データベースシステム 90−10
（1992. 9. 11）

自動索引システムと情報検索システムの評価用共通データベースの事例

木本晴夫

ＮＴＴ情報通信網研究所

自動索引システムと情報検索システムのための評価用データベースの事例について述べる。従来はこれらのシステムの評価は、開発元で個々に準備したデータに基づいていた。このために評価のベースがバラバラで、各システム相互の精度比較ができず、客観性に乏しかった。システム開発も多くなされるようになり、研究者や開発者が共通に利用できる、評価用共通データベースの必要性の認識が高まってきた。本稿では、Ｅ．Ａ．Ｆｏｘのレポートに従って、アメリカなどで評価用データベースの種類と概要を紹介し、特にＣＡＣＭの評価用データベースについて、サブベクトル付き文献ＤＢ、オリジナル文献ＤＢ、適合性判定ＤＢ、自然文検索式ＤＢ、ＡＮＤ・ＯＲ検索式ＤＢ、適合性の判定方法、などの例を示す。

TEST COLLECTIONS FOR EVALUATIONS OF AUTOMATIC INDEXING SYSTEMS AND INFORMATION RETRIEVAL SYSTEMS

Haruo Kimoto

NTT Network Information Systems Laboratories

1-2356, Take, Yokosuka-shi, Kanagawa, 238-03 Japan

This paper describes test collections for evaluating automatic indexing systems and information retrieval systems. These systems were evaluated using individual collection made by each developer. So, it was difficult to compare one another. As the number of systems grew large, the standard test collection became much more necessary. Here, test collections in America and in Europe are introduced, such as CACM collection, including the document collection with subvector, the query collection and so on.

１．まえがき

本稿では、自動索引システムと情報検索システムのための評価用データベースの事例について述べる。従来はこれらのシステムの評価は、システムを作成した開発元で個々に準備したデータに基づいておこなわれていた。このために評価のベースがバラバラで、各システム相互の精度比較ができていなかった。最近は自動索引システムや情報検索システムの開発も多くなされるようになり、各システムの実現精度の相互比較や客観的評価のために、研究者や開発者が共通に利用できる、評価用共通データベースが必要であるとの認識が高まってきた。また、評価用のデータベースを作成するためには、データを集めることはもちろん、索引付けや文献検索の適合性の判断はすべて人手にたよらねばならず、時間と労力と費用のかかる仕事であり、このような理由からも、誰もが利用できる信頼性のある評価用データベースが研究開発をスムーズに進めるためにも必要で不可欠である。

一方、情報検索の分野の研究の先進国である、アメリカやイギリスでは、評価用データベースが整備されている。まず、歴史的にはＡＤＩやＣＲＡＮなどの評価用データベースは30年前に既に作成されている。分野としてはコンピュータ科学(ＣＡＣＭデータベース)、情報科学(ＩＳＩデータベース)から雑誌(ＴＩＭＥデータベース)まで主たる分野はカバーされている。また、評価用データベースの数としては、約10個に近い。かつおのおののデータベースを構成する文献数も1,000から3,000で豊富である。最近では、ＤＡＲＰＡのＴＩＰＳＴＡＲプロジェクトにおいてギガバイト単位の評価用データベースの構築が計画されている。量的には、少なくとも、数百の文献のデータベースでは評価用としては、不十分だと考えられている。データベースのデータ構造は統一された標準的なものが確立している。そして、研究者や開発者の間で広く、手軽に利用できるように、マグネティックテープやＣＤ－ＲＯＭのメディアで提供されていて、なおかつダウンロード可能で、無料である。アメリカやイギリスで作成されたので、当然のことながら、言語は全部、英語である。

Ｆｏｘはコーネル大学において、これらの評価用データベースについてまとめ、かつＣＡＣＭの評価用データベースとＩＳＩの評価用データベースについて、データの内容を詳しく分析して、報告している[1]。本稿は、日本でのこのような評価用データベースの構築のよりどころとするために、Ｆｏｘのレポートに従って、アメリカ等の評価用データベースについて紹介する。本資料のデータはＦｏｘのレポートに基づいている。

２．　アメリカなどでの事例

２．１　評価用データベースの種類

本節ではアメリカなどでの評価用データベースの種類を紹介する。それらを表１に示す。

表1　評価用文書データベースの種類と概要

名称	主題分野	文献数	1文献あたりのｷｰﾜｰﾄﾞ数	ｻﾌﾞﾍﾞｸﾄﾙの種類	期間
ＣＡＣＭ	ｺﾝﾋﾟｭｰﾀ科学	3204	40.1	au, bi, bc, cc, cr, ln, tm	1958-1979
ＩＳＩ	情報科学	1460	104.9	au, cc, tm	1969-1977
ＡＤＩ	ﾄﾞｷｭﾒﾝﾃｰｼｮﾝ	82	27.1	tm	1963
ＭＥＤ	医学	1033	55.8	tm	-1969
その他：ＣＲＡＮ(航空機工学)、ＬＩＳＡ(情報科学)、ＲＩＲＤ、ＴＩＭＥ(雑誌ＴＩＭＥ)などが有る。					

【注】　au：著者サブベクトル　　cc：共引用サブベクトル
bi：書誌サブベクトル　　cr：分類サブベクトル
bc：書誌結合サブベクトル　　ln：リンクサブベクトル

2．2　検索式と適合文書の特性

代表的な4種類の評価用データベースでの、検索式とその個々の検索式に対する適合文書数に関する統計を表2に示す。

表2　検索式とその適合文書数の統計

評価用データベースの名称	検索式の数	検索式の平均長さ[注]	ひとつの検索式に対する適合文書数		検索結果の上位10位の中での適合文書数	
			平均の数	対全文献比率の平均	平均の数	対10文献比率の平均
ＣＡＣＭ	５２	１１．４	１５．３	０．５	１．９	１９
ＩＮＳＰＥＣ	７７	１６．０	３３．０	０．３	２．０	９
ＩＳＩ	３５(76)	８．１	４９．８	３．４	１．７	４
Ｍｅｄｌａｒｓ	３０	１０．４	２３．２	２．３	３．８	１８

[注]Ｃｏｓ検索式の平均長

2．3　評価用検索式データベースの種類

評価用検索式データベースの種類とその概要を表3に示す。

表3　　評価用検索式データベースの種類と概要

名称	自然文検索式		AND・OR検索式	
	数	作成者または起源	数	作成者または起源
1．CACM	52	コーネル大学	52	コーネル大学の大学院生
2．ISI	76	ADI, ISPRA, SIGIR Forumの抄録	35	ADIに同じ
3．ADI	35	ハーバード大計算機学科学生	35	ハーバード大計算機学科学生と図書館員
4．MED	30	NLMファイル	30	NLMのサーチャ

3．評価用データベースを構成する概念

3．1　拡張ベクトルモデルと多重概念型

拡張ベクトルモデル(Extended Vectors Model)とは、従来のモデルでは文献をタームだけのベクトルで表現するが、このタームサブベクトルのほかに、著者サブベクトル、分類サブベクトル、書誌サブベクトル、書誌結合サブベクトル、共引用サブベクトル、リンクサブベクトル等のその他のサブベクトルをつけ加えて文献を表現するように拡大したものである。おのおののサブベクトルについてまとめて表4に解説する。

拡張ベクトルを導入する利点は以下のとおり。例えば、書誌結合サブベクトルについていえば、それを統計分析して評価用データベース内の書誌結合の数が大きい場合は、いくつかの特定の文献が、そのほかの多くの文献から共通して引用されていることが分かる。例えば、文献集合内に古典的な文献があれば、これにあたる。このように書誌結合特性の分析によって評価用データベースの内容のより詳細な分析が可能となる。この拡張ベクトルモデルは、多重概念型(Multiple Concept Types)[1]の考え方に基づいている。

3．2　サブベクトルの統計

ISIのサブベクトルの長さ統計を表5に、また、CACMのサブベクトルの長さ統計を表6に示す。このような統計を紹介する理由は、これらの統計情報を分析することによって、データベースの性格が明らかになるからである。例えば共引用（cc）サブベクトルについて、ISIの場合とCACMの場合を比較してみると、ISIの場合はCACMに比べてはるかに値が大きい。これは、ISIデータベースの各文献の相関性がかなり高いことを示している。また同じくCACMの場合についていえば値が小さいので、各文献間の相関性が平均して小さいことが分かる。

表4　サブベクトルの解説

サブベクトル名 (英語略称、英語名)	解説
著者サブベクトル (au:author)	著者名のベクトル。
書誌サブベクトル (bi)	掲載雑誌名、掲載年月等。
書誌結合サブベクトル (bc:bibliographic coupling)	ある文献を、本文献とともに引用している別の文献がある場合に、その別の文献のベクトル。
共引用サブベクトル (cc:co-citation)	本文献を別の文献とともに引用している文献がある場合、その引用側の論文のベクトル。
分類サブベクトル (cr:computing reviews categories)	分類コードのベクトル。
リンクサブベクトル (ln:links)	参照(reference)または引用されている文献のベクトル。

表5　ＩＳＩのサブベクトルの長さ統計

統計項目	サブベクトル		
	ａｕ	ｃｃ	ｔｍ
平均	１．４	５４．０	４９．６
メディアン	1	４０	４７
最小	1	1	8
最大	7	２７６	１７９
標準偏差	０．８	４６．４	２１．５
概念の総数	１２５５	１４６０	７３９２

表6　　ＣＡＣＭのサブベクトルの長さ統計

統計項目	サブベクトル					
	ａｕ	ｂｃ	ｃｃ	ｃｒ	ｌｎ	ｔｍ
平均	１．３	４．２	３．７	１．２	２．７	２５．０
メディアン	1	0	0	0	2	１５
最小	1	0	0	0	1	1
最大	7	１８３	１１１	２８	７４	１６８
標準偏差	０．７	１０．８	１０．７	１．９	３．１	２２．７
概念の総数	2647	3204	3204	200	3204	10446

４．評価用データベースの内容例

ＣＡＣＭ評価用データベースの内容例を図１から図３に示す。それらは、サブベクトル付き文献データベース(図１)、オリジナル文献データベース(図２)、適合性判定データベース(図３)の各内容例である。

なお、ＣＡＣＭの自然文検索式データベース、ＡＮＤ・ＯＲ検索式データベース(作成者Ａ、作成者Ｂ)は第５節にて紹介する。

５．ＣＡＣＭの評価用データベースについて

５．１　自然文検索式とそのＡＮＤ・ＯＲ検索式

Ｆｏｘらは1982年にコーネル大学で、かなり大きな検索式の集合とそれらの検索式を使ってＣＡＣＭの文献データベースを検索したときの適合性判定のデータベースを構築した。この目的のために、コーネル大学の教職員と学生をはじめとして、他のアメリカの計算機学科のメンバが、計算機科学の分野での何らかの興味を示す自然文検索式を提出するように求められた。そして、それらをもとにしてＦｏｘらは52の検索式とその適合性判定データを構築することができた。自然文検索式のサンプルのいくつかを図４に示す。

```
.I 3096
.T
An Optimal Method for Deletion in One-Sided Height-Balanced Trees
.W
A one-sided height-balanced tree is a binary tree in which every node's
right subtree has a height which is equal to or exactly one greater than
the height of its left subtree. It has an advantage over the more gener-
al AVL tree  in that only one bit of balancing information  is required
(two bits are required for the ACL tree).  It is shown that deletion of
an arbitrary node of such a tree can be accomplished in O(logn) operati-
ons, .....
.B                                  .C
CACM June, 1978                     3.73 3.74 4.34 5.25 5.31
.A                                  .N
Zweben, S.H.                        CA780601 DH February 26, 1979 12:48 PM
McDonald, M.A.                      .K
                                    Balanced, binary, search, trees
.X
2839 4 3096          3096 4 3096          3009 5 3096
3009 4 3096          3163 4 3096          3065 5 3096
3042 4 3096          3163 4 3096          3096 5 3096
3042 4 3096          3163 4 3096          3096 5 3096
3065 4 3096          2839 5 3096          3096 5 3096
3096 4 3096          2889 5 3096          3163 5 3096
3096 4 3096
```

図１　　ＣＡＣＭのサブベクトル付き文献データベース

```
title--An Optimal Method for Deletion in One-Sided Height-Balanced Trees
abstract--A one-sided height-balanced tree is a binary tree in which ev-
ery node' s right subtree has a height which is equal to or exactly one
greater than the height of its left subtree.  It has  an advantage over
the more general AVL tree in that only one bit of balancing information
is required (two bits are required for the ACL tree).  It is shown that
deletion of  an arbitrary node of  such a tree  can  be accopmlished in
O(logn) operations, where n is the number of nodes in the tree.  Moreov-
er .....
journal--CACM June, 1978
author--Zweben, S.H.; McDonald, M.A.
keys--Balanced, binary, search, trees
categories--3.73, 3.74, 4.34, 5.25, 5.31
end--CA780601 DH February 26, 1979 12:48 PM
```

図２　　ＣＡＣＭのオリジナル文献データベース

```
1    1410 0 0.0          3    2290 0 0.0          4    3127 0 0.0
1    1572 0 0.0          3    2923 0 0.0          4    3128 0 0.0
1    1605 0 0.0          4    1749 0 0.0          5     756 0 0.0
1    2020 0 0.0          4    1811 0 0.0          5    1307 0 0.0
1    2358 0 0.0          4    2256 0 0.0          5    1502 0 0.0
2    2434 0 0.0          4    2371 0 0.0          5    2035 0 0.0
2    2863 0 0.0          4    2597 0 0.0          5    2299 0 0.0
2    3078 0 0.0          4    2796 0 0.0          5    2399 0 0.0
3    1134 0 0.0          4    2912 0 0.0          5    2501 0 0.0
3    1613 0 0.0          4    3043 0 0.0          5    2820 0 0.0
3    1807 0 0.0          4    3073 0 0.0
3    1947 0 0.0          4    3082 0 0.0
```

図３　　ＣＡＣＭの適合性判定データベース

```
(1)What articles exist which deal with TSS (Time Sharing System), an op-
   erating system for IBM computers?
   .N   1. Richard Alexander, Comp Serv, Langmuir Lab (TSS)
(5)I'd like papers on design and implementation of editing interfaces,
   window-managers, command interpreters, etc.  The essential issues are
   human interface design, with views on improvements to user efficiency,
   effectiveness and satisfaction.
   .N   5. Pavel Curtis (editing interfaces)
```

図４　　自然言語検索式の例(ＣＡＣＭ)

また、Ｆｏｘらはこれらの自然文検索式に対応するＡＮＤ・ＯＲ検索式を作成した。図４の例に対応するＡＮＤ・ＯＲ検索式の例を図５に示す。

```
Searcher 1:
#q1 = #or ('tss', #and('ibm', #and('time','sharing')));
#q5 = #and ('editing', #and(#or('human','user'),
                      #or('satisfaction',
                        'efficiency')));


Searcher 2:
#q1 = #or (#and('ibm','tss'),
       #and('ibm','time','sharing','system'));
#q5 = #and( #or('design','implementation'),
       #or( 'human','satisfaction','user'));
```

図５　　ＡＮＤ・ＯＲ検索式の例(ＣＡＣＭ)

５．２　適合性の判定方法

52個の検索式があって、それらの3204個の文献に対する適合性判定データを得るのは膨大な仕事である。通常は検索式を作成したユーザは、全文献データに対する適合性判定は、その作業の膨大さの理由から、これをしたがらない。一方、ユーザによる適合性の判定は、第３者によるものと比べて非常に重要である。そこで、Ｆｏｘらは折衷案として、第１段階として、検索システムを使用して適合していそうな文献を洗い出して、次に第２段階として、それらを検索式の作成者に示して、最終的な適合性判定をしてもらう方式をとった。

まず、30個の文献の検索をベクトル検索方式でおこなった。つまり、すべての文献の自動索引をおこなう。次に、文献と検索式のキーワードのベクトルの類似度を計算して類似度の大きいものから30個の文献を選択した。

次にこの30個とは別に、70個の文献を検索して選ぶ。これらの70個の文献は以下に示す、７種類の検索の結果を次の(1)、(2)の方法でマージして得た。

(1)７種類の検索の、おのおのの、上位７個の文献は無条件に選ばれる。

(2)次に７種類の検索結果をすべて集めて、その中で検索された頻度の上位の文献が選ばれる。

この７種類の検索相互の間では、検索される文献の重複がほとんど無いと考えられる。それは、異なった形式の検索式による検索結果の間では、重複した検索結果が少ないという考え方に基づいている。７種類の検索は以下のとおり。

(1)ＡＮＤ・ＯＲ検索式(２人の学生のうちの１人が作成したもの)

(2)ＡＮＤ・ＯＲ検索式(２人の学生のうちの別の１人が作成したもの)

(3)フィードバック検索式(タームサブベクトルを利用)

(4)フィードバック検索式
(タームサブベクトルとリンクサブベクトルを利用、
それらの比は、１：４)

(5)フィードバック検索式
(タームサブベクトル、書誌結合サブベクトルと共引用サブベクトルを利用、それらの比は、１：３：３)

(6)フィードバック検索式
(タームサブベクトル、著者サブベクトルと分類サブベクトルを利用、
それらの比は、１：３：３)

(7)フィードバック検索式
(タームサブベクトル、著者サブベクトル、書誌結合サブベクトル、共引用サブベクトル、分類サブベクトルとリンクサブベクトルを利用、
それらの比は、すべて同等)

以上の30個の文献と70個の文献を併せた100個の文献をユーザに提示して、適合性の判定をしてもらって適合性のデータを得た。

５．３　ＣＡＣＭの評価用データベースの内容の詳細分析

ＣＡＣＭの評価用データベースの内容の詳細分析の項目を以下に示す。分析結果については、文献[1]を参照されたい。これらの項目とその分析は評価用データベースの内容を明らかにするのに有効である。

詳細分析の項目

①年ごとのＣＡＣＭ発行文献数
②年ごとの１文献あたりの平均引用文献数
③年ごとの１文献あたりの平均書誌結合数
④年ごとの１文献あたりの平均リンク数
⑤年ごとの１文献あたりの平均共引用数
⑥ＣＡＣＭの分類サブベクトルの長さの頻度分布
⑦ＣＡＣＭのリンクサブベクトルの長さの頻度分布
⑧ＣＡＣＭの書誌結合サブベクトルの長さの頻度分布
⑨ＣＡＣＭの共引用サブベクトルの長さの頻度分布

６．まとめ

Ｆｏｘ[1]のレポートに基づいて、アメリカなどでの、自動索引システムと情報検索システムのための評価用データベースについて紹介した。今後は、日本でも自動索引や情報検索の研究開発の発展のために、評価用データベースの構築とその共有化が望まれる。

参考文献

[1]Fox, E.A. :"Characterization of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts", Technical Report 83-561, Cornell University Department of Computer Science, Ithaca, New York, Sep. 1983.