投稿論文

リファレーションデータベースによる知的検索の可能性

浅井勇夫*

テーブル型データベースにおける完全マッチの検索には限界がある。これを解決する一つの方法はリンク情報の付与である。ここでは参照文献データから得られるリファレーションを用いる。検索の概念が完全マッチの検索からリンク数の多い文献の検索に変わる。このリファレーション検索システムの導入により，関連の高い文献の検索が可能になる。ウェイト付きの検索結果は関連の高い順に表示される。その上，文献群に対する検索や分析が可能になるので，知識構造が明確になる。リファレーション検索はまさに知的検索とみなせる。

キーワード：参照文献，被参照文献，リファレーション，リファレーションデータベース，APTSコード，関連性測度，リファレーション検索，ランキング，知識構造，知的検索

1．はじめに

文献データベースは検索を目的とする抄録型を中心に発展してきた。抄録型の規模は大きく，その社会的な役割は大きい。この抄録型をとりまく問題として，知的検索，利用者の情報環境の変化，そして全文型の台頭がある。

オンライン情報検索は手がかりになるキーを完全マッチさせる検索と検索結果のブール演算とを組み合わせて行う。しかし，用語にあいまいさが入るために完全マッチの検索は難しい。また，文献数の増加に比例してキーワード数や分類コード数が増加しないため，データベースの規模が大きくなればなるほど検索は難しくなる。これらの問題を解決するために多数の研究がなされているが，まだ標準的な手法は確立していない。そのため専門家に検索を依頼する場合が多い。情報の共有化が進行するなかで少数の専門家が利用するシステムはやがて存続できなくなる。

次に，高性能なパソコンやネットワークなど，利用者の情報環境はこの数年間で大きく変化し，今後も変革し続ける。20年前に開発されたオンライン情報検索システムは情報処理機器を持たない利用者を前提に設計されていた。その後，メインフレームの発展で検索システムは改善しサービスは向上しているが，電子媒体の検索結果を自由に活用できる状況になっていない。

最近，全文型文献データベースが流通するようになり，一次資料に直接アクセスできるようになってきた。全文型といっても，ハードウエアの制約で文献のテキスト部分だけをデータベース化したものが多い。今後，ウインドウズの動くパソコンが普及すれば，文献中の図表や式を含む完全な形の全文型が提供されるようになる。この全文型が普及すると抄録型は大きな影響を受ける。

ここでは知的検索について扱う。抄録型（速報型）データベースに文献間のリンク情報を追加したリファレーションデータベースを提案し，新しい知的検索の可能性について述べる。本論にはいる前に情報間のリンクについて考察する。

2．情報間のリンク

知的データベースを構築するには情報間のリンクを如何に表すかが問題になる。まず，いろいろなデータモデルとそこで扱われる情報間のリンクについて考察する。表1はデータモデルの特徴を簡単にまとめたものである。

表1　各種のデータモデルの特徴

	テーブル型データベース	知識ベース	ハイパーテキスト	リファレーションデータベース
内容	書誌・抄録	知識	情報	文献・情報
性質	客観的	主観的	主観/客観的	主観/客観的
規模	大規模	小規模	小規模	大規模
タイプ	テーブル型	階層型	ネットワーク型	テーブル型
関係	なし	条件付き	あり（複数）	あり（複数）
単位	属性（複数）	1センテンス	1パラグラフ	属性（複数）
計算機処理検索	大型 手続き マッチ	WS 推論 連続リンク	パソコン 手続き 離散リンク	パソコン 手続き 関連/類似性

* あさい いさお　大阪府立大学工学部経営工学科
〒593 大阪府堺市学園町1番1号
Tel. 0722-52-1161
（原稿受領　1993.05.13）

データモデルとして，テーブル型，階層型，ネットワーク型がある。テーブル型は最もシンプルな構造をしており，リレーショナル・データベースとして発展している。これはデータ部分から属性の定義部分を独立させ，選択，射影，結合などの関係演算を可能にしたものである。

オンライン情報検索で利用する抄録型文献データベースはテーブル型であり，文献検索は選択，検索結果のフォーマット出力は射影とみなせる。しかし，検索結果を再利用する結合は考慮されていない。

キーワードや分類コードを共通にもつ文献群は互いに関係が深い。しかし，これは文献とカテゴリーとの関係を示すものであって，文献と文献との関係を示すものではない。

データや情報を直接データベース化して利用するものに，ファクトデータベース，経済統計データベース，知識ベース，そしてハイパーテキストがある。前2者は，テーブル型のデータ構造である。ファクトデータベースは属性の定義が難しく，経済統計データベースは時間軸を含む3次元テーブルである。いずれも文献データベースに類似している。

エキスパートシステムにおける知識ベースは，従来あまり使用しない階層型やネットワーク型のデータ構造を扱う。知識ベースの構築には，知識とその関係の入力が必要である。IF～THEN～ルールで蓄積したデータへのアクセスを推論部と呼ぶが，与えられた条件に応じて関係するリンク先を探索する。関係が複雑になると探索が困難になるのと，すべての組み合わせのデータを入力しないために論理的な整合性は保証されない。専門知識ベース以外に“常識知識ベース”を構築する必要があり，データの作成に非常に負荷のかかるシステムである。

最近，ハイパーテキストが開発されている。これは情報管理用の簡易言語の一種で，情報それ自体をデータベース化する。情報は見出しと内容の2つの属性から構成する。百科辞典における見出しとその解説は代表的な例である。この場合の情報間のリンクは内容の中に含まれるキーワードと他の見出しとを1対1に対応させるものである。この1対1のリンク先を決めるのに知的作業が必要である。そのため小規模な対象にしか適用できない。自動的にリンク先が決まる手法の開発が望まれる。

抄録型の検索問題を解決する一つの方法がリンク情報の付与である。それには二通りの方法がある。一つはキーワード間のリンク情報を付与して検索を支援し，完全マッチの制約を少なくすることである。もう一つは文献の属性として文献間のリンク情報を付与して関連検索を行うことである。ここでは，後者の方法を採用する。

3．参照文献の利用

参照文献は非常に重要な情報源である。文献に関係の深い文献が列挙されており，しかもそのコメントまで記されている。それを手がかりに一次資料を求めるケースが多い。網羅性はないが最も安価で簡単な遡及検索法である。

一般に，文献Aが文献Bを参照すれば，文献Bは参照文献（References）である。文献の末尾にリストされるのは，このBの集合である。このAとBとの関係を逆の視点からみると，文献Bは文献Aに参照されたとみなすことができる。このAの集合が被参照文献（Citations）である。

参照文献は文献間の関係を表わすデータであり，文献属性の一つとみなすことができる。しかし，参照文献をデータベース化した事例は少ない。その理由として，(1)簡単に得られるため価値がない，(2)抄録誌に不採用のため重要でない，(3)データとして信頼性に欠ける，(4)データベース化が難しい，などが考えられる。現在，オンライン検索可能なISI社のSCI（Science Citation Index）は参照文献を入力したデータベースである。

SCIはGarfieldが1955年に提案した引用索引の概念をもとにしている。そして，1961年に発行された613種類の科学技術雑誌に掲載された140万編の参照文献データをコード化し，コンピュータ処理した冊子体を発行したのが始まりである。SCIは引用索引を使った文献検索，文献や雑誌の評価，あるいは引用分析を使った計量的な情報科学現象の解明などに利用されている。

従来，参照文献や被参照文献は別々に扱われていた。SCIは被参照文献が主体であり，現在のSCIの参照文献コードでは参照文献と被参照文献とを統合することはできない。筆者は小規模な参照データを作成して参照文献と被参照文献の特性を研究してきた。その結果，参照文献，被参照文献，そして文献自体を一つにまとめた新しい概念を提案し，それをリファレーション（Referations）と名づけた[1)~7)]。この文献自体を含める考え方は筆者独自のものであり，後に述べるリファレーション検索の際に有効に働く。

図1は1981年発行の文献番号428のリファレーションを示したもので，3編の参照文献と5編の被参照文献があることを示す。参照文献は簡単に得ることができ，年月が経過しても増えることはない。しかし，被参照文献は簡単に得られない。被参照文献は参照文献

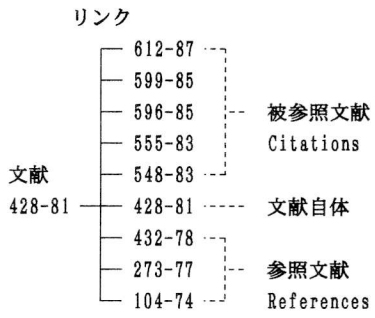

図1　文献428のリファレーション

を多数集め，それを転置することにより得られる。被参照文献は文献がどのように利用され発展してきたかを示す。

リファレーションの概念を記号を使って厳密に定義する。ある分野の文献総数をnとすれば，参照文献行列Rはn行n列の

$$R = \{r_{ij}\}$$

と表すことができる。ここで，

$r_{ij} = 1$　　文献iが文献jを参照する

$r_{ij} = 0$　　文献iが文献jを参照しない。

そして，Rを転置して得られる被参照文献行列Cは

$$C = \{c_{ij}\} = R^T$$

である。これらを使って，リファレーション行列Fを次のように定義する。

$$F = \{f_{ij}\} = C+E+R$$

ここで，Eは単位行列，Fは左右対称行列である。なお，この行列表現は概念を示すものであって，実際のソフト上の処理は別の形をとる。

図1で明らかなように，文献のリファレーションは文献と文献とのリンク情報を示している。われわれはリファレーションを得るだけでも価値がある。さらにリファレーションを使った関連検索をすれば，従来とは異なった知的検索が可能になる。これを述べる前に，参照文献を如何にデータベース化するかについて述べる。

4．リファレーションデータベース

抄録型文献データベースのデータ構造はテーブル型である。このテーブル型だけでは検索に限界がある。これを打開する一つの方法はリンク情報の付与である。テーブルがレコード（文献）と属性（著者，標題，出典など）から構成されている場合，リンク情報はレコード間の関係，すなわち文献間の関係とみなすことができる。前節では文献間のリンクとして参照文献を発展させたリファレーションを用いることを述べた。

リファレーションデータベースの簡単な例を図2に示す。図2のリファレーションファイルは，各文献間の関係を表わす。例えば，文献3は文献1，文献3，そして文献6と関係が深いことを示す。非常にシンプルなデータ構造で，従来型のデータベースにリファレーションファイルを追加しただけである。新しい文献の参照文献が追加されるたびにリファレーションファイルは更新され，文献間の関係は絶えず最新の状態に保たれる。

参照文献の入力は簡単なようで非常に難しい。参照文献の書誌項目を全部入力すれば情報量は増えるが，一方で冗長さが増して文献を同定することが難しくなる。約30年前に開発されたSCIの場合の参照文献コードは著者と出典に関する書誌項目を一定のルールで作成する。一例を示すと，

SMITH LC (LIBRARY TRENDS, V30, P83, 1981)
LAWANI SM (J AM SOC INFORMATION, V34, P50, 1983)。

1つの文献を30～50バイトで表し，逐次刊行物だけでなく，単行本やレポートなどすべての参照文献を入力する。SCIを使って被参照文献をカウントして研究者や

文献ファイル（テーブル型）

No	著　者	表　題	出　典	発行
1	Luhn, H. T.	Keyword-In-Content I	Am. Doc. 11-	1960
2	Small, H. G.	The Structure of Sci	Sci. Stu. 4-	1974
3	Vickery, A.	A Reference and Refe	J. Doc. 43-1	1987
4	Lancastar,	Information Retrieva	John Wiley	1968
5	Garvey, W. D	Research Studies in	Inf. Stor. 8	1977
6	Chubin, D. E	Citation Classics An	J. ASIS, 35-	1984

＋

リファレーションF

No	Ref.
1	1, 3
2	2, 5
3	1, 3, 6
4	1, 4
5	2, 5
6	3, 6

図2　リファレーションデータベースの簡単な例

研究機関を評価する際には第1著者しかないことや冗長なために文献の同定が難しいことを考慮しなければならない。

大規模なリファレーションデータベースの開発には参照文献の入力が必要である。多くの参照文献に点在した文献を同定するための新しい文献同定コードAPTSを開発している[7]。そのねらいは，誰でも，どこででも，簡単なルールで文献をコード化することにある。

文献の基本的な書誌データを使って文献コードを構成するとコード化は容易になる。以下，英文文献の場合のコード化ルールについて述べる。

1）APTSコード（16桁）は著者，発行年，標題，出典から構成する。

2）著者コード（4桁）は姓（前2桁と後1桁）と名（前1桁）から構成する。

Bradford, S.C. → BRDS
Yu, C.T. → YUUC
Spack Jones, K. → SPSK

3）発行年コード（4桁）は発行年を西暦年号（4桁）で表す。

(1934) → 1934

4）標題コード（4桁）は4単語の頭文字（3字以内は除く）で構成する。なお，ハイフンとスラッシュは空白とみなす。

Sources of information on specific subjects
→ SISS

5）出典コード（4桁）は次の3つの場合にわける。

・逐次刊行物は巻（下1桁）と開始頁（下3桁）。
Engineering, 137, 85-86 → 7085

・編集物は記号「=」と開始頁（下3桁）。
In AFIPS Conference Proceedings, 34, 435-446
→ =435

・単行本は発行所や大学名の主要な単語（前4桁）。
Academic Press → ACAD。

以上のルールに従って文献をコード化すると，次のようになる。

Bradfod, S.C. (1934). Sources of information on specific subjects. Engineering, 137, 85-86. → BRDS 1934 SISS 7085。

1時間に約100編の参照文献をコード化できる。1文献当たり約20～30編の参照文献があるとして1時間に約3文献のコード化が可能である。データベース産業は入力産業といわれているが，抄録型データベースの作成費用に数%の費用を追加するだけで参照データが得られ，新しい情報検索システムが構築できる。

最近の傾向として，一つの文献が複数のデータベースに重複して入力されている。書籍や雑誌がコード化されたように，今後は文献を同定するためのコード化が必要になってくる。このAPTSコードはその一つの試みである。

5．文献間のリンク

文献間の関連性を測る方法に書誌結合や共引用がある。書誌結合は参照文献だけのリンクをもとにしているため，新しい文献同士に対して有効である。また共引用は被参照文献だけのリンクをもとにしているため，古い文献同士に対して有効である。ここでは参照文献や被参照文献をリファレーションに拡張する。リファレーションを使うと，文献間の発行年による差がなくなるため，すべての文献に適用できる。

図3は2つの文献227と434のリファレーション間に共通な文献を示したものである。参照文献だけを用いる書誌結合は2編，引用文献だけを用いる共引用は2編であるが，リファレーションを使うと8編になる。直接のリンク（文献227と434）や参照文献と被参照文献とのリンク（文献32と31）が見つかり，単独で用いるよりも情報量は増える。リファレーションの定義に文献自身を含めた理由は直接のリンクをカウントするためであった。

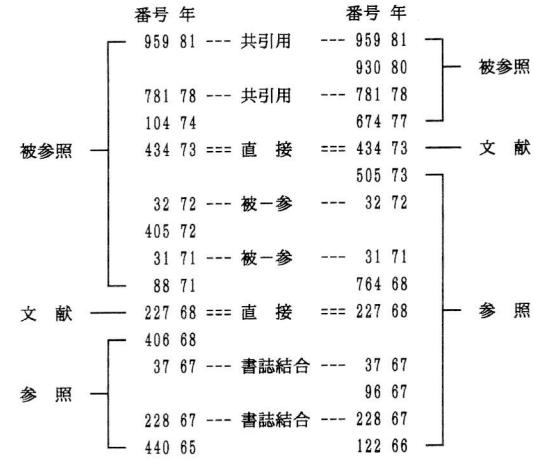

図3　文献227と文献434のリンク数のカウント

関連性の測度としてリンク数だけを用いるものやリファレーションの総数で調整するものなどがある。代表的なものとして，

1）関連度数　$x = c$

2）関連係数　$y = c/(a+b-c)$

ここで，aとbは文献AとBのリファレーション数，cは文献AとBの共通のリンク数である。

文献のリンク数は簡単なアルゴリズムで求めることができる。その処理はデータベースの文献数に比例せず，文献のリファレーション数に比例する。そのためデータベースの規模の大小にかかわらず検索時間はほぼ同じである。

従来の情報検索は多数のレコードの中から検索式にマッチしたものだけを検索した。知識ベースの推論は条件に合うリンクを1つ1つたどりながらリンク先の内容を検索する。これに対してリファレーション検索はリンク数をもとに検索する。

知識ベースやハイパーテキストではリンクを直接使うため，データ収集は厳密さが要求される。一方のリファレーション検索はリンクを間接に使うため，データ収集は前者よりも柔軟である。

6．リファレーション検索による知的検索

リファレーションをベースにした検索の全体の流れ図を図4に示した。新しい検索システムは多様で，検索結果は関連の高い順にリストされ，そして文献群に対する検索が可能になる。

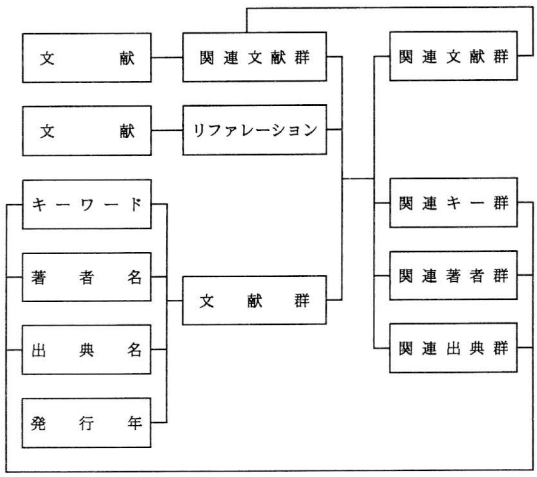

図4　リファレーション検索システム

検索パターンごとに詳しく説明すると，

a）文献から関連文献群の検索

ある文献に関連の高い文献群を検索し，関連の高い順にリストする。文献のリファレーションと異なる文献が見つかる場合がある。

b）文献からリファレーションの検索

ある文献のリファレーションを検索する。発行年順にリストされ，発展の過程がわかる。

c）キーワード，著者名，出典名，発行年から文献群の検索

従来型の完全マッチの検索で，検索結果のウェイトはない。

d）上記a）～c）の文献群から関連文献群，関連キーワード群，関連著者群，関連出典群の検索

複数の文献に対してリファレーション検索を行う。文献以外に，キーワード，著者，出典などを関連の高い順に検索する。

e）検索結果を使ったフィードバック検索

上記a）～d）の検索結果を使って検索する。

リファレーション検索の特徴をまとめると，

1）関連検索：関連の高いものを検索するので，検索キーにマッチしないものも検索する。

2）ランク検索：関連の高い順に検索する。

3）ウェイト表示：検索時のウェイトを表示する。

4）多様な検索：関連の高い文献群，キーワード群，著者群，そして出典群が検索できる。

5）グループ検索：複数の文献，キーワード，著者，出典に関連したものを検索する。

6）知識構造の表現：SCIの被参照文献による引用分析をリファレーションに応用すれば，専門分野の知識構造をクラスターリングやマッピングして表すことができる。

7）最新の関係：各文献のリファレーションは更新され，最新の関係状態で検索や分析を行う。

8）主観的なデータ不要：キーワードのリンクやウェイトなどの主観的なデータを必要としない。

リファレーション検索を利用すると，従来の完全マッチ検索における多くの問題は解決する。データベースを操作して，何か新しい関係を導き出すことができる。知的検索の定義は難しいが，リファレーション検索は知的検索と言える。

7．おわりに

参照文献データを利用したリファレーションデータベースを構築すると，新しい知的検索の世界が出現することを示した。このような情報検索システムが実現すると，研究の効率は促進する。特に，ランク情報や知識構造は中級クラスの研究者のレベルに匹敵する。

抄録型の検索は難しいが，冗長さが増える全文型はもっと難しくなる。その点，リファレーション検索は検索問題の多くを解決するだけでなく，知識構造まで抽出する。まさに21世紀を指向したデータベースとみなせる。

謝　辞

1992年度情報科学技術協会の「研究発表賞」対象文献を最新のものに書き直した。このような萌芽的な研

究を評価していただいた情報科学技術協会，ならびに表彰選考委員の方々に感謝いたします。

参 照 文 献

1）浅井勇夫．パソコンによる引用文献データベースの開発．第21回情報科学技術研究集会発表論文集．Vol.21，p.21-31（1985）
2）浅井勇夫．Referationを用いた特定専門分野情報の分析．第22回情報科学技術研究集会発表論文集．Vol.22，p.135-142（1986）
3）浅井勇夫．インテリジェント・リサーチのためのリファレーションDBの開発．日経コンピュータ．1987-2-16号，p.137-149（1987）
4）Asai, I. "Referation" Database for Document Information Analysis. Proc. the ASIS Annual Meeting. Vol. 24, p. 1-5 (1987)
5）Asai, I. Referation Search : A New Method of Intelligent Information Retrieval. ASIS Mid-Year Meeting. Abstracts. p. 33 (1988)
6）浅井勇夫．リファレーション型データベースによる知的検索の可能性．第19回ドクメンテーション・シンポジウム予稿集．p.36-40（1989）
7）浅井勇夫．大規模なリファレーションDBの開発：文献同定のためのAPTSコードの導入．第29回情報科学技術研究集会発表論文集．Vol.29，p.273-278（1993）

Possibility of intelligent search using referation database, Isao ASAI (University of Osaka Prefecture)

Abstract : Full matched search in table type database has its limits. One way out of the difficulty is to add linking data between documents. This paper use the referation madeup of references. The concept of information retrieval changes from full matched search to highly linked search. Introduction to referation search system leads to select of highly linked documents. The results with weighted data are listed in linked order. Furthermore, search and analysis to grouped documents is feasible, so the structure of knowledge is revealed. Referation search regard just as intelligent search.

Keyword : references/citations/referations/referation database/APTS code/similarity measure/referation search/ranking/knowledge structure/intelligent search

情報の科学と技術　Vol.43 No.10（1993.10）抜刷

投稿論文

リファレーションデータベースによる知的検索の可能性

浅　井　勇　夫