平成 20 年度 卒業論文

論文題目

# 検索エンジンはどのようにして有用なサイトと有用でないサイトを判別しているのか

神奈川大学 工学部 電気電子情報工学科

学籍番号 200502716

林 潤

指導担当者 木下宏揚 教授

# 目次

# 第１章

# 序論

近年、インターネットの普及に伴いGoogleやYahooなどの検索エンジンを利用するユーザは日々増えている。
そのユーザに企業のホームページや販売サイトに来てもらうためには検索エンジンの上位に表示されることが必要である。
実際、あるキーワードで１ページ目（1～10位）にそのサイトが表示されるだけで、そのサイトへのアクセス数は劇的に違い、検索エンジンの結果次第でサイトへの集客力が決まるといってもいい過ぎではない。こうしたことから、現在は各企業が、検索結果の上位に自社サイトが表示されるようにしのぎを削っている。
このように、ある特定の検索エンジンを対象として検索結果でより上位に現れるようにウェブページを書き換えることを検索エンジン最適化[参考文献１]と呼ぶ。

現状として検索エンジンは、数あるサイトの中から有用なサイトと有用でないサイトを判別して順位付けを行っているのだが、正式な判別基準の情報は公開されていない。最も近い情報として、Googleページランクの論文[参考文献２]やGoogle創始者の論文[参考文献３]が存在するだけである。実際に検索エンジン最適化と検索して調べてみても判断基準の情報はサイトによって違い、それが正しいのかどうかは不明である。

そこで本稿では、既に上位表示されているサイトなどから検索エンジンの判別要因を予想し、最も集客できるとされる、あるキーワードでの検索順位の１ページ目（1位～10位）以内に自分で作成したサイトが表示されることを目標に研究を行う。

具体的には、日々検索エンジンのアルゴリズムの変化を追い、上位に表示されているサイトの特徴をまとめたものから、検索エンジンが有用かどうかを判別するための要因をまとめる。（第３章、第４章）さらに、実際にそれに沿ったサイトを作ることでyahooやgoogleの上位に表示させる。（第５章）

こうすることによって検索エンジンに好まれるサイトと好まれないサイトの傾向が分かり、どのようにして検索エンジンが有用なサイトと有用でないサイトを判別しているのかが理解できる。

以下、第２章では、検索エンジンについての基礎知識について述べる。また、第３章で検索エンジンが判別を行っている内部要因について述べ、第４章で検索エンジンが判別を行っている外部要因について述べる。さらに、第５章でこれらの要因に基づいて作成した試験的なページを元に実験と考察について述べ、最後に第６章でまとめる。

# 第 2章

# 基礎知識

## 2.1　検索エンジンの目的

検索エンジンの目的は「快適な検索体験の提供」である。

完璧な検索エンジンを構築するには、

①ユーザーが何を探し求めているのか、その検索意図（インテント）を正確に把握すること。
②一つ一つのウェブページが何について記述されているか、その内容を正確に把握すること。

が必要になる。[参考文献４]

世界で最も利用されている検索エンジンのGoogleでは、Webサイト上で、次のようなミッション・ステートメントを掲げている。

「Google の使命は、Google 独自の検索エンジンにより、世界中の情報にアクセスを可能にし、Web 上の検索経験をより実りのあるものにすることである。」

（英語：Google's mission is to deliver the best search experience on the Internet by making the world's information universally accessible and useful. ） [参考文献５]

つまり、ユーザーが探している情報に対する的確な答え・情報を提示すること、検索キーワードに対する検索結果の関連性を最大限に高めること、これが検索エンジンの最終ゴールであり目標である。


図. 1　Yahooのトップページ


図. 2　Googleのトップページ

## 2.2　CrawlerとIndex

Crawlerとは

・Web上を自動的に巡回してWebページを収集する検索ロボットプログラムのこと。
一般にクローラは、既知のHTML文書の新しいコピーを要求し、文書中に含まれるリンクをたどり別の文書を収集するという動作を繰り返す。新しい文書を見つけた場合はデータベースに登録する。また、既知のファイルが存在しないことを検出した場合はデータベースから削除する。[参考文献６]

Indexとは

・Crawlerが収集したWebページを解析し、検索に使用するための情報を集めた保存庫のようなもの。[参考文献７]

以上を踏まえると検索エンジンの動向は以下のようになっている。（図３）

（ｉ）Crawlerが常にサイト内のリンクを伝って巡回している。
（ii）Crawlerは、その情報をまとめてIndexを作成する。
（iii）検索エンジンは、ユーザが検索したキーワードに対して、Algorithmにしたがっての情報から「関連性の高い」順に表示する。

検索エンジンは、「快適な検索体験の提供」つまり、ユーザの要求（キーワード）に対して適している全てのサイトを関連性の高い順に表示することが目的なので（ｉ）～（iii）の動作を常に繰り返し行っている。

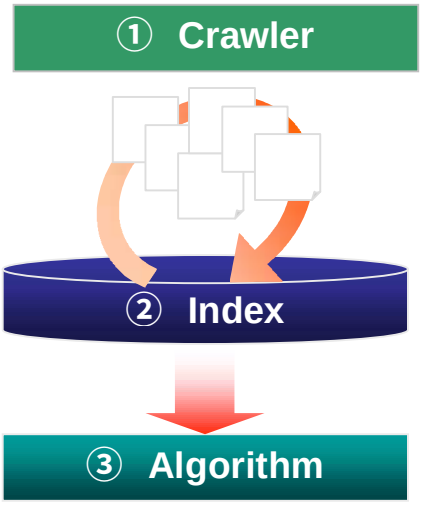



図.３　検索エンジンの動向

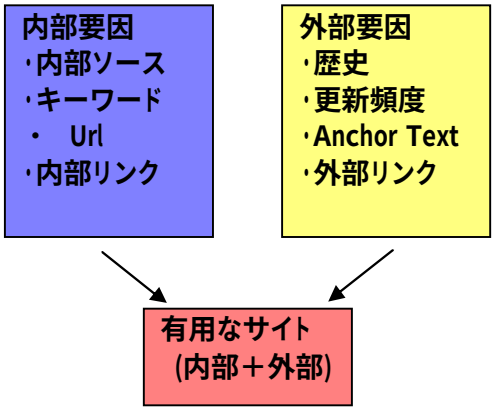



図.４　内部要因と外部要因

## 2.3　有用なサイトの判別要因

検索エンジンのAlgorithmが有用なサイトと有用でないサイトを判別するには、内部要因（on the page factors）と外部要因（off the page factors）の２つの要因がある。内部要因と外部要因の概要は以下のようになると予想される。

内部要因は、HTMLタグ（内部ソース）、キーワード含有率（キーワード）、Domain名（URL）、内部リンク構成によって判別している。一方、外部要因は、ページができてからの経過年数（歴史）、更新頻度、被リンクの際のText（Anchor Text）、外部リンクの質と量によって判別している。（図４）

これらの要因を、第３章では内部要因について、第４章では外部要因についてそれぞれ説明していく。

# 第 3章

# 内部要因

## 3.1　内部ソース

内部ソースには、検索エンジンのCrawlerに読み取ってもらうためのHeadとページに来たユーザーに見てもらうためのBodyの２つに分けられ、どちらもHTML形式で記述する。

ここでいう内部ソースとはHead情報のことである。

有用なサイトと判別されるためにはHeadを以下のように記述する。

```
1・・・・・<head>
2・・・・・<META http-equiv="Content-Type" content="text/html; charset=Shift_JIS">
3・・・・・<meta name="robots" content="all">
4・・・・・<title>タイトル</title>
5・・・・・<meta name="Keywords" content="キーワード１,キーワード２,キーワード３,キー
6・・・・・ワード４,キーワード５・・・">
7・・・・・<meta name="Description" content="サイト説明">
8・・・・・</head>
```

ここで行ごとの解説を行う。

２・・・・・ページがシフトJISで作られている場合、このタグでShift_JISを指定する。この文字コードを指定することによって文字化けを防ぐ。（日本語のサイトのほとんどがシフトJIS。）

３・・・・・Crawlerへの指示をするタグである。このタグは、<meta name="robots" content="index,follow">と同じ意味で、
「このページはindex可能で、このページからのリンク先ページもindex可能」
というのを検索ロボットに指示している。

４・・・・・タイトルタグといい、この部分が検索エンジンの検索結果のタイトルになる。
このタイトルタグは、検索エンジンがどのようなサイトであるかを判断するための１番重要なタグである。

５、６・・・・・メタキーワードといい、サイトの内容に適したキーワードを記述する。
このキーワードは、
キーワード１＞キーワード２＞キーワード３＞・・・
のように、キーワード１から徐々に優先度が下がっていく。

７・・・・・メタディスクリプションといい、サイト説明を記述する。この部分が検索エンジンの検索結果の説明部分になる。

## 3.2　キーワード

検索エンジンは、キーワード出現頻度を重要視する。
これはそのページ内のテキストとキーワードの関連性を測る最も基本的な指針だからである。

例えば「検索エンジン」というキーワードに対して、例１、２のような文章があったとき、

例1)「検索エンジンにはディレクトリ型とロボット型があります。ロボット型はGoogleのようなもので、ディレクトリ型はYahooのようなものです。」

例2)「検索エンジンにはディレクトリ型検索エンジンとロボット型検索エンジンがあります。ロボット型検索エンジンはGoogleの ようにロボットが自動で集めてきたデータの中から検索する検索エンジンです。ディレクトリ型検索エンジンはYahooのようにエディタによって手動で登録される検索エンジンです。」

同じ内容の文章でもキーワード出現頻度が異なるため、例２の文章ほうが検索エンジンに重要視される。

さらに分かりやすく記述すると、以下のようになる。

例1)
検索 | エンジン | に | は | ディレクトリ | 型 | と | ロボット | 型 | が | あります | ロボット | 型 | は | google | の | よう | な | もので | ディレクトリ | 型 | は | yahoo | の | よう | な | も | の | です。

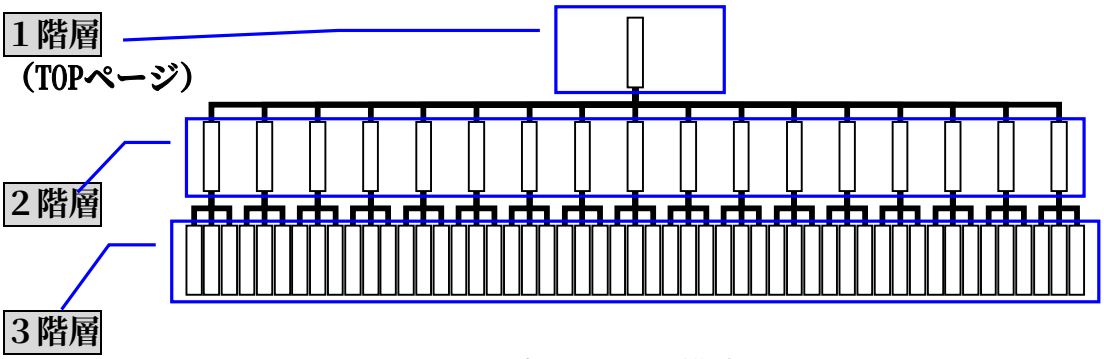
例2)
検索 | エンジン | に | は | ディレクトリ | 型 | 検索 | エンジン | と | ロボット | 型 | 検索 | エンジン | が | あります | ロボット | 型 | 検索 | エンジン | は | google | の | よう | に | ロボット | が | 自動 | で | 集めて | きた | データ | の | 中 | から | 検索 | する | 検索 | エンジン | です | ディレクトリ | 型 | 検索 | エンジン | は | yahoo | の | よう | に | エディタ | によって | 手動 | で | 登録 | される | 検索 | エンジン | です

実際にキーワードの出現率を算出して見ると、例１の方は「検索」「エンジン」ともに文章中に3%あるのに対し、例２の方は「検索」が14%、「エンジン」が12%あり、例２のほうがキーワードの出現率が高いことが分かる。

## 3.3　内部リンク

サイトを作成する際は、図５にある理想のリンク構成にあるように、１階層、２階層、３階層というように階層をつけて内部リンクを作ると有用なサイトだと判別されやすい。この内部リンクは、リンクが多ければ多いほど有用なサイトと判別される。

階層について説明すると、例えば、１階層目が、
http://1.com
だった場合、２階層目というのは、
http://2.1.com　または、 http://1.com/2/
になる。
同様に、３階層目というのは、
http://2.1.com/3/ または、 http://1.com/2/3/
になる。




図．５　理想のリンク構成

さらに、これを図６にあるように、どの階層を見てもリンクが張り巡らされているメッシュ（網目状）リンク構造にするとCrawlerに認識されやすく、有用なサイトと判別されやすい。
（図６では省略しているが、３階層目からTOPに戻れるリンクもつける。）

具体的には、メッシュリンク構造を構築するために、全てのページに「ランキング」「特集」「オススメ」といったナビゲーションリンクを用意することによって全てのページから、直接リンクを飛ばす。（少なくとも３クリック以内に全てのページに辿り着ける構成にする。）

また、このとき「ランキング」を「総合ランキング」と「ジャンル別ランキング」の２つ作るなど工夫することによってさらにメッシュリンク構造を細かくすることができる。

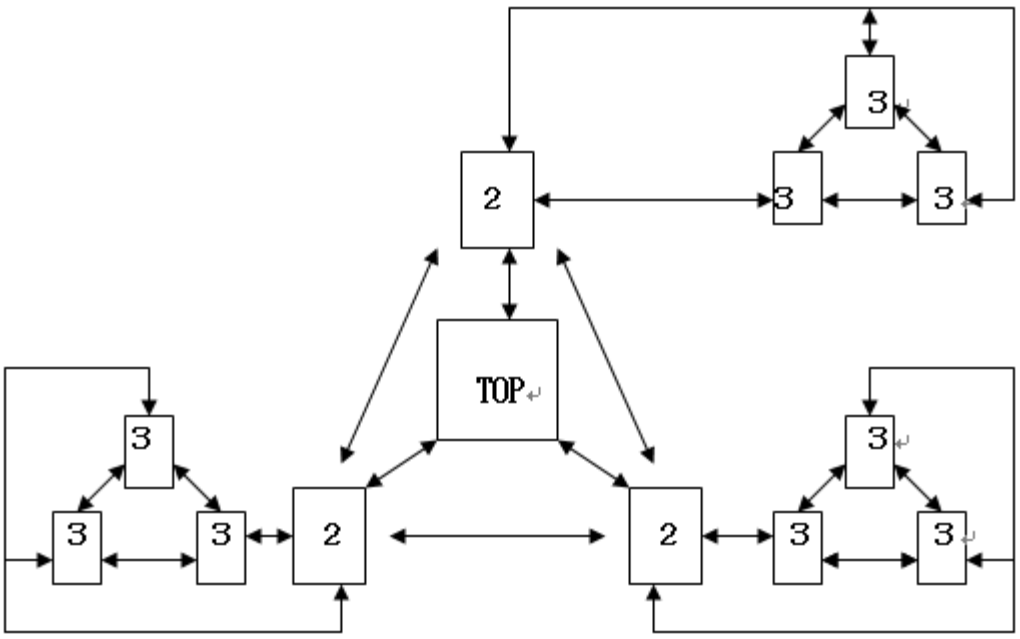



図.６　メッシュリンク構造

## 3.4　URL

検索エンジンが有用なサイトかどうかを判断する要因で１番重要なのはタイトルタグであるが、次に重要なのは、URLといわれている。
例えば、神奈川大学のサイトを作成する時は、
http://kanagawa-daigaku.co.jp（実際は、http://www.kanagawa-u.ac.jp/）
のようになるべくキーワードと同じ意味のURLにしたほうが有用なサイトだと判別される。

最近では、日本語URLというのが普及してきていて、ウィキペディア、amazonは既に日本語URLを導入している。（図７）


図. 7　日本語URLの例

# 第 4章

# 外部要因

## 4.1　歴史

新しいドメイン（URL）を取得してすぐにGoogleやYahooにサイト登録しても、一定期間は上位に掲載されない現象をエイジングフィルタというように、ドメインには歴史が関係している。
これは単純に、古いドメインほど上位に表示されやすいことを意味している。

もし、新たにサイトを開設する予定がある場合は、あらかじめドメインだけを取得しておき、空ファイルでアップロードして検索エンジンにクロールさせておくと良い。

## 4.2　更新頻度

サイトは「更新頻度が高いほど上位表示される」といわれている。
具体的には、文章やデザインやページ数など、コンテンツの追加の更新をしていくうちに有用なサイトだと判別される可能性は上がる。

しかし、タイトルやメタ情報の変更のみを行ったり、更新頻度が多すぎるなど、明らかに意図的に検索結果の順位を上げようとしていることが分かる更新は、それが原因で順位を落とすことがある。

検索エンジンは「明らかな順位対策を行っているサイトにはペナルティを与える」ように出来ているので、更新する際にはこのことに注意する必要がある。

## 4.3 Anchor Text

Anchor Textとは、リンクが設定されている部分に書かれた文字のことである。
ページ同士の結びつきは、このリンクによって成り立っているので、Crawlerがリンクを辿ってページを探す際、リンク先のページがどのようなページであるか認識する際に、このAnchor Textを参考にします。
このAnchor Textの内容とページの内容が一致しているほど、これらの結びつきが強いと検索エンジンは判断します。
つまり、ページの内容と一致したAnchor Textによって貼られているリンクの数が多いページほど評価の高いページとして認識されます。

内部リンクを構成する際は、このAnchor Textに注意して構成する必要がある。

具体的には、トップページへのAnchor Textは「トップ」や「Home」などにせず、キーワードをいれて、「キーワードのトップへ戻る」のようにする方が良い。

※　このAnchor Textは外部要因と記しているが、内部リンクで記述することができるので、内部要因ともいえる。

このAnchor Text について分かりやすい例がある。
YahooやGoogleで「出口」と検索すると、YahooやGoogle自身が１位に表示されるが、これは、このAnchor Textが原因で起こる現象である。

図８のような成人向けのサイトの場合、入口と出口を設け、「出口」というAnchor TextにはGoogleやYahooにリンクしているサイトが多く存在するため、出口→Yahoo、出口→Googleというリンクは、検索エンジンには「出口」というサイトなのだと認識されてしまう。

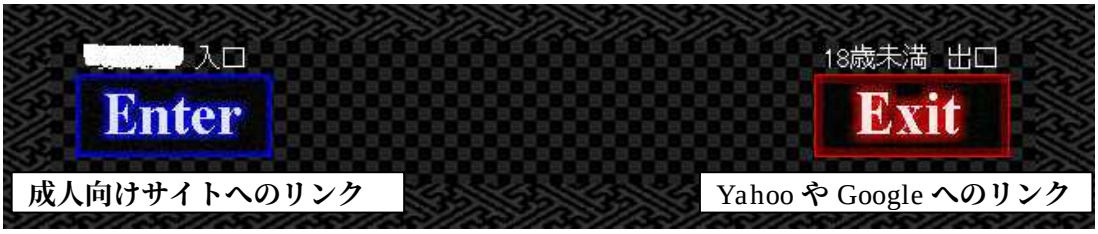



図.８　成人向けのサイトの例

## 4.4　外部リンク（ページランク）

ロボットが、有用なサイトかどうかをサイトの内容だけから判別することは困難である。
そのため、質の高い検索結果を返すために、外部リンク（ページランク）を使って判別している。
「多くの良質なページからリンクされているページは、やはり良質なページである」
というGoogleの基本概念はこの外部リンク（ページランク）のことである。

ページランクは、次の３つの観点によって測っている。

①被リンク数
②ページランクの高いページからのリンク（Yahoo!など）
③リンク元の被リンク数

ページランクの概念図は図９のようになる。

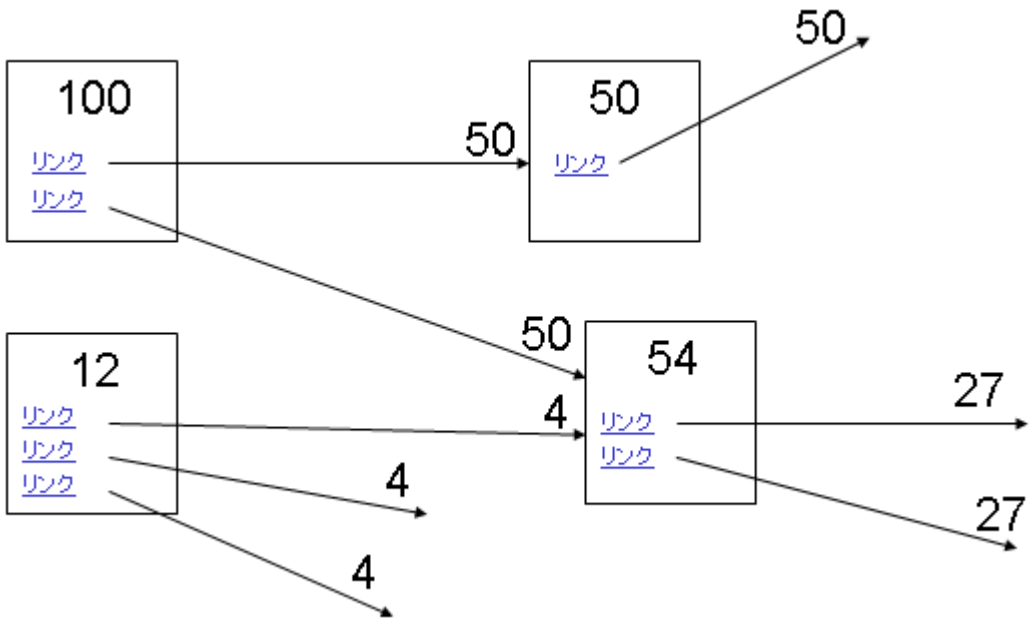



図.９　ページランクの概念図

ページランクは以下のようにして作られた。

①ネットサーファーを考える。
②Web上のリンクをランダムに進んでいく。
③決して前のページには戻らない。
④時折全く関係のないページへ飛ぶ。（図１０）

つまり、ページランクとは、ネットサーファーが、あるサイトを訪れる確率である。

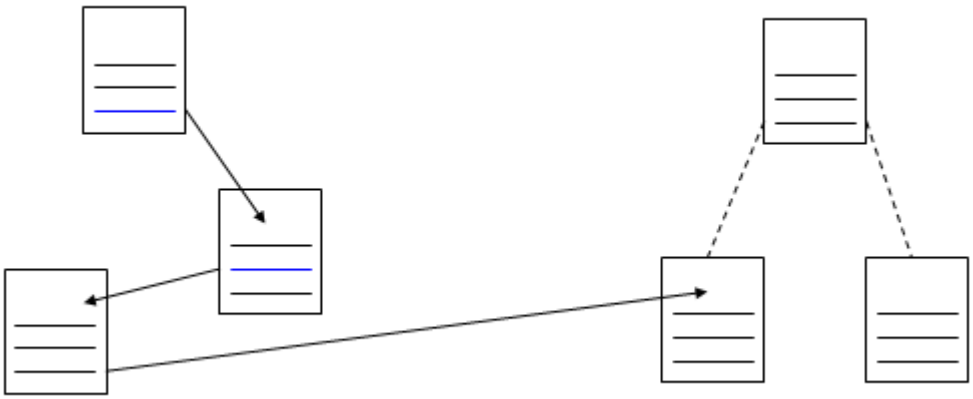


図.１０　突然無関係なページへ移動

# 第５章

# 実験と考察

第３章と第４章で述べた、内部要因と外部要因に沿って「大学偏差値ランキング」というキーワードで上位表示を狙ってサイトを作成（2008/5/21）した。
その結果、2008/1/25時点で、
「Google９位、Yahoo３３位、Googleモバイル１位、Yahooモバイル２位」
に表示された。

内部要因対策を施した時点では検索エンジンにIndexさえされていなかったが、外部要因対策を施した直後にIndexされたので、検索エンジンは外部要因対策を施さないとCrawlerによってIndexされないことが分かった。
また、PCとモバイルで検索エンジンの順位の違いが生じたのは、PCとモバイルでは、外部要因と内部要因の判断基準が違うからだと考えられる。
本実験では、内部要因対策は十分に行えたが、外部要因対策の内、歴史などは期間的に短く不十分であった。
このことから、PCの検索エンジンは外部要因を重視するのに対し、モバイルの検索エンジンは内部要因を重視することが順位から推測できる。
また、GoogleとYahooで順位の違いが生じたのも、Googleは比較的内部要因を重視するのに対し、Yahooは比較的外部要因を重視することが順位から推測できる。

# 第６章

# まとめ

「検索エンジンはどのようにして有用なサイトと有用でないサイトを判別しているのか」というテーマでこれまで研究を行ってきたが、有用なサイトというのは人間と似ている部分が多くあると感じた。

例えば、人気のある人というのは、内面も外面も優れている。
人に関して、内面は「性格、家柄、人脈」、外面は「顔、評判」などがあるが、これらは、
「性格→内部ソース、家柄→Urlや歴史、人脈→内部リンク」・・・内部要素
「顔→サイトデザイン、評判→外部リンクやAnchor Text」・・・ 外部要素

のようにサイトの内部要素と外部要素に置き換えることができる。
検索エンジンは常に進化し続けているが、いつの時代になっても人間が内面と外面両方の観点から判断されるのと同じように、検索エンジンも内部要因と外部要因の２つから判別されていくと考えられる。

ただ、今後検索エンジンは、個人によって異なる検索結果を表示するパーソナライズ検索、検索した場所によって異なる検索結果を表示するGPS連動検索、動画や地図や画像などさまざまなコンテンツを表示するユニバーサル検索なども用いて有用なサイトかどうかを判別していくと予想されるのでそれらについても今後検討していく必要がある。

# 謝辞

本研究を進めるにあたり, 終始熱心に御指導していただいた木下宏揚教授と鈴木助手に心から感謝致します。また、良き研究生活と良き研究環境を提供して頂いた木下研究室にも心から感謝致します。

# 参考文献

[1] Wikipedia「検索エンジン最適化」
(http://ja.wikipedia.org/wiki/%E6%A4%9C%E7%B4%A2%E3%82%A8%E3%83%B3%E3%82%B8%E3%83%B3%E6%9C%80%E9%81%A9%E5%8C%96)

[2] Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen “Combating Web Spam with TrustRank”

[3] Sergey Brin and Lawrence Page “The Anatomy of a Large-Scale Hypertextual Web Search Engine”

[4] Google という検索エンジンの目的は？
(http://faq.sem-research.jp/1/20040229130324.html)

[5]Google株式会社概要 (http://www.google.co.jp/intl/ja/profile.html)

[6] Wikipedia「Crawler」
(http://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AD%E3%83%BC%E3%83%A9)

[7] 検索エンジンの仕組み(http://www.seo119.com/measure/system.html)

# 質疑応答

・ 検索エンジン最適化（SE0）との違いは？（能登先生）

内部要因と外部要因に自分でカテゴリ化して，まとめたことが違います.

・ どういう応用が考えられるか？（豊嶋先生）

企業サイトや商用サイトで同じようにサイトを作成してアクセスアップ（集客数アップ）を狙うことができます.