WWWサーバのグループ化による分散型情報検索

1AC－7

中村 勇一郎† 川上 豊† 秋山 幸生‡ 満永 豊†

†NTTアクセス網研究所 ‡NTT技術協力センタ

1 はじめに

WWWの膨大な情報リソースから目的の情報を検索するために，サーチエンジンと呼ばれる検索データベースが提供されている．現在のサーチエンジンにおける検索リソースの収集は，探索ロボットと呼ばれるソフトウェアを用いて，WWW内を自動的に探索・収集する方法[1]と，サーバの管理者から寄せられる情報から手動で情報をカテゴリ分けする方法に大きく分けられる．

しかし，サーチエンジンの問題点として，一般に集中型の膨大な検索データベースから，キーワードが情報に含まれるかのみを判定するので，利用者にとって適切な情報への絞り込みが難しく，検索の質の向上が困難であることが挙げられる．さらに，別の問題点として，現在の情報リソースの急速な拡大，情報の頻繁な変更に対して，検索データベースの更新が対応できず，スケーラビリティの点で問題がある．スケーラビリティの改善の試みとしては，Ingrid[2]が提案されているが，検索の質の向上については考慮されていない．このように，前に挙げた2つの問題を同時に解決する方法が示されていないのが現状である．

本研究では，1つのWWWサーバ(以下，サーバ)の，1つのカテゴリに属する情報リソースを1人の専門家にみたて，人が専門家へ問合せる行為と，専門家相互の情報交換を，WWWの検索サービスにモデル化する．すなわち，検索サービスを専門化し，分散配置することにより，高い質を持ち，スケーラビリティのある検索環境を提供する．

2 提案する検索技術

同じカテゴリに属する情報リソースをグループ化し，1.人の専門家への問合せと専門家間での連携，2.専門家同士の情報交換の2つの行為をモデル化する(図1)．ここでは，同じカテゴリの情報を持つ，もしくは興味を持つ，サーバとクライアントの集合を，コミュニティと呼ぶことにする．コミュニティにおいては，サーバ間で検索要求を連携して検索を行い，またサーバ間で検索に必要な情報リソースのインデックスを交換する．

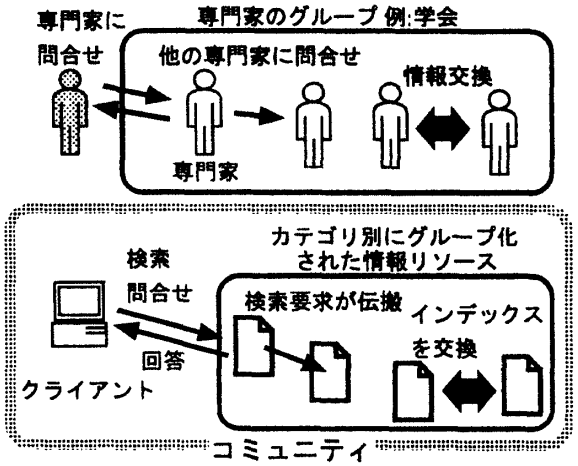

図1: コミュニティ内の情報検索

以上により，以下の特徴が期待できる．

- 検索のインデックスが分散し，情報リソースの変更が各ローカルなサーバ内のインデックスの更新のみで済み，スケーラビリティがある
- 情報をカテゴリごとにグループ化し，専門化するので，適切なコミュニティに参加し，検索を行えば，より高い検索の質が期待できる

以下，3章では，このような検索技術を実現するための具体的な課題について述べる．

3 課題

3.1 サーバが持つ情報を動的に交換

本節では，あるカテゴリに属する情報リソースを連携し，コミュニティを形成する手順について述べる．まず，カテゴリごとにサーバにインデックスファイルを置く．このインデックスファイルは，コミュニティを記述するものであり，コミュニティ内の情報リソースの所在を表し，キーワー

Distributed Information Retrieval System at Grouped WWW Servers.

Yuichiro NAKAMURA†, Yutaka KAWAKAMI†,
Yukio AKIYAMA‡, Yutaka MITSUNAGA†
†NTT Access Network Systems Laboratories
‡NTT Technical Assistance & Support Center
Tokai-Mura, Naka-Gun, Ibaraki-Ken 319-11 Japan

ドとその作成日付，優先度，URLによって記述されるインデックス情報部，コミュニティ自体を表し他のインデックスファイルの所在をURLで記述するポインタ部，に大きく分けられる．このとき，具体的な手順は，つぎのようになる．

1. 情報リソースの管理者は，ローカルに持つインデックスファイルのポインタ部で，近い情報を持つ情報リソースのインデックスファイルをいくつか指定する．それを繰り返すことにより，コミュニティが自然に形成される．
2. 情報リソースのインデックス情報を，各ローカルのインデックスファイルに記述する．そのインデックス情報を，コミュニティ内で1で指定したインデックスファイルから取得する．
3. コミュニティ内をインデックス情報が伝搬するにつれ，優先度を下げる．ローカルのインデックス情報は，日付の古いもの，または優先度の低いものを削除することにより管理する．

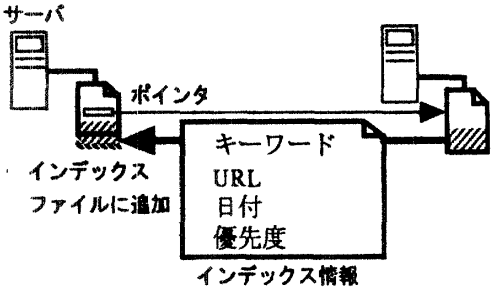

図 2: インデックス情報の流通

これらの手順により，コミュニティが自動的に生成され，コミュニティ内で情報の共有がなされる(図2)．このとき，新しいインデックスが優先的に残り，また，優先度の導入により，関連性の低いインデックスは検索時に参照されにくい特徴を持つ．

3.2 検索動作の伝搬と管理方法

本節では，コミュニティ内でデータを連携する方法および，それを利用した検索について述べる．ここで，検索で用いられる検索データは，クライアントの指定する検索キーワード，最大ホップ数，検索履歴からなり，検索データをサーバ間で連携し，情報検索を行う．したがって，クライアントが検索する際に，データの連携を意識する必要はなく，1つのサーバのみにアクセスすればよい．具体的な検索手順は，つぎのようになる．

1. クライアントは検索データをコミュニティ内のサーバのうちの１つに送る．
2. サーバは，自分の持つインデックスファイルのインデックス情報部から検索する．検索要求とインデックス情報のキーワードが終了条件を満たした，結果を返して終了．が終了条件を満たしたら，結果を返して終了．
3. 検索データに履歴を追加し，インデックス情報の優先度，日付と，部分マッチの点数より決まるつぎのサーバに検索データを送り，2へ．

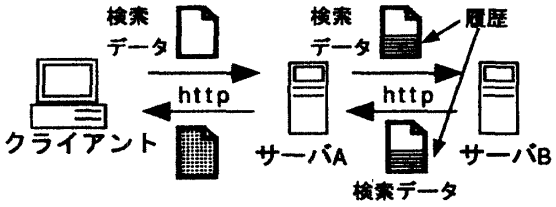

図 3: 検索動作の伝播

ここで終了条件は，完全にマッチすること，全くマッチしないこと，クライアントの指定する最大ホップ数を越えたことのいずれかを満たすことをいう．以上により，コミュニティ内で検索動作が伝搬し，分散したインデックスを実現できる(図3)．また，検索は各ローカルなサーバ内で行われるため，情報の変更に対しても直ちに対応できる．

4 まとめ

本発表では，検索の質を高め，各ローカルなサーバ内の操作のみで分散情報検索を実現する枠組みを示した．今後考慮すべき点としては，検索の質をより高める為に，インデックスファイル間でリンクを張る際のガイドラインを明確に定めること，分散型の検索では，検索に時間がかかることが挙げられる．以上を考え合わせると，本システムはイントラネットのような比較的小さいコミュニティに適したシステムであるといえる．今後の課題として，シミュレーションにより，ガイドラインをより明確化すること，検索の履歴を利用することにより，各インデックスファイル間のリンクを再構成し，検索の効率化を図ることが挙げられる．

参考文献

[1] 林良彦,"探索ロボットに基づくWWWサーチエンジン", 電子情報通信学会 情報・システムソサエティ大会講演論文集, pp.580-581, 1996

[2] ポール フランシスら,"次世代情報検索インフラストラクチャIngrid", NTT R&D, Vol.45, No.2, pp.159-165, 1996