人工知能学会研究会資料
SIG-SWO-A303-10

Web コンテンツの機能に着目した検索結果の構造化に関する提案
－アンカテキストを用いた Web コンテンツ形式の推定－

川前 徳章　　高橋 克巳

ＮＴＴ情報流通プラットフォーム研究所　〒180-8585　東京都武蔵野市緑町 3-9-11

E-mail:　{kawamae.noriaki,takahashi.katsumi}@lab.ntt.co.jp

あらまし　本研究はユーザの情報検索を効率化することを目的として、情報検索システムの検索結果を構造化するための手法を提案する。一般に Web 空間における大半のコンテンツは構造化されておらず、その機能も異なっている。それにも関わらず、従来の情報検索システムは、構造化されたデータの検索を前提としたキーワードマッチング技術を用い、ユーザの入力したクエリに対し、そのクエリを含むデータの一覧をランキングという一次元の属性で表示していた。その結果として、それら Web コンテンツの集合である情報検索システムの検索結果もまた構造化されないため、ユーザは検索結果から必要なコンテンツを探す負担が生じ、情報検索は非効率になるという問題がある。本研究は、この問題を解決するために、検索結果の構造化を実現する手法を提案する。この提案では、Web コンテンツのカテゴリ、機能、形式、作成者、作成日時の五つの属性を持つように構造化し、これらの属性を抽出する。提案手法は、リンク元の Web コンテンツはリンク先の Web コンテンツに対してのアンカテキストを付けている Web 空間のハイパリンク構造に着目し、Web コンテンツの形式の推定を行う。この形式及び Web コンテンツの作成者、作成日時の抽出によって検索結果が構造化されるため、自身の目的に合った Web コンテンツの情報検索が効率化されるだけでなく、Web コンテンツを活用した情報抽出が容易になることが期待できる。

キーワード　情報検索，情報抽出，非構造化データ，ハイパリンク，アンカテキスト

Construction of Search Results Based on the Web Content's Function
－Estimation of Web Content's Type from Anchor Text－

Noriaki KAWAMAE Katsumi TAKAHASHI

NTT Information Sharing Platform Laboratories 3-9-11,Midori-cho Musashino-shi Tokyo 180-8585 Japan

E-mail:　{kawamae.noriaki,takahashi.katsumi}@lab.ntt.co.jp

Abstract　We propose the novel method can construct search results to improve user information retrieval. Existing search engines return search results including user query as the answer of user query by the keyword matching method. The most of Web contents are not constructed data and variety function, these engines utilize the keyword matching method performed for constructed data nevertheless. Therefore the search results, collected of web contents under the same query, are also non constructed and leading to prevent user from information retrieval activity. To solve this problem, we propose the novel method that can construct search results. Our proposed construction is composed of five factor web content category, function, type, author and date of update. Our proposed method focus on the hyperlink structure like Web, Web contents place an anchor text and refer to other contents, and estimates the content's type by using the anchor text, and construct search results based on their function. Because the constructed search results help user to seek contents meet user need, this method can not only improve user information retrieval but also simplify to extract information from web contents.

Keyword　Information Retrieval，Information Extraction, Non-structured Data. Hyperlink, Anchor Text

1. はじめに

現在、Web 空間に存在するコンテンツはサイト単位では 360 万、ページ単位では 8 億ぐらいあると見積もられている。これらのコンテンツは言語、対象とするユーザや提供する内容やサービスといったコンテンツとしての機能も異なっている。その違いによって様々なユーザの検索の目的を満たしている。ユーザの検索の要求もまた言語、内容やその専門性や目的なども異なっている。情報検索システムは我々がこのようなコンテンツから構成される Web 空間から必要なコンテンツを捜すために用いられている。

しかし、現在の情報検索システムは大部分のユーザの検索の目的や Web コンテンツの機能を区別して処理していないため、ユーザの検索効率が阻害されている問題がある。データベースシステムにおいてデータは構造化されているため、検

索結果は網羅性が高いが、Web コンテンツは構造化されていないため網羅性が低くなる。さらにその機能も異なっているので検索結果にはユーザが必要とするコンテンツとそうでないコンテンツが含まれる問題がある。データの構造が異なるにもかかわらず、現在の情報検索システムはデータベースと同じ技術を用いているため、検索結果もまた構造化されていない。例えば、ユーザが“ Linux のセキュリティ対策について ”の技術情報を Web から検索すると、検索結果には目的の技術情報だけでなく、書籍、製品、セミナーの案内なども含まれてしまう。ユーザは必要な Web コンテンツを探すためにこれらの検索結果を何度も見直さなければならずユーザの情報検索が非効率的になる。

情報検索におけるユーザの検索効率を上げるために、我々は検索結果の構造化を実現する手法を提案する。提案する構造化手法は Web コンテンツをカテゴリ、機能、形式、作成者、作成日時の五つの属性を用いて表現する。この構造化に用いる五つの属性のうち、本稿では Web コンテンツの形式、作成者、作成日時の三つの属性を抽出する手法を提案する。まず、Web コンテンツの形式を抽出するために、Web 空間のハイパリンク構造を利用する。我々はハイパリンク構造において、Web コンテンツのリンク元がアンカテキストにおいてそのリンク先の Web コンテンツの機能、形式や内容を要約していることに着目し、それを利用して Web コンテンツの形式を推定し、それを機能別に整理する。また、Web コンテンツの作成者や作成日時は直接ファイルから取得することができる。

本稿では、この構造化を実現する為に、以下の提案を行う。

・検索結果の構造化を実現する為のシステムの構成とその処理方法

・検索結果の構造化に必要な情報の抽出と処理方法

・アンカテキストの抽出方法

・各コンテンツの形式を推定する為のコンテンツ形式 DB の作成とその利用方法

次にこの提案をシステムとして実装し、その有効性を実験により確認した結果を示す。

本論文の構成は、二章で既存研究の問題点について触れ、三章で提案手法を実現するシステムの概要と技術の詳細について触れ、四章でその手法の評価実験を行い、最後にまとめとする。

2. 既存研究

情報検索システムの最終目的の一つは、ユーザの情報検索要求に直接回答することを実現することである。その実現のために様々なアプローチがなされている。そのアプローチの一つに Semantic Web などの情報抽出の自動化手法がある。我々は情報抽出の自動化の前段階として、情報検索システムがユーザの情報に対する要求を意味的に処理することと、検索対象のコンテンツを意味的に処理することが必要であると考える。

ユーザの情報に対する要求を意味的に処理する研究に[3]がある。検索対象のコンテンツを意味的に処理するには、PLSA[5]などのアプローチがある。これらの手法がコンテンツ内の単語の出現頻度を用いているのに対し、Web 空間の特徴であるハイパリンク構造に着目した研究には[1]がある。[1]はアンカテキストを用いてリンク先の Web コンテンツのカテゴリを決定するものである。コンテンツのカテゴリでなく、その情報を分析したものに[7]がある。この研究では、Web コンテンツの分類に分類ルールを作成している。

情報抽出の自動化のためにはこれらの処理を自動化し、有機的に結合することが必要である。またユーザの検索目的に合ったコンテンツの発見のためには、コンテンツの持つ機能について明らかにしておく必要がある。

3. アンカテキストを用いた Web コンテンツの機能分析

3.1 システムの概要と検索結果の構造化

提案手法を実装するシステムは、ユーザと従来の情報検索システムの間に位置し、ユーザのクエリと情報検索システムの検索結果を利用して検索結果の構造化を行うものである。システムの概要を図１に示す。

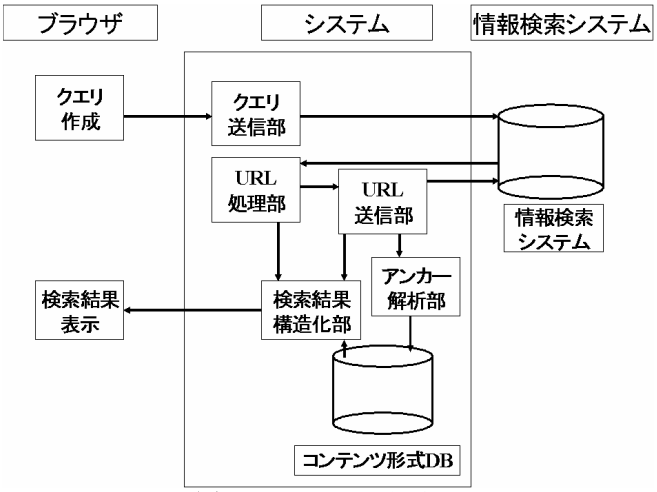

図１．システムの構成図

システムはブラウザからユーザのクエリを受け取り、クエリ送信部から情報検索システムに送信する。クエリに対する結果を元に URL 送信部、URL 処理部、アンカ解析部、コンテンツ形式 DB と検索結果構造化部が処理を行うことで検索結果の構造化を行う。以上の処理によって Web コンテンツが単にクエリを含むという属性だけでなく、コンテンツの機能、形式、作成者と作成日時という属性を持つことで、検索結果は、コンテンツの機能＞形式＞作成者＞作成日時という項目で構造化ができる。

3.2 システムの構成技術

クエリ送信部：ブラウザからユーザのクエリを受け取り、クエリを含む Web コンテンツを要求として情報検索システムに送信する。

URL 処理部：情報検索システムの検索結果に含まれる URL を DB に格納し、URL 送信部に URL のリストを渡す。その際にリンク先かリンク元かのフラグを立てる。

URL 送信部：URL 処理部より渡されたリストを受け取り、リンク先のフラグがあれば HTTP サーバと情報検索システムに送信する。HTTP サーバに対して GET コマンドを送信し、ヘッダ情報に含まれる Last-Modified をそのコンテンツの作成日時とする。作成者が抽出できない場合はどのドメイン名を作成者として代用する。情報検索システムに対しては、着目する URL をリンク元として含む Web コンテンツを要求として情報検索システムに送信する。リンク元のフラグがあれば HTTP サーバに URL を送信する。その結果はリンク元の URL と併せてアンカ解析部に渡す。

アンカ解析部：URL 送信部から渡された結果から、リンク元の URL のアンカテキストを形態素解析によって処理する。形

態素解析の後、各 URL をキー、その URL を要素としたリストを作成し、コンテンツ形式 DB を参照し、各 URL の Web コンテンツの形式の判定処理を行う。
コンテンツ形式 DB：コンテンツ形式 DB は各単語、URL のトークンの各コンテンツ形式における確率値を保持している。この値を用いてアンカ処理部から渡されたリストに対して各機能に対する確率値を算出して最も高くなるコンテンツ形式を、その URL のコンテンツ形式として URL に付加して検索結果構造化部に渡す。
検索結果構造化部：各処理部から渡された情報（URL とそのコンテンツの形式、作成者、作成日時）を元に検索結果を構造化してブラウザ側に渡す。

3.3　ユーザの検索の目的とコンテンツの形式・特性・機能

3.2 の示したシステムにおいて、提案手法を実現するのはアンカ解析部の処理とコンテンツ形式 DB である。Web コンテンツの機能とは各コンテンツが持つ特性によって、ユーザの検索の目的合致するものであり、ユーザ検索の目的や情報源選択の基準となるものである。

Morrison[5]らは Web におけるユーザの検索の目的を次のように分類している。ただし、これらのユーザは必ずしも情報検索システムを用いた検索を前提としていない。

1. 発見
   a. ダウンロード　2%
   b. 事実の確認　15%
   c. 文書の取得　6%
   d. 製品　2%
2. 比較・選択　51%
3. 理解　24%

ユーザがこれらの目的を実現するために得る Web コンテンツには、次の２通りの機能が存在すると考えられる；
・Web コンテンツそのものを得ることがユーザの目的を満たす機能、
・Web コンテンツが示す書籍、製品、セミナーの案内などの Web 空間に限定されない情報によって目的を満たす機能
我々は Morrison らの定義する目的の「.発見」に関しては前者の機能が必要であり、「比較・選択」及び「.理解」に関しては後者の機能が必要であると考えられる。なぜなら後者において、ユーザは実際に幾つかのサイトを訪問することとなるが、もし一つのサイトでそれを完結しようとすれば、それは検索の目的の「発見」

表 1.　コンテンツの形式・特性・機能とユーザの検索の目的との対応関係

クエリとの関係	一次的関係						二次的関係	
コンテンツの形式	論文	解説	ニュース	掲示板	日記	サイト	リンク集	ガイド
作成者	著者	筆者	記者	掲示板の投稿者	コンテンツの作者	サイトの管理者	管理者	管理者
内容の完結性	○	○	○	×	×	△	○	○
内容の客観性	○	△	○	-	-	-	-	-
検索の目的								
1.　発見								
a ダウンロード						○		
b 事実の確認	○	△	○					
c 文書の取得	○	○	○	○	○	○		
d 製品						○		
2.　比較							○	○
3.　理解							○	○

に包含されると考えられるからである。本稿では Web コンテンツ自身がユーザの情報検索の目的を満たす機能を持つ場合、クエリとの関係を一次的関係、Web 空間以外に存在するコンテンツあるいはイベント、行事や場所など、ユーザの情報検索の目的を満たすもの照会する機能を持つ場合、クエリとの関係を二次的関係と呼ぶ。一次的関係であるコンテンツには論文、解説記事、ニュース、掲示板、日記、サイトがあり、二次的関係であるコンテンツにはリンク集、ガイドなどがある。ここで、ガイドとはリンク集と異なり、Web コンテンツ以外のものを参照しているものを指す。表１においてコンテンツの形式・特性とその機能とユーザの検索の目的との対応関係を示す。なお、表１における検索の目的のダウンロードはコンテンツそのものダウンロードでなく、アプリケーションや実行ファイルなどのダウンロードである。コンテンツの特性のうち、完結性は、そのコンテンツ自身で機能が完結しているか否かを表し、客観性はその内容の客観性である。客観性は、コンテンツの作成者に依存する。
表２にコンテンツの作成者の一覧を示す。

表２．コンテンツの作成者

形態	コンテンツの作成者
営利	ポータル
	営利法人
	報道機関
	出版社
	学校
非営利	非営利法人・コミュニティ
	個人
	政府・官公庁・公共団体

表１に示したコンテンツの機能、形式、表２において示したコンテンツの作成者とコンテンツの作成日を利用し検索結果を構造化することによって、ユーザは自身の検索の目的や情報源を選択し、その結果、検索の効率化を実現することができる。

3.4　アンカテキストの利用

我々は検索結果の構造化の観点にはコンテンツの機能とその形式、コンテンツの内容、コンテンツに対する評価があると考える。本稿ではコンテンツの機能とその形式に着目して構造化を行う場合について考える。図２にハイパリンク構造とアンカテキストの例を示す。

図２．ハイパリンク構造とアンカテキスト

アンカ処理部は以下の手順で URL のリストを作成する。
・各 URL をリンク先として含む Web コンテンツから、その URL のアンカテキストを抽出し、言語解析ツールで形態素解析を行い、品詞が名詞、未定義語である単語だけを各 URL をキーとしたリストに加える。
・また該当する URL をディレクトリ毎に分割し、そのトークンも同様に加える。
・同一 URL に対して複数の Web コンテンツがリンク元である場合、同一リストに同じように単語とトークンを加える。

3.5　コンテンツ形式 DB の自動構築

Web コンテンツの分類にルールベースを用いるものが多い。ルールベースの問題は精度の高い分類の出来るルールの抽出コストが高く、スケラビティの点で問題がある。このような問題に対応するために、本手法では Web コンテンツのコンテンツ形式の判定を、各単語の各コンテンツ形式に対する所属確率を計算し、それを用いて行う。所属確率の算出にはベイズ推定と EM アルゴリズム[2]を用いる。ベイズ推定によって算出された確率を用いた場合、高精度のルールを低コストで抽出できるだけでなく、複数のコンテンツ形式を持つ Web コンテンツを分類することも容易にできる。以下に単語 w_i がコンテンツ形式 c_k に対する所属確率を求める式を次のように定義する。

$$P(c_k|w_i)=\frac{P(w_i|c_k)\rho(c_k)}{\sum_{k=1}^{Z}P(w_i|c_k)\rho(c_k)} \quad (1)$$

ここで Z はコンテンツ形式 DB で設定したコンテンツ形式別の総数であり、$\rho(c_k)$はコンテンツ形式 c_k に関するギブス分布であり、$\rho(c_k)$は確率密度関数 $P(c_k)$に関する損出関数 $L(c_k)$をエネルギー関数として次のように定義する。

$$L(c_k)=-\sum_{j=1}^{m}N(w_i,d_j)\log P(d_j|c_k)-\sum_{i=1}^{n}N(w_i,d_j)\log P(w_i|c_k)-\log P(c_k) \quad (2)$$

$$\rho(c_k)=\frac{1}{Z}\exp(-\beta L(c_k)) \quad (3)$$

ここで $N(w_i,d_j)$ はコンテンツ d_j に w_i が出現する回数、m はコンテンツの総数、n は単語の総数、そして β は逆温度定数である。

次に式（１）を解くために EM アルゴリズムを用いる。EM アルゴリズムは E ステップと M ステップより構成され、両ステップを交互に繰り返すことで解くことができる。ここでは式（１）が E ステップに相当し、M ステップの式を次のように定義する。

$$P(w_i|c_k)=\frac{P(c_k|w_i)P(w_i)}{\sum_{i=1}^{n}P(c_k|w_i)P(w_i)} \quad (4)$$

EM アルゴリズムは（１）の値が収束したところで停止する。EM アルゴリズムによって（１）の値が求められ、次にコンテンツ d_j の各コンテンツ形式に対する所属確率を求める式を次のように定義する。

$$P(c_k|d_j)=\left(\prod_{i=1}^{l_j}P(c_k|w_i)\right)^{\frac{1}{l_j}} \quad (5)$$

ここで l_j はコンテンツ d_j に含まれる単語の総数である。この値が最も高くなる c_k をコンテンツ d_j のコンテンツ形式であると判断する。この値によっては表１の箇所で述べたように複数のコンテンツ形式を持つコンテンツを識別することも可能となる。

4. 評価実験

4.1　実験概要

提案した構造化手法の評価を行うために実験システムを構築し、検索結果の精度について評価を行う。構築したシステムは perl 5.8.0、茶筅[5]と Postgres7.3.0 を使用した。今回の実験で情報検索システムとして google を利用した。実験で利用した単語の数は 10 で、各単語に付き 1000 件の URL を抽出した。更に各 URL に対してその URL を含む Web コンテンツの上限を 100 と設定した。コンテンツ形式 DB はこれら抽出した URL のアンカテキストを用いて作成する。今回の実験では式（１）における Z を８、式（3）における β を 0.8 とした。更に構造化の有用性をユーザの検索の目的との合致性より評価した。

4.2 構築されたコンテンツ形式 DB

提案手法は検索結果の構造化のためにコンテンツ形式 DB を利用するため、構造化の精度はコンテンツ形式 DB の性能に依存する。表 3 にクエリを Perl としたときに、コンテンツの形式ごとにシステムが作成した単語を示す。これらの単語が。コンテンツ形式の判定に利用される。なお、表３に示した単語は、所属確率の高いもの上位５件である。この DB は 609 のコンテンツに対して 16676 のリンク元となるコンテンツがあり、１コンテンツあたり平均 27 のリンク元となるコンテンツが存在した。

4.3　構造化された検索結果の評価

次にコンテンツ形式 DB を用いて検索結果を構造化したときの評価結果を示す。評価の基準には再現率、適合率を用いる。評価の仕方は、各クエリに対しての検索結果からランダムに 100URL 抽出し、それぞれの URL がコンテンツ形式 DB で判断したコンテンツ形式と一致するか否かで行う。この実験結果を表４に示す。

表 3．クエリが Perl であるときのコンテンツ形式 DB

クエリとの関係	コンテンツの形式	単語
一次的関係	論文	PDF、著者、Tex、学会、検索
	解説記事	CPAN,script,download、正規表現、解説
	ニュース	発表、印刷、更新、news、編集
	掲示板	cgi,カウンタ、script、サンプル、管理人
	日記	月、日、天気、メモ、バックナンバー
	サイト	発売、価格、応用、簡単、発売
二次的関係	リンク集	リンク、集、サイト、カテゴリ、自由
	ガイド	発行所、作成、フリー、初心者、帯

表 4. コンテンツ形式 DB の精度評価

クエリとの関係	コンテンツの形式	再現率	適合率
一次的関係	論文	85.7%	92.3%
	記事	45.3%	57.3%
	ニュース	82.6%	87.9%
	掲示板	92.5%	91.3%
	日記	91.3%	87.5%
	サイト	51.8%	65.7%
二次的関係	リンク集	52.3%	57.4%
	ガイド	41.2%	43.5%

表 4 よりコンテンツのタイトルやアンカテキストなどで容易にそのコンテンツの形式が特定可能であることから論文、ニュース、掲示板、日記などは再現率、適合率が共に高いことが分かる。

次に Morrison[5]らのユーザの検索目的の分類とその割合を用いて、提案手法の有効性を評価し、その実験結果を表 5 に示す。表 5 において全体に占める割合は、ユーザの全行動のうち、目的別の行動が占める割合であり、コンテンツ形式 DB による一致率は、それぞれの目的に対し、コンテンツ形式 DB を用いて判断した Web コンテンツが合致した率である。この計算には、表 4 のデータを用いた。全体における一致率はこれら二つの値をかけることで求める。

表 5. ユーザの検索目的に対するコンテンツ形式 DB の有効性

ユーザの検索の目的	全体に占める割合	コンテンツ形式 DB による一致率	全体における一致率
1. 発見			
a ダウンロード	**2%**	**65.7%**	**1.31%**
b 事実の確認	**15%**	**79.2%**	**11.88%**
c 文書の取得	**6%**	**80.3%**	**4.82%**
d 製品	**2%**	**65.7%**	**1.31%**
2. 比較	**51%**	**50.5%**	**25.76%**
3. 理解	**24%**	**50.5%**	**12.12%**
合計	**100%**	**-**	**57.20%**

表 5 の結果より、提案手法によって検索結果の構造化を行うと、検索結果からユーザが自身の検索の目的に合った Web コンテンツを選択する適合率は 57.20%であることが分かる。

5. 考察

提案する検索結果の構造化の目的は、検索結果においてユーザが自身の検索の目的に合ったコンテンツをコンテンツの機能と情報源の選択を可能とすることにより、ユーザの情報検索を効率化することである。この目的に対し、クエリが技術用語や専門用語である場合、コンテンツ形式の判定精度が高く、それ以外では低くなる。これはクエリが前者であれば該当するコンテンツが満たす機能が一つに限られるのに対して、後者であると一コンテンツで複数の内容と機能を持つ場合があるためであると考えられる。機能の判定精度を高めるには一コンテンツを同じ機能を持つセクションに分割する必要があると考えられる。

また、情報検索システムはユーザの主として検索目的の発見を目的としているが、検索結果の構造化で比較や理解についても構造化することによって、より多くのユーザの検索を情報検索システムは支援することができると考えられる。

6. まとめ

本研究では情報検索システムを用いたユーザの情報検索を効率化するために、検索結果の構造化を実現する手法を提案した。提案手法は検索結果の構造化を Web コンテンツのカテゴリ、コンテンツの機能、形式、作成者、作成日時の五つの属性を用いることを提案した。手法の中で Web コンテンツの機能の抽出のために Web コンテンツのリンク元がそのリンク先のコンテンツの機能、形式や内容を要約していることに着目し、その情報からコンテンツ形式 DB を作成し、検索結果の構造化を行った。提案手法の新規性は、構造化を Web コンテンツの機能、形式、その作成者と作成日時によって行ったことと、その構造化に用いた技術にある。今後は構造化に Web コンテンツのカテゴリを加え、アンカ文字列から Web コンテンツの評価を抽出すること、構造化された検索結果から情報抽出を実現する方式を検討する。

文　献

[1] G. Attardi, A. Gullì and F. Sebastiani.: Theseus: Categorization by Context, 8th Word Wide Web Conference, Toronto, Canada, 1999.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin.: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B, 39(1): 1-38, 1977.

[3] Noriaki KAWAMAE Hideaki SUZUKI and Osamu MIZUNO.: Collaborative Filtering Via Bayesian Statistical Model(To appear)

[4] J. B. Morrison, P Pirolli and S. K. Card.: A Taxonomic Analysis of What World Wide Web Activities Significantly Impact People's Decisions and Actions, Conference on Human Factors in Computing Systems,2001.

[5] T. Hofmann.: Unsupervised learning by probabilistic latent semantic analysis. Machine Leaming Joumal, 42(1):177-196, 2001.

[6] 茶筅:http://chasen.aistnara.ac.jp/index.html.ja

[7] 松平正樹，上田俊夫，大沼宏行，森田幸伯.: Web コンテンツの分析に基づくオントロジー構築および情報整理の試み,人工知能学会, 第４回セマンティックウェブとオントロジー研究会, 2003