知識集約型データベース検索システムの構築

服　部　　忍*

Construction of knowledge intensiveness type data base exploration system

Shinobu HATTORI

In this paper, the knowledge intensiveness type data base exploration system is handled. The proposed data base exploration system constructs it as a search system corresponding to the data base that the common background knowledge exists between the database construction person and the database retrieval person. This search system assumes use such as TLO (technology licensing organization) as the liaison support system. In this research, the method of consolidating the knowledge of the data base registrant and use is examined, and it proposes the knowledge intensiveness type data base exploration model that can be retrieved effectively and efficiently in a specific group.

キーワード：キーワード検索，置換え可能語，一致率，知識集約型，ＴＬＯ

１．は　じ　め　に

データベースの情報検索において，様々な検索モデルが用いられている．データベースの情報検索では，必要な情報を有する情報源を対象とするデータベースから，いかに漏れを無くし網羅的に検索するか，また，無駄な情報をいかに少なくするかが最も重要な事柄となる．このことが検索の信頼性や効率化に大きく影響し，効果的な検索の実現を左右する，

しかしながら，現在用いられている検索モデルのほとんどは，利用者を十分に満足させているわけではない．実用的な検索モデルである全文検索モデルとキーワード検索モデルにおいても，必ずしも検索者の意図する検索ができるとは限らない．検索の結果に満足できずに，異なる検索語（検索する語句）を用いて複数回検索を実行しても，望む結果が得られないことも多い．

本研究では，データベース登録者の知識を集約して利用する方法について検討し，特定集団において有効かつ効率的に検索できる知識集約型データベース検索モデルを提案する．

*電子制御工学科助教授

原稿受付　2002年5月17日

２．各種検索法とその課題

現在，代表的な情報検索システムのモデルは，全文検索モデルとキーワード検索モデル（Booleanモデル）である．また，最近，実用化されたベクトル空間モデルも有力な検索モデルである．

２－１　全文検索モデル

全文検索モデルでは，テキストそのものを検索対象としている．検索では，複数の検索語を論理演算子（ＡＮＤやＯＲなど）で関係づけて入力とし，検索結果は検索語を含むテキストの集合として出力される．テキストは計算機内の情報表現であるため，検索用の特別な前処理が不要である．そのため小規模なシステム処理系で実現でき，維持費等においても低コストで済む大きな利点がある．

一方，この全文検索システムの課題としては，検索適合率が低いことが挙げられる．全文検索システムでは，検索したい語句を文字列として扱い，検索キーとしている．そのため意味をなさない文字列とも照合する結果，無駄な照合が生じる．この対策としては，全文検索の結果を言語解析する方法が採られている．また，検索語と一致する文字列がテキス

トに現れないと検索できないため再現率も低いため，その対策として，シソーラスを用いた同義語等によって照合する方法が行われている．さらに，全文検索は文字列の一致のみで検索するため主題と関係の無い部分も検索されてしまう場合が多いという欠点も有する．それぞれの課題に対する改善策も検討され実施されているが，そのために，全文検索の利点である小規模なシステム処理系での実現や維持費等の低コストが損なわれている事は否定できない．

2-2 キーワード検索モデル

キーワード検索モデルは，テキスト内容を代表するキーワードをキーとして検索を行うモデルで，情報検索システムで従来から多く用いられている．検索語を入力する方法は全文検索と同じであるが，検索結果はキーワードが付与されたテキストの集合である．キーワード検索モデルは，キーワードごとにそのキーワードが出現したテキストを記録する転置ファイルを作る方式を用いているため，高速検索が可能である，また，キーワード設定時にテキスト内容に適したキーワードを人手あるいは自動的に言語解析やテーマ解析で設定できるため，高い検索精度が得られる．

しかしキーワードを付与するために，人手の場合は膨大な労力が必要であり，キーワード自動抽出システムを利用した自動化の場合はシステムに言語解析系を組み込む必要がある．高精度な言語解析系システムの構築には多くの困難が伴っている．いずれにしてもメンテナンスや運用維持費で高コストになりやすい欠点がある．

また，キーワード検索モデルを改良したものに，検索語に重要度を付加する拡張キーワード検索モデルがある．このモデルは，検索の一致度や重要度などの付加価値を加えたもので，システムの利便性を大幅に向上させることができる．

2-3 ベクトル空間モデル

ベクトル空間モデルは，テキストを単語や概念のベクトルとして表現するもので，設定された単語(キーワード) 等に重要度等の情報を割り当てて，より高い精度の検索を実現しようとする検索システムである．検索語の入力は自然言語入力でベクトルに変換され，検索結果は登録テキストの入力検索語に対する一致度(類似度)順のリストとして出力される．このモデルでは，キーワードの重要度をどのように設定するかがポイントになる．代表的な設定法は，あるテキストにおけるキーワードの重要度を，そのキーワードがテキスト中に出現する頻度と全テキストにおける頻度の逆数とする方法である．

このモデルの問題点は，精度や効率の面を考慮してベクトル表現に用いる基底単語（概念）の次元や単語等を最適に選択することの難しさにある．また，テキストごとに単語の重要度を保持しなければならないため，索引ファイルが大きくなる欠点がある．

3．本データベース検索システムの基本構造

本研究で提案するデータベース検索システムは，データベース構築者とデータベース利用者との間に共通の背景知識が存在するようなデータベースに対応する検索システムとして構築する．本検索システムはＴＬＯ（技術移転機関）等におけるリエゾン支援システムとしての利用を想定している．

3-1 基本コンセプト

本システムは，次の基本コンセプトに基づき構築されている．

①高い検索精度を有すること．
②利便性の高いシステムであること．
③システム構築や維持のコストが低いこと．
④利用が簡単であること．
⑤システムの更新等が自動化できること．

これらの基本コンセプトに基づいて本検索システムのアウトラインを策定する．

前節の検索モデルの特徴を考慮して，これらの基本コンセプトを満足するシステムを構築していく．まず①および②より，本システムの基本構造として，一致度のような付加情報を持つ拡張キーワード検索モデルを基本として採用する．そしてそれぞれの登録情報ごとにモジュール構造として構築する．

また，通常ではデータベースの作成には莫大な時間と労力を必要とするが，本システムでは，データベース作成に，ほとんど労力は必要とならない．それは本システムでは，データベースにおける情報の蓄積は基本的には情報登録者が行う．そして情報登録者が行う作業も，情報の提供とそれに付与するキーワードおよびキーワードの置換可能語の登録だけとする．このようにすることで，システム構築や維持のコストを低く抑えることができる．また，データベース管理者に専門的知識を特に必要としないシステムとすることができる．こうしてコンセプト③および④は満足される．

情報登録者の拡大や，情報量の増加や更新があっ

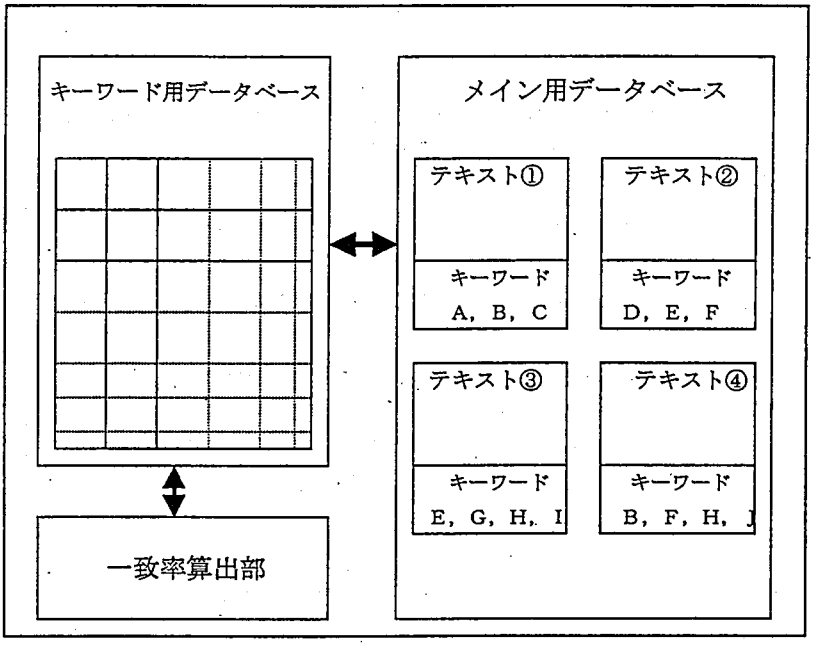

図１　本検索システムの基本構造

ても，モジュール構造であるため追加や変更などの更新が容易である．また，自動更新システムとして構成し，コンセプト⑤の要求を満たす設計とする．

３－２　基本構造

本データベース検索システムの基本構造を図１に示す．本システムの基本構造は，主要部分として，メイン用データベースと，キーワード用データベースと，キーワード用データベースを基に一致率を算出する一致率算出部と，の３つを有して構成されている．

メイン用データベースは，図１に示すように，情報登録者が登録した情報とそれに付与された複数のキーワードとからなるモジュール構造である．情報登録者は，情報本体と，その内容を的確に表す数個のキーワードと，そのキーワードごとに対応させた置換可能語があれば併せてキーワードとともに登録する．このように，登録情報は，情報内容ごとのモジュール構造となってメイン用データベースに蓄積される．

キーワード用データベースは，情報登録者が登録した情報に付与されたキーワードとそのキーワードと置換可能語を抽出して，配置変更して再構成した後，データベースが作成される．キーワード用データベースの構造を図２に示す．図２に示すように，キーワード用データベースは，行列形式の構造を持っている．まず，メイン用データベースに格納された内容のうち，キーワードと，それに付与された置換可能語を抽出する．そして抽出した語をソートして，行と列に配列する．次に行をキーワード，列をそのキーワードに対応する置換可能語とみなし，キーワードに付与された置換可能語の個数をカウントする．多数の情報登録があるため，当然キーワードが重複する場合が生じるが，その場合は重複した頻度をキーワードの情報として記録しておく．行のキーワードと，列の置換可能語とのマトリックスの交点の値は置換可能として登録された数（以後，置換頻度数という）となっている．

一致率算出部は，情報検索者が検索語として入力したキーワードに対し，データベースに登録された情報に付与されたキーワードとの一致の度合いを算出するもので，情報登録者がキーワードに対する置換可能語として登録した語句も考慮して算出する．算出された一致率は，キーワード用データベース上に併せて登録される．そのため，利用者が検索を行った場合には，即座に対応するキーワード同士の一致率を取り出すことができ，トータルの検索一致率の算出に利用される．

４．一致率の算出方法

データベースの登録状態が表１に示すようになっている場合の，一致率算出方法について述べる．なお，一致率の計算では，キーワードと置換可能語と

表1 キーワード用データベースのマトリックス配列

		置換可能語												
キーワード	重複数	A	B	C	D	E	F	G	H	I	J	K	・	・
A	4		2	1	3	1	1	1	2		1			
B	2	1		1	1			2				1		
C	1				1	1				1				
D	2			1		2		1			1			
E	1						1		1	1				
F	1				1					1	1	1		
G	2		1	1			1	1				1		
H	1				1				1	1				
I	1					1			1			1		
・														
・														
・														
・														

は，相互参照の関係にあるものとして取り扱う．

4－1 一致率の考え方

キーワードAとキーワードBとの一致度は，次のように算出される．

表1において，Aの行とBの列との交点の値を見ると置換頻度数は2，このときのキーワードAの重複度は4である．また，置換可能語としてAを見ると，キーワードBとの交点の置換頻度数は1であり，そのときの重複度は2である．キーワードAとキーワードBとの相互の一致率をα（A：B）で表すと，

α（A：B）＝0．5(2/4＋1/2)＝0．5

として計算され，キーワードAとキーワードBとの一致率は５０％である．

また，α（A：C）は同様に，

α（A：C）＝0．5（1/4＋0/1）＝0.125

と計算され，一致率は１２．５％である．一致率算出部は，このような計算のもと，すべての場合の一致率を算出する．

例えば，検索語としてAを入力した場合に，キーワードとして，Aが無くBとCが付与されている情報に対しては，一致率はそれぞれの一致率の和，６２．５％となり，この情報を一致率６２．５％の情報として検索されることになる．

4－2 一致率の一般的算出方法

キーワード用データベースが表1のように表されている場合，キーワードiとキーワードjとの置換頻度数を，行列 M（i, j）と表す．また，キーワードiの重複度を，ベクトル m(i)と表すと，

一致率(i：j)は，

$$\alpha(i:j)=\frac{1}{2}\left(\frac{M(i,j)}{m(i)}+\frac{M(j,i)}{m(j)}\right)\cdots \quad (1)$$

と，表される．

4－3 一致率表の作成

キーワードの一致率の計算例として，キーワード用データベースの頻度表が表2の(a)の場合について，以下のように計算される．

キーワードに対する置換可能と判断する頻度の割合を，すべてのキーワードについて計算すると，表2の(b)のように計算される．本システムではキーワードと置換可能語は相互参照の関係にあるとして，一致率の計算においては区別していない．そこで，式（1）に従って，一致率を計算すると，表2の(c)のようにキーワード間の一致率表が作成される．

4－4 検索一致率の算出

次に，検索時の検索一致率の算出について述べる．ここでは，検索語が複数のｐ個ある場合のキーワードの一致率を求める．

表2　キーワード相互間の一致率計算表

		置換可能語											
キーワード	重複数	A	B	C	D	E	F	G	H	I	J	K	・
A	5		2	2	1	2	1	2			2	1	
B	3	1		2		1		1			1	2	
C	2	1	1			1					1	1	
D	4	2		1		2	1	1		1	1	2	
E	1		1				1						
F	2	1		1	1				1				
G	1					1			1	1			
H	1							1					
I	3	1	1	1	2	1		1	1		1	1	
J	2		1		1		1			1			
K	1			1		1				1			
・													

(a)

		置換可能語											
キーワード	重複数	A	B	C	D	E	F	G	H	I	J	K	・
A	5	0	0.4	0.4	0.2	0.4	0.2	0.4	0	0	0.4	0.2	
B	3	0.333	0	0.667	0	0.333	0	0.333	0	0	0.333	0.667	
C	2	0.5	0.5	0	0	0.5	0	0	0	0	0.5	0.5	
D	4	0.5	0	0.25	0	0.5	0.25	0.25	0	0.25	0.25	0.5	
E	1	0	1	0	0	0	1	0	0	0	0	0	
F	2	0.5	0	0.5	0.5	0	0	0	0.5	0	0	0	
G	1	0	0	0	0	1	0	0	1	1	0	0	
H	1	0	0	0	0	0	0	1	0	0	0	0	
I	3	0.333	0.333	0.333	0.667	0.333	0	0.333	0.333	0	0.333	0.333	
J	2	0	0.5	0	0.5	0	0.5	0	0	0.5	0	0	
K	1	0	0	1	0	1	0	0	0	1	0	0	
・													

(b)

		置換可能語											
キーワード	重複数	A	B	C	D	E	F	G	H	I	J	K	・
A	5		0.367	0.45	0.35	0.2	0.35	0.2	0	0.167	0.2	0.1	
B	3			0.583	0	0.667	0	0.167	0	0.167	0.417	0.333	
C	2				0.125	0.25	0.25	0	0	0.167	0.25	0.75	
D	4					0.5	0.125	0.125	0.167	0.125	0.625	0.25	
E	1						0.5	0	0.333	0.25	0	0	
F	2							0	0.25	0.25	0	0	
G	1								1	0.667	0	0	
H	1									0.167	0	0	
I	3										0.417	0.667	
J	2											0	
K	1												
・													

(c)

まず，p個の検索語が，キーワード用データベース上の何番目のキーワードにあたるのかが調べられ，対応ベクトルqが作成される．

例えば，3個の検索語で検索する場合，それぞれの検索語が，キーワード用データベース上で，何番目のキーワードに相当するのかが調べられる．
ここでは，次のように表されたとする．

1番目の検索語は，5番目のキーワード
2番目の検索語は，13番目のキーワード
3番目の検索度は，16番目のキーワード

対応ベクトルをq(i)で表すと，

$$q=\begin{bmatrix}5\\13\\16\end{bmatrix} \qquad \cdots\cdots \quad (1)$$

となる．

メイン用データベースの各情報に付与されたキーワードとそれぞれの検索語との一致率が抽出される．この例の場合は，各情報に付与されたキーワードと3つの検索語との間の一致率が抽出され，すべての検索度について合算される．

p個の検索語を使った検索では，検索一致率Eは，次のように求められる．

$$E\{q(i)\}=\sum_{i=1}^{p}\sum_{j=1}^{n}\alpha\{q(i),j\} \quad \cdots \quad (2)$$

となる．

ただし，nはキーワード用データベースでのキーワードの総数，αはキーワードの一致率である．

5．本システムの動作説明

本システムの動作例を，データベースの作成から順を追って説明すると，次のようになる．

5・1 データベースの作成

データベースの作成では．情報登録者が情報内容と，内容を的確に表す数個のキーワード，およびキーワードごとの置換可能語を一括して，メイン用データベースに登録する．するとシステムは，図2(a)に示すように，キーワード用データベースを生成し，登録情報がある場合は自動的に更新される．

例えば，情報登録者①が情報①を登録すると，その情報①に付与されているキーワード，A，B，C，および，キーワードAの置換可能語，D，E，F，キーワードBの置換可能語G，H，キーワードCの置換可能語I，J，Kが抽出され，ソートの後，キーワード用データベースに配列される．

次に，情報登録者②が情報②を登録すると，その情報②に付与されているキーワード，F，I，および，キーワードFの置換可能語，A，K，M，キーワードIの置換可能語N，O，が抽出され，ソートの後，キーワード用データベースに配列される．

同様に，情報登録者が登録作業を行う毎に，自動的に上述の作業が行われる．そして，メイン用データベースに情報が蓄積されると同時に，キーワード用データベースの更新が自動的に行われる．

5・2 一致率の算出

キーワード用データベルが作成され，あるいは更新された時には，キーワード等の相互間の一致率が計算，あるいは再計算される．それらの計算は図2(b)に示すように，キーワード用のデータベース上に併せて記録され，利用者が検索をした際に利用され，時間の短縮に役立っている．

6．おわりに

本データベース検索システムは，情報登録者がキーワード付与や置換可能語を登録する際，この情報登録者の知識を集約する形で利用する．個々の情報登録者にとっては別段，検索に関する知識を提供している認識はほとんど無いと思われるが，キーワードとその置換可能語に対する認識等が情報登録者全体で集約されると，強力な検索ツールとなる．これを利用することで，検索精度の向上や，システムの小規模化や，システム管理者の負担が軽減されるなど大きな効果を有している．それ故，本システムは知識集約型データベース検索システムと称している．

本検索システムは，TLO用のリエゾン支援システムへの適用に適していると考えられ，この方面で広く活用されることを期待する．

参考文献

1)徳永健伸：情報検索と言語処理(言語と計算-5)，東京大学出版会(1999)