社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (19)
2008－ＦＩ－ 91 (19)
2008／6／20

Wikipedia を用いた Web 閲覧履歴からのキーワードプロファイル抽出とその応用

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

Keyword Profile Extraction Using Wikipedia from Web Browsing History, and its Applications

Mitsumasa KONDO, Tetsushi MORITA, Akimichi TANAKA, Tadasu UCHIYAMA
NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation

1. はじめに

Web の世界が進歩するにつれて，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，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 に不慣れなユーザだけでなく，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なく，今晩のテレビ番組や面白いブログ，動画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直接的に推薦するのでなく，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が可能である．従来の推薦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，検索クエリレベルのキーワードを推薦する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. キーワードプロファイル

本稿では，検索クエリとして利用可能なユーザの興味キーワードの集合をキーワードプロファイルと定義する．本稿で考える検索クエリとして利用可能なキーワードとは，キーワードから連想されるユーザの検索意図と検索システム側の結果が，大きく異ならないキーワードである．ここにおける検索システムのキーワード検索の精度は，入力されたキーワードを人間が判定しても解釈の近いものであるものとする．別の表現で表すと，キーワードの実体を一意に表すことのできるキーワードともいえるだろう．このようなキーワードの例を挙げると，固有表現と呼ばれる，人名（例：中村俊輔，ハニカミ王子），地名（例：アメリカ合衆国，横須賀市），組織名（例：日本電信電話，NTT ドコモ）であったり，事件や話題を一意に表す話題語であったりする．そのため，ジャンルや年代を問わずキーワードを体系的に網羅している百科事典をベースにキーワードを抽出することが理に適っていると考え，本稿ではオンライン百科事典である Wikipedia の見出し語をキーワード候補として使用する．キーワードプロファイルを用いたアプリケーションの一例を図１に掲載する．

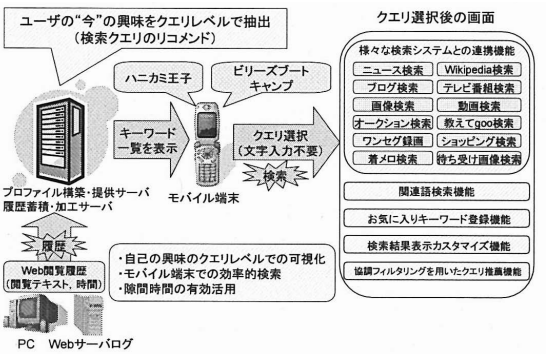

図１：キーワードプロファイルを用いたアプリケーション例

3. 提案手法

本手法は，ユーザの Web 閲覧履歴テキスト中に現れたキーワード候補から，ユーザが興味をもつと思われる順にキーワードをランキングする．提案手法では，語の出現頻度や出現分布，Web ページの閲覧時間だけでなく，キーワード本来の人気度、知名度といった重要度を考慮したキーワード固有重要度を用いることで，テキスト中に含まれるキーワードの中でも，ユーザにとってより重要なキーワードを上位に位置づける手法を提案する．キーワード固有重要度は，Wikipedia の特徴的な構造とリンク構造を考慮した解析結果から算出する．評価実験の結果，tf・idf と閲覧時間を用いたキーワードランキング手法と比べ，本提案手法が優れていることが確認された．

参考文献

1) 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，第 19 回データ工学ワークショップ論文集，2008.