検索エンジンの検索アルゴリズム

兼宗 進

概要

WWW(World Wide Web) 上の文書を検索する検索エンジンは、インターネットを利用する上で不可欠な存在である。検索エンジンは従来の情報検索技術を基礎としながら、独自の発展を遂げてきた。しかし、内部の検索アルゴリズムが十分に公開されていないことから、検索エンジンは、中の見えないブラックボックスとして手探りの使い方をされることが多い。そこで本稿では、検索エンジンの検索アルゴリズムを構成する「適切なページの収集手法」「ノイズや漏れのない検索を高速に行う手法」「適切にランキングして表示する手法」について述べる。

1 はじめに

現代の検索エンジンに必要とされるのは、適切なページを検索結果の先頭に表示する検索アルゴリズムである。検索結果を重要な順に並べ替える処理をランキングと呼ぶ。現代の検索エンジンの評価はランキングで決まると言っても過言ではない。

図 1 に、検索エンジンにおける検索の流れを示す。

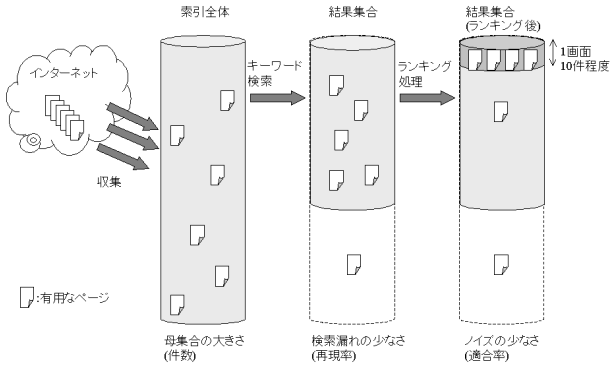

図 1 検索エンジンの検索の流れ

検索エンジンは、収集処理で WWW を巡回し、あらかじめ多くの WWW ページを収集して索引に蓄えておく (本稿では、WWW で公開されている状態のページを WWW ページと呼び、内部に蓄えられたページと区別する)。ユーザーが検索を行うと、索引から検索語を含む検索処理が行われ、結果集合が作られる。結果集合にはランキング処理が行われ、重要度の高いページから結果が表示される。

従来の情報検索では、結果を全件見ることを前提に、小さな結果集合を作ることに主眼が置かれていた。検索エンジンによる検索では、検索結果の適切な並べ替えに重点が置かれている。

2 何が検索エンジンに求められているか

2.1 典型的な使われ方

検索エンジンを使う人は、どのような検索を行っているのだろうか。海外の検索エンジンである Infoseek[13] では、「2 個程度の単語で検索し、先頭の 10 件を見る」のが典型的なユーザーの使い方であり、「検索式を使った検索は全体の 10% 程度であり、正しく記述されないことが多い」、「上級検索を使うのは全体の 1% 程度であり、ほとんど使われることはない」ことが報告されている。国内の検索エンジンである ODIN[19] では、「1 語による検索が 70% 以上である」ことが報告されている。

これらの報告から、ユーザーは「1,2 個の思い浮かんだ検索語を入力するだけで、現在必要としている情報が掲載されたページが先頭 10 件に表示される」ことを検索エンジンに求めていることがわかる。

表 1 に、検索エンジンの使われ方を示す。比較のために、情報検索の使われ方と対比する。

表 1 検索エンジンの使われ方

項目	従来の情報検索	検索エンジン
ユーザー	サーチャー	初心者
検索語	吟味した検索式	思い付いた 1,2 語
結果の閲覧	全件	先頭 10 件
絞り込み	する	しない
結果集合	数十〜数千件	数万〜数百万件
求める情報	網羅的	数件

検索エンジンを使うユーザーの多くは、検索の専門家ではないし、検索のための専門の訓練も受けていない。そのため、検索語をあらかじめ吟味してから使う

ことはないし、複雑な検索式を使うこともほとんどない。その結果、思い浮かんだ 1,2 個の単語による検索では、結果として数十〜数百万件ものページがヒットすることが多い。(試しに Google で「情報検索」を検索すると約 36 万件が、「Information Retrieval」を検索すると約 375 万件がヒットした)

このように多くの検索結果が得られた場合、網羅性を重視する従来の情報検索では、全件を閲覧できる大きさまで結果集合を絞り込むことが行われていた。一方、検索エンジンによる検索では、結果集合の件数が多いことは問題にならず、ランキングによる表示を重視している。

2.2 評価

情報検索システムの性能は、一般に再現率 (検索漏れの少なさ) と適合率 (検索ノイズの少なさ) で評価される。

図 1 において、収集処理ではできるだけ多くの WWW ページを収集する必要がある。検索エンジンでは、インターネット上で公開されているページでも、索引に登録されていないページは検索されないためである。索引からの検索では、検索語が含まれているページを漏れなく検索することが必要である。これらは、再現率の観点から重要と考えられる。

ランキング処理では、ユーザーが見る先頭 10 件とそれに続くページに重要度の高いページが集められている必要がある。これは適合率の観点から重要と考えられる。

表 2 に、検索エンジンの検索アルゴリズムに求められる性質をまとめる。

表 2　検索アルゴリズムに求められる性質

1. 検索語を含むページを検索する
2. 多くのサイトから検索する
3. 新しい情報を検索する
4. 先頭に有用なページを表示する

以下では、検索エンジンがこれらの性質をどのように実現しているのかを順に見ていく。

3 検索データの生成

よい検索を行うためには、検索用のよいデータを準備する必要がある。質の悪いデータから質のよい検索結果を得ることはできないし、そもそも多くのページを蓄えておかなければ、そこからよいページを選ぶこともできないからである。ここでは、検索用のデータを作る処理のうち、検索に影響を与える部分を中心に解説する。

図 2 に、検索エンジンの構成図を示す。上半分は索引生成処理であり、インターネットからページを収集して検索用の索引を生成する。下半分は検索処理であり、ユーザーからの検索リクエストに対して検索を行い、結果を表示する。検索処理のうち、全文検索とランキング処理については後述する。表示処理の詳細は他の文献 [15][18] を参照されたい。

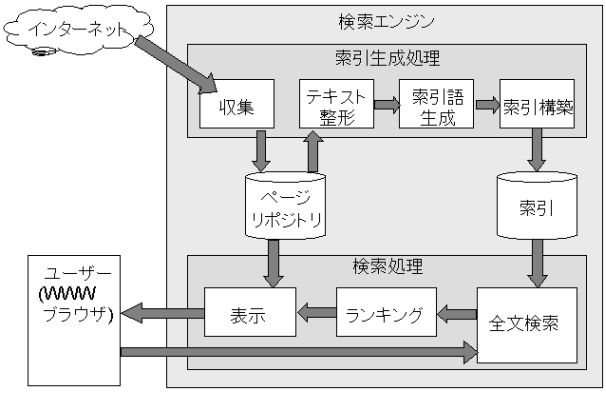

図 2　検索エンジンの構成

表 3 に、索引生成処理で行われる処理の概要を示す。

表 3　索引生成処理

処理	内容
収集	WWW を巡回し、ページを収集する
テキスト整形	文字の表記を加工する
索引語生成	テキストを分割し、索引語を作る
索引構築	索引に索引語を格納する

3.1 収集

収集処理では、ロボット (robot)[9] と呼ばれるプログラムを実行し、WWW を巡回してページを収集する。ロボットは、自律的に WWW ページを閲覧して回ることから、スパイダー (spider) やクローラー (crawler) と呼ばれることもある。

どのページを収集し、どれだけ多くのページを蓄えるかは、検索結果に影響を及ぼす。収集されたデータは検索の母集合を決定し、収集されたページだけが検索される対象になるからである。収集ページ数は検索エンジンによって異なるが、Google では約 33 億ページである。

WWW ページは 1 度収集すればよいものではない。

更新が行われてページの内容が変化することもあるし、新しいページが追加されたり、存在したページが削除されることがある。そのため、ロボットは過去に収集したページに対して、定期的に巡回を繰り返す。巡回の間隔は短いことが望ましいが、検索エンジンが収集対象とするページは現在数十億ページに達しており、すべてのページを数週間単位で巡回することは不可能である。そこで、「ページの更新頻度に応じて巡回する」、「ページの重要度に応じて巡回する」などの巡回戦略が用いられている。

3.2 テキスト整形

WWW ページを収集した後は、HTML(Hyper Text Markup Language) からテキストを取り出す。どの部分のテキストを取り出して使うかは、検索結果に影響を及ぼす。索引に登録されないテキストは、検索の対象にならないからである。

表 4 に、WWW ページからテキストを取り出す個所を示す。

表 4　テキストを取り出す個所

テキスト	内容
本文	画面に表示されるテキスト
非表示部分	画像の alt タグに含まれる代替テキストなど
ヘッダ部分	title タグに記述されたページタイトル
	description タグに含まれる説明文
	keywords タグに含まれるキーワード
HTML 以外	PDF(Portable Document Format)、パワーポイントなどのファイルに含まれるテキスト

HTML のヘッダ部分には、ページの内容を示すメタ情報を記述することができる。本来は、ページ作成者が記述した「タイトル」、「著者」、「説明文」、「キーワード」などを本文とともに検索することは有効と考えられるが、実際にはタイトル以外は重要視されていない。これは、メタデータを記述しているページが全体の一部に過ぎないことと、後述する SPAM の影響を受けやすいためである。

取り出したテキストには、索引語に処理しやすいように前処理を施す。表 5 に、前処理で行なわれる処理の例を示す。

3.3 索引語の生成

大量のデータから高速に検索を行うために、WWW ページから取り出したテキストは、分割して索引に登録される。どのようにテキストを分割するかは、検索結果に影響を及ぼす。分割方法によって検索の漏れやノイズが変化するためである。代表的な索引語生成手法として、形態素解析 [17] と N-gram[16] が存在する。これらの詳細は後述する。

表 5　テキストの前処理

前処理	内容
HTML 表記	HTML 中に特別な表記で埋めた記号を戻す (例 ”>” →”>”)
ストップワード	英字の ”the” などを除く
特殊文字	検索に使われない特殊な文字を除く
助詞の除去	日本語の助詞を除く
表記の正規化	大文字と小文字、全角と半角、語尾変化処理 (ステミング)、「眞」と「真」など新字と旧字、「べ」と「ヴェ」など仮名の表記、「メール」と「メイル」など

3.4 索引構築

全文索引は、前処理によって整形された語を登録し、高速にテキスト検索を行う。索引手法としては、転置索引が広く使われている。他に、シグネチャ索引や SuffixArray 索引、Patricia Trie 索引などが存在する [17][21]。

4 全文検索

ユーザーから検索語を受け取ると、検索エンジンは索引を用いて検索を行い、結果集合を作る。

結果集合は、前処理や索引の仕組みによって質が変わることがある。

4.1 形態素解析

形態素解析では、文の構造を解析し、文を意味を持つ最小限の単語に分割する。単語を単位とするために、質のよい検索が可能である。

英語など、分かちされている言語は語の区切りが明確だが、日本語など分かちされていない言語では、テキストを検索の単位となる語に区切る処理が必要になる。

形態素解析では、図 3 に示すように、検索語によっては漏れが生じる可能性がある。形態素解析は、alltheweb[2]、goo[4]、Google[5]、infoseek[6] などで採用されている。

4.2 N-gram

N-gram では、文字単位で文を分割する。文を解析しないため、単語分割による検索漏れが生じない。

N-gram では、テキストから 1 文字ずつずらしながら

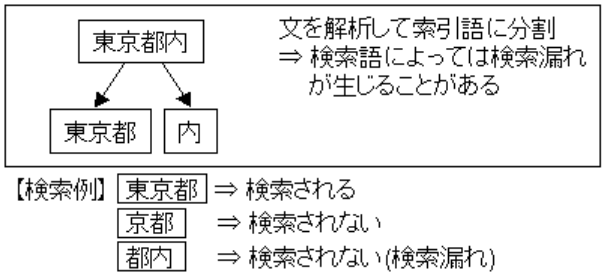

図 3　形態素解析と検索例

N 文字を切り出す形で索引語を生成する。N は通常 1 ～3 程度であるが、漢字、ひらがな、カタカナ、英字、数字などの字種によって N を可変とする方法もある。

N-gram では、図 4 に示すように、検索語によってはノイズが生じる可能性がある。N-gram は、AAA!CAFE[1]、AltaVista[3] などで採用されている。

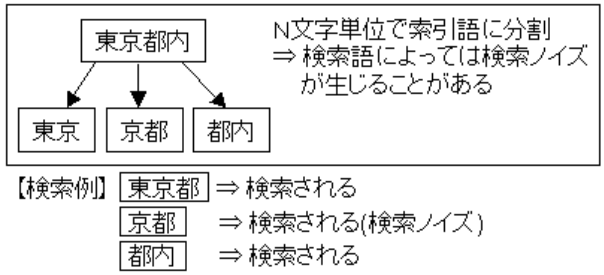

図 4　N-gram と検索例

5　ランキング

ユーザーは、検索の結果として、質のよいページを求めている。

仮にランキングのない検索を考えてみると、全文検索の結果がそのまま表示されることになる。結果集合は数百万件にも達することがあるため、その中から有用なページを探し出すことは大きな困難を伴うことになる。ランキングのない検索は、平積みのない本屋と考えることができる。どちらも大量の資料を提示するが、その中から資料を選択する行為をユーザーに投げることで、資料の評価を放棄してしまっている。

ランキングを行うために、検索結果に重み付けを行う処理をスコアリングと呼ぶ。表 6 に、検索エンジンで使われている代表的なスコアリング手法を示す。ここではスコアを計算する根拠によって、手法を 3 種類に分類した。

1 つ目は、データの特性を利用した計算方法である。

表 6　代表的なスコアリング手法

基になる情報	スコアリング手法	類似の概念
データの特性	出現頻度、タグ、出現位置、近接度	カーナビゲーション、鉄道経路探索
ユーザーの行動	クリック人気	ベストセラー情報
ユーザーの推薦	リンクポピュラリティ	文献の引用情報

ページ中に含まれるキーワードの数や更新日時など、データそのものが持つ特性を利用している。類似の概念としては、目的地までの行き方を示す経路探索がある。最適な経路は、路線や道路が持つ距離、運賃、所要時間などの特性を基に、アルゴリズムによる計算で求めることができる。

2 つ目は、ユーザーの行動記録を利用した計算方法である。検索エンジンの表示から、実際にどのページへのリンクがクリックされたかをカウントし、スコアに反映する。類似の概念としては、出版などのベストセラー情報がある。多くのユーザーが利用するデータは、他の人にとっても有用であることが多い。

3 つ目は、ユーザーの推薦を利用した計算方法である。HTML のリンクを他のページへの推薦とみなし、リンクされたページのスコアに反映する。類似の概念としては、文献の引用情報がある。多くの文献で引用される文献は、他の人にとっても有用であることが多い。

従来は、ページの特性によるスコアリングが用いられていた。現在は、ユーザー行動やユーザー推薦など、人間による評価をスコアリングに取り入れる流れが一般的である。検索エンジンは、これらの手法をさまざまに組み合わせて使用しており、スコアリング手法の改良が続けられている。

5.1　キーワード出現頻度

重要な役割を果たすキーワードは、ページ中に何度も出現することが多い。この考えに基づいた重み付けがキーワード出現頻度である。

単純にはページ中にキーワードが出現する回数を数えるが、ページサイズで補正したり、少ない文書だけに現われるキーワードを優先して扱う補正処理を行うことがある。

5.2　タグごとの重み付け

HTML はタグによって構造化されている。重要なタグには重要なキーワードが記述されることが多い。この考えに基づいたスコアリングがタグごとの重み付けである。図 5 に、HTML と表示例を示す。

```
<html lang="ja">
<head>
<title>ページタイトル</title>
<meta name="keywords" content="形態素解析,N-gram,索引">
<meta name="description" content="検索アルゴリズムのページです!">
</head>
<body>
<h1>検索アルゴリズム</h1>
<p>重要な概念は、ページや段落の先頭付近に書かれる場合が多いものです。</p>
<p>本文中では、重要な概念が<em>強調</em>や<strong>より強調</strong>されたり、
<a href="link.html">ハイパーリンク</a>で示されていることがあります。</p>
</body>
</html>
```

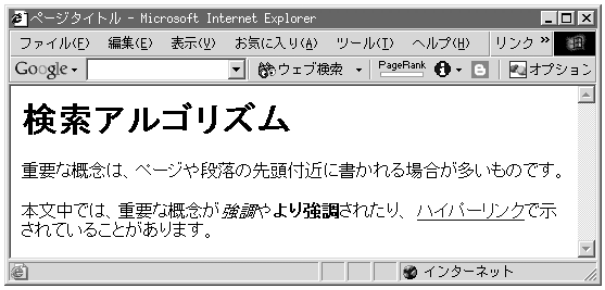

図 5　HTML と表示例

HTML のヘッダ部には、そのページに関する各種のメタ情報が記述されている。title タグには、そのページのタイトルが記述されており、多くの検索エンジンが検索の際に重視している。meta タグには、そのページを表す説明文 (description)、キーワード (keywords) などが記述されており、以前は利用する検索エンジンが多かった。しかし、メタ情報が記述されているページは全体の 10% 程度に過ぎないため、有効な検索手段として活用するのが難しいのが現状である。

HTML の body 部には、Web ブラウザが表示する内容が記述されている。h1〜h6 タグは、節の見出しを表す。em, strong や b, i, font などのタグは、テキストを強調したり他の部分と異なる書式で表示する。a タグは、リンクを表す。これらのタグには重要な内容を表すテキストが記述されることが多いことから、重視する検索エンジンが多い。

5.3　出現位置

重要な概念はページや節の先頭付近に存在することが多い。この考えに基づいたスコアリングが出現位置による重み付けである。

ページの先頭付近、h1〜h6 タグの直後などを重視することが多い。

5.4　キーワードの近接度

重要な複数のキーワードは、互いに近い位置に存在することが多い。この考えに基づいたスコアリングが近接度による重み付けである。

複数のキーワードが指定された場合、ページ内でキーワードの出現位置が近いほど高いスコアが与えられる。Google などが採用している。

5.5　クリック人気

多くの人に参照されるページは、他の人にとっても有用であることが多い。この考えに基づいたスコアリングがクリック人気による重み付けである。

検索結果の表示画面から、ユーザーがページへのリンクをクリックした回数を記録し、スコアに反映する。goo などが採用している。

クリックされた回数だけでなく、そのページを閲覧している滞在時間をスコアに反映する方式が提案されている。Teoma[7] などが採用している。

5.6　リンクポピュラリティ

多くのページから参照されるページは、有用であることが多い。この考えに基づいたスコアリングがリンクポピュラリティである。

以前は、ページ間のリンク数を単純に集計する手法が提案されていたが、現在では、参照元のページによってリンクに重み付けをする Google の PageRank[14][20] などの手法が主流である。使われているキーワードなどからサイトのテーマを推測し、リンクの重み付けに反映する WiseNut[8] の WiseRank などの手法も提案されている。

6　今後の課題

検索エンジンの検索アルゴリズムは改良が続けられている。

6.1　公正なランキングの維持

検索結果の先頭ページは、ユーザーの目に触れる確率が高いため、商品価値が生まれている。

商用ページは、先頭ページに表示されるように SEO(Search Engine Optimization) 技術を使い HTML に修正を施すことがある。必ずしも有用でないページが上位にランクされるようになると検索の質が損なわれることから、検索エンジンではアルゴリズムの改良や特定サイトへのペナルティ (スコアを下げる措置) などにより、不当な SPAM サイトへの対策を行っている。

6.2　概念による検索

ユーザーが指定した検索語が、有用なページに必ず含まれているとは限らない。

一部の検索エンジンでは、関連するページを検索する試みがある。ひとつは同義語検索である。検索に同義語辞書を適用することで、「コンピューター」で「電

子計算機」を検索できるようになる。もうひとつは概念による検索である。リンク解析などにより、検索語や同義語が含まれないページを含め、関連するページを検索することが可能である。

6.3 メタ情報の活用

ランキングにおいて人間の判断や行動の利用が有効であるように、検索においてもデータそのものだけでなく、人間の判断が有効と考えられる。

従来も Yahoo のようなディレクトリ階層を利用した WWW の分類が提供されていた。現在では、それに加えてインターネット資源を記述する Dublin Core[12] などの書誌フォーマット、RSS(RDF Site Summary)[11] によるサイト情報の公開、セマンティック Web[10] によるメタ情報の有機的なネットワーク構築などが進められている。今後の動向に注目したい。

6.4 国内の動向

検索エンジンの発展には、健全な競争関係が不可欠である。

世界的に検索エンジンの再編が進んでいる。国内においては、今まで特色のある開発を行っていた大手の検索エンジンが次々と撤退を表明し、検索サービスで使われる検索エンジンが Google 一色になりつつある。すでに、NETPLAZA(NEC)、Infonavigator(富士通)、ODIN(NTT) の検索エンジンが姿を消し、2003 年 12 月には goo も自前の検索エンジンを終了する予定である。しかし、競争のないところに進歩はない。

明るい話題としては、一部で評価の高かった kensaku.org の検索エンジンが、AAA!CAFE[1] として復活したことがある。索引語の生成に N-gram を採用し、漏れのない検索を実現している。今後収集したページ数が増えてくれば、他の形態素解析を用いた検索エンジンと組み合わせた利用法も考えられる。

7 おわりに

以上、検索エンジンの裏側で使われている検索アルゴリズムを解説した。

検索アルゴリズムの性能は、検索に必要なページを収集し、ノイズや漏れのない検索を高速に行い、適切にランキングして表示することで成り立っている。

検索エンジンの検索アルゴリズムは、情報検索の技術をベースに発展を続けている。ランキング技術を中心に、情報検索へのフィードバックも行われつつある。今後の発展に期待したい。

参考文献

[1] AAA! CAFE. http://www.aaacafe.ne.jp/.

[2] alltheweb. http://www.alltheweb.com/.

[3] AltaVista. http://www.altavista.com/.

[4] goo. http://www.goo.ne.jp/.

[5] Google. http://www.google.com/.

[6] infoseek. http://www.infoseek.co.jp/.

[7] Teoma. http://www.teoma.com/.

[8] WiseNut. http://www.wisenut.com/.

[9] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. *ACM Transactions on Internet Technology*, Vol. 1, No. 1, pp. 2–43, August 2001.

[10] Tim Berners-Lee. What a semantic can represent. http://www.w3.org/DesignIssues/RDFnot.html.

[11] RSS-DEV Working Group. RDF Site Summary (RSS) 1.0. http://purl.org/rss/1.0/ spec/.

[12] ISO. 15836:2003(E) information and documentation — the dublin core metadata element set. http://www.niso.org/international/SC4/n515.pdf.

[13] Steven Kirsch. Infoseek's experiences searching the internet. *SIGIR Forum*, Vol. 32, No. 2, pp. 3–7, 1998.

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[15] 浅井勇夫. Google に隠れた秘密あり！ KWIC 方式の紹介文. 検索デスク. http://www.searchdesk.com/view/vptc323.htm.

[16] 小川泰嗣. N-gram 索引における複合検索条件の効率的な処理方法. Technical report, 株式会社リコー研究開発本部, 1999. http://www.ricoh.co.jp/rdc/techreport/No25/index.html.

[17] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002.

[18] 久野高志, 安形輝, 上田修一. 情報検索システムとしてみたサーチエンジン. 第 49 回日本図書館情報学会研究大会, 2002.

[19] 原田昌紀, 佐藤進也, 風間一洋. 索引篩法 - 大規模サーチエンジンのための高速なランキング検索法. In *DEWS2003*. 情報処理学会, 2003.

[20] 馬場肇. Google の秘密 – pagerank 徹底解説. http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html.

[21] 山本毅雄, 橋爪宏達, 神門典子, 清水美都子. 全文検索. 丸善, 1998.