tf-idf 法の改良による Web 情報検索結果の表示法に関する研究

能登研究室　　阿部剛仁 (16002)

1　はじめに

World Wide Web の情報空間の内容の充実と拡大に伴って，情報を探す手段としてのサーチエンジンの重要性はますます高まっている．しかし検索するページ数が膨大であった場合，ユーザが求めている内容のページがページのトップに出現するとは限らないなどの問題点もある．

本研究ではユーザの求めているページを検索結果から排除しユーザにとって見易い検索結果を表示させることを目的とする．

2　サーチエンジン

サーチエンジンには，大きく分けてディレクトリ検索とロボット検索の 2 種類が存在する．前者は人の手によってホームページの内容を登録し，カテゴリからページを絞り込んでいく検索で，Yahoo!などに代表される．後者は WWW サーバ上をロボットが常に巡回してページのソースを自動収集し，ユーザが入力したキーワードが入っているページを出力するものであり，goo などに代表される．

サーチエンジンは WWW の情報を検索するのに有効であるが，ユーザが求めている情報がページのトップに出てこないことが多いようである．これは HP 作成者がアクセス数を増やすためにページに手を加えてあるというのが原因の 1 つとして考えられる．

本研究では検索結果が膨大な量になってしまうロボット検索について研究を行う．

3　問題点の改善

ユーザが目的のページを探しにくいのは検索結果が膨大な数になってしまうからである．よって先にあげた問題点を解決することで見易い検索結果の表示をすることができると考え，それを改善する方法を以下に提案する．

3.1　tf-idf 法

ページ内で繰り返し使われている語句はそのページにおいて重要と考えられる．よって出現回数（term frequency: tf）をランク付けに使用する．しかし多くのページで使用されている語句は索引語としての有用性は低い．そこで出現ページ数の逆数（inverse document frequency: idf）も使用する．具体的には tf と idf を掛け合わせることで重みを定めることとなる．この手法を tf-idf 法と言う．

3.2　問題点の改善方法

あるキーワードをサーチエンジンで検索したとき，キーワードと全く関係のないページが検索されたことはないだろうか？

もし背景と同じ色で文字が書かれていた場合，ユーザの閲覧するページには何の支障もなくキーワードをそのページに埋め込むことができる．サイトへのアクセス数が一定数を越えることでプロバイダから宣伝料が支払われるなどの金銭的な問題もあり，アクセス数を増やすために HP 作成者がこのような行為をすることは少なくない．これが検索結果にキーワードと関係のないページが出力される原因の 1 つである．これを改善するためには背景と同じ色で書かれた文字を出現文字として換算しないような tf-idf 法を開発すれば良い．

通常，検索を行う場合，ページのソースをデータベースに蓄積する．そして HTML 言語などの命令文字を排除した状態で検索を行う．だが，命令文字を排除した時点で背景と同色の文字も排除し検索を行えば，このような問題点を改善できると考えられる．

具体的には HTML 言語において背景を設定する BGCOLOR=“XXXX” とフォントを設定する FONT COLOR=“XXXX” の両者の XXXX の部分が一致したとき，そのフォントで書かれた文字を tf として換算しないというプログラムになる．

しかし色の設定には “black” や “white” などのように英語で書かれたものと “＃ 000000” や “＃ fffff” などのように光の三原色を 16 進数で表したものがある．この “black” と “＃ 000000”,“white” と “＃ fffff” はそれぞれ同じ色を表示することとなでどちらにも対応するようなプログラムを構築する必要がある．

4　実験

今回は既存の「goo」を使って実験を行った．キーワードとして使われることが多い「MP3」というキーワードを使用し，またページに手を加えているのはアダルトサイトに多いと言うことで,「MP3」と「アダルト」の AND 検索をする．その結果，検索結果数は 1126 件が検索され，実際にトップページに表示された 10 ページ（1 ページは存在せず）を開いたところ，背景と同じ色で書かれたページが 6 ページ存在した．開いたのはトップページの 10 ページだけだが残りの 1116 件のページにトップページと同じ割合でこのようなページが入っていると考えると背景と同じ色で書かれたページ数 X は

$$X = A \times \frac{M}{10 - N}$$

ただし
A：検索結果数
M：トップページ内で背景と同じ色で書かれた文字が存在したページ数
N：トップページ内で存在しなかったページ数
となり全検索結果の中でこのようなページ数は 750 件存在すると予想される．半数以上のページが該当することからこれによってキーワードと関係ないページのランクを下げることができ，ユーザの見やすい検索結果が表示できると考えられる．図 1 は goo の検索結果ページの種類の分類である．

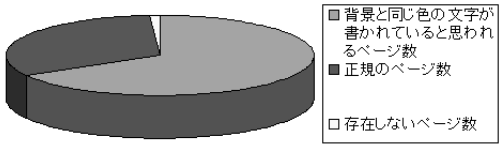

図 1: 実験結果

5　おわりに

本研究では，サーチエンジン「goo」を用いて背景と同色で書かれた文字を出現数として換算しないシステムを提案し，ユーザにとって有用な検索結果を表示できることがわかった．今後は，ホームページの URL から ソースを取得し，そこに本稿で述べたようなシステムを適用してランク付けを行えるサーチエンジンのプロトタイプシステムの構築が必要である．また，既存のサーチエンジンと比較し，このシステムが有効的であることを示す．