Web 検索結果におけるランキング変動に着目したキーワード支援システム

望月　祐臣†　　東　基衛†

早稲田大学大学院　理工学研究科　経営システム工学専攻†

1　はじめに

近年、利用者の増加によりインターネットは情報源として重要な地位を獲得している。このような状況において、情報検索者（以下、ユーザ）が膨大な Web ページの中から求める内容にあったページをより素早く見つけるためのツールの１つに検索エンジンがある。しかし、検索エンジンを用いた Web 検索はユーザの求める内容に合った Web ページを見つけられないことも多い。

そこで本研究では検索エンジンによる効率的な Web 検索を実現するシステムの研究を行った。

2　研究目的

現在提供されている検索エンジンの多くは Google のようなキーワード型検索エンジンである。それらを用いた Web 検索の際、ユーザは検索結果の上位だけを閲覧して、求める内容に合った Web ページがない場合は検索式の変更を行う。つまり、ユーザの求める内容に合った Web ページが検索結果として出力されていたとしても、それが検索結果の上位になければ閲覧される可能性は低い。

そこで本研究ではユーザの求める内容を把握し、その内容に合った Web ページを検索結果の上位に表示させることを研究目的とし、ユーザの適合・不適合判定を用いたキーワード支援システムの研究を行った。

3　現状分析と問題点

キーワード型検索エンジンを用いた Web 検索でのキーワード支援の問題点を以下にまとめた。

<問題点①>適合ページを検索結果の上位に表示させるとは限らない

検索エンジンのランキングアルゴリズムは、単語の出現頻度だけではなく、リンク構造や html タグなどの要因を総合的に評価している[1]。よって、適合ページにおける出現頻度の高い単語で検索したとしても、適合ページを検索結果の上位に表示させることができるとは限らない。

A Support System for Finding Keywords based on Ranking Change in Web Retrieval Result

Masaomi Mochizuki, Motoei Azuma, Dept of IMSE, Graduate School of Sci. & Eng. Waseda Univ.

<問題点②>AND 検索にしか対応していない

検索エンジンでは、代表的なものとして「AND 検索」、「OR 検索」、「NOT 検索」を使うことができる。しかしキーワード支援の従来研究において、AND 検索で用いることが前提となっている場合がほとんどである。

4　研究アプローチ

4.1　システム要件

以上よりシステム要件とそれを実現するための技術的な課題についてまとめた。

<要件①>適合ページを検索結果の上位に表示させること

検索エンジンのランキングアルゴリズムに対応した支援を行うために、検索結果内の情報を考慮してユーザに対する支援を行う必要がある。

<要件②>AND 検索以外にも対応させること

入力した検索式または選択した Web ページとの関連性以外の尺度を用いて、単語の重み付けを行う必要がある。

4.2　研究アプローチ

本研究で行った関連キーワード抽出に対する研究アプローチである検索結果におけるランキング変動の考え方を図１に示す。

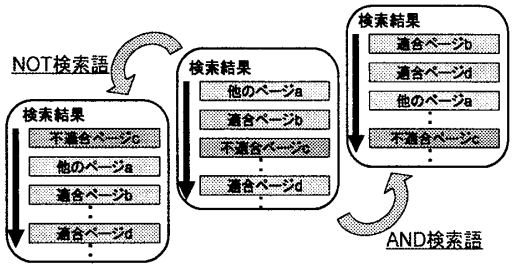

図１　検索結果におけるランキング変動

初期の検索結果では、ユーザにとっての適合ページ・不適合ページ・どちらでもないページが混在して表示されている（中央の状態）。この状態から、検索式に単語を追加することにより、検索結果におけるランキングが上昇・下降する。

このとき、ある単語を追加した場合に、適合ページのランキングが上昇し、不適合ページのランキングが下降する（右の状態）のであれば、その単語はユーザの求める内容に合った単語で

あると考えられる。よって AND 検索語として支援する。逆に、適合ページのランキングが下降し、不適合ページのランキングが上昇する(左の状態)のであれば、その単語はユーザの求める内容に合っていない単語であると考えられる。よって NOT 検索語として支援する。

以上のアプローチにより、適合ページのランキング変動に着目しているため、システム要件①を満たすことができる。また AND 検索、NOT 検索で用いるキーワードをそれぞれ支援しているため、システム要件②を満たすことができる。

5 提案システム

5.1 システム概要

提案システムは図 2 に示すように 4 つのコンポーネント、データベース、ユーザインターフェイスから構成される。

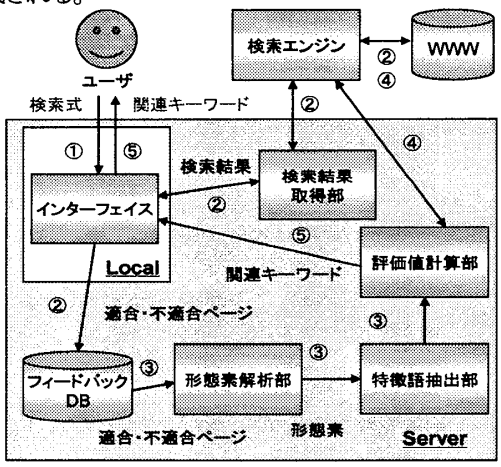

図 2 提案システム概要図

ユーザはインターフェイスにおいて、検索式の入力と検索エンジンの選択を行う(①)。システムはユーザが選択した検索エンジンを経由して、検索結果を取得・出力する。ユーザはその検索結果から適合・不適合ページを選択する(②)。システムは適合・不適合ページを形態素解析し特徴語を抽出する(③)。すべての特徴語に関して、選択された検索エンジンを経由して評価値を計算する(④)。その評価値から関連キーワードをユーザに提示する(⑤)。

以上の処理手順から、ユーザは選択した検索エンジンに対応した初期検索式の関連キーワードを得ることができる。この関連キーワードを追加して再検索を行うことで、従来システムより効率的な Web 検索を実現することができる。以下にそれぞれのモジュールの詳細を述べる。

5.2 インターフェイス・検索結果取得

インターフェイスにおいて、ユーザが検索式の入力と、利用する検索エンジンの選択を行う。適合・不適合ページ判定後は、関連キーワードが表示される。インターフェイスでの入力を基に検索結果取得部が既存の検索エンジンを経由して WWW から検索結果を取得する。

5.3 形態素解析・特徴語抽出

適合・不適合を判定したページの本文を取得する。本文は html タグなどを除去し形態素解析を行う。形態素解析結果の名詞の中から、出現頻度の高いものを特徴語として抽出する。

5.4 評価値計算

抽出したすべての特徴語に関して、評価値を計算する。評価値は式(1)で求められる。

$$value(Q,w)=\frac{1}{|T^+|}\sum_{t\in T^+}\log\frac{R_t(Q)}{R_t((Q)AND(w))}+\alpha\frac{1}{|T^-|}\sum_{t\in T^-}\log\frac{R_t((Q)AND(w))}{R_t(Q)} \quad \cdots (1)$$

$R_t(Q)$:検索式Qで検索したときのページtの順位

T^+:適合ページ数　T^-:不適合ページ数

α:重み付けのパラメータ

この評価値は、適合ページのランキングを上げ、不適合ページのランキングを下げる単語ほど高くなる。よって評価値の上位 N 語を AND 検索語、下位 N 語を NOT 検索語として提示する。

6 評価実験と考察

本研究では提案システムの有効性を検証するために、プロトタイプを利用した評価実験を行った。評価はキーワードの妥当性、検索結果上位での適合率について Google を用いて行った。キーワードの妥当性評価を図 3 に示す(手法①:提案手法、手法②:従来手法)。

表 1 キーワードの妥当性評価

		適合率	
		A	A+B
手法①	AND	0.419	0.747
	NOT	0.403	0.706
手法②	AND	0.347	0.663
	NOT	0.288	0.55

これらの評価項目に関して従来手法と比較した結果、提案システムの有効性が示された。

7 おわりに

本研究では効率的な Web 検索の一手法として Web ページに対するユーザの適合・不適合判定を用いたキーワード支援システムを提案した。検索結果におけるランキング変動に着目することにより、ユーザの求める内容を反映できるキーワードを支援することができた。

今後、さらなる抽出精度向上のためにキーワード抽出手法の改良などを考える必要がある。

参考文献

[1] 神崎洋治, 西井美鷹: “体系的に学ぶ検索エンジンのしくみ”, 日経ＢＰソフトプレス, 2004