DEWS2004 I-3-03

個人オントロジを基にしたWeb情報検索支援に関する研究

大島　裕明†　　田中　克己††

† 神戸大学大学院自然科学研究科情報知能工学専攻
〒 657–0013 神戸市灘区六甲台町 1–1
† 京都大学大学院情報学研究科社会情報学専攻
〒 606–8501 京都市左京区吉田本町
E-mail: †{ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，まず，個人が日々取り扱う文書などから個人オントロジを半自動的に生成・更新する．そして，ウェブ上で検索を行う際に，個人オントロジの情報を基にして，ウェブ情報検索を支援する手法について提案を行う．個人オントロジは大きく，(1) 分類された文書群，(2) 概念の階層構造，から成り立ち，個人が現在どのような知識を持っているか，どのような概念の階層構造を持っているかを表す．ユーザのローカルコンピュータ上での各種文書の分類方法や，電子メールのメールクライアント上での分類方法は，個人の知識を表していると考えられ，このような知識を個人オントロジで表現する．そこで表現された知識を基にして，ウェブ情報検索時に，すでに持っている情報を排除したり，現在関心の高い情報の優先順位を上げるなどによって，ユーザを支援することが可能となる．
キーワード　個人オントロジ，Web 情報検索，知識処理

Supporting Web Information Retrieval Based on Personal Ontology

Hiroaki OHSHIMA† and Katsumi TANAKA††

† Department of Computer and Systems Engineering,
Graduate School of Science and Technology, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, 657–0013 Japan
† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606–8501 Japan
E-mail: †{ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　We have developed an ontology which depends on each person, called ”Personal Ontology”, for supporting Web Information retrieval. A person using a computer is dealing with a lot of documents on it. These documents and his/her methods to classify them include much information which expresses what kind of information the user has and how he/she thinks of things. But these information is not understood by a computer, so we developed Personal Ontology. Personal Ontology consists of (1) classified documents, (2) a hierarchical conceptual structure. Any kinds of documents like emails, MS-Word documents, pdf documents are classified in Personal Ontology. When the user uses a search engine on the Web, Personal Ontology can help it by removing already acquired information from the search results or by re-order the serarch results.
Key words　personal ontology, Web information retrieval, knowledge processing

1. はじめに

個人はローカルコンピュータでさまざまな文書を扱っている．ワープロ文書や，電子メール，ローカルコンピュータに保存した Web ドキュメントなどである．それらに記述されている情報はユーザに知識として吸収される．つまり，各種文書はある意味ではコンピュータ上で表現されている知識，ということができる．

しかし，いくら多くの有益な文書を持っていたとしても，そこに含まれている知識を，ウェブ検索に利用することはできな

い．例えば，すでに自分がどのような知識を手に入れていて，どのような目的でウェブ情報を探しているのか，ということを検索エンジンに伝えることはできないのである（図 1）．

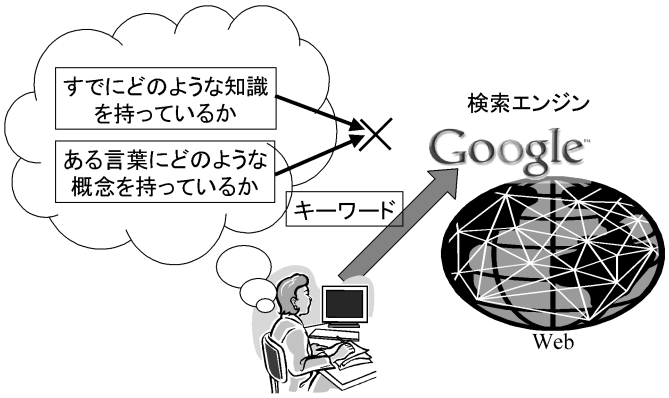

図 1　現在のウェブ利用環境

そのようなことを行うためには，知識をコンピュータが理解可能な状態にしなければならない．コンピュータに理解可能な知識を記述することは現在セマンティックウェブの分野で行われているようなオントロジを用いることで可能である．これまでも，あるグループ内で，個人個人もしくは全体の知識を表現する，ということは行われているが，その目的は，そのグループ内で知識化された情報の共有を行うことがほとんどであり，ウェブ情報を取得するために個人の知識が表現されることはなかった．

本研究では，まず，個人の知識を表す個人オントロジを作成する．個人オントロジは，

- 分類された文書群
- 概念の階層構造

から成り立つ．

分類された文書群の部分では，個人がコンピュータ上で扱っている文書を統一的に管理し，それらを内容や時間など，多視点から分類することによって得られる様々な知識，をコンピュータに利用可能な状態にする．

概念の階層構造の部分では，一般的なオントロジのように，概念どうしの関係などを記述する．

オントロジの作成や維持というものは非常に手間がかかることであるため，それをサポートするしくみも必要である．

まず，文書群を分類するにあたって，すでに存在しているカテゴリーに対して，特徴ベクトルから求められる類似性をもとに自動分類することと，オープンディレクトリ [13] の情報を利用した，一般的なカテゴリーにおける文書の自動分類によって，サポートを行う．

概念の階層構造の構築にあたっては，ユーザが簡単に概念の階層構造を作成可能なシステムを作成するとともに，シソーラスを用いることによって，一般的な言葉に関してはユーザは何も行わなくてもかまわないようにした．

さらに，このように表現された個人オントロジを用いて，ユーザがウェブ情報検索を行う際の支援として，

- 検索キーワードの意味拡張
- 検索結果から既得の情報を排除する
- 検索結果の再ランキングを行い，ユーザが現在興味を持っている情報の優先度を高くする

を行う．

以下，2 章で関連研究と関連事項について，3 章と 4 章で個人オントロジの概念について，5 章で個人オントロジを用いたウェブ情報検索支援について，6 章で本研究のまとめと今後の課題について述べる．

2. 関連研究と関連事項

2.1 個人の知識の表現

まず，個人の知識を表す研究について述べる．

Haystack [1], [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 RDF [8] で管理する．本研究では，Haystack と同様に，さまざまな個人的な情報を管理するが，蓄積された情報そのものをより効果的に利用しようとする Haystack とは異なり，新たな情報を獲得する際に，蓄積された情報を利用することを目的としている．

WorkWare++ [3], [4] 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や，ある事柄を知っていそうな人が誰であるかなどの情報を取得可能である．本研究では，知識としての情報管理を各ユーザが行うとともに，それらを現在の Web の利用のために利用することを目的としており，WorkWare++で行っている，グループによる情報共有や，蓄積された知識の獲得とは異なる．

Hyperclip [5] 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．本研究では，現在の Web にある情報を検索エンジンなどを用いて利用するときに，自分の既得の知識を利用できるようにすることが目的であり，Hyperclip とは目的が異なる．

湯川ら [6] 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．本研究と同様に，個人が所有する文書をもとに個人の概念を表しているが，その目的がピア・ツー・

ピア型のネットワークで共有することであり，Web 情報の自動取得を目的とする本研究とは異なる．

2.2 オントロジの記述

次に，オントロジ記述について述べる

オントロジの記述言語は，セマンティックウェブ [7] の分野で盛んに研究されており，RDFS [9]，DAML+OIL [10], [11]，OWL [12] などが挙げられる．RDFS は概念の階層構造の記述が可能で，DAML+OIL や OWL はそれに加えて，クラスやプロパティに対する制約や，概念やオントロジどうしのマッピングなどを記述することが可能となっている．

本研究の個人オントロジでは，OWL の記述能力の範囲をサポートしながら，さらに分類された文書群を扱うことで，より簡単に個人の知識を表現することを目指す．

3. 個人オントロジの概要と概念の階層構造

3.1 個人オントロジによって表現されるもの

個人が持っている知識は，コンピュータには理解できる形になっておらず，ユーザはコンピュータを利用する際のノウハウという形で，それらの知識をコンピュータに伝えている．例えば，ウェブ上で検索エンジンを利用する際に，ユーザは自分が求める情報を得るためにはどのようなキーワードが適切かを考え，そのキーワードによって得られた検索結果から，すでに自分が持っている情報や，含まれているキーワードが自分の考えている使われ方ではない情報などを排除して，最終的に求める情報を得る，ということを行っているのである．

そのような，個人がすでに持っている知識や，ある概念に対して個人がどのような考え方を持っているか，というものを表現するものが個人オントロジである．

3.2 個人オントロジの構成

個人オントロジは，

- 分類された文書群
- 概念の階層構造

という二つの部分によって成り立つ．これらによって，オントロジが持つべき機能である，

- ものごとの分類体系
- 推論ルール

という二つの役割を表す．

分類された文書群については 4 章で詳しく述べる．以下では，まずオントロジがどのように記述されるかについて述べる．さらに，概念の階層構造による知識の表現についてと，シソーラスによる個人オントロジの情報の補完について述べる．

3.3 個人オントロジの記述方法

個人オントロジの記述は，基本的に OWL の表現にならったものとする．

3.3.1 クラスとインスタンス

まず，オントロジ内では，すべての概念はクラスやそのインスタンスとして表現される．例えば，一本の鉛筆があったとすると，まず,「鉛筆」クラスが作成され，そのインスタンスとしてその鉛筆が表現され，また,「鉛筆」クラスは「筆記用具」クラスのサブクラスであり,「筆記用具」クラスのサブクラスには「ボールペン」クラスも存在する，というように，様々な概念が表現されるのである．以下は，OWL で上記のことを記述した例である．

```
<owl:Class rdf:ID="筆記用具"/>
<owl:Class rdf:ID="鉛筆">
    <rdfs:subClassOf rdf:resource="#筆記用具"/>
</owl:Class>
<owl:Class rdf:ID="ボールペン">
    <rdfs:subClassOf rdf:resource="#筆記用具"/>
</owl:Class>
<鉛筆 rdf:ID="私のお気に入りの 1 本の鉛筆" />
```

OWL では，すべてのクラスの基になるクラスとして Thing クラスを持つ．個人オントロジでも，OWL の表現にならい，Thing クラスを持つものとする．

3.3.2 プロパティ

プロパティはあるクラスにおいて，別のクラスとの関係やリテラルなデータとの関係を表したものである．

プロパティには，

- データプロパティ
- オブジェクトプロパティ

の 2 種類がある．例えば,「書籍」クラスには，データプロパティとして「書籍価格」というリテラルなデータを持つプロパティが，オブジェクトプロパティとして「著者」という,「人」クラスと関連づけられたプロパティが，それぞれ作成されると考えられる．以下は，OWL で上記のことを記述した例である．

```
<owl:ObjectProperty rdf:ID="著者">
    <rdfs:domain rdf:resource="#書籍"/>
    <rdfs:range rdf:resource="#人"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="書籍価格">
    <rdfs:domain rdf:resource="#書籍" />
    <rdfs:range  rdf:resource="&xsd;positiveInteger" />
</owl:DatatypeProperty>
```

3.3.3 ラ ベ ル

通常，オントロジはセマンティックウェブという，すべてコンピュータで理解可能な世界で利用することが想定されており，rdf:ID によって認識されれば良い．しかし，それは逆に通常のウェブや人間には理解できない形である．本研究では，ユーザの知識を個人オントロジで表現して，それを通常のウェブで利用することを考えているため，何らかの形で個人オントロジの中の情報を外部で利用できる形にしなくてはならない．本研究では，個人オントロジ内のすべてのクラス，プロパティ，インスタンスに，RDFS の label 要素を付加することで，オントロジ外での最低限の表現とする．

3.4 個人オントロジにおける概念の階層構造

3.4.1 一般的なオントロジとの違い

一般的に，オントロジは，複数のユーザである分野の概念に対して共通した認識を持つために，作成，利用される．しかし，本研究で作成する個人オントロジは，完全に個人的に利用するためのオントロジである．そのため，共通概念よりもむしろ，

ユーザの好みや，造詣の深い分野に関する，より細かい概念の階層構造を記述することが主となると考えられる．

また，オントロジには，オブジェクトデータベース的な部分があり，個人オントロジでは，例えば，自分のゴルフのスコアや，蔵書の情報などを管理することが考えられる．

3.4.2 個人オントロジの実現

個人オントロジの作成のため，現在，オントロジエディタの試作システムを作成中である．本システム上では，クラスの作成，プロパティの作成，インスタンスの作成，など，RDFS でサポートされているレベルのオントロジの記述を完全にサポートしており，今後，制約などの OWL でサポートされている記述を行えるよう改善を行っていく．

図 2 で示したのは，クラスを作成する画面である．本システムでは，個人オントロジの情報は，現在はすべて RDB で保存管理している．

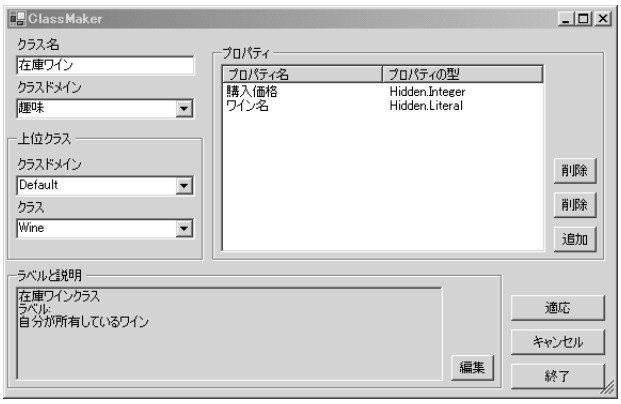

図 2　試作システムのイメージ

4.　分類された文書群による知識

一般的なオントロジでは，普通の文書をそのまま取り扱うようなことはなく，必ず，概念レベルで扱う．しかし，個人レベルで概念レベルのオントロジを作成することは，個人レベルにおいては非常に困難なことである．そこで，個人オントロジでは文書を分類することから，知識を得る手法を提案する．

4.1　個人が保有する文書群から理解できること

ここでは，個人が保有する文書群から，どのような知識が抽出できるかということと，知識の抽出のためには，どのような管理をすべきかということについて検討する．

個人が扱う文書には，

- ローカルに保存されたウェブドキュメント
- ワープロドキュメント
- PDF ドキュメント
- プレゼンテーションファイル
- 電子メール

などがある．これらの文書の管理は，ファイル管理システムやメールソフト上などで行われていたわけだが，その際に，ファイル管理システム上では，例えば，内容ごと，プロジェクトごと，一定期間ごと，などの基準に従ってディレクトリを作成し，ファイルの分類・管理が行われることが多く，メールソフト上では，送信者ごとにメールが分類され管理することがよく行われる．

そのように分類され管理された文書を，他人が見たときには，その人がどのような考えや知識を持っているか，ということをある程度判断することができる．つまり，そこには何らかの知識が存在するといえる．

そこで，このような，個人が保有する文書群に含まれている内容とその分類方法を個人オントロジに取り込むことで，様々な知識を得る．

このような文書群から統合的な知識を取り出すためには，個人オントロジにおいて

- 一元的な文書群の管理
- 多視点による文書群の分類

を行う必要がある．

まず，一元的な文書群の管理であるが，これは，様々な形式の文書として保存されている知識を統一的に扱うために必要なことである．これまで，さまざまな文書はワープロやメールクライアントなどの専用のソフトで管理されてきたが，そこからテキスト情報を抜き出すことによって，一元的な管理を実現する．

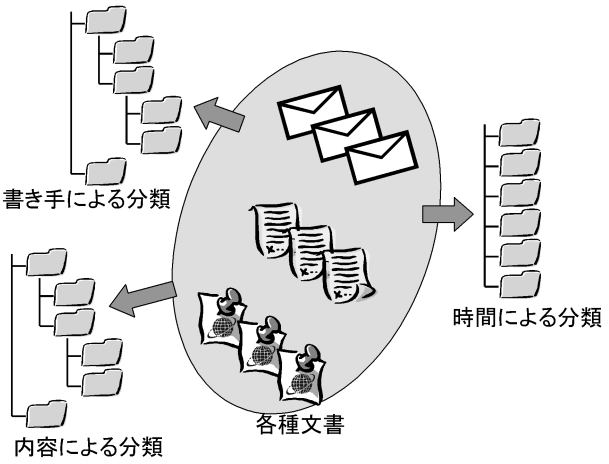

図 3　多視点による文書群の分類

次に，多視点による文書群の分類であるが，図 3 のように，文書群は本来，内容や時間など様々な視点から分類が可能である．例えば，内容による分類，文書の作成日時や更新日時による分類などはほとんどの文書群において同時に行える．また，例えば，メールからは送受信の時間や送信者の情報などをメタ情報として抽出することが可能であり，これらも分類を行う際の基準にすることができる．また，自分のライフスタイルに合わせた文書群の分類ということ考えられる．例えば，仕事のための文書と，娯楽のための文書というのは，異なり，これも多視点による文書群の分類ということで，表現可能になる．

4.2　分類された文書群のためのクラス

ここでは，分類された文書群を，個人オントロジのクラス表現の中でどのように実現するかについて述べる．

4.2.1 Document クラス

まず，文書を扱う基本的なクラスとして，Document クラス

を用意する．ユーザは新しく文書を手に入れると，その文書を個人オントロジに登録する．その際に，個人オントロジ内では，文書を Document クラスのインスタンスとして扱うのである．

ワープロ文書や電子メールなどを扱うためには，Document クラスからサブクラスを作成し，WordDocument クラス，EmailDocument クラスなどとする．それぞれには，付加可能なメタ情報をプロパティとして付加する．例えば，電子メールから得られる，送信者や送信日時といったメタ情報が，EmailDocument クラスのプロパティとして付加される．

4.2.2 DocumentDirectory クラス

次に，文書を分類してまとめるクラスとして，DocumentDirectory クラスを用意する．DocumentDirectory クラスは，ファイルシステムにおけるディレクトリ（フォルダ）に相当する役目を持つ．

DocumentDirectory は，ファイルシステムのディレクトリと同様に，ツリー構造を構成する．ただし，多視点による文書群の分類を行うため，個人オントロジには，複数の DocumentDirectory のツリー構造が作られることになる．

4.3 文書群の個人オントロジへの取り込み

個人オントロジを生成する際に，ファイルシステムやメールソフトから，分類情報もそのまま取り込むことで，ユーザがわざわざ個人オントロジのために分類を行う必要は無くなる．

多視点による文書群の分類を行うとはいえ，個人オントロジで扱う前の文書群の分類方法というものは，当然，重要視すべきである．個人差はあるが，文書の分類において重視されるのは，タスクやイベントごとによる分類であり，次いで，時間による分類であることが多い．タスクやイベントによる分類は，次節で述べるような自動分類では行いにくい分類であり，有意であると考えられる．

次節では個人オントロジ上での文書群の分類手法について述べる．

4.4 文書群の分類手法

ファイル管理システムや，メールクライアントで扱っている文書を個人オントロジに取り込む際にはその分類構造を DocumentDirectory として取り込み，読み込んだ文書から作成された Document をその DocumentDirectory と関連づけることでもとの分類を個人オントロジ内で維持することができる．

もとの分類の他にも，例えば，PDF 文書は内容ごとに分類していて，メールは送信元ごとに分類していたときに，個人オントロジ内では，メールを PDF 文書で使っていた内容ごとの分類方法で分類することができる．

個人オントロジ内で Document を DocumentDirectory に分類する方法は，

- ユーザによる手動分類
- 既存のカテゴリーへの自動分類
- オープンディレクトリを利用した自動分類
- プロパティの評価による自動分類

が存在する．それぞれについて説明する．

4.4.1 ユーザによる手動分類

これは，ユーザが自分で内容やメタ情報をもとにして，手動で該当する DocumentDirectory に分類する方法である．

4.4.2 既存のカテゴリーへの自動分類

ある程度の文書が各 DocumentDirectory に格納されるようになると，文書の特徴ベクトルの類似度を用いた手法によって自動分類することが考えられる．本研究では，特徴ベクトルの作成に TF-IDF 法を用いる．

文書 d におけるある単語 t の出現回数が TF 値で，$TF(d,t)$ と表される．IDF 値は，複数の文書が存在するときに，ある単語 t_i に対して，

$$IDF(t_i) = log(\frac{\text{総文書数}}{\text{単語 } t_i \text{が出現する文書数}})$$

として定義される．

文書 d における特徴ベクトル $F(d)$ を，出現単語 $t_1, ..., t_n$ を各基底とする重要度 $T_1, ..., T_n$ として求めると，

$$F(d) = (T_1, ..., T_n) = \frac{1}{N}(f_1, ..., f_n)$$

として表される．ここで，$f_j, j \in (1, ..., n)$ は，文書 d における単語 t_j の TF 値，N は d の総単語数を表す．すなわち，

$$T_j = TF(d, t_j)$$

である．

次に，ある DocumentDirectory によって形成されるカテゴリーの特徴ベクトルを求める．カテゴリー c 内に含まれている文書を $d_1, ..., d_p$ とすると，カテゴリー c における特徴ベクトル $Fd(d)$ を，出現単語 $tc_1, ..., tc_q$ を各基底とする重要度 $Tc_1, ..., Tc_q$ として求めると，

$$Fc(c) = (Tc_1, ..., Tc_q) = \frac{1}{Nc}(fc_1, ..., fc_q)$$

として表される．ここで，$fc_j, j \in (1, ..., q)$ は，カテゴリー c における単語 tc_j の TF 値と，カテゴリーごとをひとつの文書と見なしたときに得られる IDF 値を掛け合わせたもの，Nc は c に含まれている文書の総単語数を足しあわせたものを表す．すなわち，

$$Tc_j = IDF(t_j) \cdot \sum_{i=1}^{p} TF(d_i, t_j)$$

である．

このようにして得られた，文書の特徴ベクトルと，カテゴリーの特徴ベクトルから，その文書の各カテゴリーに対する類似度をコサイン相関値を用いて計算する．すなわち，

$$Similarity(F, Fc) = \frac{F \cdot Fc}{\|F\| \cdot \|Fc\|}$$

が類似度となる．

この類似度が最も高いカテゴリーに対して文書の分類を行うことで，自動分類が可能となる．このとき，多視点による分類のため，DocumentDirectory のツリー構造は複数存在するが，それぞれにおいて，このような分類が可能である．

4.4.3 オープンディレクトリを利用した自動分類

既存のカテゴリーに，すでに文書が分類されている場合は，上述した自動分類を行うことができる．しかし，あまりこれま

で知らなかったような分野の文書を手に入れたときは，どこに分類すればよいかわからないことも多い．

そのようなときのために，DocumentDirectory のツリー構造のひとつとして，オープンディレクトリのカテゴリー分類を利用する．

オープンディレクトリには，カテゴリーの情報として，下記ような情報が記述されている．

```
<Topic r:id="Top/World/Japanese/ビジネス/薬品・バイオテクノロジー/薬品">
  <catid>1150881</catid>
  <d:Title>薬品</d:Title>
  <d:Description>
    薬品・医薬品やその業界に関するサイトを掲載します.
  </d:Description>
  <lastUpdate>2003-10-28 20:39:58</lastUpdate>
  <narrow r:resource="Top/World/Japanese/ビジネス/薬品・バイオテクノロジー/薬品/卸売"/>
  <narrow r:resource="Top/World/Japanese/ビジネス/薬品・バイオテクノロジー/薬品/小売・調剤薬局"/>
  <narrow r:resource="Top/World/Japanese/ビジネス/薬品・バイオテクノロジー/薬品/団体"/>
</Topic>
```

また，各カテゴリーに登録されたページの情報として，下記のような情報が記述されている．

```
<ExternalPage about="http://www.suzuken.co.jp/">
  <d:Title>スズケン</d:Title>
  <d:Description>
    医療用医薬品，診断薬，医療機器，医療材料,
    医療食品，大衆薬などを扱う.
  </d:Description>
</ExternalPage>
```

オープンディレクトリの中の日本語を対象とした部分には，約 9 千のカテゴリーと約 8 万件のページが登録されており，ありとあらゆる分野についての文書を扱うことができる．

このオープンディレクトリによる分類においても，既存のカテゴリーへの自動分類と同様の方法で自動分類を行うことができる．

4.4.4 プロパティの評価による自動分類

プロパティの評価による自動分類とは，メールソフトで行われているメールの自動振り分けなどと同様に，あらかじめ，分類のルールを記述しておいて，そのルールに該当する文書を決められたカテゴリーに分類する方法である．

この方法による分類には，静的なものと，動的なものが存在する．

静的なものの例としては，

- メールの送信者ごとの分類
- 文書の作成者ごとの分類

などがあげられる．

動的なものとしては，時間による分類があげられる．例えば，最近 2 週間で新たに得た文書を自動的に分類して，カテゴリーを作成することが可能である．

4.5 分類された文書群からの知識抽出

このように作られた，個人オントロジの分類された文書群の部分から，どのような知識を実際にコンピュータ上で利用できるかについて述べる．

4.5.1 ある語に対する概念ベクトル

概念ベクトルとは，個人が保有する文書の中での語の共起関係に基づいて作られるベクトルで，ユーザのある語に対してどのような考えを持っているかを表す．

個人オントロジで扱う文書全体の中で，ある語 k が出現する文書の特徴ベクトルを合成した，

$$Fk(k) = \frac{1}{n} \cdot \sum F(d_u)$$

（ただし，d_u は語 k が出現する文書，n は語 k が出現する文書の総数,）が概念ベクトルである．

例えば，スポーツという語とテニスという語が共起する文書を多く持っている人にとって，スポーツとテニスは関連が深いものである，ということが理解できる．この概念ベクトルによって，検索キーワードの意味的な拡張や検索結果のフィルタリングの精度を上げることが行える．

4.5.2 カテゴリーの特徴ベクトル

ユーザにとって，あるカテゴリーに属するような文書を探す，ということは頻繁に行う行動である．個人オントロジでは，カテゴリーの特徴ベクトルを得ることができるため，ある文書が特定のカテゴリーに属するかどうかを，類似度を計算することにより判断することができる．

また，カテゴリーの差を見つけることも，対象となる複数のカテゴリーで TF-IDF 値を用いて特徴ベクトルを作成することで可能になる．これによって，例えば，時間によるカテゴリの時間変化に伴うユーザの扱う語の変化などをとらえることが可能である．

4.5.3 既得文書かどうかの判断

ある文書の情報がすでに持っているかものかどうかを判断することも，文書どうしの特徴ベクトルの類似度を計算し，ある閾値以上であるかどうかを調べることによって可能になる．

5. 個人オントロジを用いたウェブ情報検索支援

ユーザがウェブ情報を検索する際には，ウェブ上の検索エンジンに対して，何らかのキーワードを入力し，返ってきた結果を見て，自分が求めていた情報かどうかを判断する，ということを行っている．そこでは，ユーザがすでにどのような情報を持っているか，ユーザが用いたキーワードに対してどのような概念を持っているか，ユーザが今どのような情報に興味を持っているか，といったことを検索エンジンに伝えることができていないのである．

個人オントロジを用いることによって，ウェブ情報検索において，

- キーワードの意味的拡張
- 既得情報の除去（フィルタリング）
- 興味分野の提示優先度向上

といったことが行えるようになる．

5.1 キーワードの意味的拡張

ユーザは情報検索の際に，キーワードを入力するが，各個人によって，そのキーワードに対して持っている概念は異なると考えられる．例えば,「パソコン」という言葉に対して，ある人は真っ先にラップトップコンピュータをイメージするが，別の人はデスクトップコンピュータをイメージする．このような概念の違いを検索エンジンに伝えることはできなかったが，キーワードを個人オントロジを用いて拡張することによって可能になる．

また，キーワードを指定しなくても，文書である Document や，分類である DocumentDirectory をキーとして検索する方法も考えられる．

キーワードを拡張したり，Document や DocumentDirectory から，検索キーワードの候補をあげる方法としては，以下のようなものが利用できる．

- キーワードの概念ベクトル（共起語）
- 概念の階層構造（シソーラス）
- Document や DocumentDirectory の特徴ベクトル

また，キーワード候補としてあまりにも一般的であったりする語を取り除くためには，以下のような方法を用いることができる．

- IDF 値
- ストップワード（不要語リスト）

5.1.1 実　　験

実際に，11 種，約 1000 通のワインに関するメールマガジンから個人オントロジを作成し，キーワードの意味的拡張の実験を行った．

「チーズ」という語について共起度を求めると以下のような結果になった．

共起回数	語
127	ワイン
98	味わい
94	イタリア
94	香り
93	ボルドー

また，シソーラスから「チーズ」の広義語と狭義語を求めると，以下のような語が得られる．

広義語	狭義語
発酵食品	エダムチーズ
乳製品	エメンタール
スプレッド	カステルマ―ニョチ―ズ
ペースト	カッテージチ―ズ
	カマンベ―ル
	クリームチ―ズ
	（他多数）

まず，個人の「チーズ」という語に対するイメージが概念ベクトルから得られるため，それらの語を検索キーワードとして用いる．さらに，

- より細かい概念の情報を得たい
- より広い概念の情報を得たい

という要求に応じて，それぞれ，狭義語，広義語を AND 条件でキーワードに付加した．ただし，共起語からは，絞り込みすぎないように上位 2 位のみを用いた．すなわち，

- ワイン AND 味わい AND (発酵食品 OR 乳製品 OR ...)
- ワイン AND 味わい AND (エダムチーズ OR エメンタール OR ...)

というキーワードを作成し，Google に対して検索行動を行った．

結果は，広義語を用いた場合については，上位 10 件で，下記のような結果が得られた．

- チーズやヨーグルトなどの乳製品を扱う商品一覧のページ 7 件
- ワインの販売のページ 1 件
- 日本酒の販売のページ 1 件
- ワインとチーズを含む様々な商品一覧のページ 1 件

狭義語を用いた場合については，上位 10 件で，下記のような結果が得られた．

- ワイン好きの個人が日記的なページ 4 件
- ほとんど無関係なページ 2 件
- ワインとチーズのマリアージュについてまとめたページ 1 件
- チーズについてまとめたページ 1 件
- ワイン販売のページ 1 件
- チーズ販売のページ 1 件

広義語を用いた場合には，ある程度まとまった結果となり，乳製品の商品を扱うページが現れた．狭義語を用いた場合は，個人の日記的なページが得られたが，これらのページは普通に「ワイン」や「チーズ」といった語による検索ではなかなか上位に現れないページであり，実際，興味深く読むことができた．

5.2 既得情報の除去

検索エンジンから返された検索結果には，すでに自分が持っている情報が含まれていることがあり，これは新しい情報を取得したい場合には不要な情報である．そこで，そのような既得情報をフィルタリングによって除去する．

まず，検索結果のウェブページの特徴ベクトルを求め，既存の文書との特徴ベクトルとの類似度を計算し，ある閾値以上であれば取り除く，という作業によって実現可能である．

5.3 興味分野の提示優先度向上

ユーザが現在，どのような興味を持っているかということによって，求める情報は変わってくるはずである．そのようなユーザの関心は，

- 時間による分類の中で最近のカテゴリー
- 最近多くの文書が分類されているカテゴリー

といったことから，どのようなカテゴリーに属するものかというかたちで，判断することが可能となると考えられる．

時間による分類の中で最近のカテゴリーの特徴ベクトルの変化からは，最近になってよく用いられるようになった語という

ものも得ることができる．これは，キーワードの意味的拡張にも利用可能である．

検索エンジンから返された検索結果は，検索エンジンの判断によって，ランキング付けされているが，最近ユーザが興味を持っていると考えられるカテゴリーに属すると類似度の計算から判断される情報は，ランキングの上位に再配置することでユーザにとって有用と考えられる情報にアクセスしやすくなる．

6. おわりに

本研究では，個人が持っている知識を表現するための，個人オントロジを提案した．個人オントロジは，

- 分類された文書群
- 概念の階層構造

から成り立ち，これらによって，個人がどのような知識をすでに持っているか，ある概念に対してどのような考えを持っているか，ということをコンピュータ上で表現する．これらは，ユーザがローカルコンピュータ上に保有している，ワープロ文書や，電子メールなどの各種文書と，その分類方法を取り込むことによって作成される．また，オープンディレクトリプロジェクトの情報や，シソーラスの情報を利用することにより，個人オントロジの作成や管理の際にかかるユーザの負担を軽減する手法を提案した．

さらに，個人オントロジを用いて，ウェブ情報検索支援を行うことを提案した．ユーザが行った検索の結果に対して，

- 検索キーワードの意味拡張
- 検索結果から既得の情報を排除する
- 検索結果の再ランキングを行い，ユーザが現在興味を持っている情報の優先度を高くする

を行うことによって，ユーザにとってより有用な情報を提示することが可能になる．

今後は，システムの実装を進め，検証を行いたい．

謝　　辞

本研究の一部は，平成 15 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：15017249，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] E.Adar, D. Karger and L. Stein, ”Haystack: Per-User Information Environment”, Proc.1999 Conference on Information and Knowledge Management, pp.413-22, 1999.
[2] Haystack: http://haystack.lcs.mit.edu/
[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] 内野寛治，津田宏，松井くにお，WorkWare:WEB を用いた文書の時間順整理の試み，情報処理学会 研究報告「情報学基礎」No.051, 1998.
[5] Hiroyuki Sato, Yutaka Abe and Atsushi Kanai, ”Hyperclip: a Tool for Gathering and Sharing Meta-Data on Users’ Activities by using Peer-toPeer Technology”, WWW2002 Workshop on Real world RDF and Semantic Web applications, 2002.
[6] 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.
[7] Semantic Web ホームページ (W3C): http://www.w3.org/2001/sw/
[8] W3C Resource Description Framework (RDF) (W3C): http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
[9] W3C RDF Schema (W3C): http://www.w3.org/TR/rdf-schema/
[10] W3C DAML+OIL (W3C): http://www.w3.org/TR/daml+oil-reference
[11] DAML.org: http://www.daml.org/
[12] OWL Web Ontology Language Reference (W3C): http://www.w3.org/TR/owl-ref/
[13] Open Directory Project: http://dmoz.org/