統計的手法によるWeb検索補助システム"Seezle"の開発

ー"触れる"検索インターフェースー

松田耕史*

1　背景

近年のWeb検索エンジンの研究動向として、検索対象ページを増加させること、および検索精度の向上が挙げられる。その他に、ユーザーの検索要求に対して大量のページがヒットすると、その中から目的の情報を探し出すのはユーザーに取って大きな負担になるという問題も存在し、これに対する対策は大きな研究課題となっている。上記のような問題が発生した場合、検索者は検索対象に対する不十分な知識をもとに試行錯誤によって適切な検索キーワードを決定しなければならない。また、検索によって得られたWebページの集合のうちの、どのWebページが目的とする情報を十分に含んだページであるかを判断するのは困難であり、手当り次第に閲覧しながら手探りで情報を探さなければならない状況に陥ることも多い。

また、現在の検索エンジンでは検索結果が上から順に順位付けされて表示されるが、本来情報の質というものは単一の指標で測ることができるものではなく、検索者それぞれの趣向や検索者が調べようとしている分野に対する考慮がなされているべきものである。

これらの問題を解決するための一手段として、検索結果として返されたWebページ群から絞り込み検索に用いることのできる語句を抽出し、ユーザーに対して提示することによって、ユーザーの絞り込み検索をサポートするというアプローチを考案した。

*宮城工業高等専門学校　材料工学科

2　目的

これらのことを踏まえ、本プロジェクトの目的を以下の2点に設定した。

1. 検索結果として示されたWebページ群から絞り込みに適すると思われる語句を抽出する手法を開発する。
2. 実際にユーザーインターフェースを開発し、上記によって抽出した語句をユーザーに対してどのように提示するかを検討する。

3　開発の内容

3.1　語句の抽出

本プロジェクトでは、絞り込みのための語句を抽出する方法として酒井ら[1]によって提案された手法を用いた。今回、検索によって得られた文書群S中のある文書sにおける単語wの重みを、複数文書中での出現頻度の差を考慮して以下のように定義する。

$$W(w,s) = tf(w,s) \times log(|S|/df(w)) \times log(dt(w)/tf(w,s)) \times log(|S| - n)$$

$tf(w,s)$:文書sにおける語wの頻度
$df(w)$:上位文書群Sにおける語wを含む文書数
$dt(w)$:上位文書群Sにおける語wの頻度
n:文書sの順位

全文書群 S における単語 w の重み $G(w)$ は以下の式で計算される。

$$G(w) = \prod_{n=1}^{|S|} W(w, s_n)$$

以上の式によって単語の重み $G(w)$ を計算し、これらの上位から任意の個数を選択し、絞り込みのためのキーワードとして提示した。

3.2 ユーザーインターフェース

ユーザーインターフェースは、前項で述べた処理によって抽出した語句を最大限に活かすように意識して設計を行った。

本プロジェクトでは、以下の特性を持つインターフェースを開発した。

- 一覧性の低下および下位文書の軽視という問題を解決するために、検索結果は２次元平面上に散在させる。
- 絞り込みのために抽出したキーワードと検索された文書それぞれとの間の関連を色分けという形で表現する。

3.3 実装

以下のソフトウェアにより構成されるシステムを構築した。

- 検索エンジンフロントエンド部
- HTTP クライアント部
- テキスト抽出、整形部
- 形態素解析機フロントエンド部
- キーワード抽出部
- XML レンダリング部
- 接続待ち受けサーバー
- Flash インターフェース部
- 起動用 CGI スクリプト
- 検索エントランスページ

サーバー側の処理部は Ruby を用いて記述し、形態素解析器には MeCab を用いた。また、クライアント側のユーザーインターフェースは Flash を用いて構築した。

開発したユーザーインターフェースのスクリーンショットを図１に示す。

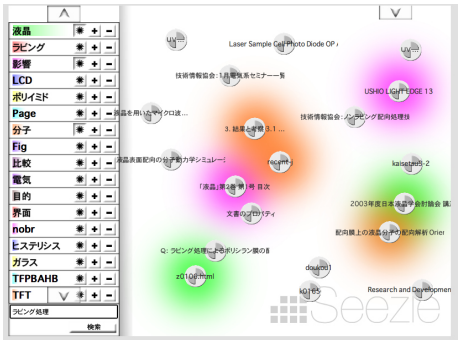

図 1: 開発したインターフェース

4 従来の技術との相違

今回開発したユーザーインターフェースは標準的な検索インターフェースの機能に加え、以下の機能を持つ。

- 検索結果を平面上に並べることができる。
- 検索結果をさらに分類する語句を表示し、語句をクリックすることで絞り込み検索を行うことができる。
- 語句を選択することで、その語に関係が深い文書に色付けを行うことができる。
- 検索結果を表すボタンは自由に動かすことができる。

これらの機能により、ユーザーは最初の一回を除き、マウス操作のみで検索を続行することが可能になる。

また、色分け機能により、ページにジャンプする前にページの内容に対する見当をつけることができ、ユーザーの余分な操作を減らすことが可能になる。

5　期待される効果

ソーシャルネットワーキングやWeb広告へと広がりを持つ、検索エンジンを中心とした技術分野は、単なる適合率や再現率にとどまらず、様々な指標での高レベルな使いやすさが求められるようになり始めている。それに伴い、少数の企業の運営するサービスの寡占的な状態から、比較的小さな企業による様々な「付加価値」を持ったサービスが多数出現しそれぞれの価値を主張する方向へ、少しずつ動きだしている。

本プロジェクトの成果も、その流れに乗り、「検索の楽しさ」を演出する検索エンジンとして、あまり普段検索エンジンを利用することのないユーザー層を中心に利用者を増やしていきたい。

6　活用の見通し

現在は検索要求を受け付けるたびに Web サーバーからページの内容を取得する構造になっている。これはサーバーマシンの処理能力上、検索サーバー内に Web 上の文書を保存してそれを利用するのは非現実的だという理由による。しかしこの構造だと、応答の遅い Web サーバーの応答時間に、検索サービス自体の応答時間が律速されてしまうという問題が生じる。現在、1 クエリーを処理するのに要する時間は 10〜15 秒ほどであり、この 7 割前後を Web ページ取得のために費やしている。

この問題を解決するために、検索サーバー内へ良く用いられる文書をキャッシュしておく方法、また、語句抽出の補助にシソーラスのような静的な言語データを用いることによって文書自体への依存度を下げるというアプローチが考えられる。

今後はこの点を重視して開発を進め、早期にサービスを提供できる状態にしたいと考えている。

7　開発者名

松田耕史 (宮城工業高等専門学校　材料工学科)
e-mail:s99937@yahoo.co.jp

8　謝辞

プロジェクトマネージャとして多数の助言をいただいた京都大学の田中克己先生、プロジェクト管理組織としてサポートしていただいた京都高度技術研究所の平家様、川上様、また、検索インターフェースという材料工学とかけはなれた題材を卒業研究のテーマとして認めてくださった宮城高専の鈴木勝彦先生に心から御礼申し上げます。

参考文献

[1] 酒井浩之、大竹清敬、増山繁; 絞り込み語提示による一検索支援手法の提案、豊橋技術科学大学、2001.

[2] 荒木良、是津耕司、角谷和俊、田中克己; リンク参照と文書構造に基づく Web ページのアスペクト抽出、DEWS2003, 2-P-03.

[3] 北研二、津田和彦、獅ヶ堀正幹; 情報検索アルゴリズム、共立出版、2002.