2P2-10

PC 上の Web 閲覧履歴からのクエリ抽出技術を用いたモバイル情報検索システム

Mobile Information Retrieval System using Query Extraction from Web Browsing History on PC

近藤 光正 Mitsumasa KONDO　森田 哲之 Tetsushi MORITA　田中 明通 Akimichi TANAKA　内山 匡 Tadasu UCHIYAMA

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
NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation

This paper proposes a mobile information system that uses the user's web history on PC without any manual effort. Similar to collaborative filtering methods, it can deal with specific domains. Query recommendation allows us to search various fields like blogs and moveis. The result of an experiment, our query recommendation method obtained significant improvements compare to baseline methods.

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 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嗜好に合った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探索する際に，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術である．しかしながら，Web が格段に進歩した現在においても，我々が必要な情報を見つける際の基本は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書籍だけではなく，今晩のテレビ番組や面白いブログ，動画，企業の株価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モバイル検索システムを提案する．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の API が公開されている．そのため，ユーザが興味を持つクエリを推薦することで，様々な検索システムと柔軟に連携できる．例えば，ハニカミ王子に関するニュースを高い頻度で閲覧するユーザは，ハニカミ王子に関するブログや動画等に興味を示す可能性が高い．推薦技術の従来手法である協調フィルタリングや類似度ベースの手法は，分野に閉じた推薦を行うものが主流であったが，クエリレベルの推薦を行うことで，分野に閉じない情報推薦システムの実現を目指す．

2. 本システムが目指す情報推薦

本稿で目指す情報推薦は，ユーザがあらかじめ特定の目的・意図を持つ情報検索のための推薦ではなく，モバイル特有の隙間時間，すなわち電車の移動・待ち時間や仕事の休憩等に楽しめるコンテンツを推薦するための情報推薦である．そのため，ユーザが時間を持て余している際に，モバイル端末を開くとユーザの嗜好に合った検索クエリが表示され，ユーザは文字入力をすることなく様々な検索システムから自分の嗜好に合った面白いコンテンツを探し出すことが可能なシステム[*1]を目指した．

近年，Yahoo!の oneSearch やモバイル goo から，複数の検索システムの結果を 1 画面の検索結果で返す検索技術が発表されている．これらの技術は，ユーザが入力したクエリの特性やユーザの好みに合わせた検索画面を提示するため将来有望な検索技術である．この検索技術は，ユーザがクエリを入力する前からは予想もしなかった検索結果，いわゆる”気づき”の情報を発見することができる．例えば，ハニカミ王子のニュース結果を期待して検索したユーザは，ハニカミ王子が出演しているテレビ番組の存在に”気づき”，ワンセグでテレビ番組を見るといった行動が”気づき”の情報である．しかしながら隙間時間上の検索において，ハニカミ王子に大変興味のあるユーザでも，検索クエリとしてハニカミ王子がすぐさま浮かび，さらにクエリとして入力するユーザは少ないと考える．モバイル端末上において文字入力という作業は比較的コストのかかる作業であることや，自分の興味を整理できているユーザは少ないからである．そこで，これらの複数の検索結果を 1 画面で提示する検索サービスと本稿で提案する PC 上の Web 閲覧履歴から抽出したクエリを推薦するシステムを組み合わせることで，ユーザはあらかじめ検索クエリを考えることなく，提示された検索クエリによって自分の興味に”気づき”，次に複数の検索結果から検索する以前からは予想もしなかった検索結果による”気づく”，2 重の”気づき”による情報検索を実現できると考える．これらの検索を実現するユーザの作業コストは，携帯端末から自分の興味キーワード一覧画面を開き，クエリを選択するだけである．

3. 隙間時間に関する調査

本研究を進めるにあたって，隙間時間に関するインターネット調査を行った．本調査では，「自分の思い通りにならずに空いてしまう時間で，つい無駄に過ごしてしまいがちなので，できれば有意義に使いたい時間」を隙間時間として定義した．調査対象は 10 代〜50 代以上の男女を調査対象とし，2148 件の有効回答が得られた．日常生活においてユーザがどのような状況を隙間時間と感じやすいかを調査した結果を図 1 に掲載す

連絡先: kondo.mitsumasa@lab.ntt.co.jp

*1 キーワード提示型の検索サービスとして，kizasi.jp や goo ランキングがある．これらのサービスは，個々のユーザ毎にキーワードの重要度を算出している訳ではなく，ユーザ全体のキーワード重要度（話題度）を算出している点が，本システムとは異なる．

る．調査方法は，隙間時間の簡単な説明を最初に行い，次にあらかじめ用意していた隙間時間が発生しそうな状況を複数個質問項目として挙げ，その状況下で隙間時間が発生するかを調査した．隙間時間の判定は，「隙間時間はない」から始まり，「5分未満」，「5〜10 分未満」･･･ の内，いずれか 1 つを選択する形式をとった．

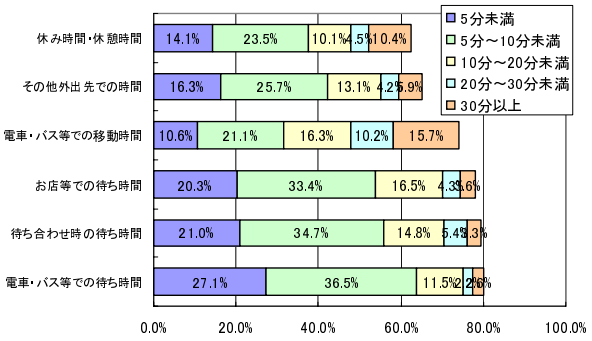

図 1: ユーザが隙間時間と感じやすい状況（上位 6 位）

ユーザが最も隙間時間と感じる状況は，電車・バス等での移動時間ではなく，電車・バス等での待ち時間であった．電車・バス等の移動時間は，まとまった時間が取れるため，読書やゲーム，メール等を行うことからこのような結果となったと予想される．5 分未満の隙間時間を考慮しない場合においては，電車・バス等の移動時間が 1 位になる．ユーザが隙間時間と感じる状況は，主に 20 分未満が全体の大半を占め，5〜10 分未満の隙間時間が最も多い．

次に，隙間時間と感じそうな状況を複数個回答して頂いた後に，1 日におけるトータルの隙間時間を回答して頂いた．そのユーザの割合が図 2 である．

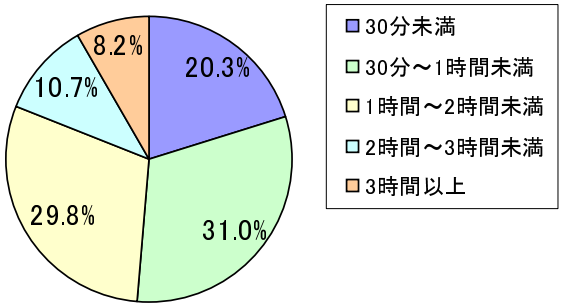

図 2: 1 日における隙間時間のユーザ割合

ユーザの隙間時間は，30 分〜1 時間未満が最も多く，次に 1 時間〜2 時間未満，30 分未満と続く．約 8 割のユーザは，毎日 2 時間未満の隙間時間を感じているようである．

隙間時間と呼ぶと，日常生活においてはあまり存在しないように思われがちであるが，半数近いユーザが 1 日 1 時間以上の隙間時間を感じており，無視できない存在であることがわかる．また，大半の隙間時間が外出中での状況で発生するため，モバイル端末における隙間時間の有効活用は重要な問題である．

4. Web 閲覧履歴からのクエリ抽出技術

4.1 クエリとして適したキーワードに関する考察

本稿では Wikipedia の見出し語をクエリ候補として用いる．日本語 Wikipedia は見出し語が約 70 万語存在する非常に大規模なユーザ参加型のオンライン百科辞典である．Wikipedia は，話題の 5W1H となりやすい固有名詞や名詞等を網羅し，さらに誰もが自由に書き込める言語資源のため，日々誕生する話題語や新語を網羅している．単名詞だけではなく，複合名詞の切り出しが出来ている点もクエリ候補として非常に評価ができる点である．また，goo サイト[*2]上の検索窓から入力された検索クエリの上位 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[*3]，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[*4]．以上のことからも，Wikipedia の見出し語をクエリ候補として用いることは適切であるといえる．

4.2 クエリ抽出手法

本稿では Wikipedia のリンク構造からキーワード固有重要度を算出することで，Web 閲覧履歴のテキスト中から特徴的なキーワードを抽出する手法を提案する．リンク構造からキーワード固有重要度を算出する手法は，Web ページのランキング手法で多く用いられる Kleinberg が提案した HITS を改良したアルゴリズムを提案する．Wikipedia の各ページにはそれぞれ見出し語が存在するので，改良 HITS による Wikipedia のページのランキングがキーワード固有重要度になる．アルゴリズムの改良点として，Wikipedia の特徴なページ構造と密なリンク構造に対応させるように改良した．

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．

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，地名やジャンル名のような広い概念を持つキーワードは自リンク数に比べて，被リンク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，最近知名度が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．そのため少ない被リンクにおいても，authority 値を高める必要がある．これらの問題を解決するために，authority 値の算出の際に自リンクの被リンクの比率を考慮する．

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，「〜年」や「〜一覧」といった様に明らかに authority とならない見出し語が存在する．これらの見出し語は，自リンクが非常に多く，被リンクも非常に多いためノイズになりやすい．そこで，明らかに authority とならない見出し語の authority 値は常に変更しない（初期値にする）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

*2 http://www.goo.ne.jp/
*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
*4 スペースで複数クエリとして分かれている場合は，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．

とで，自リンクは多いが hub として質の低いページの hub 値を下げる重みを考慮する．

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 が存在する．例えば,「イチロー」には,「鈴木一朗」,「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はなく，キーワードの被リンク数に大きな影響を持つため，redirect キーワードを親ノードにまとめることで，異表記のキーワードの重要度を算出し被リンク数の問題も解決する．

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 $p' \in P'$ pointing to p,

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

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

$$h(p) = \frac{\sum_{p'} \log\left(a(p')+1\right)}{count(p)} \cdot \sum_{p'} a(p') \qquad (2)$$

次に上記の式で算出した authority の値を昇順にソートし，以下の 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．

$$Link\_Score(k) = \exp\left(\frac{\log(R+1)\cdot(total(K)-rank(k)+1)^a}{(total(K))^a}\right) - 1 \qquad (3)$$

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．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．

最終的なキーワード重要度

最終的なキーワード抽出手法は，以下の式で表される．

$$Score(k) = \frac{1}{(count(p_k))^a} \cdot \sum_{p \in P} \sum_{k_p \in K_p} tf(k_p) \cdot WebIDF(k_p) \cdot Link\_Score(k_p) \cdot Time(p) \qquad (4)$$

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．$count(p_k)$ 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．WebIDF は Wikipedia の見出し語を検索エンジン投入し，その結果得られた検索 HIT 数から算出した IDF 値である．

5. 提案システム

本稿で提案する個人適応型モバイル検索システムを図 3 に掲載する．PC の Web 閲覧履歴の取得は，ユーザの PC に履歴取得ソフトウェアをインストールすることによって実現する．ユーザの Web 閲覧履歴は履歴蓄積・加工サーバに自動送信され，そこでクエリレベルのキーワードが抽出される．そして，ユーザはモバイル端末からプロファイル提供サーバにアクセスし，自分の旬キーワード一覧を得る．旬キーワード一覧は，1

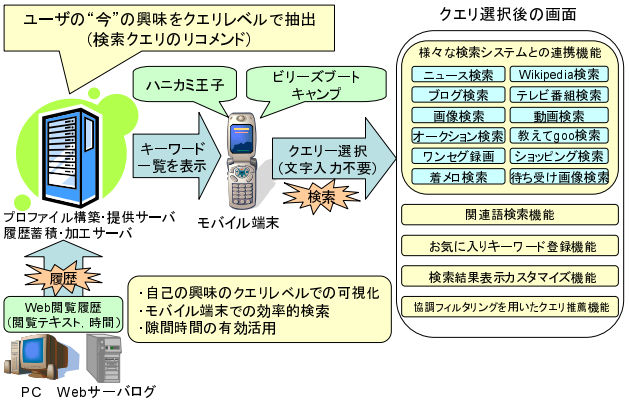

図 3: PC の履歴を用いた個人適応型モバイル検索システム

日間隔や 6 時間間隔といった様に一定の時間間隔に分けて表示される．このように表示することで，様々な時間帯における自分の旬ランキングを作成することができ，自分の Web 閲覧における興味キーワードがクエリレベルで可視化できる．昨日見たニュースの話題で今どうなっているのかを調べたい場合や，そのキーワードのより詳しい情報を調べたい場合に活用できる．そしてユーザがクエリを選択後に，次にどの検索システムで検索するかのリンク一覧を表示し，さらにニュースやテレビ番組等の高い頻度でユーザが興味を持ちそうな検索結果も同時に表示する．ユーザによっては欲しい情報が異なると予想されるため，検索画面はユーザがカスタマイズ可能な仕様にした方がよいだろう．また検索画面では，ユーザが選択したクエリの関連語も表示する．例えば，ハニカミ王子の関連語としては，ゴルフやぽっちゃり王子等が提示される．関連語を選択すると，その関連語の検索画面が提示され，さらにその関連語の関連語が表示される仕組みになっている．ユーザは興味のあるキーワードの検索結果が多少ずれていたとしても，そのキーワードの関連語を検索したり，ニュース本文から抽出されたキーワード等のコンテンツ先で提示されている検索クエリを選択することで，当初の検索意図とは異なる他の興味情報に文字を入力することなく辿り着くことが出来る．お気に入りキーワード登録機能は，自分の旬ランキングでこれからも気になりそうなキーワードを保存する機能である．RSS リーダ等でのお気に入りキーワード登録と異なり，システムが推薦した中からキーワードを選んで登録するといった文字入力なしの半自動キーワード登録が可能になる．

画面の狭いモバイル端末においては，ニュース記事のタイトルの様な文や動画のサムネイル等を複数個同時に表示した場合，視認性の低い画面表示になってしまうが，本提案システムの様に，興味クエリの一覧を表示し，クエリ選択後に検索エンジンを選ぶという工程ならば，視認性が高く効率の良い画面表示と情報検索が可能になると考える．

6. クエリ抽出技術の評価

ユーザの Web 閲覧履歴から検索クエリを抽出し，評価を行った．被験者の総数は 6 人である．

6.1 評価方法

各被験者は，goo のニュースサイトを起点とし，Web 閲覧を 10〜20 分程行う．そして Web 閲覧終了後に，システムが抽出したクエリ (キーワード) を評価する内容である．実験は 2 日に分けて計 2 回の評価実験を行う．各日程において，過去の履歴は引き継がないものとする．評価内容は，以下の 2 つ

表 1: クエリ抽出精度（上位 10 位）

評価項目	手法 1	手法 2	手法 3	提案手法
1. 検索したいキーワードの抽出精度 (Q1)	0.075	0.308	0.250	**0.467**
2. 以前から知っていて，かつ検索したいキーワードの抽出精度 (Q1⋂Q2)	0.075	0.283	0.233	**0.367**

の質問から構成され，それぞれ 3 段階の評価を行う．

- Q1. このキーワードを検索クエリとして使用したいか？
  - 使用したい．
  - どちらとも言えない．
  - 使用したくない．
- Q2. このキーワードは以前から知っていたか？
  - 知っていた．
  - どちらとも言えない．
  - 知らなかった．

そして Q1，Q2 の質問結果から，2 通りの評価を行う．Q1 の質問結果から評価をする「1. 検索したいキーワードの抽出精度」と，Q1 と Q2 の質問結果から評価を行う「2. 以前から知っていて，かつ検索したいキーワードの抽出精度」の 2 通りの評価である．評価項目 1 はユーザが純粋に検索したいキーワードの抽出精度で，評価項目 2 はユーザの興味クエリを抽出できたかの精度である．評価項目を分けた理由を補足すると，検索したいか否かの評価項目 1 は，自分が知らないキーワードだけどキーワードの詳しい内容が知りたい場合にも正解となり，ユーザプロファイルとしての興味クエリを抽出できていないため評価項目を分けた．

ベースラインとして，形態素の名詞をキーワード候補とし，TF・IDF，時間を用いてキーワードを抽出する手法 (手法 1)，ＩＲＥＸで定義された固有表現（人名，地名，組織名，固有物名）をキーワード候補とし，手法 1 と同じく，TF・IDF，時間を用いてキーワードを抽出する手法 (手法 2)，Wikipedia の見出し語をキーワード候補とし，手法 1 と同じく，TF・IDF，時間を用いてキーワードを抽出する手法 (手法 3)，以上 3 つの手法をベースラインとして用いた．なお，形態素解析器には MeCab，固有表現抽出器には CaboCha を用いた．クエリ抽出の精度は以下の式から算出し，抽出されたキーワードの上位 10 位までを評価対象とした．

$$\text{精度} = \frac{\text{正解キーワード数}}{\text{システム出力のキーワード数}}$$

6.2 評価結果

評価実験の結果を表 1 に示す．評価の結果，評価項目 1，2 共に提案手法が最も良い結果となった．TF・IDF と時間を用いて，キーワード定義を変化させて比較した手法 1〜3 の評価結果は，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り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候補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，TF・IDF と時間のみを用いた手法において，最も精度の高い結果となったが，IREX の固有表現はユーザが興味を持つキーワードをすべて網羅しているわけではないため，今回評価できなかった再現率が低いと予想される．Wikipedia の見出し語を用いた場合は，ユーザが興味を持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既知のため，あらかじめ WebIDF を計算出来ることや，リンク構造から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，これらを用いた場合に有効である．

7. まとめ

本稿では，PC 上の Web 閲覧履歴からのクエリ抽出技術を用いて，モバイル特有の隙間時間に着目したモバイル情報検索システムを提案した．提案したクエリ抽出技術は，TF・IDF と時間を用いた従来法と比較して高い精度で抽出できることが確認できた．今後は，よりリッチで直感的なユーザインターフェースの構築・評価と，クエリ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．

参考文献

[1] A.Broder, M.Fontoura, V.Josifovski, and L.Riedel. A semantic approach to contextual advertising. *Proceedings of the 30th annual international ACM SIGIR Conference(SIGIR '07)*, 2007.

[2] A.Das, M.Datar, and A.Garg. Google news personalization: Scalable online collaborative filtering. *In Proceedings 16th World Wide Web Conference (WWW '07)*, 2007.

[3] G.Adomavicius and A.Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, Vol. Vol 17, No. No.6, June 2005.

[4] J.S.Breese, D.Heckerman, and C.Kadie. Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI '98)*, 1998.

[5] J.Teevan, S.T.Dumais, and E.Horvitz. Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th annual international ACM SIGIR Conference(SIGIR '05)*, 2005.

[6] Z.Dou, R.Song, and J.Wen. A large-scale evaluation adn analysis of personalized search strategies. *In Proceedings 16th World Wide Web Conference (WWW '07)*, 2007.

[7] 近藤光正, 森田哲之, 田中明通, 内山匡. 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. 電子情報通信学会第 19 回データ工学ワークショップ論文集, 2008.

[8] 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.