テキストマイニング技術を融合したウェブブラウザの開発

1. 背景

ウェブブラウザの多機能化はユーザの情報収集の効率化に不可欠である。しかし実際に情報収集を行うとき、検索エンジンを駆使して候補を絞り込み、ページを見て内容を確認するのは結局ユーザの仕事である。特にコンピュータの扱いに慣れていないユーザにとってこの負担は大きく、作業効率の改善が求められる。

2. 目的

本プロジェクトでは、ウェブページからキーワード等の有用な情報を抽出し、ユーザのウェブ閲覧を補助する機能を搭載することで、情報収集の効率を改善するウェブブラウザを開発する。このために必要なのが、高速で高品質なキーワード抽出と簡単に使えるインタフェースである。本プロジェクトではこの両方を新たに開発し、単体で利用可能なウェブブラウザとしての完成を目指す。

3. 開発の内容

キーワード抽出技術とウェブブラウザのインタフェースに分けて説明する。

3. 1 キーワード抽出

高速かつ高品質なキーワード抽出を実現するため、データ構造レベルからの見直しを行い、独自の手法を開発した。抽出のための索引として用いる**単語 N グラム木**(図 1 右)は、文書中の N 以下の任意の単語 N グラム(N 個の単語のつながり)の出現回数および位置を高速に検索可能な、**接尾辞木**をもとにした木構造である。任意の単語 N グラムを考慮することで、映画のタイトル等を含めた柔軟なキーワード抽出が可能である。また、「東京都知事」の中の「京都」のような意味上は存在しない単語を扱わないため、ノイズが出にくいという利点がある。さらに、「未踏」の後に必ず「ユース」が現れる場合、「未踏」を省略して「未踏ユース」のみを扱うため、メモリ効率が良いほか、キーワードのランキングのためのスコア計算にかかる時間も節約することができる。

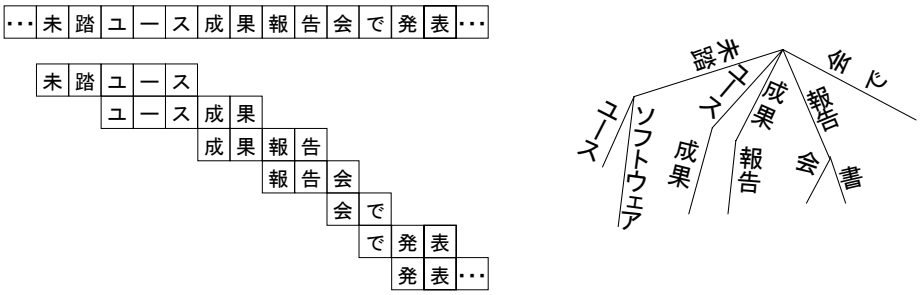
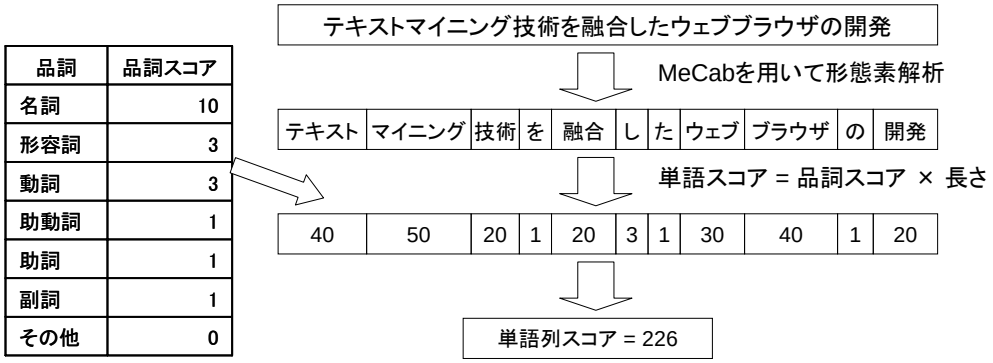


図 1 単語 N グラム(N=2)と索引構造の例

キーワードのスコア付けには、その文字列から定まる**単語列スコア**と、文書中の出現回数から定まる**頻度スコア**を用いる。単語列スコアは、各単語から求まる**単語スコア**の和で図 2 のように計算される。頻度スコアは、文書中のキーワードの出現回数の log をとった値

とする。単語列スコアと頻度スコアの積が、キーワードのスコアである。つまり、名詞などの重要な品詞を多く含み、より長く、より多く出現するほど高いスコアが与えられる。



品詞	品詞スコア
名詞	10
形容詞	3
動詞	3
助動詞	1
助詞	1
副詞	1
その他	0

図 2 単語列スコアの計算例

また、ユーザが直近に見たページ(10 ページ程度)を保存しておき、以降のキーワード抽出に役立てる技術を開発した。これは、「ユーザは関係の深いページを見ているので、過去に見たページに現れるキーワードは現在のページでも重要である」という考えに基づき、このようなキーワードのスコアを補正するものである。この補正により、100～200 文字程度の短い文章からのキーワード抽出品質を大幅に改善することができる。

ここまでの手続きでは、「未踏ソフトウェア事業」と「ソフトウェア事業育成」のように、キーワードの各単語が重なってしまい、一覧性が悪い。そこで、より上位のキーワードに既に現れている単語の割合等を考慮し、明らかに重複したキーワードを取り除くことで、より自然な結果を出力できるようにした。キーワード抽出の例を図3に示す。

以上の抽出手法を単純に実装すると、(テキスト長×N)に比例した時間がかかり、より長いキーワードを抽出したいときには非効率的である。そこで本プロジェクトでは、単語 N グラム木の構築とスコア付けを同時に行い、スコアの高いキーワードを順に出力する高速なアルゴリズムを開発した。日本語テキストによる実験では、Pentium M 1.3GHz の一般的なノート PC において、N=∞としても約 1M バイト／秒の速度でキーワード抽出が行えることを確認した。また、MeCab による日本語の形態素解析と合わせても 400K バイト／秒程度で処理できる。実際のウェブページはせいぜい数 K バイトから数十 K バイト程度であるから、ユーザのウェブ閲覧を妨げないような実時間処理を可能にしたといえる。

選別前のキーワード	選別後のキーワード
弟子の僧	弟子の僧
弟子の僧の	内供
内供	鼻
鼻	禅智内供
供	自分
弟子	池の尾
弟子の	顔
弟子の僧は	中童子
内供は	上唇の上から顋の下まで
供は	木の片

図 3 芥川龍之介「鼻」に対する上位 10 個のキーワード

3．2　ウェブブラウザのインタフェース

3．1のキーワード抽出機能をウェブブラウザに統合し、キーワード表示・検索機能、関連サイト表示機能、自動キーワードハイライト機能、要約機能、検索語入力支援機能という 5 つのインタフェースを実装した。これらのインタフェースは、見るだけで内容の理解を助けることができ、またできる限り簡単に、自動またはワンクリックで利用可能であることを基本として設計した。

開発したウェブブラウザのスクリーンショットを図4に示す。これらは IE コンポーネントと .NET Framework を利用し、C++/CLI 言語により実装した。

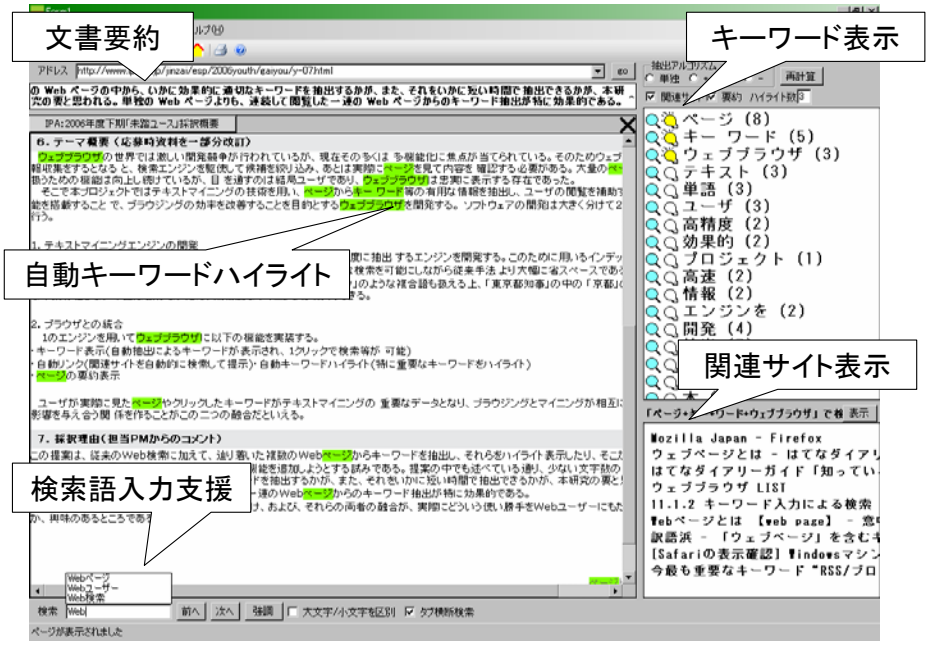


図 4　開発したウェブブラウザのスクリーンショット

以下に各機能の概要を示す。

・キーワード表示

キーワード抽出によって得られた上位のキーワードをスコア順に表示する。各キーワードはクリックでページ内検索・ウェブ検索・ハイライト表示が可能である。これにより、文書全体を俯瞰するだけでなく、マウスによる選択・コピーやキーボードからの入力といった面倒な作業を軽減することができる。

・関連サイト表示

上位のキーワードを用いて Google でウェブ検索することで、関係の深いと思われるページを抽出する。検索の手間を削減できるだけでなく、ユーザが想定しなかったページを発見できる可能性がある。

・自動キーワードハイライト

文書中に現れる上位のキーワード数個をハイライト表示することで、ユーザの閲覧を補助する。

・文書要約

上位のキーワードが多く含まれる部分を抜粋し、表示する。

・検索語入力支援

ユーザが検索語を入力する際、現在閲覧している文書を用いて、次に入力する単語を予測し提示する。

4. 従来の技術（または機能）との相違

日本語に対応したキーワード抽出ツール自体が少ない中、任意の単語 N グラムを対象にした高速なキーワード抽出手法を開発した。さらに、ユーザのウェブ閲覧という条件に特化し、これまでに閲覧した文書の情報を利用することで、短いウェブ文書からも高精度な抽出を可能にした。

また、これを実際に情報収集に役立つインタフェースに利用したことも、本プロジェクトの大きな特徴である。テキストマイニングにより、いわばコンピュータが閲覧の一部を肩代わりすることで情報収集を補助するアプローチは、多機能化によりユーザ自身の閲覧を補助する従来のものとは一線を画している。

5. 期待される効果

今回開発したキーワード抽出技術は、名詞や形容詞などに限定せず任意の単語 N グラムを柔軟に抽出可能であり、自然言語処理やテキストマイニングへの応用が考えられる。また高速に動作することから、実際のアプリケーションに利用することもできる。

6. 普及（または活用）の見通し

この半年間で、日本語の形態素解析やキーワード抽出など、関連の深いウェブサービスがいくつか発表された。本プロジェクトで開発したキーワード抽出技術も、ウェブブラウザ上での利用にとどまらず、ウェブサービス等への応用が可能である。

ウェブブラウザとしては、各機能は基本的にクリックするだけの直感的な操作が可能であり、コンピュータの扱いに慣れていないユーザが利用した場合にも高い効果が得られると期待されるので、開発成果を広く公開したいと考えている。

今後は開発を継続する一方で、プラグインやライブラリとしての提供、またウェブサービスとしての提供などを視野に入れつつ展開していきたい。

7. 開発者名（所属）

上村卓史(北海道大学大学院情報科学研究科)

（参考）開発者URL

http://www-ikn.ist.hokudai.ac.jp/~tue/