

ネットワークアクセス行動の DB 化と WWW 検索への応用

西山 顯[†] 川越 恭二^{††}

[†]立命館大学大学院理工学研究科

^{††}立命館大学理工学部

概要

WWW 技術の普及により、様々な情報を即座にしかも大量に入手できるような環境が存在する。そして、膨大な情報の中から必要な情報を探し出す情報検索技術の重要性は増大し、数多くの研究が行われている。特に、WWW 特有のリンク構造を用いた情報検索技術が検索精度の向上が期待できるために大きく注目されている。しかし、従来の情報検索技術では、公共的な信頼性のある、あるいは常識的なページが検索されることが多いが、ユーザがこのような情報を求めず個人的な特定領域の情報を求めている場合には満足いく情報が得られないものとする。そこで、本稿では、ユーザ自身が WWW 上で行う行動をデータベースに格納し、それを他の利用者が情報検索の際に利用することで、ユーザによる情報の取捨選択を効率化する方法を提案する。

WWW Access Logs Databases and its Utilization for Improving WWW Retrievals
Akira Nishiyama and Kyoji Kawagoe
Ritsumeikan University

Abstract

Recent Information Technology development enables us to quickly obtain various kinds of information among an enormous amount of WWW data set. There is much research work on the information retrievals technology with which you can search necessary information from such data set. The current technology in general aims at searching the reliable, public or common kinds of information useful for a number of users. However, for an individual who needs some information in its specific field and special only for the individual, he or she needs a lot of WWW accesses or obtains no satisfying information. In order to solve such a problem, we propose a new WWW Access Support method. In this method, we store WWW user access logs in an access log database in a structural way. Then with sharing such databases, a WWW access user first get access to the databases in order to find appropriate access process, then get access to the WWW sites in such a way as shown in the obtained access process. It is also described that the number of WWW accesses can be reduced by the preliminary experiment.

1 はじめに

WWW 技術の普及により、様々な情報を即座にしかも大量に入手できるような環境が存在する。そして、膨大な情報の中から必要な情報を探し出す情報検索技術の重要性は増大し、

数多くの研究が行われている。特に、WWW 特有のリンク構造を用いた情報検索技術が検索精度の向上が期待できるために大きく注目されている。

現在 WWW の情報を検索する場合、一般的には `google` などの検索エンジンを使用する。検

検索エンジンは、ユーザから入力された検索語とページの索引語を比較し、類似度を求め、これをページの評価として結果を出力するが、ページの評価をそれに含まれる単語のみで求めているためにユーザは検索結果に満足できないことが多い[2]。そこで近年、WWW 特有のリンク構造に着目してページの評価を求める方法が提案され、検索精度の高さなどから注目されている[3][4]。

しかし、これらの方法は基本的に、多くのユーザが必要としているページがより重要であるという考え方に基づいていると考える。したがって、その検索結果は公のページや権威的なページの検索に偏る傾向がある[5]。これにより、従来の検索エンジンを使用したWWWアクセスでは、ユーザが真に必要とするページにたどり着くには、検索後のユーザ自身による情報の取捨選択作業が重要となっている。特に、ユーザが個人的な特定領域の情報を求めている場合には満足いく情報が得られないものと考ええる。

本稿では、ユーザ自身が行う取捨選択作業をデータベースに記録し、それを構造化した上で共有化し、他ユーザがこのデータベースを検索・活用することにより、ユーザが行う情報の取捨選択作業の軽減と情報検索の効率化を図るWWWアクセス支援方法を提案する。

2 行動 DB によるWWWアクセス支援システム

2.1 基本的考え方

本稿で提案する行動 DB による WWW アクセス支援システムでは、ユーザが普段行うネットワークに対して行う操作の行動履歴を利用する。ここで操作の行動履歴とは、具体的には、ユーザが WWW 利用時にブラウザ上で行うクリック、戻る、検索ボタンを押す等の操作の履歴を指す。

システムはこれらの行動履歴をユーザ各自が導入するプロキシサーバから取得し、データベースに保存する。そして、保存されたユーザの行動履歴をモデル化した行動プロセスモデルを作成する。

ユーザが行った行動履歴は、それぞれのページについてその内容を示すキーワードを抽出し、そのキーワードから各ページ間での類似度を算出する。算出した類似度を用いて、一定以

下の数値を持つ部分を行動履歴から切り離すことにより意味的にまとまりのあるアクションクルー(Action Clue)を得る。

これらのアクションクルーを他のユーザが検索することにより、従来の検索における検索結果の取捨選択作業を軽減し、求める情報に対して短時間に効率よく到達できると考えられる。

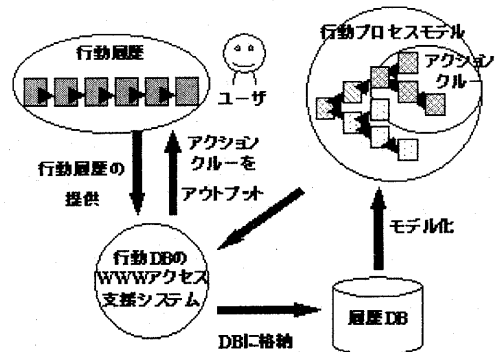


図 1: 全体概要図

2.2 ユーザの行動履歴取得方法

WWW アクセスにおいて、ユーザの行動履歴を得る方法は以下に示す client logs, network logs, server logs, proxy logs の 4 つに分類できる[6]。

- (1). ユーザの利用する Web ブラウザ操作の監視(client logs) :
これは、ユーザが利用する Web ブラウザが残す履歴情報や、Web ブラウザを改造、機能付加して行動履歴を得る方法である。
- (2). 通信パケットの内容の解析(network logs) :
ユーザが利用する PC とサーバ間や、その他様々な部分で HTTP のポートの通信内容をモニタすることで、ユーザの行動履歴を得る方法である。
- (3). Web サーバのログファイルの利用(server logs) :
Web サーバの分析ツール開発を容易にするために、ログファイルの共通形式として common logfile format が定義されている。これを利用してユーザの行動履歴を得る方法である。
- (4). プロキシサーバのログの利用(proxy logs) :

ユーザの利用するプロキシサーバのログを分析し、ユーザの行動履歴を得る方法である。

本システムでは、上記の4種類の方法の内、プロキシによる方法を使用する。これは、WWW 上でのユーザの行動履歴を得ることに關して以下の点を考慮したためである。

- Web ブラウザの履歴を利用する場合には、Web ブラウザの種類、バージョンごとに行動履歴を取得するプログラムを対応させなければならない。また、Web ブラウザの仕様変更にも対応していかなければならず、広範囲にユーザを獲得する上で障害となることが考えられる。
- 通信パケットの分析と Web サーバのログを利用する方法では、自分自身のログしか残らないために、複数ドメインにまたがる行動履歴を得るためには数多くの Web サーバの協力が必要で、現実的ではないと考える。
- プロキシサーバを利用する方法は、最も広範囲にユーザの行動履歴を集められるものと考えられるが、プロキシサーバを利用している全てのユーザが、自身の WWW 上での行動履歴の利用を認めているわけではなく、プライバシーの問題が発生する可能性がある。

すなわち、これらの問題を解決するために、本システムを利用するユーザは、ユーザ自身のクライアントマシンにプロキシサーバをインストールし、それを利用してユーザの行動履歴を得る。これにより、自身の行動履歴の利用を許可するユーザのみが本システムを利用することとなりプライバシーの問題は解決できる。

2.3 履歴情報の DB 化

本システムは、ユーザの行動履歴情報として以下の4つを1つのログレコード(log record)としてデータベースに格納する。

- (1). アクセスしたユーザの host 名(host name)
- (2). アクセスしたページの URL(access URL)
- (3). アクセス元ページの URL(referrer URL)
- (4). ページにアクセスした時間(access time)

すなわち、ユーザの行動履歴 $LOG=\{hi\}$, hi は以下のような要素からなる4ツ組でログレコードと呼ぶこととする。

$LOG=(Host, URL_{to}, URL_{from}, Time)$

本システムでは、この4つの情報を利用し、以下のアルゴリズムにより、1次元である履歴情報を階層的に視覚化したモデルである行動プロセスモデルに変換する。行動プロセスモデルはアクセスしたページとページの間を移動した履歴を元に作成した階層構造である。すなわち、行動プロセスモデルは複数のノードから構成される階層構造である。個々のノード nk には $access\ URL \in URL_{to}$ の情報を対応させる。

Step1: 行動履歴を host name 別に分割する。

$LOG=LOG1 \cup LOG2 \cup \dots$

LOG_i は i 番目の host name(ユーザ)に対応した行動履歴である。

Step2: 行動履歴 LOG_i のログレコードを時刻の昇順に取り出す。以降、 k 番目の時刻 T_k のログレコードを $Log(k)$ とする。 $Log(k)$ の $referrer\ URL \in URL_{from}$ によって以下の処理を実行する。

Step2-1:

$referrer\ URL$ が $Log(k)$ に存在しない場合には、 $Log(k)$ の $access\ URL \in URL_{to}$ を行動プロセスモデルの第一階層に設定する。

Step2-2:

$referrer\ URL \in URL_{from}$ が $Log(k)$ に存在する場合、以下の処理を実行する

- a. $Log(k)$ の $access\ URL \in URL_{to}$ が行動プロセスモデル上で $referrer\ URL$ が設定されている階層より上位の階層に存在する場合には、 $Log(k)$ を廃棄する。これは、ユーザは Web ブラウザの「戻る」操作で戻らず、戻るためにクリックしたと判断できるためである。
- b. $Log(k)$ の $access\ URL$ が行動プロセスモデル上で $referrer\ URL$ が設定されている階層より上位の階層に存在しない時には、その行動履歴の $access\ URL$ を行動プロセスモデル上の $referrer\ URL$ の下位階層に設定する。
- c. $Log(k)$ の $access\ URL$ が $referrer\ URL$ と同一の場合、その履歴は破棄する。

例を表1に示す。表1は、行動履歴のサンプルであり、この表の行動履歴を行動プロセスモデル化する。

表 1：行動履歴サンプル

番号	access URL	referrer URL
1	Page A	なし
2	Page B	Page A
3	Page C	Page B
4	Page D	Page A
5	Page A	Page D
6	Page F	Page A

まず、番号 1 の行動は、Page A にアクセスした行動である。referrer URL が設定されていないので行動プロセスモデルの第一階層に設定される。

次に番号 2 の行動では、アクセスした Page B のreferrer URL は Page A であるため、Page A のリンクをクリックして Page B に訪れたと考えられ、Page B は Page A の下のノードとなる。同じく番号 3 の行動では Page C は Page B の下のノードとなる。

番号 4 の行動では、referrer URL が Page A となっている。これはユーザが Page C を閲覧後、ブラウザ上の「戻る」ボタン等を利用して Page A に戻った後、リンクをクリックして Page D を訪れたと考えられる。よって、Page D は Page A の下のノードとなる。

番号 5 の行動は、Page A へ Page D からのリンクをクリックしたことを示す。しかし、Page D の上位ノードに Page A がすでに存在するため、ブラウザでの「戻る」行為に相当するため、この履歴は無視される。

したがって、表 1 の例から出力される行動プロセスモデルは図 2 に示す結果となる。

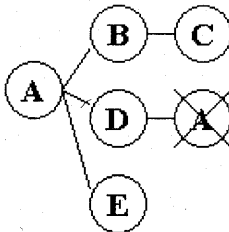


図 2：出力される行動プロセスモデル

2.4 行動プロセスモデルのアクションクルー一抽出

2.3 のアルゴリズムでは、ユーザから得られる行動履歴情報を階層構造に変換した行動プロセスモデルを得る。しかし、この行動プロセスモデルは、ユーザがある事柄について調査した時に訪れた全てのページを表しているの、調査したい目的に対して、多くの冗長が含まれている。

このため、他のユーザが同じ事柄について調査する際には、行動プロセスモデルをそのまま提供するのではなく、意味的にまとまりがあると考えられる部分で行動プロセスモデルを分割したアクションクルーを提供する。このことで、ユーザの取捨選択作業を軽減することができる。

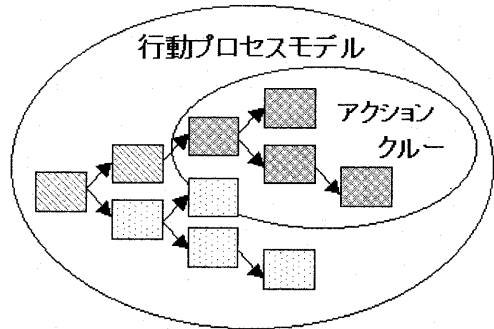


図 3：行動プロセスモデルとアクションクルー

行動プロセスモデルをアクションクルーに分割するために、以下の方法を使用する。まず、階層構造を構成するノード間の類似度を算出する。次に、階層構造をノード間の類似度がある閾値以下となる部分で分割する。このことで、意味的にまとまりのあるアクションクルーを生成することができる。

詳細な処理を以下に示す。

Step1:

階層構造を構成し直接接続されている 2 つのノード n_i と n_j に対応するページの類似度を算出する。

Step2:

類似度の値が閾値 ε 以下であれば、階層構造におけるこの 2 つのノード n_i, n_j 間の接続関係を切断する。

以降に類似度の算出方法を示す。

2.5 類似度の算出

ユーザが移動したページに対応したノード間の類似度を算出する。以下の3つのステップから構成される。

Step1: 各ページのキーワード抽出

Step2: 抽出されたキーワードの重み付け

Step3: ノード間の類似度計算

Step1: キーワード抽出

まず、ノードに設定されているページの内容から話題を特徴づけるキーワード群を抽出する。まず、日本語の文章を形態素分析によって品詞に分解し、品詞の基本形や読み、活用などの情報を出力する。

キーワード群抽出のためには、キーワードのための品詞を選択する必要がある。キーワードは話題を特徴づける目的を持つため、動詞や形容詞などではなく、名詞が適切であると考えられる。また、名詞の中でも代名詞はキーワードとして適当ではないので使用しない。

また、このステップではキーワード抽出のみでなく、ステップ(2)での重み計算の為に出現頻度等も同時に算出する。

Step2: キーワードの重み付け

ステップ(1)で得た各キーワードに関してTF-IDF法^[7]を用いて重み付けを行う。

ここで、TF-IDF法を簡単に説明する。

● TF-IDF法

TF(Term Frequency)は、ある文書 d におけるキーワード k の出現頻度である。各文書から出現するキーワード数のばらつきを正規化するために、出現頻度をキーワードの総出現数で割った値である。これを $t(k, d)$ とする。繰り返し出現するキーワードの重要度を上げるためにTFを使用する。

IDF(Inverse Document Frequency)は、分析対象とするページの総数 n と、キーワード k が1回以上出現する文書の数 $d(k) (=1, 2, \dots, n)$ とによって、式(1)で定義される値である。

$$i(k) = \log(n/d(k)) \quad (1)$$

n が一定値であるため、 $d(k)$ が小さいほど式(1)の右辺は、大きな値をとる。これは、特定の文

書にしか出現しない語はキーワードとしての重要度が高い、という前提があるためである。

TF-IDF法では、上記の2つの値の積を用いる。すなわち、キーワード k の、文書 d における重みを $w(k, d)$ とすると、重みは、式(2)のように定義される。

$$w(k, d) = t(k, d) \cdot i(k) \quad (2)$$

Step3: 類似度計算

ページに対応したノード間の類似度を求めるためにベクトル空間モデルを使用する。すなわち、文書を多次元空間上のベクトルとして表現し、類似関数 SIM によって2つのベクトルの類似度を求める。類似関数 SIM は余弦関数 $\cos \theta$ で表すことができる。

類似度が0の場合、2つの文書間に共通するキーワードが全く現れなかったことを表しており、類似度が1の場合、2つの文書で全く同じキーワードが同じ頻度で出現した、ということを表す。

文書全体から出現するキーワードの集合の要素数を m とすると、ある文書 d_i の持つ重みベクトル $\vec{d}_i$ は次の式(3)で定義される。

$$\vec{d}_i = [w(k_1, d_i), w(k_2, d_i), \dots, w(k_m, d_i)] \quad (3)$$

このとき、文書 i と文書 j の類似度 e_{ij} を以下の式(4)で定義する。(θ は $\vec{d}_i$ と $\vec{d}_j$ のなす角。)

$$\begin{aligned} e_{ij} &= SIM(\vec{d}_i, \vec{d}_j) \\ &= \cos \theta \\ &= \frac{\vec{d}_i \cdot \vec{d}_j}{|\vec{d}_i| \times |\vec{d}_j|} \\ &= \frac{\sum_p w(k_p, d_i) \times w(k_p, d_j)}{\sqrt{\sum_p w(k_p, d_i)^2} \times \sqrt{\sum_p w(k_p, d_j)^2}} \end{aligned} \quad (4)$$

2.6 アクションクルーの検索

以上のプロセスで生成されたアクションクルーは、意味的にまとまりのある行動履歴と考え、それぞれに複数のキーワードを設定して保存する。

ユーザがこのアクションクルーを用いて情報を検索する手順を以下に示す。

Step1: まず, ユーザが必要な情報を得るためにキーワードを指定する。

Step2: 指定されたキーワードに対して行動履歴 DB を検索し適切なアクションクルーを得る。

Step3: ユーザは視覚的に提供されたアクションクルーの各ノードをたどることにより必要な情報を得る。

このように, 従来の検索結果の取捨選択作業を行うことなく行動履歴 DB を用いることで目的とする情報に到達することができる。

3 実験

3.1 実験方法

行動履歴 DB の利用による WWW 検索方法の有効性を調べるために, 本システムのシミュレーションプログラムを作成した。予め設定した検索目的に対する検索行動を行動履歴 DB に格納しておき, その DB と上記シミュレーションプログラムによりその情報への到達が容易になることを示す。シミュレーションプログラムの構成を図 4 に示す。

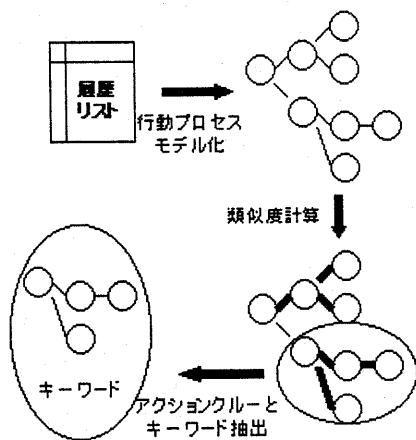


図 4: シミュレーションプログラムの概要図

ページのキーワード抽出を行うためにフリーソフトウェアである茶笥(ちゃせん)[8]を用いた。茶笥は, 様々な日本語の文章を形態素分析によって品詞に分解し, 出力オプションを指定することによって, 品詞の基本形や読み, 活用などの情報を選択し, 出力することができるプ

ログラムである。なお, 品詞のなかでキーワードとして利用する品詞は, 文書的话题を特徴づける物である表 2 に示すような一般名詞, 固有名詞を利用する。

表 2 キーワードとして利用する品詞

品 詞 番号	品詞の階層	例
2	名詞, 一般	北西部, 山
4	名詞, 固有名詞, 一般	広辞苑, ウィンドウズ
9	名詞, 固有名詞, 組織	荏原製作所, 環境庁, 朝日新聞
11	名詞, 固有名詞, 地域, 一般	ランカウイ島, 神戸市
12	名詞, 固有名詞, 地域, 国	アメリカ, 日本, カナダ

3.2 実験データ

実験では著者の一人が実際に「ランカウイ島でスキューバダイビングを体験するのに大体どれくらいの費用が必要か?」という情報検索目的に対して WWW 上を検索した行動履歴を使用した。表 3 にそのデータの一部を示す。

表 3: 実験に使用したデータ(一部省略)

番号	ページタイトル
1	Yahoo! JAPAN
2	Yahoo! JAPAN Search Results
3	Untitled Document1
4	Untitled Document2
5	マレーシア・旅行記 1//ma5
:	:
30	{beach 浜辺 浜 水辺 渚 海岸 shore}+{resort 別荘地}
31	Island of Legends LANGKAWI 伝説の島ランカウイ
32	ランカウイアクティビティ

上記の情報検索の目的に合致する情報に到達するために, 32 種類のページにアクセスし, 43 回の移動を行った。最終的に「ランカウイアクティビティ」というページで目的を達成し

た。その間、既存の WWW 検索エンジンに対して 5 回の検索要求を出した。

3.3 実験データの行動プロセスモデル化

表 3 のデータを、2.3 のアルゴリズムを用いて行動プロセスモデル化した。その結果を図 5 に示す。

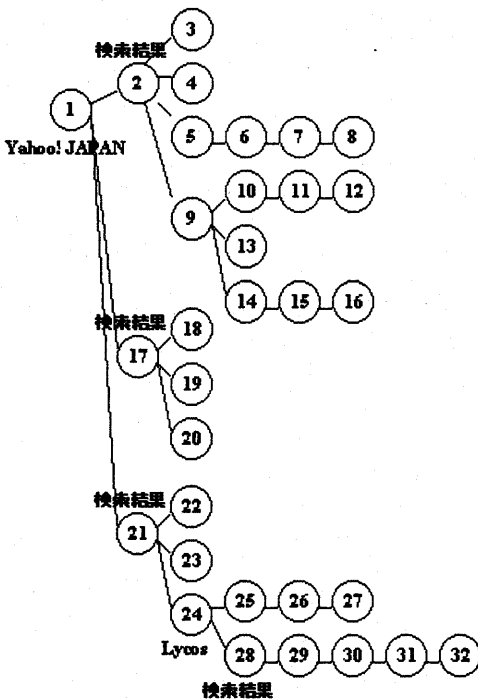


図 5 : 実験データの行動プロセスモデル化

3.4 ページ間の類似度計算

3.3 で述べた行動プロセスモデルに対応したページ間の類似度を 3.1 で示した方法により算出した。なお、実験では各ノード間の類似度が 0.07 以下の部分で行動プロセスモデルを分割し、アクションクルーを作成した。情報検索目的でアクセスした 32 ページの行動履歴は 4 つのアクションクルーに分割された。その結果を図 6 に示す。

3.5 評価

実験において作成したアクションクルーを用いて、他利用者による行動が効率化されることを示す。

まず、最終的に得たアクションクルーにおける 2 及び 21 の内容を表 4 に示す。

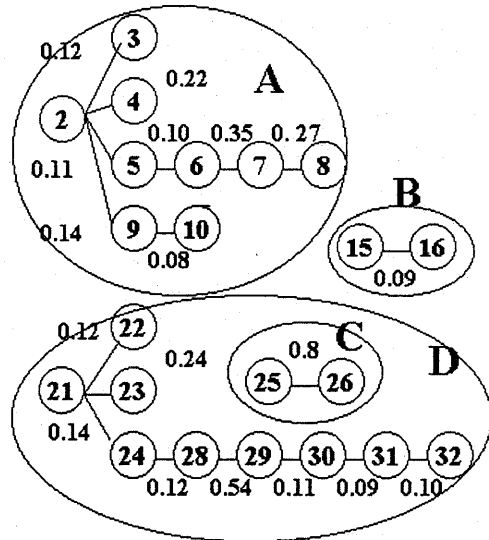


図 6 : 作成されたアクションクルー

表 4 : A の 2 番と D の 21 番のページ内容

番号	ページ内容
2	Yahoo! JAPAN で「ランカウイ、スキューバ」で検索した結果のページ
21	Yahoo! JAPAN で「ランカウイ、アクティビティ、料金」で検索した結果のページ

また、アクションクルー A、D に関してアクションクルー内に含まれるノードに設定されたキーワードを表 5 に示す。

表 5: アクションクルー A とアクションクルー D のキーワード

グループ	キーワード
A	島, 南, タイ, スキューバ, ランカウイ, ビーチ, 情報, 楽園, 情報
D	スポーツ, ツアー, ランカウイ, 料金, 島, トラベル, スキューバ, ビーチ

本システムでは、表 5 に示すアクションクルーがキーワード検索によって選択される。今回

の情報検索の目的である「ランカウイ島でスキューバダイビングを体験するのに大体どれくらいの費用が必要か?」を他利用者が求める場合には A あるいは D がまず出力される。仮に D を選択した場合には該当するページを見つけるまでのアクセスは、最大でアクションクルー D に含まれる総ページ数である 9 ページである。また、A を選択した場合には最大でアクションクルー A に含まれる総ページ数である 9 ページと、アクションクルー D に含まれる総ページ数である 9 ページとの合計 18 ページである。一方、従来の WWW 情報検索方法では、2 番、21 番等の他の検索結果ページや、9 番以下のリンクページに訪れなければならない。

以上の結果から、WWW 検索において、本システムを用いることにより、ユーザの検索結果の取捨選択作業が軽減され、効率の良い情報検索を行える可能性があると考ええる。

4 おわりに

本稿では、ユーザの行動履歴をモデル化し、WWW 検索に応用することで、ユーザの検索結果の取捨選択作業を軽減し、効率良く情報検索が行える行動 DB による WWW アクセス支援システムを提案した。

本システムと類似のシステムとして Blink! [9] などがある。これはよく使用される“お気に入り(ブックマーク)”の共有化、DB 化を行い、これによって検索の効率化を図ろうとするものである。本システムではユーザの行動した履歴を洗練化し抽象化し DB 化したものであり、より効率的に必要な情報を得ることができると考える。しかし、以下に示す課題をさらに解決する必要がある。

- 行動履歴のデータを格納するデータベースの設計
- 類似するアクションクルーの統廃合
- 行動プロセスモデルの分割に用いる類似度の規定値の検討
- 履歴内のゴミの処理
- ユーザへのアクションクルー提供の為にインターフェイス設計

これらの課題に取り組みつつ、本システムを実装する予定である。

参考文献

- [1] "goo",
<http://www.goo.ne.jp/>
- [2] 佐藤 光弘 他:「WWW における情報検索技術の動向」, 電子情報通信学会, Vol.82 No.12, pp1237-1242, Dec 1999
- [3] "Google",
<http://www.google.com/>
- [4] "jCentral",
<http://www.ibm.com/developer/java/>
- [5] 森 幹彦, 山田 誠二:「ブックマークエージェント:ブックマークの共有による情報探索の支援」, 電子情報通信学会論文誌, Vol.J83-D-1 No.5, pp.487-494, May 2000
- [6] Abrams, M. and Williams S.: "Complementing Surveying and Demographics with Automated Network Monitoring", World Wide Web Journal, Vol.1, No.3, 1996
- [7] 徳永 健伸:「言語と計算 5:情報検索と言語処理」, 東京大学出版会, 1999
- [8] 日本語形態素システム,
<http://chasen.aist-nara.ac.jp/index.html>
- [9] Blink.com
<http://www.blink.co.jp/>
- [10] 佐藤 進也 他:「サーチエンジンへの問い合わせの解析」, 情報処理学会, 研究報告 自然言語処理 No.136-018, pp135-141, Mar 2000
- [11] 大野 潮満 他:「参照重要度に基づく WWW 検索」, 情報処理学会, 研究報告 自然言語処理 No.135-1, pp1-8, Jan 2000
- [12] 三浦 信幸 他:「WWW サーバアクセス履歴からのユーザモデル構築」, 情報処理学会, 第 52 回 全国大会 Vol.37 No.11, pp1060-1061, Nov 1996
- [13] 風間 一洋 他:「WWW のユーザ操作履歴による HTML 文書の相関関係の解析」, 情報処理学会論文誌, Vol.40 No.5, pp2450-2458, May 1999
- [14] Barrett, R., et al.: "How to Personalize the Web", Conf. on Human Factors in Computer Systems, 1997