Journal of National Fisheries University　54 (2) 15−24 (2006)

シソーラスとカテゴリ情報を用いた水産物ウェブ検索

楫取和明*・瓜倉茂*・青木邦匡*

Web searching of fishery products by using thesaurus and category information

Kazuaki Kajitori*, Sigeru Urikura*, and Kunimasa Aoki*

Using web search engines is not an easy task when users are asked to input complex boolean expressions in order to express their search needs. Since web users usually input just a few keywords to search engines, it would be convenient to users if a system expands user's keywords to an appropriate boolean expression. Although it is not easy to construct such a system for all purposes, restricting search areas to fishery-related areas seems to make such a construction near at hand. In this paper, we investigate effectivities and possibilities of such systems.

Key words : Information services, Fishery products, Thesaurus, Dictionaries

1　はじめに

ブーリアン検索式(OR, AND, NOTなどの論理記号を含む検索式)をサポートする検索サイトにおいてウェブ情報を検索するとき、効果的な検索式を書くのは簡単ではない。たとえば、「鯖の料理方法」を調べたいとき、まず「さば」や「サバ」なども検索対象にしたければ、「鯖 OR さば OR サバ」(「鯖」か「さば」か「サバ」を含むという意味)という検索式を書かねばならない。しかもこの選言形の書式は検索エンジンごとに異なる。さらに、鯖に関するページにはショッピングや釣りなどのページが多く存在するから、「料理方法」に限定するようにしなければならない。そこで「レシピ OR 材料 OR 作り方」のような条件を検索式に加えることで意図するページ群にできるだけ限定するようにしなければならない。検索エンジンにおいてユーザが指定するキーワードは平均して2語未満といわれている[1]。そこでユーザの与える少ないキーワード群から出発して効果的な検索式を作成できれば便利であろう。分野を限らない汎用の検索に対しこのようなサービスを構築するのは膨大な手間がかかるであろうが、水産分野などに限定すれば試作も比較的容易であると思われる。参考文献[2]ではウェブ検索の研究開発の動向として、全世界のウェブを対象にした汎用的な検索を目指す方向性と、ジャンルやドメインを限定するといった方向性に分かれるようになってきたとしている。そこでは、ジャンルやドメインを限定したウェブ検索の例またはその研究例として情報系分野の学術論文の検索サイトや主観的な情報を活用する研究などがあげられている。

本編は、ユーザが検索エンジンを使って水産に関するウェブ検索を行うとき、少ない条件指定から効果的な検索式を作成できるようなシステムを試作し、水産というジャンルに特徴的な点を探りながらそのようなシステムの有効性・可能性を考察することを目的とする。

2　試作システムについて

以下、論理ORは∨、論理ANDは∧、論理否定は¬で表す。

「鯖」という検索キー指定に対し、「さば」「サバ」という表現を加えて「鯖∨さば∨サバ」という形に展開して検索を行う方法をシソーラス（類義語辞書）を用いた検索という。すなわち、ユーザが指定した検索キーワードをシソーラスを用いて同義語などの関連語を含めた選言形式に展開し、検索をかける方法である。シソーラスを用いた検索の研究はいろいろなされているほか、特定分野の文献検索などに実用的に用いられている。

2005年4月13日受付. Received April 13, 2005.

＊　水産大学校水産情報経営学科（Department of Fisheries Information and Management）

上の例で「料理方法」に対し「レシピ∨材料∨作り方」という条件を考えたところは、興味対象の「鯖」の「どういうこと」を調べたいのかを決める部分といえる。この部分もたとえばユーザに「レシピ」というキーワードを考えさせ、それから「レシピ∨材料∨作り方」という条件をシソーラスを用いて導くことが考えられる。あるいは、「料理情報」「買物情報」「釣り情報」などいくつかのカテゴリをあらかじめ設定して、カテゴリごとに検索条件を決めておくという方法も考えられる。後者はユーザーがキーワードを考える手間を省くという意味でメリットがある。また水産という分野に限った場合カテゴリはとりあえず数個でスタートできるであろうし、シソーラスとしては水産関係のシソーラスを用意すればよいことになる。

そこで試作するシステムとしては、 水産のカテゴリを数個あらかじめ用意して水産に関するシソーラスを自作するものとする。

カテゴリとしては、「グルメショッピング」「レシピ」「フィッシング」「アカデミック」「その他」を用意した。

ユーザの与えたキーワードをシソーラスで展開するとき、カテゴリが何かによってシソーラスを変えた方がよいことが多い（このことはシソーラスによる検索一般に認められる[6]）。たとえば、「アカデミック」カテゴリに関しては魚名に対して学名をシソーラスに登録しておくと効果的であるが、他のカテゴリに関しては学名はむしろマイナスである。したがって、シソーラスはカテゴリごとに作ることにする（「その他」カテゴリに関してはシソーラスは作らない。）。

各カテゴリにおいてウェブ上に非常に豊かな情報が存在するような対象がある。たとえば、「グルメショッピング」カテゴリに関しては「まぐろ」のショッピングページが非常に多くある。こういうものはユーザに豊富な情報が存在することを気づかせるためにディレクトリ（集めたウェブ情報を分類・整理して保管し提供する仕組み）に検索式を登録して提示することが考えられる。また、ウェブ上に豊富な情報がありながらユーザキーワードのシソーラス展開＋カテゴリキーワードによる検索式ではうまくいかないものもある。こういったものもよりうまく検索できる検索式をディレクトリに登録しておくといいだろう。したがって、カテゴリ毎にディレクトリを作ることにした。ディレクトリの各項目にはユーザが検索語を入れる欄を設ける（以下の図参照）。したがって、各カテゴリのディレクトリはサブカテゴリと見ることもできる。

システムはウェブサーバアプリケーションとし、linux上に構築した。ウェブサーバはapacheを利用した。よって、システムは以下の部分からなる。

- カテゴリ毎のシソーラス
- カテゴリ用のキーワード辞書
- ディレクトリ用の検索式辞書
- ウェブブラウザをユーザインターフェイスとするためのスクリプト
- 検索エンジン

「シソーラス」と「カテゴリーキーワード辞書」と「ディレクトリ検索式辞書」は、RDBであるPostgreSQLを利用して構築した。

ユーザインターフェイスの外見は以下のようである。

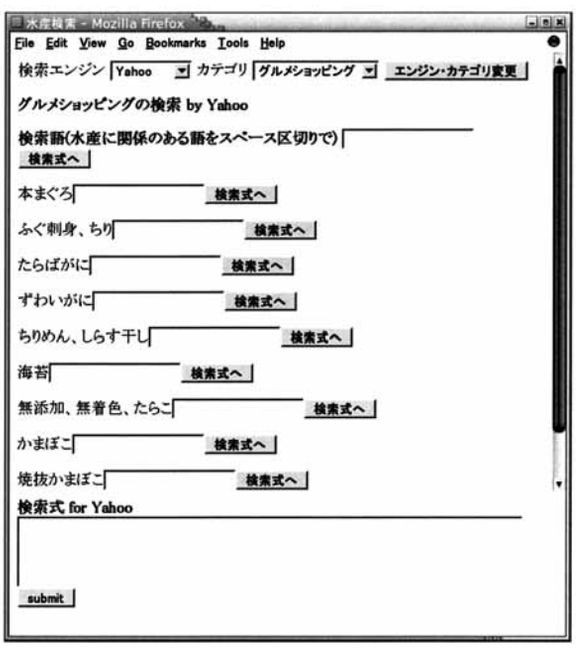

Fig. 1 .

使い方は以下のようになる。ユーザはまず検索エンジンとカテゴリを選択する。検索エンジンの選択肢については、Fresheye、およびYahooのページ検索を用意した[1*]。「グルメショッピング」のカテゴリに対しては、「楽天」における検索も加えた。

カテゴリキーワード辞書は、二つのテーブルcategory, ncategoryからなる。categoryテーブルはカテゴリに属するページが（それらから少なくとも一つ）含んでいるべきキーワード群を格納している。たとえば、

1＊ 定評のあるGoogleは使用条件が厳しいのとブーリアン検索式のサポートが弱いので避けた。

Table. 1 .

category

category	keywords
グルメショッピング	ショッピング, 買い物, 注文
グルメショッピング	食, グルメ

という行があれば、「グルメショッピング」を選ぶと「(ショッピング∨買い物∨注文) ∧ (食∨グルメ)」(「ショッピング」か「買い物」か「注文」を含み、かつ「食」か「グルメ」を含む) という検索条件がつくことになる。

ncategoryテーブルはカテゴリに属するページが含んでいるべきでないキーワード (除外キーワードということにする)を格納している。たとえば、

Table. 2 .

ncategory

category	keywords
グルメショッピング	掲示板, 日記, サイト, 化粧品

という行があれば、カテゴリ「ショッピング」を選ぶと「¬(掲示板∨日記∨サイト∨化粧品)」(「掲示板」「日記」「サイト」「化粧品」を含まない) という検索条件がつくことになる。よって、上のcategoryテーブルからのキーワードと合わせると、

(ショッピング∨買い物∨注文) ∧ (食∨グルメ) ∧¬(掲示板∨日記∨サイト∨化粧品)

という検索条件で「ショッピング」カテゴリを判別することになる。

検索語の入力欄には、1個か空白で区切った複数個のキーワードを入力できる。システムは各キーワードにシソーラスを適用してOR(∨)で展開する。

例えば、ユーザが「ショッピング」カテゴリを選択して、「いか ちりめん」と検索語を指定したとすると、システムはこれらを展開して、

(いか∨イカ∨烏賊) ∧ (ちりめん∨ちりめんじゃこ∨縮緬∨しらす干し) ∧ (ショッピング∨買い物∨買物) ∧ (食∨グルメ) ∧¬(掲示板∨日記∨サイト∨化粧品)

なる検索条件を生成し、これ (をユーザが選択した検索エンジンの書式で書いたもの) をユーザ画面に表示する。ユーザはこれを確認しsubmitボタンを押すと、スクリプトはこれを検索エンジンに送り、検索結果を受け取って別ウィンドウに表示する。

一般に、

- $a_{11},\ldots,a_{p1}$をユーザが指定した検索語、
  $a_{11},\ldots,a_{1m_1}$=..=$a_{p1},\ldots,a_{pm_p}$をそれらの類義語群
- $b_{11},\ldots,b_{1n_1}$=..=$b_{q1},\ldots,b_{qn_q}$をユーザが選択したカテゴリのキーワード群
- $c_1,\ldots,c_r$をユーザが選択したカテゴリの除外キーワード

とすると、これらに対して生成する検索式は、

$(a_{11}\vee\ldots\vee a_{1m_1})\wedge\ldots\wedge(a_{p1}\vee\ldots\vee a_{pm_p})\wedge(b_{11}\vee\ldots\vee b_{1n_1})\wedge\ldots\wedge(b_{q1}\vee\ldots\vee b_{qn_q})\wedge\neg(c_1\vee\ldots\vee c_r)$

となる。

3 試作システムによる実験

3.1 辞書の構築

試作システムを使うにあたって

1. シソーラスにどのような類義語を登録するか。
2. カテゴリキーワード辞書、カテゴリ除外キーワード辞書にどのようなキーワードを登録するか。

を決める必要がある。

まず、どのような類義語、キーワードを使っても、ブーリアン検索式による全文検索により求めるページ群をアーカイブの中で正確に判別するのは無理がある。

関係するページをすべて見つけようとすると無関係なページも多く含まれることになる。また、無関係なページを除こうとすると関係するページまで除いてしまうことになりがちである。また通常、ユーザは検索エンジンの返す膨大な検索結果のうち最初の方しか見ない。一方、各検索エンジンはそれぞれの方法でページのマッチ度、重要度を評価し評価の高い順で表示しているので、この表示順を利用することができる。よって、検索エンジンが返す結果の上位数十ページに求めるページ群のめぼしいものが集まっているという状況を目指すことにした。

シソーラスについては、水産物に対し、カテゴリごとにシソーラステーブルを作り、ひらがな名、カタカナ名、漢字名、英語名、学名などを登録した。たとえば「まいわし」の「グルメショッピング」カテゴリ用のシソーラスエントリは、

まいわし、真鰯

「アカデミック」カテゴリ用のシソーラスエントリは、

マイワシ、"Sardinops melanostictus"

といったぐあいである。

語xに対しyを含むページを検索するべきときシソーラス

テーブルにタプル(x,y)を登録する（この意味でyはxの類義語であるということにする）。このとき(x,x),(y,y),(y,x)も登録すべきであろうか。また、(x,y)と(y,z)が登録されているとき(x,z)も登録すべきであろうか。すなわち、シソーラステーブルTを同値関係と見なすならばTは次の同値関係の条件を満たしているべきであろう。

- 反射性：If $(x,y)=T$ for some y, then $(x,x)=T$.
- 対称性：If $(x,y)=T$, then $(y,x)=T$.
- 推移性：If $(x,y)=T$ and $(y,z)=T$, then $(x,z)=T$.

しかしながら、たとえばカテゴリ「レシピ」に関して「イカ OR 烏賊」の方が「いか OR イカ OR 烏賊」よりWEBの検索結果がよいとする。すると、「いか」に対する類義語は「イカ、烏賊」とすることになる。シソーラス辞書には(いか, いか)はないので反射性は崩れることになる。また、シソーラス辞書に(いか, イカ)はあっても(イカ, いか)はないので対称性も崩れる。

そこでシソーラスは同値関係としてではなく以下のようにして構築する。上の例でいえばまずイカに関して見出し語「いか、イカ、烏賊」をとり上げ、それらの各々に対して「イカ、烏賊」を類義語として登録する。

カテゴリ辞書の内容に関していくつかの検索エンジンに対してシステムを試した結果、以下のような内容にひとまず落ち着いた。

Table. 3 .

category

category	keywords
グルメショッピング	ショッピング, 買い物, 注文
レシピ	レシピ, 材料, 作り方

Table. 4 .

ncategory

category	keywords
グルメショッピング	掲示板, 過去, バックナンバー, 日記, 地域情報, 化粧品, 釣り, ペット, 観光, 旅行, 感想
レシピ	出版社, 著, 掲示板, 過去, バックナンバー, 日記

「フィッシング」「アカデミック」カテゴリについては、適当なカテゴリキーワードがないと判断しカテゴリ辞書には登録せずディレクトリのみとした。

辞書を構築するに当たっては検索エンジンに実際に試して得られた以下の知見を参考にしている。

1．シソーラスにある語の類義語を数多く登録してしまうと、より多くの類義語を含むカテゴリとはあまり関係のないページが上位に表示されてしまう。たとえば、「魚のさばき方」を「魚のさばき方∨魚のさばきかた∨魚の捌き方」に展開するのは必ずしもよい結果をもたらさない（「魚のさばき方」のページが主であるから）。
2．あるカテゴリのカテゴリキーワードも多すぎると数多くのキーワードを含む特定のページ群が偏って上位に表示される傾向があるので、多ければいいというわけではない。たとえば、「レシピ」カテゴリで「煮る」「炒める」「焼く」などのキーワードを含めると、いろいろな調理法の載っているページ、いろいろな料理が多く載っているページが上位に来て、単一の料理について詳しく載っているようなページが後へ追いやられる。
3．「ショッピング」カテゴリの判別はカテゴリキーワード群として「ショッピング」「買い物」「注文」(のOR結合)が水産物に関してはかなり有効である。しかし、OR結合して食品販売のページをうまく取り出せるキーワード群は見い出せない。食品販売のページに特有の語群が見当たらないのである。たとえば、「食」「海産物」「グルメ」のいずれかを含むとすると、多くの食品販売のページが除かれてしまう。
4．同様の理由で「水産」関連のページを効果的に抽出できるキーワード群も存在しない。よって、すべてのカテゴリで、ユーザの与えるキーワードが水産関連の語でない場合は、水産関連のページに絞り込めないことになる（よってその旨を検索ページに記している）。
5．検索エンジンごとに結果は違ってくる。特に大きいのは検索エンジンが結果を並べる順番である。たとえば、Yahooだと「いか」のような他語の一部と区別がつかないキーワードでも「イカ」とOR(∨)結合することで「いか」「イカ」をともに含むページが上位にくる。Googleで同様の検索式を試してみると「いか」ショッピングの検索で「いかり」とか「イカロス出版」とかが上位に出てきてしまう[2*]。

2＊ YahooはHITS[3],GoogleはPageRank[4] というようにリンク構造解析によるページ評価を表示順に反映させているといわれているが,Googleの方がリンク構造解析を優先させているように見える。この辺は確かな情報がないのではっきりしたことはわからない。

6．ショッピング関係のページではとくに美容関係が目立つところに置かれているので、美容関係のページを除くことで間接的に食品販売ページを選別する効果がある場合がある。また、ネット上では食品やショッピングの話題を取り扱う掲示板、日記のたぐいも多いのでこれらを除くことも効果的である場合がある。除外キーワードは常に効果があるわけではない。

3.2 各カテゴリ毎の調査

構築された試作システムについて以下の調査をした。

「その他」を除く各カテゴリ毎にいくつかの検索語を選び、それぞれについて、

- 検索語のみによる単純検索
- 検索語のシソーラス展開によるシステム検索1
- 検索語のシソーラス展開＋カテゴリキーワードによるシステム検索2
- 検索語のシソーラス展開＋カテゴリキーワード＋カテゴリ除外キーワードによるシステム検索3

を行い、

- (精度)上位50ヒットのうち検索語およびカテゴリに関係あるものの数
- (再現性)上位50ヒットのうちYahooカテゴリにあるものの数

を記録する[3*]。

検索エンジンにはYahooページ検索を用いた。Yahooページ検索での論理記号は、∨が＋、∨が*、∧¬が＃である。

精度も再現性も上位50ヒットを対象にしたのは、検索結果のうちユーザが実際に見るページ数がこのくらいと想定したからである。

精度の指標としての「検索語およびカテゴリに関係のあるもの」には、指定された検索語とカテゴリに直接関係あるリンクが含まれるページも該当するものとする。

再現性の指標として「Yahooカテゴリにあるものの数」としたのは、検索エンジンにYahooページ検索を用いたからである。

「グルメショッピング」カテゴリの調査結果

再現性に使ったYahooカテゴリは、主に「ショッピング＞マグロ」(登録数15) である。

上位50ヒット中の精度と再現性の調査結果を以下の表に示す。

Table. 5.

	単純検索		システム検索1		システム検索2		システム検索3	
検索語	精度	再現性	精度	再現性	精度	再現性	精度	再現性
本まぐろ	34	5	20	0	37	4	45	4
たらばがに	50	5	49	3	46	5	50	10
ふぐ	22	13	21	14	33	3	48	9
しらす	16	5	19	1	26	2	35	3
海苔	39	23	49	33	46	13	50	9

3＊ ちなみにYahooを含む主要検索エンジンが表示する「ヒット数」は当てにならないので評価の指標にはしない。

「本まぐろ」は「本まぐろ，クロマグロ，本マグロ」、「たらばがに」は「たらば，タラバ」、「ふぐ」は「とらふぐ，フグ」、「しらす」は「ちりめん，しらす干し，ちりめんじゃこ，縮緬」、「海苔」は「海苔，のり」にシソーラス展開される。

販売のページは非常に多い。例えば「かに」という指定では多すぎるので「たらばがに」の指定にしたが、このぐらい限定してもまだ多い（数百は下らない）。

「オークション」をカテゴリキーワードに含めるとリンク切れで無関係数が増す。オークションはネットの特徴でもあるし数も多いのであるが含めなかった。含めなくてもある程度はヒットする。

「本まぐろ」「たらばがに」「海苔」は単純検索でもかなりの精度を持っているが、全体的にシステムの効果が出ているといってよいだろう。シソーラス展開のみでは効果が薄いがカテゴリキーワード、除外キーワードと合わせて精度を向上させている。システム検索では再現性が多少落ちているきらいがあるが、これはYahooカテゴリで使用される語とシソーラス展開が合わないためもあると思われる。シソーラス展開でより広い範囲を探る効果はこの調査では十分見ることはできない。

システム検索１とシステム検索２では、日記、weblogのたぐいが上位に多く出てくるのが単純検索との違いである。カテゴリに合わせたシソーラス展開にしているにもかかわらず、シソーラス展開がそういったページにヒットしやすくしている。

「レシピ」カテゴリの調査結果

上位50ヒット中の精度と再現性の調査結果を以下の表に示す。

Table. 6 .

	単純検索		システム検索１		システム検索２		システム検索３	
検索語	精度	再現性	精度	再現性	精度	再現性	精度	再現性
鯖鮨	5	1	3	0	14	0	15	0
鯖のおろし煮	29	0	43	0	40	0	47	0
鰤大根	25	0	23	0	46	0	48	0
釣魚料理	34	0	22	0	-	-	46	0
魚のさばき方	35	1	-	-	-	-	45	5

「鯖鮨」は「さばずし，さば鮨，鯖鮨」、「鯖のおろし煮」は「さばのおろし煮，鯖のおろし煮」、「鰤大根」は「鰤大根」にシソーラス展開される。

「釣魚料理」はディレクトリ検索「＝(釣魚料理+=釣り魚料理⇉＃(amazon＋書籍＋出版＋日記＋バックナンバー＋過去＋西東社)」であり、カテゴリキーワードは用いていない。「魚のさばき方」はディレクトリ検索「＝(魚のさばき方)＃(出版＋著＋書籍＋日記＋教室＋受講＋学習＋学ぶ)」であり、カテゴリキーワードは用いていない。

他に、「魚のおろし方」という「魚のさばき方」と同種のものもディレクトリに登録している。「魚の下ろし方＋魚のさばき方」という検索より別々に検索した方が明らかに結果がよいからである。

「フィッシング」カテゴリの調査結果

上位50ヒット中の精度と再現性の調査結果を以下の表に示す。

Table. 7.

	単純検索		システム検索１		システム検索２		システム検索３	
検索語	精度	再現性	精度	再現性	精度	再現性	精度	再現性
投げ釣り, 釣り方	30	2	-	-	-	-	45	1
めばる, 釣り方	22	1	-	-	-	-	43	1
釣魚図鑑	45	4	-	-	-	-	49	5

すべて、ディレクトリ検索で、「投げ釣り，釣り方」は「＝(投げ釣り＋釣り方)＃(価格＋定価＋日誌＋日記＋リンク集＋情報＋投票)」、「めばる，釣り方」は「＝(メバル釣り)＃(釣行記録＋釣果情報＋出版社＋日記＋日誌)」、「釣魚図鑑」は「＝(釣魚図鑑)＃(出版＋著者＋リンク集)」である。

単純検索では「釣り方」を「投げ釣り*釣り方」「めばる*釣り方」と*でつないでいるが、ディレクトリ検索では＋でつなぐか、全く用いないでいる。また、釣りでは「めばる」より「メバル」の方が多く使われる。それらの効果が精度の面で出ている。ページの多様性の面でも「釣り方」で制限しない方がよいようである。

ウェブ上の釣魚図鑑は個人が釣った魚を掲載しているような小規模なものが多いが、単純検索でもかなりの精度はある。

「アカデミック」カテゴリの調査結果

上位50ヒット中の精度と再現性の調査結果を以下の表に示す。

Table. 8.

	単純検索		システム検索1		システム検索2		システム検索3	
検索語	精度	再現性	精度	再現性	精度	再現性	精度	再現性
水産生物(まいわし)	17	0	32	0	41	0	47	0
魚の図鑑	39	4	-	-	-	-	50	12
無脊椎動物図鑑	12	1	-	-	30	3	40	3
水産資源学	17	0	28	1	30	1	33	1
水産物流通	24	0	-	-	46	0	46	0

すべてディレクトリ検索との比較。「水産生物」ディレクトリの「まいわし」は「マイワシ＋“Sardinops melanostictus”」と展開される。結果、生物学的、資源学的なページにヒットした。「まいわし」のみでは、ショッピング、うんちく系、流通などのページが入り交じる。「魚の図鑑」は「＝(魚図鑑)＃(価格＋出版社＋書籍)」。「魚類図鑑」単独ではヒット数が少なく、＋で「魚図鑑」に加えても単に精度が落ちる。「無脊椎動物図鑑」は「＝(無脊椎動物図鑑＋エビ図鑑＋カニ図鑑＋甲殻類図鑑)＃(価格＋出版社＋書籍)」。「水産資源学」は「＝(水産資源学)*(研究)＃(募集＋学ぶ＋コース)」。水産資源学のページが多くないためか精度はこれ以上上がらないと思われる。「水産物流通」は「＝(水産物流通)*(研究)」。「水産物流通」のみの単純検索では精度は落ちるが流通統計年報のページにいくつかヒットする。

4　課題と考察

まず前節の調査についていえば、どのカテゴリの調査においても検索語によって差はあれシステムの効果をある程度デモンストレーションできているといってよいかと思われる。しかしながら、上の調査ではヒットしたページの質の評価方法、再現性の評価方法は全く不十分である。また調査に使った検索語はシステムの有効性を示す程網羅的ではない。このような検索システムの評価方法をもっと吟味する必要がある。

また、システムを実用レベルで効果的に使えるようにできるまでにはいろいろ問題点がある。以下そういった問題点とその解決に向けての課題について述べる。

- **辞書の構築方法について**　辞書はトライ＆エラーによりヒューリスティックに作成していて構築法はシステマチックではない。作成の手間と将来インターネットの状況が変化した場合への適応などの点で問題を残している。辞書作成のノウハウはカテゴリの設け方と検索論理式の構造に大きく依存するが、これらについてはまだ深く吟味できている段階ではない。ターゲットとなるコーパスを分類した上で検索式を作成することまでをどの程度自動化できるのか、今後の課題としたい。
- **検索条件について**　除外キーワードは全体的に効果は小さいが検索語によっては大きい効果があるので一定の存在理由はあるだろう。また、サポートしたブール論理式において論理否定(¬)の入り方が限定的すぎるきらいはある。主要検索エンジンのサポートする論理式はかなり限定的であるが、命題論理式のより完全に近いサポートも視野に検索条件をどのような形式で表現するのがよりよいのかの吟味が必要である。

主要検索エンジンの表示する結果は検索に使ったブール式を厳密に反映していないことがある。いずれにせよYahooなどの検索エンジンの用いているプログラムロジックは詳らかにされていないので検索手法を評価する土台としては問題がある。

また、論理式$A \vee B$の扱いがA，Bをともに含むページ優先になっている検索エンジンが多いようであるが、これではキーワードの数を限定せざるを得ず本編の方法は十分効果が発揮できない。$A \vee B$に関してA含むページをBだけを含むページより優先する条件をサポートするとか、検索条件指定のきめ細かな工夫も必要である。

これらの問題に対しては目的に最適化されたオリジナルの検索エンジンを使うことが根本的な解決に必要であろう。

- **水産分野への限定、カテゴリの設け方に関して**　キーワードによるブーリアン検索式ではカテゴリや分野を判別することが不可能なことが多い。分野を特徴づけるキーワードがないために網羅的に関連キーワードをOR結合して増やすとかえって結果が悪くなることがほとんどなのは、上述のようにキーワードの優先順位など複数のキーワード群に関する細やかな検索条件の指定ができないことでキーワード群に込めた意図が反映できないことも一因である。

「フィッシング」「アカデミック」ではカテゴリ検索をあきらめてディレクトリ検索のみにしたように、もっときめ細かい分類によるアプローチが必要と感じた部分が多くあった。インターネット上に情報量の多いもの(カテゴリ)に限定してみてもいいかもしれない。しかし、インターネット上に情報が少なかったりしてうまくいかないカテゴリも将来はどうなるかわからない面もある。数年前まではインターネット上にはコンピュータに関する情報以外にはあまりめぼしい情報はなかったのである。本研究のような取り組みでは分野を近視眼的に限定するのではなく創造的な見極めが必要とされているように思われる。

まぐろのショッピングページなど非常に情報量の多いものは、さらに情報量を絞る工夫が欲しい。「評判」をコーパスから取り出し[5]評価基準に入れて上位の

ものから並べるなど、カテゴリに合わせた価値基準で自動で計算できるような評価方法があるといいが、水産物ショップの評判がネット上に存在しているのはごく少数しか確認できていない。

テキストから単語単位ではなくパターンなどの構造的言語情報を自動的に読み取る技術が研究されているが、ウェブページから価格や重量その他アピール点などの製品情報を読み取って評価するメカニズムを作ることも考えられる。しかし水産物は電化製品などとは違い同じ「たらばがに」でも大きさ、産地、鮮度、加工技術の他に目利きを必要とされる微妙な個体差まで客観的に証明しがたい多種の要素が製品の質を左右しており、少なくとも対重量価格だけで比較できるものではない。しかし、価格などのウェブページから読み取れる情報に一定の利用価値はあるであろう。

本編のディレクトリ検索のような形でいわゆる目利きによる選別をした上で提示するのは、すべての分野のエキスパートを揃える必要がないから分野を限定するメリットが生かしているともいえる。自動的に収集できるような情報を生かしつつ目利き的な評価を提示できればよいのかもしれない。

- **オリジナルの検索エンジンについて**　クローラー、全文検索およびページ評価プログラムを目的に合わせて作ることの必要性を述べてきた。インターネット上のウェブページを広く収集し、膨大なコーパスから全文検索をかけるにはそれ相当のコンピュータ資源が必要である。分野を限った検索エンジンであれば収集する範囲もリンク情報などを元に限定できるかもしれない。分野の絞り方がここでも重要になる。

参考文献

1）J. K. Wen, J. Y. Nie, H. J. Zhang, Clustering User Queries of a Search Engine. In Proc. of WWW 10, pp. 587-596, 2001.

2）江口浩二，自然言語による情報アクセス技術：2.Web検索の技術動向と評価手法，情報処理学会誌，Vol. 45, No. 6, pp. 569-573, 2004.

3）J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

4）L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking : Bringing Order to the Web, http://www-db. stanford. edu/~backrub/pageranksub. ps, 1998.

5）藤村滋，豊田正史，喜連川優，Webからの評判および評価表現抽出に関する一考察，情報処理学会研究報告 2004-DBS-134(II), pp. 461-468, 2004.

5）徳永健伸，情報検索と言語処理，東京大学出版会，p145, 1999.