社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－DD－67（12）
2008 / 7 / 25

特徴語の共起性を利用した携帯電話メールの関係性検出手法

仁野 裕一† 野田 潤† 中尾 敏康†

†NEC サービスプラットフォーム研究所 〒630-0101 奈良県生駒市高山町 8916-47
E-mail: †{y-nino@ab, j-noda@cw, t-nakao@bp}.jp.nec.com

あらまし 筆者らは，携帯電話に蓄積される各種情報の関係性を検出することによって，アプリケーションの操作性向上を目指す研究を進めている．その第１歩として，携帯メールに含まれる特徴語の共起性を利用して，トピック関連性が高いメールを自動検出する手法を開発した．これは特徴語を場所・時・人に関係する単語とし，時間的に近い同一宛先のメールを統合した上で，この特徴語の有無を多次元特徴ベクトルとして類似性が高い順にメールを検出する．本手法により，文章が短く，トピックキーワードが含まれにくい携帯メールにおいてもトピックに関するメールの検索が高精度となる．本手法を実際の携帯メールに適用した結果，トピックキーワードの検索結果に比べ，再現率が 10%～50%向上することを確認した．

キーワード 携帯電話，メール，トピック，関係性

A proposal of the method to detect the relation of cell phone mails using co-occurrence of feature words

Yuichi NINO† Jun NODA† and Toshiyasu NAKAO†

† NEC Service Platforms Research Laboratories 8916-47 Takayama-cho, Ikoma, Nara, 630-0101 Japan
E-mail: †{y-nino@ab, j-noda@cw, t-nakao@bp}.jp.nec.com

Abstract We develop the method to automatically extract topic-based relation of cell phone mails using co-occurrence of feature words. This method combines cell phone mails into the threads based on their sender/receiver address and time of origin/receipt, and generates feature vectors whose components are the words of location, time, person that contain in the mails, and analyzes the relational threads. It makes it more accurate to search the topic-related mails from cell phone ones whose sentences are generally short and lack of feature words related to the topic. We applied this method to real cell phone mails and found that it improves recall by 10-50% compared to topic key word search.

Keyword Cellular phone, email, short message service, topic, relation

1. はじめに

メーラ／スケジューラなどの PIM(個人情報管理)アプリケーションは，アプリケーション間の連携はあまり進んでいない．例えば，スケジューラに記載された予定に関するメールを参照する／その予定の参加者の連絡先を知るためには，メーラやアドレス帳など別のアプリケーションを起動して確認しなければならない．PC においては，それらのアプリケーションを同時に起動すればよく，それほど操作性は問題とならない．しかし，携帯電話の場合，画面の大きさの制約からアプリケーションを複数同時起動することが難しく，起動しなおすのであれば，操作性が大きく損なわれる．

また，携帯電話には GPS など個人の行動を捕捉することができる各種センサが搭載されてきており，ライフログなどの様々なアプリケーションが検討されて来ている[1]．これらセンサデータを PIM アプリケーションと連携して，操作性向上を実現できる．例えば，帰宅メールをよく送信している場所や時間帯に到達すると，帰宅メールの下書きがユーザにポップアップして示されればユーザのキー入力が少なくて済む．

このような PIM アプリケーションの操作性を向上させるためには，各 PIM アプリケーションが管理しているデータ間，PIM アプリケーションが管理しているデータとセンサデータ間の関連付けが有効である．なぜなら，異なる PIM アプリケーションのデータを参照する際には，ユーザはデータが関連するものを参照する傾向にあるし，ユーザが明示的に操作しなくても得られるセンサデータと PIM アプリケーションのデータとの関連性から，ユーザが状況に応じて必要な PIM アプリケーションのデータを自動で推定できるためである．

筆者らは，以上の目的から，これらの関連性を得る研究開発を実施しているが，その第一歩として，トピック関連性の高いメール検出技術を開発したので報告する．

従来，このようなトピック関連性の文書分類技術は，企業内のメールやブログなどで多く行われてきた．例

えば，平野[2]らは，あらかじめ 91 個のトピックをもとにサンプルブログから特に関係する単語(名詞，動詞，形容詞)の組を学習し，各ブログに含まれる単語が上記単語をどの程度含むかをもとに確率的に分類を実施している．

ところが，携帯電話のメールを分類する場合，(1)文章が短いので，分類に適した特徴語を含むことが難しい，(2)携帯電話のメールを分類するトピックは多様で事前に定められないので，あらかじめトピックの特徴語を決めて学習することが難しいという特有の問題があり，従来手法を単純には適用できない．

2. 提案手法

本手法は，このような携帯メール特有の問題を解決するため，以下のアプローチをとる．まず，同一送受信者が一定期間内に送受信したメールは同じ話題が継続されていることが多いという携帯電話メールの性質を利用して，(a)同一送受信先のメールのうち，一定期間内に送受信されたものを統合してスレッドとする．そして，PIM アプリケーションで使われるトピックでは場所・時・人に関係する単語が重要と考えられることから，(b) スレッドから抽出された場所・時・人にあたる名詞を特徴語として，その特徴語の共起性からメール関係性を検出する．

(a)により，短いメールをスレッドとして統合して長い文章とすることができるので，(1)の問題を解決できる．また，(b)により，多様なトピックに対しても良好に特徴語を抽出できるので，(2)の問題を解決できる．

本手法の処理の流れを図 1 に示す．

まず，受信したメールについて形態素解析を行う．これは，ChaSen[3]などの既存の手法を用いればよい．つぎに，場所・時・人に当たる単語を特徴語として抽出する．これは，辞書を使うなど様々な方式が考えられるが，筆者らは，辞書とルールを併用して検出した．詳細は次章に示す．そして，同一送受信先のメールのうち，一定期間内に送受信されたものをスレッドとして統合する．

つぎに，スレッドごとに含まれる特徴語をもとに，

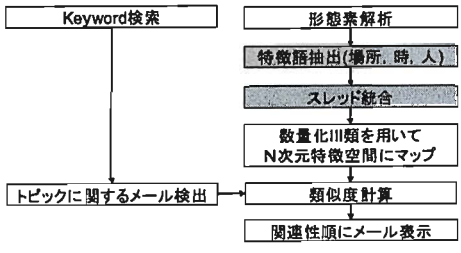

図 1 本手法の処理の流れ

スレッド間の関係性を検出するための特徴ベクトルを導出する．これは，各スレッドに対し，すべての特徴語の有無を(true, false)の値として記載した多次元特徴ベクトルを作成し，その多次元特徴ベクトルを数量化 III 類の処理によって次元圧縮したものである．このようにして得られたスレッド毎の特徴ベクトル間のユークリッド距離をもとに，スレッド間類似度を計算する．

つぎに，ユーザがトピックに関するキーワードを入力すると，そのキーワードを含むトピックに関するメールを抽出し，そのメールを含むスレッドを求める．

最後に，トピックに関するメールを含むスレッドに近いスレッドを類似度が高い順に検出し，スレッドをメールに戻した上で関係性順にメールを表示する．これはキーワードを含むメールを上位に示した上で，類似度が高いスレッドを類似度順に検出し，そのスレッドに含まれるメールを出力するものである．

3. 提案手法の実装

本手法の有効性を検証するため，Windows PC 上でプロトタイプを実装した．以下，図 1 の特徴語抽出と，スレッド統合の処理を説明する．

3.1. 特徴語抽出

3.1.1. 場所に関する単語抽出

本抽出処理は，形態素解析ツール ChaSen[3] ver.2.4.2，辞書として UniDic[4]の 2007/10 時点の最新バージョンを利用し，品詞として「地名と固有名詞」を抽出した．固有名詞を利用したのは，地名として，「マクドナルド」「ディズニーランド」などの組織名が含まれることが多いためである．ただし，固有名詞を含むことにより，地名と無関係な単語も拾われてしまうことがあるので，今後，辞書を改良することにより精度向上を図る予定である．

3.1.2. 時に関する単語抽出

本抽出処理は，"時"に関する単語として日付に関する用語を抽出し，同じ日付を指しているものについて共通のフォーマットに置き換え，特徴語として抽出する．この置き換えのための日付変換ルールの例を表 1 に示す．このルールは，メール 512 通から日付に関する標記を全部洗い出すことによって作成したものである．適宜ルールを追加し，精度向上を行っていくために，ルールは設定ファイルに記載するだけで追加可能になっている．

3.1.3. 人に関する単語抽出

“人”に関する抽出処理は，本実装ではメールの From/To アドレスのみの抽出を実施した．今後，本文

表 1 日付変換ルールの例

単語	変換ルール
今日／きょう	$(送受信日)
昨日／きのう	$(送受信日-1)
明日／あした，あす	$(送受信日+1)
一昨日／おととい	$(送受信日-2)
明後日／あさって	$(送受信日+2)
○月○日	$(送受信年の○月○日)
○日	送受信日＜○日なら$(送信月の○日) 送信日＞○日なら$(送信翌月の○日)
○/○	$(送受信年の○月○日)
○/○/○	$(○年○月○日)
日曜日/日曜	$(送受信日の直近の日曜日)
月曜日/月曜	$(送受信日の直近の月曜日)
火曜日/火曜	$(送受信日の直近の火曜日)
水曜日/水曜	$(送受信日の直近の水曜日)
木曜日/木曜	$(送受信日の直近の木曜日)
金曜日/金曜	$(送受信日の直近の金曜日)
土曜日/土曜	$(送受信日の直近の土曜日)
土日／週末	$(送受信日の直近の土，日曜日)
来週の○曜	$(送受信日の次の週の○曜日)
再来週の○曜	$(送受信日の次々週の○曜日)
先週の○曜	$(送受信日の前の週の○曜日)
(この)前の○曜	$(送受信日の前の○曜日)
来月の○日	$(送受信日の次の月の○日)
再来月の○日	$(送受信日の次々月の○日)
先月の○日	$(送受信日の前の月の○日)

* $()は括弧内の日付を共通のフォーマットに変換したもの

中に出てくる名前についても，電話帳に書かれている名前などを利用することによって特定していくことも予定している．

3.2. スレッド統合

本処理では，共通の話題を議論していると考えられるメールをスレッドとして統合する．PC のメールでは，メールトピックを subject に記載し，それに返信した場合は subject に「Re:」(返信であることを示す記号)が付与されるので，それを手がかりにスレッドを見つけることができる．しかし，携帯電話のメールにおいては subject に何も記載しないことが多いため，subject の情報は単に「Re:」「Re:Re:」という文字列が多く，スレッドを見つけるのに有用ではない．

一方で，携帯電話のメールは，chat のように 1 つの内容に関して短時間で連続的に特定の相手とやり取りされる傾向が高いため，それらを 1 つのスレッドとみなすことができる．そこで，同一のアドレスで送受信しているメールのうち，前の送受信から一定時間内に送受信したメールは同じスレッドで議論しているメールとみなして，統合する．この一定時間は，様々なメールデータから評価した結果，1 日の区切りで行うことが最適であることが分かった(詳細は 4.1.3 節で述べる)．さらに，多様な職種・世代のユーザのメールから 1 日の区切りに有効な時刻を評価した結果，午前 4 時が最適であることが分かった．そこで，この時刻を基準に，午前 4 時から翌日の午前 3 時 59 分までの同一送受信者のメールをスレッドとして統合する．

4. 本手法の評価

本手法の評価は，特徴語抽出・スレッド統合の個別処理の精度評価と，本手法を利用したメール関係性検出の精度評価の 2 種類を行った．

4.1. 個別処理の評価

4.1.1. 場所に関する単語抽出

3.1.1 節に示した場所に関する単語抽出処理の評価にあたり，512 通のメールを利用し，評価を行った．その結果，再現率は 88.9%，適合率は 76.1%であった．適合率の低下要因は，主に地名とは無関係な固有名詞が検出されたことによる．また，再現率の低下要因は，一般名詞と地名の両方で使われる名詞(例.王子)が一般名詞として検出されたこと，複数の一般名詞の組み合わせと地名との区別が難しい単語(例.十条)で複数の一般名詞の組み合わせとして検出されたことによる．適合率を改善するには辞書の改良で解決できるが，再現率を改善するには形態素解析の精度向上が必要である．

4.1.2. 時に関する単語抽出

3.1.2 節に示した場所に関する単語抽出処理の評価にあたり，512 通のメールを利用し，評価を行った．

その結果，再現率は 88.9%，適合率は 100.0%であった．抽出に失敗したのは，ルールベースで定義するのでは非常に難しい場合である．これは，日付を数字だけで記載した場合と，「曜」を省略して曜日を記載した場合の 2 通りがあった．数字だけで記載した場合，これは日付のみを指しているのか，他の意味を指しているのか（例えば，上記 500 通の例では，「31」をアイスクリーム店の意味で使っている場合もあった）を形態素解析と文字の前後関係から判断するのは難しい．また，曜日を省略した場合，それぞれの文字(月，火，水，木，金，土，日)は別の意味も持つので，その区別も形態素解析の結果だけでは判断が難しい．今回，曜日の二文字を組み合わせたもの(例えば，土日)はサポートすることにしたが，単に一文字のみで表わされたものについてはサポートしなかった．これらの場合において日付を抽出するには，文脈から数字の意味・月～土の文字の意味の検出が必要であり，今後の課題である．

4.1.3. スレッド統合

スレッド統合の有効な時間範囲を評価するために，前のメール受信時刻からの経過時間・時間帯(午前中・午後0時～午後5時・午後5時以降)による区切り・1日の区切りのうち，どれがよいか様々なメールデータから実験した．その結果，表2のように，経過時間・時間帯による分離・1日の区切りでは誤検出率(違うトピックのメールがスレッドに含まれてしまった数/総メール数)は殆ど変わらず，統合失敗率(同じスレッドのメールなのに別スレッドとして認識された数/総メール数)は1日の区切りが他に比べてずっと少ないことが分かった．統合失敗率が大きいとスレッドの文章が短くなり，特徴語数を減少してしまうという悪影響が大きいため，本手法では，誤検出率が少々増えても統合失敗率が低いことを優先し，1日の区切りで行っている．

なお，表2の経過時間には1時間(1H)しか記載していないが，その間隔を短くしても統合失敗数が増加するだけで，誤検出数の改善は殆ど見られない．また経過時間を長くしても統合失敗率が1日の区切り以上の精度にならなかった．同様に，時間帯の区切りを細かくしても統合失敗率が増加するだけで誤検出率の減少にそれほど寄与せず，逆に荒くしても1日の区切りより統合失敗率の精度が上回ることがなかった．

表 2 スレッドの区切り時間の評価

	前メールを受信してからの経過時間(1H)	時間帯による区切り	1日の区切り
誤検出率	9.1%	10.2%	11.3%
統合失敗率	9.1%	8.0%	1.2%

つぎに，512通のメールをこの処理でスレッド統合した結果，同じトピックのメールを80%程度で検出していることを確認した．なお，失敗するケースとしては，仲良しからのメールで話題転換があった場合が多かった．この話題転換は表2に示すように前のメールを受信してからの経過時間とは無関係に行われており，話題転換を検出するにはメール内部での意味的な解析を行う必要がある．また，男性は用事だけをメールするユーザが多く，話題が変化することが少ないが，女性は雑談に近い内容が特定ユーザ間で何通もやりとりされる傾向があるため，女性で多く失敗する傾向が見られた．

4.2. 携帯メールを利用した本手法の評価

本手法の評価のため，実際の携帯メールから関係性検出を行った．実験に利用したメールは，女子高校生の一般ユーザから収集した154通と30代社会人男性の377通のメールである．これらのメールは，スレッド統合処理をせずにトピックに関する正解を人手で与えたものを利用し，適合率と再現率で評価した．まず，女子高校生のメールと30代社会人男性のメールにおいてトピックに関するキーワード検索して得られた結果をそれぞれ表3，表4に示す．

表3，表4のキーワード検索では，各トピックを表すのに最適な単語を評価者が選び，検索した．その結果，適合率についてはほぼ100%であった(トピックキーワードを含んでいても関係ないメールが含まれていた場合があり，全てが100%にならなかった)，社会人男性のトピック1を除いて，再現率は7%～50%と低い値にとどまっている．これは，トピックに関係するメールでもキーワードを含まないメールも相当数存在するためである．これを，同じキーワードを入力し，本手法で抽出した結果をそれぞれ図2，図3に示す．

図2，図3では，横軸は本手法により関係性が高い順に検出したメール数(検出メール数)，縦軸はこのメール数だけ検出した場合の適合率・再現率を示したものである．

メールをランダムに検出した場合，適合率は検出メール数とは無関係に低い値に留まり，検出メール数が増えるのにつれ，再現率はなだらかに上昇する．

表 3 女子高校生のメールのキーワード検索結果

	ﾄﾋﾟｯｸ1	ﾄﾋﾟｯｸ2	ﾄﾋﾟｯｸ3
トピック内容	お祭り	ある場所での遊び	ある人物の別れ話
キーワード	場所名	場所名	人物名&喧嘩
各トピックに関するメール数（正解数）	24	18	14
キーワード検索で検出したメール数	4	5	1
└うち正解数	4	4	1
再現率	0.17	0.28	0.071
適合率	1.00	1.00	1.00

表 4 社会人男性のメールのキーワード検索結果

	ﾄﾋﾟｯｸ1	ﾄﾋﾟｯｸ2	ﾄﾋﾟｯｸ3
トピック内容	テニス	入院	CDの話題
キーワード	テニス	病気箇所 and 入院	アーティスト名 or アルバム
各トピックに関するメール数（正解数）	38	28	8
キーワード検索で検出したメール数	28	5	4
└うち正解数	27	5	4
再現率	0.72	0.18	0.5
適合率	0.96	1.00	1.00

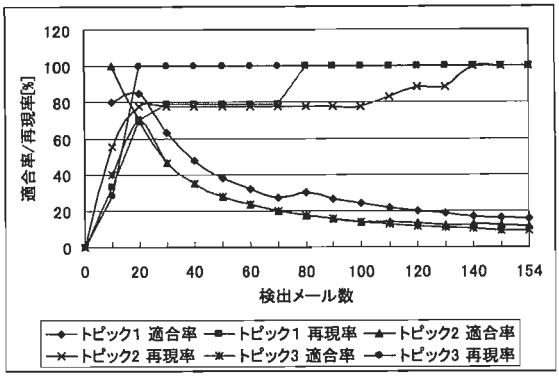

図 2 女子高校生のメールの本手法による検出結果

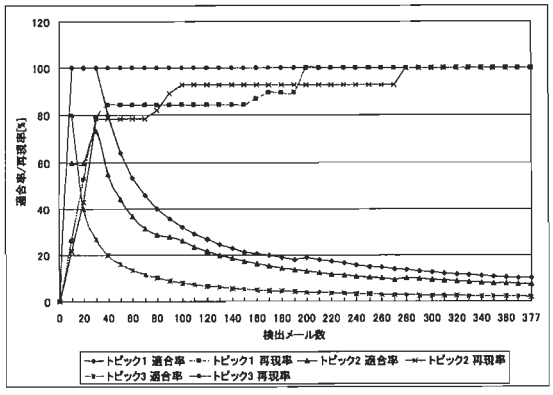

図 3 社会人男性のメールの本手法による検出結果

一方，本手法の結果を見ると，適合率はトピックの正解数近辺まで 60%～80%程度と高い値を示し，それを超えるとなだらかに減少する．また，再現率はトピックの正解数近辺まで急激に上昇して 70%～80%程度に到達し，以降はなだらかに上昇する．これは，トピックの正解数近辺までは検出されたメールに高い確率で正解が含まれており，以降は序々に正解が含まれにくくなっていることを意味している．したがって，本手法では上位にトピックに関するメールを適切に検出できていることが分かる．

5. 人間関係検出への本手法の適用

3 章までに示した手法は，メールに含まれる場所・時・人の特徴語の共起性からメールの関係性を検出するものであったが，送受信者のメールに含まれるこれらの特徴語の共起性から送受信者間の人間関係に近いものが得られると考えられる．なぜなら，同じ場所・時・人に関する特徴語がメール中に現れる人物は，リアル世界で同じ場所に居合わせる可能性が高いことを意味するからである．

本手法をスレッドではなく，送受信者ごとにメールを統合した上で同様の処理を行い，一定の閾値以上の類似性が見られる送受信者を検出した．図 4 はこの検出した人間関係を可視化ツールで表示した例である．このツールでは，閾値以上の類似性があると判定された人物間にリンクが張られ，各人物を表す頂点をクリックすると，リンクが張られた頂点が別色で表示される．また，ここに示した人物属性データ(father,mother など)はユーザからあらかじめヒアリングして得たユーザとの関係を示したもので，評価用に表示している．

これを 4 章に示した社会人男性のメールに適用し，可視化した例を図 5，6 に示す．図 5，6 では白黒印刷でも分かるように，別色で表示された人物に枠を入れ，人物間のリンクを太線で強調して表示している．図 5 では，弟(brother) に関係が近い人物を可視化ツールに

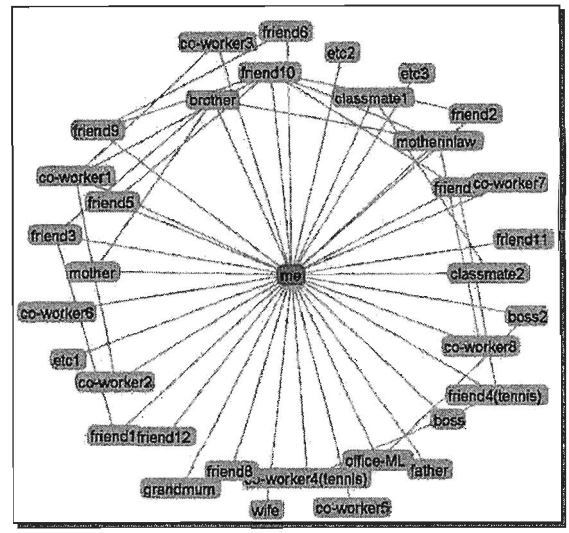

図 4 人間関係の可視化例

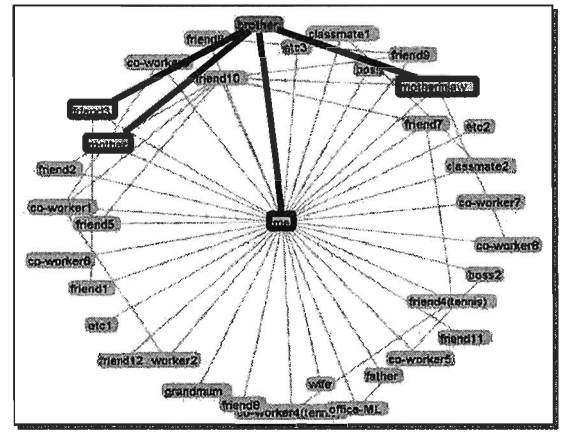

図 5 弟(brother)に関係が近い人物の例

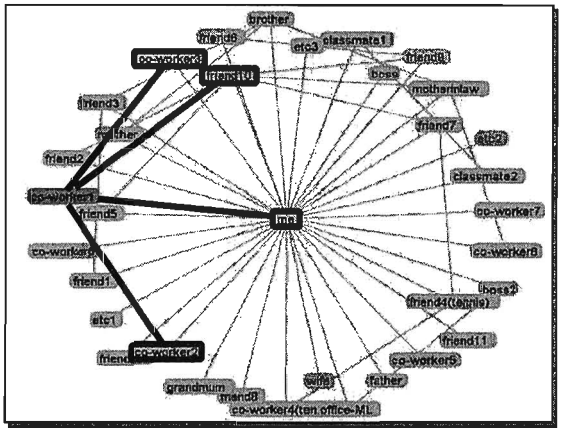

図 6 同僚(coworker1)に関係が近い人物の例

て表示させたものである．検出された人物を図中の線で囲んだ領域で示したが，母親や義理の母親など家族に関係する人物が中心である．また，図 6 では，ある同僚(coworker1)に関係が近い人物を可視化したものである．このときは，他の同僚(coworker2,3)など会社関係の人物が中心に正しく検出された．

これより，本手法により検出される送受信者間の人間関係は，実際の人間関係に近いことが分かった．従来，このような人間関係を検出する研究は，松尾ら[5]の研究のように，論文共著者・所属・プロジェクト名など，直接的な情報から人間関係を検出する研究が多かった．しかし，人間関係を表す直接的な情報がなくても，メールから場所・時・人に関する特徴語を検出することによって，ある程度の関係性を検出できることが確認できた．

6. おわりに

本稿では，関連性の高い携帯メールの検出手法を提案した．本手法の特徴は，場所・時・人にあたる名詞を特徴語として関係性を検出すること，一定期間内に送受信されたものをスレッドとして統合することにより短文から構成される携帯電話メール間の関係性検出を実現したことの 2 点である．

さらに，送受信者ごとのメールを統合し，本手法を適用すると，送受信者の実際の人間関係に近い関係性も検出できることを確認した．

今後は，本手法の特徴語検出を高精度化するばかりでなく，PIM アプリケーションに記載された場所・時・人情報と GPS などのセンサ出力との対応づけを行い，PIM アプリケーションの操作性を向上していくための研究開発を進めていく．

文　献

[1] 森川他，“携帯電話を利用したライフログ取得と SNS への適用に関する一検討”，電子情報通信学会総合大会，B-15-1，p.612，Mar. 2007

[2] 平野，古林，高橋，“日本語圏ブログの自動分類，”情報処理学会研究会報告(自然言語処理)，2005-NL-170, pp. 21-26, Nov. 2005.

[3] ChaSen, http://chasen.naist.jp/hiki/ChaSen/

[4] UniDic, http://tokuteicorpus.jp/dist/

[5] 松尾，友部，橋田，中島，石塚，“Web 上の情報からの人間関係ネットワークの抽出“，人工知能学会論文誌, 20 巻 1 号 E, pp.46-56, 2005