情報抽出とユーザの行動履歴に基づく電子メールのランキング*

5Q－4

長谷川 隆明†　高木 伸一郎
NTT 情報通信研究所‡
E-mail: hasegawa@isl.ntt.co.jp

1　はじめに

インターネットの普及により、電子メールをコミュニケーションの手段として用いる人が増えている。電子メールの持つ同報性のため、一度に多くの人に情報を伝達できるので送信側にとっては都合が良いが、そのために受信する電子メールは多量になり、どの電子メールが本当に重要なのかがわからなくなるという点で受信側にとっては問題になっている。現状では、メールヘッダに記述されている差出人やタイトルに含まれる文字列によって電子メールをフィルタリングしたり、ヘッダに明示的に優先度を記述することによって電子メールに優先順序を与えたりする方法が用いられている。しかし、これらの方法では、メールヘッダの情報しか利用できず、送信側が設定した優先度がそのまま受信側に反映されてしまうので、一人一人異なる個人情報を持ったユーザ指向の電子メールの優先度を表すことができない。

本稿では、イベントの通知と返信の依頼を含む電子メールからイベントのカテゴリと差出人および返信の期限の情報を抽出し、ユーザの電子メールに対する行動履歴から得たイベントのカテゴリへの関心と差出人との関係と、抽出された返信の期限までの残り時間とに基づいて優先度を算出することによって、ユーザ毎に異なる優先度を受信した電子メールに与え、ユーザ指向で電子メールをランキングする方法を提案する。

2　電子メールの優先順位付け

2.1　電子メールの特徴抽出

電子メールに優先順位を付けるために、メールヘッダだけでなく、電子メール本文に記述されている情報を利用するためにメール本文からの情報抽出を行う[2]。抽出する情報は、通知されるイベント名、返信の期限である。電子メールから抽出したイベント名は多岐にわたるため、そのままでは再利用が難しい。そのため、似たようなイベントは一つのカテゴリにまとめた方が再利用しやすい。そこで、イベント名から通知されるイベントのカテゴリを分類する。分類の仕方は、イベント名に含まれているキーワードを用いる。

通知された電子メールがどのくらい緊急を要するのかを判断するためには、メールヘッダだけではなく、電子メールの本文を見なければわからない。電子メールの本文には返信の期限の情報が含まれていることがあるので、これを利用することによって緊急度を算出する方法を提案する。どのくらい緊急かということは、指定された期限までどのくらいの時間が残されているかと置き換えることができる。残された時間が少ししかない通知は緊急度が高く、反対に残された時間が多くある通知は緊急度が低いと考える。このような緊急度を求めるには、残り時間を定量化しなければならない。残り時間を計算する基準となる時刻は、メールヘッダに記述されている Date: フィールドを用いる。すなわち、残り時間は電子メール本文に記述されている期限とメールヘッダの Date: フィールドの差分とする。緊急度は、残り時間に応じて0～1までの間に割り当てられる。

2.2　ユーザの行動履歴からの学習

優先順位が高い電子メールは、ユーザの関心事や対人関係がそれぞれ異なるためにユーザ毎に異なる。電子メールに対するユーザの行動履歴を観察することによって、計算機が自動的にユーザの関心事や対人関係を考慮して電子メールの優先順位を判定できるようにすることが望ましい。我々は、通知される電子メールに対してユーザが返信を行うかどうかを観察し、これをユーザの行動履歴とすることを提案する。電子メールの行動履歴の特徴として、イベントのカテゴリと差出人を考える。

電子メールで通知されるイベントのカテゴリを特徴にすることにより、どのカテゴリの電子メールにユーザは関心を持っているのかを推定することができる。受信した同じカテゴリの電子メールのうち、何通に返信を行ったかという割合（以下、カテゴリ返信率）を求めることで、ユーザのカテゴリに対する関心を定量化する。また、電子メールの差出人を特徴にすることにより、どの差出人からの電子メールを重要だと判断しているかを推定することができる。特定の差出人から受信した電子メールのうち何通に返信を行ったか

*Ranking of E-mails Based on Information Extraction and User's Action History
†Takaaki Hasegawa and Shin'ichiro Takagi
‡NTT Infomation and Communication Systems Labs.

表 1: カテゴリ返信率と差出人返信率

種別	対象	受信数	返信数	返信率
カテゴリ	ゴルフ	18	16	0.89
差出人	horii@ntt.co.jp	50	40	0.8

という割合（以下、差出人返信率）を求めることにより、差出人との対人関係を定量化する。表 1に返信率の様子を示す。

あるカテゴリ i に対するカテゴリ返信率 c_i とある差出人 j に対する差出人返信率 s_j から、カテゴリ i について差出人 j からの電子メールに対するユーザの選好 $t(c_i, s_j)$ を (1) 式により定量化する。

$$t(c_i, s_j) = \frac{w1 * c_i + w2 * s_j}{w1 + w2} \quad (1)$$

ただし、$w1$,$w2$ は重み係数とする。

受信したカテゴリ i と差出人 j の電子メールに対してユーザが返信を行えば、カテゴリ i の返信数と差出人 j の返信数を 1 つずつ増やして、カテゴリ返信率 c_i と差出人返信率 s_j をそれぞれ更新する。これにより、常に新しいユーザの選好を維持することができる。過去にない差出人から電子メールを受信した場合は、新しい差出人を差出人のテーブルに追加更新する。

2.3 優先度の算出

受信した電子メールの本文から抽出したカテゴリ i と差出人 j と緊急度 e を用いて、受信した電子メールの優先度を計算する。過去のカテゴリ i と差出人 j に対する電子メールの選好 $t(c_i, s_j)$ と緊急度 e を用いることにより、ユーザの電子メールに対する優先度 $p(t_{(c_i,s_j)}, e)$ を (2) 式により定量化する。カテゴリが抽出されなかった場合は、カテゴリ返信率は 0 とする。算出した優先度の大きさに応じて、電子メールの優先順序付けを行う。

$$p(t_{(c_i,s_j)}, e) = \frac{w3 * t(c_i, s_j) + w4 * e}{w3 + w4} \quad (2)$$

ただし、$w3$,$w4$ は重み係数とする。

3 議論

本手法が有効であることを検証するためにシミュレーションを行った。受信する電子メールを図 1とし、受信する日付を 1997 年 11 月 12 日であると仮定し、カテゴリ返信率と差出人返信率は表 1とする。簡単化のため、(1)(2) 式の重み係数はすべて 1 とした。

図 1の電子メールから情報抽出を行い、イベント名として「第 2 回小池部長杯ゴルフコンペ」が、差出人として「 horii@ntt.co.jp 」が、返信の期限として「11 月 30 日」が抽出される。イベント名からイベントのカテゴリは「ゴルフ」と判定される。メールヘッダの Date: フィールドに記述されている日付「11 月 12 日」から、残り時間は 18 日で 3 週間以内なので 0.3 となる。表 1よりカテゴリ返信率は 0.89、差出人返信率は 0.8 である。(1) 式により図 1に対するユーザの選好度は 0.84 と算出される。(2) 式により図 1の電子メールの優先度は 0.57 と算出される。

```
Subject: GOLF-COMPE
From: horii@ntt.co.jp (Motoyuki Horii)
To: chishinbu@ntt.co.jp
Date: Wed, 12 Nov 1997 12:30:26 +0900 (JST)

堀井です。
以下の日程で「第２回 知信部 小池部長杯 ゴルフコンペ」を開催することになりました。
    日時：平成９年１２月１９日（金）
    場所：未定（千葉方面を予定、決まり次第別途周知）
多数の方の参加をお願いいたします。参加希望者は私（horii@ntt.co.jp）まで連絡をください。期限は 11 月 30 日までとします。
```

図 1: 電子メールの例

本稿で提案した方法は、発信側が電子メールに優先順位をつける方法や電子メールの特徴を対象として memory based reasoning を用いる方法 [1] に比べて、以下の点において有効であると考えている。(1) 徐々に学習していくユーザの関心事や対人関係というプロファイルを用いることによって、受信側で優先順位を判断することができる。(2) 電子メールの本文中に記述されている指定された期限を優先順位に反映させることができる。(3) 返信率の算出や (1)(2) 式の計算は 1 次の演算であり、学習に必要な時間が要らない。(4) 常に電子メールに対する返信率を更新するので、逐一変化するユーザの関心事や対人関係に追随できる。

4 おわりに

本稿では、受信した電子メールからイベントのカテゴリと差出人および返信の期限とを抽出し、カテゴリと差出人について過去のそれらに関する受信総数と返信数から返信率をそれぞれ算出し、二つの返信率から計算された電子メールに対するユーザの選好と期限から求めた残り時間を定量化した緊急度とから、電子メールの優先度を計算することによって、ユーザ指向で電子メールのランキングを行う方法を提案した。

参考文献

[1] Maes, P. : Agents That Reduce Work and Information Overload, *Communication of the ACM*, Vol. 37, No. 7, 1994, pp. 30–40.

[2] 長谷川 隆明, 高木 伸一郎：電子メールコミュニケーションにおけるスケジュール情報抽出, 情報処理学会自然言語処理研究会 NL123-10, 1998.