自然言語処理技術を利用した電子メールのデータベース化に関する提案

喜名 眞魚† 片岡 信弘†

インターネット上のサービスの一つとして電子メールがあり，その重要性はますます高くなってきている．本稿では，電子メールの問題点に着目し，ユーザの負担が少なくかつ効率的な電子メールデータベース化システムの提案を行う．本提案では，自然言語処理技術と次世代 Web であるセマンティック Web の技術を用いての電子メールへのメタデータ付加とメタデータでのメール管理データベースの提案をする．

The proposal of E-mail processing using natural-language-processing technology

Mao KINA† Nobuhiro KATAOKA†

There is an E-mail as one of the services on the Internet, and the importance is becoming still higher. This paper focuses on the problem of the E-mail and proposes an efficient database system with few burdens of a user. By this proposal, mail management by the metadata addition to the E-mail using technologies natural-language-processing and the Semantic Web that is the next generation Web system.

1． はじめに

インターネットの普及，社会の情報化に伴い個人・企業を問わずに電子メールでの情報交換や情報入手が一般的に行われている．電子メールは，一過性の情報ばかりでなく技術的な情報，各種トラブル情報，客先情報など蓄積し利用されるべき情報も多い．蓄積された膨大な電子メール本文の中には有益な情報も多いが，それを人間が検索して探し出すことは意外と手間がかかることがある．またスパムメールと総称される受け手の意図にそぐわない電子メールが受信箱を圧迫し，電子メールユーザのストレスになりつつある．本稿では，電子メールの検索について取り上げる

一方，Web 上のドキュメントやディジタルデータ化されたテキストファイルに対して，自然言語処理技術の実用化が行われるようになってきた．

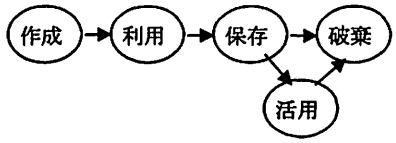
本稿では，電子メールの情報を元に自然言語処理とセマンティック Web の技術を用いデータベース構築する方法を提案する．

これによりユーザは，膨大な情報の中から必要な情報をすばやく見つけ出すことが可能となる．

†東海大学工学研究科電子工学専攻
Department of Electronics
Graduate School of Engineering, Tokai University

2. 電子メールシステムの現状の問題点

電子メールの基本的な技術は1970年代に考案されて以来，大きな変更は受けていない．個人の受信箱にメールを取り込みそこで処理を行う．メールの保存には内容ごとに分けたフォルダを作成しその中に溜めて行くことがほとんどである．電子メールが使われ始めた当初は一日に数通のメールしか受け取らなかったと考えられるため，フォルダを使ってのメール整理もストレスにはならなかったと考える．しかし，現在では一日に何十通ものメールを受け取ることは珍しくなく，これらのメールをフォルダに分ける作業に時間



理想とする情報のライフサイクル

現状の電子メールシステムのライフサイクル

図1 情報ライフサイクル

を費やすことがある．また，メールの量が増えるに従いフォルダの数も増え，フォルダが階層構造を持つようになる．階層化されたフォルダの中に入れられたメールはたとえそれが有用な情報を持っていたとしても，目に付くことはなくなる．メールソフトには検索機能も備わっているが，一般的には単純なパターンマッチングによる検索である．パターンマッチングでは、同義語や類似との理解できないため、絞込みが不十分であったり、検索漏れがあったりといった問題が生じる。つまり，現状ではフォルダに入れられた時点で電子メールのライフサイクルは終わっているのではないかと考える（図1参照）．結果としてフォルダに溜められた電子メール情報は活用されていないケースが大部分であると考えられる．

3. 提案内容

3.1 自然言語処理技術を用いてのアプローチ

このような限界に近づいた電子メールの管理方法に変わる手法として自然言語処理技術を利用しての管理方法を提案する．

自然言語で書かれている電子メール本文に対して以下の(1)-(5)の手順で処理を行うことにより，コンピュータ可読なメタデータを作成する．

(1) 前処理（メール本文から引用符のある行やシグネチャを取り除く）
(2) 形態素解析をして，「名詞」「未知語」を抜き出す
(3) TF/IDF 法を用いての，単語の重み付けを行う
(4) ベクトル空間法を用いて，類似メールの分類処理を行う
(5) 得られたデータをメタデータとして各メールに付加する

次に各処理の詳細についての内容を述べる．

3.1.1 前処理

前処理として，引用文・シグネチャ・その他解析に必要のないものを取り除く．

本文各行の先頭に引用記号がある場合，その行は引用文であるとみなし，取り除く．本文の最終行から数行さかのぼった所に多数の改行，または一定の記号の並びがあった場合，それ以降をシグネチャとみなし，取り除く．

3.1.2 形態素解析処理

形態素とはそれ以上分解してしまうと意味を消失してしまう最小の文字列のことである．形態素解析により形態素に分類されたものの中から，「名詞」と「未知語」をぬきだす．「未知語」とは形態素解析時に使用された辞書内になかった形態素であり，それを含む電子メールの特徴になる．

3.1.3 TF/IDF 法での処理

TF/IDF 法は，単語の出現頻度に基づいての重み付けをする手法である．この手法では TF と IDF という二つの指標を用いて単語の重み付けをする．

TF(Term Frequency)法
・ひとつの文書内に繰り返し現れる単語はその文書を特徴付けるため重要となる

IDF(Inverted Document Frequency)法
・特定の文書中にしか現れない単語はその文書を特徴づけるものになるため重要となる

これら二つを式で表す一例を示す．
tf(w,d) = 文書 d に語 w がどれだけ出現するか
idf(w) = 全文書数 / 語 w が出現する文書数

上 2 式を用いて語 w の文書 d における重要度を表すと，以下のようになる．
重要度 = tf(w,j)log(idf(w))

メール本文を形態素解析することで得られる「名詞」「未知語」の重要度を TF/IDF 法で求め，ある閾値を設定することでそのメールの特徴を決定付ける．

重要単語とその単語重要度の決定は，システムの設計において重要な点である．重要度の設定の仕方によって次項での特性ベクトルの向きが大きく変わってくる．
TF/IDF 法は比較的文書長のあるものに対しては正確なスコアリングが行われる．しかし，電子メールのような文書長が様々のものに対しては，文書長の正規化を行う必要がある．

文書長の正規化にはコサイン正規化やピボット正規化があるが、ここではピボット正規化を用いる．コサイン正規化は単純な式による正規化で比較的よく用いられている．しかし，文書長が長いほどその文書内に含まれる出現回数の高い単語の重要度を小さく見積もり，文書長の短い文書に含まれる単語により重みをおく傾向がある．ピボット正規化は，コサイン正規化での短い文書を優先してしまうという点を解消している．

3.1.4 ベクトル空間法での処理

TF/IDF 法によって求められた単語の重要度に基づき，メールの類似度を判定する．各メールの特性ベクトル M は，TF/IDF 法で求められた語 Wi の重要度を wi とすると次のように現される．

M = (w1,w2,...,wi,...,wn)

ここでは，2 つの特性ベクトルのなす角のコサイン値を類似度として用いる．[2]

$$\text{類似度}(Ma,Mb) = \frac{|Ma||Mb|}{Ma \cdot Mb}$$

3.1.5 メタデータ

電子メール本文から抽出された，そのメールを特徴付ける語（重要度の高い語），類似度の高いメールの情報，類似度の低いメールの情報を，送り主の名前やメールアドレスとともにメタデータとしてその電子メールに付加する．このメタデータは電子メールの意味情報を持つことになる．メタデータの記述には RDF(Resource Description Framework)をもちいる．RDF は主語（リソース），述語（プロパティ），目的語（プロパティの値）の 3 要素の組み合わせで成り立つ．リソースは記述するメタ情報の対象，プロパティは記述するメタ情報の内容・項目である．プロパティの値が新たなリソースになることもある．

表 2 電子メール解析結果

mail A の解析結果	
送信者	someone@test Mao KINA
重要単語（重要度）	KW1 (1.0) KW2 (0.89) … KWn (0.7)
類似度の高いメール	mail D
類似度の低いメール	mail E

RDF のモデルはラベル付有向グラフで表され，その実装には XML を用いる．

表２のような，ある mail A に対する解析結果が得られたとする．　この結果を RDF の有向グラフで表すと，図３のようになる．

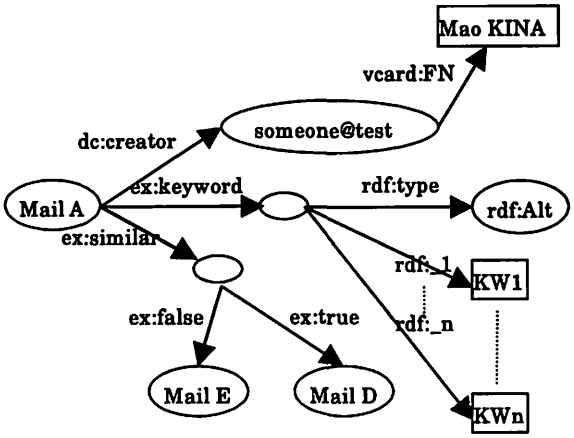


図３　RDF 有向グラフ

```
<rdf:RDF ...    >
 <rdf:Description rdf:about="mail A">
  <dc:creator rdf:resource=someone@test
              vcard:FN="MaoKINA" />
  <ex:keyword>
   <rdf:Alt>
    <rdf:li ex:score="1.0" ex:keyword="KW1" />
            ....
    <rdf:li ex:score="0.7" ex:keyword="KWn" />
   </rdf:Alt>
  </ex:keyword>
  <ex:similar ex:true="mail D" ex:false="mail E" />
 </rdf:Description>
</rdf:RDF>
 (XML 宣言，NameSpace 宣言省略)
```

図４　RDF の XML 表記

図中の接頭辞 dc は Dublin Core の，vcard は vCard の，ex は独自に定義した XMLNamaSpace を表す．

この RDF モデルの XML 表記は次のようになる（図４参照）

3.2　メタデータによるメール管理

リソース（電子メール本体）の意味内容記述と，リソース同士のつながりを RDF メタデータであらわせれば，それをもとにメールの管理を行う．まず，類似度の高い電子メールを自動で振り分けられる．これは，フォルダ分けのように固定的な振り分けではなく，類似度の閾値を指定することで柔軟な振り分けが可能になる．

つぎに，RDF スキーマ・オントロジを用いることにより，更なる意味づけが行える．RDF スキーマは，プロパティやより一般的なリソースについての基本的な枠組みを作ることにより，語彙の定義をする．オントロジは，リソースを体系ごとに分類・関連付けて，推論を可能にさせる「辞書」である．オントロジを用いることにより RDF に記述された電子メールを特徴付けている語句同士の，論理和・論理積・論理差などといった組み合わせや，一致・反対といった関係を表現できる．これらを用いることで，重要単語同士の意味をつなげる事ができる．

従来のパターンマッチングによる検索では，検索キーワードと完全に一致している文字列を持つ電子メールしか探し出すことができないが，RDF とオントロジをもちいることで，キーワードの意味を理解して，同義語や類似語を考慮しての検索が可能となる．また，同意の重要単語をもつ電子メール同士を集めての自動分類も可能である．

4.　全体構築設計

4.1　対象ユーザ別の実装方法

4.1.1　個人ユースの場合

個人ユーザが電子メールのデータベース化を行う場合，MUA（Mail User Agent）の機能として実装する．MTA（Message Transfer Agent）より受信した電子メールと送信メールに対して本稿 3.1 の処理を行い，メタデータ作成する．電子メールとそのメタデータを対とし，メールデー

タをハードディスク内に蓄積してゆく．この際，類似度の高い電子メール同士を自動的に同じフォルダに入れたり，従来と同じくユーザの任意のフォルダに入れたりすることで，メールの管理を行える．

ユーザが電子メールの検索を行う場合，電子メールの意味情報である，メタデータを検索の対象とする．検索時はオントロジを用いて検索キーワードと電子メールの重要単語の意味を拡張し，より精度の高い検索を行う．これを図5に示す．

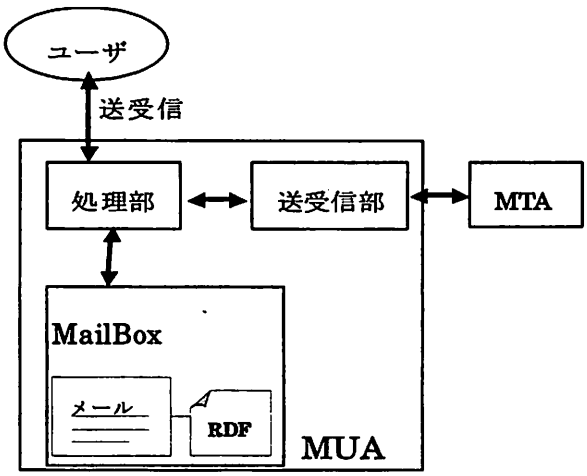


図5　個人ユースの場合の概略図

4.1.2　グループユースの場合

グループ内のユーザの電子メールをデータベースとして活用する場合，MTA内に本稿3.1の処理機能を組み込む．ユーザはユーザ登録を行うことでこのシステムを利用することができる．ユーザ登録の方法は，電子メールアドレスによる登録と，ドメインを登録してのグループ全体の登録を考える．ユーザ登録の時，パスワードを発行し、このパスワードを用いてシステムにログインする．グループでの利用の場合には，ひとつのMTAの中に複数個のデータベースを構築することが可能である．

このシステムはMTAを通るすべてのメールに対して処理を行う．処理された電子メールとそのメタデータはデータベースに蓄積されてゆく．他人に公開したくない内容の電子メールは，送信者が電子メールのサブジェクト欄に特定の文字列を入れることにより、フィルタリングされ，処理部に送られないようにする．蓄積された電子メールの検索はブラウザから行うほうが汎用性が高いと考える．そのため，電子メールを蓄積する際，HTMLデータに変換する．

ユーザが電子メールを送受信する際は，通常の場合と同様に，MTAからMUAを使っての受信，MUAからMTAへの送信依頼というかたちで行う．

検索は，データベース上の電子メールデータをWebベースの検索システムを用いてブラウザから検索する．この時，検索キーワードと電子メールの重要単語はオントロジを用いて意味拡張をし，精度の高い検索を行う．これを図6に示す．

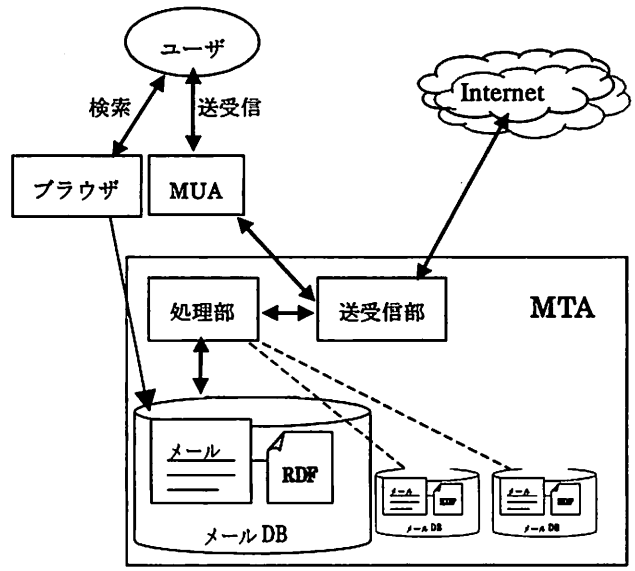


図6　グループユースの場合の概略図

4.2　評価方法

本稿での提案の評価には，重要単語の抽出とその重要度の設定の的確さと，電子メールの検索にかかる時間を検証する．

重要単語の的確さに対する評価は，200通程度の内容のある電子メールやメーリングリストでやり取りされる電子メールに対して，本稿での方法での重要単語の抽出と，人手による抽出を行い，その重要単語を比較して評価する．

検索に対する評価は，従来のパターンマッチングでの検索で目的のものを探すまでの時間と，本稿での提案システムによる検索にかかる時間の比較を行う．また，蓄積されたメールの増加に対して、システムにかかるオーバーヘッド増加の割合の評価も行う

5. セマンティック Web 技術を用いたメールシステムへの拡張

セマンティック Web は Web コンテンツなどに意味情報をもたせ，コンピュータによる解釈・自動処理を目的とするものである．RDF・オントロジはセマンティック Web での基盤となる技術である．

本提案の拡張として，セマンティック Web 技術を用いたメールシステムの一例を示す．使用環境は，企業等の集団，あるいは部門ごとにおいてやり取りされた電子メールのすべてを蓄積し，それを知識データベースとして活用するというものを想定する（図 7 参照）．

メールサーバを通るすべてのメールを対象に本稿 3.1 での手順に従って RDF データを付加してゆく(タグ付け)．タグ付けされたメールはデータベース内に蓄積してゆく．ユーザはエージェントを介してオントロジを使い蓄積されたメールの中から，有用なものをマイニングすることが可能になる．

エージェントとはセマンティック Web において人に代わって推論を行うプログラムの総称である．

エージェントは自立的にメールに付加されたメタデータを検索し，メールを特徴付けている重要単語とオントロジに基づく同義語・類似語をユーザの指定する検索キーワードと比較し目的のメールを探し出す．

6. まとめ

本稿では，電子メールの問題点についての検討をもとに，自然言語処理とセマンティック Web の技術を用いての電子メール処理手法の提案を行った．

今後は今回提案した内容の実装と評価・検証を行う予定である．また，次のステップとして本稿 5 で述べたセマンティック Web 技術でのエージェントの開発に取り組んでいく予定である．

参考文献

[1]上田宏高 他 3：電子メールの傾向分析への知識獲得手法への適用，情報処理学会論文誌 Vol.41 No.12 2000 pp 3285-3294

[2]野口進祐，木下哲夫，白鳥則朗：参照情報を利用した文書特徴量抽出方式，GW, Vol.2000 Num.45 pp 103-108

[3] 小倉　弘敬，他 5 ：セマンティック Web の応用システム
情報処理 No.43,pp742-750,2002

[4]松本　裕治：形態素解析システム「茶筌」
情報処理 No.41,pp1208-1214,2000

[5] Semantic Web(W3C)
http://www.w3c.org/2001/sw/

[6] INTAP セマンティック Web 委員会
http://www.net.intap.or.jp/INTAP/s-web/index.html

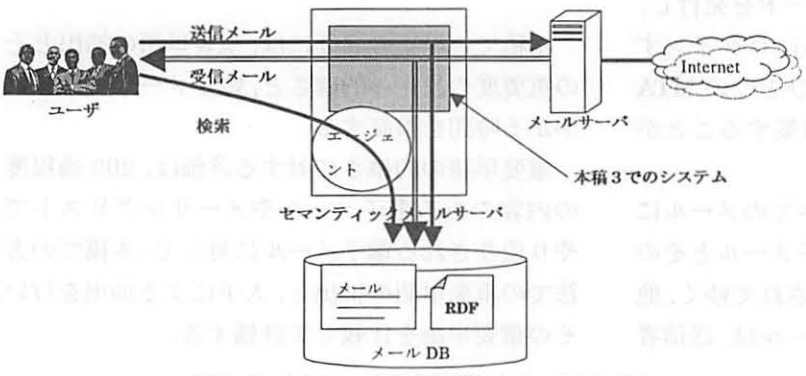


図 7 セマンティック Web を用いたメールシステム