2006－DBS－140（Ⅱ）（73）
2006／7／14

コンテキストを用いたメールの情報補完

河　重　貴　洋†　大　島　裕　明†　小　山　　聡†
田　島　敬　史†　田　中　克　己†

ある電子メールを閲覧している際に，これまでに受信したメールや，インターネット上の情報など，そのメールの中には含まれない情報がユーザに必要となることがある．これまで，そのようなユーザが必要とする情報は自分で検索する必要があった．そこで本研究では，閲覧中のメール，および，これまでに受信したメールを解析することで，閲覧メールの情報を補完する情報を持ったメールを自動的に検索・提示する手法を提案する．提案手法においては閲覧中のメールを補完するメールを抽出するために，メールデータ中のある日時区間に受信したメールにおけるキーワードの出現数に着目し，話題となる語の特定を行う．そして，同じ話題を持つメールを解析することで，その話題語に対する詳細語の発見を行う．この詳細語に基づいて補完するメールを検索し，ユーザに補完情報の提示を行う．

Complementary Information for Email Messages using Context

Takahiro KAWASHIGE,† Hiroaki OHSHIMA,†
Satoshi OYAMA,† Keishi TAJIMA† and Katsumi TANAKA†

When reading some Email message, we often need some complementary information which is not included in that Email but in other Emails or some Web pages. In such a case, we have to search the mail boxes or Web by ourselves. To help such an activity, we propose a method of automatically retrieving Emails including some information complementary to the given Email. In our approach, we first detect the keywords representing the topic of the given Email based on the term frequencies in the Emails in some time span. By using those keywords, we find Emails that have the same topic, and also find the description keywords of that topic from those Emails. Then, the system automatically retrieves the complementary Emails by using those description keywords.

1. はじめに

インターネットやPCの普及に伴い，Webページに代表される大量の情報にアクセス可能な状況になり，PC上に蓄積される情報の量も増加の一途をたどるのみである．特に，電子メールのやり取りをする機会が多くなり，日々大量の電子メールを閲覧，返信するようになった．ある一通のメールを閲覧している際に，そのメールの情報だけでなく，過去に受信したメールの検索や，検索エンジンを用いてWebページの検索を行うことで，閲覧中のメール以外の情報を求めることが多い．その際，求める情報を得るための検索質問を考える手間がかかる．また，過去のメールを検索しようとしても，大量のメールに埋もれた，ある特定のメールを検索するための検索質問を上手く作ることができず，結局目的のメールを見つけられないことも多々ある．そこで，本研究では閲覧中のメールを補完する内容のメールを自動的にユーザに提示する手法を提案する．本提案手法においては，最初に閲覧中のメールから話題語となる語を抽出する．次に過去に受信したメール群を解析し，同じ話題について書かれたメールを抽出する．そして，それら同じ話題について書かれたメール群からその話題に対する詳細語の抽出を行う．この詳細語を元に，現在閲覧中のメールを補完する内容のメールの発見・提示を行う．本論文で述べる提案手法の流れは以下のようになる．

(1) ユーザがある1メールを選択
(2) そのメールの話題となる語を抽出
(3) その話題語の詳細語をこれまでに受信したメール群の中から抽出
(4) 閲覧メールでの話題語，詳細語と他のメールでの話題語，詳細語をもとに，補完する内容のメールを検索・提示

以下の章で各ステップにおける手法の詳細について記す．

† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

2. 関 連 研 究

2.1 情 報 補 完

馬ら[1)2)3)] は話題構造の抽出に基づく補完情報検索の手法を提案している．文献[1)] ではテレビ番組の字幕データのようなテキストストリームを用い，補完情報検索を行う手法を提案している．馬らの手法では字幕データをセグメントに区切り，各セグメントから話題構造を抽出し，話題構造から構造化質問を作成する．その構造化質問で Web 検索を行い，検索結果を補完度を基に再ランキングを行う．話題構造の抽出では，語の共起関係を基に話題グラフを作成し，この話題グラフを基に検索結果の各ページの補完度を求めている．この手法と本提案手法を比べると，補完情報を求め方が異なっており，本研究では 2.4 章で述べる小山ら[4)] の詳細語発見の手法や，メールの時間的区間に注目して特徴語を発見する手法を用いている．

2.2 流行度・新鮮度

宮崎ら[5)] は Web ページの変更・追加が行われた際の変更ページの価値の尺度として流行度、新鮮度を定義している．新鮮度や流行度の計算においては，追加または変更されたページと比較対象のページとの内容的な類似度や時間的な距離に基づいた計算を行っている．時間的な要素を重みの計算に用いている点では本提案手法における特徴語の抽出方法と同じであるが，本提案手法ではメールの件名といった情報を用いている点がこの新鮮度，流行度の計算方法と異なっている．

2.3 メール閲覧時の情報提示

メール閲覧時に関連情報を提示する研究としては Mitchell[6)] の研究がある．この研究では電子メールのデータを解析し，そのメールのプロジェクトに関係のある人や，ファイルといったものを抽出する．この情報を元にして，メール閲覧時に関連情報の提示を行っている．我々の提案手法ではメール閲覧時に情報を提示する点は同じであるが，閲覧中のメールを内容的に補完する情報をユーザに提示することを目的としている点でこの研究とは異なっている．

2.4 詳 細 語

小山ら[4)] は検索エンジンの各ページの要約文や，検索ヒット件数といった情報を元に，あるキーワードが与えられた際に，そのキーワードの詳細語を発見する手法を提案している．あるキーワード A に対してキーワード B が詳細語であるかどうかの判定を行う際，A がタイトルに含まれる際に，そのページの本文中に B が出現する確率 $p(B|intitle(A))$ を (1) 式で求める．ただし，A が Web ページのタイトルに含まれることを $intitle(A)$ とし，それを満たすページの数を $DF(intitle(A))$ とする．

$$p(B|intitle(A)) = \frac{DF(intitle(A) \bigwedge B)}{DF(intitle(A))} \quad (1)$$

次に A と本文中に出現する場合に，B も本文中に出現する確率 $p(B|A)$ を (2) 式で求める．

$$p(B|A) = \frac{DF(A \bigwedge B)}{DF(A)} \quad (2)$$

そして，$p(B|intitle(A))$ が $p(B|A)$ よりも有意に大きくなっている場合に，キーワード B がキーワード A の詳細語であると判定を行う．メールデータは件名と本文という構造をもった文書であり，この点で Web ページの構造と同じであるため，メールデータにおいても，この詳細語発見の手法は適用可能である．本研究では，閲覧中のメールを含む複数のメール集合から，ある話題の詳細語を抽出する部分において，メールの件名と本文に対し，この詳細語発見の手法を用いている．

3. 特徴語の特定

3.1 時間的な一区間を用いた特徴語の特定

ここでは，ユーザがあるメールを選択したと仮定し，その閲覧メールの解析を行う．現在閲覧中のメールの話題語を特定するための最初の段階として，現在閲覧中のメールを特徴付ける語の抽出を行う．メールデータは送信時間がついたデータである．そこで，その時間情報を利用することを考える．ある話題について何通かのメールを受信する場合を想定すると，それらのメールは時間的に離れて受信されるわけではなく，ある一定の日時に偏って受信するものであると考えられる．例えば，DBWS の論文投稿に関するメールは，論文投稿の告知のメールが開催の数ヶ月前に送られ，そこからワークショップ開催の後少しの間までに集中して受信される．そこで，特徴語を抽出するために，メールデータを時間区間に区切り，その時間区間内でのキーワードの出現数に着目する．ここで $TF-IDF$ 法と同様の考え方を用いる．$TF-IDF$ 法ではある文書に出現するキーワード k のその文書での重みを (3) 式の $TF(k)$，および (4) 式の $IDF(k)$ を掛け合わせることで求める．

$$TF(k) = \text{キーワード k がその文書に出現する数} \quad (3)$$

$$IDF(k) = log(\frac{\text{文書総数 } N}{\text{キーワード k が出現する文書数}}) \quad (4)$$

$TF-IDF$ 法ではある文書に頻出，かつ他の文書にはあまり出ないキーワードの重みが高くなるような重みの計算方法となる．本提案手法においては，時間的な

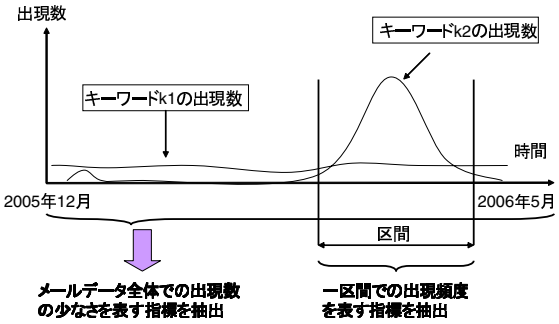

図 1 キーワードの出現数

ある区間 (2006 年 5 月 1 日から 5 月 31 日など) で出現頻度が高く，なおかつ他の区間ではあまり出現しないようなキーワードを特徴語と考えることとする．図 1 にキーワードの出現数の例を示す．このキーワード $k1$ は全体的に少し出現するキーワードである．キーワード $k2$ はある一区間において頻出するが，他の区間にはあまり出現しないキーワードである．本提案手法では，これらのキーワードを比較した場合，全体的に分散して出現する語よりもある一区間に局所的に出現する語の方が特徴のあるキーワードであると考えるので，特徴語としてはキーワード $k2$ のほうがキーワード $k1$ よりもふさわしいとみなす．

特徴語の抽出を行うために，閲覧メールの本文および件名について形態素解析を行い，名詞のみを抽出する．名詞以外の単語は話題語としてふさわしくないと考えられるからである．実装に際し，形態素解析には茶筌[7] を用いる．この形態素解析を行う前後にいくつか不必要な情報を除去する必要がある．

3.2 不必要な情報の除去

ここで，メール本文を解析するにあたり，その本文中に含まれるいくつかの不必要な文字列を除去する方法について述べる．

- 署名部分の除去
  メールの署名部分には差出人の名前，電話番号，所属組織，メールアドレス等が含まれるが，これらの情報はメール本文の話題の内容とは関係がないと考えられる．これらをそのままにして解析を行えば，多くのメールを受信している送信者の署名情報が多数出現するため，特徴語として選ばれてしまうということが起こる．そこでメール本文から署名部分を除去する必要が生じる．しかし，署名部分の書き方は人によって異なっており，一定の形式は存在しない．本文を引用して返信を行っていないメールにおいては，署名はメールの最下部に存在するが，本文を引用して返信を繰り返している場合には，本文中のあらゆる箇所に署名部分が存在する可能性がある．そこで署名部分を大まかに除去するために，メール本文を改行コードで分割し，その分割された部分において以下の 2 条件を満たす箇所を署名であると考え，除去するものとする．
  – メール本文の半分以降に出現
  – 句読点 (“。”，“、”，“，”，“．”) を含まない
  この方法ではすべての署名部分を除去できるわけではなく，署名部分でない箇所をいくつか除去してしまう可能性もあるが，署名部分をそのままに解析を行う場合の問題の方が大きいと考えるため，本提案手法においてはこの方法で署名部分の除去を行う．
- メールアドレスや長いアルファベット列の除去
  本文中に現れるメールアドレスや URL といったアルファベット列は，ある話題に関する特徴語となることはないと考えられる．よって，アルファベット列のうち “@” や “.” 等の記号を含むものをメールアドレスまたは URL であると判定し除去する．また，申し込み番号等で使われることのある長い英数字列についても同様の理由で除去を行う．また，送信者情報に含まれるメールアドレス以外の文字列は送信者の名前を表している．この名前情報も不必要な情報として本文から除去することにする．
- ストップワードの除去
  本提案手法においては，形態素解析において茶筌[7] を用い，名詞の抽出を行っている．その解析の結果得られた名詞の中には “あちら” や “向こう” といったメールの特徴語としてはふさわしくないと考えられる語がいくつか含まれることがある．そこで，それらの名詞をストップワードとして登録し，それらの語を得られた名詞集合の中から除去する．

3.3 重みの計算

3.3.1 計 算 手 法

こうして，特徴語の候補となる名詞集合が得られた．次にこれらの候補語の重み付けを行うことで特徴語としてふさわしい語の選択を行う．キーワードが特定の時間区間に頻出，かつ他の区間にはあまり出現しなければ重みが高くなるような重み付けを行う．まず，特定の時間区間に頻出であることを表す値を求める．これは $TF-IDF$ 法において TF に対応する．特定の区間に頻出であることを表す指標として以下の (5) 式，(6) 式が考えられる．

$TF_{区間}(k)$ = k が区間において出現する数 (5)

$DF_{区間}(k)$ = k が区間において出現するメール数(6)

次に特定区間以外の区間にあまり出現しないことを表す値，つまり全体的に出現することを表す値を求める．これは $TF-IDF$ 法において IDF に対応する．そのキーワードが局所的に出現することを表す値としては以下の (7) 式，(8) 式，(9) 式が考えられる．

$$\frac{1}{DF_{全体}(k)} = \frac{1}{全体でkが出現するメール数} \quad (7)$$

$$IDF_{全体}(k) = log(\frac{文書総数\ N}{DF_{全体}(k)}) \quad (8)$$

$$\frac{1}{(DF_{全体}(k))^2} \quad (9)$$

これらの TF と IDF にあたる二種類の値を掛けあわせることでキーワードの重みを計算し，その重みの高い語を特徴語と考える．これらの組み合わせは複数考えられるため 3.3.2 章で比較実験を行い，用いる手法の決定を行う．

3.3.2 実 験 結 果

実際のメールデータを用いて特徴語の抽出を行った．実験は以下の条件で行った．

- メールデータ
  2005 年 10 月 27 日から 2006 年 5 月 23 日までの 4000 通程のメール
- 時間区間
  5 月 1 日から 23 日まで
- 閲覧メール
  5 月 15 日に受信した DBWS の投稿に関するメール

表 1 に各計算方法で抽出されるキーワードとその値を記す．

表 1 を見ると，局所性を表す値として全体での IDF を用いた $TF_{区間} * IDF_{全体}$ および $DF_{区間} * IDF_{全体}$ において，“発表”，“多数”，“会場” 等の一般的なキーワードが抽出されている．このような一般的なキーワードが抽出されてしまう原因としては，メールの総数が 4000 通程度と多いため，出現数の大小が対数をとると，あまり差が出なくなってしまうためであると考えられる．他の 4 手法を比較してみると，あまり明確な差異は現れなかった．閲覧メールを変更して実験を行ってみたが明確に優越を決めることができなかった．よって以降では $DF_{区間}/DF_{全体}$ を候補語の重みの計算に用いることとする．今後実験を繰りかえし，適切な重みの計算手法を検討していきたいと考えている．

4. 話題構造の抽出

4.1 閲覧メールの話題の決定

3 章において閲覧中のメールから候補語を抽出し，特徴語を表す重みが計算された．次のステップとして，特徴語の重みの高いキーワードの中から閲覧中のメールの話題語を求め，この話題語のもつ話題構造の抽出を行う．まず，この特徴語の中から閲覧中のメールの話題となっている語を選択する必要がある．ここでメールの件名に注目する．メールの件名はメールの送信者がそのメールが何について書かれたものか，わかりやすいように考えてつけられたものである．そこで件名に含まれる名詞のうち，3 章で計算された特徴語の重みの最も高い 1 語をそのメールの話題語であると考える．

4.2 同トピックのメールの収集

4.1 章で閲覧メールの話題語が決定された．次にこの話題語の持つ話題構造の抽出を行う．そのために，現在閲覧中のメールの話題と同じ話題について書かれたメールを集める必要がある．ここでもメールの件名に注目し，全メールの中から閲覧中のメールの話題語をタイトルに含むメールを集める．これらの同じ話題について書かれたメールの中から話題構造の抽出を行う．

4.3 詳細語の抽出

4.3.1 概　　要

収集された同じ話題に関するメールの中から，話題構造の抽出を行う．メール群を解析することで，話題構造として話題語に対する詳細語となるキーワードの発見を行う．この話題語とそれに対応する詳細語を元に，閲覧メールの補完する内容のメールの抽出を行う．

4.3.2 候補語の取得

詳細語発見の最初の段階として，詳細語の候補となる語の抽出を行う．詳細語の候補はその話題について書かれたメールの本文中に出現すると考えられるので，集めてきた同じ話題語のメール群の本文および件名を解析し，含まれる名詞をすべて抽出する．この際，3 章で述べた特徴語抽出の際と同様の理由で，3.2 章の手法を用いて不必要だと思われる箇所の除去を行う．こうして得られた候補語の中から詳細語の選択を行う．

4.3.3 重みの計算

こうして得られた候補語の中から詳細語を発見するために，候補語それぞれの重みの計算を行う．まず 2 章で述べた小山ら[4] の詳細語発見の手法をメールの場合に適用する．つまり話題語 A，詳細語候補 B に対して (10)，(11) 式を計算する．

	$DF_{区間}/DF_{全体}$		$TF_{区間} * IDF_{全体}$		$TF_{区間}/DF_{全体}$	
	語	値	語	値	語	値
1	ホテル線慶	1.0	発表	350.3	dbws	1.4
2	月岡	1.0	原稿	282.5	原稿	1.1
3	dbws	0.9	論文	254.1	ホテル泉慶	0.8
4	アブストラクト	0.5	データベース	223.9	月岡	0.7
5	dbjapan	0.3	締切	210.8	アブストラクト	0.6
6	温泉	0.3	dbws	183.4	データベース	0.5
7	恒例	0.3	開催	169.3	dbjapan	0.4
8	データベース	0.2	タイトル	136.2	論文	0.3
9	ワークショップ	0.2	申込	125.7	発表	0.3
10	原稿	0.2	会場	109.8	タイトル	0.3

	$DF_{区間} * IDF_{全体}$		$DF_{区間}/(DF_{全体})^2)$		$TF_{区間}/(DF_{全体})^2$	
	語	値	語	値 (*10-2)	語	値 (*10-2)
1	dbws	111.6	月岡	33.3	月岡	22.2
2	発表	103.3	ホテル泉慶	16.7	ホテル泉慶	13.9
3	データベース	100.8	dbws	5.5	dbws	9.0
4	締切	99.2	恒例	4.1	アブストラクト	3.9
5	多数	83.6	アブストラクト	3.1	原稿	2.9
6	程度	75.9	dbjapan	1.9	dbjapan	2.3
7	申込	73.0	温泉	1.7	恒例	2.0
8	論文	72.6	原稿	0.5	温泉	1.0
9	ワークショップ	72.5	ワークショップ	0.3	データベース	0.6
10	検討	63.9	新潟	0.3	タイトル	0.2

表 1 抽出される特徴語

$$p(B|intitle(A)) = \frac{\text{A を件名，B を本文に含むメール数}}{\text{A を件名に含むメール数}} \quad (10)$$

$$p(B|A) = \frac{\text{A，B をともに含むメール数}}{\text{A を含むメール数}} \quad (11)$$

そして (10) 式が (11) 式よりも大きくなるようなキーワード B が詳細語であると考えられるため，ここでは (10) 式を (11) 式で割った値を重みの計算に利用する．ここで，この手法のみを用いた場合，一般的なキーワードはどのようなメールにも出現するため，重みが高くなってしまうという問題が起こる．そこで 3 章で話題語の決定に用いた重みを掛け合わせることで，この問題の解消を図る．つまり，話題語 A に対するキーワード k の重みを (12) 式によって求める．

$$w(k) = \frac{DF_{区間}(k)}{DF_{全体}(k)} * \frac{p(k|intitle(A))}{p(k|A)} \quad (12)$$

4.3.4 実　　験

ここで，閲覧メールでどのような詳細語が含まれているか実際に計算を行った．閲覧メールとしては DBWS の投稿募集のメールを用いた．得られた詳細語と重みの値を表 2 に示す．

“ホテル泉慶” といった開催地に関する情報や，論文投稿に関するキーワードが詳細語として取得できた．

	キーワード	重み
1	ホテル泉慶	1.19
2	月岡	1.19
3	アブストラクト	0.89
4	dbjapan	0.56
5	恒例	0.51
6	温泉	0.35
7	データベース	0.32
8	ワークショップ	0.23
9	原稿	0.30
9	タイトル	0.165

表 2 話題語 “DBWS” に対して得られた詳細語

5. 補完メールの提示

4 章で閲覧中のメールの話題語を求め，その話題語に対する詳細語候補の重み付けが行われた．ここではこれらを用いて，閲覧中のメールの内容を補完するメールの提示を行う．まず詳細語候補の重み上位数十件を詳細語であるとみなし，閲覧文書の詳細語および，同じ話題のメールの詳細語をそれぞれ求める．閲覧文書の持つ詳細語を全く含まないメールは，閲覧メールの内容と関係のない内容になっている可能性があるため，補完メールとしては選択しないことにする．そして，同じ詳細語を持つメールの中で閲覧メールに含まれる詳細語以外の詳細語をより多く持つものが補完度合いが高く，ユーザにとって有用であると考える．図 2 のように各メールが詳細語を持っていた場合，メール 2

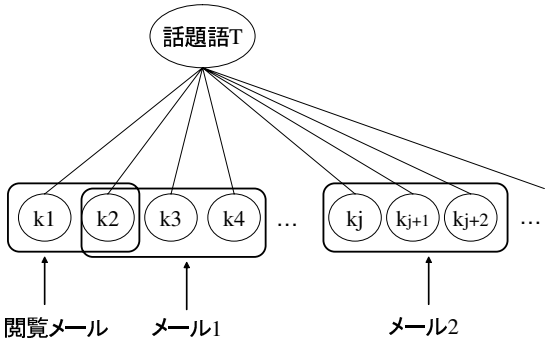

図 2 各メールが含む詳細語例

	得られる詳細語
閲覧中のメール	鎌原，原稿，アブストラクト
メール 1	月岡，ホテル泉慶，dbs，朱
メール 2	日本大学，中辻，成澤，霜田，三好
メール 3	日本大学，中辻，成澤，霜田，三好
メール 4	日本大学，中辻，成澤，霜田，三好
メール 5	補完，コンテキスト，アブストラクト
メール 6	faculty，アブストラクト
メール 7	intitle，intext，補完，コンテキスト，アブストラクト
メール 8	不老，鎌原，淳三，月岡，ホテル泉慶

表 3 得られた詳細語

は閲覧文書の詳細語を全く含んでいないため，メール 1 が閲覧メールを補完するメールとして適切であると考える．

ここで，提案手法を用いて実際に閲覧中のメールから話題語を抽出し，同じ話題のメールを収集し，詳細語の抽出を行った結果を示す．候補語の重みの高い上位 50 件を詳細語とした．閲覧メールとしては，これまでと同様に DBWS の投稿募集の案内のメールを用いた．その結果，話題語として “DBWS” というキーワードが取れた．そして “DBWS” を件名に含むメールを収集したところ 8 メールが集められた．そして，これらのメールの中から話題語 “DBWS” の詳細語の抽出を行った．閲覧中のメール，および集められたメールの詳細語の上位 5 件までを表 3 に記す．

閲覧メールおよびメール 8 は DBWS 論文投稿に関するメールであり，“アブストラクト” や ‘ホテル泉慶‘” といった論文投稿やワークショップの開催地の情報が詳細語として含まれている．メール 2，3，4 の詳細語として並んでいる語は “DBWS” にあまり関係のない語が抽出されている．これはメールは DBWS の話題について書かれているものであるが，そのメールを送る際に，過去に送られてきた “DBWS” とはあまり関係のないメールに返信する形で送られたメールであるため，引用の形で関係のない話題が含まれてしまったためである．このように，メールを送る際に話題とは関係ない過去のメールに返信するという形をとることがあるので，今後，こうしたメールにおいては，話題に関係のある部分と関係のない部分の判定を行ったうえで，詳細語の抽出を行っていきたい．また，学会等の論文投稿の案内という話題語，詳細語がわかりやすい例であれば問題はないが，たとえば研究室の研究会に関するメールではタイトルに一般的なキーワードしか含まれず，適切な話題語が選択できない場合が存在する．そのような場合には，タイトルに含まれるキーワードのみから話題語を決定すると，話題語としては広すぎる内容になってしまうため問題である．この場合，タイトルに含まれる語だけではなく，本文中のキーワードも組み合わせる事で，複数語での話題語を今後考える必要がある．表 3 の 8 つのメールを閲覧メールの詳細語を含み，閲覧メールに含まれない詳細語を多く持つ順に並べ替えると表 4 のようになる．

順位	メール	閲覧メールと共通の詳細語数	異なる詳細語数
1	メール 8	4	5
2	メール 7	1	4
3	メール 1	2	2
4	メール 5	1	2
5	メール 6	1	1
-	メール 2	0	39
-	メール 3	0	39
-	メール 4	0	36

表 4 メールをランキング

メール 2，3，4 は閲覧メールの詳細語と共通の詳細語を持たなかったためランキングから除外された．こうして得られた結果をユーザに提示することになる．たとえば，順位の最も高いメール 8 を選択すると，ユーザはホテル泉慶という開催地の情報を得ることができる．

6. プロトタイプ

本提案手法を実装したプロトタイプを作成した．実行した際の実行画面は図 3 のようになる．図 3 の左側の画面は普通のメーラと同じように，受信したメールの一覧画面，および一覧画面のうち選択したメールの本文を表示する部分から成っている．画面の右半分に現在のメールを補完するメール情報の表示を行っている，右上部の画面において同じ話題のメールの持つ詳細語をメールごとに表示している．また補完メールのリストは 5 章で求められたランキングの順に並べらる．ユーザがこの詳細語のうち必要であると思われるものを選択すると，右下部の画面に選択されたメールの本文が表示されるようになっている．選択されたメールを表示する際に，閲覧メールに含まれる詳細語と補完メールに含まれる詳細語を色分けして表示することで，わかりやすくなるようにしている．

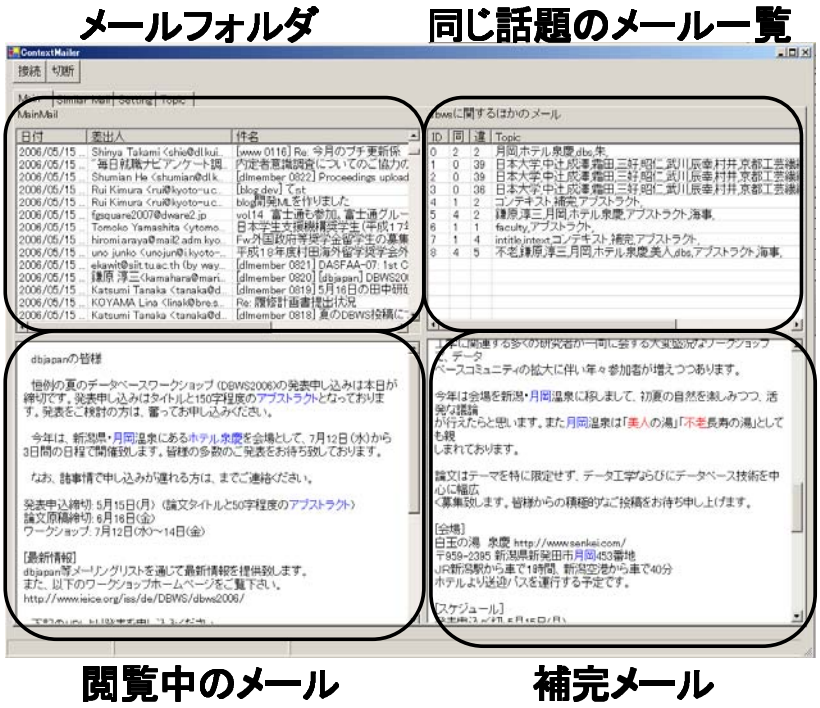

図 3 実行画面

7. まとめ及び今後の課題

本論文では，ユーザが閲覧しているメールを補完するメールの提示手法を提案した．閲覧メールから時間的な区間でのキーワードの出現数に着目し，話題語の抽出を行った．そして，得られた話題語を話題とするメールを集め，その話題語の詳細語を抽出した．そして，この詳細語を元に補完メールの提示を行った．

署名部分の特定および除去手法に関しては，現在簡単な方法で大まかに署名部分の特定を行っている．この手法では除去できない署名が存在する上，無関係な部分も除去してしまうことがあるため，今後改善を図りたいと考えている．また，関係のないメールに引用するという形で，新しい話題のメールを書いた場合の無関係な部分の特定を行うことで，より適切な詳細語の発見を行っていきたい．

現在は話題語がメールの件名に含まれていると仮定しており，タイトルにその話題語が含まれていなければ，たとえ本文中でその話題についてふれていても，ユーザに提示されることはない．タイトルにはその話題後は含まれないが，内容的にはその話題について書かれているメールを発見することで，より正しい詳細語の発見や，必要な情報の提示が行えると考えられる．今後，この同じ話題のメールの特定手法について考える必要がある．

現状では差出人の情報等は全く利用していないが，誰からどのような話題のメールが来ているという情報を利用することで，ユーザの必要とする情報の特定に役立つのではないかと考えられるため，今後利用を検討したい．また時間区間の対象時期および長さについても今後考察を行う必要があると考えている．

閲覧中のメールを補完する情報は過去に受信したメールのみにあるのではなく，Web ページの情報にも存在する．そこで今後，補完する Web ページをコンテキストを基に検索を行い，メール情報と共に提示することで，よりユーザの役に立つ情報を提示していきたいと考えている．

謝　　辞

本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（代表：田中克己），および，平成 18 年度科研費特定領域研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」（課題番号：18049041，代表：田中克己）によるものです．ここに記して謝意を表すものとします．

参 考 文 献

1) 馬強, 田中克己：“テキストストリームの文脈を考慮した補完情報検索.”, 電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005).
2) 馬強, 田中克己：“話題構造に基づく放送と Web コンテンツの統合のための検索機構.”, 情報処理学会論文誌： データベース (TOD23).
3) 馬強, 田中克己: “補完情報の検索に基づくコンテンツ統合.”, 情報処理学会研究報告、2004-DBS-134.
4) S.Oyama and K.Tanaka: “Query modification by discoverying topics from web page structures”, Proceedings of the 6th Asia Pacific Web Conference (APWeb2004).
5) 宮崎慎也, 馬強, 田中克己：“WebSCAN:Web サイトの変更発見と放送型変更通知.”, 情報処理学会研究報告, Vol.2000, No.69, 2000-DBS-122-68.
6) T.Mithcell: “Computer workstations as intelligent agents”, SIGMOD 2005 Keynote Talk (2005).
7) 形態素解析システム茶筌 `http://chasen.naist.jp/hiki/ChaSen/`.