A Knowledge Representation Model for Analytico-Synthetic Classification

M.A. Gopinath and A.R.D. Prasad, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, India

Abstract : This paper presents the preliminary results of the research work carried out by the authors at Documentation Research and Training Centre (DRTC), on automatic assignment of class numbers to book titles in the field of Library and Information Science. It should be made clear that it is not attempted to build a system that generates a classification schedule. On the other hand, we have attempted to translate the Colon Classification schedule into a kind of facts in a Frame based Expert System. The objective is to develop a Knowledge Representation (KR) model for analytico synthetic classification scheme. The system that is developed on this model is named PROLOGOMENA.

Methodology

1. About 500 book titles in Library and Information Science are collected.
2. A parser is developed using the programming language PROLOG (Programming in Logic). The Parser generates syntactic structure in order to identify noun-phrases in document title.
3. The Non-phrases are then passed on to a Frame-based Expert System for semantic analysis, i.e., identification of fundamental categories of the constituents of a title to facilitate construction of a class number.
4. The class numbers thus generated are compared against manually worked out class numbers.

Hypotheses

This work has been carried out basing on the following hypotheses.

- Noun-phrases play a significant role in document titles. This is in contrast with Schank's [3] approach, where it is verb-oriented.

- In subject classification, each noun-phrase (NP) plays a definite role in semantic analysis irrespective of the position of NP in a given title. In other words, a particular noun-phrase belongs to one only one fundamental category. In fact, it is rigid approach, which has to be reconsidered in the future work.

Advances in Knowledge Organization, Vol.4(1994), p. 320-327

Components of the System

The system is basically a two layered one with the following sub-systems.

1. Natural Language Processing (NLP) component.
2. Frame-based Expert System component.

NLP Component

It includes an interface program which reads a document title and converts into a list structure. For example, the title *Cataloguing of pamphlets in research libraries in India* is transformed into the following list structure

[cataloguing,of,pamphlets,in,research,libraries,in,india].

A top-down parser is developed using Definite Clause Grammars (DCG), developed by Pereira and Warren [2]. An illustration of the grammar is given below:

1. sentence –> noun-phrase, verb-phrases.
2. noun-phrase –> proper-noun.
3. noun-phrase –> determiner, noun.
4. verb-phrase –> intransitive-verb.
5. verb-phrase –> transitive-verb, noun-phrase.

Given a sentence, the parser identifies the syntactic cataegories (using a lexicon, *please see* Appendix-B) of constituent words and attempts to see whether a given sentence is a valid sentence according to the predefined rules of grammar and builds a syntactic structure for the given sentence. (The DCG rules for the document titles are given in Appendix-A). Thus the title of our example gives rise to a syntactic structure.

It is believed that in subject classification, noun phrases play an important role in semantic analysis. Thus, the parser also includes a procedure to identify the noun-phrases from the syntactic structure and passes it on to the KR component for semantic analysis.

The noun phrases identified from the syntactic structure are:

cataloguing
pamphlets
research library
India

KR Component

Basically there are three knowledge representation models, viz., Rule-based Expert System, Semantic nets and Frame-based Expert system. Of these frame based, KR model is chosen as it allows the hierarchical structure of the schedule and also allows procedural attachment for computing the class numbers.

The Knowledge engineering aspect of the present work involves representing the fundamental categories, viz., PMEST, the isolates and their class numbers in terms of facts. In other words, it is attempted to develop a frame for each isolate in the schedule and express the frame in a kind of logic notation so that it can be represented as PROLOG facts. The conceptual graph surrounding the node *university library* can be represented in the frame, logic and PROLOG notations as follows:

FRAME NOTATION
university library
a-kind of : library
category : personality

LOGIC NOTATION
a-kind-of(university library, academic-library)
category(university library, personality)
number(university library, J3)

PROLOG NOTATION
value(university library, a-kind-of, academic library).
value(university library, category, personality)
value(university library, number, J3)

Inference Mechanism

However, one of the advantages of frame based KR is that, we do not have to explicitly state the category of each and every isolate. If the category of the first order isolate is stated, the second and other order isolates inherit the category value from the first order isolate. Thus while constructing the frame, we mention the category of the isolate *library*; the isolate *academic library* inherits the category value and passes it on to *university library*. The Colon Classification schedule for Library and Information Science is translated into facts in frame based knowledge representation. (*see* Appendix-C).

Another advantage of frame based KR is that it allows thesaurus entries to be incorporated in the systems. Thus if *ranganathan classification* is synonymous to *colon classification* and the latter is the standard term, it can be reprsented in PROLOG notation as

value(ranganathan classification, use, colon classification).

Thus, the system replaces the non-standard terms by standard terms and activates the inference mechanism.

Another feature of the system is that, if there is no class number for a particular term, the system looks for the immediate super-ordinate term and assigns the number allocated to the super ordinate term. Here, the thesaurus entries representing broader term and narrower terms relation come handy to the inference engine.

Examples

1. Title : Cataloguing of pamphlets in research libraries in India.

NLP Parser generates the following noun-phrases

catgaloguing
pamphlets
research library
India

facts in the frame:

```
value(library, basic_subject, library science).
value(library science, numnber, 2)
value(personality, symbol,',').
value(matter_property, symbol, ';').
value(cataloguing, number, 6).
value(cataloguing, category, matter-property).
value(research library, number, K)
value(research library, a-kind-of, academic library).
value(academic library, a-kind-of, library).
value(library, category, personality).
value(pamphlets, a-kind-of, document)
value(document, category, matter-property)
value(pamphlets, number, 5z43).
value(India, category, space)
value(India, number, 44).
```

Thus the class number is: 2,K,5z43;6.44

Note: It should be noted that the facts need not be in any particular order. In fact, the Prolog backtracking mechanism takes care of the search. However, the order in which the class number to be computed is decided by the inference engine, where the rules are constructed in accordance with the facet formula.

2. Circulation of photostat material in unversity libraries

NL parser identifies the following Noun-phrases

circulation
photostat material
university library

facts in the frame:

value(university library, a-kind-of, academic library).
value(academic library, a-kind-of, library).
value(library, category, personality).
value(university library, number, J4).
value(circulation, category, matter-property).
value(circulation, number, 8).
value(photostat, a-kind-of, document).
value(document, category, matter-property).
value(photostat, number, 5z17).

Thus the Class number is: 2,J4,5z17;8

Conclusion

The present system is limited in many aspects. We have yet to try many of the features of analytico shynthetic classification. The system can compute class numbers if the given title falls in one and only one discipline. That is, it can implement facet relation within the basic subject in lihrary science. However, phase relation mechanism is yet to be implemented. Of the devices, it can handlecompound isolate device, other devices like geographical, chronological device are to be worked out in the future. We believe, that future research will yield bvetter results in this direction.

References

Periera, F.C.N.; Warren, D.H.D. (1980). Definite Clause Grammars for Natural Language Analysis. Artificial Intelligence, 13.

Ranganathan, S.R. (1987). Colon Classification, Ed.7. Bangalore(India): Sarada Ranganathan Endowment for Library Science.

Schank, R.C.; Childer, P. (1983). The Cognitive Computer: On Language, Learning and Artificial Intelligence. New York (USA): McGraw Hill.

Appendix–A:
GRAMMAR

The programs and the facts presented in the appendixes are developed using the Arity-Prolog Interpretor/Compiler on IBM PC compatible machines.

```
s(NP1:NP2)  --> np(NP1), pp(NP2).
s(NP1:VP:NP2 ) --> np(NP1), v(VP), np(NP2).
s(NP1:VP:NP2)  --> np(NP1), v(VP), pp(NP2).
s(GE:NP) --> gerund(GE) , s(NP).
s(GE1:GE2:NP) --> gerund(GE1), conj, gerund(GE2), s(NP).
s(NP:NP2) --> S(NP1), conj, s(NP2).

np(NP) --> det, np(NP).
np(N) --> n(N).
np(ADJ^N) -->  adj(ADJ), n(N).
np(ADJ1^ADJ2^N) --> adj(ADJ1), adj(ADJ2), n(N).
np(ADJ1^N:ADJ2^N) --> adj(ADJ1), conj, adj(ADJ2), n(N).
np(ADJ^NP1:ADJ^NP2) --> adj(ADJ), n(NP1), conj, n(NP2).
np(GE1:GE2:N) --> gerund(GE1), conj, gerund(GE2), s(N).

np(PN1)  --> pn(PN).
np(PN1^PN2) --> pn(PN1), pn(PN2).
np(PN1^PN2^PN3) --> pn(PN1), pn(PN2), pn(PN3).

PP(NP:VP:NP) --> prep, v(VP),s(NP).
pp(NP)  -->  prep,  s(NP).
pp(NP) --> [].

gerund(GE) --> [GE], {gerund(GE)}.
v(VP) --> [VP], {v(VP)}.
```

Appendix– B:
LEXICON

```
% ADJECTIVES

adj(subscription).
adj(council).
adj(children).
adj(commercial).
adj(private).
adj(expansive).
adj(sanskrit).

% CONJUNCTIONS

conj(and).
conj(as).
conj(between).
conj(with).

% DETERMINERS

det(a).
det(an).
det(its).
det(other).
det(some).
det(the).

% GERUNDS

gerund(abstracting).
gerund(building).
gerund(choosing).
gerund(generating).

% NOUNS

n(section).
n(recataloguing).
n(mutilation).
n(difference).
n(pamphlet, pamphlets).

% PROPER NOUNS

pn(dewey).
pn(kingdom).
pn(melvil).
pn(ranganathan).
pn(states).
pn(united).

% PREPOSITIONS

prep(at).
prep(by).
prep(for).
prep(to).
prep(towards).
prep(with).

% VERBS

v(applicable).
v(based).
v(build).
v(compared).
```

Appendix–C:
FACTS

```
value('public^library^system', kind_of, world^library^system).
value('public^library^ksystem', number, 'A*Z').
value(world^library^system, number, 'A').
value('public^library^system', category, '1p1).
value('1p1', symbol, ',').
value('public^library^system', basic_subject, library^science).
value(library^science, number, '2').
value(tamilnadu, category, space).
value(tamilnadu, number, '4411').
value(space, symbol, '.').
value(inter_library_organisation, category, c_matter).
value(inter_library_organisation, number, '1').
value(c_matter, symbol, ';').
value(legislation, number, '13').
value(legislation, category, c_matter).
value(classification, basic_subject, library^science).
value(classification, number, '5').
value(classification, category, c-matter).
value(sanskrit, kind_of, language^division).
value(sanskrit, number, '9B15').
value(sanskrit, category, '1p1').
value(language^division, number, '9B).
value(language^division, basic_subject, library^science).
```

The synthetic languages are the languages where the grammatical meaning expresses with the help of the endings, affixes, alternations (or simply the sound changing), suppletion (for example, in some Slavic languages there is imperfective and perfective form of words). Letâ€™s look at the example of suppletion in the Ukrainian language.