Workshop on
Information Resource Management
13th-15th March,, 2002
DRTC, Bangalore

Paper: DG

Use of a Knowledge Base of Subject Relationship for Enhancing Retrieval Performance in the Digital Environment

K.S.Chudamani
JRD Tata Memorial Library
Indian Institute of Science
Bangalore –560 012

and

A.Y.Asundi
Dept. of Library and Information Science
Bangalore University
Bangalore – 560 056

Abstract

The current article discusses problems related to subject analysis. The study proposes for creation of a knowledge-based system of subject relationships on the basis of user subject need navigation. Emphasizes on class number as given in Physics Abstract to develop a "toy" knowledge-based system. Demonstrates potency of knowledge-based system of subject relationship to enhance information retrieval efficiency.

1 Introduction

Organisation of knowledge has been a vocation of human being since its birth. Knowledge is being organised since thousands of years since it is being acquired. In this task several tools and techniques have been evolved from Notational Classification to the concept subject analysis, and the latter has become a tool for organising knowledge or information particularly at the tangible level in information systems. In order to retrieve it from this store of organised knowledge, many surrogates like author, title, subject catalogues have been created by libraries. Subject analysis, used in the creation of subject catalogues, is defined by Vickery "as the study of the whole text to select words that indicate those aspects of its information content that are likely to interest the users of the resulting subject index". The indexer is responsible for the Subject Index should be knowledgeable about the extensions and intensions of the subject(s) area(s) he is dealing with, and also its relationship with other subjects. Though the disciplines are defined and ordered for the purpose of organizing knowledge, but in the contemporary era the interdisciplinary nature of subjects is increasing the complexity of defining and ordering knowledge and information. The computers were employed to intervene in retrieving information from huge storage systems and improve the effectiveness. But it has remained static, as far as confined to precision and recall parameters. This can be attributed to limitations in the present retrieval models, which still confine to match the user model with system model with traditional search methods. New statistical methods of analysis have been attempted by TREC experiments. But the results are not found to be satisfactory. All this has been attributed to the lack of semantic inference on both the user and the system. In this regard a suggestion to create knowledge-based systems of subject relationship is made which can be used as such, the user subject need navigation maps. In this paper a proposal for the creation of a knowledge based system of subject relationships is made and its augmentation in retrieval through user subject need navigation is demonstrated with a "toy" model.

2 Mapping subject relationship

A subject is defined as a systematised body of knowledge whose extensions and intensions fall within the comprehensions of an individual. A relationship as defined by the Oxford University Press dictionary is "the state or the fact of being related - a connection, an association". In order to derive these relationships, subject relationship maps are drawn. Mapping, again according to the Oxford English Dictionary is "the action of drawing the map", in the present context; it is the act of associating subjects with one another. For example in a Library a document classified by DDC may not be able to represent the relationships the subject of the documents has with other related subjects as per its content. Whereas using either UDC or CC the Class Number may represent all the facets of the subject in relation with, in a linear order by means of some syntactic methods. DDC will have only one class number for a document of multiple subject relationship whereas, UDC or CC will have a class number that can synthesize subject relationships and display them in a linear order. They are shown as disjuncts in an Abstracting periodical like Physics Abstract. However, some of these aspects are also shown in a Document Description tool like a Cataloguing system, either CCC or AACR and which uses a system of subject headings to represent the semantic relationship among the subjects and their surrogates. In Some cases the subject headings can also be derived with the help of class numbers as done in DDC and CC using Sears' List and Chain Indexing repsectively. Now the Library of Congress Subject Headings is also following this procedure. All the main class numbers assigned have associated isolate of numbers from each main class associated classes and from the associated classes main classes can be identified. Similarly it is the case with INSPEC Physics abstract. The emphasis in this paper is on class number as given in Physics Abstract can be used to navigate hierarchically, aggregates and disaggregates, both the main subject and related subject. Such a map can be drawn for a subject, with its related subjects, and it turns out be a knowledge-base of subject relationships for that subject. A knowledge base, in the general sense is the knowledge of the specialists in a specialised area of study. In the context of information studies, it represents the knowledge of the terminology and their interrelationships. The creation of a knowledge-base of subject relationship involves knowledge engineering practices involving a two step process:

1. Identification of the problem domain;
2. Knowledge acquisition

As a first step, a problem domain is identified for creating a knowledge - base. In this paper the authors have "superconductivity" as the area of study or problem domain because of its multifaceted subject relationship. The study of the development of the subject "superconductivity" reveals that the subject has developed over a period of 80 years, as the superconductivity phenomena was observed by Cammerlingh Onnes in 1911, but its actual development as recorded in physics classification scheme of Physics Abstracts starts in 1965.

3 Present study

In order to acquire the knowledge of subject interrelationship, usually experts are consulted. Also, co-citation, co-word, co-classification maps can be used in this context. It is specifically pointed out by Todorov(1989) that Co-citation maps generated at macro level are difficult to validate using traditional quantitative approaches. Same is the case with co-word maps. Hence, co-classification analysis of a subject based on a classification scheme is considered more effective. The source of data for such co-classification is Physics Abstract. The subject chosen is superconductivity. Subject relationship data is collected for the period 1985-93 from the printed abstract. From this, a “toy” knowledge based system for superconductivity subject relationship is developed and is used to demonstrate how Knowledge based system of subject relationship can be used in enhancing retrieval efficiency.

4 Knowledge based system of subject relationship

A knowledge-based system is a computer system, that embodies the Knowledge-base, the representation and the inference engine. Knowledge-base is the knowledge and heuristics of an expert within a specialised domain and the inference engine provides the necessary manipulation mechanism for answering queries.

The knowledge-base of knowledge based system of subject relationship, consists of a subject, its class number, the related subject, its class number. This knowledge-base can be generated using the classification scheme, or a Subject heading list like, the Library of Congress Subject Headings list etc. In this paper the data for knowledge-base of subject relationships for the superconductivity is obtained using the INSPEC - Physics Abstract (Print Version). The data was collected for 1985-1993 by using its classification scheme. A sample of data is given below to show the subject and its subdivisions related in the context of superconductivity.

65 Thermal properties of condensed matter
65.40 Heat capacities of solids
65.50 Thermodynamic properties and entropy
65.70 Thermal expansion and thermo-mechanical effects

Totally a data for about 51 main classes and about 400 subclasses was generated for the period covered under the study. In order to display these subjects to the user , in order to enable them to select a suitable related class and subclass, menus can be devised and used. From the classification scheme of Physics Abstracts to represent main class (Superconductivity) and penumbral classes (Subject relationships) of superconductivity are recorded and identified as core, non-core patterns respectively. This heuristics can be acquired with the help of the expert or by the experienced classifier. It was also generated by a bibliometric analysis of cross references as proposed in a earlier paper by Chudamani and Asundi. In this paper it is reported that such a pattern has been ventured. In this way the menu displays can be handled more intelligently. Once a division or a class has been selected from the core and non-core classes, in the next step only related subclasses are displayed for further selection in order to narrow down the search.

Representation the other component of the Knowledge-based system is the mode of encoding the acquired knowledge in the memory of the computer for further manipulation. The system creates a frame -based representation method and thus

different chunks of knowledge can be related and used for manipulation. Frames can be considered as collection of related facts that can be treated as single units. In one sense a frame can be considered as a record with fields or slots that can be filled in with specific values as presented in the section on knowledge base.

Only one subject has been used for data collection purpose. Hence only related subjects of it have been recorded in the data file. Three elements, namely the main class, the related class, and its name are recorded in a file in Dbase. But when such a file is created for all subjects the main class which is related to other classes should also be recorded in the data file.

Though there are two representations available namely, semantic nets and predicate calculus, which are more specific in application, hence it is felt that context dependent information rather than weighting can help in retrieval.

Inference engine is the most important part of an knowledge based system. It is a program that can fire when certain facts are available, while searching answers for some queries. Conflict resolution is most important. When two rules are likely to fire, it has to judge the rule to fire to seek an answer to the query. It works on the principle that fact1, fact2, and fact6 yield the conclusion 1 and so on. Facts are data and manipulation of this data by the inference engine leads to conclusion which is again data.

5 The present system

From the foregoing discussion it is clear that even to develop a "toy" knowledge based system it becomes essential to go through the steps as enumerated below in creating the system for representing subject relationships.

Step 1: selection of a broad subject from Physics Classification scheme

Step 2: identification of its related subjects

Step 3: establishing related terminology

Step 4: designing a representation for recording relationships

Step 5: development of a program for generating inference from the system

As already described earlier "Superconductivity" has been chosen as the subject for mapping subject relationships. It is highly interdisciplinary and has may associate relationships. It is important to study the subject from the point of view of its applications in many areas. In order to identify its related subjects from the source document as indicated before, a bibliometric analysis of the literature for the period 1985-93 was conducted to obtain data and heuristics. The results of the analysis has been used to develop a menu for use in the system to present subject associations. The type of heuristics used are :

- List subjects for display depending on their degree of relationships to the selected subjects
  (in this case superconductivity related subjects: core and non-core)
- If subject selected is superconductivity and material science then list subdivisions of material science to select related subdivisions. Then a general heuristic would be if subject selected is **x** and related subject selected is **y** display the related subdivisions of **y.**

As required, a data file for retrieval based on terminology data analysis has been created for **'ceramic superconductors'**. It has been created in **dBASE III +**. Abstract number and keywords for 365 articles have been entered in the file. The representation is frame based. The slots have been filled with keywords and class numbers.

In the system, first a list of broad subjects as given in the Physics abstract is presented to the user. From this menu user selects a subject of his choice as a broad subject. In this case only 74 can be selected; otherwise the system will give a message that this program is limited to superconductivity alone and terminates it. Once superconductivity is selected, a second menu appears asking the user to select a closely related subject. This menu is created using the core related subjects. Actually here options for browsing hierarchically subordinate subjects or related subjects can be given. But, this program gives only closely related subjects. From this menu, 81(material science) has to be

chosen, otherwise it terminates again with a message. The message is that - it responds to class 81 only. Once 81 is chosen, only, selection of 81.20L is permitted in menu3 as the system is tailored to terms belonging to ceramics alone in relation to superconductivity. Once 81.20L is chosen, keywords can be entered and the data file is searched for keywords and pertinent record number and keywords are displayed.

6 Advantages of the system

The system has following advantages in information retrieval and enhances its efficiency.

- Data mining is possible
- One can enter in one subject and reach any related subject
- Data can be aggregated and disaggregated if notation is used
- Large database can be created
- Large knowledge based systems can be created

7 Limitations of the system

- Class number assigned may be wrong and lead to unnecessary noise
- User perception and system perception do not match, then retrieval may not be satisfactory.
- Software, Hardware and Architecture, needs to be developed

8 Conclusions

The creation of a "toy" knowledge-based system for associative subject relationship has been described. It indicates the possibility of designing front ends to assist retrieval of references pertinent to queries in classification based systems. This has been demonstrated for **'ceramic superconductors'** in this 'toy" model. But to create real-time systems one has to think of more suitable software tools that deal with knowledge manipulation and the sophisticated hardware and architecture

8 References

1. **Vickery (B. C.).** Analysis of information. *IN* Allen Kent [ed.], Encyclopedia of Library and Information science, Marcel Dekker : New York, 5, 1970, p. 356.
2. **Alberico (Ralph) and Micco (Mary).** Expert systems for reference and information retrieval, Meckler : Westport, 1990.
3. **Chudamani (K. S.) and Asundi (A .Y.).** A bibliometric analysis of cross references in a subject field: A case study of superconductivity, *PAPER ACCEPTED FOR Poster Session at the 4th International conference on Informetrics*, Bibliometrics, and Scientometrics. Rosary College, Chicago, Ill., USA. 1995.