4Q-2

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ついての提案と検証

喜名　眞魚†　　片岡　信弘†
東海大学工学研究科電子工学専攻†

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．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．

本研究では，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．

2．自然言語処理技術を用いてのアプローチ

2.1　メールの処理手順

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．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，同義語や類似語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．

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 RDF メタデータとして各メールに付加する．

The proposal of E-mail processing using natural language processing technology
†Mao Kina　Nobuhiro Kataoka
Graduate School of Electronics, Tokai Univ.
3aeem014@keyaki.cc.u-tokai.ac.jp

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．さらに，RDF スキーマ，オントロジを用いることで語句の意味づけや語句同士の関係を表現できる．これをもとに，自動的にフォルダ分けや高精度な検索を行う．

2.2　形態素解析

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 p「未知語」を抜き出す．

2.3　TF/IDF 法

単語の出現頻度に基づいて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行う．TF 法は，一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に重要であるとする．IDF 法は，特定の文書中にしか現れない単語はその文書特徴付けるため重要であるとする．これらを式で表すと

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

この 2 式を用いて語wの文書ｄにおける重要度を表すと

重要度 = tf(w,d)*log(idf(w))

となる．

2.4　メタデータ

電子メールにその意味情報として付加す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 では，リソースのプロパティや，リソース間の関係のみしか定義することができない．リソースのカテゴリーを定義するためには，RDF スキーマを用いる．RDF スキーマは仕様で用意されている基本クラス，基本プロパティを用いて必要なクラスを定義し，RDF のリソースをそのサブクラスとして導くことが可能である．これにより，同じ性質を持つリソースを一つのカテゴリとして定義できる．本提案においては，電子メール管理における最低限のフォルダ概念として RDF スキーマの基本クラスを用い，そのサブクラスに TF/IDF 法で抽出されるようなキーワードを配置した．

３．全体構築設計

3.1 実装方法

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２の処理を行い，メタデータ作成し，電子メール本体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カテゴリに分類したり，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．メールの分類には RDF スキーマを用いる．電子メールから抽出された重要単語が複数の RDF スキーマの基本クラスに分類される場合はその分布の割合を管理することで，フォルダによる管理では実現できないフォルダ同士の重なりの表現や，階層化の防止を実現する．

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．検索時の検索キーワードを RDF スキーマで拡張し，その基本クラス以下の語句を検索キーワードに関連するものとして検索に活用する．また，検索により選び出されたメールと抽出された重要単語が類似しているメールや，類似した複数の RDF スキーマの基本クラスを持つメールを検索結果の候補として提示する．これにより，従来のパターンマッチングによる検索では実現できない同義語や類似語を考慮した検索が可能となる．これを図１に示す．

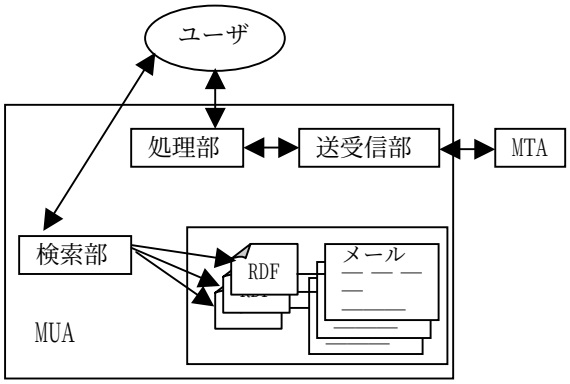

図１　実装概略図

3.2 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，150 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

４．検証

本提案での重要単語抽出方法についての評価検証を行った．評価手順として，まず 1000 通（約７MB）ほどのメールをもとに，各単語の出現するメール数を求めこれを IDF 値算出のための頻度表とした．これら 1000 通のメールの中より 150 通のメールを取り出し，評価者にカテゴリー分けをしてもらった．各カテゴリーに当てはまる語句の集合を作りこれを RDF スキーマの代用とした．今回，この語句の集合は人手で作成した．

次に，150 通の各メールに対し形態素解析と TF/IDF 法を施し，重要名詞と重要名詞・未知語の集合を得た．これら語句の集合をもとに RDF スキーマを用いてメールを分類し，先の評価者の分類との比較を行った．これを表１に示す．

表１　各カテゴリーの抽出単語正解率

カテゴリ	名詞のみ	名詞・未知語
特定話題１	56.8%	70.5%
特定話題２	61.7%	73.3%
添付ファイル	75.0%	75.0%
広告メール	0.0%	0.0%
一過性メール	38.7%	12.9%

これより，特殊な話題を扱うメールからの単語抽出については，ある程度的確性のある結果が得られた．また，名詞のみではなく，形態素解析時に一般的な辞書に含まれない「未知語」がメールを特徴付ける語句になっていることもわかる．

一方，広告メールや一過性のメールといったものは本提案手法では判断が難しいといえる．

５．まとめ

今回は，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を行った．今後は今回の手法を拡張し，グループユースでの電子メールのデータベース化について実装と評価･検証を行う．また，次のステップとしてセマンティック Web 技術でのエージェントの開発とそれを用いての自動処理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