共同研究報告

学生のオンライン行動による社会力の評価

Evaluation of social skills of students through their online activities

共同研究メンバー
○出原至道*、久保田貴文*（○代表、執筆者）

Keywords：Social Skills

1. 背景と目的

本研究の目的は、学生の送受信したメールの送信先・送信元から構成されるネットワークを分析することにより学生の社会力の評価指標を導出することである。

学生の社会性は、対話的で実践的な教育を通じて向上していくと期待できるが、これを客観的・定量的に評価することは困難である。この評価に利用可能な指標が定義できれば、さまざまな手法の比較や、学生個人の成長段階の判定に利用可能になると期待できる。

この目的で、本研究は、学生のメール送受信履歴に着目する。

現在､チャットサービスやメッセージングサービスが一般に提供されるようになっているが､それでもなお、社会人の間では電子メール（メール）によるコミュニケーションが主流として利用されている。これは、メールが RFC によるオープンな議論に基づいた仕様となっており、信頼性・継続性・検証可能性の点で、クローズドな他の通信手段に比べて優れているためである。一方、大学生の多くは、きちんとしたメール送受信ソフトによるコミュニケーションをほとんど経験しないで入学するようになっている。

このことから、学生が社会人として成長していく過程で、メールの送受信履歴に変化があることが期待できると考えた。

一方で、メールは、極めて個人的な情報を含む。このため、本研究では、通信の相手方・通信内容を一切参照することなく、メール間の関連性と送信時刻だけを参照する。

今年度の研究成果は、(1) メールの自動受信・データベース化システム、(2) メールを返信するまでの時間の算出、(3) メールツリーの深度の算出である。

2. 手法

本研究では、メール本文ではなく、メールヘッダに含まれる情報だけを利用して、メールの関連性とメール送信行動の追跡を行う。システムの開発には、Python 言語を使用し、データベースには mysql を使用した。

* 多摩大学経営情報学部　School of Management and Information Sciences, Tama University

メールヘッダは、RFC2822[1] によって定義されている（図 1）。このうち、本研究の収集対象は、メール固有の ID（Message-ID）、返信先メール ID（In-Reply-To）、送信時刻（Date）、受信者と送信者のトップレベルドメイン（From/To の最後のドメインのみ）として、プライバシーに配慮した。

さらに、メッセージ ID に所属団体名が入ることが多いため、このデータを匿名化したい。この目的で、Message-ID, In-Reply-To については、sha256 によるハッシュ化を行い、保存される文字列からもとのデータが復元できないように配慮した。ハッシュ関数の性質から、この状態でもメールの相互間の関連性を復元することができる。

```
MIME-Version: 1.0
Date: Thu, 7 Mar 2019 22:17:48 +0100
References: <CAHxd11UVf33uL-xyaV1M8jRsf9871iK10Rs=zETaZ6p1E-????@mail.gmail.com>
In-Reply-To: <CAHxd11UVf33uL-xyaV1M8jRsf9871iK10Rs=zETaZ6p1E-????@mail.gmail.com>
Message-ID: <CADaiGvRSOYNYDCFvpiR_ERHBhVANTbY9tJOf5LW-gq*********@mail.gmail.com>
Subject: 問い合わせ
From: <idehara@tama.ac.jp>
To:  <someone@example.com>
Content-Type: multipart/alternative; boundary="0000000000005d34d1058387a3dc"
```

図 1：メールヘッダの例

本大学のメールシステムは Google 社の gmail を採用している。このため、当初は、各学生にアプリケーションを配布し、匿名化された情報だけをデータベースサーバに送信することを目標としてアプリケーション開発を開始した（図 2）。これによって、実際にメールヘッダを受信してから匿名化するまでの処理を各個人のデバイスの中で完結させ、分析システム側からは匿名化前の情報を全く参照しないことが可能となる。

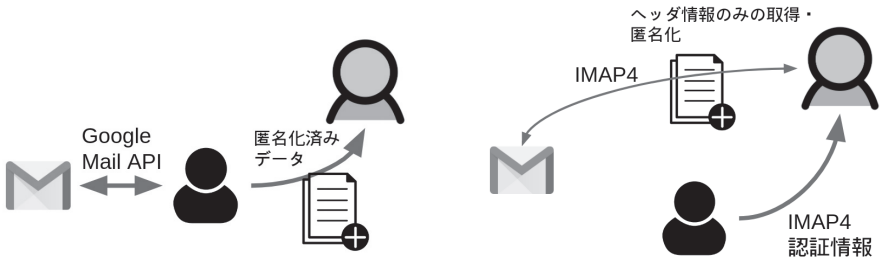

図 2：システム設計（左：Google API　右：IMAP4）

しかし、Google のプライバシーポリシーの厳格化をうけて、このような動作をするアプリケーションに対して API の利用認証許可の取得が困難となった。このため、ユーザから認証情報の一時的な提供を受けて、分析サーバから直接メールサーバにアクセスすることでメール情報を取得するようにシステムを変更した。この際、POP による接続では、gmail の仕様によって取得可能なメールが約 500 件に限定されることが判明したため、IMAP4 方式による接続を利用した（図 2）。

gmail に IMAP4 で接続する場合、SSL 接続が必須である。この接続方法は、通信経路における傍受に対して耐性があり、セキュリティ面からもこの方式が好ましい。また、IMAP4 では、接続先に要求する情報を限定することができる。今回の実装では、取得要求を「ヘッダ情

報」に限定することで、メール本文を取得してしまうことを回避した（図 3）。

```
with IMAP4_SSL('imap.gmail.com', ssl_context=ssl.create_default_context()) as imap:
        imap.login(myAddress,myPassword)
        imap.select()

        typ, data = imap.search(None, 'ALL')
        for num in data[0].split():
                typ, data = imap.fetch(num, '(RFC822.HEADER)')
```

図 3：IMAP4 によるヘッダのみの取得

メールヘッダのうち、メールの送信時刻を表す Date については、RFC に準拠しない記述が古いメールに散見された。これを全て正しく時刻型に変換することは困難であったため、email.utils ライブラリの parsedate_tz 関数 [2] で変換できない時刻を持つメールについては、分析の対象外とした。

3. 指標の検討

匿名化されたメールの送受信データから算出可能な指標として（1）メールを返信するまでの所要時間、（2）返信したメールのツリー内での深度をとりあげた。システムが実験段階であることから、実際の学生のデータを取り扱うことは避け、教員のデータを対象として指標の算出実験をおこない、有効性の推定を行った。

対象としたメールは、共同研究責任者である出原（idehara@tama.ac.jp）の 2010 年以降の送受信メール 125,428 通（うち、送信メール 3,536 通）である。なお、メールが削除された場合、さかのぼってメールの関連性を追跡することは困難であるが、gmail の特性上、ユーザは「メールの削除」を行わないと期待できる。

メールを返信するまでの所要時間は、In-Reply-To (re フィールド) に値を持つレコードについて、その値の Message-ID (mid フィールド）を持つレコードとの送信時刻 (timestamp) の差によって表される。月ごとの所要時間の平均値を算出することで、返信までの所要時間の指標とした（図 4）。

```
select
       year(r2.timestamp),month(r2.timestamp)
       ,avg(TIMESTAMPDIFF(hour,r1.timestamp,r2.timestamp))
from rels as r1
       inner join rels as r2 on (r1.mid = r2.re)
where r2.issent
group by year(r2.timestamp),month(r2.timestamp);
```

図 4：平均返信所要時間の算出

このデータは、社会性がほぼ変化していないと考えられる教員を対象としているため、指標に大きな変化は現れないと期待した。しかし、長期出張の入る時期（3 ～ 4 月、9 ～ 10 月）を中心に、返信所要時間に変化が観察できる（図 6）。これは「急な返信を必要としないメールに対して、返信を出張後に伸ばした」「ネットワーク環境が悪く、返信ができなかった」などの理由が考えられる。メール受信者の時間的な余裕によって、数値が揺らぐ可能性があるため、

指標として採用するには検討が必要である。また、2015 年 5 月の異常値は、1 年前のメールへの返信の形で新たな情報を提供したことによるもので、このような異常値を排除するアルゴリズムが必要である。

次に、送信したメールのツリー内における深度について算出した。各レコードについて、深度（level）を 0 で初期化した後、返信メール（re フィールドが空でないメール）について、元メールの深度＋ 1 で level を更新する（図 5）。これを、level の更新が起きなくなるまで繰り返した。再帰的に level を更新することも検討したが、データ件数が多いことから、計算負荷が大きくなることが予想され、今回は原始的な手法によった。

```
update rels as r2
       inner join rels as r1 on r1.mid = r2.re
       set r2.level=r1.level+1;
```

図 5：メールツリー深度の算出

メールツリー深度は、送受信が 1 往復して終了するようなメールではなく、繰り返しメールのやり取りを行って議論していることを表していると想定される。このうち、送信メールの最深のツリー深度を図 7 に示す。図 6 の返信までの所要時間に比較すると、安定した指標となっていることが観察できる。

一方で、最深深度だけでは、他の人がやり取りした後に 1 通だけメールを送るような行動と、そのツリーで頻繁にメールをやり取りした行動が同じ評価となる点が課題である。

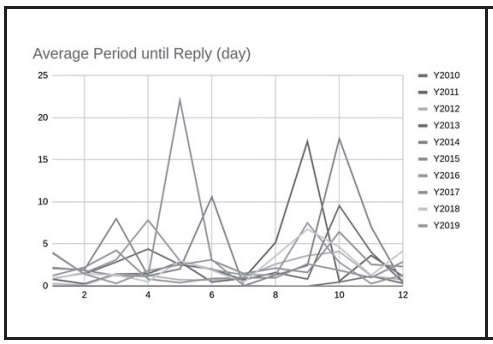

図 6：メール返信に要した平均時間（日）

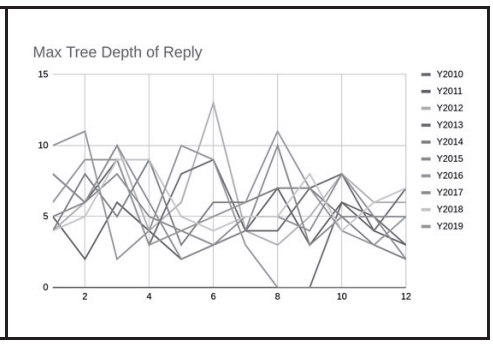

図 7：送信メールの最深ツリー深度

4. 考察と課題

今年度は、匿名化されたメールの送受信履歴データだけを用いて、指標の候補を算出することができた。メール返信所要時間は、送信者の状況によって変動が大きいため、指標としては有効でないことが示唆された。また、メールツリーの深度については、単に最深深度を採るのではなく、ツリーの中での議論への貢献度を反映する指標を検討する必要がある。

今回、教員のデータのみを使用して指標の検討を行ったが、今後、学生本人の許諾を得た上で、学生のデータに基づいた分析を行っていく。

参考文献

[1] https://www.ietf.org/rfc/rfc2822.txt

[2] https://docs.python.org/3/library/email.utils.html