P2Pの真実

Peer to Peer

インターネットのこれからの通信スタイルを変える

川崎裕一
Jnutella.org代表
URL http://www.jnutella.org/

第4回　GoogleとP2P分散検索エンジン

ナプスターから始まったP2P型のファイル共有は、ADSLをはじめとする高速常時接続のインフラの普及で、ここ日本でも爆発的に利用が増えている。すでにインターネットのトラフィックの主役は今後HTTPではなくP2Pになるかもしれないとも言われている。しかし、P2P＝違法ファイル交換という図式から抜け出せないでいると、P2Pの本質を見失うことになり、技術の進化の機会を逃しかねない。
今回は最強の検索エンジンとして君臨するGoogleとGoogleの抱える問題、そしてその問題を異なるアプローチで解決しようとする分散検索エンジンの可能性と課題を考えてみたい。

現時点では最強の検索エンジンGoogle

今インターネットで最も使われている検索エンジンがGoogleであることは誰もが認める事実だろう。Googleは、検索結果がすばやく返ってくること、検索の幅広さ、検索の正確さによって多くのユーザーを獲得し、もはやWWWと同義語になりつつある。それは次のようなデータからも裏付けられる。

- WWW検索要求の75パーセント、1日2億回の検索要求を処理
- 検索エンジンの利用シェアでは55.2パーセントを占める（21.7パーセントで2位のYahoo!はGoogleのエンジンを使用しているので、合計するとGoogleエンジンのシェアは実に76.9パーセントとなる）URL 01
- Googleは、200か国、88の言語で検索される
- インターネット広告の40パーセントがGoogleの扱う広告で費やされる
- Googleの抱える広告主は10万社
- 2002年度の売り上げが4億〜7億ドル（予測）
- 2003年度の売り上げは20億ドルになると言われている

Googleをしてもできないこと

前述したように検索エンジンビジネスで支配的な地位を固めるその一方で、Googleが把握できているドキュメントはウェブに転がる情報のほんの一握りにすぎないのである。

Googleが現在インデックスしているのは30億ページ強だが、ウェブ上には5500億ページものドキュメントが存在すると言われているURL 02。

この大きな隔たりは、Googleの検索スパイダーで情報を集められないドキュメント

が数多く存在することによる。具体的には、次のようなドキュメントは、そもそも検索できない、検索精度が悪い、検索頻度が追いつかないなどのために、Googleが検索を苦手とするドキュメントである。

a. 個人ユーザーのローカルマシンの中のドキュメント全般

Googleがスパイダーを巡らせているウェブサーバーは、基本的に常時接続されていていつでもドキュメントを提供可能な状態になっている。これらに対して、個人ユーザーがISPから割り当てられたIPアドレスとダイナミックDNSを使い、少人数を対象にしているような、極めて限られた目的での情報は、Googleにとって見つけるのは極めて苦手だ。また当然と言えば当然だが、インターネットにつながっているけれどもローカルマシンのドキュメントを公開する設定にしていなければそもそも検索対象にはできない。

b. 映像、音楽など、メタデータがドキュメントの意味を効果的に示さないドキュメント

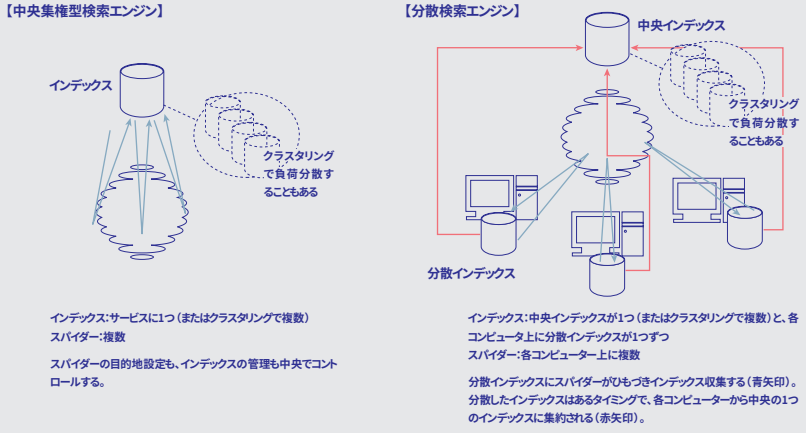
これに関しては検索の仕組み自体に課題が存在する。たとえば60分のMPEGの動画の中で、ある野球選手がホームランを打っているシーンを見つけ出すというのは非常に困難である。ここでは、「ある選手」「ホームランを打っている映像」「ホームランが打たれた時間」などの映像情報がテキスト形式で記述されていなければ、Googleで行われているようなキーワードによる検索では引っかからない。

本当に検索精度を上げるのであれば、セマンティックウェブで提案されているような「意味情報」などを組み合わせなければいけない。

c. Blogのような更新頻度の高いドキュメントや、あまり有名ではない個人のウェブサイト内のドキュメント

これに関しては、GoogleのPyra Labs買収URL03がポイントとなるだろう。RDF情報をGoogleの検索技術と組み合わせることで更新頻度の高いBlogのような情報に対する検索精度の向上を図っていくものと思われる。2003年5月9日には、GoogleのCEOであるEric SchmidtがJP Morgan Technology and Telecom conferenceの中でBlogをGoogleで検索できるようにすると発言しているURL04ということなので、実現の日も近いだろう。

検索エンジンの情報収集とインデックス作成



cに関してはともかく、aやbに関してはGoogleのアプローチでは根本的な問題解決は図れないのかもしれない。そこで、その他の検索アプローチが必要となってくる。

その1つのアプローチがP2Pを利用した分散検索エンジンである。

新しいアプローチ:P2P分散検索エンジン

分散検索エンジンを一言で表すなら、昨今話題となっている分散コンピューティングやグリッドコンピューティングの分散処理の考え方を検索エンジンのインデックス作成に利用するというものである。前述したGoogleに代表される従来の中央集権型検索エンジンでは、スパイダー(またはロボットやクローラー)と呼ばれるプログラムリンクをたどってウェブ上のドキュメントを巡回し、その結果が中央のインデックスにまとめられるという手順を繰り返している。ユーザーは検索を行う際に、このインデックスを検索し、そのインデックスに書かれているURLからウェブに飛んでいくというわけである。

分散検索エンジンのアプローチでも、基本的な検索エンジンの仕組みとして検索対象のインデックスを作り出すことには変わりはない。変わるのは、そのインデックスの作成方法とインデックスの集め方である。

分散検索エンジンでは、各ピアが各々のスパイダーを持ち、そのピアスパイダーは、ウェブを回って任意に決められた自分の検索範囲や検索のロード(負荷)に基づいて、インデックスを拾ってくる。そしてそれぞれのスパイダーが、自分が所属するピアの中にインデックスを作り出す。これが分散インデックスである。この分散インデックスを定期的に1つにまとめることで、1つの大きなインデックスが作り出されることになる。

分散検索エンジンは もう動き出している

GrubはP2P分散検索エンジンのオープンソースプロジェクトだが、LookSmartがその技術を2003年1月に130万ドル相当の現金と株式で買収したURL05。MSNなどにディレクトリーサービスを提供していたLookSmartがGrubを買収したことにより、その検索サービスに関する事業戦略がどのように変化するかは注目に値する。

多くの検索エンジンでは、蓄積されているインデックス上の各ドキュメントは1か月に1回ほどしか情報が更新されない。これは主にコンピュータリソースの限界によると言われている。しかし、P2P技術を利用すれば、協力的なエンドユーザーのコンピュータリソースを使ってもっと頻繁にインデックス上の各ドキュメントの情報を更新できるのだ。

もちろんGrubのほかにも、さまざまなプロジェクトによってP2Pによる分散検索エンジンの可能性が研究されている。また直接検索には関係はないが、Googleが、ユーザーが分散コンピューティングに参加できる機能をGoogle Toolbarに搭載していることは興味深い。

分散検索エンジンの 現状と課題、今後

さまざまな分散検索エンジンが主に米国を中心として提案、実装されつつある中で、これらを評価する研究論文も徐々に発表されてきている。UCバークレーの研究者により行われた分散検索エンジンの実現可能性に関する論文である「On the Feasibility of Peer-to-Peer Web Indexing and Search」URL06から、分散検索エンジンの現状と課題、今後を紐解きたい。

この論文の中で、研究者は当時Googleが持っていたインデックスよりもやや多い30億ページを持つインデックスを分散検索エンジンで実現することを前提に置いている。

60テラバイトの分散ハッシュテーブル

分散検索エンジンの基盤となる技術が分散ハッシュテーブル(DHT:Distributed Hash Table)である。DHTは、従来のハッシュテーブルを分散させたものであり、代表的なものとしてMITのChord URL07や、NYUのKademlia URL08などがある(ハッシュテーブル、DHTに関しては本稿では詳細な解説は行わない)。

このDHTを分散インデックスとして用いた分散検索エンジンで、30億のウェブページのインデックスを管理するには、分散検索ネットワーク全体として60テラバイトのディスクスペースが必要とされる。そしてこれは、分散検索ネットワークに参加する各PCが自分のディスクスペースを1Gバイトずつ提供しても、6万台のPCが協力する必要がある計算である。

検索を支える帯域幅

次にこの分散検索ネットワークに必要となる帯域幅を考えてみる。1999年度の米国のインターネットを支えるバックボーンが100ギガビットであった。仮に分散検索エンジンにおいて1秒間に1000クエリーの検索命令が行われ、これが前述したバックボーンの10パーセントを使うとすると、この検索ネットワークを支えるクエリー当たりの通信コストは約1メガバイトになる。

分散検索エンジンには、2つの検索手法がある。1つはGnutellaやKaZaAが実装している文書単位分割(Partition by Document)である。それは、「MP3」や「MPEG」などのファイルタイプを重視する検索である。もう1つはGoogleのような

分散検索エンジンへの発展が期待されるLookSmart

Grubでは常に1000台以上のP2Pクライアントが稼働している

キーワードで検索できるキーワード単位分割（Partition by Keyword）である。同論文によると、文書単位分割の場合、6万台のピアに100バイトのクエリーを投げるとすると6メガバイトが必要となる。これは先の仮説である米国インターネットのバックボーンの10パーセントを用いた際のさらに6倍が必要となる。一方でキーワード単位分割の場合、530メガバイト～4000メガバイトが必要になるということである。これは530倍から4000倍ということなる。

この帯域コストは非常に高く、今すぐに実用に堪えられるかというと疑問だ。しかし、検索の最適化を行うために、キャッシング、予測計算、圧縮、クラスタリングなどを行うことでこの帯域コストは圧縮できる可能性があるため、研究の進捗によって大きな改善が見込める可能性がある。

段階的に進む分散検索エンジン

いずれにせよ、インターネットどころか、ウェブの規模から見ても、P2Pネットワークはまだまだ小さいものである。最大のネットワークであるKaZaAのP2Pネットワークでも、ユーザー数は437万人にすぎない（2003年5月9日現在）。

LookSmartは、Googleなどの検索エンジンと真っ向勝負というよりは、これらに対してGrubを通じて作り出された安価、かつ更新頻度が高い新鮮なインデックスを提供し、補完する役割を果たそうとしているように見える。つまり、既存の検索エンジンを裏で支える負荷分散ソリューションを提供するビジネスモデルである。これによって、LookSmartはGoogleを強大な競争相手とするのではなく、ソリューション提供先としてGoogleから収益を上げることが可能になるのだ。

そうではなく既存の検索エンジンに真っ向から挑戦を挑むならば、ビジネスとしてはGoogleが現在大きな収益を上げているキーワード広告モデルを踏襲するだろう。実際、Altnetでは「Top Search」というキーワード広告をKaZaA上で展開している URL09。KaZaAに代表されるファイル共有アプリケーションでは、ユーザーは求めるファイルを検索する。その検索結果画面の上位に広告主のコンテンツを配置し、ユーザーのコンテンツ購入やコンテンツサイトへのトラフィック流入を発生させるわけである。ファイル共有アプリケーションでは、コンテンツへの欲望が非常に強く、コンテンツをダウンロードするためにリンクをクリックする率もGoogleと同じか、それ以上と想定されるために、高い広告効果が期待できる。

Googleを超えるP2P分散検索エンジンが生まれ、他のP2P事業分野よりも堅牢なビジネスモデルを持って登場する可能性もないとは言えないのではなかろうか。

P2P分散検索のプロジェクト

ジョージア工科大学大学院生のTodd Miller氏によるプロジェクト「Hyperbee」	URL http://www.hyperbee.com/
Jnutella.orgのコアメンバーでもあるSam Josephが運営する「Neurogrid」	URL http://www.neurogrid.net/
プリンストン大学の「FASD」	URL http://cse.ogi.edu/~krasic/cse585/kronfol_final_thesis.pdf
JXTAを用いている「Anthill」	URL http://www.cs.unibo.it/projects/anthill/
スタンフォード大学の「Routing Indices」	URL http://dbpubs.stanford.edu:8090/pub/2001-48
スタンフォード大学の「HyperCup」	URL http://www-db.stanford.edu/~schloss/hypercup/
LimeWireで提案されている「Query Routing」	URL http://www.limewire.com/developer/query_routing/keyword%20routing.htm
GPLでソフトウェアが公開されている「ALPINE Network」	URL http://www.cubicmetercrystal.com/alpine/
ラドガーズ大学の「PlanetP」	URL http://www.panic-lab.rutgers.edu/Research/planetp/

検索エンジンの利用シェアで、Googleが圧倒的首位に

URL01 http://internet.watch.impress.co.jp/www/article/2003/0514/one.htm

JEP: The Deep Web

URL02 http://www.press.umich.edu/jep/07-01/bergman.html

Googleが“Blog”作成ツール開発元のPyra Labsを買収

URL03 http://internet.watch.impress.co.jp/www/article/2003/0218/google.htm

The Register: Google to fix blog noise problem

URL04 http://www.theregister.co.uk/content/6/30621.html

Grub's Distributed Web Crawling Project

URL05 http://www.grub.org/

On the Feasibility of Peer-to-Peer Web Indexing and Search

URL06 http://www.cs.berkeley.edu/~boonloo/papers/search_feasibility.pdf

Chord

URL07 http://www.pdos.lcs.mit.edu/chord/

Kademlia

URL08 http://kademlia.scs.cs.nyu.edu/

Altnet

URL09 http://www.altnet.com/

［インターネットマガジン　バックナンバーアーカイブ］ ご利用上の注意

このPDFファイルは、株式会社インプレスR&D（株式会社インプレスから分割）が1994年～2006年まで発行した月刊誌『インターネットマガジン』の誌面をPDF化し、「インターネットマガジン　バックナンバーアーカイブ」として以下のウェブサイト「All-in-One INTERNET magazine 2.0」で公開しているものです。

http://i.impressRD.jp/bn

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、URL、団体・企業名、商品名、価格、プレゼント募集、アンケートなど）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真の撮影者、イラストの作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は収録されていない場合があります。
- このファイルやその内容を改変したり、商用を目的として再利用することはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用する際は、出典として媒体名および月号、該当ページ番号、発行元（株式会社インプレスR&D）、コピーライトなどの情報をご明記ください。
- オリジナルの雑誌の発行時点では、株式会社インプレス R&D（当時は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

このファイルに関するお問い合わせ先
株式会社インプレスR&D
All-in-One INTERNET magazine 編集部
im-info@impress.co.jp