キーワード抽出による情報検索支援システムの研究

能登研究室 **前原拓哉 (20046284)**

1 はじめに

近年、計算機の急激な性能の向上とインターネットの普及により膨大な情報が計算機上でアクセス可能になりつつある。その一方で求める情報を的確に検索して欲しい情報を得ることは容易ではない。主に、インターネットのユーザが Web ページ上の情報を検索するとき、検索結果はユーザが入力したキーワードを含むページを出力する。よって、非常に多くのページの中からユーザの望む情報だけを得ることができない。ユーザが検索対象として存在する Web ページの内容を知らないことや、度忘れなどによってユーザが検索語をすぐに思いつかないこと、あるいは検索語として適当な単語をもとから知らないことが原因として起こる。また、語彙の不一致などが考えられる。

本研究ではキーワードを抽出することによりユーザの情報検索を支援するシステムを提案する。抽出されたキーワードによってユーザは望む情報を得るための検索式を入力することが可能である。

2 提案システム

検索を行うユーザは、ブラウザ上のインターフェースから検索式を入力する。入力された検索式で検索を行い、検索されたページを実際にユーザが閲覧する。ユーザ自身が自分の求めている欲しい情報が得られなかった場合には検索結果の上位 50 の文書群により語の重み付けを行い、キーワードを抽出する。そして、ユーザは抽出されたキーワードを眺め、検索要求に適した語を自由に複数選択することで再検索を行なう。これらの操作を繰り返すことによってユーザの望む情報を得る事ができる。

本研究の検索方法は情報が一つだけ欲しいとき (天気など) ではなく、多くの情報（論文検索など）が欲しい時や調べたい事柄があいまいの場合に有効に働くのである。また、検索語の不足を補うので、抽出されたキーワードにより検索がスムーズに行なうことができ、Web ページにおけるユーザの興味とのギャップを埋めることができる。システムの流れを図 1 に示す。

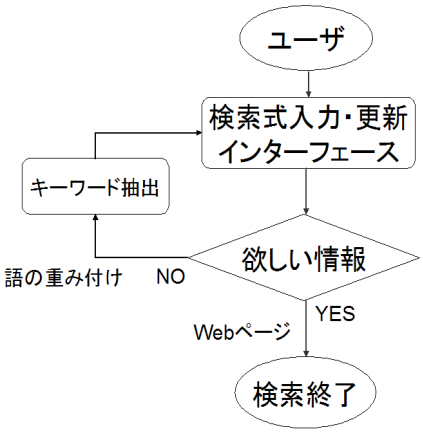

図 1: システムの流れ

3 tf×idf 法・語の重み付け

tf (term frequency) は、対象となる文書においてその単語がどれくらいの頻度で出現するかをあらわしたもので、tf が大きいほど単語がその文書をよく特徴付けているといえる。しかし、tf だけではその単語が文書を特徴付けているとはいいきれない。また、idf (inverse document frequency) はその単語が出現する文書が少ないほど、その単語がよく特徴付けていると考えるものである。idf を用いるといくつかの比較する文書の中で、その単語が出現する文書が少なければ、その単語に対する重みを大きくするという考え方である。tf×idf はそれら二つを掛け合わせたもので、すなわち、tf が大きく、df が小さい (idf が大きい) ならば、その単語が文書を大きく特徴付けるといえる。

語の重み付けには tf×idf 法を参考にする。具体的には、上位文書群 S の文書 s に含まれる複合名詞と、カタカナ、英字表記の語、地名、組織名を語とし、その語に以下の式で重み $W(w,s)$ を付ける。

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

$tf(w,s)$：文書 s における語 w の頻度,
$df(w)$:上位文書群 S で語 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.

こうして算出された重みが高い語から順に検索者に提示することにより Web 上に存在するユーザの望むべき文書に到達するための語として有効であると考えられる。

4 考察

本システムはユーザの検索語不足を補い、ユーザが望む情報を得るための検索支援を行う。評価実験には複数の被験者に複数の検索課題を与えて実際に検索を行なってもらいキーワードなしの場合とありの場合での比較を行なうことで目的の文書を検索する場合の検索のしやすさを検討した。本システムでは抽出されたキーワードにより望むべき文書に確実に到達できるという点では有用であると考えられるが、WWW 上で検索する際に検索式を入力するには個人差があり、万人に使いやすいものではないと考えられる。

5 おわりに

本システムは、抽出されたキーワードをユーザ自身が検索式を更新しなければならない。よって、今後の課題としてはユーザが入力、更新しやすいインターフェースを作成することや、より検索しやすいシステムの作成が挙げられる。さらに、ユーザ自身の求める情報を具体化して的確な検索式作成を行うことのできるシステムの作成が課題である。