インターネット研究：情報検索について
On information retrieval

言語・情報講座　教授　舟　阪　　晃
funasaka@post01.osaka-gaidai.ac.jp

0　まえがき

人間の知の源泉は、以前は、知識の蓄積そのものであったが、現在は、蓄積した知識をいかに効率よく、また、適切に活用できるかに依存している。コンピュータが出現する以前は、一般的には、知識の蓄積が増えれば増えるほど、その有効な活用は困難になるといわれていた。つまり、手元の知識の蓄積のなかに、関連のある (relevant) 情報があることはわかっていても、それを効率よく見つけ出すことは容易なことではない。ある文筆家が、書斎のなかの混乱した書籍・文献のなかから関連のあるものを瞬時に見つけ出すことを「名人芸」といわれるのは、理由のあることである。現在は、コンピュータの強力な検索技術が名人でない極く一般的な人々も利用できるようになり、事態は大きく変わってきている。それにともなって現在注目されるべきキーワードのひとつは、「情報検索」で、本稿ではその諸相について考察を加えるつもりである。

知識の蓄積に関して、最初に想起されるのは図書館の存在である。膨大な情報の蓄積があり、年々蔵書の数は増えることはあっても減ることはない。しかし、コンピュータの検索システムが導入されるまでは、検索技術のレベルは非常に低いレベルであったといえる。図書館では、文献ごとに分類番号を

つけ整理されている。同時に、図書カードが作成され、書名別に、また、著者名別に分類が行われていることが多い。そもそも、分類するというのは検索のための最初の、また、一番初歩的な段階であるといえる。いかに利用しやすい分類方式をとるかということは、いかに検索を効率よく行うかということと関連している。図書館によっては司書が配置されているが、その役割のひとつは、図書館の検索機能の補強ということもできよう。しかしながら、図書館のこれまでの機能をみてみると、基本的には、情報の蓄積能力は大きいが、検索機能のレベルは低かったといわざるを得ない。その後のコンピュータの導入による検索能力の高度化は、図書館の機能を革命的に改善したといえよう。

図書館とは情報量がちがうが、一冊の図書も知識を蓄積したものと考えることができる。図書の場合も、知識の蓄積が中心で、検索は二の次にされている。検索に貢献するのは、目次と索引であるが、とくに検索にとって重要な索引は、「遠慮がちに」最後のページに付け加えられており、索引項目もあまりにも不十分なものが多い。筆者の個人的な見解でいえば、図書の最初の部分に索引表があるような図書が出版されてもいいと思う。一方、辞書、事典の類は、項目のアルファベット順、アイウエオ順の配置が見られ、検索を前提にしているが、紙製のものでは検索時間が長くなり効率が悪い。最近は、いわゆる電子辞書が、紙製のものに取って代わる傾向が見られるのも当然であるといえる。

個人的なレベルでも、住所録、名刺ホルダ、スクラップブックなど、情報の蓄積と検索は日常的に人々の注意の対象になっており、アルファベット順、アイウエオ順、年月日順、事柄順、業務別など、効率のいい検索のための分類は苦労する点である。

コンピュータの出現とその検索能力の活用は、情報検索に革命的な変化をも

たらしたといえる。とくに、検索のスピードは、いろいろの分野に影響を与えている。たとえば、言語学の語法研究で、膨大な文献のなかからある特定の語法例を収集する仕事は、コンピュータが出現する以前は、長時間の、しかも骨の折れる仕事が必要となり、それ自体ひとつの業績に勘定されたといえるが、いまや、そのような仕事はコンピュータにまかせれば、ほんの一瞬で結果が入手できる。コンピュータの出現により、どちらかといえば、蓄積を重視してきたこれまでの知的な作業は、検索重視に方向転換をすることになる。情報検索についての考察が重要である所以である。

情報検索には、明確な問題意識と、適切なキーワード、豊かな想像力が必要とされる。このような必要条件は、すべての人によって充足されるものではないので、将来的には、「情報検索士」というような人の存在が必要とされ、そのような資格が認められることになろう。情報検索士は、クライアントの依頼をうけ、場合によっては素人には思いもつかないようなキーワードを駆使し、検索を行い、不必要なノイズを除去し、素人にも読めるかたちで情報を提供したり、答えを提供したりし、それに対する報酬を受けることになろう。ちなみに、1993年から、情報科学技術協会が、「情報検索基礎能力試験」という資格試験を実施しているが、どちらかというと情報検索の知識についての試験のように思われる。一方、実用的な検索能力については、雑誌社などが企画する「検索の鉄人」などがある。今後は、この両面をまとめたような公的な資格試験が必要になるであろう。

1　情報について

情報ということばは、多くの人によって、いろいろの状況で、いろいろの意味で用いられる。三省堂の『大辞林』（第二版）によれば、つぎのように定義されている。

（1） 事物・出来事などの内容・様子。また、その知らせ。

（2）　ある特定の目的について、適切な判断を下したり、行動の意思決定をするために役立つ資料や知識。

（3）　機械系や生体系に与えられる指令や信号。

（4）　物質・エネルギーとともに、現代社会を構成する要素のひとつ。[引用おわり]

（1）は、極く漠然と使われる情報で、（2）は、気象情報、交通情報などが、（3）は、遺伝子情報などが、それぞれ、該当する。（4）は、最新版にはあるが、旧版にはない語義で、情報という言葉の語義の流動性が明確になり興味深い。

情報の価値は、当事者によって決定される。つまり、ある情報がすべての人にとって同じような価値があるということは考えにくい。たとえば、交通渋滞情報は、該当する道路を同方向に走行するドライバーにとっては価値のある情報であるが、その渋滞に関係のないドライバーにとっては価値のない情報である。また、台風が襲来することを予測する情報は、瓦の葺き替え中の住宅にとっては好ましくない情報であるが、水不足で節水を余儀なくされている地域にとっては朗報であるといえる。まとめていえば、情報には何らかの潜在的価値はあるといえるが、それが実際上どういう価値と認定できるかは、情報の受け手の主観的・相対的価値判断によることになる。

情報の検索という観点からすれば、上の情報の定義の（1）から（3）まで、情報の中にふくまれ、その中から、いかにして価値ある情報を効率よく引き出すことができるかが重要になる。検索の対象となる情報は、規模の大小の差はあるが、何らかの形のデータベースを構成している。住所録、名刺のリストなどは目的の特化した規模の小さいデータベース、図書文献目録などは、規模の大きいデータベース、インターネットは、汎用の規模の大きいデータベースといえる。同時に、これらの情報は、文字に限らず、画像情報、音声情報も、場合によっては、含んでいる。

2　データベースについて

データベースの定義は、関係する分野により違いがある。最も簡単な定義は、「組織的に蓄積されたデータのこと」(『世界大百科事典』) となるが、『情報検索の基礎』には、つぎのような定義とそのまとめが与えられている。

著作権法の定義：論文、数値、図形そのほかの情報の集合物であって、それらの情報を電子計算機を用いて検索することができるように体系的に構成したものを言う。
JIS の定義：１つ以上のファイルの集まりであって、その内容を高度に構造化することによって、検索や更新の効率化を図ったもの。
通商産業省のデータベース台帳に関する規則における定義：データを整理・統合し、電子計算機による検索を行いうる形態にした集合体をいう。
まとめ:大量の情報をコンピュータ処理できるようにシステム化して蓄積し、必要なときにコンピュータですぐに検索できるように整理・統合したもの。
(情報科学技術協会（編）1997：25-26)

上の定義を参考にして、データベースの成立のための重要な要因を上げてみるとつぎのようになる。

（1）　基本的には、効率のいい検索のために、情報・データの集合を体系的に整理し構造化したものとなるが、インターネットにみられるように、効率のいい検索手段（たとえば browser）があれば、体系化・構造化は必ずしも絶対的な条件ではない。
（2）　情報・データのメディアは、文字情報のほかに、数値、画像、音声を含みうる。
（3）　コンピュータによる検索が中心になるが、それ以外の検索法をとってもよい。

データベースはいくつかの観点から分類できる。まず、一方に、特定の目的に特化されたデータベース、他方、汎用のデータベースが、さらに、両者の間に連続的に変化する程度差が認められる。たとえば、特化されたデータベースとしては、個人が作成した住所録・電話帳などがわかりやすい。電話局が作成した電話帳は、電話という特化は働いているが、個人のそれよりは汎用の方向に向いている。現在一番汎用データベースとしてわかりやすいのはインターネットであろう。データベースは、特化されればされるほど検索がしやすくなる。個人の住所録・電話帳などは、登録されている人数によるが、検索は容易である。一方、インターネット上の情報は玉石混交で、いわば体系化・構造化されていないといえるが、その検索はブラウザーを活用することにより実現される。ここに、ブラウザーの存在理由がある。

3　検索について

検索は、コンピュータの出現により飛躍的に効率のいいものになったが、それ以前にも検索という行為はあった。たとえば、カードに穴やくぼみをつけ、それを頼りに同種類のカードをまとめるという検索方式が1940年代に始まっている。(情報科学技術協会（編）1997 : 3)

検索の目的は、大きく二つに分けられる。一方は、ユーザやクライアントの問題に対して最終的な解答を与えるのを目的とし、他方は、問題解決のための材料を提供するものである。前者に該当するものとしては、体の健康状態についての質問にこたえると、それにもとづいて、病名を知らせてくれるようなエクスパートシステムが一例としてあげられる。この場合は、コンピュータが、情報に基づいてある種の判断を行ったといえる。後者の方は、種々の程度差が認められるが、最終的な結論は、提供された情報をもとにユーザまたはクライアントが行うことになる。もちろん、両者の区別が判然としない場合もある。たとえば、データベースとしての百科事典により何らかの情報

をえた場合、それで問題のすべてが解決されることもあるが、また、それを判断材料のひとつとして、ユーザまたはクライアントが最終的結論を出すということもある。

検索が最終的な結論を提供する場合は、すべてコンピュータにまかせておけばいいが、結論を引き出すための材料が提供されるだけという場合は、「検索精度」が問題になる。「検索精度」には、検索結果の「的確性」と「包括性」が含まれる。実際に検索を行ってみると、入力したキーワードが、内容的にまったく関係のない「ノイズ」を含むことがある。このノイズを取り除くことにより「的確性」を高めることができる。一方、特定のキーワードがしかるべき情報をすべてカバーしておれば「包括性」が高くなる。ここでむつかしいのは、「包括性」を高めるためには、検索の範囲を広くとらねばならないので、必然的にノイズが多く含まれることになる。ノイズを如何に少なく、的確な情報をいかに多く入手するかということが重要になる。「情報検索士」の腕の見せ所といえよう。

検索には、通常、AND、OR、NOT のブール代数（Boolean algebra）の演算子が用いられる。(X AND Y) は、情報の中に X と Y の両方が含まれるもの。X と Y の順序は不問で、また、X と Y の間にべつの項目が入ることは許される。つぎに、(X OR Y) は、X または Y が含まれるもの。さらに、(X NOT Y) は、X は含まれるが、Y は含まれない。実際の検索の際には、通例は、AND を多用し、OR は、「ノイズ」が多くなるので、ヒット数が極端に少ないと予測される場合以外ではほとんど利用しない。NOT は、それが付与された項を排除して検索するときに用いられる。一例として、「仮想」、「大学」をキーワードとして検索した結果を示めすと次のようになる。丸括弧内は、随意的に選択される任意の記号列を含むが、ゼロの場合もある。

仮想 AND 大学:「(...) 仮想 (...) 大学 (...)」、「(...) 大学 (...) 仮想 (...)」
仮想 OR 大学:「(...) 仮想 (...) 大学 (...)」、「(...) 大学 (...) 仮想 (...)」、

「(...) 仮想 (...)」、「(...) 大学 (...)」、

仮想 NOT 大学：「(...) 仮想 (...)」　(...) は「大学」を含まない。

これらの演算子は、サーチエンジンにより反応がちがう場合がある。たとえば、Yahoo では、AND、OR は上のとおりの機能を果たすが、NOT は、上記の機能では使えない。一例をあげれば、「仮想 AND 大学 NOT アメリカ」で検索をすると、NOT は AND としての機能しか果たしていない。したがって、「仮想」、「大学」、「アメリカ」が含まれるサイトが検出される。また、「仮想　大学」のように、キーワードの間にスペースを入れた場合は、AND 検索と同じ機能になるが、「仮想大学」というキーワードを含むサイトが優先的に表示される。

4　キーワードの表記と検索

キーワードの入力に際して、その表記の仕方により検索結果がちがってくることがあり、検索精度を高めるという観点から、重要な問題が生じてくる。いくつかの代表的なサーチエンジンに、表記のちがうキーワードを入力し検索結果の違いを調査してみる。キーワードは、日ごろの検索の際問題を感じたものを恣意的に選んでいる。エンジンには、大きく分けて、ディレクトリ型検索とロボット型検索があるので、数値が大きな差を示すこともあるが、同一のエンジン内の数値を比較することにより、検索語の取り扱い方策を垣間見ることができる。下表の数値は、2001年9月現在のものである。

4.1 同義異形（英語）

	Yahoo Japan (件)	goo (pages)	Google (件)	infoseek (件)
Second language acquisition	1540	4982	477000	1943
Second-language acquisition	1110	1924	116000	1004
2nd language acquisition	334	1213	637	383
L2 acquisition	379	848	30400	375

一般的に、英語の場合は、比較的異形が少ないので、日本語の場合よりは問題は少ないが、上表のようなヒット数のばらつきがみられる。最初のふたつの違いはハイフンの有無だけであるが、goo と Google では、両者の間にヒット数の大きな違いが認められる。最後のふたつは、相対的にヒット数は少なく、Google を除けば、ほぼ同じようなヒット数を示している。

4.2 同義異形（日本語）

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
こども	97/698	13553/43228	559000	300000over
子ども	92/1771	32352/198456	1240000	300000over
子供	86/609	91669/412328	3050000	300000over

英語の場合は問題にならないが、日本語による検索に際しては、文字種のちがい、異体字などに注意が必要である。KODOMO で検索をするとき、上表

のように、KODOMO の文字種により検索の結果がちがう。(注：大文字アルファベットは文字種を無視した表記を表すものとする)　Yahoo Japan では、「こども」「子ども」が比較的多いが、goo、Google では、逆に、「子供」の方がヒット数が多い。

あるサイトで検索をしたところ、実在する『子どもたちの言語獲得』は、『子供たちの言語獲得』や『こどもたちの言語獲得』ではヒットしない。検索に際して利用者の注意が必要とされるところである。

4.3 同義異形

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
デジタル	46/1706	42399/224296	1400000	300000over
ディジタル	5/34	4953/18516	113000	300000over
デズニー	916pages	130/160	1060	50356
ディズニー	8/157	7681/19127	213000	50356

「デジタル」と「ディジタル」は前者が、「デズニー」と「ディズニー」は後者が、それぞれヒット数が多くなり、国語辞典などの表記と一致する扱いになっているが、検索の際には注意が必要である。「デズニー」、「ディズニー」の場合、Yahoo Japan では、ヒット数集計の単位にちがいがある。つまり、「ディズニー」は、範疇やサイトで集計されているが、「デズニー」は出現するページが勘定されているだけである。また、infoseek では、表記の違いを無視している点が特徴的である。

4.4 同義異形

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
ビタミン	2/42	14203/43510	301000	84017
ヴィタミン	1190pages	40/65	1160	84017

この結果は予想通りである。ベートーベン、ヴェートーヴェンの場合も含めて国語辞典などの表記と一致している。infoseek では区別をしていない。

4.5 同義異形（長音記号１）

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
メイル	4/13	5385/20217	139000	300000over
メール	30/2487	188884/1208384	9350000	300000over

4.5から4.8までは長音記号を扱っている。「メール」、「メイル」については、前者がふつうであるといえるが、infoseek では両者の区別をしていない。

4.6 同義異形（長音記号２）

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
コピー	3/303	38770/119786	968000	300000over
コピ	6/323	1229/1968	512000	12833

4.7 同義異形（長音記号３）

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
ユーザー	28/490	49591/277094	1500000	300000over
ユーザ	30/572	22764/165840	658000	173643

4.8 同義異形（長音記号４）

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
コンピューター	209/414	28691/74735	603000	300000over
コンピュータ	486/2241	56873/291143	1860000	300000over

4.6から4.8までは、単語の最後の長音記号の扱いを調べている。「コピー」と「コピ」、「ユーザー」と「ユーザ」については、Yahoo Japan では、長音記号がない方がヒット数が多く、その他のエンジンは逆の結果をだしている。表記上のちがいをあまり重視していないように思われる infoseek が、この点では区別をしていることは特徴的である。また、「コンピューター」と「コンピュータ」については、infoseek を除いて、すべてのエンジンで長音記号のない方がヒット数が多くなる。文字数が多くなるにつれて、最後の長音記号は省略される傾向にあるといえる。

4.9 同義異形

	Yahoo Japan (pages)	goo (sites/pages)	Google (件)	infoseek (件)
アルビン・トフラ	20	2/3	20	0
アルビン＝トフラ	20	0/0	20	0

アルビントフラ	5	1	5	12
アルビン　トフラ	20	2/3	21	0

英字の人名を日本語のカタカナでどう表記するかは悩むところであるが、表記の仕方によって検索に影響がでてくる。上表のように、Yahoo Japan と infoseek は、同じ原則にたっているようで、日本語で多用される表記法はカバーしている。一方、goo では、アルビン＝トフラではヒットしないことになる。また、infoseek では、「アルビントフラ」のみが認められている。なお、「アルビン」と「トフラ」の間に空間をあけた場合は、「アルビン・トフラ」と等価になる。

4.10 類語

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
言語習得	1/9	390/826	4810	14618
言語修得	1/0	32/47	195	4691
言語獲得	1/0	159/336	1450	14351

通例、language acquisition の日本語訳は、上表のように、必ずしも唯一的に決まっているわけではなく、検索に際しては、この事実を知っている必要がある。

4.11 類語

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
米国	8/263	30005/170582	1280000	300000over
米	72/1914	35209/171575	3740000	300000over
アメリカ	89/1769	72806/360564	2390000	300000over
合衆国	4/52	2696/5912	107000	34802
USA	2/1962	8438/40317	32600000	172407
US	544/24590	18695/125875	174000000	181838
アメリカ合衆国	2/43	4976/9915	64500	15889

上表のように、ひとつのことを表現するのに多くの類似した表現があり、検索の際、工夫が必要になることがある。このようにヒット数が多い場合は、どのように絞り込むかということが重要な問題になる。

4.12 類語

	Yahoo Japan (cats/sites)	goo (sites/pages)	Google (件)	infoseek (件)
PDA	2/75	4003/20910	2980000	31826
個人情報端末	0/336	24/33	359	1661
携帯型情報機器	0/1	38/91	525	245

上表は、新しく出現した英語表現に対する日本語訳がまだ決まっていない例を示している。PDA は Portable Digital Assistant の頭字語である。

5　あとがき

われわれの周辺の情報の蓄積は日々増加している。問題は、その情報のかたまりから関連のある適切な情報をいかにして効率よく入手できるかということである。情報検索は、適切な問題意識、適切なキーワード、豊かな想像力を必要とするが、素人のだれもができることではない。「情報検索士」が必要とされる所以であるが、検索士の機能には、クライアントの問題に対して最終的な解答を与えるものと、必要な判断材料を提供してクライアントの判断を助ける働きがある。本稿では、検索に際しての表記上の問題の一面に焦点をあて、具体的な問題点をあきらかにした。(2001.11.30)

参考文献

『大辞林』三省堂　第二版 1995.
『世界大百科事典』日立デジタル平凡社 1998.
情報科学技術協会（編）『情報検索の基礎』 1997.
その他インターネット上の情報