

P2P ネットワークを基盤とした分散 Web 検索システムの実装と評価

豊田 正隆† 勅使河原 可海†

創価大学大学院工学研究科† 創価大学工学部†

〒192-8577 東京都八王子市丹木町 1-236

E-mail: {mtoyoda, teshiga}@soka.ac.jp

同人誌即売会に参加するサークルは、即売会の直前に発刊情報をサークルのサイトで公開する場合がある。しかし、一般のサーチエンジンでは個人サイトで公開されて間もない情報を検索できない。我々は、個人の所有する計算機資源を利用して、簡便に構築・運用が可能な分散 Web 検索システムの研究を行っている。試作システムを用いて実験を行い、既存の商用サーチエンジンと比較しても、最新情報の検索に有利であることを確かめた。

Implementation and Evaluation of Distributed Web Retrieval System Based on Peer-to-peer Network

Masataka Toyoda† Yoshimi Teshigawara†

Graduate School of Engineering, Soka University†
Department of Engineering, Soka University†

1-236 Tangi-cho, Hachioji, Tokyo, 192-8577 Japan

E-mail: {mtoyoda, teshiga}@soka.ac.jp

Some circles which participate a comic convention sometimes publish what they will distribute just the day before the convention. However, a general search engine which crawls the whole web cannot retrieve instantly such newly published information because of a trade-off relation between crawling range and frequency. In this background, we designed and developed a distributed web retrieval system focused on personal web sites. This system has features of utilization of individual computing resources and easiness to construct and to operate on the peer-to-peer network infrastructure. We evaluated by an experimental use and found that this system achieve better performance for retrieving fresh information than a conventional general search engine.

1. はじめに

1.1 研究の背景

近年、World Wide Web (以下、Web) の発展により、個人が Web 上でサイトを運営することによって情報発信を行う事例が増加している。これら個人サイトの運営者のうち、経済的・社会的にも大きな影響を与えていると考えられる対象に、マニア消費者層が挙げられる。野村総合研究所のニュースリリースによると、マニア消費者層はインターネット利用率と情報発信能力が高く社会的影響力が強いとし、また、購買意欲が高いだけでなく、コミュニティ形成の核となつてきている[1]。この中でも同人誌即売会に参加する層と分類される消費者は、マニア消費者層の1/3以上を占めている。この層には同人誌を執筆する者も含まれており、その多くは発刊予定等の自身の活動を各自が運営する個人サイト上で公開している。

しかし、これらのサイトで公開される発刊情報は、同人誌が発刊される即売会の開催前日に公開されるようなこともままある。これらの Web 上の情報を発見する一般的な手段はサーチエンジンであるが、一般のサーチエンジンは個人サイトにおいて公開された情報が即座に検索できるようにはならない。これは、サーチエンジンでは Web ページを収集し(文書収集)、高速に検索可能な形式に変換する(索引化)という作業が必要であり、索引化された情報しか検索できないためである。巨大な Web 全体に対して行う文書収集にはかなりの時間がかかり、個人サイトに関しては数週間毎にしか索引が更新されない。

検索の対象範囲と索引の更新頻度はトレードオフの関係にある。そのため、文書収集の範囲を即売会に参加するサークルのサイトに絞れば、索引の更新頻度を高めることができる。しかし、現状では特定の即売会のためだけ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新たに構築することは、多大な費用と運用の手間を必要とする。

1.2 研究の目的

我々は同人誌即売会を対象として、その即売会に参加するサークルが公開する情報を、即座に発見可能な検索システムの簡便な構築を目的として研究を行ってきた。

本システムは、その即売会に興味を持つ個人の所有する計算機資源を利用した分散検索システムである。すなわち、個人が家庭で利用する PC に専用ソフトウェアをインストールしてもらうことで、本システムの構成要素(ピア)として動作させる。資源提供者は、資源を提供する代わりに即売会に参加するサークルの最新情報を検索できるようになる。これにより、検索システムの構築に必要な物的コストをゼロとすることができる。

2. 関連研究

本章では、Web 上の新鮮な情報の検索を目的とした研究について述べる。

2.1 PRSM

PRSM (Public Robot Server Manager) [2]は早稲田大学の村岡らによって 1998 年から 2000 年にかけて研究開発された分散協調型クローラ(文書収集ソフトウェア)で

あり、全てのJPドメインのWebページを24時間以内に収集することを目的としている。

全体を管理するPRSMと、個々のクローラであるPRS (Public Robot Server) から構成される。PRSとWebサーバ間の転送速度が全体の収集時間を決定する大きな要因となることから、PRSMではPRSとWebサーバ間の距離を考慮した分配を行っている。PRSMは、PRSに対して担当Webサーバの分配や、各WebサーバとPRS間の距離計測を指示する。一方、PRSで新規に発見されたWebサーバの情報や、距離計測の結果はPRSMに送られる。このように、PRSはPRSMからの指示に基づいて、互いに重複しないWebサーバを担当してWebページを収集する。収集されたWebページは、最終的にSSS (Search Service Server) に再配布され、検索のための索引作成が行われる。

日本国内の103箇所のWebサーバを対象として行われた収集実験では、7箇所のPRSに無作為に分散することによって、一箇所で集中して収集する場合と比較して2.6～10.6倍の高速化を実現した。また、各PRSの負荷が均一になるように分散させた場合、5.5～22倍の高速化が可能であることを示した。図1にPRSMの概念図を示す。

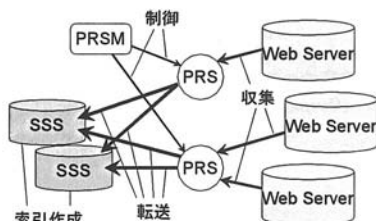


図1 PRSMの概念図

PRSMの例から、クローラを分散させることが文書収集の高速化に極めて有効であることが分かる。しかし、PRSMは文書収集を分散させるためのシステムであり、索引を分散させたまま検索を行うということに関しては考慮されていない。

2.2 CSE

CSE (Cooperative Search Engine : 協調サーチエンジン) [3]は、東洋大学の佐藤らによって研究されている分散サーチエンジンであり、イントラネットにおいて索引の更新間隔を短縮することを目的としている。

検索対象となる文書が格納された各Webサーバに、局所的な文書収集、索引作成、検索を行うサーチエンジンであるLSE (Local Search Engine) を配置し、これらを専用のメタサーチエンジンであるLMSE (Local Meta Search Engine) で統合することにより、1つの大域サーチエンジンとして動作する。どのLSEがどのような情報を保持しているかという情報はLS (Location Server) によって管理される。

LSEはHTTPデーモンを介さずに文書収集を行うため、文書収集にかかる時間を大幅に短縮することができる。また、LSEを配置することができないWebサーバであっても、外部からHTTPデーモンを介して文書を収集することが可能である。これらの機能によって、更新間隔を大幅に短縮することが可能となっている。多くのサーチエンジンでは通信量を抑えるために、検索結果の上位一部のみを検索者に提示し、検索者から要求があった場合にのみ、残

り(の上位一部)を提示するという形をとっており、これを継続検索という。分散サーチエンジンは検索クエリを転送して検索結果を得ているために応答時間が長くなるが、CSEではCache Server (CS) に検索結果をキャッシュさせることによって、継続検索時の応答時間を短くしている。

CSEでは、各LSEがどのような索引を保持しているかという情報をLSが一括管理している。そのため、検索対象とできるWebページの数はLSが管理できる数までであるという制限がある。また、LSに障害が発生するとシステム全体が利用できなくなってしまう。CSEの概要図を図2に示す。

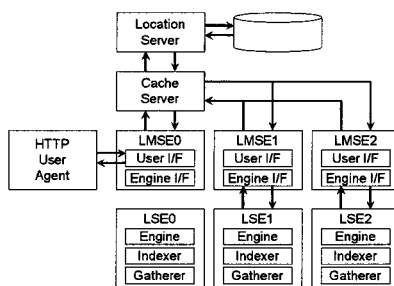


図2 CSEの概要図

2.3 News&Blog Search

1999年頃から、多くのブログ開設用の無償ソフトが登場してきた。それ以来、ブログサイトというサイトの形式が急速に広まった。News&Blog Search[4]は、これらのブログサイトと一般のニュースサイトの最新情報を検索することを目的としたサーチエンジンである。

多くのニュースサイトやブログサイトでは、RSS (Rich Site Summary) やAtomによって記述されたサイトサマリ情報を提供している。サイトサマリ情報とは、そのサイトの概要や各ページのURL、更新時刻等を人間にも計算機にも可読な形式で記述したものである。News&Blog Searchは、これを利用して更新されたWebページのみを収集することにより、文書収集の効率化を図っている。また、特定のサイトのみを文書収集の対象とすることで、さらに文書収集にかかる時間を短縮している。これによって、ニュースサイトは15～30分、ブログサイトは10～15分という短い間隔で索引を更新することを可能としている。

News&Blog SearchではRSSを利用しているため、何らかの理由によってRSSが利用できないサイトについては検索することができない。

3. 想定環境と課題

本システムは、同人誌即売会のための新鮮情報検索に利用されることを想定している。即売会に参加するサークルのサイトの特徴として、以下が考えられる。

- (1) サイト運営者のWebサーバへの権限は限られている
これは、個人によって運営されるサイトの多くが、ISP等によって提供される無料ホームページスペースを利用していることによる。無料で利用できるホームページスペースでは、単純なファイルのアップロードしか行えない、動的なページの生成ができないというように、サイト運営者に制限された権限しか与えられない場合が多い。このため、対象のWebサーバにソフトウェアのインストールを行う等の手順を踏むようなシステムでは、想定される環境で利用することができない。

(2) サイト運営者は専門的知識を持たない

個人サイトの運営者はサイト運営の専門家ではない。そのため、Web サーバやそれに関連する技術について詳しい知識を持たない場合が多い。そのため、サイト上でのサイトサマリ情報の提供等を、サイトの運営者に求めることはできない。

(3) ジャンルとスペースによって分類できる

即売会では、サークルが主に扱うジャンルに基づいて、参加サークルを分類する。また、類似したジャンルを会場の近いスペースに配置する。そのため、ジャンルやスペースによってサークルのサイトを分類することができる。

(4) 対象サイト数は1,000サイト程度である

即売会に参加するサークルの数は、規模によって数十から数万になる。しかし、参加サークル数の多い即売会ほど扱われるジャンルも多い。日本最大規模の同人誌即売会であるコミックマーケットでは参加サークル数 35,000、ジャンル数 43 となっている[5]。したがって、各ジャンルを扱うサークルは1,000 程度と考えられる。そのため、本研究では1,000 程度のサイトを検索できることを目標とする。

また、本システムでは即売会に興味を持つ人によって提供される個人 PC を、システムを構成するピアとして利用する。以下に示す個人 PC の特徴は、システムに影響を与えないと考えられる。

(5) マシンの性能が低い

既存のサーチエンジンで用いられるようなサーバと比較して、個人 PC は性能が低い。そのため、PC が何らかの処理を行う場合にかかる時間が長くなってしまふ。また、個人 PC を利用していることから、必要以上の負荷は提供者の PC 利用に影響を与える可能性がある。したがって、個々の PC が行うべき処理は可能な限り少ないことが要求される。

(6) ネットワークの速度が遅い

一般に、Web で情報を探索際にユーザが我慢できる時間は2秒までとされている[6]。しかし、個人 PC が接続されているインターネット接続回線は、特に上りの速度が遅く、このことが検索時の応答時間を長くしてしまうことが考えられる。分散システムの性質上、個々の PC 間でメッセージをやり取りすることは避けられないが、可能な限り検索時の PC 間メッセージ転送を少なくすることが望ましい。

(7) 常時起動が前提でない

個人 PC は、提供者によって停止される可能性がある。そのため、一部の PC の停止がシステム全体に影響を与えない仕組みを組み込む必要がある。

(8) インターネット側からのアクセスが困難である

個人 PC が設置されている家庭の多くでは、ISP によって割り当てられる IP アドレスは非固定である。そのため、インターネット経由で個人 PC にアクセスするためには、DDNS 等を利用する必要がある。また、複数の PC をインターネットに接続している家庭では、NAT 装置が利用される。この場合、NAT 装置にポートマッピングの設定を行わなくては、個人 PC にインターネット経由でアクセスすることはできない。DDNS の利用には、DDNS クライアントのインストールだけでなく、DDNS サーバとの契約が必要である。また、NAT 装置は機種によって設定の方法が異なる。したがって、これらの作業は定型化できないため、手作業で行う必要がある。一般のインターネット利用者である PC 提供者にこの負担を強いること、提供 PC 数の減少につながってしまう。

4. 提案システム

4.1 システムの概要

本システムは、事前に登録されたサイトを対象として文書収集と索引化を行い、要求に応じて検索を行う、分散検索システムである。本システムの概念図を図3に示す。

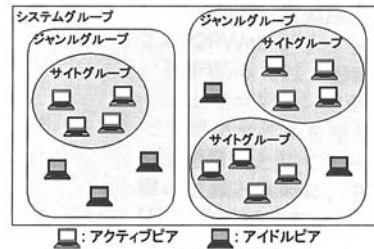


図3 システムの概念図

専用ソフトウェアが動作している個人 PC (提供された計算機資源) は、本システムを構成するピアとなる。各ピアは P2P ネットワークで結ばれている。

システムは階層的なグループ構造を持つ。各グループは、一意なグループ名を持つ。システムグループは全てのピアが属するグループである。ジャンルグループは1つのジャンルの検索を担当するピアの集合である。サイトグループは同じサイト群の検索を担当するピアの集合である。1つのサイトグループは4台のピアによって構成され、それらのピアは常に同期された索引を保持する。そのため、ピアが停止しても、そのピアが保持していた情報が失われることはない。これにより、個人 PC の常時起動を必要条件とせずに済む。サイトグループに属するピアをアクティブピア、属さないピアをアイドルピアと呼ぶ。サイトグループのピア数が4となっているのは、索引の同期にかかる時間を考慮した結果である[7]。以下に、ピアが持つ機能を示す。

4.2 ピア情報の配布

各ピアは、以下の情報を定期的に P2P ネットワーク上に配布する。

- ・自身の ID
- ・自身へのアクセス手段
- ・自身の属するシステムグループ名
- ・自身の属するジャンルグループ名
- ・自身の属するサイトグループ名 (アクティブピアのみ)
- ・自身が担当するサイトの URL (アクティブピアのみ)
- ・過去 48 回の文書収集、索引化、索引の同期にかかった時間の平均 (アクティブピアのみ)

この情報を利用して、各ピアは常に(1)自身が属するサイトグループの全てのピアと(2)自身が属するジャンルグループ内の全てのサイトグループのいずれかのピアと(3)全てのジャンルグループのいずれかのピアにアクセスできる状態を保つ。

4.3 文書収集と索引化 (アクティブピアのみ)

前章の(1)で述べた通り、対象のサイトが存在する Web サーバに特殊なソフトウェアがインストールされていなくても文書収集が可能でなければならない。したがって、本システムは単純な HTTP を用いた文書収集を行う。文書収集の際には、サーバから返される Last-Modified 応答ヘッダフィールドの値を保持しておく。次回以降の収集には、保持しておいた値を If-Modified-Since 要求ヘッダフィールドに指定することで、更新されていない文書は収集

しない。また、文書ファイルのみを収集対象とする。

最新情報を検索するためには、検索対象のサイトに対して高頻度で文書収集を行う必要がある。アクティブピアは、30 分毎に文書収集を行う。News&Blog Search では最低でも 30 分以内に公開された情報を検索できるようにしていることから、本研究でも 30 分を目標とした。収集した文書の索引化は文書収集が終了次第行われる。索引化が終了した後、次節の「索引の複製」を行う。

4.4 索引の同期（アクティブピアのみ）

前節において索引を作成したピアは、同一サイトグループに属するピアに作成した索引を転送することで、索引を同期する。同時に、文書収集にかかった時間も共有し、文書収集のから索引の同期までにかかった時間を保持する。

4.5 存在確認（アクティブピアのみ）

アクティブピアは、同一サイトグループのピアに対して 60 秒間隔で動作確認を行う。ピアの停止を検出した場合、次節の「アイドルピアのアクティブ化」を行う。これにより、サイトグループのピア数を常に 4 に保つ。

4.6 アイドルピアのアクティブピア化（アクティブピアのみ）

存在確認によってサイトグループ内のピアの脱落を検出した場合、そのサイトグループのピアのうち、最も ID が大きいピアが、自身と同じジャンルグループに属するアイドルピアを無作為に 1 つ選び、以下の情報を渡すことでサイトグループに引き入れる。

- ・ 索引
- ・ 検索対象サイトの URL
- ・ サイトグループ名
- ・ 同一サイトグループに属するピアの ID

4.7 検索要求の受け付け

ピアに受け取る検索要求には、「そのピアが保持する索引に関する検索（検索要求 A）」「そのピアが担当するジャンルのみに関する検索（検索要求 B）」「システム全体に対する検索（検索要求 C）」の 3 種類がある。検索要求 C は検索要求 B を、検索要求 B は検索要求 A を処理過程で呼び出す。ユーザが直接ピアに与えるのは検索要求 C のみであり、その他の要求はピアからピアへと与えられる。

各要求を受け取った際のピアの動作を以下に示す。

(a) 検索要求 A（アクティブピアのみ）

検索要求 A は、要求を受け取ったピアが保持する索引から検索した結果を得るための要求である。パラメタとして検索クエリを持つ。この要求を受け取ったピアは、渡された検索クエリを用いて保持している索引に対して検索を行い、検索結果を返す。

(b) 検索要求 B

検索要求 B は、要求を受け取ったピアが担当するジャンルに属するサイトから検索した結果を得るための要求である。パラメタとして検索クエリを持つ。この要求を受け取ったピアは、同じジャンルグループに属する各サイトグループから、ピアを無作為に 1 つずつ選ぶ。選んだピアに対し、受け取った検索クエリをパラメタとして検索要求 A を送り、各ピアからの応答を受け取る。これらの応答をマージしたものを、自身の応答として返す。

(c) 検索要求 C

検索要求 C は、システム全体から検索を行うための要求である。パラメタとして検索クエリとジャンルグループ名を持ち、検索結果を応答として返す。この要求を受け取ったピアは、指定された全てのジャンルグループから、ピア

を無作為に 1 つずつ選ぶ。選んだピアに対し、受け取った検索クエリをパラメタとして検索要求 B を送り、各ピアからの応答を受け取る。受け取った検索結果を更新日時順にマージしたものを応答として返す。

4.8 サイト登録要求の受付

ピアが受け取る登録要求には、「そのピアへの登録（登録要求 A）」「そのピアが属するサイトグループへの登録（登録要求 B）」「そのピアが担当するジャンルへの登録（登録要求 C）」「システムへの登録（登録要求 D）」の 4 種類がある。検索要求と同様に、各要求は階層的に呼び出される。ユーザが直接ピアに与えるのは登録要求 D のみであり、その他の要求はピアからピアへと与えられる。

各要求を受け取った際のピアの動作を以下に示す。

(a) 登録要求 A

登録要求 A は、要求を受け取ったピアにサイトを担当させるための要求である。パラメタとしてサイトのトップページの URL を持つ。本システムでは、以下の条件を満たすページの集合を 1 つのサイトと定義する。

- ・ トップページが存在するディレクトリを示す URL と同じ文字列を、URL の先頭に持つページ
- ・ トップページからハイパリンクを辿って、上の条件を満たしたページのみを経由して到達できるページ

要求を受け取ったピアは、上のように定義されるサイトを担当する。すなわち、次回以降の文書収集および索引化の対象とし、検索要求 A を受け取った際の検索対象とする。

(b) 登録要求 B

登録要求 B は、要求を受け取ったピアが属するサイトグループにサイトを担当させるための要求である。パラメタとしてサイトのトップページの URL を持つ。この要求を受け取ったピアは、同一サイトグループに属する全てのピアに対して、受け取った URL をパラメタとして登録要求 A を送る。

(c) 登録要求 C

登録要求 C は、要求を受け取ったピアが担当するジャンルにサイトを登録するための要求である。パラメタとしてサイトのトップページの URL を持つ。この要求を受け取ったピアは、同じジャンルグループに属する各サイトグループのうち、以下の条件を満たすグループを選ぶ。

- ・ 登録されているサイト数が 10 以下
- ・ 文書収集、索引化、索引の同期にかかった時間の平均が 15 分以内

この条件に適合するサイトグループを無作為に 1 つ選び、そのグループに属するいずれかのピアに対して、受け取った URL をパラメタとして登録要求 B を送る。

上記の条件で、サイト数を 10 以下としているのは、索引化の処理のためにかかる負荷を制限するためである。また、時間を 15 分以内としているのは、ページ数の多いサイトの登録等によって、文書収集にかかる時間が 30 分を越えてしまうことを防ぐためである。

上記の条件に適合するサイトグループが存在しなかった場合、自身と同じジャンルグループに属するアイドルピアを 4 つ選ぶ。それらのピアに、受け取った URL と新たに生成したサイトグループ名、4 つのアイドルピアの ID を渡し、アクティブピア化させる。これにより、新たなサイトグループが構築される。

(d) 登録要求 D

登録要求 D は、サイトをシステムに登録するための要求である。パラメタとしてサイトのトップページの URL

と1つのジャンルグループ名を持つ。この要求を受け取ったピアは、指定されたジャンルグループからピアを無作為に1つ選ぶ。選んだピアに対し、受け取ったURLをパラメータとして登録要求Cを送る。

5. プロトタイプの実装

本章では、前章で述べたシステムのプロトタイプの実装について述べる。図4に、プロトタイプの概略図を示す。

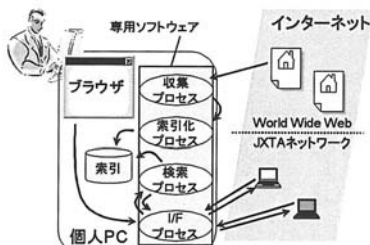


図4 プロトタイプの概略図

個人PCを本システムのピアとするためにインストールする必要がある専用ソフトウェアは、以下の4つのプロセスによって構成される。

(a) 収集プロセス

アクティブピアである場合、担当しているサイトに対して30分毎に文書収集を行う。収集した文書は、索引化プロセスに引き渡される。Javaによって実装している。

(b) 索引化プロセス

収集プロセスから受け取った文書を索引化し、格納する。日本語全文検索システムNamazu[8]を用いている。

(c) 検索プロセス

JXTAプロセスからの要求に応じて、格納された索引を用いて検索を行う。同じく、Namazuを用いている。

(d) I/Fプロセス

利用者からの要求の受け付けおよびP2Pネットワークを介して他のピアとのメッセージのやり取りを行う。利用者からの要求は、Javaで実装したHTTPデーモンが受け取る。これにより、ブラウザを用いたWeb閲覧の延長として本システムを利用することができる。また、P2Pネットワークの実装にはJXTA[9]を用いている。

JXTAでは、ピアはIPアドレスではなくJXTAIDと呼ばれる識別子で区別される。また、アクセス手段はパイプと呼ばれる仮想伝送路によって提供されている。これは、インターネットに直接接続されているピアが、FWやNATで分断されたネットワークに存在するピアに代わってメッセージをリレーすることで、物理ネットワークの環境を隠蔽して直接通信することができるというものである。これにより、非固定IPアドレスやNAT装置を利用している個人PCであっても、特別な設定を必要とせず、本システムに参加することが可能である。

JXTAには、アドバタイズメントと呼ばれるピアが持つリソースを表すメタデータを配布・発見を行う仕組みがある。これには、ピアのIDやパイプの情報だけでなく、任意の情報を記述することができる。これを利用して、各ピアは4.2節で列挙した情報を配布する。

図5に、実際のアドバタイズメントの例（一部省略）を示す。このアドバタイズメントは以下の内容を含む。

- ・システムグループ名は「comio_treasure9」
- ・ジャンルグループ名は「toho」

- ・サイトグループ名は「11253823」
- ・サイトグループにはこのピアのみが存在
- ・サイト「http://www.aoboo.jp/html/tre9/」を担当
- ・文書収集、索引化、索引の同期にかかった時間の平均は129296ミリ秒

```
<?xml version="1.0"?>
<!DOCTYPE jxta:MSA>
<jxta:MSA xmlns:jxta="http://jxta.org">
  <MSIDurn:jxta:uuid-E8BF191D52274EE4B40DB8856A3FB8AD
    DA40FFB38E742AC836DF0683A8F056806</MSID>
  <Name>JXTASPEC:JXTA-HOON_SEARCH:comio_treasure9:toho:
    11253823</Name>
  <jxta:PipeAdvertisement xmlns:jxta="http://jxta.org">
    <Idurn:jxta:uuid-59616261646162614E5047205032503B
      FA137BDF2C452CABCEE309C7985EA04</Id>
    </jxta:PipeAdvertisement>
    <Param type="group_info">
      <name>11253823</name>
      <peers>
        <peerurn:jxta:uuid-E8BF191D52274EE4B40DB8856A3F
          B8A0DA40FFB38E742AC836DF0683A8F056806</peer>
      </peers>
      <target>
        <www.aoboo.jp%2Fhtml%2Ftre9%2F/>
      </target>
      <time>129296</time>
    </Param>
  </jxta:MSA>
```

図5 アドバタイズメントの例

6. 実証実験

プロトタイプに関して、以下の項目を評価するための実験を行った。

- (1) 他の検索システムよりも新しい情報を検索できることによる有効性が得られること
- (2) P2Pの導入によるメリットが示されること

(a) NAT装置等を介していても、システムへの参加手順が煩雑でないこと

(b) 常時起動でなくても動作すること

(1)の評価のために、本年1月に開催された同人誌即売会で頒布された新刊の情報を、イベントの10日前にGoogleとプロトタイプでの両方で検索し、その結果を比較した。検索の手順は以下の通りである。

- ・プロトタイプには、即売会の公式サイトで公開されていた参加サークルのサイト、721サイト分を登録した。
- ・検索クエリは「(イベント名) ∧ 新刊」とした。
- ・Googleの検索結果から、721サイト分のURLのいずれもURLの先頭に持たないページを除外した。
- ・両方の検索結果から、同じサイトのページに関して、最も新刊の情報に近いページ以外を除外した。

以上の手順を経た結果、検索結果はGoogle:20件、プロトタイプ:25件であった。この結果の上位20件のページについて、内容を「新刊の有無と内容が確定」「新刊を発刊予定」「即売会に参加する」の3つに手作業で分類した。前のものほど新しい情報であると言うことができる。

表1に分類結果を示す。

表1 分類結果

内容	Google	本システム
新刊が確定	3	7
新刊を予定	10	9
参加する	7	4

Googleに比べ、本システムのほうがより新しい情報(新刊が確定したという情報)が検索できていることが分かる。気づいた点として、即売会の公式サイトで公開されているURLとは異なるURLを持つブログをサイトの一部と

して公開し、ブログ上で新刊情報を公開しているサークルもあった。検索の手順の途中で除外されてしまったが、これらのページは Google でしか発見できなかった。この問題は、ブログの URL もシステムに登録することで、解消可能であると考えられる。

(2)-(a)について検証するために、PC をシステムに参加させるのに必要な手順を、一昨年の 5 月に行った前回の実験で使用したプロトタイプ[10]と比較した。前回のプロトタイプでは、P2P ネットワークを利用していなかった。比較結果を表 2 に示す。

表 2 比較結果

操作		前回	今回
ソフトウェアのインストール	専用ソフトウェア	要	要
	Namazu	要	要
	JXTA	不要	要
	DDNS クライアント	応必要	不要
ネットワーク周辺の設定	NAT 装置の設定 (ポートマッピング)	応必要	不要
	DDNS の契約	応必要	不要

前回のプロトタイプでは、非固定 IP アドレスを利用している場合には DDNS クライアントのインストールと DDNS の契約が、NAT 装置を利用している場合には NAT 装置の設定が、それぞれ必要であった。第 3 章の(8)に示したとおり、この作業は定型化しづらく、誰でも簡単にシステムに参加するというわけには行かなかった。

今回のプロトタイプでは、上記の設定が不要である代わりに、JXTA のインストールが必要になっている。JXTA のインストールに必要な手順は、JXTA のライブラリファイルをコピーするだけであり、定型化できる。そのため、前回の試作システムと比較して導入の難易度は極めて低い。したがって、(2)-(a)は満たされていると言える。

(2)-(b)について検証するために、研究室内のイントラネット (100Mbps) に接続された 2 台の PC (ピア) でサイトグループを構成し、ピアを交互に脱落・復帰させても、索引が失われないかどうかを確かめた。表 3 に、この検証のために用いた索引の詳細を示す。

表 3 索引の詳細

担当サイト数	1 サイト
サイトのページ数	24 ページ
サイトの容量	255KB
索引ファイル数	65 ファイル
索引の容量	603KB

この条件において、ピアを 5 分毎に脱落・復帰させても、索引は失われず、検索を行うことができた。

ピアがもう一方のピアの停止を検出するまでにかかった時間は 90 秒程度であった。これは、ピア間の生存確認が 60 秒間隔で行われることと、生存確認を行うためのパイプのタイムアウトが 60 秒に設定されていることによる。また、ピアの脱落を検出した後に行われるアイドルピアのアクティブピア化にかかった時間は、10 回の試行の平均が 12.6 秒であった。そのため、ピアの脱落からアイドルピアのアクティブピア化完了までにかかる時間は 2 分弱となり、5 分間隔でピアを脱落・復帰させても索引が失われることがなかったと考えられる。

また、同条件において検索にかかる時間を計測したところ、10 回の試行の平均が 402 ミリ秒であった。これは、Web 上での情報検索の際にユーザが我慢できる時間である 2 秒を大きく下回っている。

7. まとめと今後の課題

本研究では、同人誌即売会等において要求があるにも関わらず、既存のサーチエンジンでは公開されて間もない情報を検索できないことを問題点として捉えた。この問題を解決するために、個人 PC を用いた分散検索システムの研究を行った。このシステムでは、P2P ネットワーク基盤を用いることで、堅牢性の向上、導入の簡便化を図っている。また、実験結果から、既存の商用サーチエンジンと比較して、最新情報の検索に関して有利であることが分かった。

今後検討すべき課題として、以下のものが考えられる。

(1) ピア選択アルゴリズム

処理能力、接続回線、システムからの脱落率等を元に、アクティブピアに適した PC を判断・選択することができれば、システム全体の性能および堅牢性の向上に役立つ。

(2) 分散収集

PRSM が行っているように、サイトグループのピア間で文書収集および索引化を分担することによって、より効率的に索引が更新できると考えられる。

(3) サイトの再振り分け手法

サイトの肥大化等によって、PC 間で文書収集および索引化の負荷に偏りが生じる可能性がある。これを防ぐために、自動的にサイトの再振り分けを行う手法を検討する。

(4) インターネット環境での実験

インターネット環境でシステムを動作させた場合、検索にかかる時間や、アイドルピアのアクティブピア化にかかる時間がどのように変化するかを確かめたい。

参考文献

- [1] 「オタク層」の市場規模推計と実態に関する調査：
www.nri.co.jp/news/2004/040824.html
- [2] 村岡洋一, 山名早人, 田村健人, 河野浩之, 森英雄, 浅井勇夫, 西村英樹, 楠本博之, 篠田洋一: Internet 広域分散協調サーチロボットの研究開発, 第 19 回 IPA 技術発表会, 2000.10
- [3] 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, 2002.2
- [4] 次世代検索サービス News&Blog Search :
blogsearch.drecom.jp/
- [5] コミックマーケット公式サイト :
www.comiket.co.jp/
- [6] Fiona Fui-Hoon Nah : A study on tolerable waiting time: how long are Web users willing to wait?, Behaviour & Information Technology, Vol.23, No.3, 2004.5-6
- [7] 豊田正隆, 山崎賢悟, 勅使河原可海: P2P ネットワークを基盤とした分散 Web 検索システムにおける索引情報の損失防止手法の検討, FIT2006 一般講演論文集第 4 分冊, pp.31-32, 2006.9
- [8] 全文検索システム Namazu : www.namazu.org/
- [9] JXTA Technology : www.sun.com/software/jxta/
- [10] 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, DICOM2005 シンポジウム論文集, pp.469-472, 2005.7