サーバ分散型キーワード検索システム Ingrid の実現

1T－3

清水 奨　神林 隆　佐藤 進也　横路 誠司　ポール フランシス
日本電信電話株式会社　ソフトウェア研究所

1　はじめに

インタネットでは、検索に対するニーズが強まっている。従来から多くの検索システムが運用されているが、検索サーバは一か所で集中管理する形態が一般的である。特に WWW を対象とした検索サービスでは、ロボットと呼ばれるソフトウェアで情報を集め、インデクシングを行ってサービスを提供することが多い。このためサーバの能力、データ内容の管理の両面から、いずれ限界に達するおそれがある。

これに対し、WAIS[1] や Harvest [2] では比較的小規模な検索エンジンを分散して置く事によりスケーラビリティを確保しようとしている。しかし、利用者が実際に検索をする際にはどのサーバにどのような情報がよく網羅されているかを知っている必要がある。また検索サーバを運営する側にとっては、他のサイトの検索ページなどに自分のサーバを登録しなければサービスを広い範囲に提供できない。

本稿では、複数の検索サーバの自律的な運営により構築され、サーバの位置を広報しなくとも検索サービスを提供できる分散型キーワード検索システム Ingrid[3] を紹介する。

Ingrid では、情報の提供者あるいは検索システム運営者が任意に検索サーバを設置すれば全体から検索可能となる。検索サーバは分散して配置されるためスケーラビリティを確保でき、利用者に対して一つの世界に見せることで必要な予備知識を少なくしている。

2　Ingrid の特徴

Ingrid の特徴として、キーワードベースであることと分散型であることの二つが挙げられる。

[キーワードベース] 内容をキーワードで示すことができ、URL[4] で情報の場所を示せれば、検索対象に制限はない。現在一般的な WWW ページ検索の他、アーカイブからのソフトウェア検索や個人のメールアドレスの検索などへの応用が考えられる。キーワードさえ与えれば、ビデオや画像、サウンドのデータベースにも適用できる。

[分散型] 検索サーバは図 1 のように分散して構成することができる。

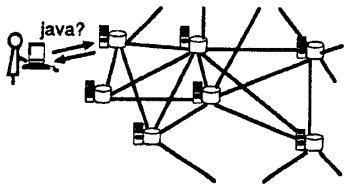

図 1: Ingrid の概念図

図 1 では、ユーザは自分のネットワークの Ingrid サーバにキーワード'java' で検索を行い、問い合わせを受けた Ingrid サーバは検索用のネットワーク (Ingrid トポロジ) をたどってキーワード'java' を持つ情報を集めて提示する。DNS[5] に似たイメージであるが、Ingrid トポロジは木構造ではなく網状であり、図 2 のような構造をしている。図 2 では情報を楕円で表わし、キーワードを中に表示している。また角丸の枠でクラスタを表わしている。

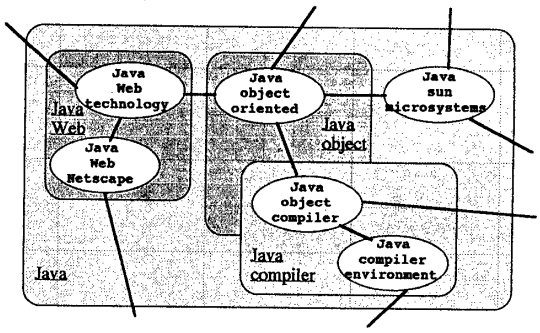

図 2: 検索ネットワーク (Ingrid トポロジ)

Ingrid トポロジは、図 2 のようにキーワードを付与した情報をノードとし、なるべく多くのキーワードを共有するノード間にリンクを張る。さらに個々のキーワードの組み合わせ毎にクラスタをなすように構成し、同じトピック内の情報検索を支援する。図 2 の例で言えば、キーワード'Java' で検索を始めたユーザは検索結果から'Java' の他に'compiler','object','Web' などのキーワードが関連しており、検索に有効であることを知ることができ、キーワードを追加して検索範囲を絞り込んだり、新たにキーワード 'Web' で検索を開始することができる。Ingrid トポロジは、各サーバに分散して保持され、自律的に維持される。

Imprementation of distributed keyword-based search infrastructure 'Ingrid'
Susumu Shimizu, Takashi Kambayashi, Sin-ya Sato, Seiji Yokoji and Paul Francis {shimizu, kam, sato, yokoji, francis} @slab.ntt.jp
NTT Software Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo 180 Japan
Special Thanks to Kazuhiro Kazama for imprementing the navigator.

3 実装

現在の実装は図3のような構成であり、主な検索対象はWWW上のHTMLで記述された情報である。

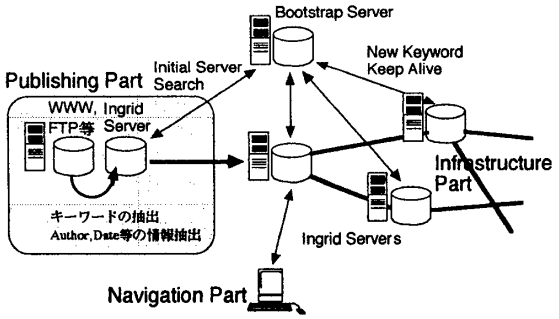

図3: Ingrid システム構成

Publishing Partは、検索対象となる実際のデータからIngridで検索するためのインデクス情報(キーワードの組、URL、タイトル、著者、日付、言語その他)を抽出し、Ingridトポロジに組み込む部分である。日本語と英語に対応しており、tf-idf法[6]に基づく自動キーワード抽出を行なっている。トポロジに組み込む際には、組み込もうとする情報が持つキーワードの組で検索を行い、見つかったノードの内、共有するキーワード数の最も多いものに対しリンクを張る。一つもキーワードが存在しなかった場合は、Bootstrapサーバに登録するだけでリンクは張らない。

Infrastracture Partは、Ingridトポロジの維持管理、及びキャッシングを行う。ノードには寿命があり、頻繁に検索されるもの以外はPublishing Partが定期的にRefreshを行わないとobsoleteとなり削除される。新しいIngrid Serverが起動した時や、新しいキーワードがトポロジに組み込まれる時はDNSのルートサーバにあたるBootstrap Serverに自動的にパケットが送られる。これには一つのキーワードにつき数台のIngridサーバが自動登録される。

Navigation Partは、Ingridトポロジ内の情報を検索し閲覧するためのユーザインタフェースである。図4に示すように、関連するキーワードと情報のタイトルを対比させ、キーワードのみで情報の関係が理解できるよう工夫されている。なお専用のインタフェースを持つNavigatorのほか、CGI経由のものとJavascriptを使うものがある。

多言語対応のため、Ingridでは内部コード系にMuleコード[7]を採用している。

4 評価実験

Ingridのスケーラビリティを評価するため、公開実験を行っている[8]。Bootstrapサーバ一台、Ingridサーバ十台、検索可能な情報数5万件程度の規模で動作することを確認しており、ボランティアを募集しながらサーバ数十台、情報数100万件規模の環境を構築中である。

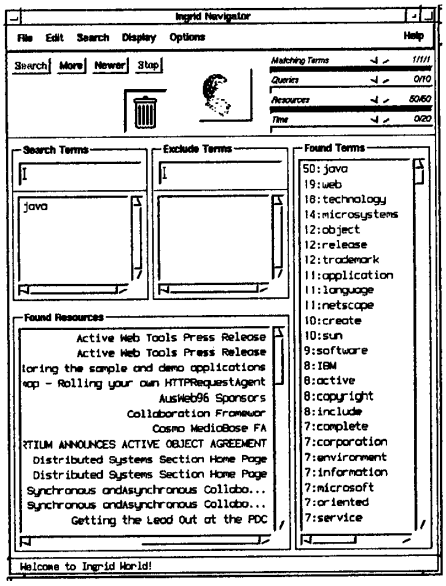

図4: ユーザインタフェース

実験はWWWページ検索以外にも、メーリングリストアーカイブの検索、検索サービスの検索、ソフトウェアアーカイブの検索など複数の応用分野で同時進行的に行なわれている。主な技術課題は、キーワードの品質、セキュリティ、サーバパフォーマンスが上げられるが、他の研究コミュニティと協調しながら問題の解決にあたっている。

5 今後の予定

今後は実験を通じ、数千万規模の情報を扱う事を目標とする。将来的には、検索の為のインフラストラクチャとして整備して行きたいと考えている。当面の計画としては、インタネットワイドな検索に加え組織・マシンに閉じたデータベースに対応することや、他のアジア系言語に対応することがある。

参考文献

[1] C. Stanfill, R. Thau. "A Parallel Indexed Algorithm for Information Retrieval" in Proc. of the SIGIR-89, ACM

[2] C. Mic Bowman, et,al. "The Harvest Information Discovery and Access System" in Proc. of the 2nd Intl. Conf. on WWW, p763-771, Oct. 1994

[3] P. Francis, et,al. "Ingrid: A Self-Configuring Information Navigation Infrastructure" in Proc. of the 4th Intl. Conf. on WWW, p519-537, O'reilly & Associates Dec. 1995

[4] T. Berners-Lee, L. Masinter, M. McCahill. "Uniform Resource Locators(URL)", RFC1738, ftp://ds.internic.net/ rfc/rfc1738.txt, Dec. 1994

[5] P. Mockapetris, "Domain names - concepts and facilities", RFC1034, ftp://ds. internic.net/rfc/rfc1034.txt, Dec. 1994

[6] G. Salton, "Automatic Text Processing", Addison Wesley

[7] N. Mikiko, H. Keninchi, T. Satoru, "Mule: MULtilingual Enhancement to GNU Emacs" Proc. of INET'93

[8] Ingrid Project, Ingrid ホームページ, http://www.ingrid.org/