WSA: Web 検索代数に基づく情報検索支援システムの実装

WSA : Implementation of Information Retrieval Support System based on Web Surfing Algebra

片山 聡一郎♠ 遠山 元道♦
Soichiro KATAYAMA Motomichi TOYAMA

Web から自分が欲しい情報を取り出すためには，その情報要求によってはいくつもの Web ページをネットサーフィンするなど複雑な検索を行わなければならない事がある．この様な検索における検索の効率化やユーザの検索における負担軽減のために，ユーザがインターネットを利用して情報要求を満たす行為を定性的に表現する数学的基盤として，Web Surfing Algebra(WSA)を提案されている．

本論文では WSA に基づく情報検索支援システムを提案する．そしてこのシステムを利用した検索例を示す事により，従来の検索手法では答えを得る事が難しかった情報要求の答えをこれまでよりも容易に得られる事を実証し，WSA 及び WSA に基づく情報検索支援システムの有効性を示す．

Web is a huge and variety information carrier. So many users have to do a complicated search that is do netsurf of many Web pages for gainning information which you want. Web Surfing Algebra(WSA) was proposed which is mathematical representation of action of collecting information with internet, so that users can do netsurf easily.

We have proposed supporting system of information retrieval based on WSA in this paper. We showed with actual example that WSA and this system are effectivity.

1. はじめに

Webは巨大かつ多様性に富んだ情報メディアであり，自分が欲しい情報を取り出すためにgoogle[1]やYahoo![2]などのWeb検索エンジンが用いられている．

Web検索においてユーザは検索クエリを通じて自分の情報要求を検索エンジンに伝えている．多くの場合，ユーザの情報要求に含まれている単語を適切に組み合わせた検索クエリを検索エンジンに1回与える事により自分の情報要求を満たす事ができる．例えば「数独の解き方を調べたい」という情報要求に対しては，ユーザは検索エンジンに「数独 AND 解き方」というクエリを与える事により返ってきた結果から，解き方を知る事ができる．

しかしながらユーザの要求は様々であり，例えば「数独のようなパズルの問題をいろいろと解いてみたい」という情報要求に対しては，情報要求に含まれる単語を組み合わせて「数独 AND パズル」といった検索クエリを検索エンジンに与えても，他のパズルの問題は検索結果にほとんど現れない．

従来この情報要求を満たすためには，一旦「数独 AND パズル」などで検索を行い，返ってきたページからパズルの具体的な名称を探し出し，その名称について改めて検索を行う必要がある．この方法を行うためには，パズルの名称を調べた結果をなんらかの形で保存しておき，その結果を利用して検索し直さなければならない．

情報要求を満たすために1回検索クエリを検索エンジンに与えるだけでは知りたい情報を得る事ができない，複数のWebページをネットサーフィンする必要がある複雑な検索を支援するシステムを構築するためには，ユーザがある情報要求を持ってから，その要求を満たすまでの検索の手順及び検索結果の履歴を管理し扱える必要がある．

そこでユーザのネットサーフィン行為に対応した代数であるWeb Surfing Algebra(WSA)が片山ら[3]によって提案された．WSAを利用する事により，ユーザのWebにおける検索の手順及び検索結果を扱える．

本論文ではWSAに基づく情報検索支援システムを提案し，それに従って実装を行なった．その結果従来の検索システムでは答えを得る事が困難であった情報要求を従来よりも容易に答えが得られる様になった事を示す．

2. Web Surfing Algebra(WSA)

Web Surfing Algebra(WSA)について紹介をする．なお詳細については文献[3]で説明されている．

2.1 Web Surfing Algebra における用語の定義

Web Surfing Algebra(WSA)とはユーザのネットサーフィン行為及び，システムで行われる処理について以下に述べる関数，演算子に対応させた代数である．

WSAにおいて用いられる語を以下に説明する．

・word：単語
名詞，動詞，形容詞など，意味を持った文字列．

・query：クエリ
検索エンジンにおける検索クエリ．

・result：検索結果
検索エンジンに対して検索クエリを与える事によって得られる検索結果．Webページのタイトル，URL，ページの要約からなる組．

・attribute：属性
名詞，形容詞など特徴を示す文字列．

・集合
0個以上の要素からなる順序つきの集合．

・WSA式
WSA式はユーザのネットサーフィンにおける情報に対する要求及びそのためのプロセスをWSAの関数，演算子によって対応付けしたもの．

2.2 WSAで使用される関数及び演算子

WSAで使用される関数及び演算子を紹介する．なお演算子と関数の違いはユーザが選択する行為が含まれているものを演算子，含まれていないものを関数と定める．

・Q(queries)：Q関数
queryを検索エンジンで検索し，検索結果の集合を得る．入力が複数のqueryからなる場合1つずつ処理していく．
入力･･･検索クエリ(queries)
出力･･･検索結果集合(results)

♠ 学生会員 慶應義塾大学大学院理工学研究科修士課程
katasou@db.ics.keio.ac.jp
♦ 正会員 慶應義塾大学理工学部情報工学科
toyama@ics.keio.ac.jp

・σ(results)：σ演算子
　入力された検索結果からユーザはURLを選択する．
　入力･･･検索結果集合(results)
　出力･･･URL集合(URLs)
・$\pi_{\{attribute\}}$(URLs)：π演算子
　入力されたURLのWebページを開きユーザに提示する．ユーザは指定された属性を持つ単語をピックアップする．
　入力･･･URLの集合(URLs)
　出力･･･指定された属性(attribute)を持つとユーザが判断し選択した単語集合(words)

WSAで使用される関数及び演算子には他にも以下のものが用意されている
・dis (queries)　/ con (queries)
・words ∪ words / words ∩ words
・words ⊔ words / words ⊓ words
・asc (words)　/ dec (words)

得られた結果の全要素の上位n件のみを出力とする場合，関数，演算子に添え字nを付ける事によって表現する．但し集合の要素数がn件に満たない場合は集合の要素全てとする．
例えばQ_{10}(query1)はquery1を検索エンジンに検索クエリとして与え，返ってきた検索結果の上位10件という事である．

3. Web Surfing Algebra に基づく情報検索支援システムの構成

本システムの構成を図1に示す．

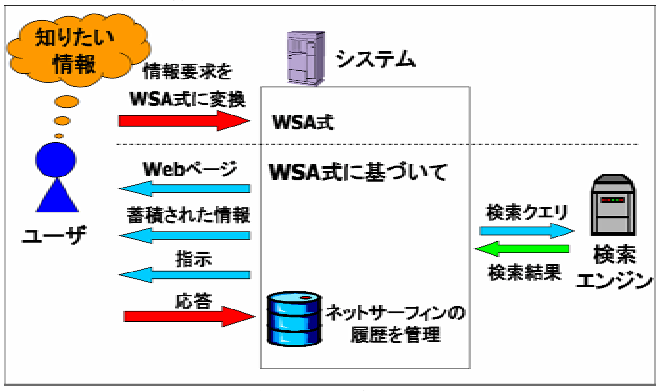

図1　WSAに基づく情報検索支援システムの構成

Fig.1　Construction of Information Retrieval Support System based on Web Surfing Algebra

システムの流れは以下の通りである．
1．ユーザの情報要求をWSA式に変換しWSA式をシステムに与える．
2．システムはこのWSA式に基づいて処理を行なう．
3．WSA式に従ってシステムが自動で処理を行える関数の部分についてはシステムが自動で処理を行い，ユーザによる選択が必要な演算子の部分ではユーザに作業の内容を指示する．
4．ユーザは指示された作業指示に応答する．
5．応答結果はシステムに履歴として管理される．
6．3．から5．のユーザとシステムとのやり取りを繰り返し，ユーザの情報要求が満たされた段階で終了する．

ユーザは始めにWSA式をシステムに与えた後はシステムからの指示に従って応答する事により情報要求を満たす事ができる．

4. 実装

現段階では式の処理を遡るための処理を除いた部分について各関数，演算子の実装が終わっており，それらを組み合わせる事により，あらかじめ定められているWSA式に対応した検索支援システムを構築する事が可能である．
各関数，演算子はPHP5.0.5及びHTMLにより記述されたWebページにより実装されている．また，システムが利用される状況を考え各関数，演算子間の入力出力のやりとりを行うためのバッファにSQLiteを利用した．
以下，各関数，演算子毎に見ていく．
・Q(queries)
図2の様な入力フォームが表示され，関数に入力されたqueryが入力フォームに入る（複数のqueryが入力されている場合には1つずつ処理する）．
"Search!"をクリックする事により，そのqueryが検索エンジンに与えられ，検索結果が関数の出力としてバッファに保存される．"完了"をクリックする事により次の関数，演算子のページに移る．
検索エンジンにはYahoo!検索Webサービスのウェブ検索Webサービスを利用した[4]．1回の検索に対する返却結果の数は50に設定した．従って，Q_{50}(queries)に対応する．

図2　Q(query)の実行画面

Fig.2　Execution Screen of Q(query)

・σ(results)
入力された検索結果を図3の様に表示する．"OK"が押されるとユーザがチェックボックスを選択したURLが上にあるものから順番に出力用バッファ格納される．全てが格納し終わったところで次の関数，演算子のページに移る．

図3　σ(results)の実行画面

Fig.3　Execution Screen of σ(results)

・$\pi_{\{attribute\}}$(URLs)
入力された属性名，入力されたURLの集合が図4の様に表示される．ここでURLを選択する事により対応するページを別窓で開く．
"メモを開く"をクリックする事により別窓に図5の様に入

力フォームが表示される．この入力フォームを介して登録された単語は出力用バッファに格納される．これがこの演算子の出力となる．"完了"をクリックする事により次の関数，演算子のページに移る．

図 4　$\pi_{\{attribute\}}$(URLs)の実行画面(1)
Fig. 4　Execution Screen of $\pi_{\{attribute\}}$(URLs)　(1)

図 5　$\pi_{\{attribute\}}$(URLs)の実行画面(2)
Fig. 5　Execution Screen of $\pi_{\{attribute\}}$(URLs)　(2)

・dis (queries) / con (queries)
入力された集合から要素を 1 つ取り出し，空の文字列に格納する．まだ集合に要素が残っている場合には，先程の文字列の後ろに OR(AND)を続け，新たに要素を 1 つ取り出しその後ろに続ける．これを集合の要素が無くなるまで繰り返す．要素が無くなったら結果を出力用バッファに格納する．格納後自動的に次のページに移る．

・words ∪ words / words ∩ words
2 つのバッファを介して 2 つの単語集合が入力される．それぞれの集合の要素を別々の配列に一旦格納する．その後その 2 つの配列に関して和集合（積集合）をとる事によって新たな配列を得る．それを出力用バッファに要素を 1 つ 1 つ順番に格納する，格納が終わったら自動的に次のページに移る．

・words ⊔ words / words ⊓ words
2 つのバッファを介して 2 つの単語集合が入力される．それらを図 6 の様に表示する．OK が押されるとユーザがチェックボックスを選択している単語に関して上にあるものから順番に出力用バッファに格納される．全てが格納し終わったら次の関数，演算子のページに移る．

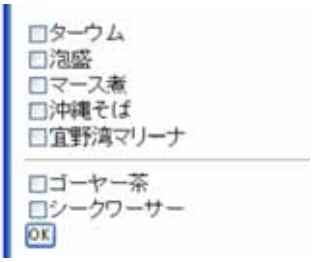

図 6　words ⊓ words の実行画面
Fig. 6　Execution Screen of words ⊓ words

・asc (words) / dec (words)
入力された単語集合の要素を 1 つずつ取り出し配列に格納する．その後その配列の要素を昇順（降順）に並べ替え，出力用バッファに配列の要素を 1 つずつ格納していく．格納が終わったら自動的に次のページに移る．

5. 実行例

・実行例 1
「スピードコントロール機能のついている MP3 プレーヤーの値段を調べたい」という情報要求を WSA 式で表すと
$\sigma(Q_{50}(\pi_{\{機種名\}}(\sigma(Q_{50}(スピードコントロール\ AND\ MP3\ プレーヤー)))\ AND\ (値段\ OR\ 価格)))$
となる．式の処理の流れ及び各々の入力出力とバッファの関係を図示したものが図 7 である．
最終的な結果として得られた URL のページから「FG200-512」「FL350-512B」「iFP-799」などの機種についての値段の情報を得られた．
この情報要求に対して「スピードコントロール AND MP3 プレーヤー AND 値段」として検索した場合には，本システムを利用した場合と比較して情報を得られた機種の種類が少なかった．その原因としては「スピードコントロール」という単語は製品の機能紹介のページに存在する単語であるが，機能紹介のページに必ずしもその商品の値段が載っているとは限らないため，「スピードコントロール」「値段」2 つの単語が存在するページは限られてしまうからと考えられる．

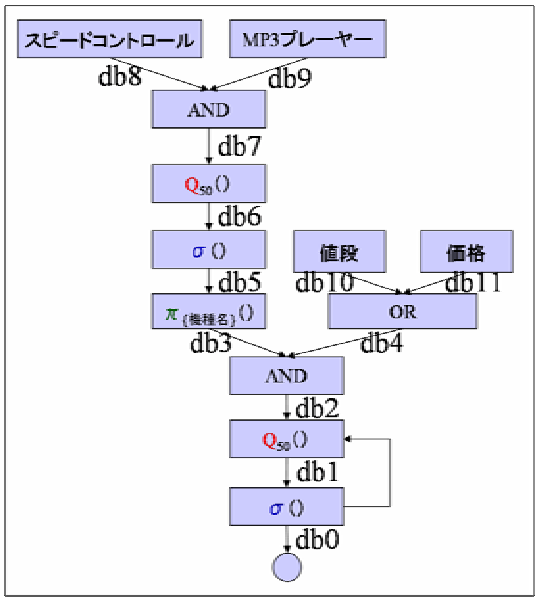

図 7　実行例 1 の式の処理の流れ
Fig. 7　Process Flowchart of example 1

・実行例 2
「宜野湾の名物を体験できる新宿の店を調べたい」という情報要求を WSA 式に変換する際，様々な式が考えられるがその一例を示すと
$\sigma(Q_{50}(dis(\pi_{\{名物\}}(\sigma(Q(宜野湾\ AND\ 名物)))\ ⊓$
$\pi_{\{食べ物\}}(\sigma(Q_{50}(宜野湾\ AND\ グルメ))))\ AND\ 新宿))$
となる．式の処理の流れ及び各々の入力出力とバッファの関係を図示したものが図 8 である．
選択インターセクションの部分で「ターウム」「泡盛」「沖縄そば」「シークワーサー」を選択した結果，一番外側の Q 関数では"新宿 AND（ターウム OR 泡盛 OR 沖縄そば OR

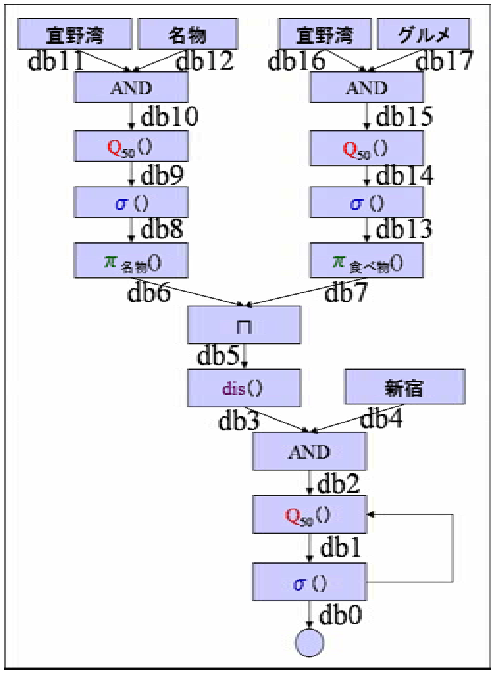

図 8　実行例 2 の式の処理の流れ
Fig.8　Process Flowchart of example 2

シークワーサー）″という query により検索が行われ，最終的な結果として得られた URL のページから「かりゆし」「島の人」といった店についての情報を得られた．

この情報要求に対して「宜野湾 AND 新宿」や「宜野湾 AND 名物 AND 新宿」と検索しても，知りたい情報はほとんど得られなかった．

6. 考察

従来の Web 検索手法との比較を行う事により本システムを利用する事による有効性を示す．

本システムが従来手法よりも優れている点として，ある検索結果を次の検索に利用できる点が挙げられる．従来手法では，ある検索結果を次の検索に利用するためには一旦検索結果を記憶するかどこかに一時的にコピーしておき，次の検索をする際，先程コピーしたものを再びコピーする必要があった．それに対しては本システムでは，検索結果から得られた単語を登録するための初めの 1 回のコピーや入力により，登録されたものを次の検索結果に利用する事ができる．

また検索結果を関数 1 つで簡単に OR，AND で結合する事ができるなど，従来の検索手法ではユーザ自身で複数の単語を組み合わせなければ作成する事ができなかった複雑な検索クエリ作成の補助をシステムで行える点も有効な点である．

一般に Web の検索エンジンに用いられる検索クエリの語数は少ない事が調査で明らかになっている[5]．この原因としては，自分の情報要求を満たすために必要な検索クエリの検索語が多くの場合 1 語で済むという理由もあるが，それ以外に，何語にも渡る検索クエリを作成するのは難しいため作成されていないという理由もあると考えられる．従って，本システムによって複雑なクエリを従来よりも簡単に作成できる事は，Web における情報探索行動の可能性を広げていると言える．

従来手法では検索結果から 1 ページずつ URL を選択して Web ページを見るのに対して，本システムでは検索結果から先にまとめて URL を選択し，後でまとめて Web ページを見ていく．この点に関しては複数のページの情報を集める必要のない情報要求の場合，本システムの様に前もって URL をまとめて選択するという方法は従来手法と比較して無駄な作業が多くなってしまうと言える．しかしながら，答えを得るために複数のページの情報を必要とする情報要求に対しては，URL を選択していくつもの Web ページを開く段階では URL を選択するだけで，従来手法の様に″戻る″ボタンを押す必要がない．

ユーザがネットサーフィンをしている最中に利用する″履歴″，″戻る″，”ホーム”などツールバーのボタンの中で最も使われているのは”戻る”であるという調査があり[6]，本システムを利用する事により”戻る”ボタンを押す回数を減らす事ができるというのは Web 検索において有効である事を示していると言える．

7. おわりに

本研究では WSA に基づく情報検索支援システムを提案した．そして WSA の関数，演算子を実装し，それらを利用して情報要求の答えを得る例を示す事により，従来の検索手法では答えを得る事が難しかった情報要求の答えを本システムを利用する事により得られる事を実証し，WSA 及び WSA に基づく情報検索支援システムの有効性を示した．

今後はこれまでよりも様々な情報要求に対応できるようシステムの実装をさらに進めるとともに、本システムのインターフェイスの改良を考えている．

[文献]

[1] L.Page, S.Brin, R.Motwani, T.Winograd: “The PageRank Citation Ranking”, Stanford Digital Library Technologies, Working Paper SIDL-WP1999, pp.414-425 (1999).
[2] Yahoo! home page, http://www.yahoo.com
[3] 片山聡一郎, 遠山元道: “Web Surfing Algebraにおける関数および演算子の性質について”, 日本データベース学会Letters Vol.4, No.4, pp.29-32 (2006).
[4] Yahoo! デベロッパーネットワーク, http://developer.yahoo.co.jp/
[5] 原田昌紀, 佐藤進也, 風間一洋: “索引篩法 - 大規模サーチエンジンのための高速なランキング検索法”, DEWS2003 (2003).
[6] E.H. Chi, P. Pirolli, J.E. Pitkow: “The scent of a site: A system for an-alyzing and predicting information scent, usgae, and usability of a web site”, Proceedings of ACM Conference on Human Factors in Computing Systems, pp. 161-168, (2000).

片山 聡一郎　Soichiro KATAYAMA
2006 慶應義塾大学大学院理工学研究科開放環境科学専攻修士課程修了．2004 慶應義塾大学理工学部情報工学科卒業．データベースの研究に従事．

遠山 元道　Motomichi TOYAMA
慶應義塾大学理工学部情報工学科専任講師．博士（工学）．1984 慶應義塾大学大学院博士課程単位取得退学．主にデータベースシステムの研究に従事．IEEE Computer Society, ACM，情報処理学会，日本ソフトウェア科学会，電子情報通信学会，日本データベース学会会員．