Vol.2011-HCI-142 No.8
2011/3/17

閲覧行動モニタリングに基づく検索意図の抽出と検索結果の分類

南翔太郎[†] 岡誠[†]

近年，WWW の普及により様々な情報を Web 上で閲覧することができるようになったことで，情報検索システムが問題解決の手段として用いられるようになってきている．しかし，検索結果に意図に合わないノイズ文書が含まれてしまうことがありユーザの作業効率を下げる要因となっている．本研究では，検索タスクでの作業効率の向上を目的とし，検索結果のフィルタリングを行う．そのため，検索タスクでのユーザの意図を，閲覧行動をモニタリングすることで抽出する手法を提案した．抽出した意図に関する情報を用いて動的にフィルタを構築する検索結果フィルタリングシステムを用いて評価を行い，提案手法により，ある程度の意図を抽出できることを示した．

Extraction of Search Intention based on User Behavior and Classification of Search Result

Shoutarou Minami[†] and Makoto Oka[†]

Recently, We often use a search engine to solve the problem with the spread of the WWW. However, Search results often contains irrelevant document to user intentions. It is factor of decrease in efficiency of search task. In this paper, the goal is to improve the efficiency of search task by to filter the search results. To achieve the goal, we propose the method to extraction of search intention based on user behavior, and evaluate proposed method using a filtering system. The filtering system filters the search result based on the extracted user intension by proposed method. We show the possibility of this method.

1. はじめに

近年，WWW の普及により，さまざまな情報を Web 上で閲覧することができるようになったことで，Google[1]などのキーワードによる情報検索システムが問題解決の手段として用いられるようになってきている．現在の情報検索システムでは，ユーザが入力したキーワードを手掛かりとして，ユーザの意図に適合する文書をデータベースから探し出し，ユーザに検索結果として提供している．その際に，リンク構造の解析など様々な手法で検索結果をランキングすることにより，ユーザは必要とする情報を効率よく集めることができるようになってきている．

しかし，ユーザが問題解決に情報検索システムを利用しようとすると，知りたいと思っていることをキーワードとして具体的に言語化しなければならないが，対象をある程度知っていなければ具体的に言語化すること自体が困難であるといったジレンマがある．また，情報検索システムの側では，ユーザの様々な要求に対してあらかじめ準備を行うことが不可能であるため，意図に関係なくキーワードを検索対象から抽出している．そのため，キーワードに忠実な検索結果を返すことができるが，検索された文書が意図に合っているかはわからないといった問題がある．このような問題から，検索結果に意図に合わないノイズ文書が含まれてしまうことがある．複数の検索結果の文書を見て，情報の確かさや新しい情報がないかを確認しながら情報を集めていく際に，このようなノイズ文書はユーザの作業効率を下げる要因となっている．

そこで，検索結果の絞り込みやユーザの意図に合わせた情報提供を行うことの必要性から情報フィルタリングの研究が行われている．情報フィルタリングでは，ユーザがどのような情報を必要としているのか，あるいは，必要としていないのかという情報をブラック/ホワイトリストとして準備する手法や，ユーザの興味を Web ページの閲覧履歴から推定する手法などがある．しかし，ブラック/ホワイトリストを用いる手法ではリストを準備する手間がかかるため，検索タスクのように毎回の意図が異なるタスクとは合わない部分がある．また，ユーザの興味を閲覧履歴から推定する手法では多くのものは 1 日～数日の長期的な興味を対象としており，検索タスクのような短期的な要求を対象としている研究は少ない．

本研究は，ユーザが問題解決を目的に情報検索を利用して複数の検索結果を確認しながら情報を集めていく際の作業効率の向上を目的とし，検索結果のフィルタリングを行う．そのためには検索タスクごとの意図に合わせてフィルタリングを行う必要がある．そこで，ユーザの Web ページ閲覧時の行動をモニタリングして検索タスクでのユーザの意図を抽出する手法を提案し，検索タスクごとに動的にフィルタを構築する検索結果のフィルタリングシステムを実装し，評価を行う．

[†] 東京都市大学知識工学部
Tokyo City University Faculty of Knowledge Engineering

2. 関連研究

フィルタリングに用いるためのユーザの興味に関する情報を取得する手法としてはユーザの手間という観点から分けて，明示的な手法と暗黙的な手法がある[2].

明示的な手法とは, アンケートや閲覧した Web ページに興味があったかの評価を付けてもらうなど，ユーザに直接問う手法である．暗黙的な手法とは，ユーザの視線やマウス挙動を基に，閲覧していたページに興味があったのか，なかったのかを判定する手法や，閲覧履歴を用いてどのページを閲覧したのかという情報を利用して，興味を推定する手法である．明示的な手法は，興味に関する情報を確実に得られるがユーザの作業を増やしてしまうという問題があり，暗黙的な手法はユーザに負担をかけずに興味に関する情報を抽出する方法として研究が行われてきている．

土方らは情報フィルタリング技術の一つである適合性フィードバックにおけるユーザの興味抽出手法として，マウス挙動を用いて暗黙的にユーザの興味のあるキーワードの抽出を行っている[3]．適合性フィードバックとは，ユーザに検索結果の中から検索の意図に適合している Web ページを選択させ, 指定された Web ページからキーワードを抽出して再検索を行い，検索結果の精度を向上させていく手法である．土方らは，適合性フィードバックにおける問題としてページからのキーワード取得の際にノイズが含まれてしまうことと，ユーザに作業負担を負わせてしまうことを指摘している．そこで，興味のある Web ページ閲覧時の特徴的なマウス挙動として，なぞり読み，リンクポインティング，リンククリック，テキスト選択を利用してユーザの興味のあるキーワードを抽出する手法を提案し，2 つの問題の解決を行っている．

ユーザの短期的な興味に関する情報を興味のあるキーワードという形で取得しているため，検索タスクの意図に対応した，検索結果のフィルタリングに応用できる可能性がある．しかし，検索結果のフィルタリングを意図に合致している適合文書と，合致していない非適合文書に分類する問題と考えると，非適合に関する情報を取得できていないことになる．

本研究では，マウス挙動にくわえ，スクロール操作や閲覧時間といったユーザの閲覧行動を利用して閲覧していたページが意図に合致した適合文書であったのか，合致していない非適合文書であったのかを判定し適合/非適合の両方の情報を閲覧された Web ページごとに蓄積する．また，この判定処理を機械学習における学習データへのラベル付けに対応させて考え，キーワード抽出技術である TF-IDF 法を利用したフィルタの動的な構築を行い，検索結果のフィルタリングを行う．

3. 提案システム

3.1 システムのイメージ

前提として利用ユーザはある程度，情報検索を用いた問題解決に慣れているものとする．ユーザは問題解決を目的に情報検索を利用する際に次のような行動をとると考える．

1. まず，ユーザは 1 つ 2 つのキーワードを検索システムに入力する.
2. 出力された検索結果を確認し閲覧したい Web ページを探す.
3. 検索結果のリンクをクリックして Web ページを閲覧する.
4. ある程度閲覧し情報を吟味してから，情報の確かさの確認や，新規の情報を探すために，検索結果に戻り Web ページを探す.
5. 検索結果の別のリンクをクリックして Web ページを閲覧する.
6. 4.5.を繰り返す.

必要があれば, ユーザは新たな検索結果を得るためにキーワードの変更を行いながら，2.～6.の行動をとる．このような流れの中で，提案システムはユーザが Web ページを閲覧している際に, 閲覧行動をモニタリングし, 閲覧していたページが適合文書であったのか，非適合文書であったのかを判定し，得られた情報を用いてフィルタを学習する. また, ユーザが検索結果に戻った際に, 学習したフィルタを用いて検索結果をフィルタリングする.

3.2 システム構成

システムの流れ図を図 1 に示す．ユーザフローは 3.1 節で挙げたユーザの行動の 1.～6.に対応する．提案システムは，閲覧行動による判定処理と特徴抽出処理，フィルタリング処理に対応する 3 つのモジュールから構成される．閲覧行動による判定モジュールではモニタリングした閲覧行動データと閲覧ページの HTML を拡張機能から受け取り，分類器を用いて判定を行い，ラベルを付けて閲覧ページの HTML を特徴抽出モジュールへ渡す．特徴抽出モジュールではフィルタに用いる特徴を抽出してフィルタを学習させる知識をデータベースへと蓄積していく．フィルタリングモジュールでは, 検索結果を拡張機能から受け取って検索結果のリンク先 Web ページを先読みする．また，蓄積された知識をもとにフィルタを構築し，先読みされた Web ページのデータをフィルタにかけ非適合文書のフィルタリングを行う.

Web ブラウザは Chrome を利用する．また，検索エンジンには Google を利用している．閲覧行動のモニタリングや検索結果の書き換えなどは Chrome の拡張，意図との適合/非適合の閲覧行動による判定処理や特徴抽出処理，フィルタリング処理は CGI（Common Gateway Interface），RPC（Remote Procedure Call）を用いて実装をおこなっている.

3.3 節で閲覧行動による判定モジュール, 3.4 節で特徴抽出モジュール, 3.5 節でフィルタリングモジュールでの処理の詳細について説明を行う.

Vol.2011-HCI-142 No.8
2011/3/17

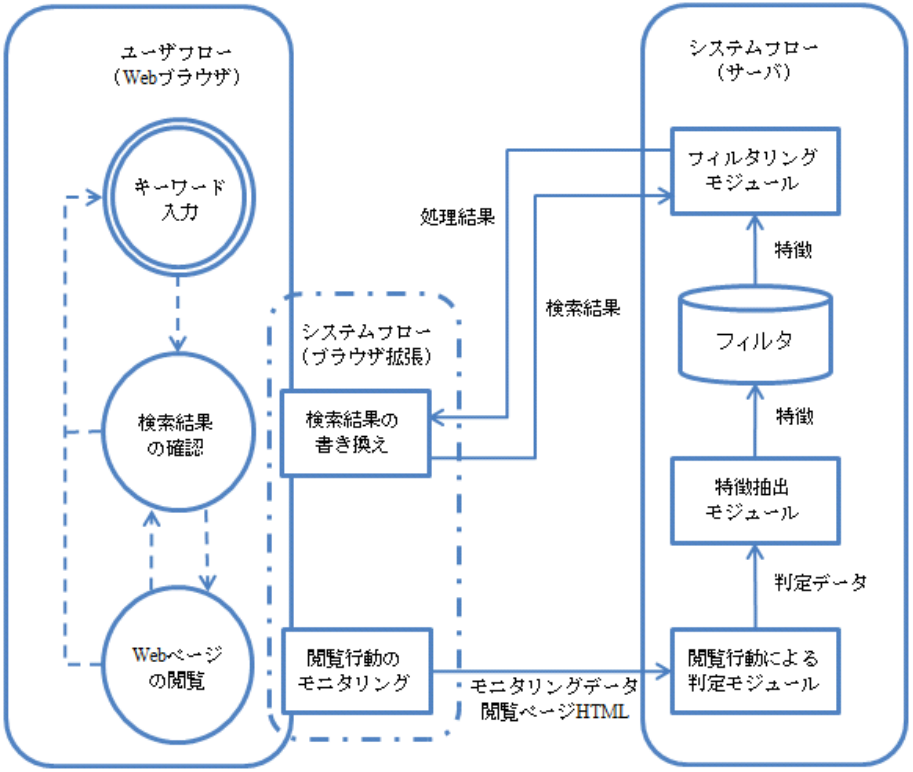

図 1 システムの構成

3.3 閲覧行動による判定モジュール

モニタリングする閲覧行動のパラメータを，予備実験を行って検証した．予備実験では，閲覧行動のモニタリング部分のみを実装したシステムを用いてユーザに検索を行ってもらい，閲覧行動のパラメータデータの収集を行った．Web ページを閲覧したときに検索の意図にあっていれば，お気に入りに登録してもらった．お気に入りに登録されたものは適合ラベル，されなかったものは非適合ラベルとして扱う．工学部研究室の学生 4 名に参加してもらい，期間は 1 日とし，316 件の事例を収集した．χ^2 値によるパラメータ選択を行った結果，次の 7 つのパラメータが選択された．

- Web ページ表示時間
  ブラウザで Web ページが見える状態で表示されていた時間
- Web ページ滞在時間
  Web ページにアクセスしてから他ページに移動，またはウィンドウから消去されるまでの時間
- 操作停止時間
  ユーザのすべての操作が行われていない時間
- マウス操作時間
  クリックや移動など，マウスを操作している時間
- なぞり読みしていた時間
  マウスを画面上のマーカとして文字をなぞり，読んでいる時間
- スクロールバー操作時間
  ブラウザのスクロールバー，または，マウスホイールを操作している時間
- スクロール距離
  Web ページをスクロールした距離

Web ページ表示時間，Web ページ滞在時間，操作停止時間はユーザが Web ページを確認していく際に，そのページが見るに値するかどうか，信頼できる情報，有用な情報がありそうかといった判断にかかるまでの時間に対応すると考えられる．また，マウス操作時間，なぞり読みしていた時間，スクロールバー操作時間，スクロール距離は GUI 操作が基本となる Web ブラウザにおいて無意識に取ってしまう行動や必須の行動に対応するパラメータとなっており，適合ラベル/非適合ラベルにおいて行動に差が出ていたと考えられる．

提案システムでは分類器を用いて適合/非適合の判定処理を行う．これは，ある程度の不確実性を考慮した柔軟なシステムを構築するためである．選択された 7 つのパラメータを入力とし適合ラベル/非適合ラベルを出力とする分類器を，予備実験で得られたデータを用いて学習し分類の評価を行った．提案システムをリアルタイムに動作させることを考慮して，分類器には計算量が少なく，実行速度が速いナイーブベイズ分類器を使用している．

表 1 予備実験　評価結果

	再現率	適合率	F 値
非適合文書	0.913	0.989	0.949
適合文書	0.833	0.366	0.508

表 1 に 10 分割交差確認法で評価した結果を示す．評価には Weka[4]を用いており，ナイーブベイズ分類器の学習を行う際に，各パラメータには MDL 基準に基づく 2 値化処理を行っている．

表 1 の非適合文書の F 値が高いことから，非適合文書の判定を良い精度で行えている．非適合文書が適合文書として判定された事例の中には，適合の正解ラベルがついているものと比較して，内容にあまり違いがないものもあり，ユーザが判断に迷っていたのではないかと考えられる事例もみられた．適合文書の適合率の低さの要因となっているものと考えられる．

予備実験の検証結果より，閲覧行動による判定モジュールではモニタリングした閲覧行動の 7 種類のパラメータデータを入力とし，適合ラベル/非適合ラベルを出力するナイーブベイズ分類器を用いて判定処理を行い，特徴抽出モジュールへ判定ラベル付きの閲覧ページの HTML を渡す．

3.4 特徴抽出モジュール

閲覧行動による判定モジュールから判定ラベル付きの閲覧ページの HTML を受け取り，フィルタを学習するために用いる特徴を抽出する．

HTML は，内容の文章に文書構造を示すタグを付けて記述されており，視覚的な情報と文章内容の情報を持っている．そこで，特徴として HTML のタグ構造に関する特徴（以下 HTML 特徴）と，文書に関する特徴（以下文書特徴）を抽出している．HTML 特徴は HTML のタグの木構造の 2 連接の部分木を特徴として利用している．例えば，

```
<html>
    <body>
        <h1>見出し</h1>
        <a>テキスト</a>
        <p><em>テキスト</em></p>
    </body>
</html>
```

という HTML からは body+h1+text や body+a+text，body+p+em，p+em+text といった特徴を抽出する．また，文書特徴は HTML からタグを除去したプレーンテキストに ChaSen[5]を用いて形態素解析処理を行い，名詞と未知語を抽出する．抽出された特徴は文書中の出現頻度を数え上げてラベルごとにデータベースに蓄積する．

3.5 フィルタリングモジュール

フィルタの学習と検索結果のフィルタリング処理を行う．フィルタリングは特徴の重み付きのパターンマッチにより行う．そのため，フィルタの学習は特徴に対して，ラベルごとの重みの計算を行う．また，マッチングを行うために検索結果の各ページをダウンロードして HTML 特徴，文書特徴を抽出する処理を行う．HTML 特徴と文書特徴は 3.4 節で挙げたものと同一のものである．

特徴に対する重みは HTML 特徴，文書特徴それぞれにおいて TF-IDF 法を用いており，片方のラベルに偏って出現する特徴に対して大きな重みを与えている．TF-IDF 法とは文書集合内のある文書におけるキーワードを抽出する技術である．TF（単語頻度）と IDF（逆文書頻度）を用いて重みを計算するアルゴリズムであり，多くの派生アルゴリズムが提案されている．本研究での重みの計算には（1）の式を用いている．

$$weight_{feature|label} = \frac{tf_{feature\,|\,label}}{tf_{label}} \left\{ log\left(\frac{N}{df_{feature}}\right) + 1 \right\} \qquad (1)$$

ただし，$tf_{feature|label}$はラベル$label$での特徴$feature$の出現頻度，tf_{label}はラベル$label$での全特徴の出現頻度の総和，$df_{feature}$は全データ中での特徴$feature$の出現頻度の総和，Nは全データ中での全特徴の出現頻度の総和である．

フィルタリングは評価対象の Web ページから抽出された特徴が，ラベルごとのデータベース内にあるかどうかを調べ，あればラベルのスコアに特徴の重みを加算していき，最終的により大きなスコアとなったラベルに判定を行う．このとき，評価対象の Web ページ内での特徴の出現頻度は考慮せず，出現の有無のみを判断に用いている．

4. 評価実験

提案した意図抽出手法を実装した，検索結果フィルタリングシステムによりユーザの検索作業の効率の向上がみられるか評価を行う．

4.1 実験方法

ユーザに穴埋め課題を提示し，検索システムを用いて情報収集を行いながら問題を解いてもらった．その際に制限として，キーワードは 2 単語までに限定し，再検索は許可しなかった．キーワードの制限はユーザ間でキーワード数による違いをなくすことを目的としている．また，再検索を許可しなかったのは，実験終了後に検索結果の全ページをユーザに確認してもらい評価をつけてもらうため，上限（20 件）を設けたためである．

制限時間は 1 問 15 分とし 2 セット行い，制限時間内に課題が終わったらそこで終了とした．課題は情報検索を用いて収集したいくつかの Web ページを参考に主観的に作成をしている．以下に問題を記す．

Vol.2011-HCI-142 No.8
2011/3/17

以下はソフトウェアエージェントについての記述です。( )を埋めてください。
ソフトウェアエージェントとはユーザの仮想的な代理（＝　　）として情報収集やソフトウェア間の連携を行い動作するプログラムを説明する（　　）である。ユーザの処理を（　　）することが目的であり、最終的には知能を持つことが要求される。プログラムどうしが通信を行い協調的にふるまう（　　を持つ）ことや自律的に動作する（　　をもつ）ことなどが特徴である。派生概念として（　　）の分野で研究されている知的エージェントなどがある。

以下はセマンティックウェブについての記述です。(　)を埋めてください。
セマンティックウェブとは（　　）なメタデータ(情報に関する情報)をＷｅｂページに付与することにより（　　）がWebページの内容を理解して処理できるようにするための技術である。現在、Webページの記述に用いられているHTMLは（　　）を表現するためのものであり、Webページの意味的な内容は表現しないためコンピュータがその内容を処理することはできない。（　　）を付与することでコンピュータがWebページの内容を処理できるようになり、言語表現の曖昧さを意味的に解釈し解決する（　　）システムや、（　　）を一つのデータベースのように扱うことなどが可能になる。

4.2 評価方法

実験終了後に検索結果として表示されていた 20 件すべてを，一つずつ確認してもらいながら次の 4 段階で評価をつけてもらった．

1. 意図に合致していない
2. あまり意図に合致していない
3. だいたい意図に合致している
4. 意図に合致している

これをユーザ評価として扱う．また，1.2.は非適合文書，3.4.は適合文書として扱う．Google の検索結果と，提案システムの出力結果とフィルタリングしたものそれぞれがどのくらい意図に合っていたのか（フィルタリングしたものについては意図に合っていなかったのか）を検証する．それぞれの出力にユーザ評価の 3.4.が多いほど意図に合っていたといえる．

4.3 実験結果

実験は，工学部研究室の学生 10 名に協力してもらい行った．実験結果を図 2 と図 3 に示す．図 2 は値が 4 に近いほど意図に合っていたといえ，1 に近いほど合っていなかったといえる．Google のユーザ評価平均に関しては，閲覧を繰り返していっても初期の出力から変化しないため値は一定である．フィルタの更新を重ねていくと，提案システムが表示しているものに関してはユーザ評価平均が増加し，非表示にしているものに関してはユーザ評価平均が下がってきていることから，フィルタの更新を重ねることでシステム自体の性能が向上していく傾向があるといえる．また，フィルタ更新回数 4 回の時点から Google のユーザ評価平均を提案システムの表示しているもののユーザ評価平均が少しずつ上回り始めており，速い段階からシステムが有効に機能し始めているといえる．

図 3 はユーザ評価 1.と 2.を非適合文書とした時に，検索結果として表示されていた非適合文書数である．また，提案システムの出力に関しては，線形近似を行った結果も記している．傾きは-0.092 である．Google の非適合文書数に関しては，図 2 と同様に，閲覧を繰り返していっても初期の出力から変化しないため値は一定である．更新初期から安定して，非適合文書が除外されていることが示されている．

図 2，図 3 より適合文書を表示に残しつつも、数件の非適合文書を表示から除外できているといえる．

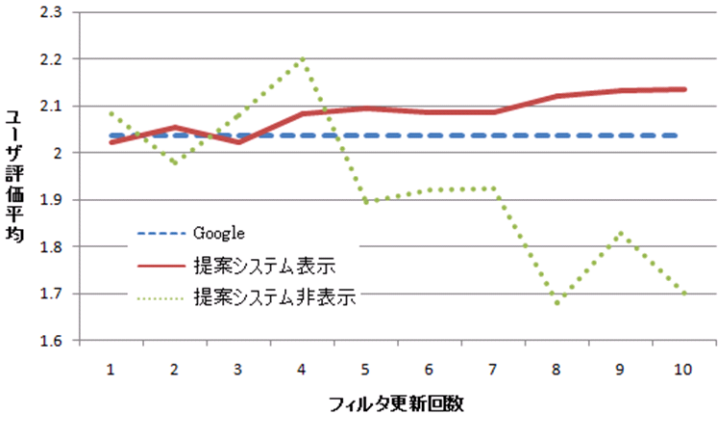

図 2　ユーザ評価平均のフィルタ更新による推移

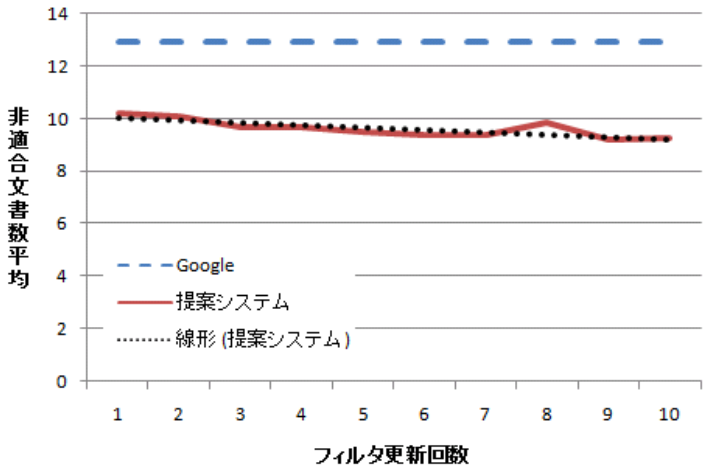

図 3　表示に含まれる非適合ページ数のフィルタ更新による推移

5. 考察

図 3 より，ユーザが直接目にする表示（Google の検索結果や提案システムの出力）に関して非適合文書の数が減少していることから，ユーザの作業効率の向上に貢献できる可能性を示せたと考えられる．検索結果にリンクが表示されており閲覧された Web ページに関して，ユーザ評価と閲覧行動による判定モジュールの評価の比較を行った結果，複数のユーザが適合の評価をしていながら，判定モジュールで非適合の判定が出ている事例が存在していた．以下事例を挙げていく．

- ポータルサイトのトップページ
  ポータルサイトのトップページは通過点としての役割が強い．サイト単位でみると意図に合っている可能性があり，ユーザ評価もその点を反映しているものと考えられる．

- 内容がほとんど同じ Web ページ
  Web 上には内容がほとんど同じページというものが存在しており，そのような Web ページは新規の情報を含まないため，ほとんど閲覧しないで検索結果に戻るという行動がとられると考えられる．

- 記述が少ないページ
  意図に合った情報を含んではいるが，ピンポイントな情報であり記述が少ないようなページ．

それぞれの Web ページとしての特性により Web ブラウザとユーザのインタラクションが少なくなってしまうことによって判定にズレが出ていると考えられる．また，適合文書を優先してフィルタしてしまっているケースが存在しており，判定モジュールのズレが影響を与えている可能性もある．Web ページの内容を考慮してユーザの閲覧行動をとらえていく必要があると考えられる．また，検索結果にリンクされた Web ページを対象にフィルタリングを行うか Web サイトを対象にフィルタリングを行うかといったことも考慮する必要があると考えられる．

6. おわりに

本研究では，ユーザが閲覧したページが意図に合っていたのか，または，あっていなかったのかを閲覧行動をモニタリングすることで判定する手法を提案した．また，提案手法により意図に関する情報を取得し，動的にフィルタを構築する検索結果フィルタリングシステムを実装し，評価を行った．

閲覧行動は意図との適合/非適合により違いがあるパラメータがあり，それらのパラメータを用いることで，ユーザの判断をある程度，推定することが可能であることを示せた．また，動的なフィルタ構築により短期的なユーザの検索意図を反映したフィルタリングを行い，ユーザの作業効率の向上に貢献できる可能性を示せた．

今後は，ユーザ評価と閲覧行動による判定モジュールの判定にズレがあった事例に対して，特徴抽出時に Web ページの類似度を考慮して，ラベルの修正をかけるなどの対応をとることが考えられる．また，フィルタリング対象を Web ページとするか，Web サイトとするかについても検証を行っていく必要がある．

参考文献

1) Google: http://www.google.co.jp/
2) 土方嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術，人工知能学会誌，19 巻，3 号，pp.365-372(2004).
3) 土方嘉徳，青木義則，古井陽之助，中島周: マウス挙動に基づくテキスト部分抽出方式と抽出キーワードの有効性に関する検証, 情報処理学会論文誌, Vol43, No.2，pp.566-576(2002)
4) Weka: http://www.cs.waikato.ac.nz/ml/weka/
5) ChaSen: http://chasen-legacy.sourceforge.jp/