

RDF を用いた電子メール管理の提案と検証

喜名 眞魚 片岡 信弘
東海大学連合大学院理工学研究科総合理工学専攻

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の提案をする。

The proposal of E-mail management using RDF meta-data

Mao KINA Nobuhiro KATAOKA
United Graduate School of Science, Tokai University

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。スパムメールに対してはさまざまなアンチスパム製品・サービスがあり、研究も行われている[1]。本稿では、電子メールの管理・検索について取り

上げる

一方、Web 上のドキュメントやデジタルデータ化されたテキストファイル、既存のデータ・コンテンツに対して、自然言語処理技術やセマンティック Web 技術の応用・実用化が行われるようになってきた[2]。

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 の技術を用いて電子メールを管理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。

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。また、インターネットが現在のように生活に密接にかかわる以前は、電子メールの利用は仕事や研究など決まった用途での物が主であった。しかし、現在でインターネットの爆発的な普及により、電子メールは個人間の利用や広告・勧誘用のメディアとして利用されることも多く、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。

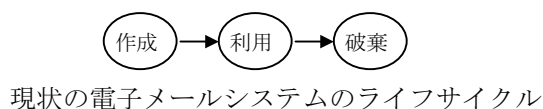
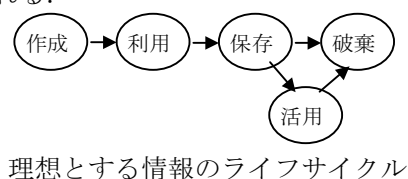


図1 情報ライフサイクル

この膨大な電子メールを処理する手法に関し

ては、さまざまな研究が行われている[3][4][5][6]。本提案では、これら既存の技術をベースにセマンティック Web でのメタデータ技術を組み合わせ、電子メールの管理手法を提案する。

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・その他解析に必要なものを取り除く。

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」をぬき出す。「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の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の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、システム的设计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 法は比較的図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図書長が様々のものに対しては、図書長の正規化を行う必要がある。

図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、図書長が長いほどその文書内に含まれる出現回数が高い単

語の重要度を小さく見積もり、図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 W_i の重要度を w_i とすると次のように現される。

$$M = (w_1, w_2, \dots, w_i, \dots, w_n)$$

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。

$$\text{類似度}(M_a, M_b) = \frac{|M_a| |M_b|}{M_a \cdot M_b}$$

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。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。図中の接頭辞 dc は Dublin Core の、vcard は vCard の、ex は独自に定義した XMLNameSpace 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。

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

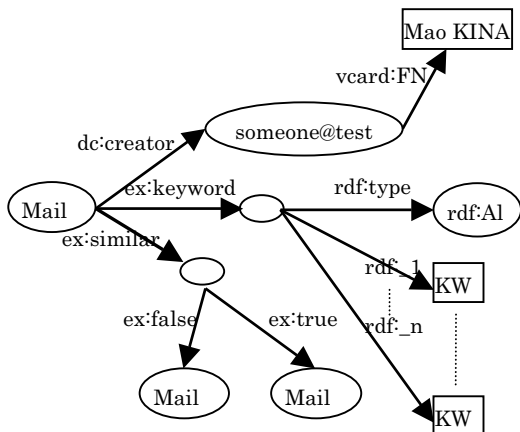


図 3 RDF 有向グラフ

```
<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

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

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、メタ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。

5. 評価方法

本稿での提案の評価には、電子メール分類の的確さと、電子メールの検索にかかる時間を検証する。

分類の的確さに対する評価は、1000~2000通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を IDF 値算出のためのコーパスとして、100~200通のメールに対して本稿での分類方法と、人手による分類を行い、比較して評価する。また、分類のためのカテゴリを設定し、各カテゴリに当てはまる語句を電子メールの中から集め、それらを元に RDF スキーマを作る。

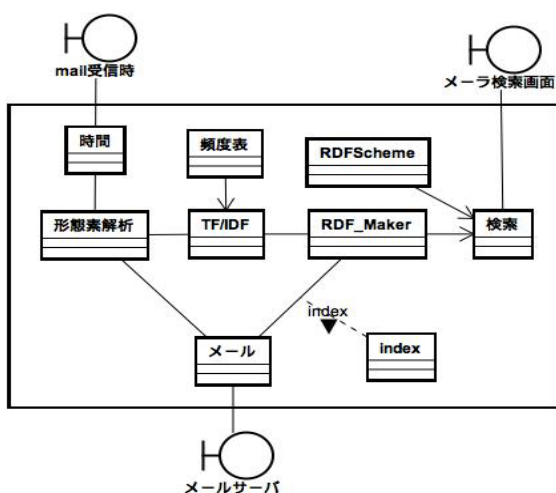


図5 設計概略図

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。

6. 電子メール分類の結果

2439通（約13.2MB）の電子メールをコーパスとし、150通（約1.0MB）の電子メールに対して形態素解析と TF/IDF 法を施し語句の集合を得た。これを元に RDF スキーマを用いて電子メール进行分类し、人手による分類との比較を行った。この結果を、表6に示す。

表6 各カテゴリへの分類正解率

カテゴリ	正解率
特定話題 1	56.8%
特定話題 2	61.7%
添付ファイル	75.0%
広告メール	0.0%
一過性メール	38.7%

7. 考察

内容の限られる特定話題に関する分類は、ある程度行えると考えられる。さらに精度を上げるには、単語の重み付け手法を今回用いた一般的な TF/IDF 法ではなく、電子メールに特化した手法をとる必要があると思われる。さらに、分配ルールである RDF スキーマの作成方法を考える必要がある。

8. 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管理手法の提案を行い、この手法がある程度分類を行えることを確認した。今後の課題として、重み付けの強化・知識処理の利用が挙げられる。

参考文献

- [1]佐々木稔 新納浩幸：文書分類を用いたスパムメール判定手法，情報処理学会研究報告 情報学基礎研究報告 2004 (93) pp 75-82
- [2]片山佳則 他 4：セマンティックグループウェア WorkWare++と KnowWho 検索への応用，情報処理学会研究報告 データベースシステム研究会報告，2003(51) pp9-16
- [3]川前徳章：構造と特徴選択に着目した電子メールの分類手法の提案－From フィールドと Jeffreys Perks を用いたナイーブベイズ分類－，情報処理学会研究報告 自然言語処理研究会報告 2005(73) pp 27-33
- [4]上田芳弘 他 5：テキストマイニングと強化学習を用いた電子メール自動分配，電子情報通信学会論文誌 D-I Vol.J87-D-I No.10 pp 887-898 2004/10
- [5]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
- [6]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
- [7]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
- [8]松本 裕治：形態素解析システム「茶釜」
情報処理 No.41,pp1208-1214,2000
- [9] Semantic Web(W3C)
<http://www.w3c.org/2001/sw/>
- [10] INTAP セマンティック Web 委員会
<http://www.net.intap.or.jp/INTAP/s-web/index.html>