

1. 検索エンジン～Web 情報の検索の概要～

インターネットで公開されている Web ページを対象とした情報検索システムを検索エンジンと呼ぶ。既に Yahoo や google などの検索エンジンを使って Web 情報を収集し、その情報を利用した経験があると思われるが、このレポート作成をとおし、複数の検索エンジンを使用することにより、検索エンジンの特徴を利用した多様な情報収集が可能になること、および英語情報にも有用な情報が存在すること、さらに英語情報の捉え方が日本語の情報と大きく異なる場合があることについての理解を深める。

検索エンジンには、キーワードを使って情報を検索するキーワード方式と、予め提示されたカテゴリと呼ばれる項目リストから必要な情報を選択するカテゴリ方式がある。キーワード方式の検索エンジンでは、ロボットとよばれる Web 情報収集ソフトにより Web サイトを巡回して Web 情報を収集し、収集した情報をデータベースにして、検索システムを提供している。収集するデータの主体は一般のデータベース同様にテキストデータである。データの収集、検索方法、検索結果の表示など検索エンジン毎に特徴を有したシステムとなっている。

2. 英語と日本語の情報

Internet Domain Survey, Jan 2009(2008)によると、インターネットのホスト数は全世界で 625,226,456(541,677,360)台である。ドメイン別でみると、1 位が net で 204,683,342(190,267,719)台、2 位が com で 123,324,475(95,448,209)台、そして 3 位が jp(日本)で 43,461,277(36,803,719)台である。国別ドメインは jp (日本)が第 1 位であるが、割合では、全ホスト数の 6.95(6.79)%に過ぎない。アメリカのホストには国別ドメインが使用されていないが、net, com, edu の大半はアメリカの情報と考えられることから、日本語の情報よりはるかに多い英語情報が Web 情報として流通しているといえる。そのため、情報収集にあたっては英語の情報も対象に含める必要がある。

年月	ホスト数
2009 年 1 月	625,226,456
2008 年 1 月	541,677,360
2007 年 1 月	433,193,199
2006 年 1 月	394,991,609
2005 年 1 月	317,646,084
2004 年 1 月	233,101,481
2003 年 1 月	171,638,297
2002 年 1 月	147,344,723
2001 年 1 月	109,574,429
2000 年 1 月	72,398,092
1999 年 1 月	43,230,000
1998 年 1 月	29,670,000
1997 年 1 月	16,146,000
1996 年 1 月	9,472,000
1995 年 1 月	4,852,000
1994 年 1 月	2,217,000
1993 年 1 月	1,313,000

図1 インターネットのホストサーバの数

出所: Internet Systems Consortium, Inc. (<https://www.isc.org/solutions/survey>)

3. キーワード方式の検索エンジンによる Web 情報検索

キーワード方式の情報検索は通常絞り込み (AND) 検索により情報検索を行う場合が多いが、多くの検索エンジンで AND, OR, NOT のブール演算子を使用することができる。検索結果はそれぞれの検索エンジンが決める Web ページのスコア得点の高い順に表示される。情報件数が膨大なための確なキーワードを選択する必要がある。複数の検索エンジンを使用すると目的とする情報を探し出すことができる場合がある。特定の情報、個別の情報、特殊な情報を探し出すのに適している。ここでは、論理式を使った検索が可能であり、かつ検索結果の件数が表示される検索エンジンを使用する。対象となる Web ページは 2005 年で 200-500 億ページと推定されており、2009/05/04 現在、cuil (<http://www.cuil.com/>)では1,244億ページ(124,426,951,803 web pages)となっている。

多くの検索エンジンが Web 上に存在するが、情報検索の授業では図 2 に示す、論理式が使用可能で、検索結果件数が表示される、データベースの基本機能を持った検索エンジンを使用する。

番号	日本語	英語
3-1	Yahoo Japan	Yahoo
3-2	Google	Google
3-3	ASK	ASK
3-4	alltheweb	alltheweb
3-5	AitaVista	AitaVista
3-6	goo	
参考	cuil	cuil

図 2 論理式が使用可能で検索結果件数が表示されるキーワード方式の検索エンジン

3-1 Yahoo

【日本語】 www.yahoo.co.jp

1994 年にアメリカで設立され 1996 年に J a h o o J a p a n が開始された。現在に至るまで国内において最も利用頻度の高い検索サイトとなっている。検索の窓に複数のキーワードをスペースで区切って入れると AND 検索となる。1 回目の検索結果が表示され、検索ボタンの隣に [条件を指定して検索](#) と表示される。条件を指定して検索をクリックすると、以下の画面となる。上から順に AND 検索、フレーズ検索、OR 検索、NOT 検索に対応する。別の検索にカテゴリ検索がある。

全て含む	AND	<input type="text" value="大学 新潟"/>	ページ内のすべて ▾
順番も含め完全に一致	フレーズ	<input type="text" value="大学生と就職"/>	ページ内のすべて ▾
少なくとも一つを含む	OR	<input type="text"/>	ページ内のすべて ▾
含めない	NOT	<input type="text"/>	ページ内のすべて ▾

【英語】 www.yahoo.com

日本語同様に 1 回目の検索結果画面にある、検索ボタンの隣に options が表示される。options をクリックし Advanced Search を選択すると論理式検索が可能な検索画面が表示される。

3-2 Google [Nasdaq:GOOG] (キーワード方式) <http://www.google.co.jp/>

【日本語】 www.google.co.jp

「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

PageRank™ のコンセプトに基づき、他からリンクが多く張られている Web ページを検索結果の上位に出力する。結果の出力順位は、①リンクが張られている数（バックリンクの数, link juice)、②リンクが張られているときの説明文（アンカーテキスト)、③Web ページ内のキーワード、④ドメインの信頼性を重視して、検索結果の順位が決められていた。結果出力表示の順位は、検索者が求める順と一致する 경우가多く、ロボット型検索エンジンでも検索者の重要度に近い検索結果が得られた。

2002 年ごろから本格的なサービスを開始したが、それ以前の検索エンジンが達成できなかった 10 億ページを越す収集件数と、検索結果の順位付けの妥当性により Yahoo を凌ぐ検索エンジンとなった。サービス開始当時①, ②, ③, ④の順のウエートで順位が決められていたが、最近は、④、②、③、①の順でウエート付けされている。①バックリンクの評価は低くなり、逆に、ドメインの信頼性が最も重要視されている¹⁾。また、登場以来、検索エンジンの中で最大の Web ページ収集量を達成していたが、2009/05/04 時点では、1つのデータベースシステムと考えた場合、Yahoo の収集件数が google を凌いでいる。

1) : <http://www.seomoz.org/blog/how-googles-rankings-algorithm-has-changed-over-time>

検索はキーワードの間にスペースを入れると and 検索となる。検索オプションをクリックすると以下の検索条件が表示され、ブール演算子による検索が可能となる。

すべてのキーワードを含む	AND 検索
フレーズを含む	連続した語の検索, " " と同じ
いずれかのキーワードを含む	OR 検索
キーワードを含めない	NOT 検索

【英語】 www.google.com 表示不可

2009/05/04 時点で、<http://www.google.com> に接続すると、自動的に <http://www.google.co.jp> に切り替わり、www.google.com は表示されない。ページの下右にあった Google.com in English も表示されない。英語で検索する場合は、www.google.co.jp で英語のキーワードを使って日本語と同様に検索する。

www.google.com は表示されなかったり、情報が検索対象からはずされる (Censorship by Google) 事例など、情報が操作され t いる場合があるが、google に限らず検索エンジンは商用ベースで運用されているので、どの検索エンジンも情報が全く公平に扱われていることはないと考えておく必要がある。

3-3 ASK <http://www.ask.com/>

【日本語・英語】 一度検索を実行すると search の右に Advanced Search が表示される。Advanced Search では以下の検索が可能となる。日本語も英語も全く同じ手順で実行できる。

all of the words	すべての語を含む (AND)
the exact phrase	検索文字の並び順通り検索する (フレーズ検索)
at least one of the words	いずれかの語を含む (OR 検索)
none of the words	除外する語 (NOT 検索)

3-4 alltheweb <http://www.alltheweb.com/>

【日本語・英語】 search の右に表示される advanced search により AND, NOT を使ったブーリン演

算子による検索が可能。Search for -は単語の AND 検索, must include は単語およびフレーズの AND 検索、must not include は単語およびフレーズの NOT 検索に対応する。OR 検索には対応していない。英語の場合は find results written in を English に指定する。

3-5 AltaVista <http://www.altavista.com/>

【日本語・英語】

Advanced Search をクリックすると以下のように AND, OR, NOT の検索が可能なページが表示される。該当する窓に、キーワードを入力し、Find をクリックすると検索が実行される。英語も全く同じ県境で実行できる。

all of these words:	すべての語を含む (AND 検索)
this exact phrase:	文字の並び順通り検索する (フレーズ検索)
any of these words:	いずれかの語を含む (OR 検索)
and none of these words	除外する語 (NOT 検索)

3-6 goo <http://www.goo.ne.jp/>

【日本語】 NTT レゾナントが運営するポータルサイト。検索ボタンの右上にある検索オプションをクリックするとブール演算子による検索が可能なページが表示される。AND, OR, NOT, フレーズ検索, 文章検索に対応している。文章で検索する (自然文検索機能) は文章からキーワードを切り出して検索する機能である。カテゴリ検索機能も有する。

キーワード	文章で検索する (自然文検索機能)
	文字の並び順通り検索する (フレーズ検索)
追加キーワード	いずれかの語を含む (OR 検索)
	すべての語を含む (AND 検索)
	除外する語 (NOT 検索)

3-7 他²⁾

・Cuil (<http://www.cuil.com/>) : 2008 年 7 月に立ち上げられた世界最大の検索エンジン。google の 3 倍の情報量を有する。

・MSN Japan (<http://jp.msn.com/>) :

4 大検索エンジン中 MSN のみは演算子による検索に対応していない。

・Excite <http://www.excite.co.jp/> :

エキサイトの検索機能には、ウェブ検索、カテゴリ検索、イメージ検索、ニュース検索の 4 種類の検索が提供されていたが、2009 年現在、演算子による検索はできない。

・Infoseek Japan :

楽天サイトに吸収されてから、従来可能であった演算子を直接入力検索ができなくなっていた。

2009 年現在、指定中の条件から論理式による検索が可能となったが、件数表示不可。

・Dogpile (<http://www.dogpile.com/>) :

4 大検索エンジンと言われている Google、Yahoo!、Ask、MSN Search を同時に検索できるメタサーチエンジン。

他の検索エンジンについては The Search Engine List²⁾ を参照してください。

2) : The Search Engine List, <http://www.thesearchenginelist.com/>

4. カテゴリ方式の検索エンジン

カテゴリ方式は、情報の内容を大項目、中項目、小項目のように階層的に分類したリストを次々にクリックすることにより容易に目的の情報に達することができる。小項目は情報数の実態に合わせて作成され内容によっては複数の階層からなっており階層の数は固定されていない。代表的なサービスに Yahoo! カテゴリがある。キーワード方式と基本的異なるのは対象とするサイトをコンピュータではなく人がサイトを選択している点である。対象となっているサイトの数はキーワード方式と比較すると少ないが (100 万件程度)、政府や大学などの情報で信頼性の高いもの、一般的な知識を得ることのできる情報、人の関心を引く情報など、多くの人にとって有用な利用価値の高いサイトの情報が含まれる。以下に示すようなカテゴリ検索サイトがあるが Yahoo の内容が最も充実しているので授業では Yahoo! カテゴリを使用する。

カテゴリ方式の検索は、検索する情報のコンセプトが明らかであれば、検索に必要なキーワードを知らなくても、選択することにより目的の情報を検索できる。情報検索の初心者にとっても有用な手段になりうることから有力な Web 情報検索の手法といえる。単純な概念だけで検索できる場合には便利だが、反面、概念を組み合わせなくては表現できないような個別テーマの検索には不向きである。なお、Yahoo の本来の検索機能はこのリスト方式であったが、google をはじめとするキーワード方式検索エンジンの機能の向上により利用頻度は減ったが、有力な Web 情報検索の 1 手段である。図 3 n にカテゴリ方式の検索エンジンを示す。日本とアメリカの Yahoo カテゴリの大分類を図 4 に示す。両者とも同一内容の 1 4 の大分類からなっている (2007/4/28~2009/05/04 現在)。

4-1 (日本語)	Yahoo! カテゴリ http:// dir.yahoo.co.jp
4-2 (英語)	Yahoo! Directory http://dir.yahoo.com/
—	goo http://dir.goo.ne.jp/
—	biglobe http://search.biglobe.ne.jp /dir/index.html

図 3 カテゴリ方式の検索エンジン

	Yahoo! カテゴリ (日本)	Yahoo! Directory (アメリカ)
1	エンターテインメント	Arts & Humanities
2	趣味とスポーツ	Business & Economy
3	芸術と人文	Computers & Internet
4	生活と文化	Education
5	教育	Entertainment
6	健康と医学	Government
7	社会科学	Health
8	メディアとニュース	News & Media
9	ビジネスと経済	Recreation & Sports
10	各種資料と情報源	Reference
11	コンピュータとインターネット	Regional
12	政治	Science
13	自然科学と技術	Social Science
14	地域情報	Society & Culture

図 4 日本とアメリカの Yahoo カテゴリの大分類

4-1 Yahoo! カテゴリ <http://dir.yahoo.co.jp/>

【日本語】階層的に並べられたエンターテインメントから始まる 1 4 のカテゴリから順に、提示されるリストの中のカテゴリを選択 (クリック) する。最後のカテゴリには Yahoo の登録サイトが表示される。この登録サイトは Yahoo が人の目で選択した重要と思われるサイトである。登録サイトををクリックすると Yahoo から離れ、登録されたサイトのページが表示される。Yahoo が独自に選んだ Web Site と一致するサイト。担当者が使う価値があると判断したサイトを検索の対象としていることから登録サイトの内容は多くの人にとって利用価値があり Yahoo が使いやすい最大の理由となっている。

Yahoo の Categories を探すために、「Yahoo! カテゴリ全体」を指定してキーワード検索を実行するとカテゴリを捜すことができるが、Categories を捜すための検索は基本的にキーワード 1 語とする。

4-2 Yahoo! Directory <http://dir.yahoo.com/>

【英語】使用方法は日本の Yahoo! カテゴリ (リスト方式) と同じ。URL が異なること、および Yahoo! カテゴリと同じ 1 4 の大分類であるが、大分類以下の分類は分野により大きく異なっている。

5. 課題（レポート2）

5-1 Web 情報検索の学習到達目標

このレポートの学習到達目標は、情報検索の体験をとおして以下の内容を理解できることである。

- ①「論理式」の概念を理解し検索式を作成できること
- ②「検索エンジン」で検索式を使った検索ができること
- ③「英語のサイト」も有用な情報源となることを理解できること
- ④「複数の検索エンジン」を使用すると効果があること

5-2 提出課題

レポート2（その1）で登録した課題について、以下に示す①，②，③，④に該当する検索エンジンを使用して Web 情報の検索を行ってください。同一の課題で4種類の検索エンジンを使って、4回検索を行うこととなります。

さらに、情報検索の結果、最初に設定した目的や目標を達成するために役立つと思われる情報を、①～④の検索毎に5サイト（ページではありません）以上選択し、得られた内容を整理してレポートとして提出してください。得られた情報の整理は、検索毎に個別のサイトの情報を100から400字程度でまとめ、最後に全体を要約した内容を提出してください。

検索結果をまとめることができたなら、最後に、4回の情報検索に使用した検索エンジンのそれぞれの特徴を整理する。

- ①日本語のキーワード方式の情報検索（図2から選択した日本語検索エンジンを使用）
- ②英語のキーワード方式の情報検索（図2から選択した英語検索エンジンを使用）
- ③日本語のカテゴリ方式の情報検索（図3からYahoo!カテゴリを使用）
- ④英語のカテゴリ方式の情報検索（図3からYahoo! Directoryを使用）

5-3 レポートの作成

4回検索を行うので計20サイト以上の情報を収集し整理する。レポートには情報を得た各サイトのURLを必ず記入する。登録した検索内容を変更する場合は、課題の再登録が必要です。カテゴリ検索で適切なカテゴリが見つからない場合は、目的・目標に近いカテゴリ情報を収集してください。

キーワード検索におけるキーワードは、登録したキーワードより適切なキーワードが見つければ変更する。最初に設定した検索式では良い結果が得られないのが普通なので、検索式の欄には良い結果が得られた場合の検索式を記入する。

件数を5件に絞るためにキーワードの数を増やすのは意味がないので、適切なキーワードで数百件程度に絞り込みそこから自分の目で内容を確認して選択するように心がけて下さい。例えば、4キーワード以上を使っても、多すぎる件数が表示される場合は、上位200件の中から、目的や目標に合う5件を選択してください。

英語の情報検索の際は、使用するキーワードを複数の意味が記述された辞書等で確認すること。英語検索の結果には日本のサイト（.jpドメインを）の情報を除き、かつ日本語の検索で利用したサイトと重複しないようにしてください。英語情報のまとめは日本語で提出する。翻訳が目的ではないので内容がわかる文章が良い。英文のままの提出や、翻訳ソフトを使った意味不明の文章は評価できません（意味がわかれば使用可）。

5 サイト全体を要約した内容には、最初に設定した目的や目標に対して有用と思われる内容を、自分の文章でまとめて記述してください。

5-5 個別のサイトの情報の記述

サイト毎に得られた具体的な内容を記述し、その後で全ての情報を全体的に要約した内容を記述する。具体的な内容とは具体的な事実や、数値データなどのことで、具体的な内容を記述することを忘れないようにまとめてください。「～について記載されていた」という類の表現は得られた内容を何も表現していないと同じなので、まとめの文章では使用しないこと。

目的や目標によって、専門的な情報を収集するのがあるいは一般知識としての情報を収集するのかなど、求める情報の量や質が変化するので、内容をまとめる際には登録した目的や目標を再確認しながら記述してください。

例：

× 韓国の文化について記載されていた。

○ 韓国の民族衣装はチマ・チョゴリで、韓国の代表的な食品はキムチである。

× パレスチナの問題が載っていた。

○ パレスチナの問題は、アメリカの権限が強いため、アメリカが協力し、世界がそれを取り巻くように協力していくことが大切である。

× このサイトでは道教とその思想的存在である道家について語られている。

○ 道家とは道教を学ぶもののことであり、その思想は自然的なものに関する物が多い。

× イチローのオリックス時代から現在のメジャーに至るまでの成績と写真が載っている。メジャーリーグのオープン戦の成績（打席の内容まで）詳しく載っていて、とてもよかった。

○ イチローのプロフィールは、ポジション右翼手、生年月日 1973 年 10 月 22 日、国籍日本、身長/体重 175cm/73kg、右投、左打で、2001 年の成績は打率.350、8 本塁打、69 打点、出塁率.381、長打率.457 であった。

5-6 使用する検索エンジン

図 2 のリストから、日本語、英語検索に使用するキーワード方式の検索エンジンを、1 つずつ選択する。カテゴリ方式は、図 3 の Yahoo! カテゴリと Yahoo! Directory を使用する。

6. レポート②（その 2）の提出

提出画面に必要事項を記入して提出してください（3 2 のみレポート BOX に提出のこと）。

3 3 項目は良く考えて記入してください。

インターネット情報検索 レポート提出画面

以下の手順でこの画面の10/10までお進みください。

協賛内容は以下のURLから確認してください。
http://rc.nsl.ac.jp/baiyo/2/2016/02_27/000000.html
 各形式別リストはレポート提出へ提出してください。

1. 名称(漢字)・キーワード
 (例: 国際会議場)
2. 年報番号(例) (1999年のように、数字で検索可能です)
 (例: 1999年)
3. 選択分野 (選択に成功した情報を検索できる分野を選択する。英語版)
4. 情報検索言語 (内訳がわかる簡単な2語-3語で検索できる言葉・英語版)
 日本語のみではなく、英語の情報も存在する認識してください。

5. から後、日本語と英語のキーワード検索の結果を記入してください。

6. キーワード方式で検索した日本語検索エンジンと英語検索エンジンの名称

日本語キーワード方式で検索したキーワードと、絞り込み検索
 (例: 検索: 4,878,807件)

7. 件数

8. 件数

9. 件数

10. 件数

絞り込み検索の件数

11. 件数

12. 件数 and 2_件数

13. 件数 and 2_件数 and 2_件数

14. 件数 and 2_件数 and 2_件数 and 2_件数

15. 件数 and 2_件数 and 2_件数 and 2_件数 and 2_件数

16. 検索に実際に使用した検索式と件数 (OR記号を使用すると上記2/5の件数と異なる)

17. 日本語キーワード方式による情報検索の結果
 情報検索で得られた5サイト以上の情報を、各サイト(100-400字で記述する各サイト)で500-1000文字、情報を適切に要約できるように記入する。画像入力せず、できるだけウェブ上で作成した文章を手の持りにコピーしてください。

英語キーワード方式で検索したキーワードと絞り込み検索の件数

