

閲覧ウェブページ内の興味範囲推定に基づく ユーザ嗜好情報の抽出

宮脇 佑介[†] 河合由起子^{††} 中島 伸介^{††}

[†] 京都産業大学先端情報学研究科 〒603-8555 京都府京都市北区上賀茂本山

^{††} 京都産業大学コンピュータ理工学部 〒603-8555 京都府京都市北区上賀茂本山

E-mail: [†]i1258140@cc.kyoto-su.ac.jp, ^{††}{kawai,nakajima}@csc.kyoto-su.ac.jp

あらまし ユーザが普段行うウェブ検索の行動から，そのユーザの嗜好情報を推定することができれば，ユーザの嗜好に沿った，精度の高い情報推薦が可能であると考えられる．そこで我々は，ユーザのウェブ検索時のマウスの動きから，そのユーザが興味を示すテキスト情報を取得し，取得したテキスト情報と wikipedia のデータベースを用いた，ユーザの嗜好情報抽出を提案する．

キーワード 情報推薦，テキストマイニング，マウス履歴，wikipedia

Extraction of the taste topic of the user based on the interest range estimate in the reading web page

Yusuke MIYAWAKI[†], Yukiko KAWAI^{††}, and Shinsuke NAKAJIMA^{††}

[†] Faculty of Computer Science and Engineering, Kyoto Sangyo University Motoyama, Kamigamo, Kita-ku,
Kyoto-City, Kyoto, 603-8555 Japan

^{††} Faculty of Computer Science and Engineering, Kyoto Sangyo University Motoyama, Kamigamo,
Kita-ku, Kyoto-City, Kyoto, 603-8555 Japan

E-mail: [†]i1258140@cc.kyoto-su.ac.jp, ^{††}{kawai,nakajima}@csc.kyoto-su.ac.jp

Abstract If users' preference information can be inferred from users' web searching behavior, information recommendation with high precision satisfying users' preference will become possible. Therefore, we propose a method which can acquire users' advanced preference information by analyzing the operation history of the mouse during the web search, moreover database of Wikipedia.

Key words information recommendation, textmining, mouse movement history, wikipedia

1. はじめに

近年，ユーザの嗜好情報に応じて適切な情報を提供する情報推薦手法が注目されている．現行のウェブサービスには，情報推薦手法を採用し，成功を収めている主な例として Amazon.com が挙げられる．Amazon.com はサイト内でのユーザの商品閲覧履歴や購入履歴からユーザの嗜好情報を読み取り，好みに沿った商品を推薦する，協調フィルタリングを採用している．しかしながら，Amazon.com のウェブサービス以外の行動履歴を十分に考慮できていないと言える．

そこで我々は，特定のウェブサイトに限定しないウェブ閲覧履歴中のユーザの振る舞いから，ユーザの嗜好情報を抽出し，それらの情報に基づく情報推薦手法の研究開発を目指す．

これまでに，先行研究としてマウスイベントを考慮したユー

ザプロファイル抽出方式について提案してきた [1]

本稿では，ユーザの普段のウェブ検索の行動履歴である，ウェブページ閲覧時のマウスの操作履歴を含めた詳細なユーザの嗜好情報の推定を目的とした，ユーザがウェブページ内でマウス操作したウェブページのコンテンツ情報を用いた，ユーザがウェブページ内で興味を示した範囲の推定を行う，高度なユーザ嗜好情報の取得手法を提案する．

2. 関連研究

これまでの情報推薦の既存研究として，ユーザの特徴やウェブページの特徴を基にしたウェブページ検索のパーソナライズ化（個人々々に適した情報推薦）が数多く行われている．

既存の研究ではウェブ検索のパーソナライズ化の手順として，はじめに，ユーザやウェブページの特徴を抽出する．次に，抽

出されたユーザやウェブページのデータを独自に解析してプロファイルとして記録する．記録されたプロファイルの要素から類似性を発見し，情報推薦へと発展させる．

以下に既存研究であるユーザやウェブページの特徴抽出や，クラスタリングの手法を挙げる．

- ウェブページのコンテンツを用いたウェブページの特徴抽出
 - － 語句の意味的階層を用いたウェブページの特徴抽出 [2]
 - － 語句の組み合わせを用いたウェブページの特徴抽出 [3]
- ユーザの身近なコンテンツを用いたユーザの特徴抽出
 - － ウェブ閲覧履歴（検索クエリー・URL・クリック数）などからユーザの特徴抽出 [4], [5]
 - － ローカルデータ（メール・ローカルファイル）などからユーザの特徴抽出 [6]

上記のほか，ユーザやウェブページの特徴抽出の結果を用いた，重要度の付加やテキストの表記ズレによる特徴誤検出などの手法として，以下が挙げられる．

- ユーザのウェブ閲覧回数を用いたウェブページの重み付け [4], [5]
- 類義語を用いたクエリーの曖昧な表記の修正 [7]?[9]
- 用語の重要度による重み付け [10], [11]

また，プロファイルを用いたユーザやウェブページのクラスタリングを行う手法として，以下が挙げられる．

- 語句のカテゴリ分類を用いたコサイン距離によるクラスタリング [12]

本研究と関連のある既存研究として，近藤ら [13] の，ユーザの閲覧しているウェブページから，ユーザが興味を示すテキストを，wikipedia を用いて取得する，MyBoom 推薦システムがある．

調査として，ユーザの視線とマウスの位置の相関関係を明確にした報告 [14]?[16] や，ユーザのウェブページ閲覧中の注目箇所推定にスクロール操作を用いる方法 [17] があり，マウス操作を用いたユーザの嗜好情報抽出が情報推薦の精度向上に繋がる可能性があることを示唆しているものもある．以下には本稿の研究と関連の深い既存研究の紹介を記す．

2.1 テキストデータから適切なキーワードの抽出

前田ら [18] は，言選 Web（日本語文献の検索を対称に専門用語抽出システム）を使用した日本語文献情報のナビゲーションシステムを作成した．これによりウェブページ抜粋のような少ないテキストデータからでも，より細かい重要度ランキングを得ることができるため，実現が容易で拡張も行うことができる検索システムが作成できるとしている．

成田ら [19] は，単語の上位概念（与えられたテキストデータからのタグ抽出）に着目している．単語 A の上位概念として存在する単語 B は，A をより一般的・抽象的・総称的に表現した単語であると考えられる．そこで，テキストデータから特徴語を抽出し，特徴語の中からオンライン辞書である Wikipedia に存在する単語の親カテゴリ（上位概念）を抽出する．これにより，従来の特徴語抽出によるタグ付けより正確なキーワードの抽出が可能としている．

近藤ら [20] は，与えられたテキストデータ内に含まれる Wikipedia の見出し語をテキストデータの重要語候補として抽出する．次に，共参照処理による人物などの略称の出現頻度修正を行い，後に BM25・TF-IDF に基づく出現頻度による重要度算出を行う．これにより高い精度で重要語を抽出し，高精度な関連コンテンツを推薦できることを示した．

2.2 ユーザのウェブ検索の行動履歴を用いた関連コンテンツの推薦

鶴原ら [21] は，ウェブブラウザの拡張機能を用いて各ユーザの閲覧履歴を収集し，それらの各ユーザの閲覧履歴（閲覧ウェブページの URL のみ）を利用して，類似度を計算することで協調フィルタリングによるウェブページの推薦を行う．

高須賀ら [22] は，鶴原ら [21] と同じく，ウェブブラウザの拡張機能を利用して収集した閲覧履歴を使用した協調フィルタリングを用いた推薦手法を評価している．ただし鶴原らの手法ではユーザの閲覧したウェブページの URL を基に協調フィルタリングを行っていたが，高須賀らは有名サイトの除外や，単語抽出を用いた検証を行うことでユーザの特徴を記したユーザプロファイルを自動生成させ，ユーザプロファイルを用いて類似度の高いユーザを発見することを明らかにしている．また，単語抽出処理には，一般性の高い単語を除外するフィルタ機能の必要性を示した．

Matthijs ら [23] は鶴原ら [21] や高須賀ら [22] のように，ユーザのウェブ検索行動をウェブブラウザの拡張機能によってデータ収集を行い，それらを解析することで，ユーザの特徴を記したユーザプロファイルを自動生成する．生成されたユーザプロファイルには前記のように，ユーザが閲覧したウェブページの URL や，抽出された単語だけでなく，ユーザが利用した検索サイトから目的のウェブページを発見するまでのウェブページの遷移経路も記録することで，より詳細な類似ユーザの発見を行うことができたとしている．

Hamid ら [24] は Matthijs ら [23] のようにユーザのウェブ検索行動をユーザプロファイルとして記録し，ユーザプロファイルを情報推薦に利用する．ただしウェブページの URL やウェブページの遷移経路といったユーザの履歴を基に情報推薦を行うのではなく，ユーザの履歴の解析から得られる，ユーザがあるキーワードでウェブ検索を初めてからの経過時間やウェブページのクリック数などを分析する．これにより得られるユーザの興味を示すキーワードとその度合いをユーザプロファイルとして記録することで，従来より精度の高い情報推薦ができた．今後の発展として，より多くのユーザのウェブ閲覧履歴を利用すれば，今以上の推薦精度も期待できるとしている．

2.3 ユーザの視線とマウスの位置に見られる相関

Chen ら [14] は，マウスの位置と視線には相関があり，マウスの位置からユーザの興味を示す部分を予測することができるとしている．

Muller ら [15] は，実験の結果から，35%のユーザはウェブページを閲覧する際にマウスで閲覧部分をなぞることがわかっている．

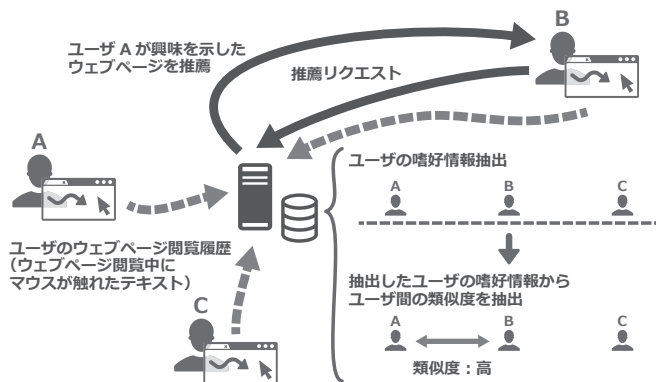


図 1 開発中の推薦システムの概要

2.4 ウェブ検索時のコンテキストを用いた情報推薦

既に我々は、インタラクションの分析に基づくソーシャルリレーションの発見手法について提案している [25], [26]

これらはユーザーのウェブ検索時の行動履歴から、コンテキストを取得することが出来れば、高精度な情報推薦やコミュニティ支援などのインタラクションの取得・分析が行えるとした。ただし提示しているのは概念に留まっており、明確に機能するシステムとしての提案まで至るには、コンテキストの定義や、インタラクションの分析に対する検討などが不十分である。

3. ユーザ嗜好情報の抽出手法

本章ではユーザー嗜好情報の抽出手法の概要や、処理内容について記す。

3.1 ユーザ嗜好情報抽出手法の概要

提案するユーザー嗜好情報の抽出手法は、ユーザーのウェブページ閲覧履歴を用いる。取得したユーザーのウェブページ閲覧履歴を基に、ユーザーの嗜好情報を抽出し、他のユーザーとの類似度を抽出する。もし、類似ユーザーを発見できれば、協調フィルタリングの手法を使いユーザーへウェブページの推薦を行う。

ユーザー嗜好情報の抽出手法を構成する重要な要素として、ユーザーのウェブページ閲覧履歴の取得がある。ウェブページ閲覧履歴を用いた従来手法には、光正ら [13] の MyBoom 推薦システムがあり、ユーザーが閲覧しているウェブページ内のコンテンツ情報を用いた手法があるが、我々の提案する推薦システムでは、ウェブページ閲覧履歴として、ユーザーがマウスで操作したコンテンツに着目する。

既に [14], [15] によりマウスの位置と視線には相関があることが確認できている。そこで、ユーザーがウェブページ内でマウス操作したコンテンツ情報を、ユーザーがウェブページ内で興味を示した情報と見なし、ユーザー嗜好情報を抽出する。

ユーザー嗜好情報の抽出手法と、推薦システムで利用する場合の概要図を 図 1 に示す。

3.2 ユーザのウェブページ閲覧履歴の取得

ユーザーのウェブページ閲覧履歴取得の一連の例を 図 2 で示す。ユーザーがウェブ検索を行なっている最中、ウェブブラウザに表示されたウェブページのコンテンツ上をマウスが移動している。この時、マウスが重なったウェブページ上のコンテンツ

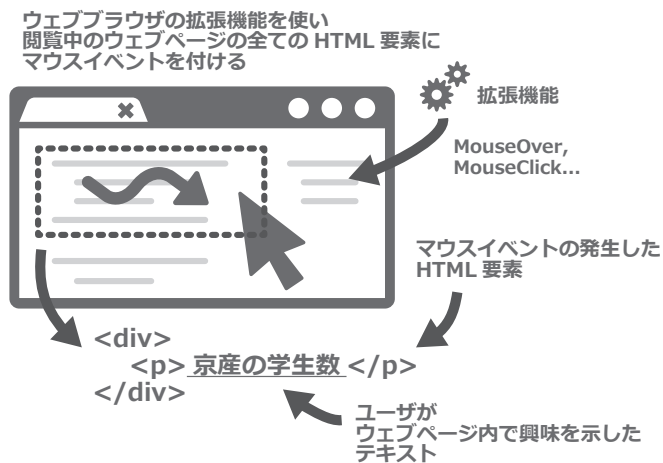


図 2 ウェブページ閲覧履歴の取得

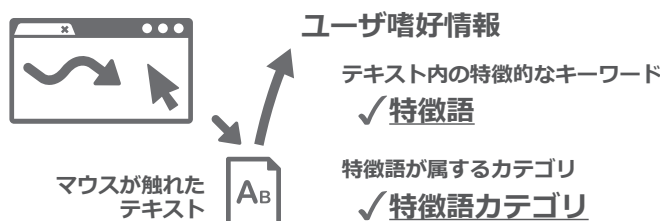


図 3 ユーザの嗜好情報

にはテキスト情報が含まれており、閲覧中のウェブページ内でユーザーが興味を示した部分と考えられる。

そこで、ユーザーのウェブページ閲覧中のマウス操作履歴を取得するために、ウェブブラウザ拡張機能を使う。任意のユーザーがウェブページを開いた場合、拡張機能はウェブページの読み込み完了次第、読み込んだウェブページの全ての HTML 要素へ、マウスイベント（MouseOver, MouseClick）を付与する。もしユーザーがウェブブラウザに表示されたウェブページに対してマウス操作を行った場合、HTML 要素へ付与したマウスイベントが発生する。そのため、マウスイベント発生元の HTML 要素を取得することで、ユーザーがマウス操作した HTML 要素のテキスト情報を取得する。

一連の動作例を示した 図 2 では、取得されたウェブページ閲覧履歴であるテキスト情報として「京産の学生数」と取得される。

3.3 ユーザの嗜好情報抽出

本節では、前節 3.2 で取得した、ウェブページ閲覧履歴を用いた、ユーザーの嗜好情報抽出手法について説明する。

ユーザーの嗜好情報の構成は、図 3 となる。

ユーザーの嗜好情報は、ウェブページ閲覧履歴から得られるテキストの特徴で構成する。ユーザーの嗜好情報を構成する要素として、特徴語と、特徴語カテゴリの2つを用いる。

3.3.1 特徴語の抽出

本項では、前節 3.2 で取得した、ユーザーのウェブ閲覧履歴から、特徴語を抽出する手法について説明する。

前節 3.2 の例で取得したテキスト情報を用いる特徴語を抽出の例を、図 4 で示す。

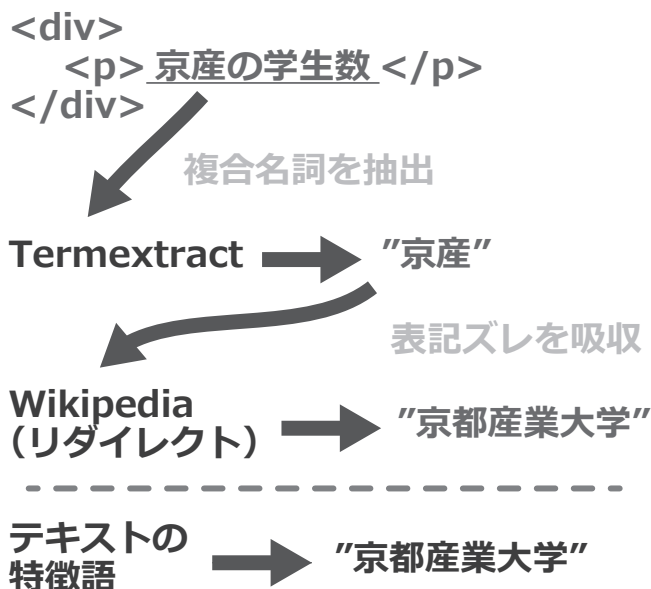


図 4 特徴語の抽出

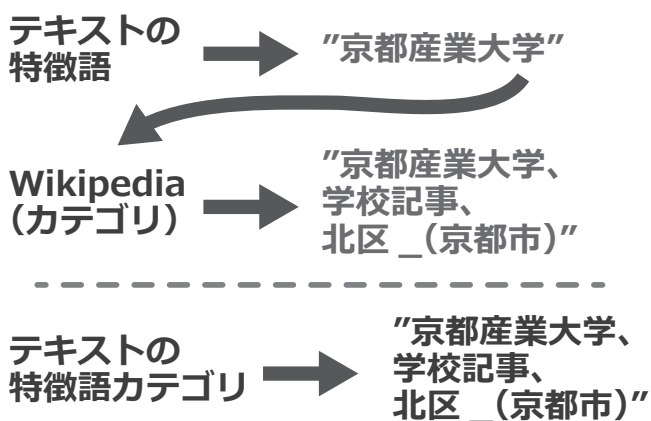


図 5 特徴語カテゴリの抽出

ユーザのウェブページ閲覧履歴には、テキスト情報が含まれている。ウェブページ閲覧履歴のテキスト情報から、特徴的な語句を抽出するために、本手法では、中川らの開発した専門用語自動抽出 Termextract [27] を用いる。

Termextract は、専門用語の多くが複合名詞であることに着目し、名詞の出現頻度や接続頻度を用いて、高い精度で専門用語の抽出を行なうシステムである。

ただし、Termextract で取得するウェブページ閲覧履歴内の特徴語には、正式名称や、略称などを含む場合がある。そのため、Termextract の出力結果を用いてユーザ間の嗜好の類似度を測る場合には、正式名称と、略称の整合性を取る必要がある。そこで、wikipedia を用いて、正式名称と、略称の整合性を取る。

3.3.2 特徴語カテゴリの抽出

本項では、前項 3.3.1 で抽出した特徴語から、特徴語カテゴリを抽出する手法について説明する。

前項 3.3.1 で取得した特徴語から、特徴語カテゴリを取得する例と結果を図 5 に示す。

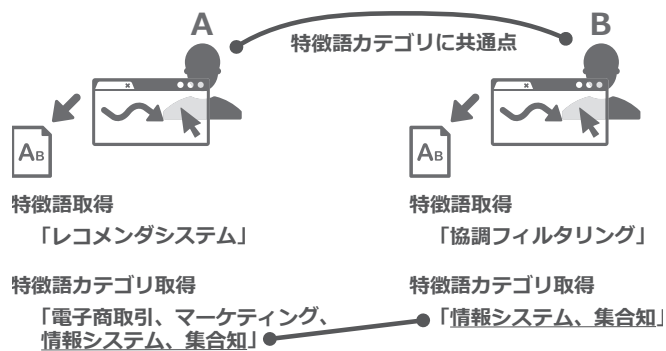


図 6 ユーザ間の類似取得

特徴語カテゴリの取得には、wikipedia のカテゴリ情報を用いる。前節 3.3 から、ユーザがマウス操作したテキスト情報「京産の学生数」を取得する。次に、テキスト情報を用いて前節 3.3 で取得した特徴語は「京都産業大学」となる。取得した特徴語を用いて、wikipedia のカテゴリ情報へ問い合わせ、特徴語カテゴリを抽出する。

3.4 抽出したユーザの嗜好情報を用いた ユーザ間の類似の取得

ウェブページ閲覧履歴から抽出した特徴語を用いて、特徴語カテゴリを抽出する必要性について、図 6 を例に説明する。既に、多人数のユーザから、ウェブページの閲覧履歴を用いて、ユーザがウェブページ内で注目したコンテンツの専門用語を抽出済みとする。この場合、多人数のユーザから、同じ情報に興味を持つユーザを発見する場合、ユーザ間で専門用語が一致している必要がある。例えば、あるユーザ A の持つ専門用語に「レコメンダシステム」が記録されており、あるユーザ B の持つ専門用語に「協調フィルタリング」が含まれているとする。この場合「レコメンダシステム」と「協調フィルタリング」では、文字列は一致しない。そのため、特徴語だけではユーザ間の類似度を測ることが難しい。そこで、抽出済みの特徴語と、wikipedia を用いて、ウェブページ閲覧履歴から抽出した特徴語が属するカテゴリを表す、特徴語カテゴリを抽出する。

特徴語カテゴリの抽出には、ユーザの持つ専門用語と、wikipedia のカテゴリ情報を用いる。上記の例に対して、wikipedia のカテゴリ情報を用いると、「レコメンダシステム」のカテゴリとして「電子商取引、マーケティング、情報システム、集合知」の 4 つが記録されており、同じく「協調フィルタリング」では、カテゴリとして「情報システム、集合知」の 2 つが記録されている。wikipedia に記録されている特徴語のカテゴリの結果を、特徴語カテゴリとする。これにより、特徴語の文字列では一致しない「レコメンダシステム」と「協調フィルタリング」は、特徴語カテゴリにより「情報システム、集合知」の 2 つが文字列で一致しているため、関係性のある単語と推定できる。以上により、抽出した特徴語カテゴリを用いて、ユーザ A と、ユーザ B は同じ事に興味を示したと推定できる。

4. まとめと今後の予定

本稿では、ユーザの閲覧ウェブページから、ユーザの嗜好情報抽出手法について報告した。今後は、抽出したユーザの嗜好情報を用いた、ユーザ間の類似度測定手法の提案と、その報告について検討する予定。

謝 辞

本研究の一部は、文部科学省科学研究費助成事業（学術研究助成基金助成金）基盤研究（C）（課題番号：23500140）による。ここに記して謝意を表します。

文 献

- [1] 宮脇佑介, 張建偉, 河合由起子, 中島伸介. Web 閲覧コンテキストおよびマウスイベントを考慮したユーザプロファイル抽出方式の提案. 第 10 回日本データベース学会年次大会, 2012.
- [2] Chirita, Paul Alexandru and Nejdl, Wolfgang and Paiu, Raluca and Kohlschütter, Christian. Using ODP metadata to personalize search. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178–185, 2005.
- [3] Fang Liu, Clement Yu, and Weiyi Meng. Personalized Web Search For Improving Retrieval Effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, Vol. 16, No. 1, pp. 28–40, 2004.
- [4] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 43–52, 2008.
- [5] Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 667–674, 2010.
- [6] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 449–456, 2005.
- [7] Mark Sanderson. Ambiguous queries: test collections need more sense. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 499–506, 2008.
- [8] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Potential for personalization. *ACM Trans. Comput.-Hum. Interact.*, Vol. 17, No. 1, pp. 4:1–4:31, 2010.
- [9] Ryen W White and Steven M Drucker. Investigating behavioral variability in web search. *Proceedings of the 16th international conference on World Wide Web*, pp. 21–30, 2007.
- [10] Susan Gauch, Jason Chaffee, and Alexander Pretschner. Ontology-based personalized search and browsing. *Web Intell. and Agent Sys.*, Vol. 1, No. 3-4, pp. 219–234, 2003.
- [11] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 525–534, 2007.
- [12] 山中努, 田中祐也, 土方嘉徳, 西田正吾. 時空間情報を伴うテキストデータを用いた状況把握支援システム. 知能と情報: 日本知能情報ファジィ学会誌: journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6, pp. 691–706, 2010.
- [13] 近藤光正, 田中明通, 内山匡. MyBoom:Wikipedia に基づく Web 閲覧履歴からの興味情報推薦システム (不均質なライフログからのデータマイニング及び一般). 電子情報通信学会技術研究報告. LOIS, ライフインテリジェンスとオフィス情報システム, Vol. 109, No. 450, pp. 97–102, February 2010.
- [14] Mon Chu Chen, John R Anderson, and Myeong Ho Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. *CHI '01 extended abstracts on Human factors in computing systems*, pp. 281–282, 2001.
- [15] Florian Mueller and Andrea Lockerd. Cheese: tracking mouse movement activity on websites, a tool for user modeling. *CHI '01 extended abstracts on Human factors in computing systems*, pp. 279–280, 2001.
- [16] 杉田賢治, 福原知宏, 増田英孝, 中川裕志. Web ページ閲覧行動分析のためのブラウザ操作ログ収集ツール. 人工知能学会, 2011.
- [17] 杉田賢治, 福原知宏, 増田英孝, 山田剛一. Web ブラウザのスクロール操作に基づくユーザ注目箇所推定に関する分析. 人工知能学会, 2012.
- [18] 前田朗, 中川裕志. 関連学術用語による日本語文献情報への簡易ナビゲーションシステム. 2010.
- [19] 成田雄一, 増田英孝. 特徴語の上位概念を表すタグの自動生成. 2006.
- [20] 近藤光正, 中辻真, 田中明道, 内山匡. 重要語抽出を用いた外部 API からの関連コンテンツ推薦. 2010.
- [21] 鶴原翔夢, 高須賀清隆, 丸山一貴, 寺田実. 独立成分分析を用いた Web 閲覧履歴の解析と Web ページ推薦への応用. 2008.
- [22] 高須賀清隆, 丸山一貴, 寺田実. 閲覧履歴を利用した協調フィルタリングによる Web ページ推薦とその評価. 2007.
- [23] Nicolaas Matthijs and Filip Radlinski. Personalizing web search using long term browsing history. *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 25–34, 2011.
- [24] H Rastegari and S M Shamsuddin. Web Search Personalization Based on Browsing History by Artificial Immune System. *Int. J. Advance. Soft Comput. Appl.*, Vol. 2, No. 3, 2010.
- [25] 宮脇佑介, 中島伸介, 張建偉, 河合由起子. インタクションの文脈理解に基づくソーシャルリレーションの発見. 第 9 回日本データベース学会年次大会, 2011.
- [26] 宮脇佑介, 中島伸介, 張建偉, 河合由起子. ユーザの Web インタクション履歴の分析に基づくソーシャルリレーション発見手法の提案. 情報処理学会関西支部, 2011.
- [27] 中川裕志, 前田朗. 専門用語（キーワード）自動抽出システム.