

HITSに基づく Wikipedia ランキングアルゴリズムとユーザ履歴を用いた 個人適応型クエリ推薦

近藤 光正[†] 森田 哲之[†] 田中 明通[†] 内山 匡[†]

[†] 日本電信電話株式会社 NTT サイバーソリューション研究所 〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: {kondo.mitsumasa,morita.t,tanaka.akimichi,uchiyama.tadasu}@lab.ntt.co.jp

あらまし 本稿では、ユーザの Web 閲覧履歴から、ユーザがクエリ(キーワード)を入力することなく興味のある情報の検索が可能になるクエリ推薦手法を提案する。従来法では、主に形態素や IREX で定義された固有表現をキーワードとして抽出していたが、形態素単位では語長が短すぎてキーワードの意味が掴み難い問題があり、また固有表現においては固有表現の定義が狭いことにより抽出できないキーワードがある問題があった。さらに、両手法共通の問題としてどのキーワードが重要であるかといったキーワード固有の重要度が算出しにくい問題がある。そこで、本稿ではオンライン百科辞典である Wikipedia を用いることでこれらの問題の解決を目指す。本手法では、Wikipedia の見出語をキーワードとして定義し、Wikipedia のリンク構造を解析することでキーワード固有の重要度を算出する。そして、ユーザの Web 閲覧履歴から得られるテキスト情報と閲覧時間から、ユーザが興味を持つと思われるクエリを推薦する。実験の結果、ベースラインとして設定した TF・IDF に基づく従来手法に比べ、本手法の優位性が高いことが確認された。

キーワード 情報推薦, クエリ推薦, Wikipedia, リンク解析, キーワード抽出, 履歴解析

Personalized Query Recommendation Using HITS-Based Wikipedia Ranking Algorithm and User History

Mitsumasa KONDO[†], Tetsushi MORITA[†], Akimichi TANAKA[†], and Tadasu UCHIYAMA[†]

[†] NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation Hikarinooka 1-1,
Yokosuka-shi, Kanagawa, 239-0847 Japan
E-mail: {kondo.mitsumasa,morita.t,tanaka.akimichi,uchiyama.tadasu}@lab.ntt.co.jp

Abstract In this paper, we propose query recommend method by using user's web history, without any inputting query. Conventional keyword extraction methods mainly extract morph or named entity defined by IREX. Morph is too short to understand keywords. And named entity defined by IREX is narrow range of a meaning and don't cover user's all interesting. And both of those method have problem which is "What keyword is important?". We will solve this problem in this paper by using online encyclopedia Wikipedia. In my method, we difine extraction keyword using Wikipedia's entry. And we analyze Wikipedia's link structure for calculating keyword importance. And then, we recommend keywords in which the user will get interested by using inspection Web text and time from user Web browsing history. The result of an experiment, our method obtained significant improvements compare to baseline methods that using morph and TFIDF.

Key words Information Recommendation, Query Recommendation, Wikipedia, Link Analysis, Keyword Extraction, History Analysis

1. はじめに

Web の世界が進歩するにつれて、Amazon.com のようにユーザのサイト内履歴から商品の推薦をするシステムや Google News [1] の様にニュースの閲覧履歴からニュース記事を推薦す

るシステムなど、ユーザの履歴からユーザの好みそうなアイテムを推薦する情報推薦技術が発達してきた。これらの技術はユーザが情報を検索する際に、クエリを入力することなく、好みのもしくは目的の情報にたどり着くことができるため、Web に慣れていないユーザだけでなく、すべてのユーザに有益な技

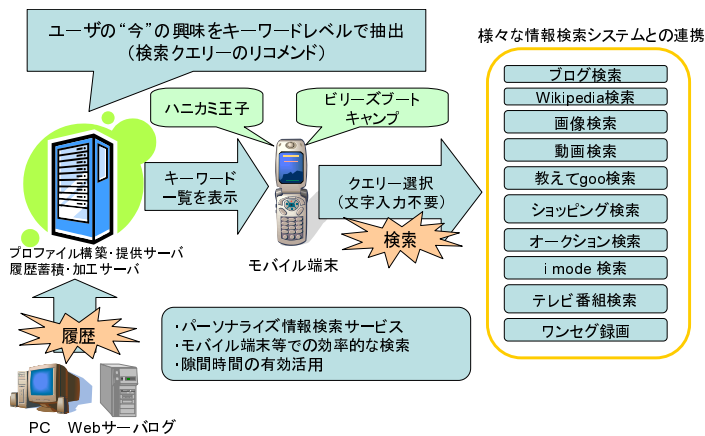


図 1 PC の履歴を用いたモバイル端末上での活用

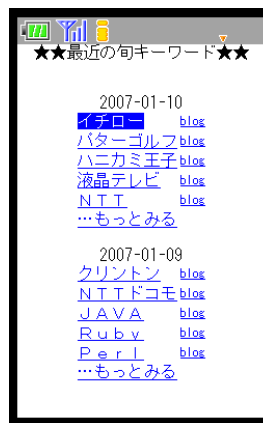


図 2 イメージ画面 1

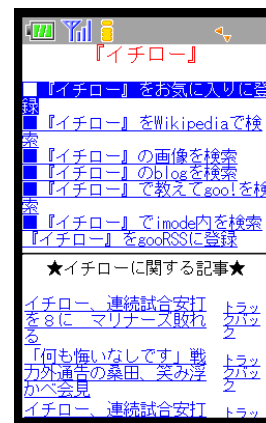


図 3 イメージ画面 2

術である。しかしながら、我々が必要な情報を検索する際の基本は、Web が格段に進歩した現在でも、検索システムの入力窓にクエリを入力することである。また、ユーザの求める情報は、最新のニュースや流行的な本だけでなく、今晚のテレビ番組や面白いブログ、さらにはオークションの出品情報や株価に関する記事等、様々なアイテムが考えられる。

そこで本稿では、ニュースや本といった特定のアイテムを推薦するのではなく、ユーザの嗜好を考慮した検索クエリを推薦する手法を提案する。現在、動画検索、Wikipedia 検索といった様々な分野に特化した検索システムが公開されている。そのため、ユーザの興味を持つクエリを推薦することで、API を通じて様々なシステムと連携できる。例えば、イチローに関するニュースを高い頻度で閲覧するユーザはイチローに関するブログや動画などに興味を示す可能性は高い。

携帯端末や PDA 等のモバイル情報端末上で、ニュース記事やブログを推薦する場合、現状のシステムではタイトルを並べて表示することで情報推薦を実現するが、画面の大きさや文字サイズの問題等から、限られた数のアイテムしか掲載できない問題がある。一方で、図 2, 3 のように、最初の画面でユーザが興味を持つと思われるキーワード一覧を画面に表示し、キーワード選択後に検索システムを選択するシステムならば、狭い画面上においても効率の良い情報推薦が可能になる。

また、従来の推薦システムは「スポーツ」や「野球」といったジャンルを用いた推薦システムが多かったが、キーワードレベルでユーザの興味を抽出することで、より粒度の細かいプロフィールが作成可能になる。例えば、野球には興味はないがプロ野球の特定の選手だけに興味があるユーザの興味をキーワードレベルで抽出することで、これらのユーザプロフィールの表現が可能になる。ユーザの興味というものは、ジャンルで分類可能な粒度以上に細かい性質を持つと予想されるため、より細かい粒度でプロフィールを作成することは重要な問題だと考える。また、キーワードレベルでユーザの興味を取得可能になれば、Web 履歴からのマーケティング調査や、ユーザの興味の可視化等の様々な応用技術が実現可能になる。

2. 関連研究

本章では、ユーザの Web 閲覧履歴や閲覧ページからキーワード抽出を行っている関連研究について述べる。土方ら [14] は、ユーザのマウスの挙動から有益なキーワードを抽出する手法を提案した。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のキーワードは重要であることが報告されている。松雄ら [11] は、ユーザの閲覧 Web 履歴から、ユーザにとって身近な語を抽出し、現在閲覧しているページにハイライト表示することで、内容を容易に理解しやすいブラウジングシステムを提案した。キーワード抽出には、ユーザの閲覧した Web 文書中における共起頻度に偏りがあるキーワードを抽出している。戸田ら [9] は、Web 文書から固有表現 (人名、組織名、地名) を抽出し、その Web 文書の左側に固有表現のクラス別にキーワードを表示し、そのキーワードの検索が可能になるラベル指向ナビゲーションを実現した。

従来のキーワード抽出技術を用いた情報推薦技術は、主にキーワードの単位を形態素や IREX で定義された固有表現を抽出していた。しかしながら、形態素単位では語長が短すぎてキーワードの意味が掴みにくい問題があり、また IREX で定義された固有表現では固有表現の定義が狭いことにより、製品名やイベント名といった抽出できないキーワードがある問題があった。関根が提案している関根の拡張固有表現を用いることも考えられるが、質疑応答システム等の情報抽出システムでの利用を考えたものであることや、またこれら进行分类するための訓練データやソフトウェア等の整備が不十分であるため使用は難しい。

3. Wikipedia の見出し語

本稿では、キーワード候補をオンライン百科事典である Wikipedia^[注1] の見出し語を用いる。Wikipedia の見出し語を用いることで形態素や固有表現といった従来法では抽出できなかった幅の広いジャンルのキーワードが抽出可能になる。さら

(注1): 本稿では 2007 年 11 月 21 日のデータを使用した。
http://download.wikimedia.org/jawiki/

表 1 各手法における抽出可能なキーワード例

キーワード	形態素	固有表現	Wikipedia
イチロー			
中村俊輔	x		
自然言語処理	x	x	
ゴルフ		x	
フィギュアスケート	x	x	
ビリーズブートキャンプ			
八木カミ王子	x	x	
おっぱっぱー	x	x	

に、Wikipedia の見出し語は詳細な説明が可能で、ユーザの嗜好対象になりやすいキーワードが多いため、嗜好を表すと言い難い一般名詞等のフィルタリングが可能になる。表 1 に嗜好を表すキーワードを例に挙げ、形態素と固有表現、Wikipedia の見出し語の内、それぞれのキーワード定義で抽出可能であるかを調べてみる。一般的に形態素は「中村俊輔」や「八木カミ王子」、「フィギュアスケート」といった複合名詞が抽出できない。また、名詞の連結規則を用いた場合においても、すべての複合名詞を抽出してしまうために、嗜好を表すキーワードのみを抽出するといった目的には適さない。IREX で定義された固有表現を用いた場合、人名や地名、組織名といった嗜好を表すと思われる固有名詞を抽出することは可能だが、「自然言語処理」や「ゴルフ」、「おっぱっぱー」といった特定の嗜好を表すキーワードが抽出できない。それに対して、Wikipedia の見出し語はこれらのキーワードを網羅し、流行的なキーワードに対してはユーザが次々と記述することから、嗜好を表すキーワードを保有しているといえ、さらに Wikipedia の見出し語は 2007 年 11 月現在で約 70 万語とキーワードとして十分な量があるため、嗜好を表すキーワード候補として実用的であるといえる。

また、goo サイト^(注2)上の検索窓から入力された検索クエリ^(注3)の上位 1 万件中(異なり数)に Wikipedia の見出し語がどれだけ含まれているかを調べた結果、全クエリ中約 66 % ものクエリが Wikipedia の見出し語と同一のキーワードであることが確認できた^(注4)。さらに、全検索クエリ(延べ数)における一致数を調べた結果、約 24 % が Wikipedia の見出し語と完全一致した。以上の理由からも、Wikipedia の見出し語をクエリ(キーワード)候補として用いることは適切であるといえる。

4. キーワード固有重要度

本稿では、ユーザ毎に変化しないキーワード固有重要度と、ユーザ毎に変化するユーザ毎のキーワード重要度の二つの視点から、最終的なユーザ毎のキーワードランキングを算出する。ユーザ毎に変化しないキーワード固有重要度を算出する理由として、同じ野球選手でも人気のある選手とそうでない選手がいるように、キーワード毎に重要度が異なることが挙げられる。

(注2): <http://www.goo.ne.jp/>

(注3): 2007 年 4 月の 1ヶ月間における goo サイト内の検索クエリログ

(注4): スペースで複数クエリとして分かれている場合は(例: 中村俊輔 ニュース, 中村 俊輔), スペース毎に分割し比較対象とした。語の揺らぎやノイズ等の変換や削除は一切行っていない。

従来手法においては、キーワードの出現頻度で重要度を算出していたが、ユーザの Web 閲覧履歴中に、人気選手 A の名前が 5 回出現し、あまり人気のない選手 B の名前が 7 回出現したという場合、そのユーザは選手 B に興味があるとは判断しにくい。人気や知名度を考慮すると、人気選手 A の方が興味のある可能性が高い。そこで、キーワード固有の重要度を算出することで、この問題の解決を目指す。本節では、HITS [4] に基づく Wikipedia ランキングアルゴリズムを用いて算出したキーワード重要度と検索エンジンの HIT 数を用いた WebIDF の算出方法について述べる。

4.1 HITS に基づく Wikipedia ランキング

主な Web ページのランキング手法として、Kleinberg が提案した HITS と PageRank が提案した PageRank [8] がある。これらのアルゴリズムは、Web ページ間のリンク構造から Web ページの重要度を算出する。本稿では、Wikipedia のリンク構造に改良した HITS アルゴリズムを適用することで、Wikipedia のページ内におけるランキングを行い、ページのランキングを見出し語の重要度として用いる。

HITS アルゴリズムは、すべての Web ページを authority (権威のあるページ) と hub (リンク集) の 2 つから構成されると定義し、良い hub から多数リンクされているページ程、良い authority であるという仮説と、良い hub は多数の authority へのリンクを持っているという仮説を再帰的に繰り返すことにより、Web ページをランキングする。HITS アルゴリズムは以下の式で定義される。

$$a(p) = \sum_{p' \subseteq P'} h(p') \quad (1)$$

$$h(p) = \sum_{p' \subseteq P'} a(p') \quad (2)$$

ここで 1 番目の式の $a(p)$ においてはページ p' からページ p にリンクが張られているものとし、2 番目の式の $h(p)$ においては、ページ p からページ p' にリンクが張られているものとする。そして、これらの結果に基づき Web ページのランキングを行う。優秀な hub と authority は相互に強化すると Kleinberg は主張している。PageRank や HITS アルゴリズムは Web ページのリンク構造をモデルにしたアルゴリズムのため、リンク構造が密な Wikipedia に適用した場合、やや難がある。そこで本稿では、リンク構造が密な中においても、優秀な hub と優秀な authority を抽出することが、正しいランキングに繋がるという仮説に基づく、Wikipedia の特徴を活かしたアルゴリズムを提案する。キーワード固有重要度を算出する主な流れは、改良した HITS アルゴリズムを Wikipedia に適用し、それぞれのページの authority 値から算出した順位を exponential loss 関数に当てはめることによりキーワード固有重要度を算出する。ページ中のテキスト量

Wikipedia の見出し語は、知名度が高くかつ内容の豊富なキーワードほど記述量が多い傾向がある。そこで、authority 値の算出の際に、自ページにテキスト量が多ければ多いほどそ

のページは重要であるといった重みを考慮する．テキスト量はリンク文字列を含まないテキスト部分の文字数から算出し，最後に全体をラプラススムージングをかけ正規化し， $text(p)$ とする．

自リンクと被リンクの比率

一般的に Wikipedia の見出し語は有名なキーワード程，自リンクと被リンクが多くなっている．しかしながら，地名やジャンル名などは多くのキーワードからリンクされやすく，このような広い概念を持つキーワードは自リンクの数に比べて，被リンクの数が圧倒的に多い傾向がある．HITS アルゴリズムは良い hub から多数リンクされている authority は良い authority であるといった仮説に基づくが，圧倒的に被リンクが多いとこれらの仮説は成り立たないと予想される．また，その一方で，最近知名度の高くなってきている流行語や有名人等の見出し語は，被リンクは少ないが，自リンクが多い傾向がある．そのため少ないリンクにおいても authority を高める必要がある．図 4～図 7 に上記に挙げたノード例を示す．

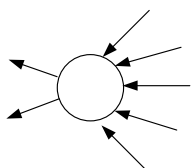


図 4 広い概念を持つキーワード

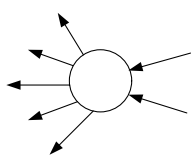


図 5 急成長・注目キーワード

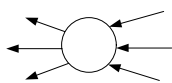


図 6 一般的なキーワード

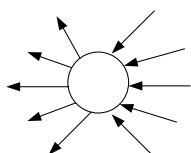


図 7 重要なキーワード

本稿では自リンクと被リンクの比率を authority の算出の際に考慮することで，この問題の解決を目指す．自リンクと被リンクの比率を用いた重みは以下のように定義する．

$$Link_Ratio(p) = \frac{\log(flink(p) + 1)}{\log(blink(p) + 1)} \quad (3)$$

ここで， $fblink(p)$ はページ p の自リンクを表し $blink(p)$ はページ p の被リンクを表す．

明らかに authority とならない見出し語の扱い

Wikipedia の見出し語には，明らかに authority とならない見出し語が存在する．例えば「～年」や「～年の～」，「～月～日」，「～一覧」，「～の歴史」，等がそれに該当する．これらの見出し語は一般的に自リンクを多数持ち，さらに被リンクが圧倒的に多い場合もあるため，ノイズとなる可能性が高い．そこで本稿では，先ほど挙げた明らかに authority とならないキーワードの authority を常に変更しない（初期値にする）ルールを追加することで，この問題に対処する．

hub の平均的なリンクの質

Wikipedia のページには，自リンクが多数あるが hub としての質の悪いページがある．そこで，リンク先ページの authority が平均的に高い hub は重要であるといった指標を考慮することで，自リンクは多いが hub として質の低いページの hub 値を下げる関数を用意する．

$$Link_Quariry(p) = \frac{\sum_{p'} \log(a(p') + 1)}{count(p)} \quad (4)$$

ここで $count(p)$ はページ p の持つ自リンクの総数を表す．

リダイレクトの扱い

Wikipedia には見出し語の異表記を解消するために redirect（転送リンク）が存在する．例えば図 8 のように「イチロー」の見出し語には「鈴木一郎」と「ICHIRO」の redirect がある．redirect は異表記のキーワードを一意にまとめる効果だけでなく，キーワードの被リンクの数に大きな影響を持つため，図 9 のように redirect キーワードを親ノードにまとめることで，異表記キーワードの重要度を算出し，被リンク数の問題を解決する．

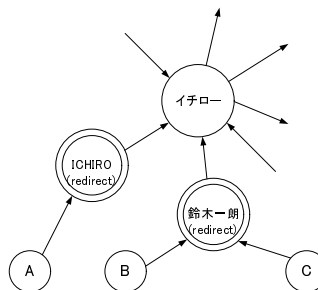


図 8 リダイレクトの例

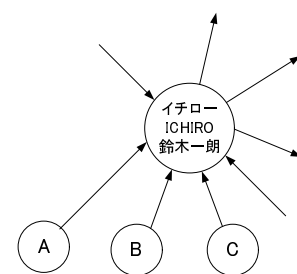


図 9 改良後の例

そして，最終的な改良 HITS アルゴリズムは以下の式で定義される．

1. For all $p' \subseteq P'$ pointing to p ,

$$a(p) = \frac{\log(flink(p) + 1)}{\log(blink(p) + 1)} \cdot text(p) \cdot \sum_{p'} h(p') \quad (5)$$

2. For all $p' \subseteq P'$ pointing to by p ,

$$h(p) = \frac{\sum_{p'} \log(a(p') + 1)}{count(p)} \cdot \sum_{p'} a(p') \quad (6)$$

次に上記の式で算出した authority の値を昇順にソートし，式 7 の exponential loss 関数に当てはめることによりキーワード重要度を算出する．

$$Link_Score(k) = \exp\left(\frac{\log(R + 1) \cdot (total(K) - rank(k) + 1)^a}{(total(K))^a}\right) - 1 \quad (7)$$

ここで， k はキーワードを表し， $total(K)$ はキーワードの総数， $rank(k)$ はキーワード k の順位， a は勾配係数とし， R はキーワード順位が 1 位の時のリンクスコアの値とする．この

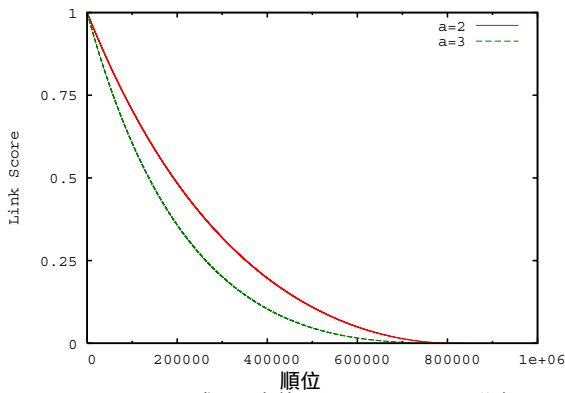


図 10 式7で定義した Link Score の分布

関数を用いた最終的なキーワード重要度の分布は図 10 の様になる。図 10 で示した様に、 a の値が大きくなるにつれて、関数の勾配は急になる。評価実験では、 $a=3$ 、 $R=1$ の値を用いる。この関数に当てはめる理由は、アルゴリズムを繰り返し実行するにつれて authority 値から算出したキーワードの順位は収束するが authority 値は収束しないことと、約 70 万語ある Wikipedia のキーワードの中でも重要なキーワードであると思われる語が約 20 万語位までであることが挙げられる。

4.2 検索エンジンの HIT 数を用いた WebIDF

Web 上に多数存在するキーワードは一般的に特徴的でないキーワードの可能性が高い。また、Web 上にはあまり出現していても、ユーザの Web 閲覧履歴に多数出現するキーワードは、そのユーザにとって重要なキーワードである可能性が高い。そこで、キーワードの Web 検索エンジンにおけるヒット数を DF とし、Web の IDF を算出する。Web 検索エンジンには Yahoo!デベロッパネットワーク^(注5)を用いた。そして本稿における Web の IDF は以下の式で定義される。

$$WebIDF(k) = \log_2 \left(\frac{N}{n_k + 1} \right) \quad (8)$$

if $N < (n_k + 1)$ then $WebIDF(k) = 0$

ここで、 n_k はキーワード k のヒット数である。 N は不要なキーワードの閾値的な値を考慮して 2160000 を用いた。そして、これらの値にラプラススムージングを適用し、最終的な WebIDF とした。

5. ユーザ毎のキーワード重要度

本節では、ユーザの閲覧履歴からユーザ毎のキーワード重要度の算出手法について述べる。

5.1 ユーザ履歴を用いたキーワード重要度

ユーザの閲覧履歴を取得し、それらを用いてユーザが興味を持っていると思われるキーワードを抽出する。具体的には、ユーザの閲覧した Web ページのテキストに含まれるキーワードを抽出し、さらに Web ページの閲覧時間から、ユーザのそのページに対する重要度を得る。すなわち、ユーザが長く閲覧したページ程、重要なページであると考えられる。ここで、閲覧した Web ページは同じページを複数回見ても 1 回だけ閲覧した

ものとし、閲覧時間はそれぞれの閲覧時間の和であるものとする。ページ p の総閲覧時間を $Time(p)$ と表す。本稿における Web 履歴取得方法は、ユーザ PC に履歴取得ソフトウェアをインストールする方法を用いた。

5.2 局所的なキーワード出現

一般的で特徴的でないキーワードは、閲覧した Web ページ集合中の多くのページに存在すると考えられる。そこで、それらのキーワードの重要度を下げ、局所的に多く出現するキーワードの重要度を上げるために大域的なキーワードの分布と局所的なキーワードの分布を考慮した重みを算出する。

$$Local_Weight(p_k) = \frac{1}{(count(p_k))^a} \quad (9)$$

ここで、 $count(p_k)$ はキーワード k が含まれるユーザの閲覧文書の数で、 a は勾配係数である。実験では $a=1.1$ を使用した。ユーザの閲覧 Web ページに広く分布しているキーワードの重要度は低くなり、局所的に多数出現しているキーワード程、重要度が高くなる。

6. 最終的なキーワード重要度

4 章で述べたキーワード固有の重要度と 5 章で述べたユーザ毎のキーワード重要度の双方を用いて最終的なユーザのキーワード重要度を算出する。最終的なキーワード重要度は以下の式で定義される。

$$Score(k) = \frac{1}{(count(p_k))^a} \cdot \sum_{p \subseteq P} \sum_{k_p \subseteq K_p} tf(k_p) \cdot WebIDF(k_p) \cdot Link_Score(k_p) \cdot Time(p) \quad (10)$$

ここで P はユーザの閲覧 Web ページ p の集合を表す。 K_p は p に含まれるキーワード k_p の集合である。

7. 評価

ユーザの Web 閲覧履歴からキーワードを抽出し、それらの評価を行う評価実験を行った。各ユーザには Web 閲覧を 10~20 分間してもらい、次に抽出したキーワードを評価する内容である。被験者として 6 人が評価実験に参加した。

7.1 評価方法

各被験者には、goo のニュースサイト^(注6)を起点とし Web 閲覧を 10~20 分程して頂いた。20 分以内に興味のある記事がなくなった場合にはそこで Web 閲覧を終了し、Web 閲覧が 20 分を経過した場合は Web 閲覧が 20 分経った事を告げ Web 閲覧を終了する。そしてその後、7.2 で挙げる各比較手法から上位 20 個のキーワードをそれぞれ抽出し、どの手法で抽出したキーワードかがわからぬようにランダムに表示し評価する。各手法間で複数個の重複するキーワードが抽出されていた場合は、1 つのキーワードとして扱った。評価作業は Web インターフェースで作成したものを使用し、Q1. このキーワード

(注5): <http://developer.yahoo.co.jp/>

(注6): <http://news.goo.ne.jp/>

表 2 各手法で用いた特徴量

手法	キーワード	時間	TF	文書 IDF	WebIDF	大域的な重み	HITS	改良 HITS
手法 1 (ベースライン)	形態素				x	x	x	x
手法 2 (ベースライン)	固有表現				x	x	x	x
手法 3	Wikipedia				x	x	x	x
手法 4	Wikipedia			x		x	x	x
手法 5	Wikipedia			x			x	x
手法 6	Wikipedia			x				x
手法 7 (提案手法)	Wikipedia			x			x	

表 3 各手法における検索したいキーワードの抽出精度 (上位 10 個)

手法	手法 1		手法 2		手法 3		手法 4		手法 5		手法 6		手法 7	
	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目
被験者 A	0.20	0.20	0.50	0.70	0.50	0.50	0.60	0.90	0.40	0.60	0.60	0.60	0.60	0.60
被験者 B	0.00	0.00	0.10	0.00	0.10	0.10	0.40	0.20	0.50	0.40	0.70	0.50	0.50	0.30
被験者 C	0.10	0.10	0.30	0.30	0.30	0.10	0.70	0.50	0.90	0.70	0.40	0.60	0.60	0.90
被験者 D	0.00	0.00	0.40	0.20	0.40	0.50	0.50	0.10	0.20	0.40	0.30	0.30	0.20	0.40
被験者 E	0.20	0.10	0.50	0.10	0.30	0.10	0.50	0.20	0.60	0.10	0.50	0.20	0.50	0.10
被験者 F	0.00	0.00	0.30	0.30	0.10	0.00	0.40	0.10	0.50	0.10	0.40	0.00	0.60	0.30
各回平均	0.08	0.07	0.35	0.27	0.28	0.22	0.52	0.33	0.52	0.38	0.48	0.37	0.50	0.43
総平均	0.075		0.308		0.25		0.425		0.45		0.425		0.467	

表 4 各手法で抽出されたキーワード中で知らなかったキーワードの割合

手法	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	手法 7
上位 5 位中で、知らなかったキーワードの割合	0.063	0.103	0.145	0.329	0.397	0.134	0.177
上位 10 位中で、知らなかったキーワードの割合	0.095	0.183	0.151	0.277	0.348	0.220	0.230
上位 20 位中で、知らなかったキーワードの割合	0.090	0.154	0.125	0.248	0.342	0.238	0.246

を検索クエリとして使用したいか? Q2. このキーワードを以前から知っていたか? の 2 個の質問を答える内容である。そして、Q1, Q2 共に 3 段階評価を行った (Q1. 検索したい, どちらでもない, 検索したくない. Q2. 知っていた, どちらでもない, 知らなかった.)。このキーワードで検索したいか? という評価については、被験者側からみればイメージし難くかつ評価し難い項目のため、「ブログで『イチロー』を検索する」といった感じの検索用アンカーリンクを複数個配置し、直接該当する検索エンジンを使えるようにした。また、各ユーザは実験当日のニュースを見ずに実験に参加して頂き、1 日間隔を空けて計 2 回実験を行った。

7.2 比較手法

本提案手法との比較には、キーワードに形態素の名詞と未知語のみを用いた手法 1 と、IREX で定義された固有表現の人名、地名、組織名、固有物名のみをキーワードとして用いた手法 2 をベースラインとして用意した。これら両手法には閲覧文書内のキーワードの TF と閲覧文書集合内における IDF 、さらに文書あたりの閲覧時間を特徴量として用いた。Wikipedia の見出し語をキーワードとして用いる手法は、形態素と固有表現の手法と同じ $TF \cdot IDF$ と閲覧時間を用いた手法 3、文書 IDF を $WebIDF$ と置き換えた手法 4、さらに局所的なキーワードの出現を考慮した手法 5、そして、Wikipedia に HITS アルゴリズムをそのまま適用した手法 6、最後に提案手法である改良した HITS アルゴリズムを適用した手法 7 を用意した。HITS を用いた手法は両手法とも式 7 の正規化を行う。Wikipedia の見

し語には、1~2 文字から成るノイズとなりやすい見出し語が存在するため、本稿では 3 文字以上の見出し語をキーワードとして使用した。また、本実験においては広告等のノイズとなる部分を除去する本文抽出等の加工は行っていない。なお形態素解析器には MeCab^(注7)、固有表現抽出器には CaboCha^(注8) を用いた。各手法で用いた特徴量を表 2 に示す。なお精度は、以下のように算出する。

$$\text{精度} = \frac{\text{上位 } n \text{ 個中の検索したいキーワード数}}{\text{上位 } n \text{ 個の出力キーワード数}}$$

評価対象となるキーワード数は、ユーザが実際に使う上での精度を考慮するために、上位 5 位、10 位、20 位とした。

8. 考察

評価実験の結果を表 3 に示し、各出力順位による精度の変化を図 11 に示す。評価の結果、Wikipedia の見出し語をキーワードとし、 $WebIDF$ 、大域的な重み、改良 HITS アルゴリズムを用いた提案手法である手法 7 が最も良い結果となった。2 番目に良い結果となったのは $WebIDF$ と大域的な重みのみを用いた手法 5 であった。

手法 5 が良かった原因を調べたところ、キーワード自体は知らないが、気になる文字列から構成されているために、どのような意味をもったキーワードであるかを調べたい検索要求が高

(注7): <http://mecab.sourceforge.net/>

(注8): <http://chasen.org/taku/software/cabocho/>

表 5 被験者 A の初日の実験結果 (手法 1~手法 4)

順位	手法 1	Q1	Q2	手法 2	Q1	Q2	手法 3	Q1	Q2	手法 4	Q1	Q2
1	goo	x		ニューハンプシャー州			シャープ			クリントン		
2	1月	x		クリントン			ニューハンプシャー州			ニューハンプシャー州		
3	液晶	x		シャープ			液晶パネル			WIRED	x	
4	州	x		民主党			ランキング	x		三八式歩兵銃	x	
5	氏	x		アジア	x		クリントン			米大統領		
6	選	x		ライフ	x		ビジネス	x		タイサイ	x	
7	シャープ			マケイン			民主党			規格争い		
8	オバマ			<政治家名>	x	x	大統領			史上最強打線		
9	東芝	x		エンタメ	x		トップ	x		VISION	x	
10	ニューハンプシャー			ハリウッド			スポーツ	x		バラク・オバマ		

表 6 被験者 A の初日の実験結果 (手法 5~手法 7)

順位	手法 5	Q1	Q2	手法 6	Q1	Q2	手法 7	Q1	Q2
1	液晶パネル			液晶パネル			液晶パネル		
2	ニューハンプシャー州			ニューハンプシャー州			ニューハンプシャー州		
3	システムLSI	x		シャープ			シャープ		
4	日立AD	x		システムLSI	x		<政治家名>	x	x
5	議員宿舎			議員宿舎			大規模集積回路	x	
6	シャープ			液晶テレビ	x		液晶テレビ	x	
7	<企業社長名>	x		ビル・クリントン			議員宿舎		
8	IPSアルファテクノロジー	x		<政治家名>	x		ビル・クリントン		
9	<政治家名>	x	x	上院議員			有機EL		
10	LGフィリップス			有機EL			ファーストレディー		

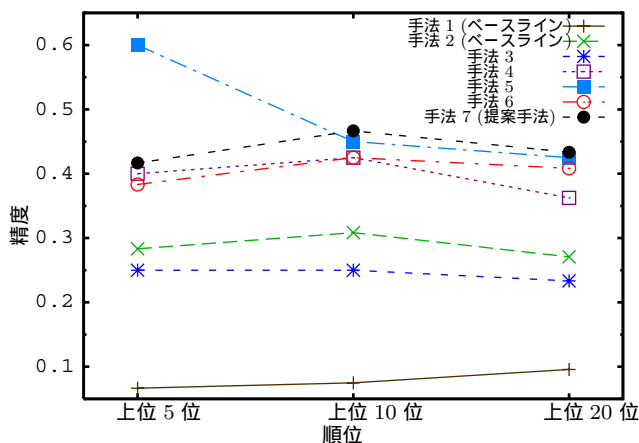


図 11 検索クエリとして使用したいキーワードの抽出精度

かったことが挙げられる。そこで、各手法による知らなかったキーワードの割合を表 4 に掲載する。WebIDF や局所的に出現しているキーワードの重要度を高くする大域的な重みを用いることで、出現頻度が非常に稀だと思われるキーワードが数多く抽出されていることがわかる。これらの重みは、一般的で特徴的でないキーワードの重要度を下げ一方、このような弊害を持つことがわかる。反対に HITS と改良 HITS アルゴリズムを用いた手法 6 と手法 7 は、WebIDF と大域的な重みを使用しても知らない割合は低く抑えられている。これは、Wikipedia を HITS アルゴリズムを用いてランキングした結果、知名度の高いキーワードほど重要度が高くなり、知名度の低いキーワードほど重要度が低くなるのが起因する。

次に、どのような意味を持つキーワードであるかを知るための検索要求ではなく、ユーザがキーワードを見る以前から純粋に興味を持っていた検索キーワードを評価するために、キーワードを知っていて、かつ、検索キーワードとして使用したいかを評価した。その結果を表 8 に掲載する。その結果、すべての実験回においても、最も精度の高い手法が手法 7 の提案手法となった。また、図 12 に示すように、すべての上位順位で手法 7 が最も良い結果を得ることが確認された。

TF・IDF と時間を用い、キーワード定義を変化させた手法 1~3 の実験結果は、固有表現、Wikipedia、形態素の順に良い結果となった。形態素は語長が短く、かつ意味的な最小単位に分割されてしまうためクエリとしては適さない。また、どのキーワードが重要であるかの絞込みが殆ど出来ないために、このような結果になったと思われる。固有表現をキーワードとして用いた手法は、人名、地名、組織名、固有物名といったキーワードの絞込みを行っているため、時間と TF・IDF のみを用いた手法においては精度の高い結果となったが、冒頭で述べたように、ユーザが興味をもつキーワードをすべて網羅しているわけではないため、今回の評価では評価できなかった再現率等の評価が悪いと予想される。Wikipedia のキーワードを用いた場合は、ユーザが興味をもつと思われるキーワードを幅広く網羅しており、さらに全体のキーワード集合が事前にわかっているため、あらかじめ WebIDF を計算できることや、リンク構造によって得られたキーワード固有重要度を用いることができるメリットがあるため、手法 4~7 の実験結果から分かるように、これらを用いた場合に有効である。今回の評価実験により得ら

表 7 知っているキーワードで、かつ検索したいキーワードの抽出精度（上位 10 個）

手法	手法 1		手法 2		手法 3		手法 4		手法 5		手法 6		手法 7	
	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目
被験者 A	0.20	0.20	0.50	0.70	0.50	0.50	0.60	0.90	0.40	0.60	0.60	0.60	0.60	0.60
被験者 B	0.00	0.00	0.10	0.00	0.10	0.00	0.30	0.10	0.40	0.10	0.60	0.40	0.50	0.20
被験者 C	0.10	0.10	0.20	0.30	0.30	0.10	0.60	0.40	0.40	0.20	0.40	0.20	0.60	0.50
被験者 D	0.00	0.00	0.40	0.20	0.40	0.50	0.50	0.10	0.20	0.30	0.30	0.30	0.20	0.40
被験者 E	0.20	0.10	0.40	0.10	0.30	0.10	0.20	0.20	0.30	0.10	0.30	0.20	0.30	0.10
被験者 F	0.00	0.00	0.30	0.20	0.00	0.00	0.10	0.00	0.10	0.00	0.10	0.00	0.30	0.00
各回平均	0.08	0.07	0.32	0.25	0.27	0.20	0.38	0.28	0.30	0.22	0.38	0.28	0.42	0.32
総平均	0.075		0.283		0.233		0.333		0.258		0.333		0.367	

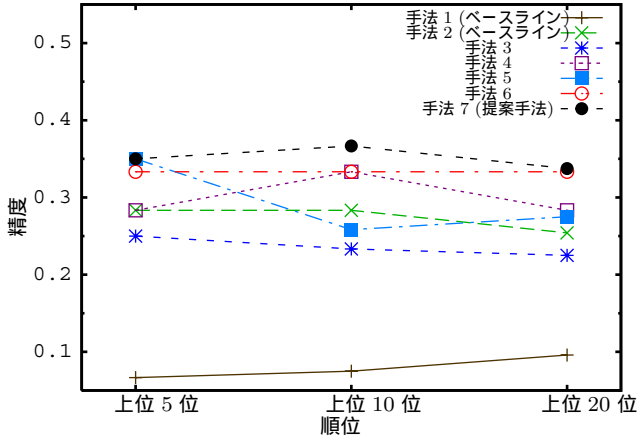


図 12 知っているキーワードで、かつ検索クエリとして使用したいキーワード抽出精度

れたその他の知見として、液晶パネルや有機 EL といった技術的なキーワードには興味があるが、製品である液晶テレビには興味が無いといったユーザや、特定のドラマには興味があるが出演者の一部には興味がないといった複雑な嗜好を持つユーザが多いことが挙げられる。キーワードレベルでユーザの興味を高い精度で抽出するためには、固有表現で定義されているクラスのように、キーワードクラスの細かい分類が必要になると思われる。ユーザの興味は我々が当初予想していた以上に複雑で粒度の小さいものであることが実験により確認された。

9. おわりに

本稿では、Wikipedia の見出し語をキーワードとし、Wikipedia の密なリンク構造に対応した改良 HITS アルゴリズムから算出したキーワード固有重要度を用いることで、ユーザの Web 閲覧履歴からのキーワード抽出において、従来手法と比べ高い精度のクエリ抽出を実現した。抽出したキーワードは、ニュースやブログ、動画検索等の様々な検索エンジンへの検索クエリとして使用できるだけでなく、RSS リーダのお気に入りキーワードのキーワード登録補助や、ニュースやブログを直接推薦する際に用いるユーザプロファイルとしての利用も可能だと考える。今後は、Wikipedia のリンクコミュニティの考慮や、改良 HITS アルゴリズムをパーソナライズ化するなどして、より精度の高いキーワード抽出の実現と、キーワード抽出技術を応用したサービスアプリケーションの検討を行いたい。

謝 辞

研究を進めるにあたり NTT サイバースリソリューション研究所、ならびに NTT サイバースペース研究所の皆様から様々なご意見・ご感想を頂きました。また、実験に参加して下さった所員の方々に感謝します。ありがとうございました。

文 献

- [1] J.Kleinberg. Authoritative sources in a hyperlinked environment. *In Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [2] S.Brin and L.Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107-117, 1998.
- [3] L.Li, Y.Shang, and W.Zhang. Improvement of hits-based algorithms on web documents. *In Proceedings 11th World Wide Web Conference (WWW '02)*, 2002.
- [4] M.Strube and S.P.Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. *In Proceedings 21st National Conference on Artificial Intelligence (AAAI '06)*, 2006.
- [5] M.Morita and Y.Shinoda. Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)*, 1994.
- [6] A.Das, M.Datar, and A.Garg. Google news personalization: Scalable online collaborative filtering. *In Proceedings 16th World Wide Web Conference (WWW '07)*, 2007.
- [7] C.D.Manning and H.Schutze. *Foundations of stastical natural language processing*. MIT PRESS, 1998.
- [8] C.D.Manning, P.Raghavan, and H.Schutze. *Introduction to Infomation Retirieval*. Cambridge University Press, 2008.
- [9] 土方嘉徳, 青木義則, 古井陽之助, 中島周. マウスの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*, Vol. 43, No. 2, pp. 566-576, 2002.
- [10] 松尾豊, 福田隼人, 石塚満. ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援. *人工知能学会論文誌*, Vol. 18, No. 4, pp. 203-211, 2003.
- [11] 戸田浩之, 中渡瀬秀一, 片岡良治. 特徴的な固有表現を用いたラベル指向ナビゲーションシステムの提案. *情報処理学会論文誌: データベース*, Vol. 46, No. SIG 13, pp. 40-52, 2005.
- [12] 山田寛康, 工藤拓, 松本裕治. Support vector machines を用いた日本語固有表現抽出. *情報処理学会論文誌*, Vol. 42, No. 6, 2002.
- [13] 竹本義美, 福島俊一, 山田洋志. 辞書及びパターンマッチングルールの増強と品質強化に基づく日本語固有表現抽出. *情報処理学会論文誌*, Vol. 42, No. 6, 2001.
- [14] 森田哲之, 倉恒子, 日高哲雄, 大浦啓一郎, 田中明通, 加藤泰久, 奥雅博. Memory-retriever: 体験獲得情報を想起させる行動検索手法. *情報処理学会論文誌*, Vol. 48, No. 3, pp. 1197-1208, 2007.