社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－171 (10)
2006／1／12

Web検索の認知モデルとその支援システム

鈴木 泉, 大里 有生（長岡技術科学大学）

インターネット検索において、人間が検索クエリーを作成する際の支援をするシステムを提案する。そのために先ず、検索行為の認知モデルを実験により明らかにした。本システムでは、人間は最初に1回だけ、クエリーに使われる可能性のある語をファジィ集合として与える。実際に検索で使用するクエリーは、人間が与えたファジィ集合に基づいて生成される。検索に成功するまでクエリーを変更してこれを繰り返すが、その際、人間が与えた指示には常に必ずしも従わないほうが、逆に良い結果をもたらすという、大局的学習法の考え方が適用できる。大局的学習法を適用することの有効性を、実際の検索を行って検証する。

A Cognitive Model of Web-Searching and its Support System

Izumi Suzuki, Ario Ohsato (Nagaoka University of Technology)

A human support system is introduced that assists human to create queries in Internet searching. For this purpose, a human recognition model of searching activity is described by performing an experiment. In the initial stage of the system, the human is requested to identify the entire word set that can be part of the queries, by means of a fussy set. Then, every searching query is generated by obeying the fuzzy set until the human successes the searching. However, the system can be even more useful if the system does not always obey the human instruction, which is outlined as the Global Learning Method. How the Global Learning Method is applied to this system, and why it can be useful are discussed by referring an experiment in which the proposing system is employed.

1. はじめに

現在、世界最大のインターネット検索エンジンが捕捉しているWebページの数は80億を超えている。この膨大な数の情報は、必要とする情報を自在に探し出すtoolさえあれば、非常に有用な情報源となり得るのである。通常、インターネット検索は検索語句を介して行われる。しかしながら探したい情報、対象物の名称、呼び名が分かる場合の検索は比較的容易であるが、そうでない場合、それらを検索することはあまり簡単ではない。例えば、街道沿いの山腹などによく設置してあるマイクロ波通信回線用の反射板を、誰が何の目的で設置しているのか解説・説明している文書を検索によって探すと仮定する。一般に呼ばれる「反射板」という名称が分かれば検索は容易であるが、そうでない場合、「白い板」「山腹」といった検索語を用い、試行錯誤による検索が行われることになる。とくに、インターネット検索に慣れていないユーザーの場合、有効な検索が出来ず、インターネット上の情報を十分に活用できない場合もあり得る。

本稿では、検索クエリーをユーザーに代って作成することで、ユーザーを支援する方法を具体的に提案する。そのために先ず、このような名称が分からない検索で、ユーザーが如何にして探したい情報に至るかを実験により確認し、その認知モデルを作成した。

2. 既存の技術および研究

2-1 インターネット情報検索

ユーザーが検索エンジンに入力し、検索の対象とする語句を *user query*（または単に**クエリー**）と呼ぶ。多くのインターネット検索エンジンは、基本的に *vector matching* という手法で検索が行われる。これは、検索語全体から成るベクトル空間を考え、個々の語の重要度の値を要素とするベクトルを検索語と文書について作成し、それらの類似度を求めるものである。検索語はは先頭に入れた語ほど重要度が高く設定される。一方、文書においては、語の出現頻度、出現位置、文字の大きさ等から重要度が設定される。クエリーの重要度を変化させることによって生ずる検索空間を探索し、ユーザーの探している情報に最も適した重要度を決めるという研究がある。[Yang (1992)] では、parallel searching である遺伝的アルゴリズム（GA）を有効に使い、個々の探索点（個体）をdocumentに対応させ、ユーザーの反応を評価に用いている。このように、ユーザーの反応を検索語に反映させる検索方法を feedback searching **フィードバック検索**と呼ぶ。このほか検索エンジンにおいては、リンク構造からページの重要度を計算する *Page-rank system* や、文書内では現れないが、その文書のキーワードとなり得る語を見つけ出す *anchor text* といった技術が用いられている。

インターネット情報検索の方式に関する研究は、インターネットが急速に普及し始めた90年代の初めに盛んに行われた。IR研究の初期段階では、ユーザーが入力した検索語にマッチする文書をいかに効率良く探すかに重点が置かれた。ところが、インターネットの普及に伴い、ill-defined targetの検索、つまり、ユーザーが自己の要求する情報がはっきり特定できない状況下での検索が注目されるようになる。そうした研究としては Oddy (1977), Belkin (1982) などが挙げられる。その後 Ingwersen (1987) は、Belkinのモデルを改良し、ユーザーの

情報検索課程を Pre-retrieval activities, Retrieval activities, and Post-retrieval activities の 3 つの段階に分けて捉えた。最近の研究では、検索の課程を problem analysis, strategy formulation, search implementation, および evaluation の 4 つの段階に分けて説明していることが多い (Sutcliffe, 1998; Saito, 1998)。

コンピューターのユーザーが作業中に、ユーザーが望むであろう情報を推測し、ユーザーに提供するシステムとしては *Adapted help systems* (または *intelligent help systems*, IHSs とも呼ばれる) が良く知られている。Adapted help systems は通常、active か passive であるか、および、static か adaptive であるかによって分類される(Brusilovsky, 1997)。Active であるが non-adaptive な help systems の例としては、"did you know" (DYK) help が挙げられる。これは、ユーザーの作業中に、作業に関連したヒントをランダムに呈示するもので、Microsoft Word など、多くのアプリケーションで取り入れられている。これとは逆に、Adaptive であるが passive な help systems は、ユーザーが help を利用している際に、ユーザーが望むであろう新たな help トピックを予測して提供する。この方式の一番の難点は、新たに提供する help トピックは、ユーザーにとって新しい知識で無ければならず、しかも、ユーザーが現在必要としている知識でなければならないことである。そこで、ユーザーの目的と、ユーザーの知識を定式化するモデルが提案されている。(Chin, 1989; Nessen, 1989; Winkels, 1990)。

2-2 大局的学習法

筆者らはかつて、active かつ adaptive にユーザーを支援するシステムとして、**大局的学習法** *Global Learning Method* (G.L.M.)を提案した (Suzuki, 1999)。これは、一般に *Active Learning* と呼ばれる、以下の手続きを繰り返すことによる、マン・マシンインターフェースの学習アルゴリズムとして分類される。

1. Renew the learning data: Add the pair (y , z) to the learning data, by receiving a response from the user, where the presented environment is .

2. Select an environment to present: Select an environment according to all the learning data from the beginning, and present it.

例えば、パソコンのマウスカーソルの移動速度の設定を考える。パソコンの使用経験の長いユーザーであれば、どのくらいの設定レベルにすれば本人にとって最も使い心地が良いか大体分かっているものだが、初心者の場合、使い心地の良し悪しにかかわらず、初期設定のままで使っていることが多い。ある初心者ユーザーの好ましい設定レベルを（何らかの方法で計測し）ファジィ集合として表わしたものが図 1 であるとする。

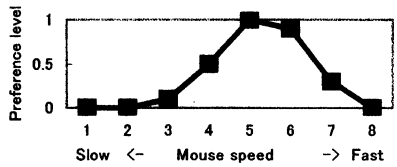


図 1 あるユーザーのマウスの移動速度の好ましさ

ユーザー本人が表明できるのは、「この設定は使い心地が良い（悪い）」あるいは、「現在の設定のほうが、前の設定より良い（悪い）」といった刺激に対する反応であり、図 1 のような尺度化された数値ではない。

いま、このユーザーが使用するパソコンにおける設定レベルは 4 にセットしてあるとする。設定レベルが 1 や 2、あるいは 8 にセットされてでもいない限り、この状態で本人から進んで「この設定は使い心地が悪い」ということを申し出るとは考えにくい。たとえ、この設定は使い心地が良いか訊かれたとしても、「特に悪いとは思わない」、つまり「良い」と答えることも十分予想される。その場合、もっと良い設定レベルであるレベル 5 や 6 があるにもかかわらず、レベル 4 で満足し、そのままパソコンを使い続けることになる。

ユーザーの作業を中断せずに、図 1 のような各レベルにおける好ましさの度合いを特定するためには、作業中に設定を自動的に変え、場合によっては設定を変えたことをユーザーに伝え、設定を元に戻すためのボタン（つまり、この設定は好ましくないことを回答するボタン）を提供する必要がある。大局的学習法は、環境をユーザーに提示し、それに対する回答をユーザーに求めることを繰り返す active learning のうち、以下の相反する 2 つの特徴を出来る限り満たすために提案した一方法である。

- ユーザーが好ましいと答える環境を出来る限り提示する
- ユーザーが好ましいと答える、答えないにかかわらず、様々な設定を幅広く試み、ユーザーの好ましい環境をファジィ集合として特定する。

マウスカーソルの移動速度の場合、設定値は 1 次元であり、しかも最も好ましい設定値の区間（以下ピークと呼ぶ）は 1 箇所であるので学習、つまり特定作業は容易である。しかし、多次元の場合や、ピークが 2 箇所以上ある場合などは、ユーザーに提示する環境を上手く選ぶ必要がある。前述の (Suzuki, 1999) では、ユーザーに提示する環境を如何に決め、ユーザーの好ましい環境を如何にしてファジィ集合として特定すればよいか、1 つの方法を提案している。

3. インターネット検索の認知モデル

3-1 実験

インターネット検索の支援システムを考えるに当たり、まず、人間がどのような手順でインター

ネット検索を行うかを知る必要がある。このような状況下での、インターネット検索の認知モデルを明らかにするために、以下のような実験を行った。

インターネット経験の豊富な10代の被験者3人を被験者になってもらい、2人1組で、1台のコンピュータを使い、与えられた課題をインターネットで検索する。検索の課程で発話された内容と、コンピュータの画面はビデオに記録する。実験後、被験者の会話と作業内容を書き起こし、分析を行う。このように、言葉で語られた内容から、発話者の認知課程を明らかにして行く方法を**プロトコル分析** *Protocol Analysis* と呼ぶ (Ericsson, 1984; 1993)。本実験は、(Saito, 1998) で報告された、ill-defined target の検索において、検索方針とユーザーの満足度の関係を調べる実験を参考にした。

書き起こされた会話の1例を以下の表1に示す。実験後の解析では、それぞれの会話が属す認知過程をカテゴリー分けして分析したが、ここでは省略する。

課題：山中を歩いていて、葉の真ん中に花が咲いている、変わった木を見かけた。それが紹介されているページを探して下さい。
（正解）花筏(ハナイカダ)という樹木

表1 実験で交わされた会話

会話	操作
A：（このキーワードでは）簡単すぎるんじゃない？	検索語「葉　真ん中　花」で検索 ヒット数 633,000 件
B：沖縄の植物ではないよね？ A：ここらでも咲いてるって言ってたじゃない	検索結果を見ている
A：山に咲いてるんでしょ B：（キーワードに）山を入れないと	
A：とりあえずこれで検索して B：山、花で出るわけないだろう	検索語を「山　花」に変更 ヒット数 57,800,000 件
A：山の花で抽象的に検索しといて、…	
B：花の図鑑みたいなのを探したほうがいいんじゃないの	検索語を「山の花」に変更（発言とは関係なく） ヒット数 478,000 件
	検索語を「珍しい　山の花」に変更 ヒット数 16,700 件
B：抽象的だよなあ A：とりあえず抽象的なものから探して、それからどんどん追い求めてゆけばいいじゃない	検索語を「謎　花」に変更 ヒット数 9,150,000 件
B：謎じゃないだろう	検索語を「謎　山　花」に変更 ヒット数 2,800,000 件
中略	
	検索語を「謎　葉の真ん中」に変更
	ヒット数 29 件
A：無いなあ A：（ページのタイトルに）桜って書いてあるじゃない	検索したページを開いてみている
A：これだ B：これじゃないの	検索した別のページを開いてみている

実験の結果、このように検索しようとする対象の名称が分からない場合、次のような行動をとることが分かった。

A1. 検索の初期段階では、思いついた検索語は必ずしも適切ではなく、（例えば「山　花」）また、その数も少ない。初回の検索で使用した検索語「葉　真ん中　花」は、課題の説明で使用した言葉をそのまま用いたもの。検索の最中に思い浮かんだ語や、検索結果から開いたページの中で使用されている語を追加する。

A2. 謎, 珍しい, 山, 山の花, 花, 葉の真ん中, 葉, 真ん中　といった検索語を試行錯誤的に組み合わせ、行き詰まると新しい検索語を追加し再び実行する。

A3. 明らかに検索結果が絞りこめないと判断されるクエリーは使用しない。

A4. 検索結果を見て、検索語が別の意味で用いられていることが多い場合は、その検索語の使用をなるべく避けている。また、求めている情報に近い検索結果が出ている場合はクエリーを少しだけ変更し、求めている情報から非常に外れている検索結果が出ている場合はクエリーを大きく変更している。（フィードバック検索）

A5. 検索結果のヒット数が非常に大きい場合は、様々な文書で多用される語句であることが多く、それらの検索語の使用をなるべく避けている。逆に、検索結果のヒット数が非常に小さい場合は、検索語がページ内の互いに離れた個所で使用されていることが多く、目的の情報を含むページを見つける可能性は低い。

A6. 数回前までの検索で使用されたクエリーは再び使用しないが、それ以前のクエリーは、一度検索したことを忘れており、再使用することがある。

一方、A1 から A5 は、モデル化しやすいように、それぞれ B1 から B5 のように解釈し直す。このように解釈する必然性は無いが、1つの方法としては妥当と考えられる。

B1. まず、検索に使用する語を出来るだけ多く（10〜20個）挙げ、これを「検索語リスト」と呼ぶ。ある語が検索に役立ちそうであれば、その語の類義語、関連語も検索語リストに含める。クエリーはこのリストの中から選び、途中で語を追加しない。

B2. 最初に、検索語リストからクエリーとして数語を選び、これで最初の検索をする。以後、語の追加、削除によって新たなクエリーを作成し、目的の情報を含むページを見つけるまで検索をを繰り返す。目的の情報を含むページかどうかは、明確に判断できるものとする。

B3 「検索語リスト」内の語は、あらかじめ２つのカテゴリーに分けておく。カテゴリーAには、それを外すと意味を絞り込むことが出来ず、検索にとって重要と思われる語を入れる。そうでないものはカテゴリーBに入れる。クエリーには、カテゴリーA の語ほど多く含むようにする。プロトコル解析実験の例では、「葉」や「真ん中」という言葉は、問題の事象を特徴付けるものであり、これらの言葉を外して、「謎」や「花」といった言葉だけで検索しても成功しないのは目に見えている。

B4. 検索の良否をフィードバックすることは、良否の判断が定式化できないため、今回は行わない。その代わりとして、およそ5回の検索につき1回はクエリーを大きく変え、それ以外では1語のみ入れ替える。入れ替えの際、除かれた語の類義語があればこれを選ぶ確率を高くする。

B5. 検索結果のヒット数が非常に大きい状態が続くときは、B4 でのクエリーを大きく変える際にクエリーに含まれる語数を増やし、ヒット数が非常に小さい状態が続くときはクエリーに含まれる語数を減らす。

人間がクエリーを与えるということは、考えうる全ての語の集合で、クエリーに含めるに語を部分集合として特定することである。一方、上記の解釈では、検索語リストの各要素を重要度に応じて２つのカテゴリーに分けることから、これはメンバーシップ値が高い、低い、または 0 のいずれかであるファジィ集合と捉えることが出来る。このことから、これ以降「検索語リスト」を**ファジィクエリー**と呼ぶことにする。つまり、ファジィクエリーの台 support が検索語リストということになる。実際に検索するクエリーは、人間が与えたファジィクエリーに準拠して生成される。

3-2 モデル化

実験で明らかにされた項目 B1 - B5, および A6 に従い、以下のように定式化する。（付録1参照）

Step 0. Initialization

検索に使用する語を出来るだけ多く（10～20個）挙げ、検索に役立ちそうであるか否かによって、２つのカテゴリーAまたはBに分ける（ファジィクエリーの作成）。カテゴリーA: 検索に役立つ見込み大、カテゴリーB: 検索に役立つ見込み小とする。また、互いに類義語の関係にある語の情報を付加する。以上は人手による作業である。

次に、ファジィクエリーから３語をランダムに選び、初期のクエリーを生成する。その際、カテゴリーAに属す語は選ばれる確率を高くする。

Step 1. Check the search result

生成されたクエリーで、検索エンジン Yahoo を使って検索する。And 検索のみ行い、検索結果として利用する情報は、上位 20 位までのページの要約表示、およびヒット数とする。要約表示されたページが探している情報に該当するか否かは、人間が判定する。目的の情報を含むページかどうかは、明確に判断できるものとする。

Step 2. Choose small change or large change

次に、次のいずれかを決める。1)クエリーの１語のみ他の語と入れ替える（small change と呼ぶ）、2)クエリーに他の語を追加または削除する（large change と呼ぶ）。その方法は、前回 large change をして以降の検索回数 T と、ヒット数の平均値 N によって、small change か、語の追加、または削除であるかが決められる（図２参照）。Small change は最大５回まで連続できる。

The Small change process

Step 2-M-1. Choose a word to be removed

クエリーから除かれる語を１つ選ぶ。カテゴリーAに属す語は、選ばれる確率を小さくし、カテゴリーBに属す語は、選ばれる確率を大きくする。

2-M-2. Choose a word to be added (Type I)

クエリーに追加される語を選ぶ。Type I では、クエリーから除かれた語と同一のカテゴリーから確率的に選ばれる。その際、クエリーから除かれた語の類義語は選ばれる確率を高くする。

The Large change process

Step 2-L-2-K. Choose a word to be added (Type II)

クエリーに追加される語を選ぶ。Type II ではファジィクエリー全体から１語を確率的に選ぶが、その際、initialization のプロセスで行ったのと同様に、カテゴリーA に属す語は選ばれる確率を高くする。当然、クエリーとなっている語は選ばない。

Step 2-L-2-I.

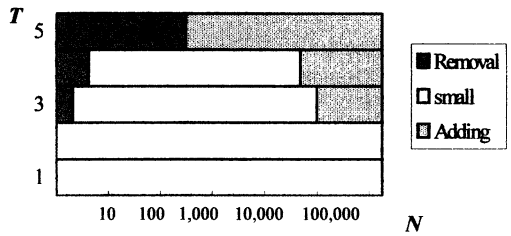
Step 2-M-1 と同様のプロセス。

Step 3. Query reasonability check

作成されたクエリーが、過去５回以内の検索で使用されたものでないかをチェックする。使用されたものである場合、１つ前のプロセスに戻る。規定回数（10回）の戻りを実行しても正当なクエリーが生成されないときは、さらに前のプロセスに戻る。

Stopping condition

探している情報に該当するページが見つかった時点で、または、規定回数（35回）の検索を実行した時点で検索は終了する。後者の場合、検索失敗とカウントされる。



T: times passed after the last large change
N: the average page number of the latest T search

図 2 Choosing small change or large change (add/removal)

4. 検証実験

提案したモデルは、人間が与えた「検索語のリスト」と各語の2段階の重要度（ファジィクエリー）から、ユーザに代ってクエリーを生成することによってユーザを支援するのであった。以下の実験、考察では、本システムを実際に使用して検索を行い、検索効率を上げるためには更にどのような改良が可能であるかを考察する。

実験で設定した問題：

以前に読んだ小説。大金を持ち逃げした犯人が、大金を持ちながら、電話をする小銭が手元に無かったことから、他人の小銭を取ろうとしてその場で取り押さえられ、犯行が明るみになる。この小説をインターネット検索によって探したい。
（正解）松本清張「百円硬貨」

問題文の下線をつけた語句を中心に、以下の検索語リストを作成した。（実験IとIIの2種類）

実験I

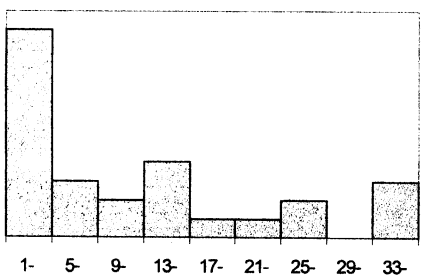
カテゴリA: 小説, 短編, 硬貨, 小銭
カテゴリB: 手元, 手元に無, 横領, 持ち逃げ, 犯行, 大金, その場で, 取り押さえられ, 電話, 電話を

実験II

カテゴリA: 小説, 短編,横領, 持ち逃げ
カテゴリB: 手元, 手元に無, 犯行, 大金, その場で, 取り押さえられ, 電話, 電話を, 硬貨, 小銭

実験IとIIの違いは、カテゴリA,B間で一部の語を入れ替えただけである。検索シミュレーションは確率の要素が入っているため、実験は30回程度実施した。最初に検索に成功するまでの失敗回数の分布を図3に示す。

実験I



実験II

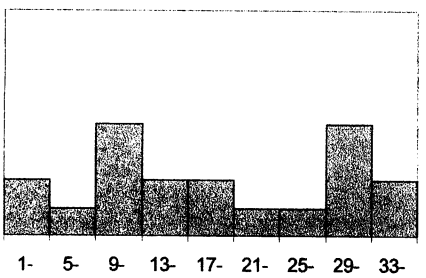


図3 最初に検索に成功するまでの失敗回数の分布
33以上には、検索失敗も含む

5. 考察

成功する確率がいずれもpであるような、独立なBernoulli試行を繰り返す場合に、初めて成功するまでの失敗の回数は幾何分布

$$G(p) = p(1-p)^x, x = 0, 1, 2, 3 \ldots$$

で与えられる

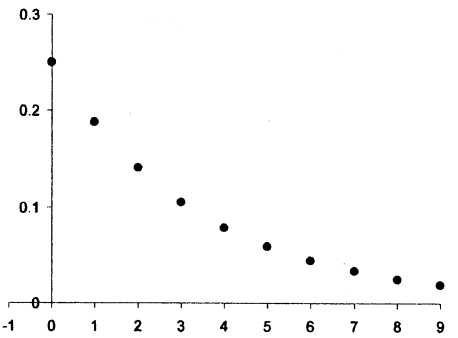


図4 幾何分布（$p = 0.25$）

実験で仮に、クエリーを検索語リストから毎回ランダムに選んだとすれば、やはりこのような分布になるはずである。実際には各試行は独立ではないので、このような分布には必ずしもならない。実験IIは、シミュレーション回数が19回と少なく、はっきりとした特徴はつかめないが、実験Iと比べ、かなり効率が悪いことは明らかである。クエリーを検索語リストから毎回ランダムに選ぶことによるシミュレーションは未だ実施していないが、おそらく、実験Iよりは効率が悪く、実験IIよりは効率が良くなるものと思われる。人間が与えるファジィクエリー次第で、このように検索効率に非常に差が出ることが分かった。

そこで、3～4回の検索で成功しなければ、人間が与えたファジィクエリーを少しだけ改変することを提案する。改変されたファジィクエリーに従ってクエリーを生成し3～4回の検索を実施する。そのごは再び元のファジィクエリーに戻す。これは、検索語リスト上で、好ましいファジィクエリーを特定するために大局的学習法を取り入れたことに相当する。人間が与えたファジィクエリーが最初から適切であれば、大局的学習法を入れることで逆に効率を悪くする可能性もあるが、最初に与えたファジィクエリーが適切ではない場合、実験IIのような結果に陥る危険を回避できる。

6. おわりに

インターネット検索における認知モデルを作成し、これに基づくユーザー支援システムを提案した。そして、筆者らがかつて提案した「大局的学習法」を本システムに適用できることを確認した。一見ユーザの意図に反すると思われる挙動が、逆に検索効率を上げることに繋がるのである。検索効率を高める方法は他にも考えられるが、当面は大局的学習法を取り入れた方法に焦点を絞るつもりである。また、本支援システムの活用方法とし

て、e-ラーニングへの応用なども視野に入れている。

参考文献

Jing-Jye Yang and Robert R. Korfhage: "Query Improvement In Information Retrieval Using Genetic Algorithms - A Report on the Experiments of the TREC Project", NIST Special Publication 500-207: The First Text Retrieval Conference (TREC-1), Gaithersburg, Maryland, pp. 31-58 (1992)

Oddy, R. N.: "Information retrieval through man-machine dialogue", Journal of documentation 33, pp. 1-44 (1977)

Belkin, N. J., Oddy, R. N. and Brooks, H. M.: "ASK for information retrieval: Part I. Background and theory", Journal of documentation, 38, pp. 61-71 (1982)

Ingwersen, P.: Toward a new research paradigm in information retrieval, Knowledge engineering, edited by I. Wormell, London, Taylor Graham (1987)

A.G. Sutcliffe and M. Ennis: Towards a cognitive theory of information retrieval, Interacting with Computers 10, pp. 321-351 (1998)

Mari Saito, Kazunori Ohmura, "A Cognitive Model for Searching for Ill-defined Targets on the Web - The Relationship between Search Strategies and User Satisfaction -", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, August 24-28 (1998)

Brusilovsky, P. and Schwarz, E.: "User as student: Towards an adaptive interface for advanced Web-based applications", Proceedings of 6th International Conference on User Modeling, Italy, pp. 177-188 (1997)

Chin, D. N.: Modeling what the user knows in UC., In User Models in Dialog Systems, Springer. pp.74-107 (1989)

Nessen, E.: User modeling in the SINIX consultant, Applied Artificial Intelligence 3, pp. 33-44 (1989) Winkels, R. G. F.: User Modeling in Help Systems. Berlin: Springer. pp. 184-193 (1990)

I. Suzuki and A. Ohsato: Fuzzy Identification of User's Requirements in Human Interface of Man-Machine System by Global Learning Method, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 7, No. 4 pp. 415-422 (1999)

Ericsson, K. A. and Simon, H. A. Protocol Analysis : Verbal reports as data. Cambridge. MA : MIT Press (1984; 1993)

付録 1 インターネット検索シミュレーションの流れ

Step 1 は人手による作業。ここで目的のページが見つかれば終了する。本文の項目 3-2 を参照。