WWW 上の知的情報検索システムに関する一考察

前田 晴美
大阪市立大学　学術情報総合センター
E-Mail: harumi@media.osaka-cu.ac.jp

近年、インターネットの普及に伴い、WWW(World Wide Web)上に誰でも簡単に利用できる膨大な情報資源が出現し、増加し続けている。今後図書館において WWW 上の電子化資料をどのように扱うか検討するために、現状の WWW の課題と動向を把握することは意義深いと考える。本稿では、現状の課題の中でも情報検索に焦点をあて、現状の検索エンジンの技術的課題を整理する。次に、試作システムが WWW 上で公開されている知的情報検索システムの研究動向を概観し、技術的課題へどのようなアプローチが行われているか考察する。

1. はじめに

近年、インターネットの普及に伴い、WWW(World Wide Web)上に誰でも簡単に利用できる膨大な情報資源が出現し、増加し続けている。今後図書館において WWW 上の電子化資料をどのように扱うか検討するために、現状の WWW の課題と動向を把握することは意義深いと考える。本稿では、現状の課題の中でも情報検索に焦点をあて、現状の検索エンジンの技術的課題を整理する。次に、従来の検索エンジンにない知的なふるまいをする情報検索システム（以下、知的情報検索システムと呼ぶ）の研究動向を概観し、技術的課題へどのようなアプローチが行われているか考察する。以下、2.では現状の検索エンジンの技術的課題について、3.では知的情報検索システムの研究動向について述べる。

2. 検索エンジンの情報検索に関する技術的な課題

2.1 検索エンジンの種類と課題

一般に、検索エンジンは、ロボットを用いてWWWページ（以下ページ）を自動収集するロボット系、階層ディレクトリを提供するディレクトリ系に分類できる。その他、複数の検索エンジンに検索に行くメタサーチ系や、特定の分野やトピックの情報を検索する専門検索エンジンなどがある。日本では、Yahoo! Japan（ディレクトリ系）と Goo（ロボット系）がよく使われる二大検索エンジン [1]である。検索エンジンの特徴や比較調査に関しては、多くの文献（例えば [2] [3] [4]）が存在するので、そちらを参照されたい。

現状の検索エンジンの課題は、情報資源をどのように収集、組織化、提供するかという情報検索面と、著作権の保護や課金をどのように行うかという運営面に大別できる。

以下では、前者に関する技術的な課題について述べる。

2.2 利用者側からみた問題

ロボット系の検索エンジンでは、情報が豊富で網羅的であるが、情報が多すぎて有用な情報にたどりつくのが難しい。逆に、登録に基づく検索エンジンでは、求めている情報が収集されていないために検索できないことがある。

キーワード検索は、一般に簡便な検索方法であるが、探したい情報に関する知識や検索式に関する知識が少ない場合や、漠然と情報を探したい場合には検索が難しい。ディレクトリ検索はこれらの欠点を補うが、運営者の作成したディレクトリ階層と利用者の考える階層が異なる場合に検索が難しい。

2.3 運営者側からみた問題

WWW上の情報が増加し続ける中、特にロボット系の検索エンジンではコンピュータ資源が不

足することが予測されている。また、ディレクトリ系では、運営者の情報の選別と組織化の負荷が非常に高くなっている。

また、WWW 上では、URL の消失・変更、文書の更新が頻発する。これらは、特にロボット系の検索エンジンにおいて情報の増大原因となっており、メンテナンスが問題となっている。

2.4 技術的課題

いくつかの問題を概観したが、最大の問題は、増加し続ける情報資源をどう扱うかという情報過多の問題である。以下に、利用者支援、運営者支援の観点から主な技術的課題項目を述べる。

(1)利用者支援

・高度な利用者検索支援 （課題 1a と呼ぶ、以下同様。）
・個人のニーズに基づく情報の収集（課題 1b）

(2)運営者支援

・価値の高い情報資源の収集、組織化（課題 2a）
・コンピュータ資源利用の効率化 （課題 2b）
・URL の変更管理 （課題 2c）

3. 知的情報検索システムの研究動向

2. で述べた技術的課題を解決するために、どのような研究が行われているのか調べる。

3.1 調査方法

最終的な目的は、WWW上の情報検索の技術的課題に関して、どのような研究開発が存在し、何がどの程度解決できているのか調査することである。しかし、この分野は、理論から応用、研究プロジェクトから商用システムまで世界中に関連する研究開発が数多く、しかも変化が非常に速いため、すべてを調べることは難しい。

そこで、第一段階として、試作システムが WWW 上で公開されている事例を調査することとした。中でも、従来の検索エンジン（ここでは Yahoo! Japan と Goo）にない知的なふるまいをする知的情報検索システムに焦点をあて、ACM、IEEE、AAAI、情報処理学会、人工知能学会などの主として情報工学系の雑誌・会議録に掲載されている研究プロジェクトを中心として調査した。以下に、98 年 10 月現在の中間報告として、テキストを対象とした事例のいくつかを、アプローチ毎に分類して述べる。各事例の元々の目的はここで述べるアプローチとは異なるものもあるが、筆者の観点で分類している。

3.2 事例

3.2.1 キーワード拡張

利用者が入力したキーワードに関連する単語を提示することにより、利用者の検索式入力を支援する。課題 1a に対応する。

・シソーラス辞書検索：あらかじめ備えた辞書（類義語、関連語）を用いて、キーワードを拡張する。
・MONDOU：文書の内容を解析して、入力キーワードに関連の高いキーワードを自動的に判定する。日本語が使える。

その他、AltaVista をはじめとして、キーワード拡張可能な検索エンジンが多数ある。

3.2.2 検索結果の自動分類（クラスタリング）

キーワード検索結果の URL を自動分類（クラスタリング）して提示することにより、利用者の検索式入力を支援する。課題 1a に対応する。

・HuskySearch：キーワード検索結果のクラスタリング表示が非常に高速である。
・クラスタリング機能付き TITAN：クラスタの数が選択可能で、日本語が使える。

3.2.3 アクセスや投票に基く情報の価値の決定

ページへのアクセスや投票統計を提示することにより、利用者の情報の選択を支援する。課題

1a、1b と 2a に対応する。
・Alexa：利用者がページを選択するたびに、関連するページやそのアクセス・投票統計を表示することにより、利用者の検索を支援する。プラグインのフリーソフト。
・Phaoks：ネットニュースでの引用統計を元に有用なページを判定し、利用者のキーワード検索に対して、点数の高いページを優先的に提示する。
・WiseWire：企業向けの商用サービスで、利用者の登録と投票に基いてより価値の高い情報提供を行う。98 年 10 月 Lycos に買収された。

3.2.4 ネットニュースグループの自動分類（カテゴリー化）
ネットニュースグループをあらかじめ決めたカテゴリーにしたがって自動分類して検索インタフェースを提供する。課題 2a に対応するが、1a、1b にも応用できる。
・Pharos：ネットニュースグループを LC Classification に従って分類する。
・NETCow：日本のネットニュースグループを、キーワード群を定義することにより多様なカテゴリーで分類する。

3.2.5 収集したページの自動分類（クラスタリング）と階層マップ表示
収集したページを自己組織化アルゴリズムを用いて自動分類して階層地図インタフェースを提供する。処理時間がかかるためどちらかといえば運営者向けで課題 2a に対応するが、課題 1a にも対応する。
・SiteMap：ある URL を起点としたページや、文書などから階層地図を生成する。
・ET-Map：娯楽関連のページから階層地図を生成する。

3.2.6 選別された情報資源からの情報の抽出
ネットニュースの記事や特定のトピックに関するページを集めて、その中から決まった項目の情報を抽出して組織化する。課題 2a に対応する。
・情報の自動編集：日本のネットニュース fj.meeting の記事から会議カレンダーの自動生成や、fj.sys.sun の記事から FAQ(Frequently Asked Question)の自動生成。
・FAQ Finder：ネットニュースの記事から FAQ を自動抽出・組織化し、利用者の自然言語質問に FAQ を提示する。
・Product Finder：登録された商品情報のページから価格や説明などの情報を抽出し、商品名の入力などに基いて、商品価格や詳細情報の一覧を提示する。Excite 提供のサービス。

3.2.7 収集したページから内容判定
・Ahoy! (Homepage Finder)：メタ検索エンジンで収集したページなどから、個人のホームページを推定する。利用者の氏名入力に基いて、ホームページと E-Mail アドレスを提示する。課題 2a に対応する。

3.3 考察
全体として、実用化または実用化に近い研究は、利用者支援が多い。多くのシステムに使われている手法は、伝統的な統計的・確率的なテキスト検索技術に加えて、(a)質問文や文書からキーワードや文を抽出するための自然言語処理、(b)利用者のプロファイルや文書の特徴などを獲得するための学習処理、(c)カテゴリー化やキーワード生成処理などに利用される知識処理である。
利用者支援に関しては、キーワード拡張に多くの研究があり、有効である。シソーラスを利用する検索エンジンは既に実用化段階である。検索結果の自動分類（クラスタリング）に関しては、結果の品質の向上が望まれる。個人のニーズに応じた情報収集に関しては、プロファイルを学習する研究事例がいくつかあるが、公開されているシステムは少ない。
運営者支援に関しては、情報資源の対象を限定（例えば、ニュースグループや、商品に関連するページ）することにより、知識の記述が比較的容易となり、情報の自動組織化に役立つことがわかる。情報資源を限定せずにロボットが収集したページから内容を判定するシステムには、個人のホームページを見つける研究があるが、逆に、他は難しいと言えるかもしれない。

今回は、システムにアクセスできるもののみを調査対象としたため、基礎的な研究、初期段階の研究、システムを公開していない研究（例えば企業の研究）については言及していない。また、知的情報検索システムを対象としたため、言語や URL の標準化指向のアプローチ（例えば XML や URN）や、分散サーバの相互接続などのアプローチ（例えば検索エンジンの協調や Z39.50 やメタデータ統合）や、既に別分野を築いている翻訳研究などは対象としなかったが、今後はこれらの研究との比較も行う必要がある。

4. おわりに

現状の検索エンジンの情報検索に関する技術的課題について整理し、これまでに試作されたいくつかの知的情報検索システムについて調査し、中間報告を行った。今後も、システムを中心とした研究動向の調査を継続して行うとともに、他のアプローチについても調べていきたい。

参考文献

1) インターネット白書'98, 日本インターネット協会編, pp.114, 1998.
2) 山名早人: WWW情報検索サービスの動向, http://www.etl.go.jp/~yamana/Research/WWW/survey.html
3) 検索デスク, http://www.bekkoame.or.jp/~asaisan/
4) Search Engine Watch, http://searchenginewatch.internet.com/

最近の技術動向に関する文献

1) 住田一男, 三池誠司: 知的情報検索の動向, 人工知能学会誌 Vol.11, No.1, pp.10-16, 1996.
2) 武田英明: ネットワークを利用した知的情報統合, 人工知能学会誌,Vol.11, No.5, pp.680-688, 1996.
3) 林良彦, 小橋吉嗣: WWW 上の検索サービスの技術動向, 情報処理学会誌, Vol.39, No.9, pp.861-865, 1998.

システム名とURL

1) Yahoo! Japan http://www.yahoo.co.jp/
2) Goo http://www.goo.ne.jp/
3) シソーラス辞書検索 http://search.kcs.ne.jp/the/
4) MONDOU http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/
5) AltaVista http://www.altavista.com/
6) HuskySearch http://zhadum.cs.washington.edu/zamir/cluster.html
7) クラスタリング機能付き TITAN http://titan.mcnet.ne.jp/index-x2.html
8) Alexa http://www.alexa.com/
9) Phaoks http://www.phoaks.com/
10) WiseWire http://www.wisewire.com
11) Pharos http://pharos.alexandria.ucsb.edu/demos/
12) NETCow http://www.ricoh.co.jp/rdc/ic/demo/netcow/
13) Sitemap http://lislin.gws.uky.edu/Sitemap/Sitemap.html
14) ET-Map http://ai2.bpa.arizona.edu/ent/entertain1/
15)情報の自動編集 http://www-sato.jaist.ac.jp:8000/people/sato/jap/research/editing/
16) FAQ Finder http://infolab.cs.uchicago.edu/faqfinder/
17) Product Finder http://www.jango.com/
18) Ahoy! http://ahoy.cs.washington.edu:6060/

（注）URL はすべて 98 年 10 月 11 日にアクセス確認した。