APPLICATION OF VECTOR BASED INFORMATION ACCESS IN BIBLIOGRAPHICAL DATABASE

R K Joteen Singh

Abstract

Developing a bibliographical database is not an easy task. While developing it the designer must think about exchange of record, minimum consumption of memory, effective retrieval of the database content, etc. With these concepts in mine, the bibliographical database should be designed and provided mandatory attributes for capturing data. Most of the Information Retrieval Systems are based on keywords of the document and retrieve a list of documents in response to a user's query. Then the user has to sequentially determine his/her relevant document. A keyword-based retrieval is a simple model that can encounter two major difficulties such as synonymy and polosemy problem. This paper introduces the application of vector based information retrieval technique in bibliographical database for effective retrieval.

Keywords: Data mining; Data warehouse; Document-term frequency matrix; Cosine distance

1. Introduction

The nations of the world are moving towards the information society. 'At every moment of our personal and professional life we are information-dependent. As the consumption of information has increased, so also has the rate of generation of information increased; this state is reflected by the term information explosion'. The major issues of information explosion are the storage of information and effective retrieval. On one side, advancement in computer technology, particularly in terms of the memory has made the problem of storage easier to handle. With the growth in the amount of electronic information, more and more documents are available on a given subject. As a result, more and more documents match the user's queries and the lists of retrieved documents more and more time consuming to be processed by the users, again the question of relevant and irrelevant of the retrieved documents. A user who is using a text retrieval system wants to retrieve documents that are relevant to his or her needs in terms of semantic content. Most bibliographical information retrieval systems support keyword-based retrieval. In keyword-based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A keyword-based retrieval is a simple model that can encounter two major difficulties such as synonymy problem and polysemy problem.

In general, bibliographical databases are based on a standard format say CCF, MARC-21, and so on. They are semi-structured in nature. However, there are some highly structured bibliographical databases, which does not include the field 'abstract/description' (tag 600 of CCF). In such environment similarity based retrieval is not effective. To make similarity-based retrieval more effective the tag 600 of CCF (table 1) is very important and mandatory field while capturing data. Some other fields are not so important from retrieval point of view. Hence, the bibliographical data from different sources are required to be processed before applying retrieval technique, thus, the idea of data warehouse becomes prominent and increasingly popular. Which is explained in section 3. Section 2 explained the bibliographical record format. Section 4 devotes to the basic idea of data mining versus normal query tools. In this paper, similar documents of a bibliographical database are grouped together by measuring the distances of term. The nucleus idea of the present paper is to explore the hidden relevant documents from a large volume of bibliographical database by constructing a 'document-term frequency matrix'. The general framework is based on semi-structured database, data warehouse, document warehouse, data mining and document-term matrix. Section 5 relates clustering of similar documents using cosine distance. The last section concludes the paper.

2. Bibliographic Record Format

The term bibliographic record refers to the collection of related attributes of a document. Gredley and Hopkinson defined a bibliographic and a bibliographic item as follows:

- **Bibliographic Record**

  A collection of data elements, organized in a logical way, which represent a bibliographic item.

- **Bibliographic Item**

  Any document, book, publication or other record of human communication; any group of documents, or part of a document, treated as an entity.

The creation of bibliographic record facilitates search and retrieval, locally as well as through electronic networks, and exchange of bibliographic information among libraries/ information centres calls for standard formats governing the process of record creation and exchange. Bibliographic formats have been created for this purpose.

The common communication format (CCF) was developed in order to facilitate the exchange of bibliographic data between organizations. The data elements prescribed in CCF for recording bibliographic and factual information in databases are presented in table 1.

001	Record identifier	088	Record to record linking
010	Record identifier for secondary segments	100	International Standard Book Number (ISBN)
011	Alternative record identifier	101	International Standard Serial Number (ISSN)
015	Bibliographic level of secondary segment	102	CODEN (for serials)
020	Source of record	110	National bibliography number
021	Completeness of record	111	Legal deposit number
022	Date entered on file	120	Document number
023	Date and number of record version	125	Project number
030	Character sets used in record	130	Contact number

031	Language and script of record	200	Title
040	Language of item/entity	201	Key title
041	Language and script of summary	210	Parallel title
050	Physical medium	230	Other title
060	Type of material	240	Uniform title
061	Type of parent document	260	Edition statement
062	Type of factual information	300	Name of person
063	Type of standard	310	Name of corporate body
080	Segment linking field: vertical relation	320	Name of meeting
085	Segment linking field: horizontal relation	330	Affiliation
086	Field to field linking	340	Countries associated with parent
400	Place of publication and publisher	510	Note on related items/entities
410	Place of manufacture and manufacturer	520	Serial frequency note
420	Place of distribution and distributor	530	Contents note
430	Address	600	Abstract/description
440	Date of publication	610	Classification scheme notation
441	Date of legal deposit	620	Subject descriptor
442	Dates related to patent	650	Services provided
444	Dates related to standard	700	Human resources
446	Dates related to thesis	705	Equipment and other resources
448	Start and end dates	710	Financial resources
450	Serial numbering and date	715	Income components
460	Physical description	716	Expenditure components
465	Price and binding	800	Nationality
470	Mathematical data for cartographic material	810	Educational qualification
480	Series statement	820	Experience of person
490	Part statement	860	Project status
500	Note		

Data stored in most bibliographic databases are semi-structured data in that they are neither completely unstructured nor completely structured. For example, the bibliographical database, based on the CCF is semi-structured, because, it consists of a few structured fields, such as title, authors, publisher and so on, also contain some largely unstructured text components, such as abstract and contents.

3. Data Warehouse

Data warehouse is a repository of data from multiple sources. For example, INFLIBNET is a national network with branches across the country. Each branch has its own megabytes of bibliographical record. Any client may put queries to provide relevant document on a particular topic. This is a difficult task, particularly the relevant data/ information are spread out over several databases, physically located at numerous sites.

Since, INFLIBNET has a data warehouse, this task is very easy. A data warehouse is a repository of information collected from multiple sources, stored under a unified scheme, and which usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, and data reduction. In the case document also the above process is necessary. An overview of the flow of document warehouse is shown in figure 1.

Data cleaning: the process of removing noise, fill in missing values and correct data inconsistencies.

Data integration: the process of combining data from multiple sources to form a coherent data store. Metadata, correlation analysis and data conflict detection contribute towards smooth data integration.

Data transformation: the transformation or conversion of the data into appropriate forms for mining.

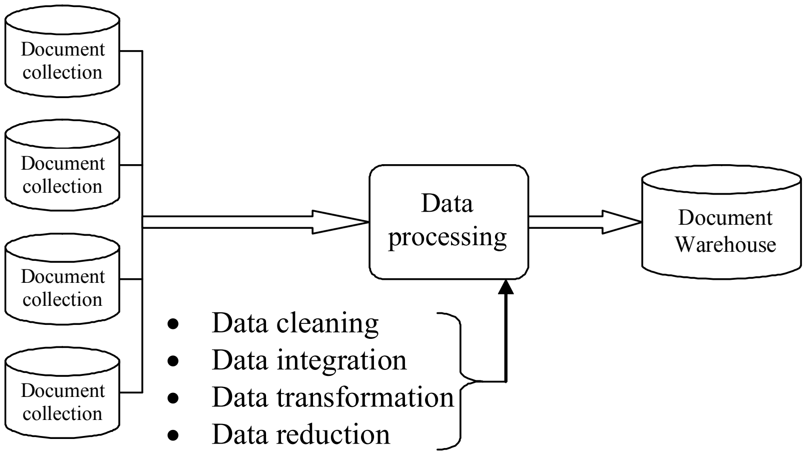

Figure 1: Overview of the document warehouse

Data reduction: the reduction of data using varieties of techniques such as dimension reduction, data compression, singular value decomposition [10] can be used to obtain a reduced representation of the data, with minimum lost of information content.

4. Data Mining

'Data mining is the process of discovering meaningful patterns and relationships that lie hidden within very large databases'. It is a technique for exploring meaningful and relevant information from a large volume of data. The area of data mining is developing very fast with a diverse application.

One of the recent development is its application in information retrieval. Traditional information retrieval techniques become inadequate for the increasing vast amount of textual data. Typically, only a small fraction of the many available documents will be relevant to a given individual or user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the terabytes of data. Hence, an efficient tool is required to compare different documents, categorized them according to their similarity. Thus, data mining has become an increasingly popular and essential.

4.1 Data Mining versus Query Tools

A data mining tool does not replace a query tool, however, it does give the user a lot of additional possibilities. Suppose that in a large file containing millions of records that describe documents consist of a number of highly relevant documents to a user's query. Most of which can be hit by firing normal queries at the database such as 'list of documents for a particular author', 'list of documents related with a keyword relativity' and so on. However, knowledge hidden in the database that is much harder to find using normal SQL. A good information retrieval system should consider synonyms when answering such queries. For example, given the keyword Internet, synonyms such as network and www should be considered in the search as well. SQL based retrieval is a model that can encounter two major difficulties such as synonymy problem and polysemy problem.

- **Synonymy problem :** A keyword, such as navigate may not appear anywhere in the document, even though the document is closely related to navigate.
- **Polysemy problem :** the same keyword, such as reaction, may mean different things in different context.

The above two problems makes disadvantage on normal SQL in certain areas. In addition to this, normal SQL does not have the idea of similarity and degree of relevance between the query and document. Hence, a model is developed to extract similar documents to a query from a bibliographical database where the abstract field is mandatory in the next section.

5. Document Vector Model

It can be assumed that a set of term qj for retrieval, $1 \leq j \leq T$, has been defined a priori. Each individual document Pi, $1 \leq j \leq$, is then represented as a term vector.

$$Pi = (pi1, pi2, pi3, \dots \dots \dots piT)$$

Where pij represents number of occurrence of the jth term in the ith document. The individual pijs are just the component values of the term vector. It is generally defined as pij = 1, if document i contains term j, and pij = 0, otherwise. In the vector space representation each term weight can be some real-valued number, e.g., a function of how often the term appears in the document, or the relative frequency of that term in the overall set of documents.

Table 2: A document-term matrix for 9 documents and 9 terms. Each ijth entry contains the number of times that term j appears in document i.

	q1	q2	q3	q4	q5	q6	q7	q8	q9
p1	20	25	32	2	1	0	0	3	0
p2	42	19	21	0	1	3	0	0	0
p3	19	14	17	4	0	2	0	1	0
p4	1	0	2	40	36	18	2	3	7
p5	3	0	5	24	18	31	0	0	0
p6	5	0	6	27	38	28	1	2	0
p7	0	0	0	3	1	2	17	32	25
p8	2	0	1	2	0	1	18	33	26
p9	0	0	0	3	0	0	45	51	40

Consider the example of table 2 with 9 documents and 9 terms, where the terms are: (q1=gene; q2=hybrid; q3=cell; q4=data; q5=cluster; q6=distribution; q7=schema; q8=normalization; q9=sort) and we got a 10x6 document-term frequency matrix U. Entry ij contains the frequency of the term j that contained in document i. We can see that the first three documents p1 to p3 contain mainly genetics terms (such as gene, hybrid, cell), the next three documents p4 to p6 contain mainly statistical terms (such as data, cluster, distribution) while the last three documents p7 to p9 contain mainly database terms (such as schema, normalization, sort).

Since, similar documents are expected to have similar relative term frequencies, we can measure the similarity among a set of documents or between a document and a query, based on similar relative term occurrences in the frequency table.

In a particular vector-space representation, the distance between documents can be defined as some simple well-defined function of distance measure techniques. One of the widely used distance measure is the cosine distance, defined as follows.

Let Pi and Pj be two term vectors, their cosine distance is defined as

$$d_c(P_i,P_j) =$$

This is the cosine of the angle between the two vectors and thus, reflects similarity in terms of the relative distribution of their term components.

Table 3: A pair-wise cosine distances for the document-term matrix U. Larger cosine values means closer angles (smaller distance) between the two document vectors and smaller cosine value corresponds to larger angles (larger distance).

	Cosine of Vectors of Values								
	q1	q2	q3	q4	q5	q6	q7	q8	q9
p1	1.000	.862	.961	.081	.145	.156	.062	.102	.051
p2	.862	1.000	.955	.060	.155	.164	.004	.056	.000
p3	.961	.955	1.000	.150	.234	.224	.045	.088	.031
p4	.081	.060	.150	1.000	.880	.952	.155	.110	.094
p5	.145	.155	.234	.880	1.000	.938	.096	.055	.024
p6	.156	.164	.224	.952	.938	1.000	.128	.088	.061
p7	.062	.004	.045	.155	.096	.128	1.000	.997	.970
p8	.102	.056	.088	.110	.055	.088	.997	1.000	.973
p9	.051	.000	.031	.094	.024	.061	.970	.973	1.000

This is a similarity matrix

It is clear from the above table 3, there are three clusters of documents such as genetics documents (p1, p2, p3), statistics documents (p4, p5, p6) and database documents (p7, p8, p9). The implication of the larger cosine of vectors of values in a particular cluster (shaded portion) of table 2, means: i) closer in angle and ii) more similar.

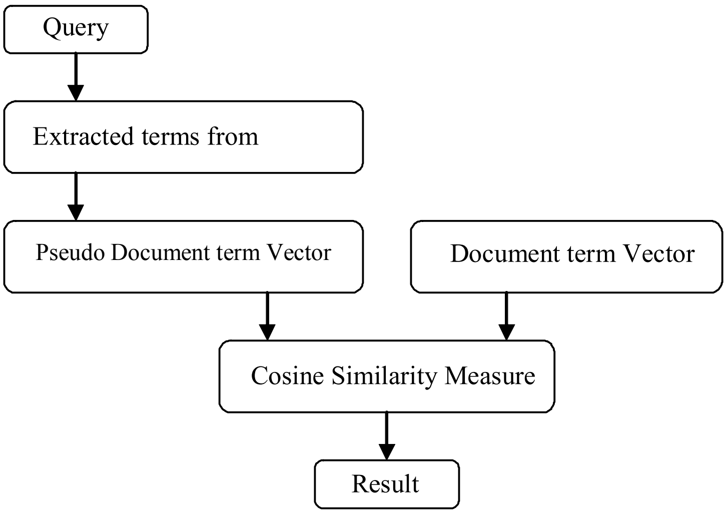

Figure 2: Matching 'pseudo term vector' and 'document term vector'

5.1 Comparing Pseudo term vector and Document term vector

A query can also be treated as a document. This document is known as a pseudo document. The pseudo document will go first through a process [1]. This process comprises of 'ignore list filter parsing', 'stop word filter', 'term extraction', etc. Finally, the pseudo document can be represented by a term-vector, called 'pseudo document term vector'. Once, pseudo document term vector is prepared the similarity between the pseudo document and every other document term vector has to be compared by using cosine distance. The flow of cosine similarity measure is shown above, figure 2.

6. Conclusion

This paper presented a model of information retrieval in bibliographical database to browse relevant documents to a user's query. It is based on document-term frequency matrix which represents the whole set of documents. The similarity between 'Pseudo document term vector' and 'set of document term vector' can be determined by measuring the Cosine distances. A sample of the cosine of vectors of values is also shown in table 3, which corresponds to the document-term frequency matrix, table 2. This model will help to find the top N relevant documents for query.

References

1. http://www-scf.usc.edu/
2. http://www-db.stanford.edu/
3. TARAPANOFF (K)., QUONIAM (L). Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil. Information Research. 7(1), October 2001.
4. http://www.arxiv.org/PS_Cache/CS/pdf/0602/0602076.pdf
5. GOLDMAN ®., WIDOM (J). DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. Proceedings of the 23rd VLDB Conference. Athens, Greece, 1997
6. ONKAMO (P)., TOIVONEN (H). A survey of data mining methods for linkage disequilibrium mapping. Human Genomics. 2(5), March 2006, p336-340.
7. MOTHE (J)., CHRISMENT ©. Information mining: use of the document dimensions to analyse interactively a document set. European Colloquium on Information Retrieval Research, 2001.
8. ANAHORY (S)., MURRAY (D). Data warehousing in the real world. Singapore: Pearson Education, 2003. p19-28.
9. HAN (J)., KAMBER (M). Data mining concepts and techniques. San Francisco: Morgan Kaufmann, 2002. p428-432.
10. HAND (D)., MANNILA (H)., SMYTH (P). Principles of Data Mining. New Delhi: PHI, 2004. p456-465.
11. CHOWDHURY (G.G.). Introduction to modern information retrieval. Great Britain: Facet Publishing, 2004. p38-40.