子供のためのウェブ情報検索支援システムの開発*

美馬秀樹
東京大学大学院工学系研究科
mima@biz-model.t.u-tokyo.ac.jp

尹泰聖
株式会社 オープンナレッジ
yoon@openknow.com

概要

In this paper, we describe a design and an implementation of the next generation web information retrieval aid system for children we developed in the IPA FY2003 Exploratory Software project. The purpose of the system is to provide cyber leaning environment which gives children intellectual interests with satisfying their spirits of inquiry. Our approach is based on 'and'-oriented query expansion technique we proposed in the web information retrieval, whereas conventional query expansion is mainly 'or'-oriented. To achieve the efficient and effective query expansion, we also developed query expansion database derived from the ontology which is also derived from textbooks, such as schoolbooks, dictionaries and encyclopedias. Namely, our system effectively reduces number of the web information retrieval results, and efficiently improves the quality of the results for children to access innocent information quickly by using appropriate existing 'knowledge'. Manually expanded keywords / terms by users in a query are also accumulated as personal query expansion data and the data is re-used in the other retrieval stage to achieve an order-made information retrieval. Furthermore, the accumulated data is expected to construct a huge knowledge database in a community to accelerate learning how to achieve efficient web information retrieval for children.

1. 背景

近年では，小学生でさえも自分の電子メールアドレスで情報交換を行うことが当たりまえになり，中には自分のホームページを立ち上げ，積極的に情報を発信している子供たちもいる．このように，情報へのアクセス環境，つまりインフラストラクチャの整備は順調に進行していると言えるが，情報の利用環境に関してはなお多くの問題が残っているのが現状である.特に深刻なのは,ウェブ検索の問題である．現在の検索エンジンは，子供が行うウェブ検索も，大人が行うウェブ検索も，検索に使う言葉が同じ場合，検索結果は全く同じである．子供に対するウェブの影響を考慮すると，子供には子供のための専用のウェブ検索環境を提供することが非常に重要である．また，教育的な観点においては，知的発達過程にある子供に対しては，検索を通じた学習までも考慮したウェブ検索システムの構築がより望ましいと考えられる．

このような問題に対して，従来，ブラックリスト方式，ホワイトリスト方式，ストップワードリスト方式等の方式を基にしたウェブ・サイトのフィルタリングが行われてきた．しかしながら，これらの方式による検索では，リストの維持管理の困難さにより，上記の二つの問題に対する根本的な解決策と成りえていないのが現状である．実際，”yahoo きっず”においては,“りんご”のキーワードによる検索でも，一般の検索エンジンと同様に，容易にアダルトサイトへ到達してしまう．

そこで本研究開発では，子供のためのウェブ検索支援システムの開発により，子供達に知的探求心をより一層満足させながら，知的刺激を与えるサイバー教育環境の構築を目指す．そのために，小学校の教科書を始めとした小学生向けの知識を分析し，有用な知識を抽出することでオントロジー情報[1]を構築する[1][2]．さらに，検索の絞込みに至った結果（検索経路）をデータベースに保存し，再利用することで，意図する情報へのアクセス性の改善を狙う．このために，これらデータベースを管理運用するためのソフトウェアを設計し実装する．

2. 関連研究

先にも述べたように，有害サイトのフィルタリングに対しては，従来，ブラックリスト方式，ホワイト

---

* Design and Implementation of a Web Information Retrieval Aid System for Children
Hideki Mima, Shool of Engineering, University of Tokyo
Taesung Yoon, Open Knowledge Corp.

[1] オントロジーの定義は，分野やドメインにより種々なされているが，本稿では，（専門）用語による概念の認知と概念間の関連性の発見による分類[3]をオントロジーとする．

リスト方式，ストップワードリスト方式等の方式を基にした方式が利用されてきた．ブラックリスト方式では，検索結果，もしくは検索対象からブラックリストに含まれるサイトを除外することで有害サイトへのアクセスを制限する．また，ホワイトリスト方式では，ブラックリスト方式とは逆に，検索できるサイトをホワイトリストにあるサイトのみに制限する．一方，ストップワードリスト方式では，検索対象であるサイトにストップワードリストにある特定のキーワードが含まれている場合に，そのサイトが有害であると判断し，検索結果から除外する．

例えば，検索サイト”yahoo”では，子供向けの検索サイトとして“ヤフーきっず”[4]を提供しているが，この検索サイトでは，ホワイトリスト方式を基にした手法により，検索内容を子供向けのものに限定することで，子供に不要と思われる情報を除いている．また”BIGLOBE”での”KIDSPLAZA”[5]や Walt Disney における子供向けの検索サービス「Disney's Internet Guide（DIG）」[6]も同様のサービスによりフィルタリングが行われていると思われる．また，ロボット系の検索サイト”goo”においては，同様に子供向けの検索サイトとして” きっず goo”[7]を提供しているが，このサイトにおいては，基本的にホワイトリスト方式，およびストップワードリスト方式の併用による有害サイトのフィルタリングが行われている．

いずれの方式にも一長一短があるため，”きっず goo”のようにいくつかの手法を併用するのが一般的であると考えられるが，基本的にリストの作成，維持は人手によるものであるため，正確，かつリアルタイムに検索要求に対応するのは非常に困難である．

また，表現の自由や，知る権利にも配慮が必要であるとの認識も存在するため，上記の問題解決にも増して，新たな方式が望まれているのが現状である．

３．　研究開発の内容

２．１ 検索結果の質と量

先にも述べたように，近年では，インターネットへのアクセス環境が向上するに伴って，小学生などの子供も簡単にインターネットの世界に入り，情報を検索できるようになっている．しかし，インターネット上の情報を検索するときに，その量と質を維持しながら探したい情報を検索することは非常に難しい．例えば，検索サイト”Google”において「リンゴ」と入力して検索した結果，約 34 万件の検索結果が提示されるが，このように何十万件もの検索結果が提示されると，多くの場合，子供はそれ以上検索を続ける意欲を失う．また，検索結果が少ない場合でも，その内容が子供に相応しいかどうかは保障できない．通常，ウェブ上の情報は一般成人を想定して作成されたものがほとんどであるため，子供のための検索システムは，先にも述べたように，子供のために特別に作ったサイトだけを検索結果とするか，あるいは，一般サイトの中から子供が見てもいいサイトだけを検索結果にするかの選択を行う必要がある（ホワイトリスト方式）．しかし，先にも述べたように、このような情報を，更新速度の非常に速いインターネットに対し，維持することは一般に非常に困難である．

一方，例えば，先の「リンゴ」の例では，まず約 34 万件の検索結果が提示されるが，その中からさらに「ビタミン C」を含む情報を持つサイトだけに絞ると約８千件となる．また「みかん」を入力して検索するとその対象は約２千件となる．その中から「原産地」を含むサイトを検索するとその検索結果は 100 件になり，さらにそのうち「京都」を含むサイトは 26 件だけとなる．この 26 件の中から「温州みかんにはガン抑制効果がある」という情報を得ることは比較的容易である．つまり，効率的な検索のためには，どの内容からどう絞って行くかがひとつの観点になる．この「絞っていく過程」をここでは「検索経路」と呼ぶ．例えば，りんご→ビタミン C→みかん→原産地→京都は，りんごから始まる一つの検索経路になる．ある概念から始まる検索経路は複数存在する．これは，検索経路は検索目的の他にも，検索者の経験，学歴，知識，時間，理解できる言語などから影響されるからである．

２．２　有害情報の遮断

有害な情報を遮断しながら，子供の知能発達過程を助けるような検索システムの提供は子供にとって大変重要な問題である．有害な情報を遮断する方法は一般的にストップワードリストによるキーワード方式，もしくはブラックボードによる禁止サイト遮断方式を使う．即ち，予め決めたストップリストにあるキーワードを含んでいる情報やサイト，もしくはブラックリストにあるサイトは子供に見せない方式である．しかし，ストップリストにないと思われるキーワードを使うか，あるいは少し変えた表現を使う情報は通過し，検索結果として子供に提示される．また，先と同様，一般にブラックリストを最新のものに維持することは非常に困難である．

以上において，本研究開発では，より積極的な方法として，教科書等の子供のために作られた知識よりオントロジー情報を抽出[1][2]し，これを基にした検索経路を子供に提供する．これにより，知識が不足しがちな子供においても，より子供に適した情報を上位の検索結果として提供することが可能となる．子供がインターネットに接続し，あるキーワードを入力してウェブ上の情報を検索する場合，本システムはそのキーワードを出発点にした複数の検索経路を提示する．子供は幾つかの岐路で，提示された複数のキーワードもしくはサイトから一つを選ぶ．結

果的に選んだキーワードやサイト群は検索経路としてまとめられ，実際の検索に再利用される．検索経路は小学校の教科内容を反映して作成するため，ウェブ検索だけではなく，小学生の学習行為や知能発達過程を支援できる．また，サイトの内容やキーワードと，教科内容，辞書の内容等との関連を積極的に提示することで，子供の知識を補い，より効率的に絞込みを進めることを支援する．

２．３　学習の一環としての検索

知能発達過程を助ける検索システムは，検索に教科課程も反映して，できるだけ子供が学習手段として検索システムを利用するのを支援するシステムである．即ち，情報検索という行為は，それ自体が学習活動の一環として行われるべきである．たとえば，ある概念から関連する概念を考えて，なぜその二つの概念が関連しているかを科学的に明示することによって，概念と概念間のネットワークが広がっていく．その結果，子供の知的能力は刺激を受け，もっと柔軟な思考方式を身に付けることができる．概念と概念間の関連付けは参考用の検索経路として子供に提供され，子供は自分の考えを追加したり，内容を修正したりしながら，検索経路を成長させる．

４．　システム構成

図１に本システムのシステム構成を示す．本システムは，以下の大機能より構成される．

１）オントロジー構築機能
２）検索知識データ提供機能
３）用語関連度計算機能

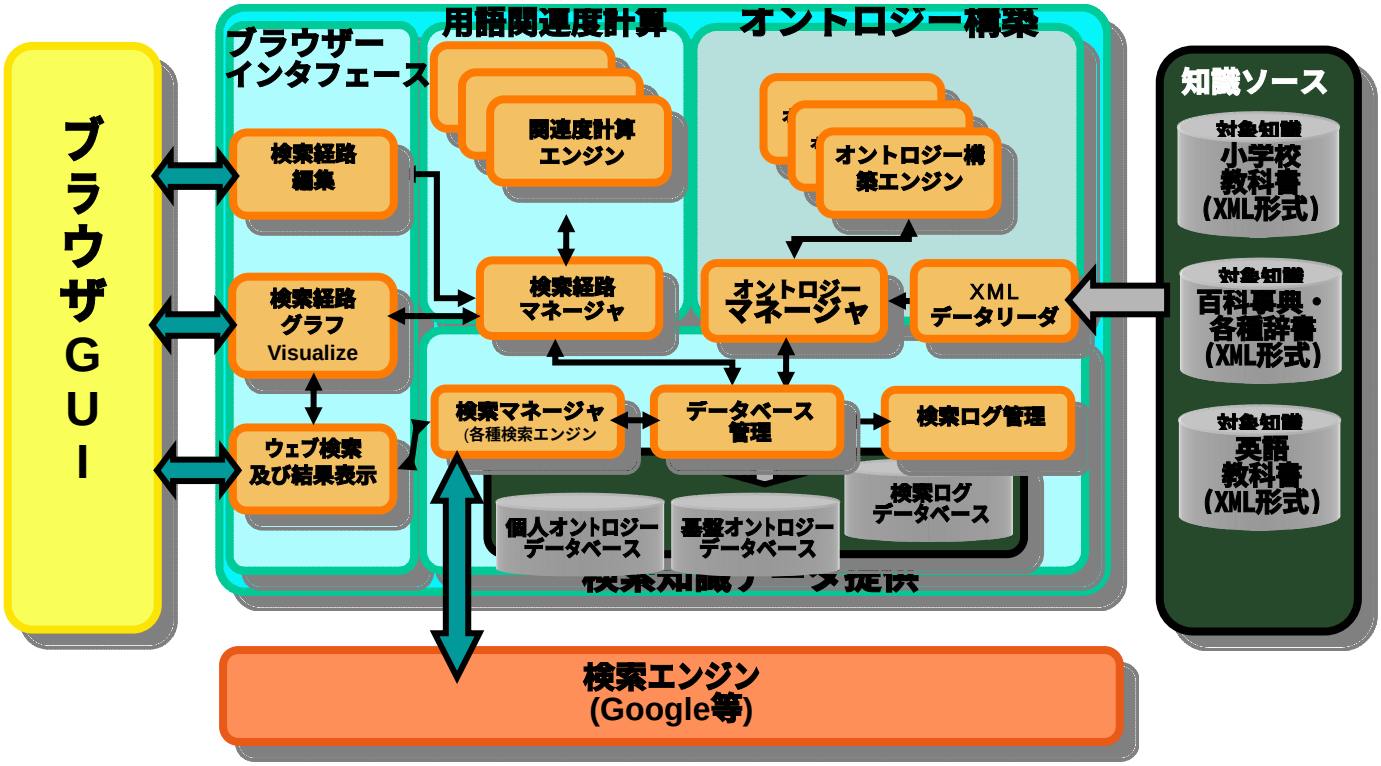

図1．システム構成

４）ブラウザインタフェース機能

以下に各機能の詳細を示す．

４．１　オントロジー構築機能

本機能は，XML 形式による教科書データを含む知識ソースを入力とし，検索経路を生成するための基礎的データとなるオントロジー情報を構築し，検索知識データ提供機能に出力する．

本機能は以下の中機能より構成される．

（１）XML データリーダ
　　XML 形式の知識ソースデータを解析し，オントロジー構築に使用するテキストを抽出する．
（２）オントロジーマネージャ

用語自動抽出，用語クラスタリングを含む各種のオントロジー構築エンジンをドライブする．

（3）オントロジー構築エンジン
テキストより用語抽出，用語クラスタリングを含む言語処理によりオントロジーを自動構築する．

4．2　検索知識データ提供機能

本機能は，「4．1 オントロジー構築機能」からのオントロジー情報を入力とし，基盤オントロジーとしてデータベースに格納する．また，本機能は，「4．3　用語関連度計算機能」からの用語要求に対し，データベースから用語のクラス，用語のオントロジー上での位置情報，用語の知識ソース上での位置情報を含む用語情報を検索し，得られた用語情報を返す．また，同「4．3　用語関連度計算機能」からの検索知識変更要求に対し，個人オントロジーデータベースにアクセスすることで用語間の関連の追加，削除，関連（度）の変更を含むデータ処理を行う．また，「4．4　ブラウザインタフェース機能」からのキーワードを含む検索要求に対し，外部検索エンジンに検索要求を出力し，検索結果を得，得られた検索結果を「4．4　ブラウザインタフェース機能」に返す．

本機能は以下の中機能，およびデータベースより構成される．

（1）　データベース管理
各データベースにアクセスし，情報の検索，および追加，削除，更新を行う．

（2）　検索マネージャ
外部検索エンジンにアクセスすることでウェブ検索を行い，検索結果を処理する．

（3）　検索ログ管理
検索ログ・データを処理することで個人オントロジーデータの更新を行う．

（4）　知識データベース
基盤オントロジーデータベース，個人オントロジーデータベース，検索ログ・データベースから構成され，検索経路生成の基盤となる知識を蓄積する．

4．3　用語関連度計算機能

本機能は，「4．4　ブラウザインタフェース機能」からの用語の入力に対し，用語を含む用語要求を「4．2 検索知識データ提供機能」へ出力し，用語情報を得る．得られた用語情報より，用語間の関連度を計算し，用語情報と共に「4．4　ブラウザインタフェース機能」へ返す．また，本機能は，「4．4　ブラウザインタフェース機能」からの検索経路編集要求に対し，ユーザ情報，用語間の関連の追加，削除，関連（度）の変更を含む検索知識変更要求を「4．2 検索知識データ提供機能」に出力する．

本機能は以下の中機能より構成される．

（1）検索経路マネージャ
オントロジー情報から関連度計算エンジンを利用して検索経路を生成する．

（2）関連度計算エンジン
検索経路の生成において，用語間の関連度を計算する．関連度の計算方式によりいくつかの計算手法を選択できるものとする．

4．4　ブラウザインタフェース機能

本機能は，ユーザからの用語の入力に対し，「4．3 用語関連度計算機能」へ用語を出力することで用語情報と関連度を得る．得られた用語情報と関連度より，検索経路を生成し，ブラウザ GUI に出力する．また，本機能は，ユーザからの用語の選択と追加，削除，関連（度）の変更を含むアクションの入力より，検索経路編集要求を生成し，「4．3 用語関連度計算機能」へ出力する．また，ユーザからのキーワード入力より，検索要求を生成し，「4．2 検索知識データ提供機能」へ出力することで検索結果を得，得られた検索結果をブラウザ GUI に出力する．

本機能は以下の中機能より構成される．

(1)検索経路編集
検索経路を個人の嗜好により編集可能とする機能．

(2)検索経路グラフ Visualize
検索経路を Visualize し，ユーザから用語を選択可能とする．ユーザの選択により検索クエリーを自動生成し，ウェブ検索の入力とする．

(3)ウェブ検索及び結果表示
ユーザからのキーワード入力，および検索経路より選択された用語の入力に対し，ウェブ検索を行い結果をブラウザ上に表示する．

5．　検索経路

従来の質問拡張の枠組みが主に，語彙のレベルに対する is-a 階層を用いた限られた推論を提供するものであるのに対し，本システムでの取り組みは，質問拡張の枠組みをオントロジーによる制御にまで拡張したものと言える．つまり，検索クエリーとして

与えられたキーワードに対し，従来の枠組みでは，シソーラス等の類義語辞書を利用して is-a の関係にある語を OR 関係により付加しキーワードの拡張を行うのみであるのに対し，本システムの特徴は，OR のみではなく，AND 関係によるクエリーの拡張を行うことにある．例えば，図２に示すように，「りんご」に対して，オントロジー上で産地により関連づけられる「青森」や，栄養素による関連の「食物繊維」等の関連語により果物としての意味的制約をかけることがこれに相当する．さらに，この枠組みにおいては，意味的あいまい性を持つ語彙の有害サイトに表れやすい意味をあらかじめ抽出することにより，明示的に NOT 関係を指定する等のクエリー拡張を行うことも考えられる．

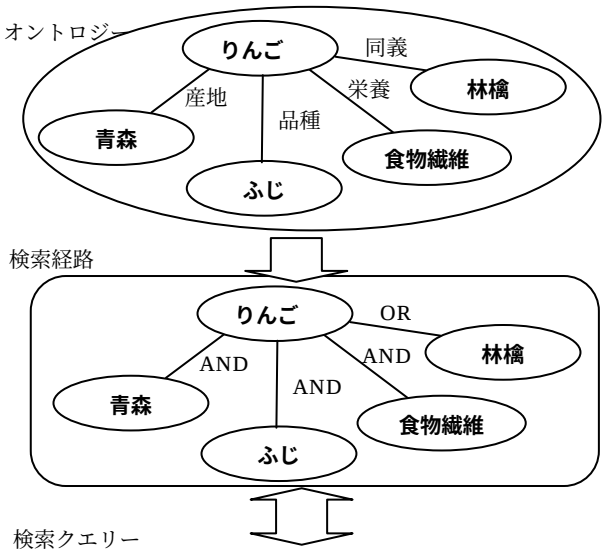

図２．オントロジーと検索経路

このような検索拡張を行うための基礎的データとして，本枠組みでは，教科書，および辞書，辞典の類のテキストから自動抽出したオントロジー情報を利用する．具体的には，

１）C/NC-value 手法による用語の自動抽出[1][2]
２）１）による用語の自動階層化クラスタリング

により，関連のある用語を，そのコンテキストの情報[2]を基に階層的にクラスター化し，クラスター階層における用語間の距離を基に類似度を計算する．これにより関連語とその関連度が得られる．また，is-a 関係は，類義語辞典，および国語辞典等より抽出したものを利用する．以上の処理により，検索経路の情報を生成，ユーザに提示する．ユーザは提示された検索経路を選択もしくは自動的な選択を利用することでクエリー拡張を行い，検索および絞込みを行うことが可能となる．

教科書等の知識ソースにおいてコンテキストに表れやすい語を検索キーワードとペアにして検索を行うということは，適合率の観点において，検索結果を，より知識ソースで表現されている内容に近い方向に絞り込むことを意図していると考えられる．また，（有害サイトのような）意図しない情報が検索される原因の一つに，指定したキーワードの多義性があげられるが，”りんご”に対する”食物繊維”のような関連を与えることは，語義のあいまい性解消のキーとなる情報を与えることに相当すると考えられる．大井ら[8]は，全文検索において，検索対象のテキストに対し，コンテキストの情報による語義のあいまい性解消手法を適応することで検索の精度が向上することを示している．

６．　考察

子供のための検索経路情報は小学校の教科課程などを反映して意図的に作ることもできるし，また，子供たちが自分からつくることもできる．本システムでは，子供のウェブ検索も学習過程の一環として捉え，教科書等の情報を反映し，意図的かつ半自動的に作成する．その結果はオントロジー情報として表現され，参考用の検索経路に変換されて子供に提供される．図３にその例を示す．子供はこの検索経路を参考にして，ある概念に関連する内容を検索していく．また，検索経路には，教科書の情報や，辞典の情報のみならず，子供のためのお勧めサイトや関連する情報も添加することが可能である．このように，本検索支援システムの提案点は，「検索経路（オントロジー情報）による，子供のためのウェブ検索」の支援にある．本方式により，いままでとは全く異なる検索環境を子供に提供できる．その結果，子供の知的発達過程はもっと活発になり，ある概念について，深く広く考えることができるようになると期待される．

また，第２の提案点として，本システムが「教科内容等の外部知識を検索過程に反映」できることがあげられる．本システムでは検索行為を学習活動の一環として捉えているため，教科内容や各種辞書，辞典等の内容を反映した検索経路（オントロジー）を作成する．この検索経路は参考用として子供や小学校などに提供される．その後，ユーザとしての子供や小学校では，独自の教科内容から新しい概念を取り出して，検索経路に追加するか既存の内容を修正することができる．また，小学校で独自に運営している検索経路を，ある地域の小学校などがその内容を統合して運営することによって，本検索経路はもっと広くて深い内容になる．

---

[2] 実験では，コンテキスト中にある用語との共起頻度をクラスタリングの入力とした．

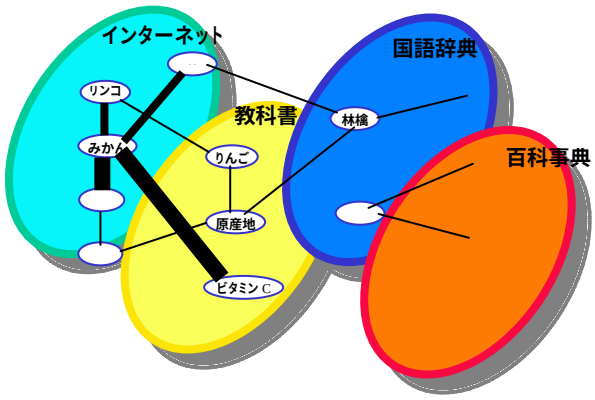

図３．検索経路の提示による検索支援

さらに，第３の提案点として，本システムによるデータが「検索経路」から「知識基盤」へ発展できることがあげられる．ユーザからの内容の追加や修正，そして，地域間の統合運営などによって，検索経路は成長する．その結果，検索経路は一つの知識基盤としても捉えられる．したがって，この知識基盤は，教育活動をもっと豊かにするための基盤として利用可能である．この知識基盤は，教育活動の支援，ウェブ上の情報検索支援，子供の知的発達過程の支援などに利用される．また，中学校や高校などの教科内容まで含めて拡張することで，この知識基盤の応用範囲はもっとも広くなると期待される．

７．　議論

７．１　ユーザインタフェース

本システムは子供を対象にするため，ユーザインタフェースに関しても大人の場合とは異なるアプローチを用いることも考慮する必要がある．特に低学年の小学生の場合，システムへのフレンドリネスを高めるためには効果的であると思われる．例えば，キャラクターを利用したり，より子供に受け入れられやすいと思われるアイコン・デザイン等を多用することも必要であろう．さらに，個人の嗜好を反映するための個人認証を前提としたシステムの場合，その認証方式は非常に重要な問題となる．例えば，一般的な BASIC 認証が子供にどこまで受け入れられるかどうかは疑問である．これに関しては，”yahoo”，”MSN”等のサイトで既に運用されているアバタ・システムや，NetPeople[9]のような agent 技術を活用することも視野に入れ，研究開発を進めている．

７．２　未知語処理

本システムは，ユーザに入力された検索キーワードより検索経路データベースを参照するため，入力されたキーワードが検索経路データベースに登録されていない場合，つまりシステムにとって未知語である場合，その扱いを考慮する必要がある．その場合，１）部分一致や文字レベルの類似度計算による類似語候補の提示，２）他の同義語や類義語の入力を促す，３）連想される関連語の入力を促す，等の処理が考えられる．また，システムの目的の一つである教育的な観点においては，先にも述べた，ユーザからの検索経路の追加により，ユーザ指向の知識（基盤）の拡充が積極的に行われることが望ましいと考えられる．しかしながら，このような場合，誤った知識の追加や，知識間のコンフリクト[10]のような問題に関しても，十分考慮する必要があると思われる．上記のような手法を含めて，ユーザを適切に支援することが期待される．

７．３　評価

システムの有用性を示すためには，システムを適切に評価する必要がある．本研究開発では，１）技術評価，２）実証実験評価，によりシステム評価を行う予定である．より具体的には，技術評価として，実際の検索エンジンと検索経路の利用による有害サイトの排除や意図した情報への絞込み効率に関する定量的評価を行う．例えば，検索経路を利用する場合と利用しない場合の比較を行い，それぞれの適合率，再現率，さらにインターバルな適合率を計測することが考えられる．これにより検索経路の利用が技術的にウェブ検索に有効であることを示す．

また，実用的な評価を目的とした実証実験評価として，実際に小学校における小学生に利用してもらい，検索ログの解析による評価，及び小学生，教師に対する聞き取り調査による評価を行う予定である．これらにより，システム利用による検索の効率性の評価と共に，ユーザのウェブ検索の学習に対する効果が評価可能となると考えられる．

一方，子供を対象とした場合の聞き取り内容の信憑性(より正確な情報の確保)や，適切なログの取得とその解析方法に関しては十分考察を行い，適切に配慮する必要があると考えている．

８．　まとめ

本稿では，子供のためのウェブ情報検索支援システムの開発について述べた．本システムの特徴は，ウェブ検索において，小学校教科書等の知識源よりオントロジー情報を生成し，それを基に作成した検索経路データを利用することにある．子供がインターネットに接続し，あるキーワードを入力してウェブ上の情報を検索する場合，本システムはそのキーワードを出発点にした複数の検索経路を提示し，子供

は幾つかの岐路で，提示された複数のキーワードもしくはサイトから一つを選ぶ．結果的に選んだキーワードやサイト群は検索経路としてまとめられ，学習により以後の検索に再利用される．サイトの内容やキーワードと，教科内容，辞書の内容等との関連を積極的に提示することで，子供の知識を補い，より効率的に検索絞込みを進めることを支援し，また有害な情報から子供をより遠ざけることが可能となると期待される．

また，それぞれの検索経路を個別に管理，再利用することで，近年，注目されているオーダーメード検索への応用も期待される．

今後の課題として，システムの実運用での評価があげられる．これに関しては，現在，複数の小学校を対象として，実際の小学生による技術評価，及び実証実験を行う計画を進めている．

参考文献

[1] Mima H., Ananiadou S., An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese, *Int. J. on Terminology* 6/2, pp. 175-194, 2001.

[2] Nenadic G., Mima H., Spasic I., Ananiadou S. and Tsujii J., Terminology-based Literature Mining and Knowledge Acquisition in Biomedicine, *International Journal of Medical Informatics*, 67, pp. 33-48, 2002.

[3] Gene Ontology Consortium, http://www.geneontology.org/, 2003.

[4] Yahooきっず, http://kids.yahoo.co.jp/, 2003.

[5] KIDSPLAZA, http://netplaza.biglobe.ne.jp/KIDSPLAZA/, 2003.

[6] Disney’s Internet Guide, http://disney.go.com/dig/today/, 2003.

[7] きっずgoo, http://kids.goo.ne.jp/, 2003.

[8] K. Oi, E. Sumita, H. Iida, Document retrieval method using semantic similarity and word sense disambiguation (in Japanese), J. of Natural Language Processing 4/3, pp.51-70, 1997.

[9] NetPeople (iNANGO Corp.), http://www.inago.co.jp/iNAGONetPeople/iNAGO_website/html/index.html, 2003.

[10] Visser P.R.S., Jones D.M., Bench-Capon T.J.M. and Shave M.J.R., An Analysis of Ontology Mismatches; Heterogeneity versus Interoperability. In AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California, USA, 1997.