日本語検索とその手法

石　井　保　廣

抄　　録

インターネット上の情報を入手する際、日本語による検索は西欧諸国語とは異なりキーワードとなる「ことば」が分かち書きされていない。本稿では、さまざまなキーワード抽出の現状と重み付けを分析し、そのキーワードの有効性について、理論的な尺度であった再現率や適合率を実際の検索から簡易に算出し、比較・評価が可能な方法を提案する。このほか IDF も併用し、具体的なサイトを使って検索結果を数値化して､日本語検索の際の特筆すべき留意点について述べる。

キーワード

有効指数、IDF、Morpheme、形態素解析、N-gram　bigram、索引化、indexing、キーワード抽出、検索語抽出、分かち書き、ChaSen、KAKASHI、MeCab、Google、Yahoo

１．はじめに

１．１　検索戦略

検索エンジンでネットワーク上の情報を入手する場合、それぞれのサイトで日本語の文字列からキーワードを抽出する方法が異なるため情報検索に関する理解や操作を困難なものとしている。キーワードの用い方如何により、検索エンジンやデータベースの検索で再現率や適合率の低下を招き、ノイズの多い情報しか入手できなかったり必要な情報を見逃す結果になることも多い[1)]。検索キーワードがサイトの特性によって、どのように抽出・索引化（indexing）されているかを検証することが検索戦略のカギとなる。

１．２　検索結果の評価とフィードバック

情報の検索では、検索結果を評価し満足し得る情報を入手できるまで適切なキーワードを与える必要がある。通常、これらのキーワードは論理式やトランケーションなどによる検索式として投入される。本論ではキーワードの評価自体が目的であるため、論点を絞りキーワードの順列及び組み合わせによる検索のみ検証し、検索式を含めた評価までは言及していない。

２．キーワード抽出の概要

検索の手掛かりとなる「キーワード」は、さまざまなポータルサイトで、インターネット上の情報源から形態素解析や N-gram などのアルゴリズムにより切り出されている。それぞれの設計

手法が異なるため、キーワードの選び方次第で適合率や再現率に差が生じ、日本語検索を一層複雑なものとしている。

２．１　形態素解析によるキーワード抽出

形態素解析は、文脈が意味を持つ最少単位（形態素）まで切り分け、それぞれの単位に品詞、活用形、ヨミなどの情報を付加する技術である。分かち書きの習慣がない日本語では、文脈から単語を正確に切り出すために形態素解析エンジンを用いるが、未知語や旧字体などを含め膨大なtable　lookup 用の辞書を使用している。さらに、「コロンビア」・「コロムビア」・「コロンビヤ」など、ことばのユレなどへの対応も必要である。

この結果、検索キーワードとして適当ではない助詞、助動詞などの付属語をストップ語として取り除く。この処理により文字列からの文脈は、検索キーワードとしての形態素が索引化される。

形態素解析のためのエンジンは、有償・無償のソフトが多数存在しており、実験的に文脈モデルを検証することは容易である。例えば、「デジタルアーカイブのための情報収集・発信モデルの展開」という文脈を例に形態素解析の比較サイト[2)]にかけてみると以下のような結果を求めることができる。形態素解析エンジンは、以下の３サイトによる解析をした。

ChaSen（茶筌）

奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室）で開発されたもので、形態素解析器として無償で配布されている。

デジタル／アーカイブ／の／ため／の／情報／収集／・／発信／モデル／の／展開

KAKASHI

kanji kana simple inverter の略で、文字列の先頭から解析し、次の単語の候補が複数のときは最も長い単語を求める最長一致法を採用している。

デジタルアーカイブ／のための／情報収集／・／発信／モデル／の／展開

MeCab

京都大学情報学研究科―日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト開発されたもので、パラメータの推定に Conditional Random Fields （CRF）が採用されている。

デジタル／アーカイブ／の／ため／の／情報／収集／・／発信／モデル／の／展開

となり、それぞれのエンジンが持つアルゴリズムにより形態素に分割されている。

さらに、詳しく調べるために、ChaSen の形態素解析エンジンをインストールし、同じ文脈を入力し、表１の結果が得られた。このことから、形態素分解（分かち書き）と品詞の分析がなされていることがわかる。

同様に、MeCab でも、まったく同一の結果が得られた。

このことから、助詞と判断された「・・・モデルの展開」の「の」は、当然ストップ語として除くことができるが、「・・・ための・・・」は、「ため」と「の」に分解され、「ため」は非自立語としているものの名詞と判断されている。また、「の」も助詞―連体化として、検索キーワードに採用される可能性が高い。

結果的に、この形態素解析システムでは、文脈からストップ語を除いたデジタル／アーカイブ／ため／情報／収集／発信／モデル／展開が検索キーワードとして、導き出されると考えてよいであろう。本論では、これらの形態素を比較のための評価キーワードとして用いる。

表１　形態素解析一覧

表層形	ヨミ	原形	品詞
デジタル	デジタル	デジタル	名詞－一般
アーカイブ	アーカイブ	アーカイブ	名詞－一般
の	ノ	の	助詞－連体化
ため	タメ	ため	名詞－非自立－副詞可能
の	ノ	の	助詞－連体化
情報	ジョウホウ	情報	名詞－一般
収集	シュウシュウ	収集	名詞－サ変接続
・	・	・	記号－一般
発信	ハッシン	発信	名詞－サ変接続
モデル	モデル	モデル	名詞－一般
の	ノ	の	助詞－連体化
展開	テンカイ	展開	名詞－サ変接続

２．２　N-gram によるキーワード抽出

N-gram は機械的に文字列を n 文字単位で切り出していく。このとき、文字片は一文字ずつ重なるようにして抽出していく。もはやこの方式では、品詞の解析は必要ではなく、単純に文字列を処理していくだけである。このため indexing は高速であるが、索引ファイルは検索ノイズを伴う冗長な索引語が膨大となる。また、検索する側からみると、bigram[3)]であれば検索キーワードは自動的に２文字単位に変換され索引ファイルを参照することになる。このため完全一致検索の漏れはなくなるが、品詞を解析しているわけではないので語尾の変化に対応できないほか、一例として、「沖縄文化」という文字列から「縄文」を切り出してしまい不要な情報（ノイズ）が多くなってしまう欠点が生じるなどの点から、規模の大きなサイトではほとんど採用されていないようである。

bigram による展開

デジ／ジタ／タル／ルア／アー／ーカ／カイ／イブ／ブの／のた／ため／の情／情報／報収／収集／集・／・発／発信／信モ／モデ／デル／ルの／の展／展開

３．検索サイトでの日本語検索

Web の検索では、検索サイトでどのように検索キーワードを抽出しているかを考慮した検索のほか、検索式をどのように組み立てているのかを把握することが大切である。ここでは、検索エンジンを比較し具体的な日本語検索の手法を考える。

一般に、検索に対する評価は再現率と適合率で示されるが、要素の一つである全文書中の適合文書数など予め用意されたテストコレクション[4)]を用意しなければならないことや、適合率の算出には、検索されなかった情報の件数が必要など理論上の要素が強く、実際の検索にそのまま応用できるものではない。

ここでは実際の検索サイトで文脈モデルをとおして、キーワードとしての有効性を模擬的に算出し、それぞれのキーワードによる検索の簡易な比較・評価方法を有効指数として提案する。

このほか、大域的重み付けも算出し、キーワード出現頻度の偏りについてその傾向も探る。

理論的な再現率・適合率 ← 再現率＝ヒット中の適合件数／サイト内に存在する全適合件数
適合率（精度）＝ヒット中の適合件数／ヒット件数

⇩

簡易な比較・評価 ← 有効指数＝abs(log(文字列のヒット数/最大ヒット件数の形態素))

・再現率・適合率

再現率におけるサイト内に存在する全適合件数を、文脈モデルによるヒット数に置き換えたものを分母とし、形態素のヒット件数を除したものである。また、適正率は主題に関係する自立語である形態素を分母とし、形態素のヒット件数を除したものである。これらの算出には、論理積による絞り込み検索が可能であることが前提になるのは言うまでもない。

・有効指数

簡易に算出することを目的に、それぞれのキーワード（文字列）を、最大ヒット件数（検索集合の上限値）をもつ形態素で除し、その絶対値を対数化したものである。この係数による評価は、求めたい情報が残っている限り、数値が大きくなればなるほど精度が上がっていると判断することができる。また、数値が実数ではなく割合となるので、各サイト間をそのまま、比較できる利点がある。

・大域的重み付け

レコード全体にわたる検索キーワードの出現頻度の偏りを表す。計算式としては、形態素を含むレコード数を検索集合の総数で除し対数化したもので、多くのレコードに出現するキーワードには、低い重みを与える。このため、IDF[5]値が高いものは目的の文献を特定する際に効果が高い（適合率が高い）と判断することができる一方、検索漏れ（再現率が下がる）が多くなることを意味する。ここでは検索集合の総数（ヒット総数）の値を、最大のヒット件数をもつ形態素として擬似的な算出法をとった。

IDF＝log（ヒット総数／形態素を含むレコード数）

３．１　形態素の分析

全文検索エンジンとしてメジャーな Google と Yahoo を採りあげ、日本語検索を検証する。

・Google

Google は、クローリングしてきたすべての web 情報は、インデックス生成過程において複雑な処理を施し検索語までを数値化し文字の配列位置も記録しているといわれている[6]。

Google は Basis　Technology 社の形態素解析システムを使用しているといわれており、実際には、検索効率の点から数値化された索引方式となっているようである。このシステムの検索文字列の形態素解析は、キャッシュやソース表示で検索語の分かち書き状況を知ることができる。

・Yahoo

Yahoo も Google 同様、web の情報をクローラーで収集したキーワードをインデクサーで抽出・索引化し検索サーバで提供しており、これを Yahoo!　Search　Technology（YST）と称している。

web 情報は、一部深層 web[7]を検索するシステムも試験運用を開始した。

Yahoo では、自社で開発の日本語形態素解析 Web　API を用いている。文字列の形態素解析

は、web 検索結果や html で検索語の分かち書き状況を知ることができる。ここでも同様の文脈モデルを検索文字列として投入した。

ここでは、確認のために文脈モデルを形態素ごとに検索し、ヒット件数とそれぞれのバラツキを見るためにメジアン[8)]を中心とした出現数の比較及び IDF を算出する。

なお、このときに付属語である「の」や「のため」は、直接主題を表すものではないので除いた。

表２　Google における形態素の重み

形態素	Google		
	ヒット数	メジアンとの比	IDF
デジタル	119,000,000	1.534	0.962
アーカイブ	70,600,000	0.910	1.189
情報	1,090,000,000	14.046	0.000
収集	34,900,000	0.450	1.495
発信	30,300,000	0.390	1.556
モデル	106,000,000	1.366	1.012
展開	77,600,000	メジアン	1.148

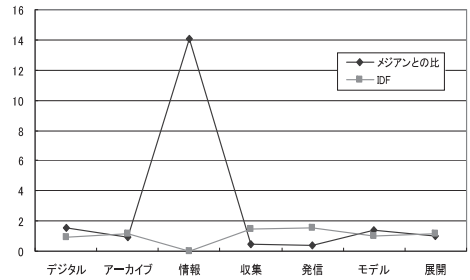


表３　Yahoo における形態素の重み

形態素	Yahoo		
	ヒット数	メジアンとの比	IDF
デジタル	375,000,000	0.949	1.020
アーカイブ	395,000,000	メジアン	0.998
情報	3,930,000,000	9.949	0.000
収集	182,900,000	0.463	1.332
発信	181,000,000	0.458	1.337
モデル	424,000,000	1.073	0.967
展開	433,000,000	1.096	0.958

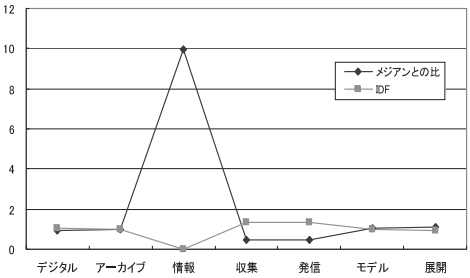


表２、表３の、Google、Yahoo における形態素ごとの件数から、偏り具合がわかる。

形態素ごとの相対的な出現傾向は、Google、Yahoo ともに同じ傾向であり、特に、広く一般に用いられている「情報」の出現件数が他の形態素に比べ約10倍と圧倒的に高いことである。

ここで注目したいのは、IDF 値である。簡易な方法として分子となるヒット総数を形態素の最大値をもつ「情報」をヒット総数としたため（このため「情報」の IDF が０となってしまった。）正確な数値はわからないが形態素間の比較はできる。IDF 値の高い「収集」や「発信」などは、情報を絞り込んでいく際に有効である。ただし、これらの形態素（自立語のみ）の中で、比較的 IDF 値の低い（キーワードとしての重み付けが高い）「展開」は、直接主題に関係するものではなく、一般語としての要素が強い。このため IDF 値だけではキーワードの重みを単純に判断はできない[9)]。

３．２　文字列の順列による検索

今回例とした「デジタルアーカイブのための情報収集・発信モデルの展開」をそのまま検索文字列として検索すると、Google、Yahoo では、次のような分割となっていた。

Google：デジタル／アーカイブ／の／ため／の／情報／収集／発信／モデル／の／展開

Yahoo ：デジタル／アーカイブ／の／ため／情報／収集／発信／モデル／展開

両システムにおける解析結果は、付属語としての助詞「の」が重出しているかどうかであり、ほぼ ChaSen や MeCab と同様であった。

「の」はストップ語とはされず、記号「・」のみがストップ語として取り除かれていた。

次に、フルタイトル、付属語を除いた分かち書き、フレーズ検索及び参考のため、「ための」や「の」検索を試みると次の結果を得た。

表４　形態素組合せによるヒット数

項番	投入文字列	システムによる分かち書き	Google	Yahoo	備考
1	デジタルアーカイブのための情報収集・発信モデルの展開	デジタル／アーカイブ／の／ため／情報／収集／発信／モデル／展開	6,920	175,000	原文のまま
2	デジタル　アーカイブ　情報　収集　発信　モデル　展開	デジタル／アーカイブ／情報／収集／発信／モデル／展開	7,280	209,000	自立語のみ
3	“デジタルアーカイブのための情報収集･発信モデルの展開”	“デジタルアーカイブのための情報収集・発信モデルの展開”	2	2	フレーズ検索
4	ための	ため　の	83,200,000	662百万	N／A
5	の	の	24.7億	81.7億	助詞

項番１は、ストップ語として記号が除かれているだけである。また、項番２では、助詞などを除き自立語のみとして検索した結果、項番１に比べ１～２割ほどヒット件数が増えている。このことから、付属語もそのままつけて検索することにより若干ノイズが減っていると推察できる。さらに、項番３では、フレーズ検索として原文の文字列のまま検索した結果、当然のことながら同一の文字列を持つ情報のみがヒットした。

この表から、１件のみヒットした項番３を最終的に求める情報と仮定して検索キーワードの組み合わせにより項番３を含む集合がどのように出現するか有効指数を用いて検証する。

①主題に関係の深い形態素のみの組み合わせ

主題に関連のあることばとして、デジタル／アーカイブ／情報／収集／発信／モデルがある。これらを組み合わせて検索した結果は以下のとおりであった。

表５　主要語のみの検索結果

項番	投入文字列	Google		Yahoo	
		ヒット数	有効指数	ヒット数	有効指数
6	デジタル／アーカイブ	653,000	3.22	29,000,000	2.13
7	デジタル／アーカイブ／情報	991,000	3.04	23,200,000	2.23
8	デジタル／アーカイブ／情報／収集	302,000	3.56	4,330,000	2.96
9	デジタル／アーカイブ／情報／収集／発信	37,000	4.47	546,000	3.86
10	デジタル／アーカイブ／情報／収集／発信／モデル	18,000	4.78	291,000	4.13

Google では、AND 検索したにもかかわらず項番６に対して項番７でヒット件数が大幅に増加

している。ホームページのヘルプでは、「Google の標準設定では、入力したキーワードをすべて含むページだけが表示されます。キーワードの間に「AND」演算子を入れる必要はありません。」としており、検索のアルゴリズムは不明である。しかしながら全体的には、検索キーワードを空白で挟んで増やすごとに AND 検索となり、無関係な情報（ノイズ）が取り除かれていると考えられる。

②すべての自立語

例文中のすべての自立語の順列を変えずに検索した結果は次のとおりである。

表６　すべての自立語の検索結果

項番	投入文字列	Google		Yahoo	
		ヒット数	有効指数	ヒット数	有効指数
11	デジタル／アーカイブ／情報／収集／発信／モデル／展開	8,290	5.12	202,000	4.29

「展開」は、主題とは直接関係していない。このようなキーワードを投入することによって、単に再現率を下げる結果になる恐れがある。このため、より多くの検索集合を期待する場合は、これらのキーワードを除き、最初から希望する情報の文脈がわかっている場合に追加するようにしたい。

③付属語を追加した組み合わせ

付属語は、直接主題には関係しないが、目的の情報が確定している場合には精度を高めるために有効であろうか。

表７　付属語を追加した検索結果

項番	投入文字列	Google		Yahoo	
		ヒット数	有効指数	ヒット数	有効指数
12	デジタル／アーカイブ／情報／収集／発信／モデル／展開／の	8,290	5.12	200,000	4.29
13	デジタル／アーカイブ／情報／収集／発信／モデル／展開／の／ため	8,000	5.13	173,000	4.36
13ａ	デジタル／アーカイブ／情報／収集／発信／モデル／展開／の／ための	7,880	5.14	153,000	4.41

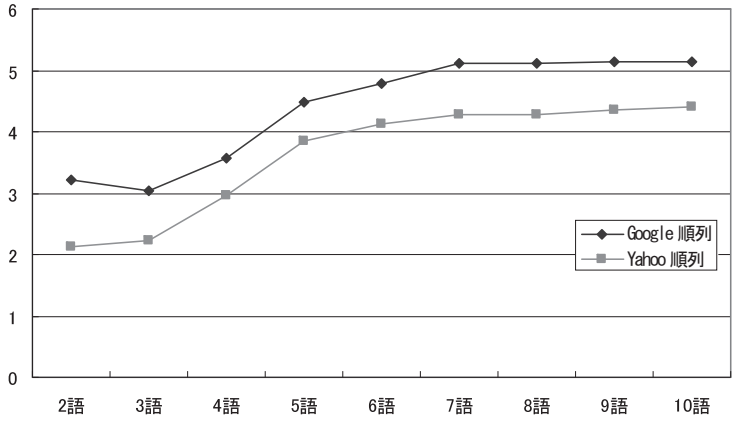
以上が、単純に文字列を順列どおりに追加した結果であるが、主要語のみでは、形態素の数に比例して有効指数が高くなっている。有効指数が高くなるということは、キーワードを増やすことによって精度が上がってきたことになる。主題に直接関係しない「展開」を追加した場合も同様である。一方、付属語を追加しても有効指数はそのままか微増であり、あまり効果がなかった。

このことから、ただちに付属語が不要とはいえずフレーズ検索や文脈どおりに投入すると精度が上がる場合も多い。

また、有効指数からみた Google と Yahoo を比較すると、ヒット件数の絶対数としては yahoo が大幅に多く、有効指数は Google が全体的に高い。このことから、精度は Google が高く、yahoo は再現率が高いがノイズも多いと推察される。

次に、キーワードの追加でどれだけの効果があるのか有効指数をグラフ化してみた。

図１　キーワードの順列追加



目盛りが対数表示なのでゆるやかな右肩上がりとなっているが、実際の上昇率はもっと大きいことを考慮する必要があるものの、キーワードの数が少ないうちの有効指数は傾きが急である。これに比べ、項番12以降の付属語の追加あたりからは穏やかな不変域となっている。

ある調査では、検索キーワードを２～３語で検索するケースが37%という。この結果は、キーワードを増やすことによって精度を上げることができることを裏付けている。しかし、これまでの検索は文脈モデルを順列に沿って逐次キーワードを投入したものであり、キーワードを有効度の高いものから組み合わせることにより、さらに効果が期待できるのではないか推察される。

３．３　IDF を配慮した検索

これまでは、文脈モデルの出現順に順列として検証してきた。付属語や主題に関係のない自立語を除き、表２や表３で得た IDF による重み付けによる組み合わせではどうであろうか。確率的には、７つの形態素から多くの組み合わせができる。IDF 値の一番高いもの、つまり文書頻度の低いものから順に追加する方法を検証する。IDF 値を順に並べかえると次のとおりとなる。

・Google：発信＞収集＞アーカイブ＞展開＞モデル＞デジタル＞情報
・Yahoo ：発信＞収集＞デジタル＞アーカイブ＞モデル＞展開＞情報

表８　重み付けによる検索結果（Google）

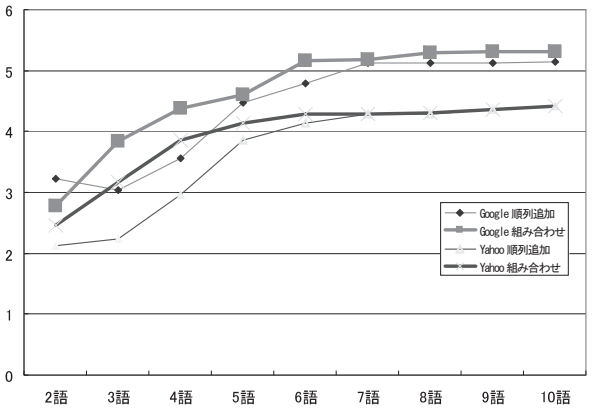
項番	投入文字列	Google	
		ヒット数	有効指数
14	発信／収集	1,840,000	2.77
15	発信／収集／アーカイブ	160,000	3.83
16	発信／収集／アーカイブ／展開	45,300	4.38
17	発信／収集／アーカイブ／展開／モデル	26,800	4.61
18	発信／収集／アーカイブ／展開／モデル／デジタル	7,420	5.17
19	発信／収集／アーカイブ／展開／モデル／デジタル／情報	7,350	5.17
20	発信／収集／アーカイブ／展開／モデル／デジタル／情報／の	5,510	5.30
21	発信／収集／アーカイブ／展開／モデル／デジタル／情報／の／ため	5,360	5.31
22	発信／収集／アーカイブ／展開／モデル／デジタル／情報／の／ための	5,320	5.31

表９　重み付けによる検索結果（yahoo）

項番	投入文字列	yahoo	
		ヒット数	有効指数
23	発信／収集	13,800,000	2.45
24	発信／収集／デジタル	2,720,000	3.16
25	発信／収集／デジタル／アーカイブ	545,000	3.86
26	発信／収集／デジタル／アーカイブ／モデル	289,000	4.13
27	発信／収集／デジタル／アーカイブ／モデル／展開	200,000	4.29
28	発信／収集／デジタル／アーカイブ／モデル／展開／情報	201,000	4.29
29	発信／収集／デジタル／アーカイブ／モデル／展開／情報／の	198,000	4.30
30	発信／収集／デジタル／アーカイブ／モデル／展開／情報／の／ため	172,000	4.36
31	発信／収集／デジタル／アーカイブ／モデル／展開／情報／の／ための	152,000	4.41

順列どおりに見ると、両システムとも同様の傾向であったが、IDF による重み付けの結果では次のグラフのようになった。なお、比較のため前述の「図１　キーワードの順列追加」も、重ね合わせてみた。

図２　キーワード IDF 順追加



形態素をキーワードとして、文脈モデルの出現順（順列）で検索したもの（細線）に比べ、Google、Yahoo ともに不変域への到達が早くなっていることがわかる。これは、限定的な形態素を先に用いることで、検索の絞込みに効果があったことを意味する。しかしながら、絞り込みが進んだことによりノイズが減るのと引き換えに再現率の低下を招くことにもなる。IDF 値を配慮して検索する場合は、再現率の低下を招かないようある程度、予測のつくタイトルを検索する場合に効果があろう。一方で、多数の情報を入手したい場合は、付属語を除いた IDF 値の低い（文書頻度の高い）主題に関係の深いものから検索することにより、適合率（精度）を落とすことなく情報を入手できる。

４．まとめ

４．１　全文検索エンジンとデータベース

今回は、全文検索エンジンのみを検証の対象とした。文献情報の検索であれば、国立国会図書館の雑誌記事索引や国立情報学研究所の CiNii などのデータベースがある。データベースの場合は、検索システムは検索文字列を自動的に分かち書きするわけではないので、分かち書きによって再現率や適合率が左右される。さらに、助詞などをストップ語とされ、キーワードとして追加することにより検索集合は０件となるなど、別途論じる必要がある。

全文検索エンジンの大規模サイトでは、一回の検索で数十万～数百万件のヒットは意外に多いが、結果を表示できるのはせいぜい数百件までで、すべてを見ることは想定されていない。しかも、検索エンジンサイトは、時々刻々と情報が更新される。本稿では、2009年１月10日前後の数日間にわたって検索したため、前半と後半でヒット数に若干の違いがあることを断っておく。

４．２　今回の検証による課題

今回の検索で、いくつかの点が見えてきた。それらを挙げて「日本語検索とその手法」のまとめにしたい。

①分かち書き

全文検索エンジンの特性として、検索文字列は自動的に分かち書きされ、それが結果表示画面で見ることができる。キーワードへの分割を意識する必要はないが、期待しない形に分割される場合がある。必要であれば、フレーズ検索を併用する。

②検索戦略

文脈モデルの順列検索や IDF 順検索をとおして、未知の情報を対象とした検索では、主題に関係ある自立語をキーワードとして検索する。また、既知の情報の検索では、付属語を含むすべてのキーワードで検索することが有効である。

③精度とノイズ

自立語は、主題への関連が強いほど、適合率が高まり、逆に一般語ほど再現率が高くなる。どちらを優先するかは、その主題の情報に対するニーズと検索結果の集合件数に左右される。

・主題に関連の深い言葉は精度を上げるが、情報を見逃す可能性も高くなる。
・一般語や付属語は絞込みに有効で、既知の情報に対する検索ではノイズを減らす作用をする。

註

1）ノイズの状況は確認できるが、必要なデータを見逃したかどうかを知る術がなく確認することは事実上不可能である。

2）形態素解析の比較サイト：http://nomadscafe.jp/test/keitaiso/index.cgi

3）bigram: N-gram は、N 文字単位で１文字ずつシフトしながら抽出する方法である。このとき、N の文字数によって、１文字＝unigram、２文字＝bigram、３文字＝trigram・・・と称される。

4）web 上で複数の英文関係のテストコレクションが公開されている。

5）IDF（Inverse Document Frequency）は検索キーワードを含む文書頻度（document frequency）の逆数として算出される。

6）実際には、検索キーワードの出現順序を変えてもほとんど変化がなかった。

7）データベースは、照会に応じて動的に情報を取り出している。このため、通常の検索エンジンでは検索することができない。これらの情報を深層 web（deep　web）と呼び静的な表層 web に対し、500倍前後の情報量があるとされる。

8）データを昇順に並べたとき中央に位置するデータ。中央値と呼ばれることもある。

9）このために、ポアソン分布を用いた残差 IDF（RIDF）があるが、レコードごとに形態素の出現回数を調査しなければならず、今回は言及しなかった。

参考文献

1）北研二［ほか］．情報検索アルゴリズム，共立出版，2003，212p.

2）京都大学情報学研究科―日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト．MeCab：Yet Another Part-of-Speech and Morphological Analyzer, 2008, http://mecab.sourceforge.net/

3）奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座．Chasen――形態素解析器，2007, http://chasen-legacy.sourceforge.jp/

4）西田圭介．Google を支える技術－巨大システムの内側の世界－，技術評論社，2008，271p.

5）佐藤雅彦．KAKASI－漢字→かな（ローマ字）変換プログラム，2004，http://kakasi.namazu. org/index.html.ja

6）新山祐介．形態素解析・構文解析入門，2008，http://www.unixuser.org/~euske/doc/nlpintro/

7）Stephen Robertson. Understanding Inverse Document Frequency : On theoretical arguments for IDF. Reprinted from：Journal of Documentation, 2004, vol. 60　no. 5，p.503－520，http://www.soi.city.ac.uk/~ser/idfpapers/Robertson_idf_JDoc.pdf

8）YAHOO ! JAPAN．日本語形態素解析，2008，http://developer.yahoo.co.jp/webapi/jlp/ma/v１／parse.html

9）山田敏弘．国語教師が知っておきたい日本語文法，くろしお出版，192p.

10）保田明夫．形態素解析と分かち書き処理．　38p，（PDF），http://wordminer.comquest.co.jp/wmtips/pdf/H15_01－4. pdf