インターネット数理科学第7回
～ネットワークのあちら側を支える数理科学その2～

2006年11月16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

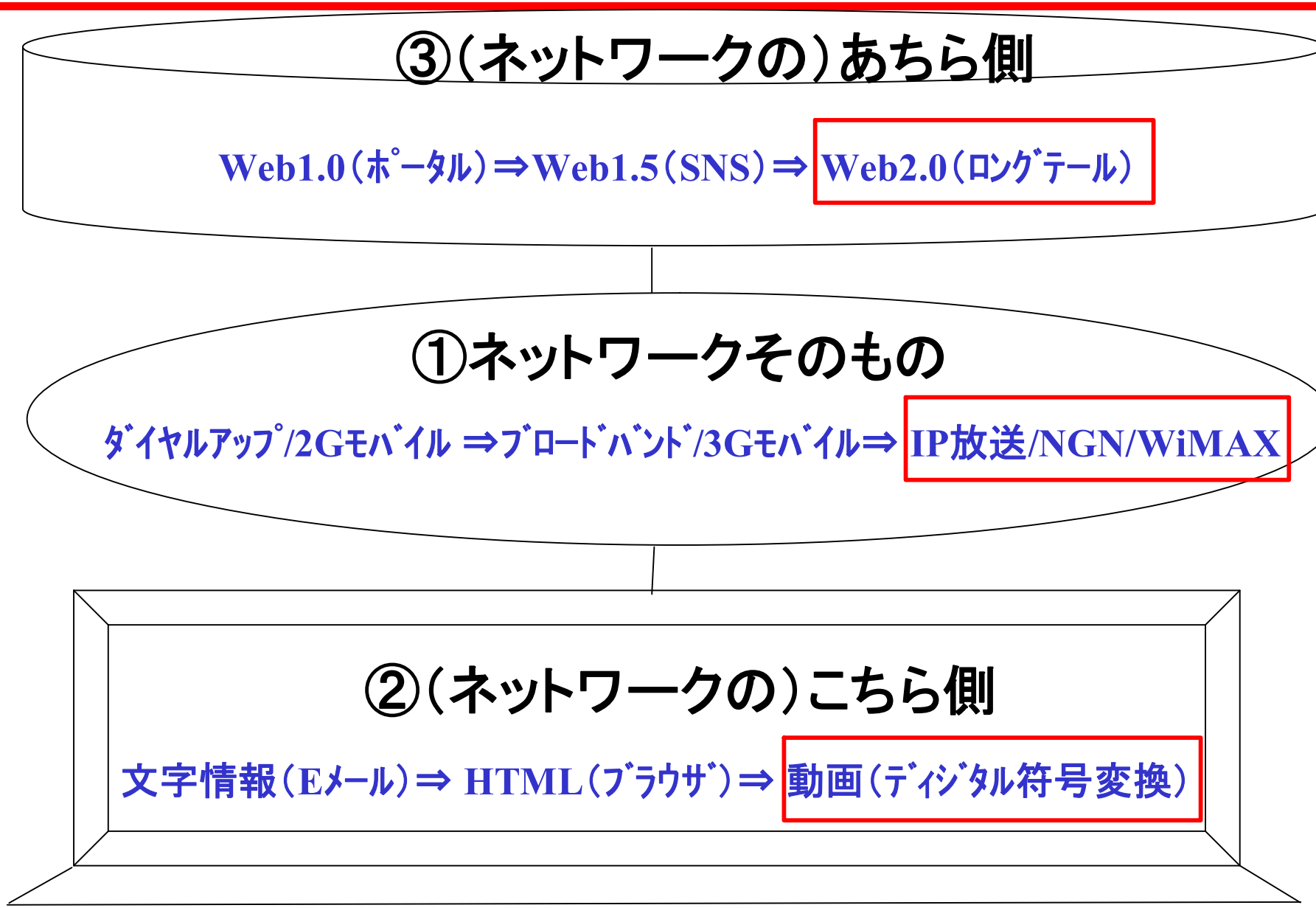
①ネットワークそのものを支える数理科学

⇒「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

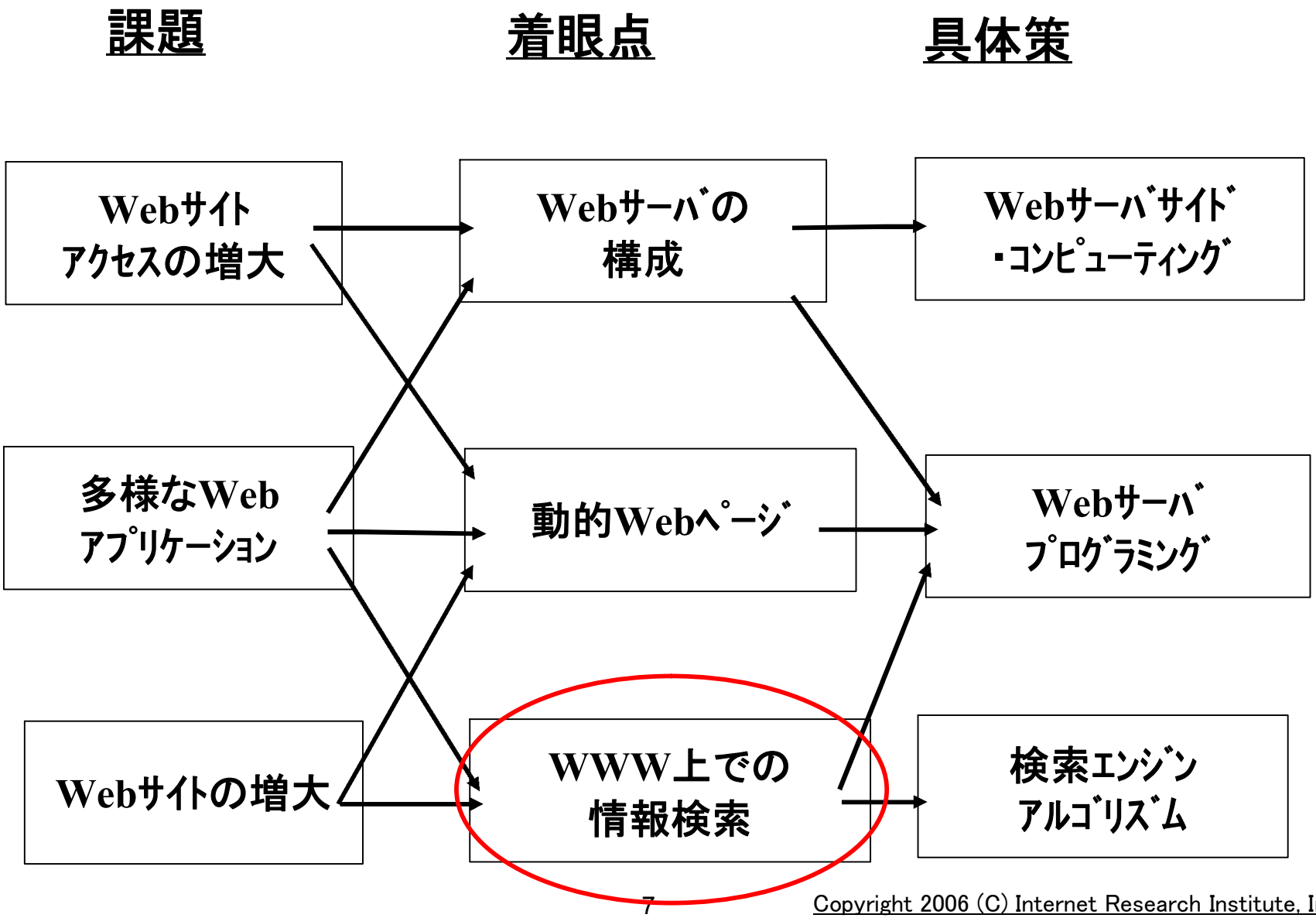
②ネットワークのこちら側を支える数理科学

⇒「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



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。

インターネットが変化した結果、出版業界は？

IRI
Ubilabs Innovations

紙メディアの週刊誌、月刊誌、書籍から成る出版産業の規模は、約2兆6000億円を記録した1996年をピークに減少を続けており、2005年時点で約2兆円。この背景を出版不況といって、業界では、読書離れ、万引き問題、携帯電話など通信費の増加、蔵書欲の減退、二次流通市場の増加、図書館における新刊本・閲覧の増加、など一見出版業界に暗い話題が多く、構造不況だとの指摘がある。日本は、本を読まない社会へ向かい、本当に業界は、衰退してしまうのだろうか？

デジタルコンテンツ白書によると、産業団体や関連省庁が発表した数値をもとに推計した2005年の「メディアコンテンツ産業の市場規模」は13兆6,811億円で、2004年の13兆5,008億円に比べ、1.3%増加。この数値にはデジタルコンテンツ市場規模も含まれるが、2005年の「デジタルコンテンツ市場規模」は2兆5,275億円で、2004年の2兆2,617億円に比べ、11.8%も増加。コンテンツ別の推計では、「映像」が前年比30.4%増の7,088億円、「音楽」が同3.1%増の7,864億円、「ゲーム」が同8.8%増の4,950億円、「図書、画像、テキスト」が同7.4%増の5,373億円。一方、デジタルコンテンツ市場の「流通メディア別割合の経年変化」を見ると、パッケージ流通がまだ大半だが、2001年の84.4%から2005年は71.7%へと縮小傾向を示しており、インターネット流通は2001年の6.8%から2005年には13.7%へ、携帯電話流通は2001年の8.9%から2005年は14.6%へと拡大傾向にある。

インターネットが変化した結果、出版業界は？

IRI Ubilabs Innovations

このように、出版を含む放送、新聞などのコンテンツ市場は、全体で1%以上の伸びを示し、特に、その要因は、10%以上の伸びを示すデジタルコンテンツ市場であり、中でも、インターネットや携帯電話で流通するコンテンツが、成長するデジタルコンテンツ市場全体を牽引していくと考えられる。出版業界は、構造不況ではなく、インターネットという技術革新にさらされ、96年を境にその影響が顕著に出始めたと考えるべき。

そこで今後の出版業界のとるべき方策は、第1に、書籍を中心とした紙メディアの流通市場が、取次店、書店からネット流通に急速に変化していることへの対応です。第2に、雑誌を中心とした出版コンテンツの紙メディアからネットメディアへの対応。言い換えれば、店頭での立ち読み型かネット評判型への変化をとらえ、「評判を創る」Web2.0型対応ともいえる。

3. 情報検索

情報検索の概要

情報検索の基本技術には、以下のようなものがある。

①データの管理および入出力のためのデータベース

②文書データ処理のための自然言語処理や計算言語学

③画像や音声を扱うためのディジテル信号処理と認知心理学を背景とするパターン認識技術

④メタデータの考え方の基本となる図書館情報学

⑤検索アルゴリズム設計

⑥情報検索システムの評価尺度理論

情報検索の過去・現在・未来

「情報検索」に関する研究が、開始されたのは、大規模に蓄積される学術文献や論文等の管理をコンピュータ上で行うために、大規模の図書館でデータの管理と検索が行われるようになった1970年代のことである 。図書館における蔵書検索や電子ジャーナル、統計資料のデータベースなどへの応用は現在でも盛んに用いられている。

1990年代からは、Yahoo!やGoogleのようなWorld Wide Web上のデータを対象にした検索エンジンが登場し、現在では一般生活にまで普及している。

今後の情報検索の課題は、以下のようなものがある。

①Deep Web（ショッピングサイト等大規模なデータベースが動的Webサイト）を対象にした検索
②より直感的で操作性に優れたユーザインタフェース
③より人間的なマルチメディア情報検索
④さまざまなメディアを統合的かつ横断的に扱うクロスメディア情報検索
⑤格納されるデータや検索入力が言語に依存しないマルチリンガル(クロスリンガル)検索
⑤P2Pネットワーク等の大規模分散データを対象にした情報検索

情報検索のシステム要素

情報検索のシステム要素を以下のようになる。

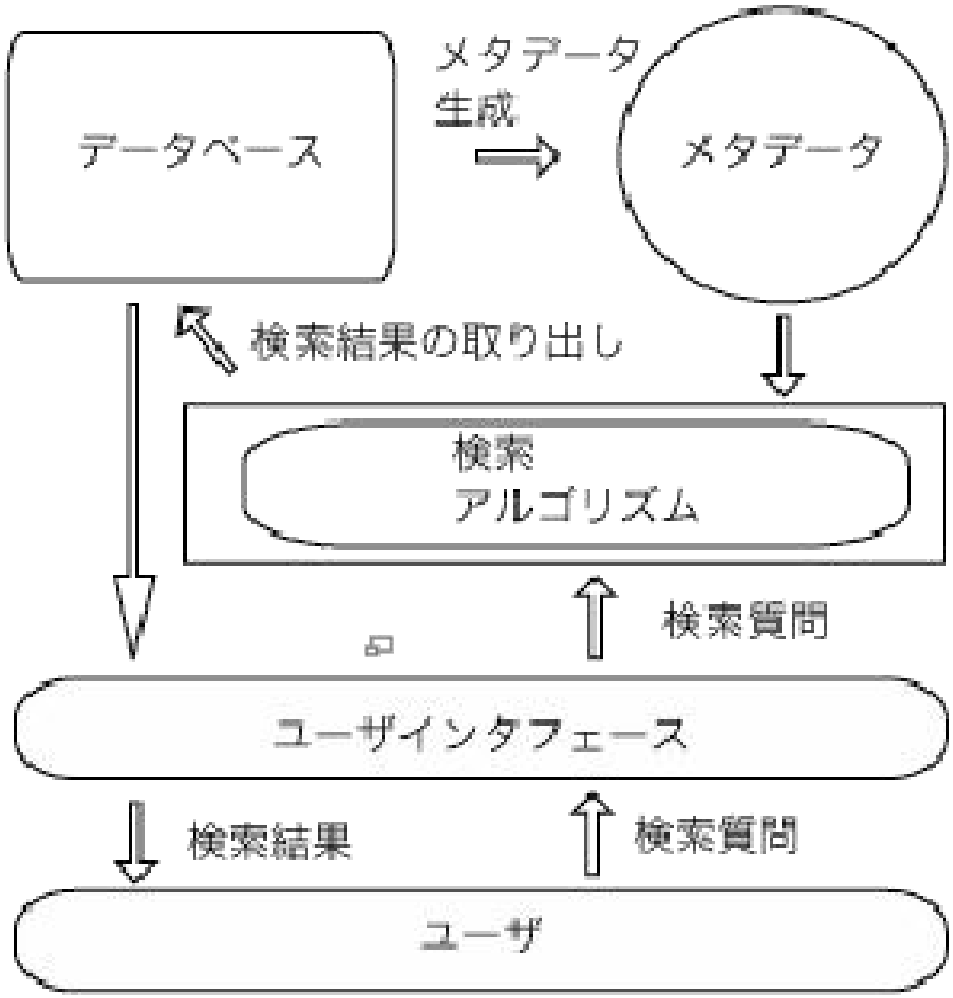
①検索対象のデータ

②データベース（検索対象データの蓄積・管理）

③索引語としてのメタデータ

④ユーザインタフェース

⑤検索アルゴリズム



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。

④検索性能の評価

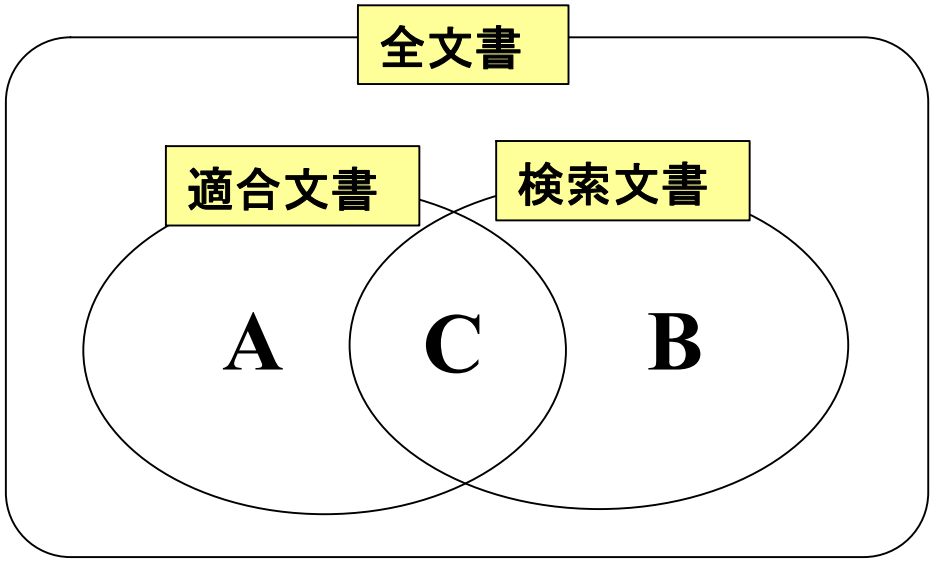
情報検索システムの検索性能の評価を行う。情報検索システムの検索性能は主に正確性と網羅性の質的な観点から適合率(precision;精度ともいう)と再現率(recall)を、処理性能の量的な観点からスループットを測定することにより判定するのが一般的。

ⅰ)適合率　P(Precision)
検索結果として得られた集合中にどれだけ検索に適合した文書を含んでいるかという正確性の指標　　検索ノイズの少なさの指標！
【適合率:P＝C/B(C:検索された適合文書数、*B*:検索結果の文書数)】

ⅱ)再現率　R(Recall)
検索対象としている文書の中で検索結果として適合文書のうちでどれだけの文書を検索できているかという網羅性の指標　検索漏れの少なさの指標！
【再現率:R＝C/A(C:検索された適合文書数、A:全対象文書中の適合文書数)】

再現率と適合率

適合率(P)をあげれば再現率(R)が下がり、再現率を上げれば適合率が下がる傾向にあるため、F値(F-measure)、E値(E-measure)という尺度も用いられる。



＊F値＝2/(1/R+1/P)によって求められる。 F値(RとPとの調和平均:逆数の平均値の逆数)が大ききほど、性能が良いことを意味。再現率と適合率双方の値(0と1の間)が大きい時に大きくなる。

＊E値＝1-(1+b^2)/(b^2/R+1/P)によって求められる。bはパラメータで再現率と適合率のどちらを重視するかを決定。b>1、b<1の時、それぞれ、適合率、再現率を重視する。小さいほど性能が良いことを意味。

情報検索技術の分類

IRI
Ubilabs Innovations

1．検索対象データの抽象度

●直接検索

メタデータを介さずデータそのものを直接計算アルゴリズム上で処理する検索方法。例としてハミングによる検索の入力を行い類似する音程の音楽を検索するもの等。実用上は、前処理としての索引の生成を事前におこなう方式も多いが、このような場合もデータに含まれる表現をそのまま用いて検索を行うため検索モデルとしては直接検索に分類される。

＊全文検索

直接検索の一種で、文書データの全文自動走査により、メタデータを作成・保管し、検索の入力に合致するデータを検索結果とする方法。

●間接検索

データベースに蓄積されたデータからメタデータを生成・保管し、検索入力時に、内部表現に変換された検索入力と保管されたメタデータを比較することで検索結果を生成する方法。

情報検索技術の分類

2. 検索入力の種類

●単語（キーワード）

単語（キーワード）を指定することによって検索を行う。最も単純な形式。

●検索言語

システム特有の検索言語を用いて検索を行う方法。論理和・論理積などのブーリアンモデルの演算を検索の絞り込みに用いる。研究者や法律・医学等の専門的な実務家など、特定の分野の専門家を対象にした検索システムなどに用いられる。SQLのようなデータベース・マネージメント・システムで標準的言語を用いることもあるが、特定分野の検索エンジン特有の検索言語を用いているシステムも多い。実現例としてはIEEE Xploreなどがある。

●直接入力

検索のパラメータとなる関連するデータを直接入力する方法。画像入力による類似画像を検索や、音楽クリップ検索などが研究されている。

●自然文

検索に関わるユーザインタフェースの研究として古くから研究されてきた。
近年ではGoo ラボによって開発された「日本語自然文検索」が話題。

情報検索技術の分類

３．検索アルゴリズム（問題を解くための手順、算法）
　一般に情報検索システムの構築時にはメタデータ生成時に索引を同時に作成し検索アルゴリズムによる検索結果の評価の際に索引を用いた最適化を行う。

①パターンマッチング
　クエリ（検索質問）として入力された表現をそのまま含む文書を検索する。単純にパターンのみを探すのではなく、活用形の変化による同義語のパターンの不一致を解消した検索を行う拡張等が行われる。パターンマッチングの詳細なアルゴリズムについては文字列探索による。

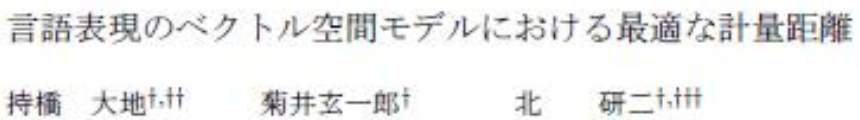
②ブーリアンモデル
　パターンマッチングに加え、メタデータの属性ごとの絞り込み条件を論理和・論理積などによって組み合わせて併用する方法

情報検索技術の分類

③ベクトル空間モデル

キーワード等を各次元とした高次元ベクトル空間を想定し、検索対象データやユーザによるクエリに何らかの加工を行いベクトルを生成。ベクトル空間上に検索対象となるベクトルを配置し、ベクトル化された検索質問とデータのベクトルの相関量（ベクトル間のコサイン、内積、ユークリッド距離など）によって検索の対象のデータと検索質問の関係性の強弱を計算する。

＊参考論文電子情報学会誌


言語表現のベクトル空間モデルにおける最適な計量距離

持橋　大地[†,††]　　菊井玄一郎[†]　　北　　研二[†,†††]

＊ベクトル空間モデルでの単語文書行列

単語文書行列とはメタデータの生成・表現法の一つであり、ベクトル空間モデルによる検索を行う際に非常に頻繁に用いられるメタデータの形式である。一般に単語文書行列は以下に示す構造を持つ。
単語文書行列:

$$\mathcal{M} = \begin{pmatrix} & d_1 & d_2 & d_3 \\ t_1 & 0 & 2 & 1 \\ t_2 & 1 & 1 & 2 \\ t_3 & 0 & 0 & 3 \end{pmatrix}$$

文書*di*に単語*tj*が*n*回出現するとき、*wij*を*n*とし、行列を形成する。単純に出現回数を利用する以外に様々なアルゴリズムによって得た重みを用いる生成方法が多用される。

情報検索技術の分類

IRI
Ubilabs Innovations

④潜在的意味索引付け

ベクトル空間モデルの応用として考案された。高次元ベクトル空間を行列として扱い特異値分解を行い、得られた直交低次元ベクトル空間上で検索する。

単純なベクトル空間モデルでの検索に比べて、同義語が用いられている文書間の関連を反映し、検索の対象のデータの内容的な偏りに影響を受けにくい検索を行うことが可能。

＊特異値分解：線形代数における、複素数あるいは実数を成分とする行列分解の一手法。行列に対するスペクトル分解定理の一般化で、正方行列に限らず任意形式の行列分解が可能。

全文検索（Full text search）

コンピュータ　において、複数の文書(ファイル)から特定の文字列を検索すること。「ファイル名検索」や「ファイル内文字列検索」と異なり、「複数文書にまたがって、文書に含まれる全文を検索する」という意味で使用される。

全文検索技術のいくつかの方法

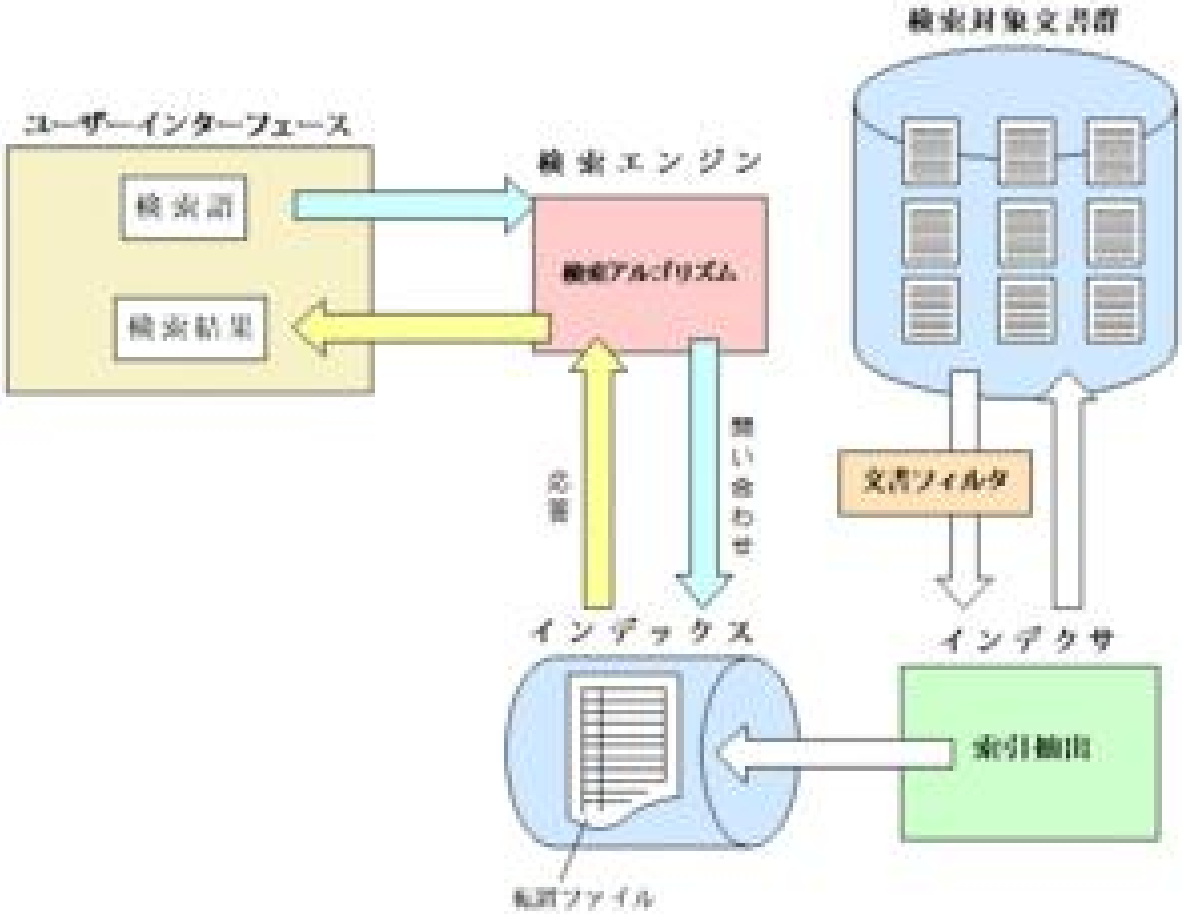
①grep型
順次走査検索、逐次検索。“grep”とはUnixにおける文字列検索コマンドであり、複数のテキストファイルの内容を順次走査していくことで、検索対象となる文字列を探し出す。
一般に“grep型”と呼ばれる検索手法は、事前に索引ファイル（インデックス）を作成せず、ファイルを順次走査していくために、検索対象の増加に伴って検索速度が低下する。“grep型”とは実際にgrepコマンドを使っているという意味ではない。

②索引（インデックス）型

検索対象となる文書数が膨大な数になる場合、grep型では検索を行うたびに個々の文書にアクセスし、該当データを逐次検索するため、検索対象文書の増加に比例して、検索時間が長くなる。そこであらかじめ検索対象となる文書群を走査し、索引データを準備することで、検索性能を向上させる手法。

事前に索引ファイルを作成することをインデクシング(indexing)と呼び、インデクシングにより生成されるデータはインデックと呼ばれ、その構造は多くの場合、「文字列 | ファイルの場所 | ファイルの更新日 | 出現頻度・・・」といったようなリスト形式（テーブル構造)を取り、文字列が検索キーとなる。検索時にはこのインデックスにアクセスすることで、大幅な高速検索が可能となる。

インデックス作成型全文検索システム



索引文字列の抽出手法

形態素解析

英文の場合は単語間にスペースが入るため、自然にスペースで区切られた文字列を抽出していけば、索引データの作成は容易となる。

日本語の場合は、単語をスペースで区切る習慣がなく、形態素解析技術を用いて、文脈の解析、単語分解を行い、それをもとにインデックスを作成する。

形態素解析を行うには解析用の辞書が必須であり、検索結果は辞書の品質に依存。また、辞書に登録されていないひらがな単語の抽出に難があるなど、技術的障壁も多く、検索漏れが生じることがある。

N-Gram

「N文字インデックス法」「Nグラム法」ともいう。

検索対象を単語単位ではなく一定のN文字単位で分解し、それの出現頻度を求める方法。

Nの値が1なら「ユニグラム(uni-gram)」、2なら「バイグラム(bi-gram)」、3なら「トライグラム(tri-gram)」と呼ぶ。たとえば「全文検索技術」という文字列の場合、「全文」「文検」「検索」「索技」「技術」と2文字ずつ分割して索引化を行ってやれば、検索漏れが生じず、辞書の必要も無い。しかし形態素解析によるわかち書きに比べると、意図したものとは異なる検索結果が多く、インデックスのサイズも肥大化しがちであることが難点。

形態素解析とN-gramの比較表

	形態素解析	N-gram
インデクシング速度	遅い	早い
インデックスサイズ	小さい	大きい
検索ノイズ	少ない	多い
検索漏れ	多い	少ない
検索速度	速い	遅い
言語依存	辞書が必要	辞書が不要

文書フィルタ

検索対象文書がプレーンテキスト以外、たとえばHTML文書ならばタグの除去等の処理を行ってテキストを抽出できるが、特定メーカーのワープロ独自形式などバイナリ形式の場合、インデクサは直接ファイルからテキストを抽出することが出来ないため、文書フィルタを利用して該当ファイルからテキストを抜き出す必要が生じる。

文書フィルタ機能は、インデクサが内包しているものもあれば、アドインなどの機能拡張によって実装する場合もある。

代表的な文書フィルタ

- **Xpdf**
  - **Xpdf**
  - **NamazuでPDF文書からテキストを抽出するために利用。**
- **IFilter**
  - **MSN サーチ ツールバー with Windows デスクトップ サーチの機能拡張**
  - **Index Service、Windowsデスクトップサーチのアドインとして各社から提供**
- **xdoc2txt**
  - **http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html**
  - **高速Grepソフトウェア「KWIC Finder」からフィルタ部分を抜き出したもの。**
  - **Hyper Estraierでは標準文書フィルタとして利用。**

転置ファイル

IRI
Ubilabs Innovations

全文検索用のインデックスには様々な形式があり、最も一般的なものは単語と、単語を含む文書ファイルのIDとで構成された可変長のレコードを持ったテーブルで、転置ファイル(inverted file　転置インデックスとも)と呼ばれる。

インデクシングや実際の検索の際には「二分探索」アルゴリズム等を使って、検索単語から文書IDを高速探索可能。

転置ファイルのデータ構造や、探索アルゴリズムは全文検索システムによって多彩であるため、これらの違いによってインデックスサイズ、検索速度、検索精度に大きな差が生じる。

転置インデックスの基本的な概念はテーブルに対して「単語Xはどの文書にあるか」といったクエリ(情報検索における明確な情報要求)の応答速度を最適化するということである。

転置ファイルの例

単語	文書ID
サーチ	1, 3, 4
デスクトップ	2, 4, 7
解析	3, 5, 6, 7
形態素	2, 6, 7
検索	1, 6
全文	1, 6, 7

＊二分探索を行うためには単語と文書IDはソート済みでなければならない

＊二分探索（にぶんたんさく、BS;Binary Search）。ソート済みリストや配列に入ったデータ（同一の値はないものとする）に対する検索を行う場合、中央値を見て、検索値との大小関係を用いて、検索値が中央値の右か左かを判断して、片側に存在しないことを確かめながら検索する。
大小関係を用いるため、未ソートのリストや配列には二分探索を用いることは不可。
n個のデータの中央の値を見ることで、1回の操作でn/2個程度（奇数の場合は(n-1)/2個、偶数の場合はn/2個または(n/2)-1個）の要素を無視できる。

全文探索の用途

IRI
Ubilabs Innovations

www検索サービス

検索サービスの中では、ホットな分野で、熾烈な競争が行われている。ウェブの初期から行われていたサービスで、技術革新の激しい分野。

企業向け社内検索サービス

社内の文書資産を高速全文検索するシステム。WordやExcelといったオフィス用ソフトウェア・フォーマットから、メール、データベースなどの多くのファイル形式に対応し、企業の性格に応じた検索結果を提供。近年、電子データの企業資産の重要性が急増したことによる急成長分野。

デスクトップ検索

個人のローカルファイル検索用アプリケーション。Word,Excel,PDFなど種々のファイル形式に対応。また、画像データなどの、個人所有のマルチメディア情報検索に特化したものもある。

全文検索エンジン例(サーバ向け)

IRI
Ubilabs Innovations

Rast

わかち書きベース・N-gramベースの選択。
単語の出現位置情報を保持し、正確なフレーズ検索が可能。フォルスドロップなし。
無償。

Namazu

わかち書きベース。
2単語によるフレーズからハッシュ値を計算し保持することによって、フレーズ検索が可能。フォルスドロップ(false drop=誤った候補)あり。
日本で広く使われている全文検索システム。無償。

Hyper Estraier

N-gramベース(N.M-gram)。わかち書き方式も併用可。
分散インデクス、Webクローラ、検索用CGIなど標準添付のプログラムが充実。
N.M-gram方式とは、N文字に続くM文字のハッシュ値を計算し保持することによって、フレーズ検索が可能。フォルスドロップあり。
大規模なインデックスも作成可。無償。

全文検索エンジン例（サーバ向け）

IRI
Ubilabs Innovations

Lucene

Apache Lucene
Analyzerと呼ばれるクラスを選ぶことにより、N-gramやわかち書き形式でのインデックス作成ができる。
Javaによる全文検索システム。
Wikipediaの検索に使われている。
大規模なインデックスも作成可。無償。

Senna

Senna 組み込み型全文検索エンジン
わかち書きベース・N-gramベースの選択
他のプログラムからライブラリとして呼び出して利用する。
MySQLの中に全文検索エンジンを組み込むパッチが提供されており、MySQLを利用しているプログラムであれば全文検索機能を手軽に実現できる。
PostgreSQLに対して、Sennaを全文検索エンジンとして組み込むためのモジュールLudiaが公開。
perlバインディングにより、perlスクリプトから簡単に利用することができる。PHP, Ruby, Pythonバインディングも提供されている。
大規模なインデックスも作成可。無償。

全文検索エンジン例(サーバ向け)

SMART/InSight 2.0

ウチダスペクトラム株式会社
形態素解析、N-gram選択可。
ActiveDirectoryなどのACL継承機能有り。
非常に高速なFastDataSearchをエンジンとして使用。
有償。

Oracle Secure Enterprise Search

(Oracle Japan / Oracle Secure Enterprise Search 10g)
N-gramベース(V-gram)。
ログインしたユーザーが参照可能な結果のみを表示するセキュア検索が特徴。
有償。

全文検索エンジン例(個人向け)

IRI
Ubilabs Innovations

FindFast(マイクロソフト)

MS Office97/2000に標準搭載されていた。
2時間おきにインデックス更新。CPU負荷が極めて高いためユーザーの評判は悪かった。
Officeファイルが対象であり、メール検索などはできない。

インデックスサービス(マイクロソフト)

Windows2000/XPに標準搭載。
アプリケーションではなくサービスとして提供されているため使いづらい。
デフォルトではオフとなっているため利用者は少ないと思われる。
ローカルディスク全体をインデックス化できるが、CPU負荷は高くなる。
メール検索には非対応。
今後はWindows デスクトップサーチをベースにしたシステムになるだろう。

Google デスクトップ(Google)

http://desktop.google.com/ja/
わかち書きベース。
利点:Google web検索と同じエンジンでローカルファイルを検索できる。
欠点:大きなファイルの場合、後半部分がインデックス化されないなどの問題。
無償。

全文検索エンジン例（個人向け）

Windows デスクトップサーチ(マイクロソフト)

http://desktop.msn.co.jp/
わかち書きベース
利点：検索対象フォルダを詳細に設定可能。ネットワークドライブにも対応。
欠点：単独でのリリースが企業向けのものしかされていない。
正式には「MSN サーチ ツールバー with Windows デスクトップ サーチ」というパッケージで配布されている。無償 。

サーチクロス(ビレッジセンター)

http://www.villagecenter.co.jp/soft/searchx/
形態素解析によるわかち書きは採用せず。
詳細なアルゴリズムは不明だが、文字種により単語を分割し、インデックスに登録していくタイプ。
ひらがなは事前に辞書登録されたもの以外は検索できない。
アルゴリズムが単純な分、インデクシングが極めて高速。有償。

コンセプトサーチ(ジャストシステム)

http://www.justsystem.co.jp/conceptsearch/
自然言語による検索が可能。
文書管理アプリケーションDocuWorks(富士ゼロックス)のバージョン6.0より「ExpandFinder」という名称でバンドルされている。有償

C2cube株式会社 / C2cube, Inc.（日本）

http://www.c2cube.com/

当社は、独自開発による自然言語認識エンジンConciergeCube™を核とした各種サービス提供を行っております。同エンジンは、一般的な言語解析手法である形態素解析とは全くコンセプトが異なる「機能素解析」アルゴリズムにより設計されております。「機能素解析」アルゴリズムは、「形態素解析」とは異なり、文法、大量の辞書、品詞情報には依存していないため、ルーズな文法や未知語に強く、ブログやメールなどに頻出する口語調のテキストの高精度構文解析において大きな強みを発揮します。

この「機能素解析」アルゴリズムをベースにした、当社運営のリアルタイムブログ評判検索サービスBuzzTunes™のみならず、BuzzTunes™データベースによる評判検索API提供サービスを利用した、「アフィリエイト自動推奨サービス」、「高精度コンテンツ連動型広告マッチングサービス」、「クチコミ地理検索サービス」、「リアルタイムCGM（Consumer Generated Media）リサーチサービス」などの用途を開発中です。

機能素解析アルゴリズムが登場

Ubilabs Innovations

3大特徴

口語調テキストに対しても高精度な構文解析を実現します

ConciergeCube™の独自アルゴリズムである機能素解析アルゴリズムを駆使した頑健な構文解析能力により、従来の自然言語処理手法で一般的な形態素解析エンジンでは困難とされていた口語調やルーズな文法の日本語テキストに対しても極めて高精度な構文解析や係り受け解析を実現します。

1. 大量の辞書を必要としません

従来の形態素解析においては、少なくとも20～30万語の辞書を必要としていましたが、ConciergeCube™はわずか数千語の辞書を使って、日本語テキストの高精度な構文解析を実現します。

これによりConciergeCube™によるテキスト処理は、システム負荷が小さく、辞書のメンテナンスも容易となるだけでなく、ネットにおける新語などの辞書にない未知語に対しても柔軟な対応が可能となります。

2. 日本語テキストの多様なゆらぎを吸収します

ConciergeCube™の頑健な構文解析能力を駆使したゆらぎ吸収標準化機能により、多様な形態で出現する日本語テキストのゆらぎを吸収して標準的な言い回しに標準化します。

これにより口語調の和文の高精度機械翻訳や方言に対する柔軟な対応が可能となります。

機能素解析アルゴリズムが登場

IRI
Ubilabs Innovations

従来の形態素解析における諸問題を全て解決

ConciergeCube™は自然言語処理手法として一般的である形態素解析における以下の諸問題を全て解決しました。

1. 単語の境界判別（分かち書き）の問題
   ConciergeCube™の機能素解析アルゴリズムによる頑健な構文解析能力により、全文ひらがなのテキストに対しても単語分解が可能となります。

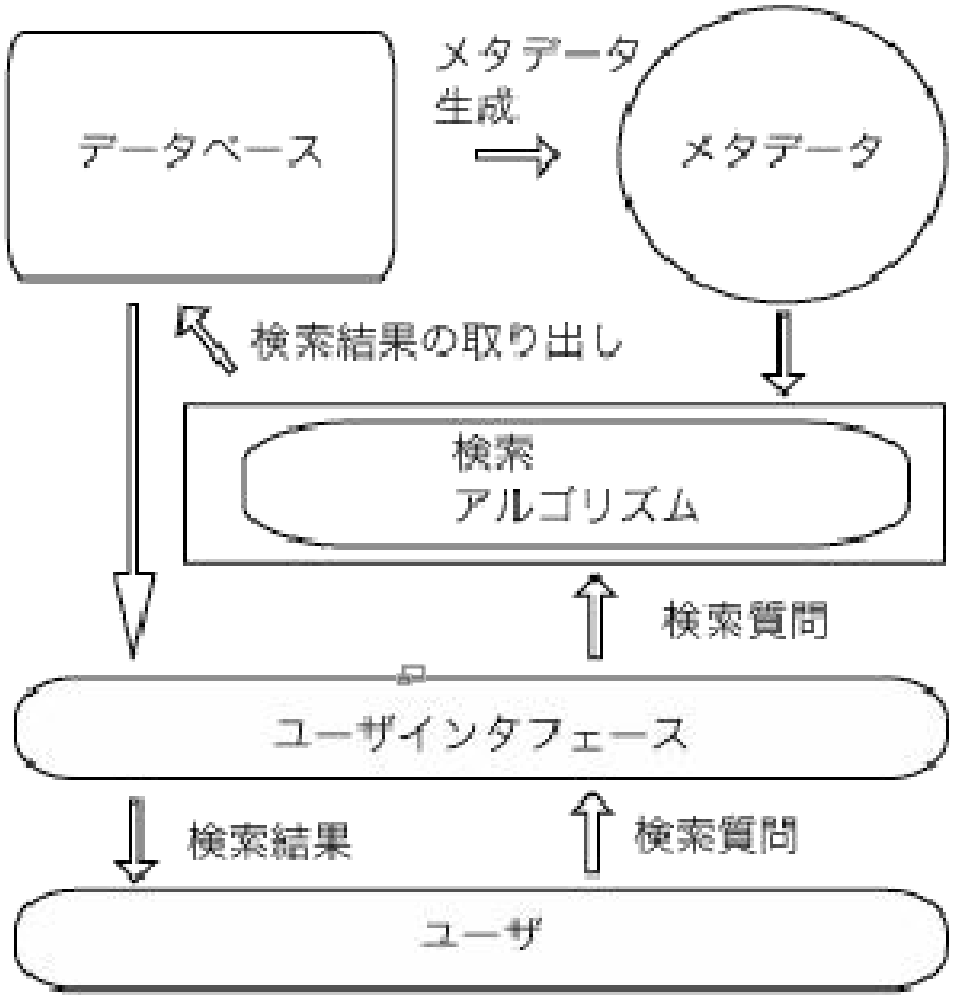
2. 品詞判別の問題
   ConciergeCube™の言語処理においては、品詞判別の必要がありません。

3. 未知語の問題
   ConciergeCube™は辞書に依存していないので、未知語に対しても柔軟な対応が可能となります。

4. ルーズな文法の問題
   ConciergeCube™のゆらぎ吸収標準化機能により、ルーズな文法や口語調のテキストに対しても頑健な構文解析を実現します。

4. メタデータとデータベース

再び相関図



←【データのデータ】

メタデータ

IRI
Ubilabs Innovations

メタデータ(metadata)＝データについてのデータ

集約的に管理されるデータを扱うための、各データが付随して持つ抽象化された、または付加的データ。例としては、検索する際により高度な条件設定をしたり、デジタルコンテンツの著作権情報などを埋め込む内容のこと。

情報検索分野におけるメタデータ

情報検索システムの検索の対象となるデータを要約したデータ。

図書館情報学分野の書誌情報

例えば文書であれば著者名や表題、発表年月日等のほか、関連キーワードなどを含めるのが一般的。また、デジタルカメラを用いて写真を撮影し、JPEGファイルとして保存した場合、Exifにそったメタデータが自動的に作成されるのが一般的。

メタデータ

メタデータとして記述される関連キーワード＝索引語

メタデータを作成するのは、検索が発行されるたびごとに検索の対象となるデータの性質を読み取り検索結果に含めるかどうかを判定するのは著しく非効率であり、あらかじめ検索を行いやすい状態に加工し、データを検索用に整理するため。

メタデータ記述の規格

メタデータの記述方法の統一は、相互運用性を高め、自動可読なメタデータの利用促進のために、様々な標準化団体や業界団体においてメタデータの記述形式が制定されている。

メタデータ利用の問題点

情報そのものとメタデータ情報との乖離問題。また、情報そのものの更新時には、メタデータも同時に変更が必要だが、時間的なズレや非同期が発生。

改竄や機械的処理による、検索の精度が低下。

情報の評価が主観に依存の場合、メタデータの有効性が喪失するため、メタデータ利用よりも情報そのものの直接検索が妥当。

データベース

●特定の課題に沿ったデータを収集管理し検索することで再利用可能にしたデータの集合。
●狭義には、コンピュータによって実現されたデータの集合でOSが提供するファイルシステム上に直接構築されるものや、データベース・マネジメント・システムを用いて構築されるものがある。
●データの再利用を高速かつ安定に実現するため、データを格納するためのデータ構造については、数理科学上の重要研究分野である。
●単純なファイルシステムには、「データ」を統一的手法で操作する機能はない。
●ファイルシステムを活用して、データ管理をするためには、データの操作機能をアプリケーション・プログラム側に持つ。データベースは、それを自ら持つことにより、アプリケーション・プログラム側でデータの物理的格納状態に依存せず、操作可能で、データの物理的格納状態に変更があった場合にもアプリケーション・プログラム側の処理に影響が及ばないことを保証する。
(プログラムとデータの独立性の保証)
●データベースをコンピュータ上で管理するためのシステムをデータベース・マネジメント・システム（DBMS）と呼ぶ。
（Oracle、SQL Server、PostgreSQL、MySQL等がある）

●データベースの例
コンピュータで管理された住所録、検索エンジン、電子カルテ、企業データベース、音楽楽曲用データベース、などがある。
広義には電子化されたもの以外も含まれるので，出版物としての電話帳、辞書、特許公報などもデータベースの範疇に入る。

●データモデル
データモデルは、データベース化するデータをどのように配置するかを論理的・物理的に規定する。
【例】
階層型データモデル 、ネットワーク型データモデル、リレーショナルデータモデル
オブジェクト指向データモデル 、カード型データモデル

●リレーショナルデータモデル
IBMのE.F.Coddによって考案された最も多く用いられているデータモデル。
複数のリレーション（関係）を基本的なデータ型とする。格納されたデータを取得するための質問（クエリ）は、関係代数演算および集合演算によって行う。

データマイニング

●データの巨大集合やデータベースから有用な情報を抽出する技術体系。

● データマイニングは、通常はデータ解析に関する用語として用いられる。

●歴史

1960年代：データマイニングの発展前史として、コンピュータの出現により、
大量のデータ蓄積が可能となったことが直接的に関係。

1980年代：リレーショナルデータベースとその操作用の言語SQLが出現し、
動的なデータ解析が可能となった。

1990年代：データ量は爆発的に増大した。データウェアハウスあるいは
データセンターがデータの蓄積に利用され始めた。

⇒データベースにおける大量データを処理するための手法として
データマイニングの概念が現れ、統計解析の手法や検索技術等が応用。

●解析手法（相関ルール抽出 ）

データベースに蓄積された大量のデータから、頻繁に同時に生起する事象同士を相関の強い事象の関係、として抽出する技術。（POS、ﾈｯﾄｼｮｯﾋﾟﾝｸﾞ等）

決定木

●決定理論の応用分野において、決定するためのグラフ理論上のグラフで、計画立案し目標達成のために用いる。

●決定木は、意志決定を支援することを目的として作る。

●決定木は木構造の特別な形である。

●機械学習の分野においては決定木は予測モデルであり、ある事項に対する観察結果から、その事項の目標値に関する結論を導く。

●内部節点は変数に対応し、子節点への枝はその変数の取得可能値を示す。葉（端点）は、根(root)からの経路によって表される変数値に対して、目的変数の予測値を表す。

●データから決定木を作る機械学習の手法のことを決定木学習、あるいは単に決定木と呼ぶ。

決定木学習

●データマイニングで良く用いられる方法。決定木は、葉が分類を表し、枝がその分類に至るまでの特徴の集まりを表わす木構造を示す。

●決定木学習は、元集合を属性値テストによって部分集合に分割することで実行。当処理は、すべての部分集合に対して再帰的に繰り返すが、分割が実行不可となった場合、または、部分集合の個々の要素が各々1つずつの分類となる段階で終了。

●決定木は、データの集合を表現したり分類化や法則化することを助ける数学的手法、計算手法。

●データは以下のような形式のレコードとして規定。

(x, y) = (x1, x2, x3…、 xk, y)

＊従属変数 y は、理解し分類や法則化の対象で、変数 x1, x2, x3 等はそれらを実行する上で参考となる変数である。

●種類

決定木には、他に2つの呼び名がある。

・回帰木： 分類ではなく実数値を取る関数近似に利用（住宅価格見積り）。

・分類木： yが分類変数の場合（例：性別（男/女））

決定木学習

●実際の例（Wikipedia）

決定木を実例で見てみよう。

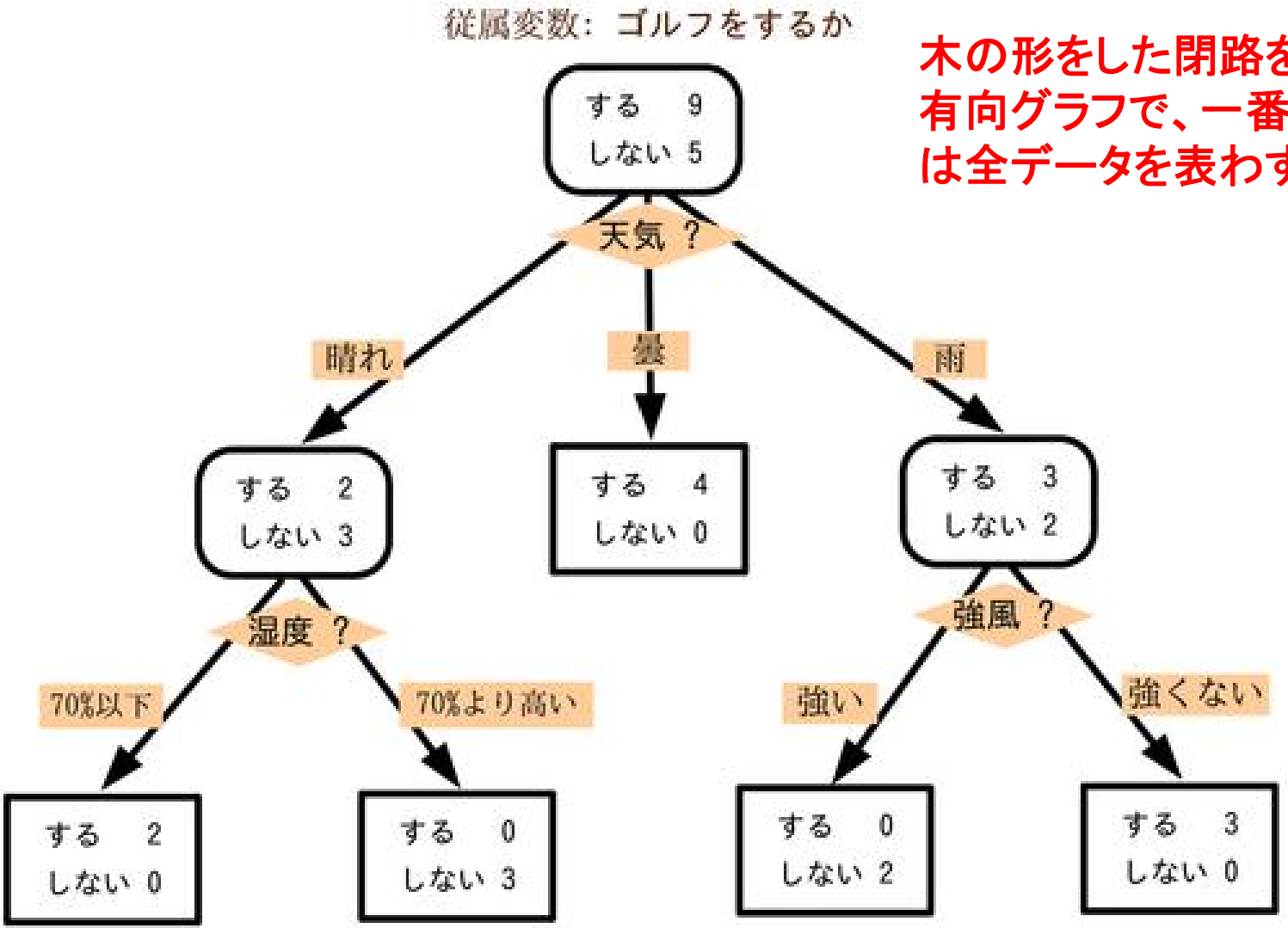
中野さんは有名なゴルフクラブの経営者。お客さんの来場状況についてちょっと悩みを抱えている。お客さんが殺到する日があり、そういう日はクラブの従業員は数が足りない。そうかと思えばお客さんがまったく来ない日もあり、そんな日は従業員は非常に暇そうだ。

中野さんは、週間天気予報に基づいて、お客さんがいつゴルフクラブにやってくるのかを予想し、従業員の勤務体制を最適化したいと思っている。そのためには、もし説明がつくのであれば、人がゴルフをやりたくなる理由を知る必要がある。
そこで2週間の間、中野さんは以下の情報を集めた。

天気（晴れ、曇、雨）、気温（度）、湿度(%), 風（強い、弱い）、それにもちろん、彼のゴルフクラブにその日、人が来たかどうか。
結果、以下のような 14 行 5 列のデータを集めることができた。

独立変数				従属変数
天気	気温 (度)	湿度 (%)	風が強いか	ゴルフをするか
晴れ	29	85	強くない	しない
晴れ	27	90	強い	しない
曇	28	78	強くない	する
雨	21	96	強くない	する
雨	20	80	強くない	する
雨	18	70	強い	しない
曇	18	65	強い	する
晴れ	22	95	強くない	しない
晴れ	21	70	強くない	する
雨	24	80	強くない	する
晴れ	24	70	強い	する
曇	22	90	強い	する
曇	27	75	強くない	する
雨	22	80	強い	しない

問題解決の決定木



木の形をした閉路を含まない有向グラフで、一番上の節点は全データを表わす。

決定木の作り方

分類木の自動生成するアルゴリズムがあり、それを上の表に示すデータセットに適用すると、従属変数である「ゴルフをするか」を説明する最も良い方法は、*変数「天気」を用いること*だという結果が得られる。実際、「天気」の値によって表を並び替えると以下の図のようになる。

独立変数				従属変数
天気	気温 (度)	湿度 (%)	風が強いか	ゴルフをするか
晴れ	22	95	強くない	しない
晴れ	21	70	強くない	する
晴れ	24	70	強い	する
晴れ	29	85	強くない	しない
晴れ	27	90	強い	しない
曇	27	75	強くない	する
曇	28	78	強くない	する
曇	22	90	強い	する
曇	18	65	強い	する
雨	22	80	強い	しない
雨	21	96	強くない	する
雨	20	80	強くない	する
雨	18	70	強い	しない
雨	24	80	強くない	する

変数「天気」の分類では3つのグループが出現

晴れの日にゴルフをするグループ、曇の日にゴルフをするグループ、雨が降っていてもゴルフをするというグループ。

ここでもし、変数「気温」を用いて分類するとどうなるかを考える。「気温」の値の昇順に表を並び替えると

独立変数				従属変数
天気	気温 (度)	湿度 (%)	風が強いか	ゴルフをするか
曇	18	65	強い	する
雨	18	70	強い	しない
雨	20	80	強くない	する
晴れ	21	70	強くない	する
雨	21	96	強くない	する
雨	22	80	強い	しない
曇	22	90	強い	する
晴れ	22	95	強くない	しない
晴れ	24	70	強い	する
雨	24	80	強くない	する
晴れ	27	90	強い	しない
曇	27	75	強くない	する
曇	28	78	強くない	する
晴れ	29	85	強くない	しない

決定木は、複雑なデータ表現を簡単化する

IRI
Ubilabs Innovations

●ある温度を境にして2グループあるいは3グループに分けるとしても、明確には分けられない。他の変数についても同様である。
●「天気」で分類すると、曇の場合に従属変数が “する” であるデータだけのグループが作れることから、最初に「天気」で分類することは適切な判断。
●全データをまずは「天気」で分類することにし、最初の結論としては、天気が曇りならば人は必ずゴルフをし、雨の日であっても熱狂的な人はゴルフをする! ということが分かる。
●さらに、晴れの日のグループを2つのグループに分ける。お客さんは湿度が70%よりも高い時はゴルフをしたがらないようだ。
●最後に、雨の日の場合を2つに分けてみると、風が強い時にはお客さんがゴルフをしに来ることはない、と分かる。
●分類木による問題の解答は、次のように結論づけられる。
①晴れでじめじめした日や風の強い雨の日にはほとんどゴルフをしに来る人はいないので、中野さんは従業員のほとんどを休ませるとよい。
②それ以外の、沢山の人がゴルフをすると思われる日には、仕事を手伝ってくれる臨時従業員を雇う。

クラスタリング

●コンピュータシステムのハードウェア構成
多数の低廉なコンピュータを、特別なソフトウェア・ハードウェアを用いて、あたかも1つの大きなコンピュータとして利用できるように接続すること。
・ソフトウェアの点からのクラスタリングは高信頼化手段。

●統計学データ解析手法の1つ（情報検索に有用）
機械学習やデータマイニング、パターン認識、イメージ解析やバイオインフォマティックスなど多くの分野で用いる。
クラスタリングではデータの集合を部分集合（クラスタ）に切り分けて、それぞれの部分集合に含まれるデータがある共通の特徴を持たせる。
共通の特徴では多くの場合、類似性や、ある定められた距離尺度に基づく近さで示す。

5. 検索エンジン

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

IRI
Ubilabs Innovations

●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。

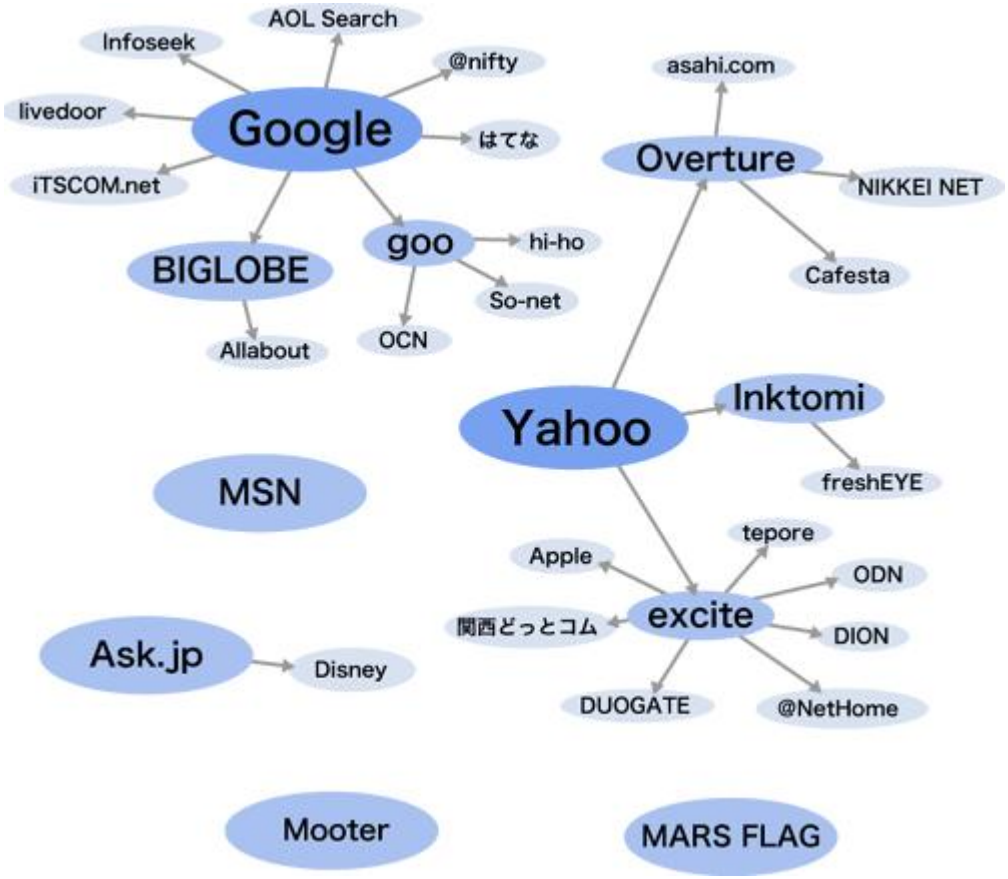
2006 年
不自然な外部リンクに対して、Googleが厳格処置。
携帯の検索サービス登場。NTTドコモ:複数エンジン、au:Google、ソフトバンク:Yahoo!。

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。

ご清聴ありがとうございました