3P-6

Web 閲覧履歴を利用したうろ覚え Web ページ捜索システム

古山　直樹†　　石川　傑也†　　小尾　哲也†　　筧　捷彦‡

早稲田大学大学院 基幹理工学研究科†　　早稲田大学 理工学術院‡

1.　はじめに

過去に閲覧した Web ページは，ブックマークや検索エンジンや閲覧履歴を利用することによって見つけ出すことができる．しかし，ブックマークをしておらず，検索ワードもわからない，ブラウザの履歴検索機能を使っても，候補が多すぎて捜しきれないということがある．

このような状況を解決するために，Web 閲覧時に独自に閲覧情報を保存しておき，それを利用することによって Web ページの捜索を効率的に行えるようなシステムを制作した．

2.　一般的な Web ページ捜索方法

一度見た Web ページを再度閲覧するための主な方法には，次の 3 つが挙げられる．

(1)ブックマークから捜す

よく見る Web ページや，そのユーザにとって重要な Web ページはブックマークされることが多い．ブックマークがされている Web ページは容易に捜すことができる．

(2)検索エンジンでキーワード検索を行う

過去に検索した時と同じキーワードで検索し，Web ページを捜索する．捜している Web ページが検索結果の下位にあったり，検索キーワードを覚えていなかったりした場合は捜すことが困難である．

(3) Web ブラウザの履歴検索機能を使う

Web ブラウザによって違いはあるが，Web ページのタイトル・URL・ページ内の文章での検索や，閲覧日時・タイトル・表示回数でのソートなどの機能が備わっている．

The vaguely-memorized web page search system using a web browsing history

† Naoki FURUYAMA, Takuya ISHIKAWA, Tetsuya OBI, Graduate School of Fundamental Science and Engineering, Waseda University

‡ Katsuhiko KAKEHI, Faculty of Science and Engineering, Waseda Universitv.

これらの特徴を考慮して，捜したい Web ページに対して，以下の 3 つの条件を満たすときを「うろ覚え状態」と定義し，うろ覚え状態のときでも効率的に Web ページを捜索できるようなシステムを制作した．

・ブックマークをしていない
・検索キーワードが分からない
・履歴検索機能を使用しても，候補となる Web ページが多すぎて捜しきれない

3.　関連研究

井倉らの研究[1]では，「閲覧時期」「経路」「閲覧目的」「メディア情報」の 4 つを指標として，対話的質疑による捜索方法を提案した．システムはユーザに対して，各指標を基にした質問を行い，その回答内容から候補となる Web ページを提示する．ユーザはうろ覚え状態であるため，回答の選択肢は「はい」「いいえ」「分からない」といった大まかなものに限定している．

本研究では，うろ覚え状態のユーザが覚えている Web ページの情報は，ユーザや状況によって異なり，そのときによって有効な情報が変わるのではないかと考えた．そこで，ユーザが条件を自由に指定できる，検索型のシステムを作成することにした．

4.　提案手法

候補となる Web ページを絞り込むための情報には，従来のブラウザの履歴機能が保存している，Web ページの閲覧日時，閲覧回数，ページのタイトル，URL を利用する．さらに，2 章で述べた検索手法にはないものとして，Web ページに辿り着くまでの経路と Web ページの属するカテゴリを利用することにした．

捜している Web ページ自体は覚えていなくても，ポータルサイトや百科事典サイトから辿り着いたといった，Web ページに辿り着くまでの経路は覚えている可能性がある．そこで，閲覧時に Web ページの遷移情報を保存し，絞り込みに利用した．

固有名詞のキーワードは，一般的な単語のキーワードに比べて思い出し難く，キーワード検索が困難だと考えられる．そのような Web ページを捜索する手法として，Web ページの属するカテゴリを利用することにした．あらかじめ Web ページの属するカテゴリを覚えておくことで，キーワードを覚えていなくてもカテゴリで検索を行うことができる．

そのために，閲覧した Web ページに対してカテゴリを自動的に推定する仕組みが必要になる．Web ページからキーワードを抽出し，その結果と階層型カテゴリを比較することによって Web ページの属するカテゴリを推定する．

カテゴリ推定は次の手順によって行う．Web ページの HTML からタグを除去し，プレーンテキストを得る．プレーンテキストに形態素解析を行った結果を使って，キーワードを抽出する．階層型カテゴリには Wikipedia で構築されているものを用い，抽出されたキーワードと比較することによってカテゴリを推定する．

5. システムの設計

システムは，Web ブラウザである Firefox[2] のアドオンとして実装した．Web ページの捜索時には，閲覧時に予め保存しておいた情報を基に，ユーザが覚えている情報と比較して候補の絞り込みを行い，捜している Web ページの候補をユーザに提示する．

5.1. Web ページ閲覧時のシステムの挙動

閲覧時のシステムの挙動を図 1 に示す．Web ページを閲覧すると，本システムを搭載した Web ブラウザはその Web ページの閲覧日時，タイトル，URL，リファラ，キーワード，カテゴリを取得し，PC 上のデータベースに保存する．

Web ページのキーワード抽出とカテゴリ推定には，処理時間がかかったり Wikipedia の膨大な情報を持っておく必要があったりするため，サーバ上で行うことにした．閲覧した Web ページの URL をサーバに送信し，サーバ上でキーワード抽出とカテゴリ推定を行い，その結果を返す．

5.2. Web ページ捜索時の挙動

Web ページを捜索したい時には，ブラウザのメニューバーから捜索システムを起動し，候補の絞り込みを行う．ユーザが検索条件を指定すると，候補となる Web ページが遷移情報を用いたツリー形式で表示される．出来上がったシステムの Web ページ捜索画面を図 2 に示す．

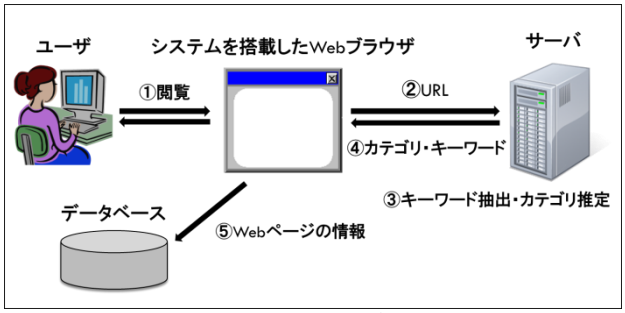

図 1　Web ページ閲覧時の挙動

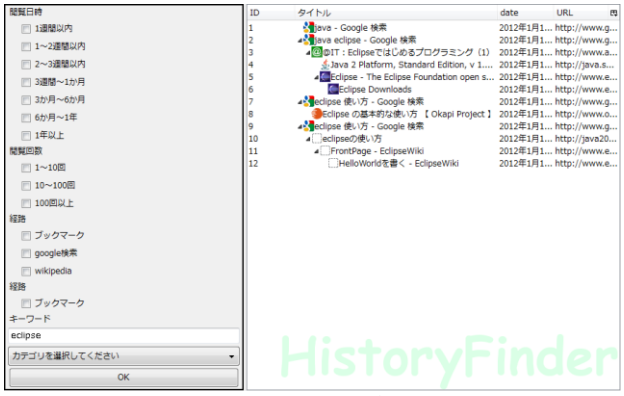

図 2　Web ページ捜索画面

6. まとめ

本研究では，うろ覚え状態のユーザが効率的に Web ページの捜索を行えるシステムを作成した．システムは，閲覧時に閲覧情報を PC 上に保存しておき，捜索に利用する．

捜索に利用する仕組みとしては，既存のブラウザが扱っている情報の他に，Web ページのリファラとカテゴリを利用した．今後は，制作したシステムの有用性について検証していきたい．

参考文献

[1] 井倉真一，近藤司，伊藤真也，原田史子，島川博光，“Web 閲覧履歴におけるうろ覚え Web ページの対話的再発見”，FIT2011(第 10 回情報科学技術フォーラム)，2011

[2] Mozilla Firefox，Mozilla，2012 年 1 月 11 日訪問，http://www.mozilla.org/