L-020

RSS の存在に依存せず新設／更新された情報を即座に収集する分散 Web 検索システムの提案

A Proposal of a Distributed Web Retrieval System to Correct New/Updated Information Independent of RSS

豊田　正隆†　　山崎　賢悟‡　　勅使河原　可海†
Masataka　Toyoda　　Kengo　Yamazaki　　Yoshimi　Teshigawara

1. 研究の背景と目的

Web 上で公開されている情報を発見する一般的な手段としてサーチエンジンがある．しかし，現在のサーチエンジンはインデックスの更新頻度が低いため，頻繁に更新される情報を発見するのには不向きである．RSS を用いることで，Web ページの新設および更新を検知することは可能であるが，RSS の URL を知る必要があるため，ユーザが存在を知らないサイトや RSS を設置していないサイトについての更新等を検知することは不可能である．そこで本稿では，Web 上に公開された様々な情報を，RSS の存在に依存することなく，即座に収集することが可能な分散型検索システムを提案する．

2. 関連研究

分散検索に関する研究として，Ingrid[1]や協調サーチエンジン(Cooperative Search Engine：CSE)[2]がある．

Ingrid は収集したリソースに付加されている Resource Profile を基にリソース間にリンクを作成する．これらのリンクに基づいて，Ingrid トポロジと呼ばれる独自のトポロジが形成される．各リソースは Forward Information(FI)サーバによって管理される．検索は Ingrid トポロジのクエリについての経路探索により行われる．Ingrid は Ingrid トポロジ上の情報を網羅的に検索することが可能だが，インデックスの更新間隔の短縮を目的としておらず，また各 Web サーバに FI サーバをインストールする必要があることが本研究との違いである．

CSE は更新間隔の短縮を目的とした分散型サーチエンジンである．各 Web サーバに Local Search Engine と呼ばれる局所的サーチエンジンを配置し，それらによって作成される局所的なインデックスのサマリ情報を Location Server と呼ばれるサーバで一括管理する．CSE は更新頻度を大幅に短縮することが可能だが，イントラネットを主な対象にしていること，全体を統括するサーバが必要なこと，専用ソフトのインストールが必要であることが本研究との違いである．

また，インデックスの更新間隔を短くすることに特化したものとして，News&Blog Search[3]がある．これはニュースサイトやブログポータルサイトを 10 分から 30 分間隔でクローリングを行うものである．しかし，ニュースサイトの整形された形式やブログポータルサイトの RSS を利用しているため，どのようなサイトにでも適用できるわけではない．

3. システムの概要

システムは検索対象となるサイト群，代表サーバ，それらの集合からなるコミュニティによって形成される．代表サーバは各コミュニティに 1 つ存在する．システムの概要を図 1 に示す．

†創価大学工学部

‡創価大学大学院工学研究科

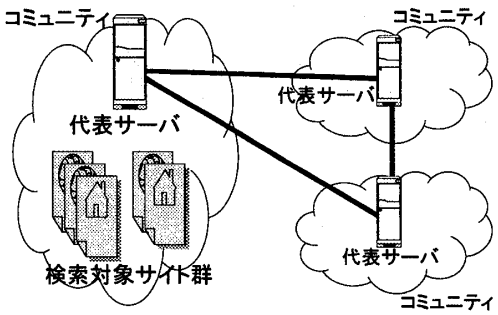

図 1. システム概要図

代表サーバは自身が属するコミュニティ内のサイト群の Web ページを収集し，インデックスを作成する．この収集の間隔を可能な限り短くすることで，インデックスを最新に保つ．また，各コミュニティが連携することで広域においての検索を可能とする．

3.1 コミュニティへの参加

サイトはトップページの URL を代表サーバに登録することでコミュニティに参加することができる．RSS が存在するサイトの場合は RSS の URL も合わせて登録する．

検索対象にしたくない Web ページがあれば，ページまたはディレクトリ単位で指定することにより，後述するクローリングから除外することができる．

3.2 クローリング

代表サーバはコミュニティに登録されているサイトのトップページを起点としてリンクを辿ることで，各サイトの Web ページを収集する．トップページより上位のディレクトリや他ホスト上のページ，コミュニティに参加する際に指定したページ，同じく指定したディレクトリ以下に位置するページは収集しない．収集した Web ページに関して，リンクの抽出，更新日時の決定，インデックスの更新を行う．インデックスはサイトごとに作成する．更新日時は，その Web 文書自体と，そのページに含まれる画像や音楽などのファイルの更新日時のうち，最新のものとする．

RSS が存在するサイトの場合は，RSS を参照することによって更新された Web ページだけを収集する．

3.3 特徴語の決定

クローリングによって作成されたインデックスから，各サイトの特徴語を抽出する．サイトから抽出する特徴語の数はサイトの規模が大きいほど多い．各特徴語はインデックスから決定される重みを持つ．

次に，コミュニティ内の各サイトの特徴語からコミュニティの特徴語を決定する．コミュニティから抽出する特徴語の数はコミュニティの規模が大きいほど多い．各特徴語は，各サイトの特徴語の重みによって決定される重み(< 1)を持つ．

3.4 コミュニティ間の連携

コミュニティ間に特徴語についての連携を定義する．ある 2 つのコミュニティが同じ特徴語を持つとき，それらのコミュニティはその特徴語について連携を持つことができる．連携には重みがあり，その重みは相手コミュニティにおけるその特徴語の重みとする．連携の例を図 2 に示す．

特徴語	重み
野球	0.8
サッカー	0.5

特徴語	重み
野球	0.8
バレーボール	0.5

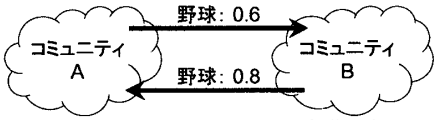

図 2. コミュニティ間の連携の例

コミュニティは自身の特徴語とその重みのリスト (特徴語リスト)を連携しているコミュニティに定期的に送信する．特徴語リストを受け取ったコミュニティは連携しているコミュニティに転送する．また，自身の特徴語と共通している語を特徴語リストに見つけた場合，その語についてまだ連携していなければ，新しく連携を作成する．また，前回の特徴語リストと比較して，削除された特徴語があった場合，その語について連携していれば，連携を削除する．連携が作成された場合，その旨を相手コミュニティに通知する．通知を受けたコミュニティは，同じように連携を作成する．

特徴語リストは Hop Count(HC)と Max HC(MHC)を持つ．HC の初期値は 0 であり，転送されるたびに 1 ずつ増加する．MHC は特徴語リストを作成したコミュニティが決定する．HC が MHC 以上になった場合，その特徴語リストは破棄される．

3.5 検索時の動作

検索者は語の組み合わせからなるクエリを任意のコミュニティに対して送ることができる．検索者からクエリを受け取ったコミュニティをルートコミュニティと呼ぶ．ルートコミュニティの代表サーバは，当該コミュニティにおける検索を行うと同時に，クエリに含まれる語について 1 つでも連携しているコミュニティにクエリを転送する．クエリを転送されたコミュニティの代表サーバは，同じように検索と転送を行う．

クエリは重みを持つ．ルートコミュニティに送られてくるクエリの重みは 1 である．クエリがコミュニティ間を転送される際，コミュニティ間の連携の重みとクエリの重みの積を転送されるクエリの重みとする．クエリに含まれる複数の語について連携しているコミュニティ間を転送される場合は，最も大きい連携の重みを採用する．クエリの重みがある閾値を下回った場合，そのクエリは転送されず破棄される．

クエリはルートコミュニティの代表サーバによって割り振われる一意な ID を持つ．同一の ID を持つクエリが複数転送されてきた場合，最も大きい重みを持つクエリを採用する．

閾値は検索者が指定することができる．また，コミュニティ間の連携の重みが 1 または 0 であるかのように検索を行うことができる．連携の重みを 1 であるとした場合，クエリの重みはサイト間を転送される際に減衰しないので，連携を辿ることのできる全てのコミュニティを検索することができる．連携の重みを 0 であるとした場合，ルートコミュニティのみを検索することになる．

クエリの転送の例を図 3 に示す．クエリは「野球」，閾値は 0.5 としている．矢印は連携，その線上の数字はその重み，吹き出しの数字はクエリの重みを表す．コミュニティ Y からコミュニティ W にクエリが転送される際に，クエリの重みが閾値を下回っているため，クエリが破棄されている．

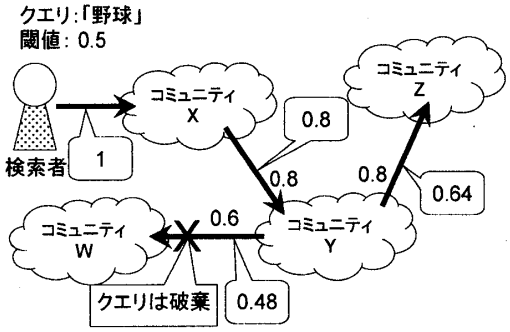

図 3. クエリの転送の例

クエリを転送されたコミュニティは，全ての転送先のコミュニティから検索結果が返送されてから，そのコミュニティの検索結果と合わせて転送元のコミュニティに返送する．その際，検索結果のスコアに連携の重みを乗算する．

ルートコミュニティは返送されてきた検索結果と自分の検索結果を合わせてスコアの降順にソートし，ユーザに提示する．必要に応じて Web ページの更新日時もスコアに反映させる．

4. 期待される効果

Web ページの収集範囲をコミュニティという限定された範囲に絞ることでインデックスを更新する間隔を短くすることが可能となる．それによって，コミュニティ内の Web ページの更新を即座にインデックスに反映させることが可能になる．さらに，コミュニティ間が連携を行うことでコミュニティに属しているサイト全体から情報を検索することが可能になる．

また，検索者がクエリを送るコミュニティを決定することで検索者が興味を持つコミュニティに焦点を当てた情報検索が可能になることが期待される．

サイトの運営者はサイトの URL を伝えるだけでコミュニティに参加することができる．RSS が存在するサイトの場合は，RSS を利用することによってクローリングを効率化することが可能である．

5. まとめと今後の課題

本稿では RSS 等を利用しなくても即座に Web ページを収集することが可能な分散 Web 検索システムを提案した．

今後はシステムの実装を行い，本手法の有効性を確認する．また，サイトおよびコミュニティの適切な特徴語の数や抽出方法，重みの決定方法等を検討する．

参考文献

[1] Ingrid : http://www.ingrid.org/
[2] 佐藤永欣，上原稔，酒井義文，森秀樹：最新情報の検索のための分散型サーチエンジン，情報処理学会論文誌，Vol.43，No.2，pp.321-331，2003
[3] News&Blog Search : http://news.drecom.jp/