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

土方嘉徳[†,☆] 青木義則[†]
古井陽之助[†,☆☆] 中島周[†,☆☆☆]

情報検索におけるユーザ分析では，ユーザが閲覧したコンテンツのどの部分に興味を持ったのかを取得することが重要となる．既存の手法でこのようなユーザの興味に関する情報を取得しようとすると，ユーザにアンケートに答えてもらうという手間をかける問題や，Webページ中の一部分というような細かい単位では自動取得できないという問題があった．本稿では，ユーザのWebページ閲覧中のマウス操作を利用して，ユーザが興味を持ったと思われるテキスト部分を全体のテキストから自動抽出する手法を提案する．本研究では，まず事前調査としてユーザのWebページ閲覧中のマウス操作の観察とインタビューを行い，どのような種類の操作がユーザの興味と関連があるのかを明らかにする．次に，これらの操作の対象となるテキスト部分が実際にユーザが興味を持った部分であるのか否かを,「TextExtractor」と呼ぶ実験システムを実装し，被験者実験を行うことで検証する．実験の結果，テキスト中におけるユーザが興味を持ったキーワードの割合は，文書全体よりも，これら各々の種類の操作が対象とするテキスト部分の方が高いことが検証された．また，これらの操作すべてを使ってテキスト部分を抽出した場合，ユーザが興味を持ったキーワードを抽出する精度は，ランダムにキーワードを抽出する方法に比べて約4倍，tf・idfに比べて約1.4倍高いことが確認できた．

Text Part Extraction Based on Mouse Operations and Evaluation of Extracted Keywords

Yoshinori Hijikata,[†,☆] Yoshinori Aoki,[†] Younosuke Furui[†,☆☆] and Amane Nakajima[†,☆☆☆]

In the area of information retrieval, it becomes important to acquire which portion of the content the user was interested in. The existing techniques for acquiring this information have the problem which forces the user to answer questionnaires or the problem which cannot carry out automatic acquisition in a fine unit like the portion in a Web page. This paper proposes a method for extracting the text parts which the user might be interested in from the whole text of the Web page based on the user's mouse operation. First, we conduct observations and interviews to discover what kind of operation is related to the user's interest. Second, we build a system called "TextExtractor" and conduct an experiment to see the effectiveness of the discovered operations. The result showed that the ratio of the keywords which the user was interested in was higher in the targeted text parts of any kind of the discovered operations than that in the whole document. When we extracted texts using all kinds of discovered operations, the precision to extract keywords of TextExtractor was about 4 times compared with that of random extraction and about 1.4 times compared with that of tf-idf.

† 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research, Tokyo Research Laboratory
☆ 現在，大阪大学大学院基礎工学研究科
Presently with Graduate School of Engineering Science, Osaka University
☆☆ 現在，九州大学大学院システム情報科学府
Presently with Graduate School of Information Science and Electrical Engineering, Kyushu University
☆☆☆ 現在，日本アイ・ビー・エム株式会社ビジネス・イノベーション・サービス
Presently with Business Innovation Services, IBM Japan, Ltd.

1. はじめに

Web上での情報獲得を支援するサービスとして検索エンジンがある．検索エンジンとは，ユーザが興味のあるキーワードを検索キーとして入力することにより，そのキーワードを含むページを推薦してくれるサービスである．しかし，検索エンジンで索引可能なページ数は数億の単位に達し[1)]，検索結果を絞り込むことが重要視されている．絞り込みの技術は情報検索の分野で多くの試みがなされているが，Web上では様々

なユーザが検索エンジンを使用するため，ユーザが検索に関する特別な知識を持っていなくても使えるような工夫が必要である．そのような絞り込みの技術として適合性フィードバック（Relevance Feedback）[2]がある．

適合性フィードバックは，検索結果として列挙されたWebページの中から，ユーザに自分の興味に適合するページを選択させて，選択されたページの中のキーワードを用いて再度検索を行う手法である．この手法では，キーワードの選択を適合するページ全体のテキストから行っている．そのため，なかにはユーザの興味に関係しないものも含まれてしまい，それらのキーワードが検索の精度を低下させるという問題がある[5]．また，ユーザに適合するページを選択させるため，閲覧操作以外の手間をユーザにかけさせるという問題もある．

本研究では，前者の問題に対する解決策として，ページ全体のテキストを利用するのでなく，ユーザが興味を持ったと推定されるテキスト部分だけを利用することを提案する．また，後者の問題に対する解決策として，ユーザに興味の有無を明示的に指摘させるのではなく，ユーザの閲覧動作からユーザが興味を持ったのか否かを推定することを提案する．具体的には，入力デバイスとしてマウスを用いてWebページを閲覧する場合を対象として，以下の手法により上記問題を解決する．

(1) Webページ閲覧時のマウス操作から，ユーザの興味と関連して発生した可能性のある操作を抽出する．
(2) 抽出した操作の対象となるテキスト部分を文や行の単位で抽出する．

このようにテキスト部分を抽出することで，以下の効果が期待できる．

(1) ユーザに自分の興味に適合するページを選択させるような手間をかけさせることがなくなる．
(2) 適合性フィードバックに使用するテキストに含まれるノイズ（ユーザの興味と関係のないキーワード）の割合を低減することができる．

本稿では，2章でユーザの興味に関する情報を取得する関連研究について触れた後，3章でユーザの興味とマウス操作の間に関連があるか否かを調べる事前調査について述べる．4章では，ユーザの興味と関連があると見られる特定の種類の操作について，その対象となるテキスト部分の有効性を調べる実験の説明を行う．また，この実験で用いたテキスト部分を抽出するシステム「TextExtractor」の説明を行う．5章で実験結果と抽出したテキストの評価について述べ，6章で実験を通じて得られた知見についての議論を行う．最後に，7章でまとめを述べる．

2. 関連研究

適合性フィードバックでは，ユーザに自分の興味に適合するページを選択させるという方法で，ユーザの興味に関する情報を取得している[2]．ユーザの興味に関する情報を取得するという観点では，情報フィルタリングにおいても多くの研究がなされている[3]．情報検索との違いは，情報検索はユーザ自らが外部から情報を取得する際に支援するのに対し，情報フィルタリングは流入する情報を取捨選択する点にある[4]．情報検索や情報フィルタリングにおいて，ユーザの興味に関する情報を取得する既存の手法には，大きく分けて次の2種類がある[5),6]．

(1) 明示的（直接的）手法：

ユーザに，(i) ユーザの興味に関してトピックやキーワードの形でアンケートに答えさせる，または (ii) 閲覧したページにどれだけ興味があったかを数段階で評価をつけさせることにより，ユーザの興味に関する情報を取得する手法である．(i) はRingo[7]やSIFT[8]，(ii) はGroupLens[9]，Syskill & Weber[10]，NewsWeeder[11]，ClixSmart[12]，AntWorld[13]といったシステムで用いられている．取得したユーザの興味に関する情報は，ユーザが直接答えたものであるため信頼性が高いという利点がある．しかし，(i) および (ii) では，アンケートに答えさせたり，閲覧後に評価を付けさせたりするという手間を，ユーザ側にかけるという問題がある．また (ii) では，ユーザが評価したページ全体のテキストからキーワードを選択しているため，選択したキーワードの中にはユーザの興味と関係ないものも含まれるという問題がある．

(2) 暗黙的（間接的）手法：

ユーザが閲覧した情報にどれだけ興味を持ったかを，(i) ユーザが閲覧に費やした時間（閲覧時間）[14]や，(ii) 閲覧中における特定のボタン操作やスクロール操作[15]，(iii) 閲覧中の視線[16),17]，といった閲覧の挙動から推定する手法である．ユーザに精神的な負荷をかけることがない点が長所である．(i) は，Webサーバ側でも取得可能であるが，ユーザがあるページを表示した後に，席を外したり他の仕事を始めたりしても，それを検出することができないという問題がある．(ii) では，記事に対して，ユーザが拡大表示するボタンを押したか否かや，スクロールをしたか否かをチェックしている．しかし，これらの操作をチェックすることでは，

ページ全体に興味を示したか否かはある程度判断できるが，ページのどの部分に興味を示したかということまでは判断できない．(iii) では，ユーザが興味を持った部分を特定することはある程度可能であるが，視線を検出するためには特殊な装置が必要となり，実用面での問題が残る．

また，情報検索や情報フィルタリングの分野では，ユーザが興味ありと評価した文書に含まれるキーワードから，検索に有効となるキーワードを選択するための手法として，tf・idf[2)]が提案されている．この手法では，文書内における個々のキーワードの出現頻度や他の文書におけるそれらのキーワードの出現する確率を基に，キーワードに重み付けを行っている．しかし，tf・idf は，ユーザが文書の一部にしか興味を示さなかった場合でも，キーワードの重み付けを文書全体の統計情報から行っている．そのため，キーワードの重み付けにユーザの興味が正しく反映されないことがある．

本研究の手法は，ユーザの興味に関する情報を，マウス操作という閲覧の挙動から推定しているため，暗黙的手法に分類される．また本研究は，(i) 一般の Web ページ閲覧中で発生し，ユーザが無意識に行うようなものまで含めたマウス操作から推定する点，(ii) ユーザが興味を示したと思われる部分をページ単位ではなく文や行単位で抽出する点，(iii) 特殊な装置を必要としない点に，既存の研究との違いがある．

3. 事前調査

Kantor らの研究[13)]では，ユーザは Web ページを閲覧するときにマウスを目で追う傾向があることを発見したと報告している．またその理由として，Web というアプリケーションでは興味のあるリンクをマウスでクリックする必要があることをあげている．しかし，Web ページ閲覧中には，具体的にどのような操作が発生するのか，またそれらの操作とユーザの興味には関連があるのかという点は明らかにされていない．

そこで我々は，ユーザの興味と関連して発生する可能性のある特徴的な操作を調査するために，31 人のユーザに参加してもらい，Web ページ閲覧の観察実験とインタビューを行った．観察実験では，ユーザに好きな Web ページを自由に閲覧してもらい，その様子を観察した．インタビューでは，Web ページ閲覧中にはどのような操作を行うことがあるか，またそれらの操作を行う理由を訊ねた．その結果，Web ページ閲覧やブラウザの機能を利用するのに必ず必要な操作だけでなく，本来 Web ページ閲覧やブラウザの機能を利用するのには必要ないが，半ば無意識に行うような操作まで，次の 10 種類の特徴的な操作があることが分かった．なお，この 10 種類の操作には，検索エンジンの検索キーワード入力フィールドに，興味のあるキーワードを入力するような直接的に興味対象を指定するような操作は含んでいない．

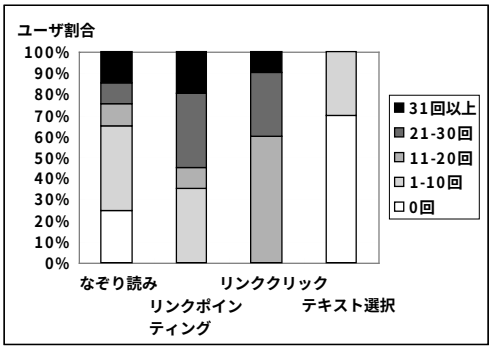

図 1 操作回数別ユーザ割合
Fig. 1 User ratio according to the number of operations.

(1) なぞり読み：マウスポインタをテキストの行に沿って左から右に動かす．
(2) リンクポインティング：マウスポインタをリンクの上に持っていくが，クリックはしない．
(3) リンククリック：リンクをマウスでクリックする．
(4) テキスト選択：マウスをドラッグすることにより，テキストを選択する．
(5) スクロール：画面を適度な速度でスクロールする．
(6) ブックマーク登録：ページをブックマークに登録する．
(7) 保存：ページを保存する．
(8) 印刷：ページを印刷する．
(9) ウィンドウ移動：画面上におけるウィンドウの位置を移動する．
(10) ウィンドウサイズ変更：画面上におけるウィンドウの大きさを変更する．

上述の操作のうち，操作対象が Web ページのテキスト中の一部となることがある操作は，なぞり読み，リンクポインティング，リンククリック，テキスト選択の 4 種類である．これら 4 種類の操作をテキスト抽出に利用できるか否かを判断するためには，どの程度の数のユーザがこれらの操作を行っているかを明らかにする必要がある．そこで，上述の観察実験とは別に，20 人のユーザを対象に 10 分間の Web ページ閲覧中の操作を観察し，4 種類の操作が発生した回数を測定した．その結果を**図 1** に示す．操作の種類によって回数にばらつきはあるが，いずれの操作も操作を行うユーザが存在することが分かった．そこで本研究では，これら 4 種類の操作の対象となるテキスト部分がユーザが実際に興味を持った部分であるか否かを，実験により調べる．

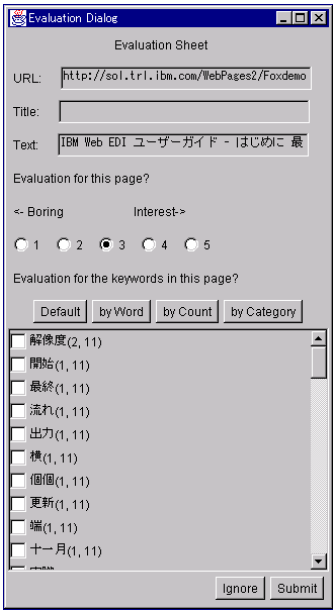

図 2 アンケート回答用ウィンドウ
Fig. 2 Questionnaire window.

4. 実験方法とシステムの実装

4.1 実験の目的と方法

情報検索や情報フィルタリングで行われる処理はキーワードを 1 つの単位とすることが多い．そのため本実験の目的は，3 章で述べた 4 種類の操作の対象となるテキスト部分を自動で抽出し，それらテキストに含まれるキーワードがユーザの興味のあるものか否かという点を検証することにある．実験方法を，以下に示す．

(1) 被験者に，あらかじめインターネット上の Web ページから閲覧したいページを探して決めておいてもらう．
(2) 被験者は，決めておいたページを自由に閲覧する．
(3) 被験者は，ページを移動するたびに，移動する直前に閲覧していたページについてのアンケートに答える．このアンケートでは，ページから抽出したすべてのキーワードを専用のウィンドウ（**図 2**）に表示し，被験者はこれらのキーワードのうち興味のあったものにのみチェックを入れる．
(4) 実験者は，被験者が指摘した興味のあったキーワードとシステムが自動抽出したテキストに含まれるキーワードとを比較して評価のための指標を算出し，抽出したテキストの有効性を検証する．

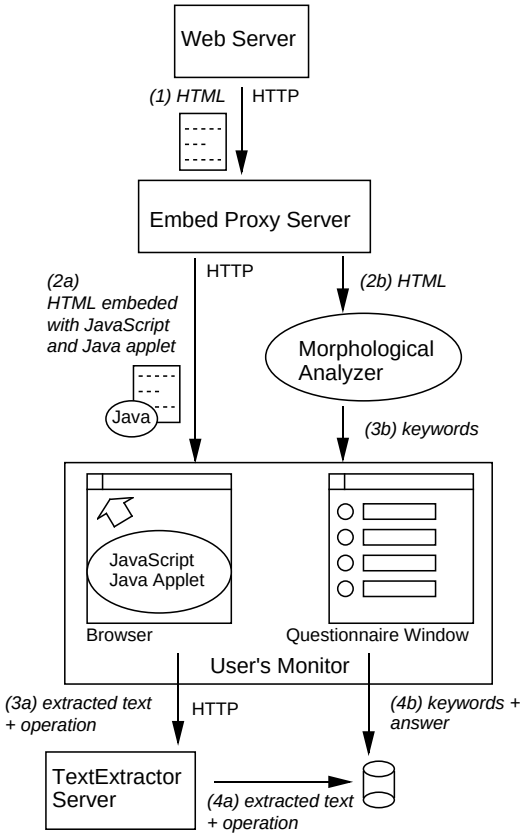

図 3 実験システム構成
Fig. 3 System structure for the experiment.

4.2 実験システム TextExtractor

本実験のために，ユーザの興味を示すと思われる操作の対象となるテキスト部分を自動抽出するシステム「TextExtractor」を実装した．本研究では，Microsoft 社の Internet Explorer のようなユーザが通常使っている Web ブラウザ上で簡単に実験を行うことができるように，また既存の Web アプリケーションに組み込みやすいという実用面も考慮して，JavaScript と Java を用いて TextExtractor を設計した．

図 3 に実験システムの構成を示す．TextExtractor の JavaScript と Java アプレットのプログラムは，プロキシサーバ（Embed Proxy Server）で Web ページの中に埋め込まれる[18]．また，図 2 のアンケートウィンドウを自動生成するため，Embed Proxy Server は，通過する HTML 文書中から HTML タグを除去して，そのテキストを形態素解析器[19),20)]に渡している．TextExtractor が抽出したテキストは，HTTP プロトコルで TextExtractor サーバに集められる．

4.3 操作抽出とテキスト抽出

TextExtractor の JavaScript のプログラムは，ブラウザ上でユーザの操作イベントを DOM（Document Object Model）[21] インタフェース経由で検出する．その後，座標などの情報を付加して通知用のフォーマットに整形し，Java アプレットのプログラムに通知している[22),23)]．Java アプレットのプログラムでは，それらの操作イベントを解析することで 4 種類の操作を抽出し，操作の対象となるテキスト部分を抽出している．操作イベント，通知用フォーマットとその記述例，具体的な操作抽出方法，および抽出用パラメータは付録に示す．

操作抽出では，いくつかの操作抽出用パラメータを用いているが，それらのパラメータは 5 人のユーザの Web ページ閲覧時の操作を観察・記録し，そのときの操作イベントを解析することで，経験的に設定している（付録・表 4）．3 章の事前調査，およびここでの観察・記録におけるなぞり読みの操作では，ユーザは正確に読んでいる行をなぞるというよりも，マウスポインタを半ば無意識に短く右方向に動かしている程度であった．そこで，短い右方向へのマウスの移動をなぞり読みの操作と認識するよう，パラメータを設定している．

テキスト抽出には，DynamicHTML[24)] の機能を利用している．テキストをマウスで選択すると，その範囲を指し示す selection オブジェクトが自動で生成されるため，テキスト選択の操作からのテキスト部分抽出にはこのオブジェクトが利用できる．また，ある座標上にある文字のページ全体のテキストにおける位置を特定することができため，なぞり読み，リンククリック，リンクポインティングの操作からのテキスト部分抽出にはこの機能が利用できる．抽出するテキスト部分は操作の直接対象となったものだけであるが，なぞり読みの操作に関しては，マウスポインタのある行とその 1 つ上の行のテキストも合わせて取得している．これは，3 章の事前調査では，主に読んでいる行に沿ってマウスを動かす場合と，読んでいる行の下にマウスポインタを持ってきて動かす場合との 2 通りが確認されたためである．

5. 実験結果と評価

5.1 実験での閲覧の様子

実験には年齢が 20 代，30 代の男女 5 人（男性 2 人，女性 3 人）のユーザが参加し，合計 120 ページ分のデータを解析に使用した．この実験での，各ユーザの閲覧の仕方，各ユーザが閲覧したページの特徴（1 ページ中の平均キーワード数も示す），各ユーザが興味ありとチェックしたキーワード数の 1 ページあたりの平均，閲覧ページ数を**表 1** に示す．

5.2 評価の焦点

本節の評価では，まずそれぞれの種類の操作の対象となるテキスト部分は，ユーザが興味を持った部分であるのか否か，すなわちユーザが興味を持ったキーワードの割合が文書全体に比べて高いか否かを検証する．その後，4 種類の操作からテキスト部分を抽出するシステム TextExtractor と他のキーワード抽出手法との比較を行う．ここでは，TextExtractor のテキスト抽出の判断基準にどの程度妥当性があるかを，ランダムにキーワードを抽出する手法と比較して検証する．また，情報検索や情報フィルタリングにおける代表的なキーワード選択手法である tf・idf とも比較する．

本実験では，評価のための指標として以下の 3 つを算出する．

(1) キーワード精度
(2) キーワード再現率
(3) ノイズ再現率

指標 (1) は，抽出したキーワードのうち，ユーザが興味を持ったキーワードの割合である．指標 (2) は，ユーザが興味を持ったキーワードのうち，抽出することができたキーワードの割合を示すものである．情報検索や情報フィルタリングでは，興味のあるキーワードのうち使用されるキーワードが少ない場合，本来ユーザの興味があるページが推薦されなくなるため，重要な指標といえる．指標 (3) は，ユーザが興味を持たなかったキーワード（ノイズ）のうち，抽出してしまったノイズの割合を示すものである．ページ全体のテキスト中のキーワードを使って適合性フィードバックを行う場合，これらのノイズが検索精度を低下させるため，重要な指標といえる．また，1 からこの値を減じると，ノイズを削減した割合（ノイズ削減率）と見ることができる．

情報検索やフィルタリングでの利用を考えたときに，指標 (1) により，抽出したキーワード群の良し悪しを評価することができる．また，指標 (1) に加え，指標 (2)，(3) により，抽出方式の良し悪しを評価することができる．具体的には，指標 (1)，(2)，(3) は，以下の式で算出する．

(1) キーワード精度 $=\mid B \mid / \mid A \mid$
(2) キーワード再現率 $=\mid D \mid / \mid C \mid$
(3) ノイズ再現率 $=\mid F \mid / \mid E \mid$

上述の式中の A，B，C，D，E，F はそれぞれ，以下の意味を持つ．

表 1 実験での閲覧に関するデータ
Table 1 Browsing data for the experiment.

ユーザ	閲覧の仕方	閲覧ページの特徴	平均チェックキーワード数	閲覧ページ数
ユーザ A	ニュースサイトが発行するメールマガジンのリンクをメールソフト上でクリックすることで各ニュース記事を閲覧.	テキストを主体としたページ．バナー広告や記事に関係する写真の画像があり．（1 ページあたりの平均キーワード数：198）	3.8	20
ユーザ B	自動車に関する専門サイトのトップページから，読みたい記事を探し出して閲覧．ニュースサイトが発行するメールマガジンのリンクをメールソフト上でクリックすることで各ニュース記事を閲覧.	多くのリンクからなるトップページと，テキストを主体とした記事のページ．バナー広告の画像があり．（1 ページあたりの平均キーワード数：351）	3.6	20
ユーザ C	個人の随筆集のページと個人のグルメ情報のページを，トップページから読みたい記事を探し出して閲覧.	多くのリンクからなるトップページと，テキストを主体とした記事のページ．記事に関係する写真の画像があり．（1 ページあたりの平均キーワード数：241）	4.7	29
ユーザ D	インデックスサービスを提供するページのリンク集から，人気歌手に関する Web サイトを選択して，各サイトでコンサート情報や掲示板などを閲覧．ニュースサイトが発行するメールマガジンのリンクをメールソフト上でクリックすることで各ニュース記事を閲覧.	多くのリンクからなるページ，いくつかのリンクからなる各サイトのトップページ，表や箇条書きによるデータを表示するページ，および掲示板のページ．画像は少なめ．テキストを主体としたページ．バナー広告や記事に関係する写真の画像があり．（1 ページあたりの平均キーワード数：157）	1.1	25
ユーザ E	旅情報を提供する市町村，および個人のサイトを，トップページからメニューを選択して閲覧.	いくつかのリンクからなるトップページと，テキストと画像からなるページ．画像は地図など大きいものもあり．（1 ページあたりの平均キーワード数：124）	2.7	26

平均チェックキーワード数：興味ありとチェックしたキーワード数の 1 ページあたりの平均

- A：抽出したキーワードの集合
- B：A のうち，ユーザが実際に興味を持ったキーワードの集合
- C：テキスト全体におけるユーザが実際に興味を持ったキーワードの集合
- D：C のうち抽出できたキーワードの集合
- E：テキスト全体におけるノイズ（ユーザが実際に興味を持たなかったキーワード）の集合
- F：E のうち抽出してしまったノイズの集合

5.3 操作種類ごとの妥当性の検証

図 4 に，操作の種類ごとに抽出したテキストにおけるキーワード精度と文書全体におけるキーワード精度を示す．どのユーザも 4 種類の操作すべてにおいて，抽出したテキストは文書全体よりキーワード精度が高いことが分かる．

また**図 5** に，それぞれのユーザが各種類の操作を 1 ページあたりどの程度の回数行ったかを示す．図 5 から，なぞり読みとリンクポインティングの操作には，その操作を行う頻度に個人差があることが分かる．また，リンククリックの操作も値にばらつきがあり，これは各ユーザの閲覧の仕方による差であることが分かった．たとえば，ユーザ A の値が低くなっているのは，ユーザ A はニュースサイトが発行するメールマガジンのリンクをメールソフト上でクリックすることで各ニュース記事を閲覧したため，Web ページ上のリンクをほとんどクリックしなかったためである．

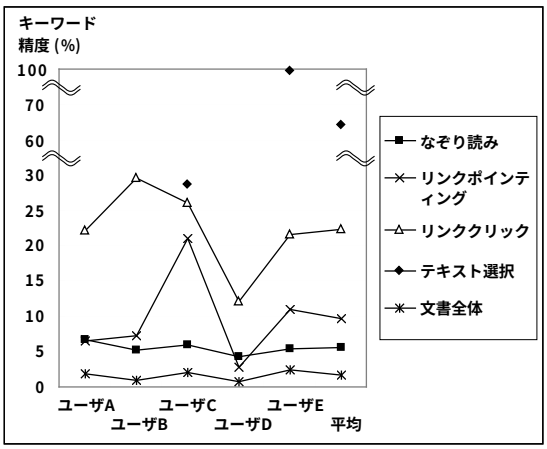

図 4 操作種類ごとのキーワード精度
Fig. 4 Keyword precision for each type of operation.

このように，操作の種類によってはその操作を行う頻度に個人差が見られたが，4 種類の操作すべてにおいて，操作の対象となるテキスト部分は文書全体に比べて，高い割合で興味のあるキーワードを含むことが検証された．

5.4 他手法との比較

ランダムや tf・idf でキーワードを抽出する場合，文

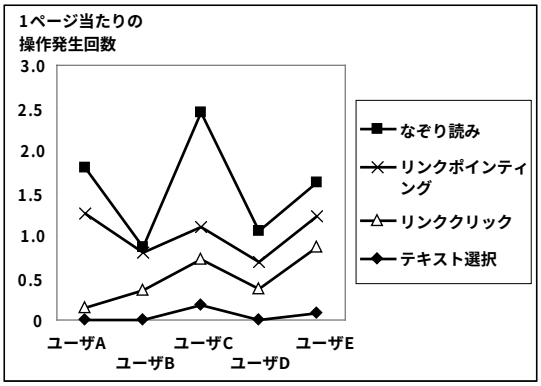

図 5 1ページあたりの操作発生回数
Fig. 5 Number of performed operations per a page.

書全体から任意の割合でキーワードを抽出することができる．一方，TextExtractorでは任意の割合でキーワードを抽出することができない．そのため，TextExtractorがページ全体のキーワードからどれだけ絞り込んでキーワードを抽出するかを表すキーワード絞り込み率を算出する．ランダム，tf・idfでもこの割合でキーワードを抽出することとし，同じ割合でキーワードを抽出した場合で比較を行う．キーワード絞り込み率は，具体的には以下の式で算出する．

キーワード絞り込み率 $=|H|/|G|$

上述の式中の G，H はそれぞれ，以下の意味を持つ．

- G：文書全体のキーワードの集合
- H：G のうちTextExtractorが抽出したキーワードの集合

ランダムでの抽出には，文書全体のテキストからランダムにキーワードを抽出した場合のキーワード精度，キーワード再現率，およびノイズ再現率の期待値を計算した．キーワード精度の期待値は，ユーザが閲覧したページ全体のテキストのキーワード精度を算出することで得られる．キーワード再現率とノイズ再現率の期待値は，TextExtractorのキーワード絞り込み率と一致する．また，tf・idfでのキーワード抽出には，事前に文書集合を設定しておく必要がある．今回の実験では，各ユーザが閲覧したページ群をそのユーザの文書集合としてベクトル空間を作成し，tf・idfの重み順にキーワード絞り込み率分のキーワードを選択した．それら選択したキーワード群において，キーワード精度，キーワード再現率，およびノイズ再現率を計算した．

表 2 にキーワード絞り込み率，**図 6**，**7**，**8** にキーワード精度，キーワード再現率，およびノイズ再現率を示す．ランダムに抽出する方法に比べると，TextExtractorはキーワード精度，キーワード再現率ともに

表 2 キーワード絞り込み率
Table 2 Keyword narrowing rate.

ユーザ	キーワード絞り込み率 (%)
ユーザ A	9.74
ユーザ B	3.50
ユーザ C	8.62
ユーザ D	7.76
ユーザ E	14.32
平均	8.78

全ユーザの平均で約4倍となっている．ノイズ再現率では，TextExtractorはランダムに抽出する方法に比べて若干低い値（つまり良い値）となっているが，その差は非常に小さいものとなっている．これは，文書全体のキーワードの98%以上がノイズであるためである（図4において，文書全体における全ユーザ平均のキーワード精度は2%以下であるため，キーワードの98%以上はノイズであることが分かる）．このことから，ランダムに抽出するよりもTextExtractorの方が多くの興味のあるキーワードを抽出することができ，さらに同程度ノイズを削減できることが分かる．

tf・idfと比べると，全ユーザ平均ではTextExtractorのキーワード精度とキーワード再現率は，tf・idfの値の約1.4倍となった．またユーザA以外は，tf・idfよりもTextExtractorの方が，キーワード精度，キーワード再現率ともに値が高くなった．ユーザAとユーザB〜Eの違いに着目すると，ユーザAはIT関連のニュース記事のページのみを閲覧していたのに対し，ユーザB〜Eは，Webサイトのトップページ，リンク集，掲示板，日記，随筆集，各種情報を表などで列挙したものなど，様々な形態のページも閲覧していた．tf・idfは，文書中における単語の出現頻度を基にキーワードの重み付けを行うため，ニュース記事のようなまとまった量の文章を含む文書に対して非常に有効な手法といえる．ユーザAの閲覧したページは，すべてニュース記事で，まとまった量の文章を含んでいたため，tf・idfのキーワード精度とキーワード再現率が高くなったと考えられる．ユーザB〜Eの閲覧したページは，必ずしもまとまった量の文章を含んでいたわけではないため，tf・idfのキーワード精度とキーワード再現率が低くなったと考えられる．一方TextExtractorは，文書中のキーワードの出現頻度に関係なく，マウスの操作からキーワードを抽出しているため，ページにまとまった量の文章が含まれていなくても，興味のあるキーワードを抽出している．したがって，TextExtractorは，tf・idfでは性能が発揮しにくい多様な形態のページに対しても，高い精度で興

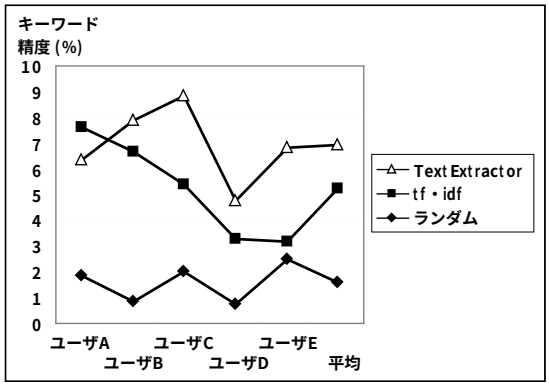

図 6 キーワード精度
Fig. 6 Keyword precision.

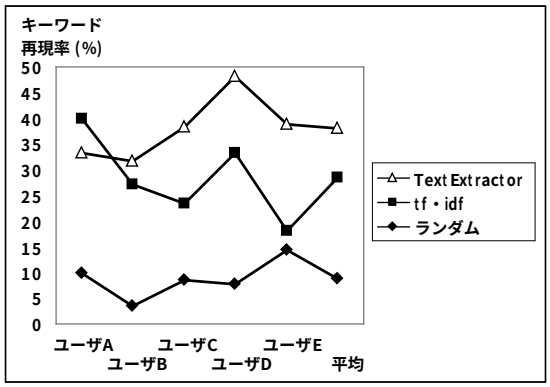

図 7 キーワード再現率
Fig. 7 Keyword recall.

味のあるキーワードを抽出することができるといえる．

5.5 結　　論

なぞり読み，リンクポインティング，リンククリック，テキスト選択の 4 種類の操作の対象となるテキスト部分は，文書全体に比べて興味のあるキーワードの割合が高いことが分かった．これら 4 種類の操作を使ってテキスト部分を抽出するシステム TextExtractor を実装した．TextExtractor を用いた実験により，tf・idf では性能が発揮しにくい多様な形態のページに対しても，TextExtractor は高い精度で興味のあるキーワードを抽出できることが分かった．そのため，TextExtractor が抽出したキーワードを適合性フィードバックに利用することにより，高精度な Web ページ検索が可能となることが期待できる．

また，実験では 5 人のユーザに一般の Web ページを普段どおりに閲覧してもらった．その結果，通常の閲覧で発生する操作を利用するだけで，上述のようにユーザが興味のあるキーワードをより高い精度で抽出

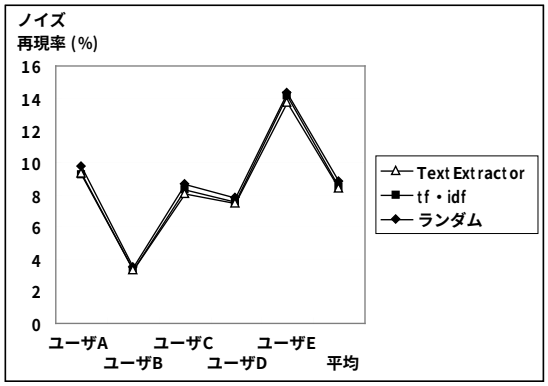

図 8 ノイズ再現率
Fig. 8 Noise recall.

することができた．このことから，ユーザに興味に関するアンケートに答えてもらったり，閲覧したページを評価してもらったりする手間をかけさせなくても，興味に関する情報が得られることが分かった．これにより，ユーザは意識してキーワードを入力したりページを選択したりすることなく，適合性フィードバックなどの検索を支援する機能を利用できるようになると期待できる．

なお，今回の実験では被験者 5 人のうち 1 人には閲覧したページの内容に偏りがあった．しかし，被験者全員において，TextExtractor はキーワードの出現頻度によらず，任意の形態のページにおいて有効であることがうかがわれたので，この被験者数でも議論が可能と判断した．

6. 議論と今後の方向性

6.1 議　　論

5 章の実験では，ユーザごとにキーワード精度，キーワード再現率，およびノイズ再現率を評価した．ここでは，4 種類の操作を操作の種類ごとに比較する．**表 3** に，操作の種類ごとに，全ユーザの抽出テキストにおけるキーワード精度，キーワード再現率，およびノイズ再現率を示す．キーワード精度に着目すると，なぞり読みやリンクポインティングのように半ば無意識に行う操作は，テキスト選択やリンククリックのような意識して行う操作に比べて精度が低くなっている．キーワード再現率に着目しても，操作の種類によって差があり，キーワード精度の高いテキスト選択やリンククリックの操作だけでは，キーワード再現率は高いものにならないことが分かる．情報検索というアプリケーションを考えた場合，数個のキーワードを使うキーワード検索には，キーワード精度の高いテキスト

表 3 操作の種類による違い
Table 3 Difference among the types of operations.

操作種類	キーワード精度 (%)	キーワード再現率 (%)	ノイズ再現率 (%)
なぞり読み	5.83	20.91	5.21
リンクポインティング	10.15	8.85	1.21
リンククリック	22.76	15.01	0.79
テキスト選択	50.00	1.34	0.02

選択やリンククリックの操作を利用することが有効と思われる．一方，多くのキーワードを必要とするベクトル空間モデルには，再現率の高いなぞり読みやリンクポインティングの操作もあわせて利用することが有効と思われる．また，ベクトル空間モデルを作成する場合，抽出に利用した操作の種類によってキーワードの重みを変更することも考えられる．このような形で，アプリケーションによって，利用する操作の種類を選択したり，操作の種類を重みなどに考慮したりすることが必要になってくると考える．

6.2 今後の方向性

5 章の評価ではキーワードに対して興味を持ったか否かを考えたが，TextExtractor の重要な特長として，ユーザが興味を持ったテキスト部分を文や行のまま抽出することがあげられる．このため，そのテキストに含まれるキーワードを取得するだけではなく，その文章そのものの意味や前後の文章を含めた文脈を分析することで，ユーザがあるトピックに対してどのように興味を持ったのかを知ることが期待できる．たとえば，同じ「Java」というキーワードでも，「Java 関連商品の市場は高まりつつある」と「Java 関連商品の性能は高まりつつある」では，そのキーワードの出現する文脈が異なる．前者の文に興味を持ったユーザは，マーケティングに興味を持っていた可能性が高く，後者の文に興味を持ったユーザは，技術に興味を持っていた可能性が高い．抽出したテキストの文脈を，自然言語処理の手法などを用いて分析し，情報検索やマーケティングにおけるユーザ分析[25]に応用することが，今後の課題としてあげられる．

また，ユーザの挙動を利用したテキスト部分抽出方法としては，視線を用いた方法も考えられる．視線を用いた方法は，特殊な装置を必要とするため実用面での問題はあるが，そのテキスト部分抽出に関する有効性に関しては，調査する価値があると考える．なぜなら，TextExtractor では取得不可能な，マウス操作に表れなかったユーザの興味を，取得できる可能性があるからである．視線を用いたテキスト抽出方法を，ランダムによる抽出，tf・idf，TextExtractor と比較実験することが，今後の課題として考えられる．

7. ま と め

本稿では，ユーザが Web ページ閲覧中に行う通常のマウス操作を利用して，ユーザが興味を持ったと思われるテキスト部分を自動抽出する手法を提案した．本研究では，まず事前調査を行い，ユーザの興味と関連がある可能性のある操作として，なぞり読み，リンクポインティング，リンククリック，テキスト選択の 4 種類の操作を発見した．これら 4 種類の操作の対象となるテキスト部分を文や行の単位で抽出する実験システム「TextExtractor」を実装し，抽出したテキスト部分が実際にユーザが興味を持った部分か否かを実験により検証した．

実験は，5 人のユーザに普段どおりの閲覧をしてもらうことで行った．その結果，4 種類すべての操作において，操作の対象となるテキスト部分は文書全体に比べて，高い割合で興味のあるキーワードを含むことが確認された．また，TextExtractor とランダムにキーワード抽出する場合の期待値と比較すると，TextExtractor は約 4 倍の精度で興味のあるキーワードを抽出することが確認された．このことから，ユーザにアンケートに答えてもらう手間をかけさせることなく，ユーザが興味を持ったテキスト部分を抽出できることが分かった．また，代表的なキーワード選択手法である tf・idf とも比較をした結果，TextExtractor は約 1.4 倍の精度で興味のあるキーワードを抽出することが確認された．また，tf・idf では性能が発揮しにくい掲示板やリンク集といった多様な形態のページに対しても，TextExtractor は高い精度で興味のあるキーワードを抽出できることが分かった．このため，TextExtractor が抽出したテキストを適合性フィードバックに用いれば，より高精度な情報検索ができるものと期待される．

今後は，抽出したテキストを適合性フィードバックに使用したときの効果について検証していきたい．また，抽出したテキストの文脈を分析して，情報検索やマーケティングにおけるユーザ分析にも応用していきたいと考えている．

謝辞 本研究の実験システムを構築するのに必要な形態素解析器のプログラムを快く提供してくださった，IBM 東京基礎研究所の野美山浩研究員に感謝いたします．また，実験に参加してくださった IBM 東京基礎研究所の所員の皆様に感謝いたします．

参 考 文 献

1) Broder, A., et al.: Graph Structure in the Web, *Proc. 9th International World Wide Web Conference, Computer Networks and ISDN Systems*, Vol.33, pp.309–320 (2000).
2) Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley (1989).
3) Resnick, P., et al.: Recommender Systems, *Comm. ACM*, Vol.40, No.3, pp.56–89 (1997).
4) Belkin, N.J. and Croftk, W.B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin?, *Comm. ACM*, Vol.35, No.12, pp.29–38 (1992).
5) 杉本雅則：情報収集システムにおけるユーザモデリングと適応型インタラクション，人工知能学会誌，Vol.14, No.1, pp.25–32 (1999).
6) Mulvenna, M.D., Anand, S.S. and Buchner, A.G.: Personalization on the Net using Web Mining, *Comm. ACM*, Vol.43, No.8, pp.123–125 (2000).
7) Shardanand, U. and Maes, P.: Social Information Filtering: Algorithm for Automating 'Word of Mouth', *Proc. CHI'95*, pp.210–217 (1995).
8) Yan, T.W. and Garcia-Molina, H.: SIFT — A Tool for Wide-Area Information Dissemination, *Proc. 1995 USENIX Technical Conference*, pp.177–186 (1995).
9) Resnick, P., et al.: GroupLens : An Open Architecture for Collaborative Filtering of Netnews, *Proc. CSCW'94*, pp.175–186 (1994).
10) Pazzani, M. and Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning*, Vol.27, No.3, pp.313–331 (1997).
11) Lang, K.: NewsWeeder: Learning to Filter NetNews, *Proc. ICML'95*, pp.331–339 (1994).
12) Smyth, B. and Cotter, P.: A Personalized Television Listings Service, *Comm. ACM*, Vol.43, No.8, pp.107–111 (2000).
13) Kantor, P.B., et al.: Capturing Human Intelligence in the Net, *Comm. ACM*, Vol.43, No.8, pp.112–115 (2000).
14) Morita, M. and Shinoda, Y.: Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval, *Proc. 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp.272–281 (1994).
15) Sakagami, H. and Kamba, T.: Learning Personal Preferences on Online Newspaper Articles from User Behaviors, *Proc. 6th International World Wide Web Conference, Computer Networks and ISDN Systems*, Vol.29, pp.1447–1456 (1997).
16) 吉田将志，吉高淳夫：Digital Reminder：ユーザの視点からの実世界指向データベースの構築とそのインタフェース，*Proc. 8th Workshop on Interactive Systems and Software* (*WISS'2000*), pp.101–110 (2000).
17) 大野健彦：IMPACT：視線情報の再利用に基づくブラウジング支援法，*Proc. 8th Workshop on Interactive Systems and Software* (*WISS'2000*), pp.137–146 (2000).
18) 古井陽之助，青木義則，土方嘉徳：Web ページに操作分析・自動実行プログラムを埋め込むための Web プロキシ，第 60 回情報処理学会全国大会，第 3 分冊，pp.421–422 (2000).
19) 奥村 学：自然言語処理関連ツールあれこれ—使えるフリーソフト，情報処理学会学会誌，Vol.41, No.11, pp.1203–1207 (2000).
20) 松本祐治：形態素解析システム「茶筌」，情報処理学会学会誌，Vol.41, No.11, pp.1208–1214 (2000).
21) http://www.w3.org/DOM/
22) Aoki, Y., Ando, F. and Nakajima, A.: Web Operation Recorder and Player, *Proc. International Conference on Parallel and Distributed Systems* (*IEEE ICPADS2000*), pp.508–508 (2000).
23) Aoki, Y., Ando, F. and Nakajima, A.: Creating Web-based Presentations by Demonstration, *IPSJ Journal*, Vol.42, No.2, pp.155–165 (2001).
24) Goodman, D.: *Dynamic HTML The Definitive Reference*, O'Reilly (1998).
25) 古井陽之助，青木義則，土方嘉徳，曾谷俊男，中島 周：Web ブラウザ操作の検出・再生技術とそのマーケティングへの応用の検討，情報処理学会研究報告，2000-DPS-97 (2000-CSEC-8), pp.73–78 (2000).

付 録

操作イベント

(1) Focus：オブジェクトにフォーカスを当てた場合に発生.
(2) Blur：オブジェクトからフォーカスが外れた場合に発生.
(3) Mouse move：マウスポインタが移動した場合に発生.
(4) Mouse down：マウスのボタンを押した場合に発生.
(5) Mouse up：マウスのボタンを離した場合に発生.
(6) Resize：ウィンドウの大きさを変更した場合に発生.
(7) Scroll：ウィンドウをスクロールさせた場合に発生.
(8) Click：オブジェクトをクリックした場合に発生.
(9) Mouse over：オブジェクトにマウスポインタが重なった場合に発生.

(10) Mouse out：オブジェクトからマウスポインタが外れた場合に発生.
(11) Select：テキストを選択した場合に発生.

フォーマット

イベント発生時刻，操作イベント種類，フレーム ID，オブジェクト ID，オブジェクト種類，座標などのデータ

操作イベント例

936332393593,blur,frame(0),7,BODY
936332407468,focus,frame(0),7,BODY
936332410218,mouseover,frame(0),7,BODY,215,0
936332410265,mousemove,frame(0),7,BODY,215,0

操作抽出手法（表 4）

(1) なぞり読み
Mouse move イベントが起こるたびに，画面の水平方向に対するマウスポインタの移動方向の角度，連続する Mouse move イベント間の時間を算出する（角度算出には，今回と n 回前の Mouse move イベントを用いる）．これらの値が，それぞれ閾値 A_r，T_r 以内であることを条件として，マウスポインタの途切れのない水平方向への移動を検出する．そのようなマウスポインタの移動が検出されれば，水平方向に移動した距離，速度を計算し，これらが閾値 L 以上と V 以下であったときになぞり読みの操作と判断する．
(2) リンクポインティング
リンクオブジェクトに対する Mouse over イベントの後に，Click イベントが起こらずに，一定時間 T_p をおいて Mouse out イベントが発生したときにリンクポインティングの操作と判断する．
(3) リンククリック
リンクオブジェクトに対する Click イベントを，リンククリックの操作と見なす.
(4) テキスト選択
Select イベントの後に Mouse up イベントが発生したときにテキスト選択の操作と判断する.

表 4 実験用操作抽出パラメータ
Table 4 Parameters for detecting operations.

パラメータ	値
角度 $A_r(\mid \tan\theta \mid)$	0.25
時間 T_r	750(msec)
距離 L	40(pixels)
速度 V	0.45(pixels/msec)
ヒストリ数 n	2
時間 T_p	750(msec)

(平成 13 年 5 月 31 日受付)
(平成 13 年 12 月 18 日採録)

土方 嘉徳（正会員）

昭和 48 年生．平成 10 年大阪大学大学院基礎工学研究科物理系専攻修士課程修了．同年日本アイ・ビー・エム（株）東京基礎研究所入所．現在，大阪大学大学院基礎工学研究科システム人間系専攻博士後期課程在籍．ヒューマンインタフェース技術，インターネット・アプリケーション技術の研究に従事．日本ソフトウェア科学会，ヒューマンインタフェース学会，IEEE 各会員.

青木 義則（正会員）

昭和 47 年生．平成 7 年九州大学工学部情報工学科卒業．平成 9 年同大学大学院システム情報科学研究科修士課程修了．同年日本アイ・ビー・エム（株）東京基礎研究所に入所．ヒューマン・コンピュータ・インタラクション，分散アプリケーション，XML の研究開発に従事．ACM，IEEE 各会員.

古井陽之助（正会員）

昭和 43 年生．平成 5 年九州大学大学院工学研究科情報工学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．現在，九州大学大学院システム情報科学府知能システム学専攻博士後期課程在籍．ヒューマンインタフェース学会会員.

中島 周

昭和 36 年生．昭和 60 年東京大学大学院工学系研究科電気工学専門課程修士課程修了．同年日本アイ・ビー・エム（株）東京基礎研究所入所．現在，日本アイ・ビー・エム（株）ビジネス・イノベーション・サービス所属．情報技術を活用したインターネット・ソリューションの構想策定，アーキテクチャ作成のコンサルティングに従事．ACM，IEEE 各会員.