M-035

個人の PC による P2P ネットワークを基盤とした 柔軟な分散 Web 検索システムの提案
A Proposal of a Flexible Distributed Web Retrieval System on a P2P Network with Private PCs

豊田 正隆　　勅使河原 可海
Masataka Toyoda　　Yoshimi Teshigawara

1. はじめに

近年，共通のテーマに沿って個人サイトを作成して情報を公開している人々と，そのテーマに興味を持つ閲覧者による，Web 上のコミュニティが増加している．それらのコミュニティに属する人々にとっては，コミュニティ内で公開される新鮮な情報を検索することの必要性は大きい．

これらの情報を発見する手段としてサーチエンジンがある．しかし，一般のサーチエンジンはクローリングの対象範囲と実行間隔がトレードオフの関係にあるため，新たに公開された情報を即座に検索することができないという問題を抱えている．

クローリングの対象範囲を特定のコミュニティに絞ることで，クローリングの間隔を短くすることが可能になる．それによってコミュニティで公開されて間もない情報を検索することができる．しかし，特定のコミュニティだけのためのサーチエンジンは存在しない．また，そのためのサーチエンジンを個人で構築するには膨大なコストがかかる．

そこで我々はこれまでに，個人の所有するマシンを利用した，個人サイト特化型分散 Web 検索システムについての研究を行ってきた[1]．実環境実験の結果，一部の PC の停止や IP アドレスの変化に対応できる柔軟性が必要であることが分かった．我々は，これを実現するためには P2P 基盤を用いることが有効であると考えた．本稿では，P2P ネットワークを用いてこれらの柔軟性を実現した，個人 PC による分散 Web 検索システムを提案する．

2. 従来システム

我々がこれまで研究してきたシステムは，個人によって提供される PC をサーバとした分散 Web 検索システムである．このシステムは登録されているサイトのみを検索対象とする．各サーバは自身が担当しているサイトに対して 30 分間隔でクローリングを行い，インデックスを更新する．これにより，30 分以前に公開された情報の検索を可能としている．また，分散システムとすることで，性能の面で劣る個人 PC であっても，大量の Web サイトに対して 30 分間隔でクローリングすることを可能としている．

また，サーバとサイトを共にカテゴリに分け，サーバにカテゴリと適合するサイトだけを担当させる．これにより，カテゴリを指定して検索を行うことで，1 つの検索要求に関して全てのサーバに検索処理を行わせずにすむようになっている．

従来システムは，サーバが固定 IP アドレスを持ち，常時外部からアクセス可能であることが前提となっていたが，そのことが PC 提供を妨げる要因となっていることが分かった．また，個人の PC は障害の発生率が高いと考えられるが，従来システムは 1 台のサーバの停止によりそのサーバが担当するサイトが検索できなくなってしまう．

創価大学大学院工学研究科
Graduate School of Engineering，Soka University

3. 提案システム

本稿で提案するシステムは，個人によって提供される PC をノードとした P2P アプリケーションとして動作する分散 Web 検索システムである．各ノードは自身が担当しているサイトに対して 30 分間隔でクローリングを行い，インデックスを更新する．ノードはカテゴリに分けられ，適合するカテゴリに属するサイトのみを担当する．

3.1 想定基盤

システムが基盤とする P2P ネットワークは以下の機能を持つとする．

(1) IP アドレス隠蔽機能
(2) NAT / FW 越え機能
(3) ノードのグループ化機能
(4) 情報をアドバタイズメントとして P2P ネットワーク上に伝播する機能
(5) 指定した条件に適合するノードを発見する機能
(6) 存在確認

(1)と(2)の機能により，固定 IP アドレスを持たないマシンや，NAT または FW 内のマシンであっても，システムに参加することが可能となる．

3.2 概要

1 つのサイトのインデックスを保持するノードが 1 つしか存在しない場合，そのノードがシステムから脱落した場合に，そのノードがインデックスを保持するサイトに関して検索が出来なくなってしまう．そこで，いくつかのノードによってグループを作成し，同じグループのノードには同じサイト群のインデックスを保持させる．同じグループ内のノードは定期的に互いの存在確認を行い，存在が確認できない場合には，グループに所属していないノードをグループに引き込む．これにより，インデックスの冗長性を維持する．検索を行う場合には，全てのグループに検索要求を送ることにより，全てのサイトを検索対象とすることができる．

3.3 構成

システムの構成を図 1 に示す．システム上には，システムを構成する全てのノードから成るシステムグループ，1 つのカテゴリを担当するノードから成るカテゴリグループ，同じサイト群を担当するノードからなるサイトグループの 3 つのグループが存在する．サイトグループは d 個のノードから構成される（図の場合 d = 2）．ノードはアクティブノードとアイドルノードに分類される．アイ

ドルノードはサイトグループに属していない，すなわちインデックスを保持していないノードである．

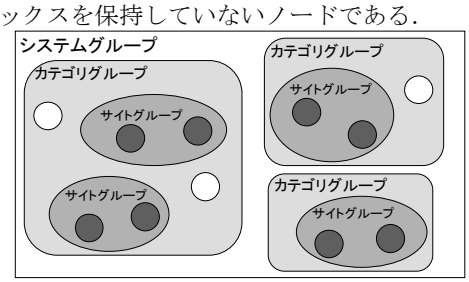

図 1 システムの構成

システムグループのノードは，各カテゴリグループのいくつかのノードの情報を保持する．また，カテゴリグループのノードは，同一カテゴリグループ上の各サイトグループのノードと担当するサイトの情報を保持する．このように，いくつかのアイドルノードと各カテゴリグループのノードの情報を常に保持しておくことで，各機能の実現を可能とする．これらの情報は，各ノードが P2P ネットワーク上に伝播するアドバタイズメントから取得する．また，各サイトグループのノードは，担当しているサイトのインデックスを共有する．

1 つのノードが保持できるインデックスの数には制限がある．そのため，各カテゴリグループに登録されているサイトの数によって適切な数のサイトグループが必要となる．1 つのノードが保持できるインデックスの数を c，登録されているサイトの数を m とおくと，必要なサイトグループの数 g は下に示す式①から得られる．ただし ceil(x)は x 以上の最小の整数を表す．

$$g = \mathrm{ceil}(m / c) \quad \cdots①$$

以下，システム上のノード数 n が d * m / c に対して十分大きいものと仮定して話を進める．

3.4 機能

システムが持つ機能についての詳細を次に示す．

(1) 検索

検索者は任意のノードに対し，検索語と検索対象カテゴリグループを含む検索要求を送る．要求を受け取ったノードは，対象カテゴリグループに自身の属するカテゴリグループが含まれていれば，全てのサイトグループに対して要求を転送する．また，対象カテゴリグループに自身の属さないカテゴリグループが含まれていれば，そのカテゴリグループに属するノードに要求を転送する．

要求を転送されたサイトグループのノードは，自身が保持するインデックスに対して検索を行い，その結果を要求元に返す．要求を転送された他カテゴリグループのノードは，同一カテゴリグループに属する全てのサイトグループに要求を転送し，その結果を要求元に返す．この時のサイトグループの動作は前述のものと同じである．検索者から要求を受け取ったノードは，全ての転送先から結果が返ってきたら，その結果を検索者に返す．各ノードは検索結果の返送前に，スコアに基づくソートと重複削除を行う．

(2) サイトの登録

登録者は任意のノードに対し，サイトの URL と登録対象カテゴリグループを含む登録要求を送る．この時，式①において g が増加する場合がある．g が増加した場合，選択されたカテゴリグループに属するアイドルノードを d 個選択して新たなサイトグループを作成する．サイトはそのサイトグループに登録される．g が増加しない場合，選択されたカテゴリグループに属するサイトグループのうち，担当しているサイト数が最も少ないものを選択し，そのサイトグループにサイトを担当させる．

どちらの場合でも，担当するサイトが変化したサイトグループはその情報を P2P ネットワークに伝播する．

(3) サイトの脱退

脱退者は任意のノードに対し，サイトの URL を含む脱退要求を送る．要求を受け取ったノードは，要求を P2P ネットワークに伝播する．要求が伝播されたノードは，自身が持つ情報から，そのサイトの情報を削除する．

サイトの登録とは逆に，式①において g が減少する場合がある．しかし，g が減少するという事象は発生しにくく，減少してもサイトの登録によって再び g が増加することはよくある．また，サイトグループを減らすためには，削除対象のサイトグループが担当していたサイトを他のサイトグループに登録しなおすというコストの高い作業が必要になる．そのため，サイトの減少によって g が減少しても，グループの再編は行わない．

(4) ノードの追加

追加者は任意のノードに対し，追加要求を送る．要求を受け取ったノードは必要な情報を追加されたノードに与える．追加されたノードはアイドルノードとなる．

(5) ノードの削除

システム上のノードは個人の PC であり，障害やシャットダウン等により，何の前触れも無くシステムから削除される．そのため，ノードの削除はサイトグループ内での生存確認によって検出される．アイドルノードの削除は検出できないが，アイドルノードがリアルタイム性を要求する機能を実現する際に必要となることはないため，検出できないことによって生じる問題は特に無い．

サイトグループに属するノードの削除を検出した場合，同一のカテゴリグループから適当なアイドルノードを選択し，新たなサイトグループのノードとする．

(6) クローリング / インデックス作成

サイトグループに属するノードは 30 分毎に担当するサイトのクローリングを分担して行う．クローリングが終わり次第インデックス作成を行い，生成されたインデックスをサイトグループ内で同期する．

4. まとめと今後の課題

本稿では P2P ネットワークを基盤とした柔軟な分散 Web 検索システムを提案した．このシステムは従来システムの機能に加え，ノードの障害を許容する機能を備えている．また，固定 IP アドレスを持たないマシンや，NAT または FW 内のマシンであっても，システムにノードとして組み込むことが可能である．

今後は，より性能の高いノードを優先してサイトグループのノードとするなど，システムの性能を高めるような動作を検討していく．また，提案システムの有効性をシミュレーション等によって評価する．

参考文献

[1] 豊田正隆，山崎賢悟，勅使河原可海：個人サイトのカテゴリ分けを利用した分散 Web 検索システムの実装と有効性の評価，DICOMO2005，5F4，2005.7