スマートサーチ：賢いメタ・サーチエンジンの開発

6Z－4

荻野　譲（早稲田大学理工学研究科）
成田誠之助（早稲田大学理工学部）

1. はじめに

現在、インターネット上では数多くの情報が乱雑に発信されている。この混乱の中から自分に必要な情報を見つけだすツールとしてサーチエンジンが近年急速に発展してきた。しかし情報量の急増という現状の前にこれらの従来型サーチエンジンでは応答速度の向上に注意が向きすぎ、効果的な情報検索の方法を欠いている。

そこでこれらの従来型サーチエンジンを用いながら、それ以上の高い関連性を持った文書検索を可能にするメタ・サーチエンジンを開発した。本稿においてはその動作原理とその効果性を検証する。

2. メタ・サーチとは

現在、インターネット上で用いられているサーチエンジンには大きく分けて三種類のものがある。一つはロボット型（AltaVista[1], Lycos[2]等）、二番目は登録型（Yahoo![3]等）、そして最後がメタ型（MetaCrawler[4]等）である。

ロボット型サーチエンジンはその名の通り、世界中のWWWサーバーに向けてロボットを放ち、公開されているHTMLファイルを集めてくる。それを自分のデータベースに溜め込み、時々更新することでサーチエンジンとしての役割を果たす。この作業すべてが全自動故、情報量では他のタイプに勝るものの、正確さでは必ずしもトップとはならない。

登録型サーチエンジンは逆に手動式である。ユーザーが自分のページを登録し、それをサーチエンジンのスタッフがチェックして回る。ほとんどの作業が人間の手でなされるため、ジャンルが間違っていることはほぼないものの、登録数が極端に少ない。

これら二種類の従来型サーチエンジンは上記のようにデータベースの構築方法には違いがあるものの、そのデータベースを検索する方法は（Excite[5]のように同義語検索を行うところもあるが）基本的にワード・マッチングである。それ故、あまり高精度な結果は期待できない。

それに対し、メタ・サーチエンジンでは自前のデータベースを持たない。従来型サーチエンジンの結果を利用することによりデータベース構築のコストを節約する。この節約により、より高度な検索技術そのものの開発に力を入れることができる。既に実用化されているMetaCrawlerなどの場合、各従来型サーチエンジンを網羅的に使用することにより情報の漏れがないことを保証する一方、これらからの結果を吟味することにより、より関連性度の高い結果を出すことに成功している[6]。

3. スマートサーチの提案

そこでこの次世代型サーチエンジンを改良し、これまでの不満を吹き飛ばすスマートサーチを提案したい。これまでの不満としては

- デッドリンクが結果に含まれている
- 同じページが複数回リンクされて現れる
- 全然関係のないページがリンクされている
- 関連性度が当たっていない

等が挙げられる。

そこで次のような要素をサーチエンジンの関連性度算出基準として用いることとした。

- タイトル・テキスト・URLに現れたキーワードの数
- 従来型サーチエンジンが振った関連性度
- 同一WWWサーバー上にあるデータの数
- 専門用語かどうかによって検索領域を限る
- どの従来型サーチエンジンが出したリンクかにより信頼性度[7]を掛け合わせる
- ホストのタイプ(com, edu, org, etc.)から情報の種類を推測
- URLの長さによって重要性を推測

これらの要素の相互関連性は未知であるので、これまでのように単に足し合わせるだけでは相互関連性を調べることができない。そこで、ニ

SmartSearch: Intelligent Meta-Search Engine
Shirabe Ogino (ogino@narita.elec.waseda.ac.jp)
Seinosuke Narita (narita@narita.elec.waseda.ac.jp)
Waseda Unversity Graduate School of Science and Engineering

[1] http://www.altavista.digital.com/
[2] http://www.lycos.com/
[3] http://www.yahoo.com/
[4] http://www.metacrawler.com/
[5] http://www.excite.com/
[6] http://www.narita.elec.waseda.ac.jp/~ogino/Project/ を参照

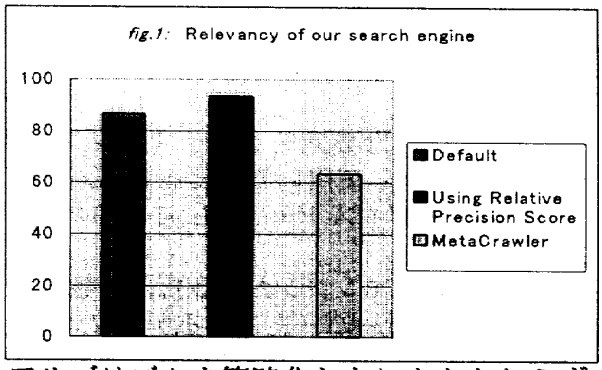
ューラルネットワークを用い、学習させることによって、相互関連性を考慮できるようにする。

4. スマートサーチの有効性の検証

実際にスマートサーチの実装に入る前にこのアルゴリズムがいったいどれだけ有効であるかを検証した。有効性の検証が目的であったのでプログラムの手間を省き、市販の表計算ソフト等の助けを借りながら手作業を中心に行った。前章で提案したアルゴリズムすべてを手作業で行うことはできないので、このシミュレーションに用いたアルゴリズムは簡略化した以下のものである。

- 従来型サーチエンジンの返した関連性度を初期値として用い、同一の URL が複数のサーチエンジンから返された場合はそれらの関連性度を足し合わせた
- デッドリンクをチェックし除去
- 前もって行った評価[7]により得られた各サーチエンジンの信頼性度を係数としてかける
- 関連性度でソートした後、同一関連性度内ではURLの長さが短いものを優先した

この作業の結果、図 1 のように有効性が証明された。左から「サーチエンジンの信頼性度を用いなかった場合」「信頼性度を用いた場合」「既存の MetaCrawler」が出力した関連性度の評価結果である。



アルゴリズムを簡略化したにもかかわらず、予想以上の有効性が確かめられた。既存の従来型サーチエンジンの成績は 60%を切っており、MetaCrawler でさえ 62%しかない中で、90%を越える成績を出していることは驚異的である。すべてのアルゴリズムをプログラムした際にはこれよりもさらに高い精度で結果を出すものと期待される。

5. スマートサーチの実装

前章で有効性が証明されたことを受けて、現在、実際のプログラムを行っている。一番ふさわしい利用形態は、各クライアントマシンが自分でスマートサーチを持つことであるが、WWW ブラウザーでの Java Applet の使用には数多くのセキュリティ問題とそれに伴う制限[8]が存在する。そのため、メインプログラムはスマートサーチ・サーバー上で動作させることとし、これらの制限を回避する。このことによりサーバーへの負荷は高まるが致し方ない。Java Applet の使用はユーザーインターフェースにとどめる。メインプログラムはどのような言語を用いても開発できるが、将来に柔軟性を残すため、現在は Java を使って開発している。

6. まとめ

今回が初めての発表となるため、スマートサーチの動作原理とその有効性を証明した。また、実装に関する報告も行った。

現在、サーバー公開のためにドメインを取得し、OCN の工事完了を待っているところである。公開後はニューラルネットの学習に必要なデータを集めるために一人でも多くの利用者が訪れることを祈っている。

http://www.smartsearch.or.jp/

[7] http://www.narita.elec.waseda.ac.jp/~ogino/Project/Results/Results_Index.html を参照

[8] http://java.sun.com/sfaq/index.html