師，1996年より，同大学院工学研究科情報システム工学専攻助教授，現在に至る．工学博士．時空間データベースおよびウェアラブルコンピューティングに興味を持つ．ACM，IEEE等7学会の会員．

西尾章治郎（正会員）

1975年京都大学工学部数理工学科卒業．1980年同大学大学院工学研究科博士課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授を経て，1992年より大阪大学大学院工学研究科情報システム工学専攻教授となり，現在に至る．2000年より大阪大学サイバーメディアセンター長を併任．この間，カナダ・ウォータールー大学，ビクトリア大学客員．データベース，知識ベース，マルチメディアシステム，分散システムの研究に従事．現在，ACM Trans. on the Internet Technology，Data & Knowledge Engineering，Data Mining and Knowledge Discovery，The VLDB Journal等の論文誌編集委員．ACM，IEEE等8学会の会員．