個人文書から抽出した語彙の意味関係に基づく Web 情報検索

Web Search Personalization based on Semantic Relationships between Terms Extracted from Personal Documents

大島 裕明[1] 小山 聡[2] 田中 克己[3]

Hiroaki OHSHIMA Satoshi OYAMA Katsumi TANAKA

ローカルコンピュータには非常に多くの個人的な文書が存在する．それらは人間にとっては理解可能であるが，コンピュータがそこに含まれる知識を抽出して利用することは行われていなかった．我々は，これまで個人的な文書とその分類構造から，語彙の意味関係を抽出する研究を行ってきた．本稿では，個人的な文書として特に，これまで扱ってこなかった電子メールに着目する．しかし，電子メールは，ファイルに対するディレクトリ構造のような分類構造が存在しないことが多い．そこで，電子メールが持つメタ情報を利用して階層構造を仮想的に取得することによって，これまでの手法を適用する．そして，抽出された個人の知識を利用してウェブ情報検索の個別化を行う手法について提案を行う．

This paper proposes a method to extract semantic relationships between terms which appear in documents in a personal computer, and its application to Web search personalization. Nowadays, a lot of personal information is stored in a personal computer. We have proposed the method to extract the semantic relationships from general documents classified in directory structures. In this paper, we treat email messages as personal information resources. As email messages are usually not classified, some classification structures are needed to be extracted with using meta data in email messages. Then, we propose Web search personalization using the extracted personal knowledge.

1. はじめに

現在，コンピュータ上では様々な個人的な文書が取り扱われている．自分で作成した文書や，興味がある事柄に関する文書，さらに，電子メールなどによって，日々多くの文書がコンピュータ上に保存され続けている．今後ますますそのような個人的な文書が増え続ける状況においては，それらをうまく活用する必要性が出てくると考えられる．

例えば，デスクトップサーチと呼ばれるものは，ローカルコンピュータ上の個人的な文書やその他のファイルをウェブの検索エンジンのように検索できるようにするツールであり，Google，Microsoft，Yahoo!などが次々と発表している．これらは，ファイルへの便利なアクセス方法を提供し，ファイルを利用しやすくしていると言える．しかし，ファイルの中にある情報を利用することは行われていない．

個人的な文書にはそのユーザに関する非常に多くの情報が入っていると考えられる．例えば，誰かが他人のコンピュータ上の文書を読んだ場合，その人がどのようなことに興味があるのか，どのようなことを知っているのか，ある事柄についてどのような意見を持っているのか，などを理解するのに十分な情報が存在していると考えられる．

我々はこれまで，文書とその分類構造に着目して，このような個人的な知識を抽出し利用すること研究を行ってきた[1]．本稿では，これまで扱っていなかった電子メールにも着目し，Web情報検索におけるパーソナライゼーションを実現する手法について提案を行う．

2. 関連研究

すでにいくつかの研究において，個人的な情報からの知識を生成するような研究が行われている．

Haystack[2]はMITが開発した，個人的な情報管理システムである．扱う情報は，e-mailやカレンダー，文書，Webページなど多岐にわたり，それらを一括してRDFで管理することができる．WorkWare++[3]は富士通研究所が開発した，会社などのグループで用いられるビジネス文書の蓄積と再利用のための情報管理システムである．さまざまな文書が登録され，その登録時には時間などのメタ情報が自動的に付加される．また，人やイベントの情報も同時に管理されている．ユーザは蓄積されたメタ情報を元に，ある研究分野に関してどのような技術が蓄積されているか，ある事柄を知っていそうな人が誰であるか，といった情報を取得可能である．Hyperclip[4]はNTTが開発した，知識流通プラットフォームである．ユーザが利用した複数のコンテンツの間の関係を表現することができ，そこで作成されたRDFをピア・ツー・ピアネットワークで共有することによって，ある文書と関連する文書を検索することができるようになる．Hyperclipで検索できる文書はピア・ツー・ピアネットワーク上の誰かによってメタ情報が付加されたものである．湯川ら[5]は，個人が所有する文書に出現する単語の隣接度合いから，それぞれの単語同士の関連度合いを表す概念ベース，パーソナル・リポジトリを個人ごとに作成した．ユーザがコミュニティーのピア・ツー・ピア型システムの他の人が保有する情報を検索するときには，エージェントが検索キーをパーソナル・リポジトリによって拡張し，他人のパーソナル・リポジトリ内でどのような情報が検索結果として適当であるかを判断することが可能になる．

これらの研究は，ある特定の環境やコミュニティーの中で利用可能な知識を作成し，それを特定の目的に活かすことを目的としており，より広い範囲で利用しようとしている本研究とは異なるものである．

3. 語彙の意味関係の抽出

3.1 意味関係抽出の考え方

我々はこれまで，ローカルコンピュータ上に保存された個人的な文書と，それらが分類されているディレクトリ構造から，個人的な知識を抽出することを行ってきた．本研究においてはそれを電子メールに対して適用する．本章ではまず，文書とその分類構造から，一般的にどのように個人的な知識

1 学生会員 京都大学大学院情報学研究科博士後期課程 ohshima@dl.kuis.kyoto-u.ac.jp

2 正会員 京都大学大学院情報学研究科助手 oyama@dl.kuis.kyoto-u.ac.jp

3 正会員 京都大学大学院情報学研究科教授 tanaka@dl.kuis.kyoto-u.ac.jp

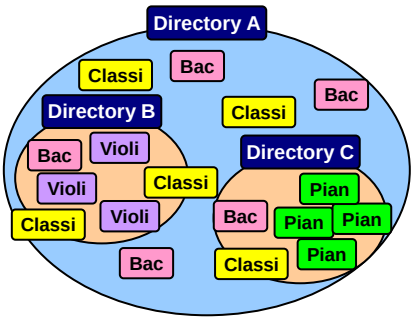

図１　あるディレクトリにおける語彙の偏りの例
Fig. 1　Deviations in term appearance in directory

を抽出するかということについて述べる．

個人がローカルコンピュータ上に保存している文書は，ディレクトリなどの階層構造の中で分類管理されていることが多い．どのような文書を持っているかということや，それらをどのように分類しているかということには，個人の知識や考え方が反映されていると考えることができる．

我々はそのような個人的な文書と分類構造から個人的な知識として，ユーザが文書中に現れる語彙どうしの間にどのような関係性をとらえているか，ということを抽出することを行ってきた．このような語彙の関係性を，語彙の意味関係と呼ぶ．文書群から得られる最も一般的な語彙の意味関係としては共起関係があげられるが，文書の分類構造を利用することによってより多くの関係性が抽出できる．その際に注目したのは，ディレクトリ構造の各階層における語彙の出現の偏りである．

図1は，あるディレクトリにおける語彙の出現の偏りを模式的に表したものである．ディレクトリAにはB，Cという二つのサブディレクトリが存在しており，その中の文書に様々な語彙が出現している．その語彙の分布を見ると，「Classic」や「Bach」は偏り無く広く分布しており，「Piano」や「Violin」は特定の場所だけに偏って分布していることが分かる．この時，2つの語彙の間にある関係性を考えると，『ClassicはPianoに対して広く使われる語である』や，『ClassicとBachは共起して使われる語である』というような関係をとらえることができる．このような関係性を，2つの語彙が対象とするディレクトリにおいて，どのような出現の偏りをもっているかによって，『広域的-狭域的』と『共起的-排他的関連』という2つの関係軸上で考える．図1の例では，ClassicはPianoに対して広域的な語であり，逆にPianoはClassicに対して狭域的な語であるといえる．図2が，2つの軸における語彙の出現の偏りを表したものである．ここでは，語彙Xに対して語彙Yがどのような関係であると考えられるかを示している．

2つの語彙の出現の偏りに違いがあった場合は，「広域的」な関係や「狭域的」な関係が考えられる．語彙の出現の偏りが，両方とも同じような場合には，「共起的」な関係や「排他的関連」という関係が考えられる．

このように，あるディレクトリ，つまり文書が分類された所における語彙の出現の偏りに着目すると，語彙の間に関係性を求めることができる．これは，文書が分類されているあらゆる所において求めることができるものである．当然，その時々によって得られる関係性も違ってくる可能性がある．例えば，あるディレクトリDにおいてXがYに対して広域的と判断されたとする．しかし，別のディレクトリD'に着目すると，XはYに対して狭域的と判断されるかもしれない．よって，

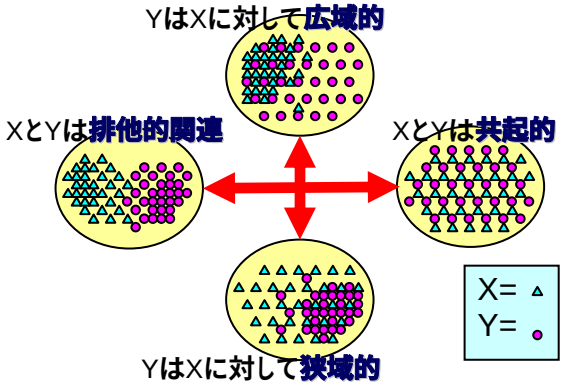

図２　語彙 X に対する語彙 Y の関係
Fig.2　Term Y's relationship to term X on two axes

語彙の偏りから関係性を求めるときには，どの部分を情報源として用いるかと言うことが重要になる．

また，対象とするディレクトリの中でそれぞれの語彙が出現する回数が多ければ多いほど関係性が深いと言うことができる．

3.2 意味関係抽出のアルゴリズムと実装

これまで述べた考え方を基に，語彙の意味関係を抽出するための基本的なアルゴリズムは，以下のようになる．(1)対象とするディレクトリDにおける，語彙X，Yの出現の偏りを求める．(2)得られた偏りをもとに，各関係においてどの程度の関係性があるかを求める．(3)語彙X，Yの出現頻度を求める．(4)計算された値を基に，語彙間の意味関係を求める．

実装には様々なものが考えられる．ここでは，現在我々が行っている実装について述べる．まず，説明に用いる記号について説明する．Dは対象ディレクトリとする．X，Yは対象語彙とし，この2語の間の意味関係を求める．$Sub_i(D)$はDまたはD以下のサブディレクトリのうちの1つを表す．$V_{関係名}(X,Y,D)$はDにおけるXとYの各関係（広域的，狭域的，共起的，排他的関連）における評価値とする．Num(D)はDに直接存在している文書の数とする．Num(X,D)はDに直接存在している文書のうちXを含むものの数とする．

始めに，語彙の出現の偏りの計算方法について述べる．現在は，ジニ係数という，主に経済学において富の偏在性などを表すのに用いられる指標を利用する．ジニ係数の範囲は[0,1]であり，完全に偏りがない場合には0になり，完全に偏っている場合には1となる．

今，着目しているディレクトリが図3におけるAというディレクトリであった場合について考えてみる．各ディレクトリにいくつかの文書が分類されていたとして，語彙Xの偏りを求めるために，まず，各ディレクトリにおいてXが出現する文書がどの程度の割合で存在するのかと言うことを求める．例えば，Eというディレクトリに文書が10存在し，そのうち8の文書にはXが出現する時には，割合は0.8になる．

図3のグラフは，語彙の出現の割合を縦軸に，ディレクトリの文書数を横軸にしたものの例である．この場合，ディレクトリAにはあまりXは出現せず，ディレクトリEにはより頻繁に出現することがわかり，ある程度偏って出現すると言うことが判断できる．このグラフに表されたデータをジニ係数計算のためのデータとして用いる．縦軸の量は，ディレクトリD以下での語彙Xの出現割合として，以下のように表される．

$$P(X, Sub_i(D)) = \frac{Num(X, Sub_i(D))}{Num(Sub_i(D))}$$

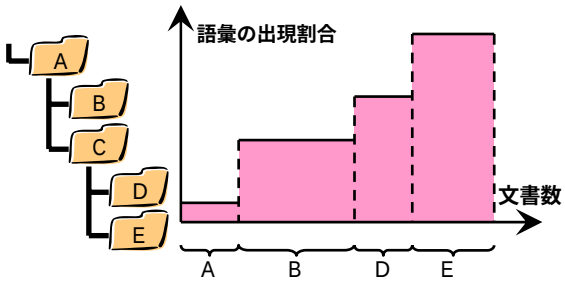

図3 ディレクトリ構造と語彙の出現割合の例
Fig.3 Example graph of deviations in the appearance of terms in a directory structure

横軸は各ディレクトリにおける文書数なので，Num(Sub$_i$(D))と表すことができる．これらを基に，DにおけるXのジニ係数GCを計算すると，以下のような式になる．

$$GC(X,D)=\frac{\sum_i\sum_j\left(\left|P(X,Sub_i(D))-P(X,Sub_j(D))\right|N_iN_j\right)}{2N_{all}(N_{all}-1)\cdot N_X}$$

ただし，N_i，N_j，N_{all}，N_Xはそれぞれ以下の式で表される．

$$N_i=Num(Sub_i(D)),\quad N_j=Num(Sub_j(D))$$

$$N_{all}=\sum_k Num(Sub_k(D)),\quad N_X=\sum_k Num(X,Sub_k(D))$$

次に，語彙の出現回数を考慮せず，語彙の偏りだけに着目したときの語彙どうしの関係性を$R_{関係名}(X,Y,D)$と表す．現実装では，Rを以下のように計算している．

$$R_{広域的}(X,Y,D)=GC(X,D)\cdot(1-GC(Y,D))$$

$$R_{狭域的}(X,Y,D)=(1-GC(X,D))\cdot GC(Y,D)$$

$$R_{共起的}(X,Y,D)=(1-GC(X,D))\cdot(1-GC(Y,D))$$

$$R_{排他的関連}(X,Y,D)=GC(X,D)\cdot GC(Y,D)$$

語彙の出現回数による重みをTとする．TはディレクトリD以下の文書で，対象とする語彙X，Yがどれだけ出現したかを表すDFを用いる．しかし，一般的な語彙は出現回数が多くなるので，全文書から求められるIDFを掛けあわせる．結果として，以下の式を出現回数による重みとする．

$$T(X,Y,D)=DF(X,D)\cdot DF(Y,D)\cdot IDF(X)\cdot IDF(Y)$$

RとTを掛け合わせたものが，関係性Vとなる．すなわち，以下のような式で，それぞれの意味関係の式が表される．

$$V_{関係名}(X,Y,D)=R_{関係名}(X,Y,D)\cdot T(X,Y,D)$$

この式によって，あるディレクトリにおいて各関係において最も関連が深い語彙のペアを見つける，というようなことができる．また，ある語が始めに与えられ，その語に対して他の語彙がどのような関係性を持っているかということを調べる場合には，対象とするディレクトリで意味関係を求めて，さらに，そのサブディレクトリにおいても意味関係を求め，最終的にそれらに重み付けして足し合わせるということを行う必要がある．

4. 電子メールのメタ情報を利用した分類構造の取得

これまで語彙の意味関係を求める手法について述べたが，語彙の意味関係を求めるためには，文書群とそれらが属する階層構造が必要である．一般的な文書はディレクトリ階層で主に内容によって分類されていることが多いが，電子メールでは，一般的な文書ほど分類がされないことが多く，分類構造が存在しても語彙の意味関係を求めるには十分ではない．そのため，何らかの方法で階層構造を取得する必要がある．メールは通常の文書よりも多くの定型的なメタ情報を持っている．例えば，送信者，受信者，日時，メッセージID，リプライメッセージID，といったものである．ここでは，これらから，分類に用いることが可能な階層構造を求める手法について考える．

送受信者の情報は潜在的に階層構造を持っている場合がある．例えば，所属するグループや学年などであり，それらの情報は，アドレス帳などには含まれている場合がある．それらを利用すれば，メールに対して自動的に階層構造を与えることができる．

メッセージIDは各メールにユニークに付けられるIDであり，メールのヘッダにおいてMessage-Idという名前で付加されている．また，あるメールに対して返信した場合には，返信される対象になったメールのメッセージIDがIn-Reply-Toという名前で付加される．これらによって，メールのツリーを生成することができ，いくつかのメールクライアントでは視覚的にこのツリーを表示するものがある．このツリー構造は小さいながらも利用できる可能性がある．

日時は細分化することによって，いくらでも構造を作ることが可能である．例えば，まず，一週間ごとにメールをまとめて最下層の分類を作り，その上位層として一ヶ月ごと，半年ごと，一年ごとなどとまとめるような方法である．その上で，最近一ヶ月に着目したときに，語彙の意味関係を求めることによって，最近ではどういう考えを持っているのか，ということを表すことができると考えられる．

このようにして，いくつかの分類構造を作ることができれば，それらを組み合わせることによってさらに深い構造を作ることも可能である．例えば，全体を送受信者によって自動的に分類し，各人のまとまりの中でさらに日時によって階層化する，といったことができる．これによって，十分な階層構造を得ることができるので，先述した方法で語彙の意味関係を取得することが可能となる．

5. Web 情報検索の個別化

我々はこれまで，一般的な文書から得た語彙の意味関係を用い，Web情報検索のパーソナライゼーションが可能なシステムを作成してきた．そのシステムでは，クエリ拡張と検索結果の再ランキングの機能を実現していた．まず，対象とするディレクトリをユーザに選択してもらい，それを基に語彙の意味関係を取得する．ユーザはクエリを入力し，そこで関連が深い語をクエリに追加するなどのクエリ拡張を行うことができる．そのクエリをもとにシステムは既存の検索エンジンを利用して検索結果を求める．求められた検索結果は，クエリで用いた語彙に対して関連が深い語が多く含まれるものが上位に来るように再ランキングを行うことができる．

今回は，一般的な文書ではなく，電子メールから得た語彙の意味階層を用いて，どのようにウェブ情報検索のパーソナライゼーションが行えるかということについて考える．メールに対する階層構造としては，4章の最後に述べたように，送受信者によって自動分類したあと日時によってさらに階層化したものを用いた．

そのような構造を用いることで，これまでの機能に加えて，誰を念頭において検索を行うのか，いつのことを念頭において検索を行うのか，ということをユーザが指示できるようになり，これまでの単なる語彙マッチングの検索よりも，より様々な意味をもった検索を行えるようになると考えられる．

実験のために用意したメールは，私の2003年4月から2005年6月までの27ヵ月分のすべての送信メール744通である．送信メールにはそのコンピュータのユーザが書いた文書が存在し，その人の知識を表すものとして非常に有用であると考えられる．

まず始めに，メールを送信先アドレスごとに分類した．これは，メールヘッダのToを用いた．今回は実験のため，同報メールの場合はToの最初に記載されているアドレスへのメールとして分類を行った．次に，送信アドレスごとの分類の中で日時を用いて，以下のようにしてさらに下位の階層構造を作成した．(1)最小の単位として一ヶ月ごとの分類を作成する．(2)それらを3つまとめて四半期の分類を作成する．(3)さらに4つまとめて一年の分類を作成する．これにより，人ごと-一年ごと-四半期ごと-一ヶ月ごと，という4階層の分類構造が作成された．

まず，ある程度長期間にわたって継続的にメールを送っている教官のメールアドレスを対象とした．ここで，「研究」という語に対して，どのような関係性を持つ語が現れるかを調べた結果が表1である．

表1 語彙「研究」と関連する語彙
Table 1 Terms related to "Research"

広域的	狭域的	共起的	排他的関連
検索	学部	内容	アブストラクト
個人	博士	概念	セマンティック
場合	計画	考え	範囲
自分	修士	利用	学生

共起的という所をみると，ユーザである私が行っている研究に必ず出てくるような語が現れている．次に，広域的という所を見てみると，もう少し一般的な時にも使うような語が出現している．狭域的では，「学部」「博士」「修士」という関連するような語が出現しており，それぞれの時期における研究の話が別々に出現していたことが予想される．排他的関連では，「セマンティック」という語が現れており，これは，本研究がセマンティックウェブと近いが異なるやり方でやっていることが現れているととらえることができる．この結果では，クエリ拡張に使うには少し一般的すぎる語が出現してしまっていて効果的ではないかもしれないが，再ランキングでは効果的に働くと考えられる．

現在興味があることや，一年前に興味があったことなどによって再ランキングを行い，自分の興味がどのように変化したかというようなことも，ある時期のメールを対象とすることでとらえることができると考えられるが，今回は行うことができなかった．他にも，得られた関係性の違いを活かして，より概要的な文書や，より詳細な文書を求めるといったような，意味づけを伴った検索を行えると考えられるので，そのようなことも含めて今後の課題としたい．

6. まとめと今後の課題

本稿では，個人的な文書から，語彙の間にある意味関係を抽出する手法について提案を行った．文書として電子メールを用いたが，電子メールが持っているメタ情報から構造を抽出し，それを利用することによって意味関係抽出の手法が適用できるようにした．そして，抽出された個人の知識を利用してウェブ情報検索の個別化を行う手法について提案を行い，実験による検証を行い，検索に意味を持たせることができることがわかった．

今後は，さらに多くの実験を行い，一般的な文書から得られる意味関係と電子メールから得られる人や時間を考慮したような意味関係が，それぞれどのような場面で役立つのかということを検証していく．また，語彙の関係性も現在の4つばかりでなく他の関係性なども考慮したい．同時に，別のアルゴリズムや実装を検討し，精度を高めていく．

[謝辞]

本研究の一部は，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17年度科研費特定領域研究(2)「Webの意味構造発見に基づく新しいWeb検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己），および，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」によるものです．ここに記して謝意を表すものとします．

[文献]

[1] Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka.: Extracting Personal Conceptual Structures from PC Documents and its Application to Web Search Personalization, Proc. International Special Workshop on Databases For Next Generation Researchers (SWOD 2005), pp.32-35 (2005)

[2] E.Adar, D. Karger, L. Stein.: Haystack: Per-User Information Environment, Proc. 1999 Conference on Information and Knowledge Management, pp. 413-422 (1999)

[3] 片山佳則，小櫻文彦，井形伸之，渡部勇，津田宏.: セマンティックグループウェア WorkWare++と KnowWho 検索への応用，情報処理学会 研究報告「情報学基礎」，No.071 (2003)

[4] Hiroyuki Sato, Yutaka Abe, Atsushi Kanai.: Hyperclip: a Tool for Gathering and Sharing Meta-Data on Users' Activities by using Peer-to-Peer Technology, WWW2002 Workshop on Real world RDF and Semantic Web application (2002)

[5] 湯川高志，吉田仙，桑原和宏.: パーソナル・レポジトリに対するピア・ツー・ピア型協調検索機構の提案，電子情報通信学会 信学技報 AI2001-48 (2001)

大島 裕明 Hiroaki OHSHIMA
京都大学大学院情報学研究科博士後期課程在学中．2004 年神戸大学大学院自然科学研究科博士前期課程修了．Web 環境におけるパーソナライゼーションの研究に従事．情報処理学会，日本データベース学会，ACM 各学生会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．