# 送受信メールのサブジェクトからのメール分類階層構造自動生成方式の提案

一色　友宏†　　本橋　洋介†　　坂上　秀和†　　内藤　圭三†

ＮＥＣサービスプラットフォーム研究所†

## 1.　はじめに

業務で PC を使用するユーザは、大量のメールやファイルに囲まれて生活しており、それを整理しておく必要がある。ユーザは手動でフォルダを階層的に生成し、そこにメールを分類したりしているが、それらを自動的に行う仕組みがあると業務の効率が上がると考えられる。

メールを自動的に分類する方式として、ベイズ理論[1]や TF/IDF 法[2]を利用したものがある。これらにおいてはサンプルデータの事前準備や常時メンテナンスが必要になる。[1]ではユーザが手動でフォルダを生成し、そこにメールを分類したものをサンプルデータとするため、ユーザの負担が大きい。[2]ではサンプルデータとして、ユーザ毎に、関連する 1000 通から 2000 通のメールを用意する必要がある。そのため、新入社員などのように、まだサンプルデータがない場合は、サンプルデータが大量にたまるまで自動的に分類をすることができない。また階層的な分類はユーザが行う必要がある。

本提案では階層的な分類構造生成、及び、そこへのメールの分類を、サンプルデータなしに、全て自動的に行う方式について提案する。

## 2.　提案方式概要

メールを管理する方法の 1 つとして、プロジェクト名のフォルダを生成し、さらにそのフォルダ内にプロジェクトに関するモジュール名や“進捗報告”などの作業内容を表すサブフォルダを生成し、メールを分類することがあると考えられる。本方式では、このような階層構造を生成し、プロジェクト(案件や、タスク)に関連するメールを分類することを目的とした。

まず、業務中に送受信したメール約 1000 通を事前に目視で調査した結果、以下の傾向があった。

①. 業務関連するメールの 80%以上には、サブジェクトにユーザに関連するプロジェクト名や案件名を示す単語が含まれていた

②. サブジェクトには①に加え、プロジェクトや案件に関係するモジュール名、サブ項目、作業内容を示す単語が含まれていた

③. サブジェクトには“本日の XYZ 開発検討会議”などのように、名詞を連結した複合語が多く含まれそれらの複合語中に①、②の単語が含まれているものがあった

④. 同じプロジェクト内では同じメンバ間でメールがやり取りされていた

そこで、我々は①、②、③、④ の特徴に着目し、以下のような方針で、メールの自動分類方式を設計することにした。まず、サブジェクトに含まれる複合語から、プロジェクト名などのキーワードを自動的に抽出し、そのキーワードを設定したフォルダを生成し、サブジェクトにキーワードを含むメールを自動的にそのフォルダに分類する。次に、そのフォルダ内のメールからモジュール名や作業項目などをキーワードとして抽出し、サブフォルダを生成し、フォルダ内のメールを再分類する。さらに、第三階層目以降も同様に再帰的にサブフォルダを生成し、メールを分類する。

## 3.　提案方式詳細

本方式の処理手順を以下に示す。

I.　送受信したメールのサブジェクトからキーワード候補を抽出する

II.　品詞を利用してキーワード候補を絞り込む

III.　キーワード候補をキーワードとするかどうかを判定する

IV.　フォルダを生成し、サブジェクトにキーワードを含むメールを自動的に分類する

V.　フォルダ内でI～IVを再起的に行うことで、フォルダの階層構造を生成し、メールを分類する

以下、各手順の詳細について述べる。

### I　キーワード候補の抽出

サブジェクトを形態素解析で単語に分割し、その単語をキーワード候補とし、また、個々の単語だけでなく、複合語もキーワード候補とする。複合語の生成方式は、まず、助詞やスペース、句読点などの区切りを表す記号を抽出し、その前後で単語をグループ化する。助詞やスペース、句読点などは除去する。次にグループ内で連続する単語を連結して複合語を生成する。例えば、“本日の XYZ 開発会議”では、“XYZ 開発”、“XYZ 開発会議”、“開発会議”がキーワード候補となる。さらに、グループに単語が 1 つしかないものは除外する。前述の例では“本日”を除外する。これは2. の③の特徴からプロジェクト名は複合語の中に含まれることが多いため、複合語にならない単語はプロジェクト名ではないとみなしたためである。

### II　キーワード候補の絞込み

Iで抽出したキーワード候補には、プロジェクト名やモジュール名に該当しない一般語が多く含まれている。一般語は複数のプロジェクトや、業務に関係のないダイレクトメールなどに含まれていて、キーワード化すると関連のないメールを集約してしまうため、除外する必要がある。本方式では、まず品詞を利用して、これを除外する。具体的には、名詞以外の単語と、名詞の中でも、“検討”、“報告”、“連絡”などのサ変接続や数詞などの単語は、キーワードである可能性がなく一般語とみなせるため除外する。また、“検討依頼”のように除外対象の単語のみからなる複合語も同様に一般語とみなせるため除外する。ただし、品詞によるフィルタでは“情報”などの普通名詞の一般語は除外できないため、まだキーワード候補として残っている。

### III　キーワード候補をキーワードとするかの判定

サブジェクトに同じキーワード候補を含むメールを一定の閾値数送受信したら、そのキーワード候補をキーワード化する。閾値は予め設定しておく。しかし、そのままではIIで除去できなかった一般語のキーワード候補も、キーワード化される可能性がある。そこで、キーワード化する直前に、2. の④の特徴を利用し、一般語であるかどうかの判定を行うことで、IIで除外できなかった一般語を除外する。To、CC、Fromに含まれるアドレスのうち、受信者自身のアドレスを除き、もっとも多くのメールに含まれるアドレスを抽出する。そのアドレスをTo、CC、Fromのいずれかに含むメール数の割合を計算する。割合が予め設定していた閾値以上であればキーワード化し、閾値以下であれば、除外する。

### IV　フォルダの生成

IIIで抽出されたキーワードを名前とするフォルダを生成する。当該キーワードをサブジェクトに含むメールは、そ

---

Method for classifying mail automatically using subject

†Tomohiro Isshiki, Yousuke Motohashi, Hidekazu Sakagami, Keizo Naito (NEC Service Platforms Research Labs.)

のフォルダに分類する。

## V　階層構造の生成

IVで生成したフォルダ内で再帰的にI～IVの処理を実行し、階層構造を生成していく。

## 4. 試作による結果と考察

本方式に基づくメール自動分類システムを試作し、4人の被験者のメールをサンプルにしてメールの自動分類を行い、想定する以下の効果について評価をした。

(1) キーワードを含む複合語をサブジェクトに持つメールにおいては、正確にキーワードが抽出され、分類される。
(2) 関連性のないメール同士を集約する一般語はキーワードとして抽出されない
(3) 階層構造の第一階層のキーワードにはプロジェクト名が抽出される
(4) 第二階層以降にはプロジェクト名＋モジュール名や、プロジェクト名＋作業内容が抽出される。

分類率、階層別キーワード数、不適切なキーワード数についてまとめたものを図 1 に示す。メールの階層構造への分類率は 38%～82%であった。次に、不適切なキーワードがどのくらいあるかを調べた。被験者 A が他の被験者の了解を得た上で、4 人の結果について、以下のキーワードを主観的に目視で調べカウントした。

(a). 複数のプロジェクトのメールを集約している
　➢"報告"や"連絡"など
(b). 一見して何のメールが分類さているかわからない
　➢"実行される問題"など
(c). 他に抽出されるべき適切なキーワードがあったのに、抽出されてしまった
　➢"ABC モジュールの進捗状況"というメールからキーワードとして"ABC"や"ABC モジュール"ではなく"進捗状況"が抽出されてしまっているもの

(a)のようなキーワードは、全ての被験者において存在せず、(2)の効果が確認できた。(b)、(c)のようなキーワードは存在したため、その原因について調査した。どちらも助詞の扱いが原因であった。2はサブジェクトが"不正にAの処理が実行される問題"など助詞の多い場合で、"不正"、"A"、"処理"が3. のIでの処理により除外されてしまったためである。3も"ABCモジュール"と、"進捗状況"が助詞で分離され、この 2 つが同時にキーワード化の閾値に到達し、システムがランダムに進捗状況をキーワードとして選択したためである。助詞の扱いについては改善の必要があることがわかった。

さらに、一番分類率の低い被験者 A についてより詳細に調査をしたところ以下のことが判明した。

- サブジェクトに複合語があり、そのなかにプロジェクト名などのキーワードが含まれるメールにおいては、95%以上分類されていた
- 複数のプロジェクトに関連するメールが混ざって分類されているフォルダや、プロジェクトに関連するメールと、ダイレクトメール等の業務に関連しないメールが混ざって分類されているフォルダは存在しなかった
- 全メールの分類率は 38%であったが、サブジェクトに被験者 A のプロジェクト名などを含むメールに対する分類率は 90%以上であった

| | | A | B | C | D |
|---|---|---|---|---|---|
| 全メール数 | | 3402 | 3154 | 1800 | 3476 |
| 分類率(%) | | 38 | 43 | 82 | 50 |
| 階層別キーワード数（不適切なキーワード） | 1階層目 | 14(2) | 18(0) | 6(0) | 11(1) |
| | 2階層目 | 22(0) | 12(0) | 17(0) | 22(7) |
| | 3階層目 | 4(0) | 7(0) | 9(1) | 2(1) |

**図 1:試作による分類結果**

さらに、図2に被験者 A の第一階層で生成されたキーワードの例を示す。プロジェクトや所属グループのメーリングリストのPrefix、所属部署名を含むキーワード、研究のコードネーム名など、被験者 A の業務に関係するキーワードが抽出され、かつ一般語はキーワードとなっていないことが確認できた。以上の被験者 A についての詳細な調査により、(1)の効果を確認でき、また、(2)の効果についても再確認できた。

| |
|---|
| "project-●●"(プロジェクトのメーリングリストPrefix) |
| "△△ML"(グループのメーリングリストPrefix) |
| "○○テーマ検討会"(○○は被験者Aの所属部署名) |
| "社内ネットワークサービス"（ネットワーク管理者向けメール) |
| "□□□□"(被験者Aの研究のコードネーム) |
| "日経Trandynet"（ダイレクトメールのPrefix） |
| "◎◎"(◎◎は被験者Aの所属グループ名) |

**図2:被験者 A の第一階層に抽出されたキーワード例**

次に 10%程度あった、キーワードを含むが、分類されなかったものについて調査したところ、以下のものが抽出されていなかったことが判明した。

- 英文のサブジェクトに含まれるキーワード
- 日本文でも単語間をスペース(空白文字)で区切って書かれたサブジェクトに含まれるキーワード
　➢ "XYZ プロジェクト_開発会議"("_"はスペースを表す)

原因は、共に単語間がスペースで区切られているので、3. のIでの、複合語にならない単語は除外する、という処理で除外されてしまったためである。このことから、スペースの前後の単語は、品詞によっては連結する、単独単語でも固有名詞や未知語などプロジェクト名になりそうなものは除外しないなどの処理を追加する必要があることがわかった。

次に図3に生成された分類階層構造の一部について示す。第一階層にコードネーム名が生成されており、第二階層にコードネーム名＋"エンジン"、"ビューア"などのモジュール名やサブコードネーム、第三階層に"進捗報告"、"開催通知"などモジュールに関する作業項目名が生成されており、(3)、(4)の効果が確認できた。

| 一階層目 | 二階層目 | 三階層目 |
|---|---|---|
| ◎◎◎ | | |
| (◎◎◎は被験者の研究のコードネーム) | ◎◎◎プロトタイプ追加機能 | |
| | ◎◎◎プロトUI変更依頼 | |
| | ×××（×××は◎◎◎のサブコードネーム） | ×××試作開発進捗報告 |
| | | ×××打ち合わせ |
| | | 会議開催通知 |
| | ◎◎◎エンジン | |
| | | ◎◎◎エンジン環境構築手順 |
| | ◎◎◎ビューア | |
| | | ◎◎◎ビューア仕様詳細 |
| | | ◎◎◎ビューア変更内容 |

**図3:被験者 A の分類階層構造の一部**

## 5. まとめ

メールの自動分類方式を提案し、実際に業務で使用しているメールをサンプルデータとして評価した。主観的な評価ではあるが、プロジェクト名などのキーワードが含まれる複合語をサブジェクトに持つメールは、想定通りに、業務関係するプロジェクト名やモジュール名がほぼ 100%キーワードとして抽出されており、期待される効果を確認できた。さらに、階層構造も、想定していたとおりに第一階層のキーワードにはプロジェクト名が、第二階層以降にはプロジェクト名＋モジュール名や、プロジェクト名＋作業内容が抽出されていて期待される効果を確認できた。

課題として、複合語を含まない、英語、スペース、助詞を含むサブジェクトについては、正しく分類されないことがあり、改善が必要であることがわかった。また、サブジェクトにプロジェクト名などのキーワード候補を含まないメールなど、本方式では分類できないメールを多く受信するユーザの全メールに対する分類率を上げるには、他の方式との併用を検討する必要がある。

今後は、課題を解決し、この方式をメーラ等に組み込んで実業務で利用できるようにすることを目指したいと考えている。

**参考文献**

[1] "POPFile" http://popfile.sourceforge.net/

[2] 喜名眞魚,片岡信弘 "RDFを用いた電子メール管理の提案と検証" 情報処理学会研究報告, 2006-GN-59-(7),2006