履歴情報を考慮したメイル文書のフィルタリング手法

6R－7

獅々堀 正幹　藤井 誠　安藤 一秋　青江 順一

徳島大学工学部知能情報工学科

1．はじめに

近年のインターネットの普及に伴い，電子メイルがコミュニケーションの1つの手段として確立された．しかし，大量のメイル文書が氾濫している現在，重要なメイルと商用メイルのような重要度が低いメイルとを分類する処理，即ち，メイル文書の重要度を判定するフィルタリング処理の実現が望まれている[1]．

メイル文書を対象にした文書処理の研究には，メイル文書やニュース記事からスケジュール情報を自動抽出する手法[2],[3]が報告されている．これらはスケジュール情報の表現パターンをルール化し，そのルールとのパターンマッチングを行うルールベースの手法である．しかしながら，重要度の判断基準は個人毎に異なるため，一意に決められたルールベースの知識を重要度の判定に適用することはできない．

本手法は各個人が受信済みのメイル文書内に重要度を判定するのための履歴情報が含まれていると考え，各個人が優先度付けした既存のメイル文書に基づいたコーパスベースの知識を構築する．本手法を用いることにより，重要度の判断基準の揺れを許容したフィルタリング処理を行うことが可能となる．

2．メイル文書の重要度とは

本手法の内容を説明する前に，まず，重要度とは何か，また，どのような要因から重要度が決定されるのかを検討する必要がある．

重要度とは，他のメイル文書よりも早く読む必要がある，または，早急に返事を出す必要があるメイル文書を重要度が高いと定義する．このような観点から考えると，メイル文書の重要性が個人毎に異なることは明白である．また，重要度を構成する要因としては，a)文書の送信元，b)文の類型，c)文のテーマ，d)時間的制限，e)過去のメイルとの関連性等が考えられる．これらを重要度決定のための属性（以後，単に属性）と呼び，各属性の内容を属性値と呼ぶ．属性値は，a)はヘッダーのFromの項目，b)は各文の助述表現，c)は各文に含まれる名詞・固有名詞，d)は時間表現や副詞，e)はヘッダーのSubjectの項目から得られる．これらの属性値に，個人毎に異なった重み付けがされ，それらが絡み合って重要度が決定されているという前提の元で本手法を提案する．

3．履歴情報を用いたフィルタリング法

図1に本手法の概要図を示す．

既存のメイル文書には，ユーザーによって優先度に関する情報がタグ付けされている．この既存のメイル文書中に含まれる属性値を調べるためのメタな知識として基本知識を用いる．基本知識には各属性値を検出

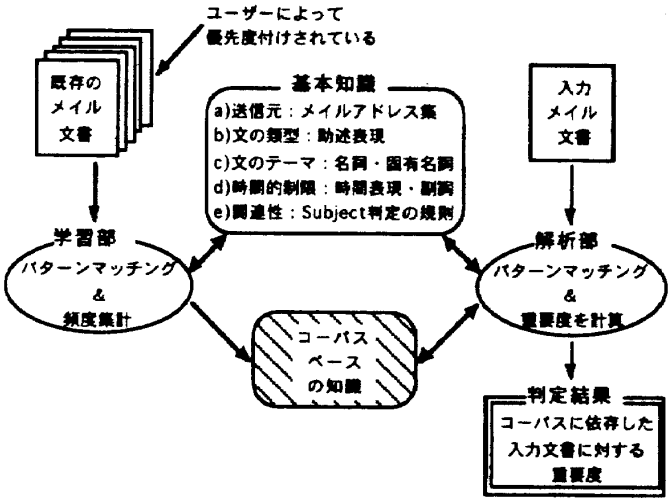

図1　本提案手法の概要図

A Mail Filtering Method Using Personal Histories.

Masami Shishibori, Makoto Fujii, Kazuaki Ando and Jun-ichi Aoe

Dept. of Information Sciences & Intelligent Systems, Faculty of Engineering, Tokushima University

表１　基本知識の例

属性	属性値	表現例
文の類型	勧告	～すること，～するように，～しないように
	依頼	～して下さい，～してくれ，～して欲しい
	条件付依頼	もし～ならば・・・下さい
	勧誘	～しましょう，～しよう，～行こう
	疑問	～でしょうか，～かな？，～ですよね
文のテーマ	ゼミ関係	ゼミ，形態素解析，青江先生，研究室，．．．
	試験関係	大学院試験，入試願書，プレテスト，．．．
	恒例行事	阿波踊り，ソフトボール大会，掃除，．．．

するためのパターンが格納されており，類似したパターンをグルーピングすることで一つの属性値を構成している．この様にグルーピングすることにより，後で説明する頻度集計の際に，集計結果のスパースさを回避することができる．例として，文の類型，文のテーマに関する基本知識の一部を表１に示す．

学習部では，まず，既存のメイル文書毎に，その中に含まれる属性値を基本知識に従って検出し，検出された属性値とそのメイル文書の優先度から成る多重組を構成する．全てのメイル文書の多重組に対して同じ内容の多重組の頻度をとり，以下のような形式のコーパスベースの知識を生成する．

$$(a_j,\ b_k,\ c_l,\ d_m,\ e_n,\ r_i) = \text{頻度};$$

但し，a_j，b_k，c_l，d_m，e_nは属性 a)～e)の各属性値，r_iは優先度を示す．このコーパスベースの知識は，ユーザーがどのような属性値の組み合わせに重要度を置いているかを示している．

解析部では，未知の入力メイル文書内に含まれる属性値を基本知識を用いて検出し，その検出結果とコーパスベースの知識を用いて，入力文書の重要度を確率的に算出する．この重要度は入力文書中に存在する属性値の条件付確率になるので，確率論的に以下の式で求めることができる[4]．

$$P(R_i|a_j, b_k, \ldots, e_n) = \frac{P(a_j, b_k, \ldots, e_n, r_i)}{P(a_j, b_k, \ldots, e_n)} \qquad (1)$$

但し，R_iは入力文書に対する重要度を示し，重要度をnランクに分類した場合，$1 \leq i \leq n$となる．従って重要度は各ランク毎の確率値として求められる．

4．評価

本手法の有効性を確認するため，机上でのシミュレーションを行った．まず，200 通の学習用メイル文書を解析し，属性値数３の属性 a)，属性値数８（表現パターン数 120）の属性 b)，属性値数５（表現パターン数 76）の属性 c)，属性値数２（表現パターン数 16）の属性 d)から成る基本知識を構築した．また，この基本知識を用いて学習用データから 108 個の多重組を有するコーパスベースの知識を構築した．

これらの知識を用いて，学習データとは別に用意した 30 通の評価用メイル文書の重要度を算出した．尚，重要度（優先度も同様）のランク数は５（ランク５が最も重要度が高い）とした．また，式(1)から得られる各ランクの確率値を次式で補正した値（補正重要度）と予め人手によって与えられた重要度を比較した．

$$\text{補正重要度} = \sum_{i=1}^{5} \left\{ i \times P(R_i|a_j, \ldots, d_m) \right\}$$

その結果，47％のメイル文書が0.5 の誤差で同じ重要度を示し，１ランク以内の誤差に収まるメイル文書は80％であった．また，コーパス数の不足により３通が判定不能となったが，今後，コーパス，知識を充実させれば十分に実用性に耐えられると思われる．

5．まとめ

本稿では，コーパスベースの知識を構築し，未知のメイル文書の重要度を確率的に計算する手法を提案した．今後の課題としては，インプリメント後のより深い実験結果の考察は勿論のこと，基本知識の充実化，より効果的な属性項目の発見，また，解析精度を向上させるために格構造解析の導入等を考えている．

参　考　文　献

[1] R. J. Hall, "How to Avoid Unwanted Email," Commun. ACM, Vol. 41, No. 3, pp.88-95 (1998).

[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資料，123-10，pp.73-80 (1998).

[3] 佐藤円，佐藤理史，篠田陽一，“電子ニュースのダイジェスト自動生成”，情報処理学会論文誌，Vol. 36，No. 10，pp.2371-2379 (1995).

[4] 北研二，中村哲，永田昌明，“音声言語処理（コーパスに基づくアプローチ）”，森北出版 (1996).