比較的少量の日本語文書に対する柔軟で高品質な検索システム

Flexible and High Quality Text Search System for Reasonable Size Japanese Text Files

高 尾 哲 康
TAKAO Tetsuyasu

1. はじめに

ここ数年のインターネットの爆発的な普及、とりわけ Web による情報提供サービスが進展するとともに、全世界で 80 億ページとも言われる Web ページから何らかの手段で情報を検索することが当たり前の時代になってきている。Web は電子メールとともにインターネットの利便性を大きく向上させた。これまでは、図書館や博物館、美術館などに行かなければ手に入らなかった情報がインターネットから効率よく得ることが可能となってきた。これにより、専門家と一般人とで情報収集手段の格差がなくなるとともに、IT 技術を使えるかどうかで新たな情報格差が生まれてくる時代となった。

現在、インターネット上の情報検索システムの代名詞ともなっている「Google」を提供している Google 社は、「世界中の情報を組織化し、全世界のユーザがそれらの情報を利用できるようにする」というミッションを掲げている。しかし、現在の「Google」は全世界の Web 情報の約 4 割の 33 億ページをカバーしているにすぎない(1)。これは、Google に匹敵する検索エンジン FAST を利用している alltheweb、AltaVista などの他の検索エンジンでも同様であり、もはやひとつの検索システムで Web 全体をカバーすることが困難になってきている。さらに新しいビジネスモデルとして広告型検索エンジンが登場してきている。Web 検索エンジンによるサービスを提供している会社がインターネット商取引に強い会社と手を結び、ユーザによる検索結果に広告を埋め込む (Google AdWords, Overture Sponsored Search など) ことで効果的な広告効果を出してきている。今後は、かつての Explorer と Netscape 間のブラウザ戦争と同様、検索エンジンどうしのシェア争いが激しくなってくると思われる。そのような時代に、Web ページによって情報発信する一ユーザ側としては、質の高い情報や新規性の高い情報を継続的に提供することが検索エンジンによるページランキングシステムで高得点を得て検索結果で上位にランクされ、広く認知されていく結果につながることになる。ページのランク値については、例えば Google Toolbar にて調べることができる。検索エンジンにおけるページランク値を上げる方法として、さまざまな裏技が考え出され、そのためのスパムサイトも存在する。しかし、現在ではそのようなごまかしがききにくくなってきていることから、一時はユーザの望む情報とは関係のないページが検索結果の上位のページになる場合など、混沌へと向かうかと思われた Web 世界も一定の秩序を取り戻しつつあるようである(1)。

一方で、各サイト側では、全文検索システム[(2)]などを導入することにより、自サイトで提供しているWebページ情報の検索を行なえるようにしているところもある。しかし、全文検索システムの維持・保守などにコストがかかるため、公開ページに関してはGoogleなどの汎用の検索エンジンを利用した「サイト検索」で検索サービスを提供し、イントラネット側でのみ全文検索システムを独自に運用することで切り分けを行なうのが一般的になってきている。

ここでは、サイト内のユーザパーティションやインターネットプロバイダのユーザホームページスペースに置ける程度の比較的少量の文書（テキストファイル）に対する柔軟で高機能、高品質な検索が行なえるシステムを実現し、検索サービスとして公開し、その結果としてどのような使われ方をしてきたかを報告する。

2.　これまでの情報検索

情報検索にはさまざまな種類がある。GoogleやalltheWebのようなWeb検索、ｅコマースにおける商品検索、画像処理を利用した顔や場所の認識などがある。今回実現したシステムはテキスト情報の検索である。テキスト情報の検索には、文字列検索、情報検索、検索インタフェースの各技術がある。文字列検索には主に2つの方法がある。一つはテキストを一次元のデータとらえ、その先頭から末尾に向かって順に検索する方法である。もう一つは検索対象となるテキストに対して検索してほしい文字列（キーワード）のインデックス（文字列とその文字列の出現位置）を前もって作成しておき、検索時にこのインデックスを利用して高速に検索する方法である。前者がシーケンシャル型検索、後者がインデックス型検索である。シーケンシャル型検索では、grepという正規表現を利用した検索プログラムを利用したものがほとんどであり、検索対象はベタテキストである。検索アルゴリズムは、KMP(Knoth-Morris-Pratt)法やBM(Boyer-Moore)法が多いが、計算量はO(n/m)～O(n)程度(m:キーワードサイズ、n:テキストサイズ)なのでテキストサイズにほぼ比例した時間がかかる。一方、インデックス型検索はWebページなど大量のテキストファイルの検索に向いており、前もってテキストファイルを解析し、キーワードとなる文字列（一般には単語）ごとにその出現位置を記録したインデックスファイルを作成しておく必要があるが、検索速度はシーケンシャル型に比べて圧倒的に高速である。例えば、Trie構造をもとにしたインデックスではO(m)程度(m:キーワードサイズ)の計算量である（図1にPentiumIII 1GHz相当のCPUによる速度比較を示す)。しかし、インデックスファイルに記録する内容で検索機能の制限を受ける。例えば、キーワードの前後の文字列との位置・順序関係や近接単語、行、節、パラグラフとの関係といった情報は記録しないのでこれらの関係情報を検索条件に指定した検索は容易にできなく

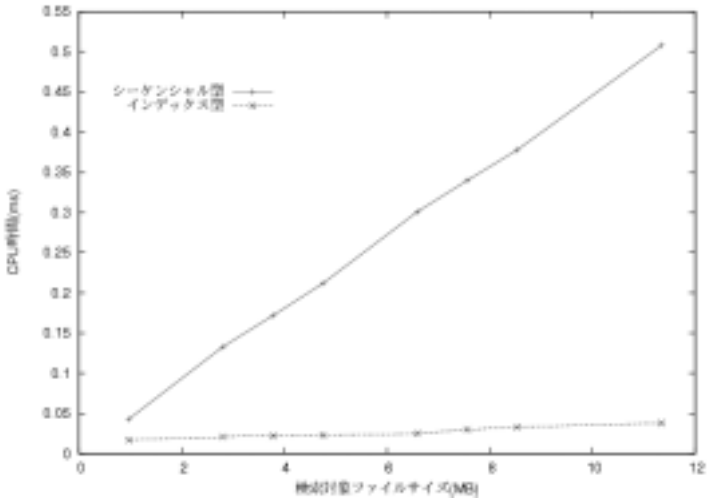

図1．シーケンシャル型とインデックス型検索の速度比較

なっている。そのため、インデックスに記録されていない情報を利用した検索を行ないたい場合は､改めて元のベタテキストを参照して検索し直す必要がある｡これらの情報をインデックスファイルに記録しようとすると、インデックス作成にかかる処理量とファイルサイズの増加を招くことになる。そのため、わざわざコストをかけて作成したインデックスがどの程度利用されるかについても考慮しながら、適度の必要を満たすレベルにとどめておくというトレードオフとなる。また、インデックスの単位となる単語をどのように決定するかも重要である。単語間にスペースがある英語テキストとは違い、日本語テキストでは単語間の区切りがない。文字種の違いを利用して区切りを識別する簡便な方法もあるが精度は悪い。単語間の区切りを決める方法には形態素解析システムを利用する方法と n-gram を利用する方法がある｡形態素解析システムを利用すれば、日本語単語辞書と単語間連接規則を利用して単語（形態素）間の接続可能性を調べることでテキストを単語ごと、つまり意味的な単位に区切ることができる。ただし、辞書に登録されていない単語がテキストに含まれていたり、テキストが詩歌や祈祷文、会話文、古語、方言など標準的な日本語文法に沿っていない文体であると解析結果の精度が悪くなる。その結果、作成されるインデックスも低品質なものとなり、検索ノイズ（検索結果に無意味な情報が混じる）が増えたり、検索もれ(検出して欲しい情報が検出されない)が生じたりする｡n-gram では単語辞書を使わず、テキストを 2〜3 文字連続ずつをそのまま区切り､それらをインデックスの単位とする｡そのため、意味のない単位で区切ることも多くなるので、検索もれはなくなるが、作成されるインデックスのサイズも大きくなり、速度面で不利になったり検索ノイズも多いものとなる。表１にそれぞれの検索方法の特徴とその比較を詳しく示す。

表１．各種の検索方法によるシステムの特徴

項目		シーケンシャル型		インデックス型	
検索対象ファイル		ベタテキスト	形態素解析済み テキスト	形態素単位の転置型 インデックスファイル	n-gram(2, 3 文字連続)単位のインデックスファイル
特徴		文字列検索	単語･複合語など意味単位での検索	単語･複合語など意味単位での検索	文字列検索
検索速度		低速	低速	高速	やや高速
検索対象テキスト適正量		少量(〜20MB 程度)	少量(〜20MB 程度)	大量	大量
前処理速度		−	やや高速 (形態素解析)	やや低速 (形態素解析＋インデックス作成)	やや高速 (インデックス作成)
インデックス容量		−	−	やや大	やや大(転置型インデックスの場合) やや小(シグネチャ型インデックスの場合)
ファイル容量		最低容量 (ベタテキストのみ)	少量 (ベタテキスト＋形態素区切り文字)	やや多量 (ベタテキスト＋形態素単位のインデックス)	多量 (ベタテキスト＋n-gram 単位のインデックス)
正規表現検索		可能	可能	一部可能 (単語内のみ)	一部可能 (n-gram 単位内のみ)
近接単語(フレーズ)検索		−	可能 (キーワードとして指定可能な任意数)	単語数に制限あり (2, 3 単語程度) 英語のみのことが多い	−
検索結果品質	検索ノイズ	大	形態素解析精度に依存	形態素解析精度に依存	大
	検索もれ	なし	小	形態素解析精度に依存	なし

3.　比較的少量の文書に対する高機能、高品質検索

比較的少量のテキストであり、繰り返し参照され、読み捨てになることがないために高機能で高品質な検索ができることが望まれる情報がある。つまり、検索速度よりも検索ノイズや検索もれがなく、かつユーザ側でさまざまな検索条件を指定できることに重点をおく場合である。例えば、辞書、事典、経典、古典といったジャンルの情報である。特に、経典や古典の場合は単なるキーワード検索だけでなく、行や節、パラグラフ内などの位置的に近接するキーワード（単語）との関係(共起関係の強度や距離など)や出現位置・順序を調べたり、文脈付き用語索引(KWIC: Key Word In Context)によりキーワードの使われ方を調べたりすることもある。このような検索を行ないたい場合は、それらの機能を提供している検索システムを利用する必要がある。このような検索システムはほとんどがスタンドアロン型の専用プログラムであり、ハードウェアや OS、操作方法などの作業環境に依存したものになる。どの端末からでもどこからでも気軽に利用できるようにするには、これらの機能を Web で提供するのが望ましい。また、ほとんどの Web 検索システムや全文検索システムでは大量の Web ページやテキストファイルを扱う必要があるためにインデックス型を採用している。このため、どこまでの情報をインデックスファイルに持つかで機能が制限される。したがって、このような検索システムの実現の方針としては、単純なキーワード検索や簡単な検索条件付きの場合は高速に、複雑な検索条件付きの場合はキーワードの出現位置をもとに元のテキストを精査してでもそれなりの時間がかかってもよいというのがユーザ側から最も受け入れやすいシステムとなる。

ここでは、形態素解析済みの日本語テキストを対象としてシーケンシャル型の検索システムをベースとして柔軟性と高機能性、高品質を確保し、また検索速度面からインデックス型検索機能も可能にしたシステムを実現した。形態素解析済みテキストを検索対象としたのは、単なる文字列検索では検索ノイズが大きすぎるためと検索ノイズをなくすために日本語単語（形態素）境界を検出するためである。また、ユーザがキーワードとして指定する文字列は単語や複合語、フレーズなど、意味のある文字列であることがほとんどであり、それならばなるべく意味単位に近い方法で検索できるのが検索システムとして望ましいためである。

図２．検索システムの実現例

4. 実験

比較的少量のテキストであり、かつ柔軟で高品質な検索機能が望まれるテキストとして、「新改訳聖書」[3]（ベタテキストで約4MB、形態素解析済みで約5MB、形態素単位で出現位置のみを記録したインデックスファイルは約 12MB）を選択し、Web サーバシステムの CGI(Common Gateway Interface)機能を利用して実験的に検索サービスを公開・提供してきた（図2参照。http://www.tuins.ac.jp/~takao/biblesearch.html)。2002年～2003年末までの約1年半の期間で12万6千件を越える検索リクエストを受け付けた（1日あたり200～300件に相当)。なお、現在、この検索ページは、Google、Yahoo、goo、BIGLOBE、Excite（以上の検索サイトはGoogleエンジンを利用)、MSN searchといった主要な検索サイトにおいて、「聖書」と「検索」をキーワードとしてAND検索すると検索結果のトップに位置し、alltheweb、altaVista、freshEYE、Infoseekなど他の検索サイトでも検索結果の5位以内に位置している。

形態素解析システムには、「茶筌(chasen)」[4]を利用した。単語辞書と形態素間連接規則表は情報処理振興事業協会(IPA)のものがベースになっており、標準状態のままでは検索対象テキストに対して解析精度がよくないので、約3,800語の単語登録と約90個の形態素間連接規則の登録を行ない、検索対象テキストが文法的に正しく解析できるようにした。追加した単語は、主に固有名詞の人名1,700語、地名1,000語で、他には専門用語と単語コスト値を低くした高頻出語などである。追加した連接規則には、標準の連接規則表の不備の部分と、検索対象テキストに特有な文体を処理できるようにするための規則がある。例えば次の規則を追加した。

```
(((((動詞 自立) * 未然形)) (((動詞 接尾) 一段 連用形 せる)) (((動詞 非自立) 五段・ラ行特殊 * なさる))) 4000)
```

この規則では、「その実を摘み取らせなさるのですか。」のようなテキストについて、「摘み取ら」（動詞-自立語『摘み取る』のラ行五段活用未然形)、「せ」(動詞-接尾語『せる』の一段活用連用形)、「なさる」(動詞-非自立語『なさる』のラ行特殊五段活用基本形）の3連続形態素をコスト値4000で連接可能にする規則である。

このように単語辞書と形態素間連接表の整備とチューニングを行なうことで形態素解析精度が格段に向上し、本検索システムでの利用に十分耐えうるものとなった。比較的少量の日本語文書に対して高機能で高品質な検索を行なう場合には、このような整備は欠かすことができない。未登録語や解析誤り箇所の抽出やコーパス（SGMLやXMLを利用して文法タグや意味タグが付けられた言語資源）作成のためのツールについては、解析システムの文法体系に依存したり、作業者の経験と勘に頼る部分もあり、現在研究中のものが多い。本実験では、未登録語抽出や解析誤りに関するヒューリスティック規則を利用することで半自動的に行なった。

表2．検索指定

検索指定			回数	割合
単一キーワード指定			106,831	84.8%
複数キーワード指定	AND	近接指定なし	9,862	7.8%
		近接指定あり	3,096	2.5%
	OR	近接指定なし	1,573	1.2%
		近接指定あり	734	0.6%
巻名章節番号指定			3,907	3.1%
計			126,003	100.0%

図３．近接指定検索の例

検索キーワード指定について､単一指定の回数､複数キーワード指定の場合は AND、OR、キーワードの近接指定の有無別に集計した結果を表２に示す。巻名章節番号指定は、直接本文の位置を指定する検索である（本実験で検索対象としたテキストには、各節ごとに、章と節番号が付けられている｡図３参照)｡全体で 126,003 件の検索のうち約 85%にあたる 106,831 件が単一キーワードで検索しているという結果が得られた。次いで同一行・節に複数キーワードを同時に含む検索が 7.8%、近接行あるいは近接節指定を行なっている検索が 2.5%あった。このような検索が少なからず行なわれていることは、複数のキーワードが使われているある程度の範囲（パラグラフレベル）をキーワードとその関係をひとかたまりのイメージ(チャンク)としてユーザが記憶しており、そのような箇所を検索したい場合があること示している。近接指定検索の実際例を図３に示す。

表３．検索キーワード

順位	検索キーワード	回数	割合
1	愛	1,414	1.01%
2	神	1,144	0.81%
3	罪	713	0.51%
4	主	655	0.47%
5	イエス	519	0.37%
6	キリスト	467	0.33%
7	666	434	0.31%
8	聖霊	400	0.28%
9	死	367	0.26%
10	光	363	0.26%
11	ヨハネ	362	0.26%
12	霊	341	0.24%
13	天	319	0.23%
14	信仰	311	0.22%
15	心	294	0.21%
16	かなかす	289	0.21%
17	モーセ	280	0.20%
18	ことば	274	0.19%
19	父	273	0.19%
20	アダム	271	0.19%
21	羊	256	0.18%
22	十字架	253	0.18%
23	人	239	0.17%
24	パン	237	0.17%
25	サタン	233	0.17%
26	いのち	231	0.16%
27	六百六十六	228	0.16%
28	御霊	223	0.16%
29	天使	221	0.16%
30	ヨブ	219	0.16%
31	神の国	211	0.15%
32	血	211	0.15%
33	ユダ	203	0.14%
34	道	199	0.14%
35	目	198	0.14%
36	蛇	198	0.14%
37	マタイ	197	0.14%
38	結婚	196	0.14%
39	水	193	0.14%
40	恵み	188	0.13%
延べ語数合計		140,670	100.0%
異なり語数合計		45,681	

検索キーワードの内容とその検索回数について集計を行なった結果を表３に示す。これらのキーワードは検索対象テキストにおける頻出語・主要語のために繰り返し検索が行なわれているとはいえ、検索の延べ語数合計から見ると多くても 1%程度とそれほど多くはなく、異なり語数合計の 45,681 語を見ても、非常に多様な検索が行なわれていることがわかる。また、数字も検索対象文書によっては重要な意味を持つ場合がある。本システムでは、数字表現のゆれを吸収するため、算術表現(「666」など)や漢数字表現(「六百六十六」など)、それらの混在表現(「14.4 万人」など)は、文字コードはもちろん、必要ならば算術演算を行ないながら正規化してから検索を行なっている。このため、数字表現を検索キーワードに指定された場合でも検索もれを生じることがないようになって

いる。今回の集計結果でも 5,350 件と全体の約 4%が数字検索となっている。一般の検索システムでは特に意味がある数字でない限り検索を行なうことはあまりないのに比べると対照的である。

さらに、検索キーワードには正規表現[5]を指定することができる。正規表現を利用することにより、表記ゆれを含むあいまいな検索、キーワードの出現順序を指定した検索、行、節、パラグラフ内のキーワードの相対位置を指定した検索が可能となる[6]。正規表現にはこれらの指定を行なうためにメタ文字を使うことでこれらの機能を可能にしている。これまでに利用された回数を各メタ文字別に分類したものを表４に示す。正規表現については、使うのに敷居が高いためか今回の集計結果では全体の 0.56%程度しか利用されていない。その中でも文字集合のメタ文字を利用する場合が圧倒的に多い。これは、主にキーワードの表記のゆれを吸収するのに利用されることが多いという結果が得られている。例えば、「マリア」と「マリヤ」の双方の可能性を考慮しながら検索する場合は正規表現で「マリ[アヤ]」とすれば表記の違いによる検索もれを防げることになる。この知見に基づき、イ段とエ段に続く「ア」、「ヤ」についてはユーザがキーワードとして明示的に正規表現を指定していない場合には自動的に正規表現を付加することにして検出失敗というユーザにとっての予想外の事態を避けることにした。他にも、「イ[ァ-ヶー]*ル」とすれば、「イ」で始まり「ル」で終わるカタカナ文字列をもれなく検索することができる。正規表現を利用した検索を行なった例を図４に示す。ここでは KWIC 検索のために正規表現を利用している。

表４．検索キーワードに使われた正規表現の種類

表記	意味	回数	割合
[…]	文字集合	600	73.0%
*	0 回以上の繰り返し	181	22.0%
?	0 回または 1 回の繰り返し	101	12.3%
.	任意 1 文字	46	5.6%
(…)	グループ	27	3.3%
^	行頭	15	1.8%
+	1 回以上の繰り返し	7	0.9%
[^…]	補集合	6	0.7%
$	行末	6	0.7%
|	選択	3	0.4%
{n, m}	n 回以上 m 回以下の繰り返し	0	0.0%
延べ回数		992	
異なり回数(キーワード全体の 0.58%)		822	100.0%

図４．KWIC 検索の例

検索キーワードに複合語やフレーズが指定された場

合は、その出現箇所の特定のために、複数形態素をまたがる検索を行なう。ただし、検索キーワードのうち、数字やカタカナ表記の部分はそれ以上に分割した形態素にマッチしないようにしている。例えば、検索キーワード「70」や「七百」ではテキストの「七百七十七」の部分にはマッチしない（なお、検索オプションにてマッチさせるようにすることもできる）。

検索を行なった場合のレスポンスであるが、実験に使用したテキストの分量では、検索そのものにかかる時間よりは、ネットワーク遅延やWebサーバからCGIプログラム呼び出しと起動、ファイルアクセスにかかるオーバーヘッドのほうが大きく、どのような検索キーワードであってもインデックス型とシーケンシャル型のいずれでも大差ない結果であった。インデックス型では、インデックスファイル自体のサイズがテキストのサイズと比べてかなり大きいのでファイルアクセスの面でシーケンシャル型に比べて不利な面がある。もう少し検索対象のテキストの容量が増えた場合にはインデックス型の効果が現われてくると思われる。これについては、計算機の能力や検索リクエスト頻度などとも関係するため、実験を重ねながら最適な選択を行なうことになる。

5. まとめ

比較的少量のテキストを対象に柔軟で高機能かつ高品質な（検索ノイズがなく、かつ検索もれもない）検索システムを実現し、約1年半にわたるWebによるサービス提供により、どのような使われ方をしているかについて分析した結果について述べた。検索の機能面については一般のユーザは正規表現を使った検索はあまり行なわず、むしろ高機能検索を利用する場合は複数のキーワードとそれらの関係を想定した検索（結果として近接行・節検索になる）を行うことが多いことが明らかになった。検索キーワードの表記のゆれを吸収するのに正規表現がかなり有効であるにもかかわらず、これを利用しなかったために望んでいる検索結果を見つけることができなかった例もかなりある。この表記ゆれについては、システム側でもっとサポートすべきかも知れない。また、検索システムについては、検索対象の文書の選択、どのような検索をどのくらいの頻度で行なうか、実現に必要なコストはどのくらいかなどによって検索方式を使い分けていくことが必要である。今後は、柔軟性や高機能、高品質を保ったままスケーラビリティと検索速度をどのように確保していくか、シソーラスなどを利用した関連語検索、表記的にも意味的にもかなりあやふやなキーワードでも適切な検索を行なえるようにすることなどが課題である。

参考文献

(1) Tara Calishain, Rael Dornfest, Google Hacks, O'Reilly (2003)
(2) 馬場 肇、日本語全文検索システムの構築と活用、ソフトバンク社 (1998)
(3) 新改訳聖書第2版、日本聖書刊行会、いのちのことば社 (1981)
(4) 松本裕治、「形態素解析システム『茶筌』、情報処理 Vol.41 No.11, pp.1208-1214 (2000)
(5) Jeffery E. F. Friedl(田和 勝 訳)、詳説 正規表現 第2版、オライリー・ジャパン社 (2003)
(6) 新田義彦、佐良木昌、正規表現とテキスト・マイニング、明石書店 (2003)