インターネット数理科学第12回

～ネットワークのあちら側を支える数理科学その3～

2007年1月18日

株式会社インターネット総合研究所代表取締役所長
東京大学大学院数理科学研究科客員教授

藤原　洋

目　　次

1．ネットワークあちら側を支える数理科学とは？

現在のインターネット数理科学の課題

③（ネットワークの）あちら側

⇒「グラフ理論」「金融工学理論」に基づくデータベース、検索
エンジン最適化、検索連動データベース、ネット金融サービス

①ネットワークそのもの

⇒「グラフ理論」による動的ルーティング、帯域制御、放送型ルーティング
「ディジタル信号処理理論」に基づく変復調技術

②（ネットワークの）こちら側

⇒「ディジタル信号処理理論」に基づくコンテンツ符号化技術

今後の「インターネット数理科学」の講義の進め方

IRI
Ubilabs Innovations

以下の3つの分野にわたって①②③⇒①②③⇒・・順に

③ネットワークのあちら側を支える数理科学

⇒「グラフ理論」「金融工学理論」に基づくデータベース、検索
エンジン最適化、検索連動データベース、ネット金融サービス

①ネットワークそのものを支える数理科学

⇒「グラフ理論」による動的ルーティング、帯域制御、放送型ルーティング
「ディジタル信号処理理論」に基づく変復調技術

②ネットワークのこちら側を支える数理科学

⇒「ディジタル信号処理理論」に基づくコンテンツ符号化技術

インターネットにおける技術革新と最先端のテーマ

IRI Ubilabs Innovations

③（ネットワークの）あちら側

Web1.0（ﾎﾟｰﾀﾙ）⇒Web2.0（ﾛﾝｸﾞﾃｰﾙ）⇒ **WebX.0**

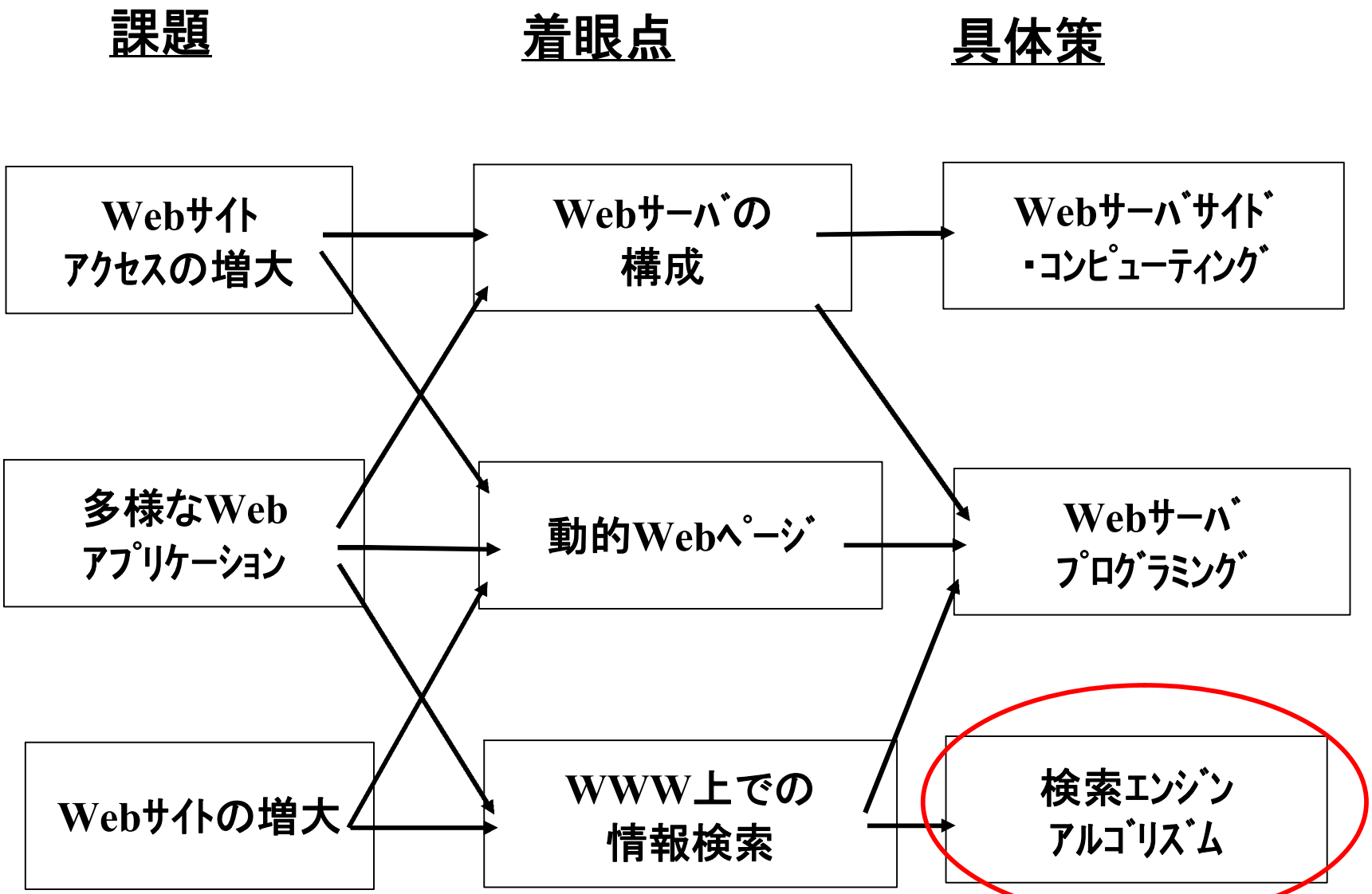
①ネットワークそのもの

ﾀﾞｲﾔﾙｱｯﾌﾟ/2Gﾓﾊﾞｲﾙ ⇒ﾌﾞﾛｰﾄﾞﾊﾞﾝﾄﾞ/3Gﾓﾊﾞｲﾙ⇒ **IP放送/NGN/WiMAX**

②（ネットワークの）こちら側

文字情報（Eメール）⇒ HTML（ﾌﾞﾗｳｻﾞ）⇒ **動画（ﾃﾞｨｼﾞﾀﾙ符号変換）**

ポータル・SNS・ロングテールなどに潜在する数理科学とは？



2. Webの登場で起こったあちら側の変化とは？

Webが登場した結果、ネットワークのあちら側の変化とは？

Web上で、キーワードから所望のURLを探し出す様々な「検索エンジン」が出現、勝ち組みが、スタンフォード大学院生だったジェリー・ヤンとデビット・ファイロが95年に設立したYahoo!社でした。このYahoo!に代表される検索エンジンを用いたディレクトリ（検索）・サービスというビジネス形態が生まれました。同社は、ポータル（玄関口）サービスへと発展しましたが、検索エンジンのアルゴリズムを徹底的に追及したのがスタンフォード大学の大学院生だったラリー・ページとセルゲイ・ブリンで、リンク数の統計から優先順位をつけて自動的に探し出すクローラー（ロボット）型検索エンジンの仕組みを作りGoogle社を98年に設立しました。

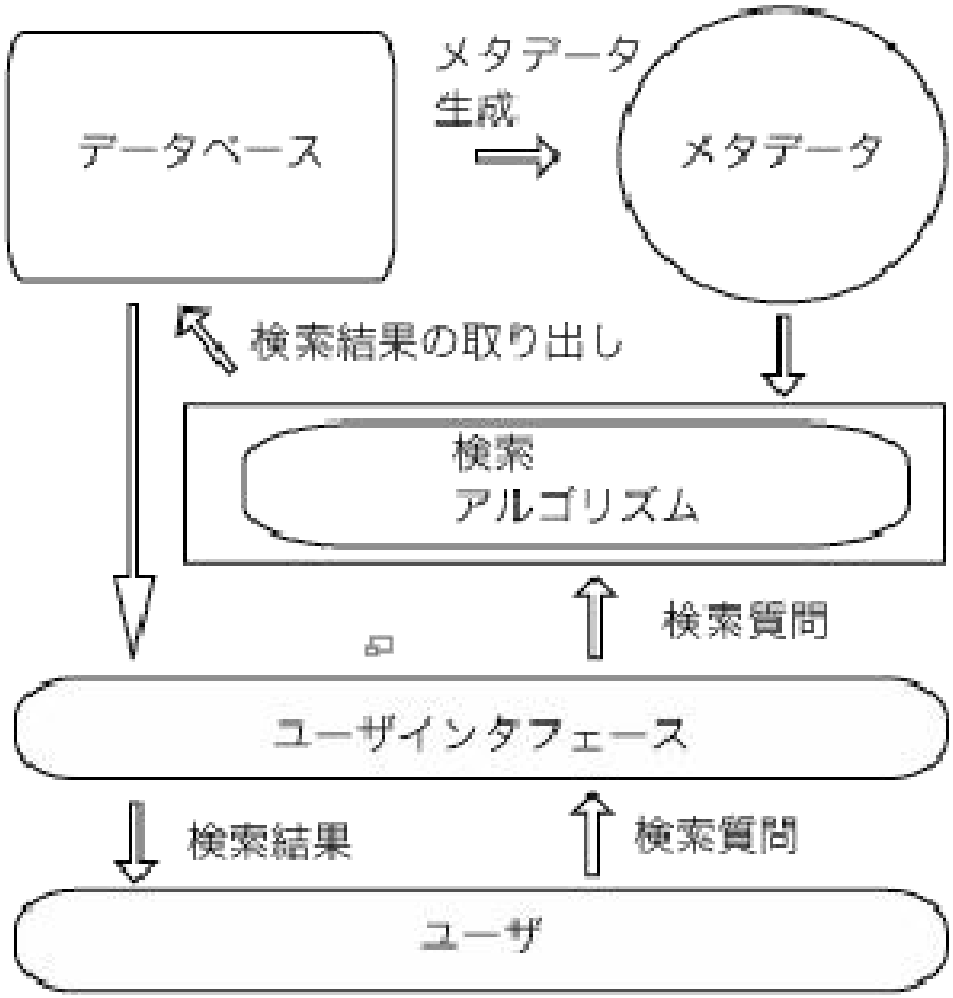
私自身は、96年にIRIを設立し、同社に注目はしていましたが、所詮Googleは、優れた検索エンジンでしかなく、Yahoo!と比較して、ビジネスとしては、難しいだろうとみていました。一方、別の技術をもったOverture社（設立時GoTo.com、現在はYahoo!の完全子会社）が、97年9月にIdealab社のビル・グロスによって設立、広告主が関連性のある特定キーワードに入札し、それが検索結果に表示を可能にするPay-For- Performance検索サービス（スポンサードサーチ）を開始しました。Googleは、2002年から高額の特許料を払い、同技術を取り入れ、「検索連動型広告」を中心にビジネスを組み立て直し一気に高収益企業となりました。Yahoo!とGoogleは、昨年地上波4大ネットワークの広告収入を上回るまでになりました。今後さらにポータルと検索エンジンが更に進化し、巨大なメディアになると思われます。

検索エンジンとウェブディレクトリの出現により、WWW は徐々にその真価を発揮し始める。数学的な理論に基礎付けられたウェブページの順位決定法を実用化することによって、検索エンジンの首座は、一気呵成に確定した。それとは対照的に、すべての分野に亘って個々の事例の集積を要するウェブディレクトリの作成は、継続的で地道な作業によって成し遂げられる辞書の編纂と似ている。前者が数学的手法に依存しているのに対し、後者は分類学的手法によっている点が対照的である。

検索エンジンとは、インターネットに存在する情報（ウェブページ、ウェブサイト、画像ファイル、ネットニュースなど）を検索する機能を提供するサーバーやシステムの総称である。インターネットの普及初期には、検索エンジンとしての機能のみを提供していたウェブサイトそのものを検索エンジンと呼んだが、現在では様々なサービスが加わったポータルサイト化が進んだため、検索エンジンをサービスの一つとして提供するウェブサイトを単に検索エンジンと呼ぶことはなくなっている。

検索エンジンには、ロボット型検索エンジン、ディレクトリ型検索エンジン、メタ検索エンジンなどに分類される。

3. 検索エンジン



情報検索システムの構築法

情報検索システムの構築は以下の各フェーズによって実行する。

①検索対象データの収集

検索対象データの収集方針の決定が重要。
World Wide Web上のハイパーテキストを収集して対象とする場合にはクローラ(ロボット、スパイダー等)を用いて自動収集するのが一般的。
Web上には、膨大なデータが存在し、データ自身が急激に変化するため、網羅的に収集することは困難。
そのため、いかにして多くの対象のデータを収集するかが 重要課題となっており、World Wide Web検索エンジンのサービスでは何ページのデータか検索が可能であるかが重要な 性能指標となっている。

②検索対象データからメタデータを作成

メタデータの形式および作成方法は、データベースの構造、検索アルゴリズム、およびデータ収集方針との関連性深い。
データ収集を広範に継続的に行うような場合、人海戦術によるメタデータ作成は、コストの大幅増大を招く。

③検索アルゴリズムの設計

作成したメタデータを用いてどのような計算によって、データを出力するか決定する。

検索エンジン（Searching Engine）

IRI Ubilabs Innovations

*狭義*にはインターネット上に存在する情報（ウェブページ、ウェブサイト、ニュースなど）を検索する機能の総称で、主として、Webサーバのソフトウェアとそれを支援するWebブラウザのソフトウで実現される。ロボット型検索エンジン、ディレクトリ型検索エンジン、メタ検索エンジンなどに分類される。インターネットの普及初期には、検索エンジンとしての機能のみを提供していたウェブサイトそのものを検索エンジンと呼んだが、現在では様々なサービスが加わったポータルサイト化が進んでいる。

*広義*には、インターネットに限定せず情報を検索するシステム全般を含む。広義の検索エンジンとしては、テキスト情報の全文検索機能を備えた全文検索システム、全文検索ではないシステム等がある。

①ロボット型検索エンジン

IRI
Ubilabs Innovations

●所定の検索アルゴリズムに従って、Webページ等を検索するサーバ、システムのこと。検索アルゴリズムは、最も単純な場合はキーワードとなる文字列のみであるが、複数のキーワードにANDやOR等のブーリアン論理式を組み合わせて指定することが多い。

●ロボット型検索エンジンの大きな特徴は、クローラ(スパイダー)を用いることにある。このクローラ機能により、Web上にある多数の情報を効率よく収集することができため、大規模検索エンジンでは、数十億ページ以上のページからの検索が可能である。

●ページ収集情報は、事前に解析し、索引情報(インデックス)を作成する。英語とは異なり、日本語などの言語では、自然言語処理機能によって生成される索引の性能が決まる。このため、多言語対応検索エンジンが今後重要となってくる。

①ロボット型検索エンジン

●検索結果の表示順は、検索エンジンにとって最も重要である。ユーザーが期待したページの検索結果の上位に表示することが重要であるため、多くの検索エンジンが、表示順を決定するアルゴリズムを非公開にし、その性能が競争状態となっている。

●検索エンジン最適化業者（SEO）の存在が、アルゴリズムの非公開要因である。

●Googleは、アルゴリズムの一部としてPageRankを公開しているが、多くの部分が非公開。

● Webページの更新時刻情報によって新しい情報に限定して検索できるものや、検索結果をカテゴリ化して表示するものなど、特長のある機能を搭載しているものもある。

●Google, Yahoo!, infoseek, Technorati,MARSFLAG, Altavista, AlltheWeb, Teoma, WiseNut, Inktomiなどがロボット型検索エンジンを利用している。

②ディレクトリ型検索エンジン

IRI
Ubilabs Innovations

●人手で構築したWebディレクトリ内を検索する検索エンジン。

●人手で構築しているため、質の高いWebサイトを検索可能。

●サイトの概要を人手で記入しており、検索結果から目的のサイトを探し易い。

●人手入力のため、検索対象となるサイト数が少ない。

●WWWの爆発的な拡大から、全Webサイトを即時にディレクトリへ反映させることが困難になり、現在では非主流。

●ディレクトリ型検索エンジンでは、ヒットするサイトがない場合、 ロボット型検索エンジンを用いる併用型が多い。

●Yahoo!, Lycos, Open Directory Project, LookSmartなど。

③メタ検索エンジン

IRI
Ubilabs Innovations

●入力されたキーワードを複数の検索エンジンに送信し、得られた結果を表示する検索エンジン。

●メタサーチエンジン、横断検索エンジンとも呼ぶ。

●“meta”とはこの場合、“beyond”（横断）の意味。

●検索毎にエンジンを使用するかを選択する「非統合型」と、検索結果を独自のアルゴリズムで総合的に判断して一つの結果として出力する「統合型」とがある。

●統合型では結果表示に広告表示が出来ないため、Google のようにメタ検索エンジンでの利用を禁止している場合もある。

検索エンジン（広義）

●全文検索エンジン
　与えられた文書群から、検索式（キーワードなど）による全文検索機能を提供するソフトウェア、システムの総称。

●非全文検索エンジン

●Webサーバに組み込んで利用される場合が多い。

●スタンドアローン環境で用いられる。特に「デスクトップ検索」と呼ばれる。

●Namazu（日本語全文検索システム）やOracle Secure Enterprise Search等。

検索エンジンの歴史その1

IRI
Ubilabs Innovations

1995 年

アメリカでヤフーとアマゾンが設立。日本のインターネット元年。検索サービスは、登録型のディレクトリサービスが中心で、アメリカで 12 月 Altavista が登場、フルテキストサーチが登場。

1996 年

日本でヤフーがサービスを開始した年だがインターネットの検索市場に動きが少ない。検索エンジンはインフォシークが登場。徐々にロボット型の検索エンジンが登場。一般人はヤフー、通は Altavista や infoseek 等の海外のロボット検索エンジンを使用。

1997 年

この年、goo がサービスを開始、同時に楽天もサービスを開始。検索のテクニックを競う「検索の鉄人」が開催。当時、検索結果に思うような結果を出すのは難しかった。鉄人大会の委員長は舛添要一氏。

検索エンジンの歴史その2

IRI Ubilabs Innovations

1998 年
スタンフォード大学の二人（ｻﾘｹﾞｲ・ﾌﾞﾘﾝとﾗﾘｰ・ﾍﾟｰｼﾞ）が Google を設立。日本で、ヤフー、MSN、インフォシーク、goo、エキサイト、ライコス、フレッシュアイが登場、90 年代後半のポータルサイトのメジャープレイヤーへ。ヤフージャパンgooと検索エンジン提携。

1999 年
2000 年問題。i モード開始。検索エンジン企業の買収活発化。ネットバブル。

2000 年
ネットバブルの頂点。Google が台頭。米ヤフーが Google を採用し、日本語サービスも開始。Google の登場でロボット検索の性能が飛躍的に向上。

検索エンジンの歴史その3

IRI
Ubilabs Innovations

2001 年
日本市場でGoogleが台頭 。ヤフージャパンの検索は、goo から Google へ転換し、Google からのトラフィックが急増。ヤフー BB 開始でブロードバンド時代へ。

2002 年
Google の AdWords 広告と Overture の Pay for Performance が日本で開始。IT不況。

2003 年
goo、infoseek が検索エンジンとして Google を採用。日本の検索エンジン消滅。SEO サービスが本格化。

検索エンジンの歴史その4

IRI Ubilabs Innovations

2004 年
Google株式市場に上場。マイクロソフトが新検索サービス発表。米国に続き、日本ではヤフーとGoogleの提携が解消。

2005 年
Yahoo!とMSNが独自の検索エンジンを採用。各検索エンジンが、多様なサービスを開始し、地図検索、パーソナライズ、デスクトップ検索などへ拡大。ブログ検索やモバイル検索エンジンなどの新サービス開始。モバイルでのインターネット利用者がパソコン利用を凌駕。

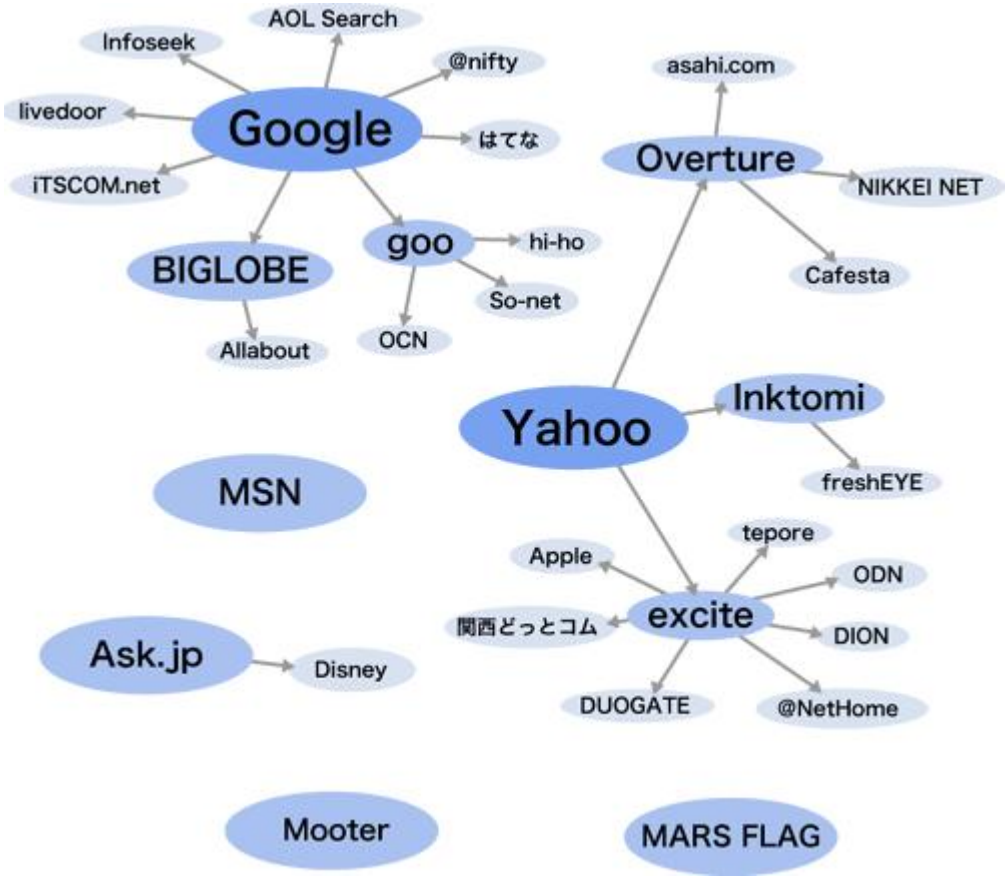
2006 年
不自然な外部リンクに対して、Googleが厳格処置。
携帯の検索サービス登場。NTTドコモ:複数エンジン、au:Google、ソフトバンク:Yahoo!。

調査方法

2004 年 6 月に実施。人気検索フレーズ 1,000 種類を選びだし、Google 及び Yahoo！Japanで検索、検索ワードごとに上位 50 位（10 件表示で 5 ページ目まで）までを解析して様々な要素にデータ化。検索フレーズによっては、リストされるURLが 50 に満たない場合もあり、実際にサンプルとなった URL の総数はGoogleが 49,964 、YSTが 49,923 。

	Google	ヤフージャパン
タイトルの長さの平均値	18.4 文字	18.9 文字
ドキュメントサイズの平均値（mean）	20.2k	27.2k
ドキュメントサイズの最頻値（mode）	2k	2k
URLの長さの平均値	34.2 文字	34.6 文字
URLの長さの最頻値	17 文字	17 文字
URLに含まれた「/」の平均値	2.3	2.3
URLに含まれた「/」の最頻値	1	1
URLに含まれた「~」の総数	4,091 (8.2%)	4,793 (9.6%)
「?」を含んだURLの総数	3,848 (7.7%)	4,126 (8.3%)
“session”を含んだURLの総数	3 (0.01%)	10 (0.02%)
“id=”を含んだURLの総数	1,155 (2.3%)	1,137 (2.3%)
キーワード広告の出現率	43.9% (439/1000)	47.1% (471/1000)
Yahoo!カテゴリへの総登録数	-	14,901 (29.8%)

日本語のロボット検索エンジン相関図



最終更新日 2006 年 8 月 21 日（調査ECジャパン）

使用言語からみたインターネット人口の割合

Internet Societyによればインターネットで用いられている言語のうち英語が占める割合は85%とされていたが、その後の進歩や各国のインターネットの普及により多言語化が進み、上表に見られるように2000年の年末には英語と非英語の言語人口が逆転し、その傾向は継続。

検索エンジンの新たな課題！

	1998年	1999年	2000年		2001年			2002年		2003年	2004年
	12月	1月	4 - 7月	12月	2月	4 - 6月	7月	1月	6 - 10月	2 - 4月	7月
英語	58%	55%	51.3%	49.6%	47.6%	47.5%	45.0%	43.0%	40.2%	36.5%	35.8%
非英語	42%	45%	48.7%	50.4%	52.4%	52.5%	55%	57.0%	59.8%	63.5%	64.2%

検索エンジンの新たな課題！

IRI
Ubilabs Innovations

●WWW検索エンジンの代表Googleでは100億を超えるWebページが登録。

●検索エンジンの利用者は、容易に検索することが困難に。

●日本語入力機能のないコンピュータを用いて日本語サイト検索は、困難。

●非英語圏の言語間の検索は中間に翻訳エンジンがないと困難に。

●インターネットの多言語化が今後も増加し、言語間障壁の克服が課題。

検索エンジンの社会的問題

IRI
Ubilabs Innovations

●膨大な情報を網羅的に調査するには有力検索エンジンを利用するしかない。
●URLがあまり知られていないサイトやドキュメントなどは検索エンジンに検索結果として表示されなければ検索可能性が著しく低下。
●表示されなくなる基準は露骨な検索エンジン最適化テクニックを使用しているサイトや各国の法律等に反しているサイトなどと考えられているが、その明確な基準は各社共に不明確で検索結果から削除される際の該当ウェブサイトへの警告はない。
●日本の有力検索エンジンは4社あり、対策としては複数の検索エンジンを使い分け検索結果の多様性を確保する方策あり。
●検索エンジンを利用したストーカー行為の事例も発生。個人の氏名で検索すると非常に詳細な個人情報が取得できるケースもあるが、個人情報の削除要請に対し検索エンジン各社は、元ページの作成者に一切の責任があるとして、応じない方針。
●中国の検索エンジンでは反政府的な内容や政府が弾圧しているといわれる宗教団体に関する情報は検索結果に表示されない。
●Googleなどは検索結果の中に「表示されている内容は一部法律に基づいて省略されている」という記述があるが、結果的に中国政府の言論弾圧に手を貸しているという批判がある。

4. P2Pの世界とは？

P2Pとは？

IRI
Ubilabs Innovations

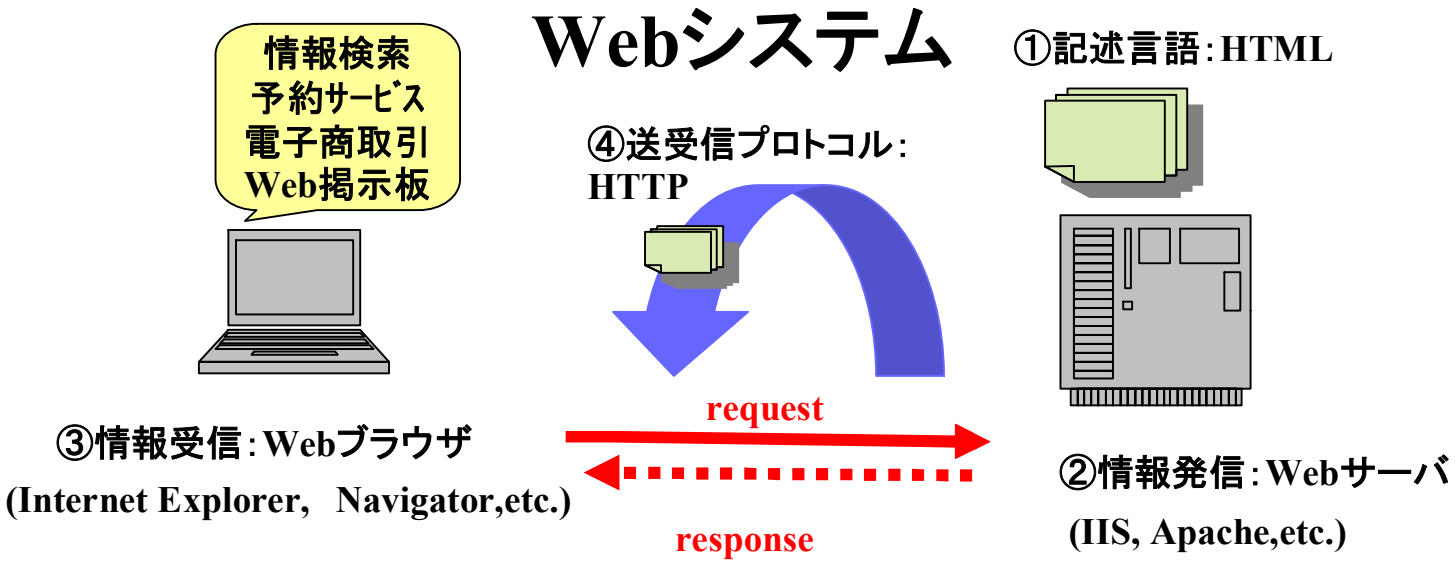
P2P＝Peer to Peer

Peer とは同等の人、対等の人、同僚、友人、仲間

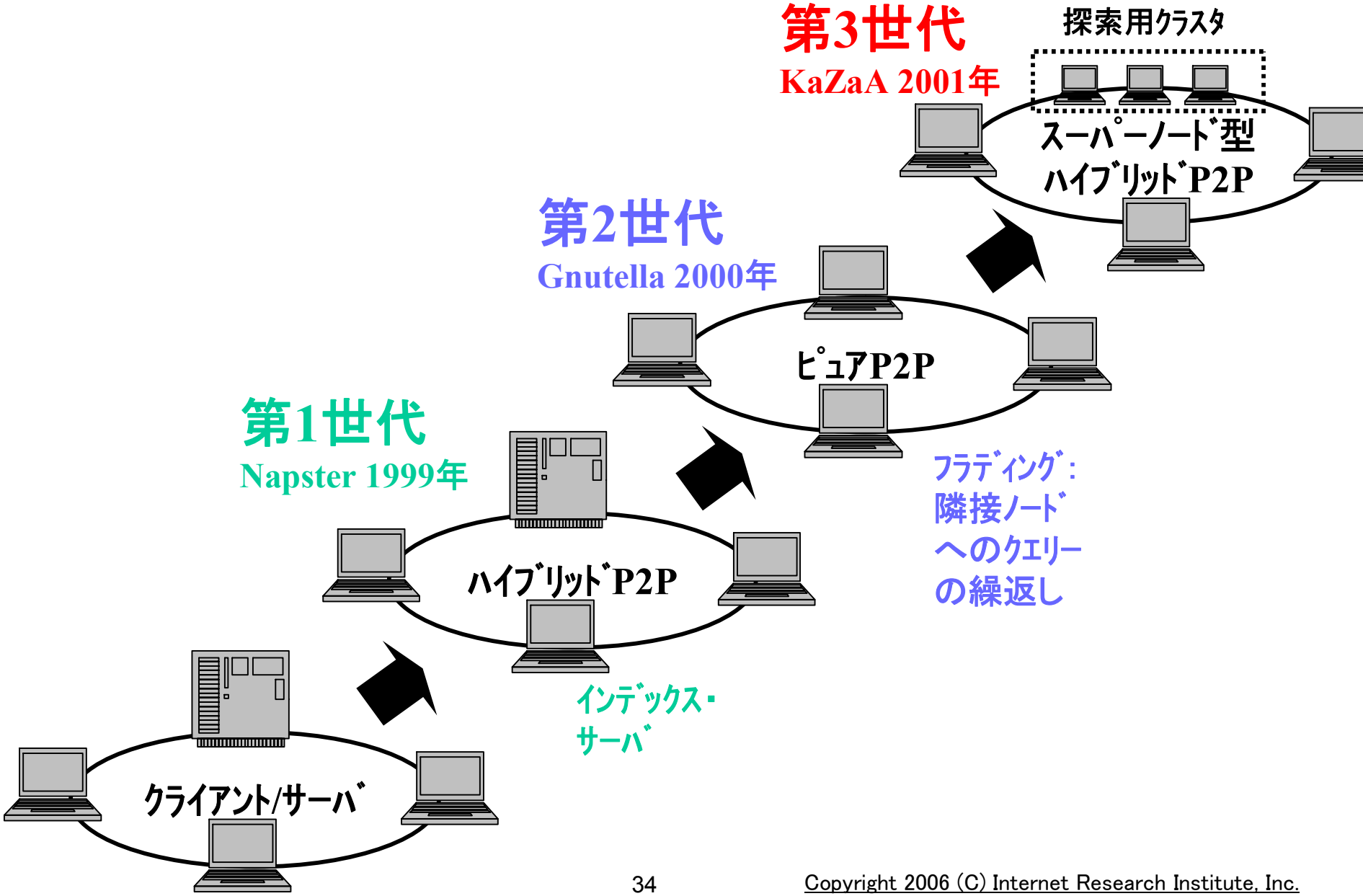
対立概念：　P2Pシステム　⇔ｸﾗｲｱﾝﾄ/ｻｰﾊﾞｼｽﾃﾑ

ｸﾗｲｱﾝﾄ/ｻｰﾊﾞｼｽﾃﾑの例＝World Wide Web

WebとP2Pとの相違



P2Pの発展経緯



P2P技術の概要

IRI
Ubilabs Innovations

1．ハイブリッドP2P（Napster：音楽MP3ファイル交換訴訟に）

・データの場所を探索するインデックス・サーバ

・データアクセスはサーバに集中しない

・Napster：　Shawn Fanning(Northeastern Univ.学生)が開発

2．ピュアP2P（Gnutella：米Nullsoft社、サービス主体不明で未訴訟）

・データの場所を探索するフラディング技術

・隣接ノードへ探索クエリーを発行（TTL：Time to Live/ Gnutella＝7）

・ Gnutella：　Justin Frankel とTom Perpper

3．スーパーノード型ハイブリッドP2P（蘭FastTrack社）

・データの場所を探索するスーパーノード・クラスター

・一般/スーパー・ノード数割合＝一定

・ライセンスビジネス⇒Skypeが利用

P2Pの特長

IRI
Ubilabs Innovations

1. 冗長性： 全ノードがバックアップ機能）

2. 拡張性： アクセス集中がない

3. オーバーレイ機能： セグメント境界の意識不要

4. 非同期アクセス機能： ローカル・データ処理機能

5. オフラインアクセス機能： 同上

6. アドホック構成： 参加者同士の合意で参加が成立

P2Pを支えるオーバーレイ・ネットワーク技術

IRI Ubilabs Innovations

1. 下位のネットワークレイヤ（IP）を抽象化する

2. IPレイヤ（IP）：ルータ、スイッチ、ファイアウォール
   などでセグメント化され、セグメント間のノード同士の
   直接接続は通常は不可

3. 通常のアクセス：サーバ名/フォルダ名/ファイル名
   【データの所在は、固定的に】
   ⇒データを複数ノードに分散設置し同一IDでアクセス
   ⇒位置透過技術

P2P技術の社会的影響

IRI Ubilabs Innovations

1. 個人利用では「インターネットの匿名性」と連動

⇒情報を匿名で入手可能

⇒情報の入手経路が特定困難

⇒P2Pファイル交換による違法コピーが流行

2. ビジネス利用では新たな可能性が増大

⇒Groove社のP2Pグループウェア：サーバ不要の企業横断型情報共有環境

⇒Kontik社、BitTorrent社のP2P型CDN (Contents Delivery Network)

次回は最終回

「P2P技術から見た」

インターネットの

そのもの/あちら側/こちら側の技術

についてお話します。

ご清聴ありがとうございました