

M-037

添付ファイル付きメーリングリストの解析 Analysis of mailing list with attached file

岡田 陽平[†]
yohei okada

片岡 亨[†]
toru kataoka

芳賀 博英[†]
hirohide haga

金田 重郎[†]
sigeo kaneda

1. はじめに

従来のメーリングリスト (ML) に関する研究では、メール本文やメールのヘッダ情報を用いて解析している研究が主なものであった。[1] しかし、添付ファイルを含む ML について解析している研究は少ない。

我々は、添付ファイルを含む ML では、最終的な決定事項など重要な事柄は、メール本文ではなく添付ファイルでやりとりされているのではないかと考え、「メール本文のみを解析するのではなく、添付ファイルも解析しなければならない」と仮説を立てた。

本稿では、上記の仮説を検証するため、添付ファイルが含まれる ML において、メール本文に出現する単語と添付ファイル内に出現する単語では、どのような特性の違いがあるかの解析の結果について報告する。

2. 単語の抽出

今回使用した ML は、XML をベースにした新しいニュース配信フォーマットである NewsML 関連の ML である。メールの総数は、4821 通であった。

単語の比較にあたって、メールの中から単語を抽出しなければならない。今回は、メールの本文だけに着目しているため、メール内のヘッダ情報、署名、返信文は対象外にし、メール本文だけを抽出した。図1は、MLでの、メールの内容の例である。そして、対象部分は網掛け部分となる。

添付ファイルについては、文書データを解析することが目的であるので Excel のファイルなどは解析する対象から排除し、Word やテキストファイルになっている添付ファイルのみ対象とした。抽出した文を、形態素解析ツール「茶釜」を用いて、形態素解析を行う。その後、形態素解析をした後の情報から単語を抽出する。しかし、専門用語は茶釜の辞書に登録されていない可能性が高い。そこで、抽出する単語は名詞と解析された単語だけではなく未知語として解析された単語も抽出することとした。抽出された単語の総数は、メール本文では 12376 語、添付ファイルでは 4579 語であった。

抽出した単語の中には一般的な単語や、その単語だけでは意味を持たないものなども含まれる。一般的な単語であると、分野に関係なく多くのメールに出現するので、メール本文と添付ファイルの比較して特徴を見つけるといった目的は達成されない。そこで、抽出した単語の中から一般的な単語などを取り除いた。取り除いた後の単語の総数は、メール本文では 5326 語、添付ファイルは 2036 語であった。

3. 単語の比較方法

抽出した単語を、以下の三種類に分類した。

[†]同志社大学大学院工学研究科

```
Date: .....
From: .....
Subject: Re: .....
〇〇の××です。皆様、昨日はお疲れ様でした。
□□となりますがよろしくお願いします。
>〇〇××です。うちの子は最近何かを頼むと
=====
〇社システム局システム部 〇〇〇〇
Tel:012-345-6789 Fax:012-345-6789
```

図1: 電子メールの例

- (1) メール本文と添付ファイルに共起する単語
- (2) メール本文のみに出現する単語
- (3) 添付ファイルのみに出現する単語

また、(1) メール本文と添付ファイルに共起する単語については、以下の三通りを抽出した。そして、抽出した単語にどのような特徴があるのかを調査した。

- (a) メール本文と添付ファイルの両方で頻度が高い単語
- (b) メール本文の方が添付ファイルよりも顕著に頻度が高い単語
- (c) 添付ファイルの方がメール本文よりも顕著に頻度が高い単語

メール本文と添付ファイルの単語の総数は異なる。そこで、抽出した単語の数をメール本文と添付ファイルそれぞれの単語の総数で割った相対頻度で比較することとした。

4. 比較結果と評価

4.1 比較結果

- (1) メール本文と添付ファイルに共起する単語

メール本文と添付ファイルの共に頻度が高い単語のリストの上位を表1に示す。

表1: 両方で頻度の高い単語

共起する単語	相対頻度の合計
NEWSML	3425.73
IPTC	1389.91
TOPICSET	1202.45
FORMALNAME	1096.59
XML	725.95

メール本文の方が添付ファイルよりも顕著に頻度が高い単語とは、メール本文に出現する単語の相対

頻度が、添付ファイルに出現する単語の相対頻度の5倍以上になっている単語である。メール本文の方が添付ファイルよりも顕著に頻度が高い単語のリストの上位を表2に示す。

表2: メール本文の頻度が高い単語

共起する単語	本文の単語	添付の単語
メール	542.20	88.14
RADIOTV-NEWSML	151.66	22.92
RADIOTV	93.22	3.53
TOPICUID	42.83	5.53
MLIST	26.72	3.53

添付ファイルの相対頻度が添付ファイルの相対頻度の5倍以上になっている単語のリストの上位を表3に示す。

表3: 添付ファイルの頻度が高い単語

共起する単語	本文の単語	添付の単語
フォーマット	89.70	532.35
TIFF	17.64	222.16
VOCABULARY	23.20	188.61
LOCAID	2.02	146.31
SERVICEID	6.05	118.10

表1, 表2, 表3から、両方で頻度の高い単語のリストに出現する単語は、メール本文または添付ファイルが頻度の高い単語のリストには、ほとんど現れない。メール本文で頻度が高い単語は、添付ファイルでも頻度が高いことがわかる。

また、メール本文の頻度が高い単語は19語、添付ファイルの頻度が高い単語が109語と少数しか抽出されなかった。

この結果から、メール本文と添付ファイルに共起する単語での出現パターンは似通っていると言える。

(2) メール本文のみに出現する単語

メール本文のみに出現する単語のリストの上位を表4に示す。

表4: メール本文のみに出現する単語

本文のみの単語	出現回数
パワーアップ	180
URBY	134
エネルギー	75
EGROUPS	72
TVPROGRAM	68

表4は出現回数が多い単語のみをリストアップしている。実際は、多くの種類の単語がリストアップされた。

しかし、リストアップした単語は、ほとんどが1度や2度しか現れていない単語ばかりであった。このことから、メール本文から抽出された単語の大部分は添付ファイルに出現する単語と共起していることが確認できる。

(3) 添付ファイルのみに出現する単語

添付ファイルのみに出現する単語のリストの上位を表5に示す。

表5: 添付ファイルのみに出現する単語

共起する単語	相対頻度の単語
DESCRIPTION	122
ELEMENTS	34
ラベル	19
SUBELEMENT	12
多値	10

添付ファイルのみに出現する単語もメール本文のみの単語と同様、ほとんどが1度2度しか現れていない単語ばかりであった。このことから、添付ファイルから抽出された単語の大部分がメール本文に出現する単語と共起しているとわかる。

4.2 評価

(1) の比較結果より、メール本文と添付ファイルに共起する単語での出現パターンは似通っていると確認できた。(2), (3) の結果より、メール本文、添付ファイル共に、大部分は共起しており、どちらかのみに出現する単語はあまり意味をなさない単語であると確認できた。

これらのことから、メール本文と添付ファイルの単語の出現パターンが似通っているという結果を得た。

我々は、「メール本文のみを解析するのではなく、添付ファイルも解析しなければならない」という仮説を立てた。しかし、今回行った、メール本文と添付ファイルに出現する単語の比較により、メール本文も添付ファイルも単語の出現パターンは変わらないという結果がでた。

単語の出現パターンが変わらないので、特に添付ファイルを解析をしなくてもメール本文を解析すれば十分である。このことから、我々の仮説は間違っていたこととなった。しかし、我々は、MLを解析していく上で添付ファイルは解析しなくても大丈夫であるという情報を得た。

5. おわりに

本稿では、添付ファイルを含むMLにおいて、添付ファイルの情報の重要性について検証した。

今回の調査結果によって、MLの解析において、添付ファイルが含まれていても従来どおりメール本文を解析すればよいという結果を得られた。

我々が、立てた仮説が間違っていたという結果になった。しかし、この結果は、ファイルを添付してメールを送ることが多くなった今日において、これからのMLについての研究において重要な情報を得ることができた。

参考文献

- [1] 上田宏高, 柳沢豊, 塚本昌彦, 西尾章治郎, 「電子メールの傾向分析への知識獲得手法の適用」『情報処理学会論文誌』(Vol41, No.12, Dec. 2000)
- [2] 長尾真 著, 『自然言語処理』(岩波書店, 1996)