Abstracts

Alexandria, VA, October 16, 1994

ISSN: 2324-9773

PROCEEDINGS OF THE 5th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

ISSN: 2324-9773

Dynamic Displays for Browsing Hierarchical Classifications

James E. Agenbroad
Systems Analyst, Library of Congress
Box 291
Garrett Park, MD 20896-0291
jage@LOC.gov

"First of all it might be well to point out that libraries have put large amounts of money and effort into classifying their collections into a meaningful shelf arrangement, but then have done relatively little to explain to their readers exactly what this principle of arrangement is." Foster Palmer said this at the 1966 Brasenose Conference in a paper on the Widener Library shelflist project. While it remains substantially true, I believe the availability of classification schedules in computer processable form could soon give readers a new way to search by letting them burrow down through the layers of classification concepts to their topic. A premise of my approach is that the notation of a classification scheme (i.e. call numbers) is cryptic data suitable only for computers and librarians but confusing to readers and best kept behind the scenes. Though the following series of hypothetical screen images use the hierarchy of the LC classification, the same method of dynamic display could be applied to Dewey, NLM or other schemes since they all deal with hierarchies of concepts. Asterisks highlight the user's choice of a topic for deeper examination.

```
                Automated Browsable Classification        (screen 1)

Move the cursor to a topic that includes yours and key enter.

                              BASIC TERMS

General works          Social sciences          Medicine
Philosophy             Political science        Agriculture
Psychology             Law                      Technology
Religion               Education                Miltary science
History                Music                    Naval science
Geography              Fine arts                Bibliography
Anthropology           Language & literature    Library science
Recreation            *Science*                 Bibliographies
```

ISSN: 2324-9773

Automated Browsable Classification (screen 2)

Move the cursor to a topic that includes yours and key enter.

BASIC TERMS	SUBORDINATE TERMS
Law	Science (general)
Education	Mathematics
Music	Astronomy
Fine art	Physics
Language & literature	Chemistry
Science ----------	*Geology*
Medicine	Natural history
Agriculture	Biology (general)
Technology	Botany
Military science	(more)

Automated Browsable Classification (screen 3)

Move the cursor to a topic that includes yours and key enter.

SUPER-ORDINATE TERMS	COORDINATE TERMS	SUBORDINATE TERMS
Law	Science (general)	
Education	Mathematics	
Music	Astronomy	*Geology (general)*
Fine art	Physics	Mineralogy
Language & literature	Chemistry	Petrology
Science ----------	Geology ----------	Dynamic & structural
Medicine	Natural history	Stratigraphy
Agriculture	Biology (general)	Paleontology
Technology	Botany	
Military science	(more)	

Automated Browsable Classification (screen 4)

Move the cursor to a topic that includes yours and key enter.

SUPER-ORDINATE TERMS	COORDINATE TERMS	SUBORDINATE TERMS
Science (general)		
Mathematics		Geology (general)
Astronomy	Geology (general) -----	Study and teaching
Physics	Mineralology	*Geographic divisions
Chemistry	Petrology	
Geology ----------	Dynamic & structural	
Natural history	Stratigraphy	
Biology (general)	Paleontology	
Botany		
(more)		

ISSN: 2324-9773

Automated Browsable Classification (screen 5)

Move the cursor to a topic that includes yours and key enter.

SUPER-ORDINATE TERMS	COORDINATE TERMS	SUBORDINATE TERMS
		Miscellaneous regions
		Arctic regions
	Geology (general)	America
Geology (general) -----	Study and teaching	Europe
Mineralology	Geographic divisions --	*Asia*
Petrology		Africa
Dynamic & structural		Australia
Stratigraphy		New Zealand
Paleontology		Pacific islands
		Antarctic regions

Automated Browsable Classification (screen 6)

Move the cursor to a topic that includes yours and key enter.

SUPER-ORDINATE TERMS	COORDINATE TERMS	SUBORDINATE TERMS
	Miscellaneous regions	Afganistan
	Arctic regions	Arabian peninsula
Geology (general)	America	China
Study and teaching	Europe	Taiwan
Geographic divisions ---	Asia ----------	Mongolia
	Africa	India
	Australia	Burma
	New Zealand	Sri Lanka
	Pacific islands	Pakistan
	Antarctic regions	Indochina
		(more)

ISSN: 2324-9773

```
                Automated Browsable Classification      (screen 7)

Move the cursor to a topic that includes yours and key enter.

SUPER-ORDINATE TERMS     COORDINATE TERMS          SUBORDINATE TERMS

                         Miscellaneous regions     (more)
                         Arctic regions            Indochina
                         America                   Thailand
Geology (general)        Europe                    Malay peninsula
Study and teaching       Asia                      Malaysia
Geographic divisions --- Africa                    Indonesia
                         Australia                 Philippines
                         New Zealand               Japan
                         Pacific islands           Korea
                         Antarctic regions         North Korea
                                                   (more)
```

Using call numbers as behind the scenes links (e.g., geology of the Philippines = QE302) an OPAC that implemented this approach could show the number of items in a category and then display their bibliographic records. Access via and display of nonhierarchical data often found in classification systems — notes, index terms and cross references — deserve exploration but are beyond the scope of this brief paper.

The above is a purely personal proposal, neither the official view nor an ongoing project of the Library of Congress. Internet comments can reach me at jage@LOC.gov; regular mail should go to Box 291, Garrett Park, Md. 20896.

ISSN: 2324-9773

Updating the Thesaurus of the Grants Databases for the University of Tennessee System

Carolyn Sellers Ashkar
3014 Forestdale Ave. #16
Knoxville, TN 37917

This brief article deals with the modification of a database thesaurus. The University of Tennessee System Funding Opportunities databases describe research grants. Originally devised for use by several government agencies such as NIH and DOE, and periodically updated, this thesaurus was still inadequate to the needs of the University research community. Presented here are some of the weaknesses of the current thesaurus, corrective measures taken, and guidelines used in recreating the thesaurus, such as locating and choosing additional terms. Chief among the radical changes made was the restructuring of the categorical index: Replacing the alphabetic arrangement of terms with a hierarchical pattern. Problems and limitations in revising the thesaurus as a whole are also discussed.

The group of grants databases, such as "funding opportunities" and "sponsored programs", at the University of Tennessee are unique. Developed and maintained by the Office of Research and Technology Development (ORTD), these databases provide a "one-stop" online search system, bringing together from disparate government agencies and other sources, information on funding available for research, and a listing of research underway in the University of Tennessee System and the sources of funding. One of the services of ORTD is that of "targeting". Those wishing to receive information on possible funding in their areas of interest can select pertinent terms from the database thesaurus. The office team sends them information based on these choices. Faculty and staff can also search the system on their own.

The *University of Tennessee Keyword Thesaurus* (UTKT) is an adaptation of Rodman's *Keyword Thesaurus* developed for use by government agencies, DOE, NIH, and others. [Despite the names, neither of these publications is based on keywords. Even the narrowest terms are classes, except in the Medical area.] Rapid changes in science and technology, major developments in the social sciences, and an ever-increasing number of interdisciplinary areas of research, all demonstrate that periodically adding a few terms to the thesaurus is insufficient to really reflect the needs of the research community.

Aside from the need to add many terms, the UTKT also required some new classifications, including the breaking down of some large groupings into two or more. Other more radical changes also seemed highly desirable. The original format divided human knowledge into thirteen major classifications. Each of these had more narrow classifications (termed "broad classifications") under them; these in turn were over more specific terms listed in alphabetic order. The number coding was based on ten digits: the first two for the major classification (also known as "the first level"), the second two for the broad classification ("second level"), and three digits each for the third and fourth levels. (Example: 0503013000 05=**Education**, 03= **Educational Modes & Psychology & Theory,** 013=**Educational & Public Television**, 000 = {**not filled, which is common for all areas except Medicine**}). If this were a static system the disadvantages of it might more easily be tolerated, but because of the continued addition of terms, the changing popularity of terms used by the research community, the format, most especially the numbering system, made

ISSN: 2324-9773

general revision awkward, deformed the pattern of the classifications, and made the numbering below the first two levels meaningless.

Among the many difficulties of the UTKT: 1) only one term for each research topic was represented, if a searcher looked for an equally valid synonym, there was nothing to indicate that the topic was given under a different term; 2) there was no hierarchy of terms beyond the first and second levels. If terms dealing with a sub-area were not together alphabetically they were scattered throughout the section, for example, while **Solar System** and **Solar Flares** were not so much of a problem alphabetically, **Coronas** would have been nowhere near the other two in a list under Meteorology. (Even the two solar terms were not situated quite right because the more narrow term would fall before the broader.) This situation was keenly unfortunate in classifications with a large number of terms and several different topics. The terms for the sub-areas were all mixed together. While the numerical index (in the back of the book) helped some, the numbering system was so inflexible, that this index was not altogether clean either.

Time, budget, and a desire to stay connected to the thesaurus used by U.S. government agencies, placed constraints on changes made to the UTKT. However, the imperative to create a more usable instrument overrode these constraints in some cases.

The three main guidelines I tried to follow throughout this project were: 1) Use terms from the thesauri of established online databases, and glean ideas for classifications from these thesauri; 2) Try to make the patterns as universal as possible, and not just follow a personal view of how the terms should be arranged; 3) Beyond drawing terms from thesauri and receiving some from the office team, discover for inclusion in the thesaurus as many "cutting edge" terms as possible. These I found in such sources as *U.S. News and World Report* science and technology articles, the university and faculty newspaper accounts of research being done within the UT University System, national newspapers, and databases such as *InfoTrac, ABIInform*, and *MedLine*. The main criterion for choosing terms was that they be pertinent to research either being done or likely to be done within the University System and that they not be too esoteric. I also consulted subject experts who were most helpful in answering questions and making suggestions.

Changes were numerous and diverse. In the major classifications there are only a few differences, such as Business was added to Management and Commerce to become **Management, Business, and Commerce**. And conceptually, the business part of manufacturing was formally positioned under the Management ...rubric. From the "Other" [i.e. miscellaneous] major classification, Library Science, with a pathetically short list of terms, was moved out into its own major classification with the name changed to **Library and Information Science**, and given a substantial number of terms. On the second level a number of new classifications were added such as **Environmental Engineering**, while some additional broad classifications were broken down into two or more, such as the new **Fitness and Health Promotion** classification which was established when it was taken in part from the Medical broad classification, **Types of Intervention.**

The greatest structural changes — these occurred primarily on the third and fourth levels — concerned introducing a hierarchy. This broader to narrower arrangement meant that the number codes had to be changed for the terms involved. Now the number code of each term reflects both the term's classification and its hierarchical relationship. [The numbering system itself was not

 ISSN: 2324-9773

changed. Some remarks on this later.] The need for hierarchy, and for related terms to be more closely grouped, meant a total rearrangement for some major classifications. The Geography section is a case in point. The UTKT had no groupings of countries and no connection between a country and the part of a continent in which it is located, such as Thailand and Southeast Asia. For some country terms this lack is not important, but for those areas where professionals disagree as to which border countries are part of a certain region, such as the Middle East, the distinction is significant. There was also no breakdown by region of such ethnically and economically diverse places as the former Soviet Union.

Less dramatic enhancements were the addition of **See** and **See also** references, increased use of major qualifiers such as **Information Retrieval (CS)*** and **Information Retrieval (LIS)**** [*Computer Science; **Library and Information Science], and finally the use of delimiters as in **Information Policies/Science Policies — Non-U.S.**

Lack of time meant that not all of the project could be completed. Some broad classifications were established just as markers for later development. In several areas in the science and technology section, not all terms were arranged in hierarchical order. No substantive changes were made to a few classifications where a small number or no terms were added. Also, while many terms were entered in the Medicine/Health Sciences/Biochemistry classification, a minimum of alteration was done to this section. This enormous area however badly needs to be thoroughly reworked. The distinction between many of the broad classifications in this area have been blurred by what seems an arbitrary placement of terms.

In the next revision, a different number coding scheme should be adopted, one that would allow greater flexibility and be easier for clients to use. Some of the broad classifications under "Other" need further development. Lastly, the Ecology areas (science, and engineering) would be more convenient if they were closer to the Biology and Chemistry classifications.

The Office of Research and Technology Development and the Center for Information Studies had planned a survey of professors to gather desired terms for the new thesaurus. The survey was belayed because of the pressure to produce an updated version as soon as possible. So this survey when taken will enhance and refine what was done under this project.

Users' experience with the new version will reveal if there are any difficulties inherent in the changes that were made, and prompt suggestions for remedies and improvements. As a dynamic work, the thesaurus will never be truly completed.

ISSN: 2324-9773

ISSN: 2324-9773

Using Feature-Oriented Classification in Software Reuse

Jürgen Börstler
Umeå University, Sweden
Department of Computing Science
Umeå University
Umeå, Sweden
jubo@cs.umu.se

The usage of libraries of software components is an obvious approach to support software reuse (/Free, 87/). Since it is not sufficient to simply store great bunches of code in a simple archive, most of the early libraries did not succeed as expected.

Since the descriptions of software artifacts are usually very complex it is useful to introduce special descriptors to index the components. In feature-oriented classification these descriptors are built by sets of features. Each of these features describes a property (e.g., the kind of a component) or attribute (e.g., lines of code, authors, ...) of a component. Features can be refined to support more detailed descriptions. In FOCS (**f**eature-**o**riented **c**lassification **s**ystem, /Börs, 93/) there exist two kinds of refinement. A view-refinement represents a distinguishable aspect of a feature, like the name, the kind, the authors, etc. of a component (see figure 1). Each component may be classified according to each view of a feature.

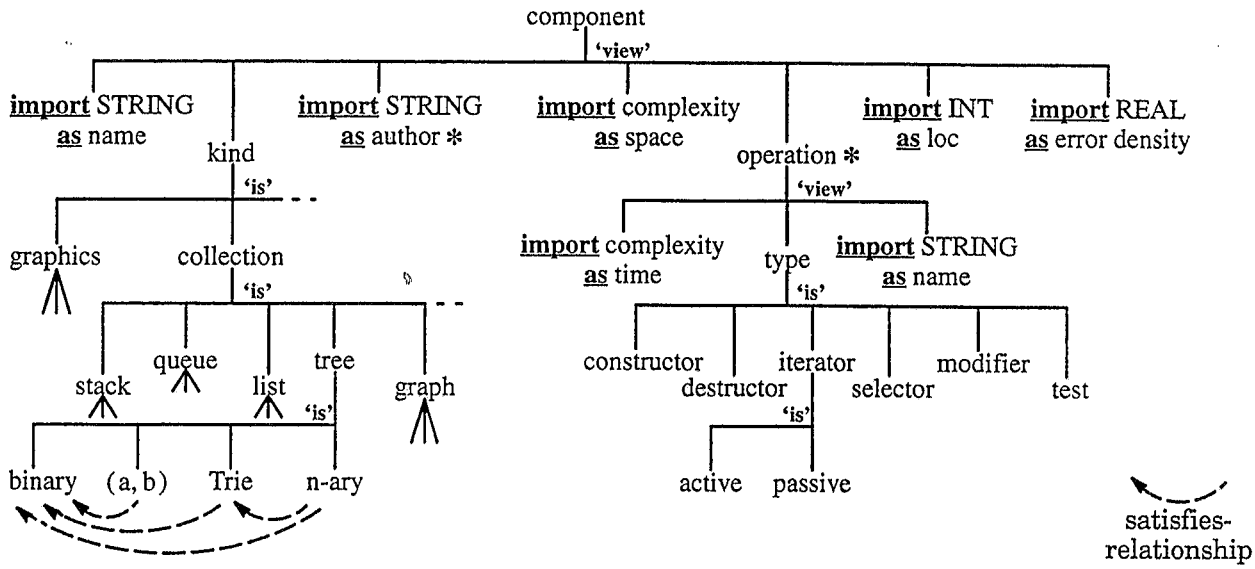

Figure 1: Excerpt of an example classification scheme.

ISSN: 2324-9773

An is-refinement represents the specialization of a feature. In our example we have is-refined the feature collection into stack, list, tree, and so on. Each component may be classified by only one specialization of a feature. Since the same feature may apply to a component in several shapes, repetition can be allowed by denoting features with a "*" (see feature operation in figure 1). The following descriptor classifies a component with name Example authored by jubo:

(name→Example, author→jubo, kind→collection→stack, space→constant,
operation→(type→constructor, name→push, time→constant),
operation→(type→destructor, name→pop, time→constant),
operation→(type→selector, name→top, time→constant),
loc→100, error density→0.01)

The component classified with the above descriptor is of kind stack and has constant space requirements. The components size is 100 lines of code (loc), its error density (number of known errors per line of code) is 0.01. The three operations of the component are (sub-) classified as of the types constructor, destructor, and selector with the names push, pop, and top, resp. All operations have constant time complexity. To avoid confusion, the set of features belonging to the same operation have to be grouped into so-called subdescriptors.

The classification language construct import allows us to (re-)use build-in schemes (INT, REAL, and STRING), as well as separately defined user subschemes (see complexity in figure 1).

Searching is also done with the help of descriptors. Users construct a descriptor containing the features the searched component should provide. This descriptor can be interpreted as a query to the database of classified components. For example, the descriptor

(author→jubo, kind→collection→stack)

will match our example component, represented by the first descriptor.

The retrieval algorithm takes into account the classification scheme to support the retrieval of similar components. Refinements of features are similar to the refined feature. The similarity of features can also be denoted explicitly by satisfies relationships. In our example classification scheme we have for example denoted that the feature n-ary satisfies the feature binary to express the fact, that n-ary trees can be used to implement binary trees. For simplicity we showed only a few satisfies relationships in our example and did not assign any weights to them. For an in-depth description of the classification language see /Börs 93/.

The similarity of components is derived from the similarity of their descriptors. The similarity of descriptors is computed from the similarity of the features they contain. A feature f matches a feature f' (by degree d):

$f \mid\!-_d f' \Leftrightarrow$
(i) $f = f'$ and $d=1$ $\vee$
(ii) f is a refinement of f' and $d=1$ $\vee$
(iii) f satisfies f' by degree d and $d \in]\,0..\,1]$
(iv) 0 else.

 ISSN: 2324-9773

The transitive closure of "|-d" is used to compute a similarity metric between components represented by their descriptors. The built-in subschemes INT and REAL have the usual '<' relation as a predefined satisfies-relationship. The descriptor

(loc→100, ...)

therefore matches also descriptors containing the feature loc→x, where x < 100. The default weights assigned to this relationship is 1. There is no predefined satisfies-relationship for STRING.

For searching purposes features can be given weights to indicate their relative importance in a descriptor. Weights can be chosen from the interval] 0 .. ∞ [. The default weight is 1. The descriptor

(kind→collection→stack: 2,
loc→100)

for example, emphasizes the feature kindÆcollectionÆstack over the feature loc→100). Therefore, components which only match feature kindÆcollectionÆstack receive a higher similarity, than the ones only matching feature loc→100).

The FOCS-library is organized as an attributed graph and stored in the special purpose database system GRAS (/KSW 93/). The classification schemes are stored together with the classified components (or rather representatives thereof) in the same graph. When a component is classified (i.e. inserted into the library) a new node in the underlying graph is created, representing this component. For each feature in the components descriptor we draw a link from the corresponding feature in the classification scheme to this node (see figure 2 for our example descriptor).

To take care of the groupings of features into subdescriptors additional nodes may be necessary. Using this library representation a query can be easily answered by collecting all component links found on the features mentioned in the query. The computation of the similarity metric can be simplified by extensive precomputation of the relation '$|-_d$'.

We have implemented the following tools in a FOCS prototype running on SUN workstations:

- An editor for the development and maintenance of classification schemes.
- An editor for the development and maintenance of component descriptors.
- A library to store all the information as an attributed graph as described above.
- A query tool to retrieve components to given descriptors.
- An integrated user interface for these tools.

The editors support syntax-directed and free text input. Additional analyzing tools check the context sensitive correctness of classification schemes and descriptors. Furthermore descriptors can be checked for their consistency with respect to a given classification scheme.

Since classification schemes tend to change as knowledge on the covered domains increase, we allow for arbitrary changes to the classification scheme. Some of these changes invalidate existing

ISSN: 2324-9773

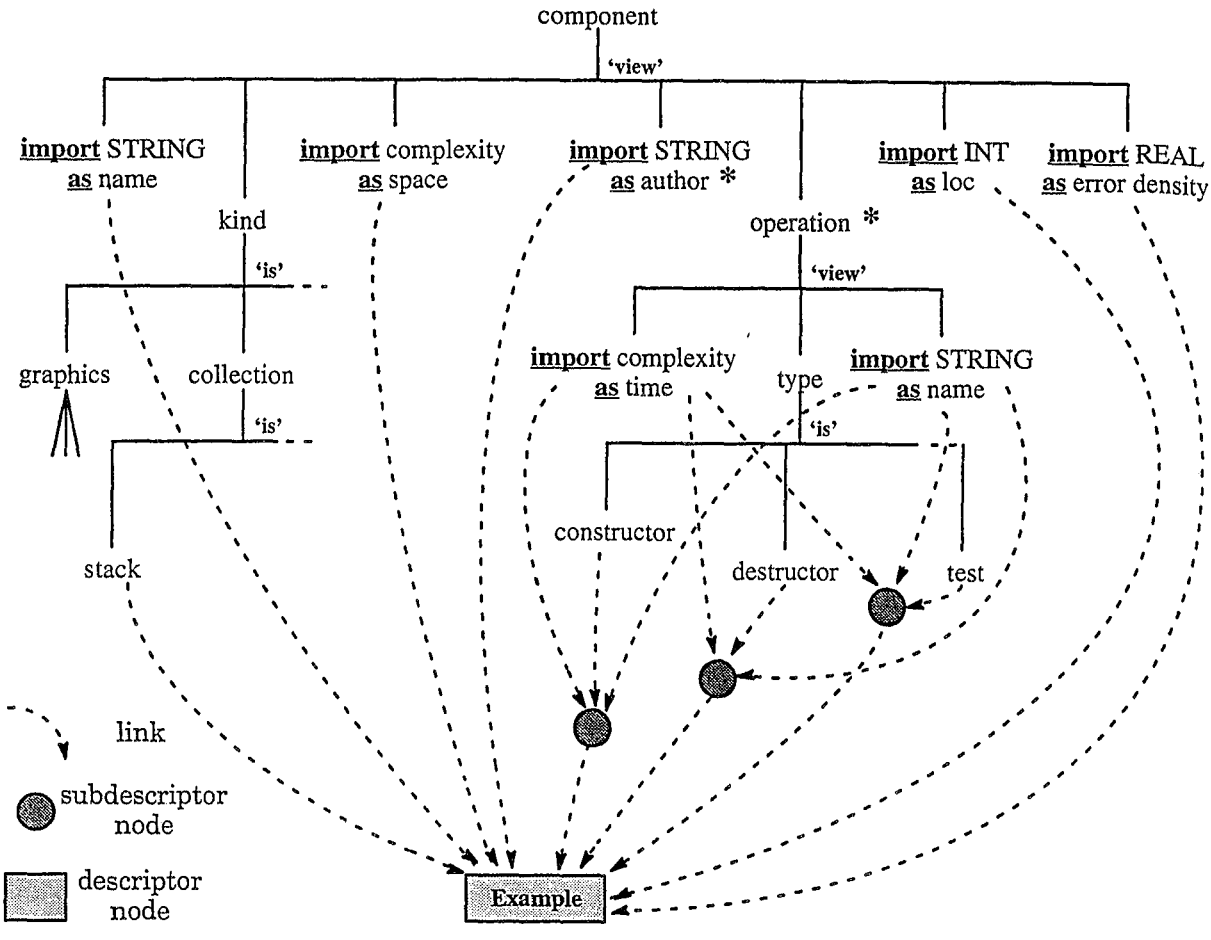

Figure 2: Example classification scheme with links to example component.

component descriptors. Our representation of the underlying classification scheme together with the descriptors as an attributed graph, allows us to access all descriptors which are affected by such a change directly (by resolving the corresponding links). FOCS can therefore cope with most of these changes without the necessity of complete reclassification (see /Börs 93/ for more details).

Although we have only applied FOCS to the domain of software components, it should be mentioned that FOCS in no way depends on this domain.

Up to now we have tested FOCS only on small libraries with classification schemes up to 150 features and 500 components. Our prototype showed acceptable performance with response times below three seconds. Since there are a lot of opportunities left for optimization, we expect acceptable performance also for realistic libraries of software components, i.e. up to 1000 features and 10000 components.

ISSN: 2324-9773

REFERENCES

/Börs 93/ J. Börstler, *Programming-in-the-Large: A Language, Tools, Reusability*, Ph.D. Dissertation, in German, Aachen University of Technology, Aachen, Germany, 1993.

/Free 87/ P. Freeman, *IEEE Tutorial: Software Reusability*, IEEE Computer Society Press, 1987.

/KSW 93/ N. Kiesel, A. Schurr, B. Westfechtel, "Design and Evaluation of GRAS, a Graph-Oriented Database System for (Software) Engineering Applications," *Proceedings CASE '93, the 6th Int. Conf. on Computer-Aided Software Engineering*, Singapore, 1993, 272-286.

ISSN: 2324-9773

 ISSN: 2324-9773

Comparison of Three Classification Systems for Information on Health Insurance

Lynn Silipigni Connaway
MaryEllen C. Sievert
School of Library and Information Science
University of Missouri-Columbia
104 Stewart Hall
Columbia, MO 65211

Newspapers today proclaim that health care reform is on its way. Insurance companies and physicians are in opposition. But, where does the consumer find information on health insurance as it has existed in the United States if he/she wants to really understand how things will differ? Libraries, public, academic or health, could provide materials on the subject. Where will these materials be found? Will the existing classification schemes put them near each other on the shelves of the library? These were the questions which prompted this research.

To test our hypotheses, the Dewey Decimal Classification System (DDC), the Library of Congress Classification System(LCC) and the National Library of Medicine Classification System(NLM) were compared. The sample was generated using the subject heading, "Insurance, Health — United States" on two databases. To find the NLM Classification System records, the first 50 titles under this heading from the CATLINE database from MEDLARS were examined. For the other two classification systems, the LC-MARC database on DIALOG was queried and the first 80 records under the heading were examined. The sample was larger in the LC-MARC database because there were a number of records with only the Superintendent of Documents Classification number available. Of these 80 records 38 had been assigned SuDoc numbers but 19 of these only contained SuDoc numbers, thus leaving a sample of 61 titles.

There were two hypotheses tested in the research. First, there would be no difference in the scatter of the three classification schemes. "Scatter" refers to the phenomenon that "as the literature on some subject grows, it becomes increasingly scattered ...and thus more difficult to identify, collect and organize."[1] Since one of the purposes of a classification scheme is the collocation of materials on the same subject, it would seem likely that most of the materials on this topic should be classified in one or, at least, only a few classes. As the number of classes increases, the subject becomes increasingly scattered.

The second hypothesis was that where there was overlap between all three schemes there would be no difference in the classes into which the subject was placed. That is, if an item appeared in the three classification schemes, it would appear in a similar class in each classification scheme, e.g.,something classified under "economics" in LCC would be classified under the corresponding number for "economics" in DDC or NLM. Eleven titles had been classified with all three systems.

The Dewey Decimal Classification System with ten large classes which attempts to classify the world of knowledge was developed in 1876. Since its scope is all of knowledge and it provides

1. Lancaster, F. W. *Indexing and Abstracting in Theory and Practice.* Champaign, IL: University of Illinois, p. 121.

ISSN: 2324-9773

only ten general classes in 1000 categories accompanied by a relative index, it would seem likely that there would be minimal scatter. In fact, the 61 titles had been assigned 13 distinct classification numbers. These 13 distinct numbers fell into 3 general classes, the 000, 300, 600. Two of the classification numbers contained cross-references to other numbers. Twenty-five records (41%) were classified in a single number, however, 362.1, "Problems of & Services to the Physically Ill," under the broad class of "Social Science" with the subclass of "Social Services." If we look at the first subclass, i.e., to the second number, e.g. the number 362.1 would be considered as belonging to subclass 360, there were seven subclasses into which the titles had been classified. Figure 1 depicts the percent in each subclass.

The Library of Congress Classification System, developed between 1899 and 1940, was based on literary warrant and was designed to classify all the titles in the library. In a sense, then, it, too, classified all knowledge but used 26 general classes. The 61 titles had been classified in five subclasses, AS, HD, HG, KF, RA, with an additional three subclasses each with a single title. Fifteen percent of the titles had been assigned to the number HD7102, "Economic History & Conditions. Social Insurance. Social Security. Pensions. US." Another thirteen percent had been assigned to the number RA410.53, "Public Aspects of Medicine. Economics of Medical Care. US. General Works." Three records had cross references to other classification numbers. Figure 2 shows the percent assigned to the subclasses with the single titles grouped under "Other."

The National Library of Medicine Classification System is the most specialized of the systems. Developed in 1948 and first published in 1951, it uses Class W of the LCC which is vacant and the medical parts of LCC's class Q. Class W is subdivided by all the other letters of the alphabet. All fifty titles had been assigned classification numbers in Class W with 13 distinct classification numbers in four subclasses. The largest portion of the sample, 24 titles (47%) had been assigned a single distinct classification number, W 275, "Medical Profession. Medical, Dental, Pharmaceutical or Psychiatric Plans, by Country." Seven numbers had each been assigned to a single title in the sample. Figure 3 shows the percent of titles in each subclass."

Eight of the eleven titles classified by all three systems had been classified in subclass RA by LCC. The titles were assigned to three subclasses in the NLM system and four subclasses in the DDC system, although 1 title was the only title in the sample assigned to subclass 610. One record from NLM and one record from DDC had cross references to other classification numbers. The remaining three titles were each in separate subclasses and the systems also did not correspond to the same subclass names for these titles. An examination of the records indicated that one LC classification number relates in class name to one NLM classification number and three other LC classification numbers relate to three DD classification numbers. The next step in the research will be to compare those records with classification numbers for both LCC and DDC.

Our data indicated that both hypotheses should be rejected. There is scatter in all three classification systems but the scatter is greatest in LCC and least in the NLM classification. When the class names are compared, LCC relates more closely to DDC than either relates to the NLM classification.

ISSN: 2324-9773

A Classification Scheme for Software Artifacts

M. R. Girardi
B. Ibrahim
University of Geneva
Centre Universitaire d'Informatique
24, rue General Dufour,
CH 1211 Geneve 4, Switzerland
girardi@cui.unige.ch
bertrand@cui.unige.ch

1 INTRODUCTION

Reuse systems that index software components manually are difficult and expensive to set up. Automatic indexing is required to turn software retrieval systems cost-effective. On the other hand, the effectiveness of traditional keyword-based retrieval systems is limited by the so-called "keyword barrier", i.e. these systems are unable to improve retrieval effectiveness beyond a certain performance threshold. This situation is particularly critical in software retrieval where users require high precision. Natural language processing techniques, for the acquisition of lexical, syntactic and semantic information from software descriptions, are potentially useful to improve retrieval effectiveness and to reduce the cost of creating and maintaining software libraries.

This paper briefly describes the classification scheme of a software reuse system [1][2][3][4][5] based on the processing of the descriptions in natural language of software components. Major requirements for the system are good retrieval effectiveness, cost-effectiveness and domain independence.

2 AN OVERVIEW OF THE REUSE SYSTEM

The current version of our reuse system consists of two major mechanisms: a classification mechanism and a retrieval mechanism.

The classification system [5] catalogues the software components in a software base through their descriptions in natural language. A knowledge acquisition mechanism automatically extracts from software descriptions the knowledge needed to catalogue them in the software base. The system extracts lexical, syntactic and semantic information and this knowledge is used to create a frame-based internal representation for the software component.

Semantic analysis of descriptions follows the rules of a semantic formalism. The semantic formalism is based on semantic relationships between noun phrases and the verb in a sentence. These semantic relationships ensure that similar software descriptions produce similar internal representations.

A classification scheme for software components derives from the semantic formalism, through a set of generic frames. The internal representation of a description constitutes the indexing unit for the software component, constructed as an instance of these generic frames.

ISSN: 2324-9773

Public domain lexicons (Webster and WordNet) are used to get lexical and semantic information needed during the parsing process.

The Knowledge base is a base of frames where each software component has a set of associated frames containing the internal representation of its description along with other information associated with the component like source code, executable examples, reuse attributes, etc.

The same analysis mechanism applied to software descriptions is used to map a query in free text into a frame-like internal representation. The retrieval system uses the set of frames associated to the query to identify similar ones in the Knowledge Base.

The retrieval system [4] looks for and selects components from the software base, based on the closeness between the frames associated to the query and software descriptions. Closeness measures are derived from the semantic formalism and from the conceptual distance between terms in the query and software descriptions. Software components are scored according to their closeness measure with the user query. The ones with a score higher than a given threshold become the reuse candidates.

As a first step the system deals with imperative sentences for both queries and software component descriptions. Imperative sentences describe simple actions that are performed by a software component and perhaps the object manipulated by the action, the manner by which the action is performed and other semantic information related to the action.

3 THE CLASSIFICATION FORMALISM

A semantic formalism establishes the rules to generate the internal representation of both queries and natural language descriptions of software components. The formalism consists of a case system for simple imperative sentences with some constraints and heuristics that are used to map a description into a frame-based internal representation.

The case system basically consists of a sequence of one or more semantic cases. Semantic cases are associated to some syntactic compounds of an imperative sentence. An imperative sentence consists of a verb (representing an action) possibly followed by a noun phrase (representing the direct object of the action) and perhaps some embedded prepositional phrases. For instance, the sentence 'search a file for a string' consists of the verb 'search', in the infinitive form, followed by the noun phrase 'a file', which represents the object manipulated by the action, and followed by the prepositional phrase 'for a string', which represents the goal of the 'search' action. In the example, the semantic cases 'Action', 'Location' and 'Goal' are respectively associated to the verb, direct object and prepositional phrase of the sentence. Semantic cases show how noun phrases are semantically related to the verb in a sentence. For instance, in the sentence 'search a file for a string', the semantic case 'Goal' associated to the noun phrase 'for a string' shows the target of the action 'search'. We have defined a basic set of semantic cases for software descriptions by analyzing the short descriptions of Unix commands in manual pages. These semantic cases describe basically the functionality of the component (the action, the target of the action, the medium or location, the mode by which the action is performed, etc.).

ISSN: 2324-9773

A semantic case consists of a case generator (possibly omitted) followed by a nominal or verbal phrase. A case generator reveals the presence of a particular semantic case in a sentence. Case generators are mainly prepositions. For instance, in the sentence 'search a file for a string', the preposition 'for' in the prepositional phrase 'for a string' suggests the 'Goal' semantic case.

4 THE CLASSIFICATION PROCESS

Morpholexical, syntactic and semantic analysis of software descriptions is performed to map a description to a frame-like internal representation.

The purpose of morpholexical analysis is to process the individual words in a sentence to recognize their standard forms, their grammatical categories and their semantic relationships with other words in a lexicon. Two semantic relations between terms are currently considered: synonymy and hyponymy/ hypernymy.

Just after morpholexical analysis, both syntactic and semantic analysis of software descriptions are performed interactively by using a definite clause grammar. The defined grammar implements a subset of the grammar rules for imperative sentences in English. The grammar supports the case system and states domain-independent knowledge of the English language through a set of syntactic and semantic rules.

A set of semantic structures is generated as a result of the parsing process, representing the internal structures of software descriptions. A language for modelling these semantic structures is shown below.

Case_frame --> FRAME Frame_name Hierarchical_link CASES Case_list.
Hierarchical_link --> IS_A Frame_name | IS_A_KIND_OF Frame_name
Case_list --> Case (Case_list)
Case --> Case_name Facet
Case_name --> Semantic_case | Other_case
Semantic_case --> Action | Agent | Comparison | Condition |
Destination| Duration | Goal | Instrument | Location | Manner| Purpose|
Source | Time
Other_case --> Modifier | Head | Adjective_modifier |
Participle_modifier | Noun_modifier
Facet --> VALUE Value | DOMAIN Frame_name | CATEGORY Lexical_category
Value --> string | Frame_name
Lexical_category --> verb | adj | noun | adv | component_id | string

The language defines a frame-like classification scheme for software components based on the defined semantic cases. The classification scheme consists of a hierarchical structure of generic frames ('IS-A-KIND-OF' relationship). Frames that are instances of these generic frames ('IS-A' relationship) implement the indexing units of software descriptions.

ISSN: 2324-9773

The generic frames model semantic structures associated to verb phrases, noun phrases and the information associated to software components, like name, description, source code, executable examples, etc.

Semantic cases are represented as slots in the frames. 'Facets' are associated to each slot in a frame, describing either the value of the case or the name of the frame where the value is instantiated ('value' facet); the type of the frame that describes its internal structure ('domain' facet) or the lexical category of the case ('category' facet). For instance, the 'Location' slot in the verb phrase frame has a 'domain' facet indicating that its constituents are described in a frame of type 'noun phrase'.

Through the parsing process, the interpretation mechanism maps the verb, the direct object and each prepositional phrase in a sentence into a semantic case, based on both syntactic features and identified case generators.

REFERENCES

[1] M. R. Girardi and B. Ibrahim, "New Approaches for Reuse Systems," *Position Paper Collection of the 2nd. International Workshop on Software Reuse (IWSR-2)*, Lucca, Italy, March 24-26, 1993.

[2] M. R. Girardi and B. Ibrahim, "An Approach to Improve the Effectiveness of Software Retrieval," *Proceedings of the Third Irvine Software Symposium (ISS'93)*, Irvine, California, April 30, 1993, pp. 89-100.

[3] M. R. Girardi and B. Ibrahim, "A Software Reuse System based on Natural Language Specifications," *Proceedings of the 5th International Conference on Computing and Information (ICCI'93)*, Sudbury, May 27-29, 1993, pp. 507-511.

[4] M. R. Girardi and B. Ibrahim, "A Similarity Measure for Retrieving Software Artifacts," *Proceedings of Sixth International Conference on Software Engineering and Knowledge Engineering (SEKE'94)*, Jurmala, Latvia, June 21-23, 1994, pp. 478-485.

[5] M. R. Girardi and B. Ibrahim, "Automatic Indexing of Software Artifacts, "*Proceedings of 3rd International Conference on Software Reuse*, Rio de Janeiro, Nov. 1-4, 1994. (To appear).

ISSN: 2324-9773

Conceptual Graphs as Semantic Representation of Noun-Noun Compounds in English and Chinese

Shaoyi He
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
HES.ILS@MHS.UNC.EDU

Conceptual graphs evolved as a semantic representation with natural language expressivity to allow the mapping to and from natural language to be simple and direct (Sowa, 1984; 1991). This project investigates conceptual graphs, focusing on their expressive power to represent the semantic structures of noun-noun compounds and their practical application to solving the problems of conceptual representation of noun-noun compounds in English and Chinese for natural language processing. Our investigation is based on the following propositions:

(1). In both English and Chinese, the syntactic structure of a noun-noun compound is the shorthand representation of its semantic structures. From the linguistic point of view, English and Chinese noun-noun compounds have the same syntactic rule, N ---> N N, which means a noun is formed by the combination of two nouns. But, such a combination means much more than the combination itself when we look at the underlying structures of a noun-noun compound. In the following examples, the meaning of a noun-noun compound is paraphrased by a clause:

English:	dog food	"the food that is for dogs to eat"
	junk food	"the food which is like junk"
	sea food	"the food which comes from sea"
Chinese:j	in-biao	"gold watch (a watch made of gold)"
	shou-biao	"wrist watch (a watch worn on wrist)"
	miao-biao	"stop watch (a watch used for special timing)"

(2). The semantic structure of a noun-noun compound is constrained by the underlying relations of the nouns in that noun-noun compound. In the above English examples, "dog food", "junk food" and "sea food" all simply have "food" as the second noun, but their underlying relationships are not simple at all. For instance, "dog food" is the "food" for a "dog" (to eat), and the underlying relations between "dog" and "food" is that of an *agent* and an *object*. So, the semantic structure of "dog food" is "an OBJECT (= food) of the to-eat EVENT with an AGENT (= dog)". But "junk food" and "sea food" have different semantic structures although they both fall into the same category of "food" as "dog food". "Junk food" is not the "food" for a "junk" (to eat) and "sea food" is not the "food" for "sea" (to eat). The underlying relationships between "junk" and "food", "sea" and "food" are not those of *agents* and *objects*. Instead, "junk" and "sea" are both ATTRIBUTES of "food" for describing its specific characteristics. Therefore, the semantic structure revealed by the underlying relationships of the nouns in "junk food" is "an OBJECT (= food) of the same quality as another OBJECT (= junk)". Also, the semantic structure constrained by the underlying relationships between nouns in "sea food" is "an OBJECT (= food) that comes from a PLACE (= sea)". These kind of constraints of the underlying relationships on the semantic structures are also influential among Chinese noun-noun compounds. For example, *jin-biao* "gold watch (a watch made of gold), *shou-biao* "wrist watch (a watch worn on wrist" and

ISSN: 2324-9773

miao-biao "second-watch = stop watch (a watch used to time events to a fraction of a second)" have different semantic structures because there are different underlying relationships between the attributive nouns *jin* "gold", *shou* "wrist", *miao* "second "and the head noun *biao* "watch". So, a *jin-biao* "gold watch" is a TIMEPIECE (= watch) made of a MATERIAL (= gold); a *shou-biao* "wrist watch "is a TIMEPIECE (= watch) worn on the BODYPART (= wrist), and a *miao-biao* "stop-watch" is a TIMEPIECE (= watch) for special TIMING (= second)". Therefore, for both English and Chinese noun-noun compounds, same underlying relationships between the nouns show same semantic structures and different underlying relationships illustrate different semantic structures.

(3). Semantic relationships in the underlying structures of a noun-noun compound can be illustrated by conceptual graphs which reveal the semantic structures of noun-noun compounds in English and Chinese. A traditional notion in linguistics considers the first noun in some certain noun-noun compounds to be functioning as an adjective and the following are some examples (Liberman and Sproat, 1992):

rubber boots	The boots are rubber.
gold medal	The medal is gold.

It is very interesting to notice that for the above English noun-noun compounds, we can find their counterparts in Chinese with the same structures and meanings:

jiao-xie	"rubber-boots"
jin-pai	"gold medal"

The underlying relationships between the nouns in this kind of noun-noun compounds can be expressed by the conceptual relation attribute (ATTR) because "(ATTR) links [ENTITY: * x] to [ENTITY: * y] where * x has an attribute * y" (Sowa, 1984):

English:	**Chinese:**	**Conceptual Graphs:**
rubber boots	jiao-xie	[BOOTS] ---> (ATTR) ---> [RUBBER].
gold medal	jin-pai	[MEDAL] ---> (ATTR) ---> [GOLD].

Other kinds of noun-noun compounds in English and Chinese can also be represented by conceptual graphs for their semantic structures:

English:

(1). circus-elephant "an elephant that performs in a circus"
[ELEPHANT]<--- (AGNT) <--- [PERFORM] ---> (LOC) --->[CIRCUS].

(2). elephant-circus "the circus that has a performing elephant"
[CIRCUS]<--- (LOC) <--- [PERFORM] ---> (AGNT) --->[ELEPHANT].

Chinese:

(1). mu-qiang "wood gun (the gun that is carved out of wood)"
[GUN] <--- (RSLT) <--- [CARVE] ---> (MATR) ---> [WOOD].

(2). hangkong-xin "airmail letter (the letter sent by airmail)"
[LETTER] <--- (OBJ) <--- [SEND] ---> (MANR) ---> [AIRMAIL].

ISSN: 2324-9773

The problems associated with noun-noun compounds have been notoriously difficult in both English (McDonald & Hayes-Roth, 1978; Finin, 1986; Lehnert, 1988; Ritchie et. al., 1992) and Chinese (Li & Thompson, 1981; Chang, 1991). There have been a variety of extensive studies on the interpretation of English and Chinese noun-noun compounds but with no satisfactory results. For example, there have been investigations on noun-noun compounds in English and Chinese, taking approaches from linguistics (Lees, 1970; Warren, 1978; Levi, 1979; Fung 1979; Li & Thompson, 1981), psychology (Hampton, 1988; Murphy, 1990), artificial intelligence (Finin, 1982; Gershman, 1979; Lebowitz, 1984), computational linguistics/natural language processing (Leonard 1984; Ritchie et. al., 1992; Sproat, 1992; Vanderwende, 1993), information retrieval (Jones 1971; Eichman, 1978; Gay & Croft, 1990), machine translation (Chang, 1991) and set theory (He, 1993). However, little attention has been paid to the conceptual graphs approach to noun-noun compounds in English and Chinese. Although the theory of conceptual graphs has been regarded as a powerful means for representing semantic structures of natural language, few people have thought it as the right tool for solving the problems associated with English and Chinese noun-noun compounds. Even Sowa himself only mentions briefly the treatment of English noun-noun compound with conceptual graphs, without further intention to formally deal with the issue in detail. Using "key employee" as an example, Sowa has given a short conceptual graph analysis:"When one noun "key" modifies another noun "employee", the modifier does not make the "employee" into a type of "key". Instead, it shows that there is some implicit relation between the concepts [KEY] and [EMPLOYEE]" (Sowa, 1984). Of cause, the whole spectrum of the semantic and conceptual relations between the nouns in noun-noun compounds is much more colorful and complex than that of "key employee", thus calling for a systematic and theoretical conceptual graphs approach to noun-noun compounds in English and Chinese. The outcome will be especially beneficial to the field of information retrieval, because noun-noun compounds remain a difficulty while playing a very important role in such areas as indexing and classification (Hutchins, 1975), subject headings and catalogues (Eichman, 1978, Coates, 1960), thesaurus construction (Jones, 1981), terminological systems (Sager, 1990) and retrieval systems (Gay & Croft, 1990; Jones 1971). In addition, it will lead to a better understanding of noun-noun compounds in English and Chinese, building up a knowledge-based machine translation of noun-noun compounds between English and Chinese, with conceptual graphs as "a universal language-independent deep structure" (Sowa, 1984). Finally, it will fill in the blank that there has been no conceptual graphs approach to any language phenomenon in Chinese.

REFERENCES

Chang, C. H. 1991. *Resolving Ambiguities in Mandarin Chinese: Implications for Machine Translation.* Ph.D. Dissertation, Northwestern University.

Eichman, T. L. 1978. Subject Heading Syntax and "Natural Language" Nominal Compound Syntax. *ASIS-1978,* 126-129.

Finin, T. W. 1982. The Interpretation of Nominal Compounds in Discourse. *Technical Reports MS-CIS-82-3*, Moore School of Engineering, University of Pennsylvania.

Fung, M. Y. 1979. A Contrastive Analysis of Word-Formation of NOUNS in English and Chinese. Babel: *International Journal of Translation.* 25:1, 131-145.

Gay, L. S. and W. B. Croft. 1990. Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management*, 26:1, 21-38.

Gershman, A. V. 1979. *Knowledge-based Parsing*. Ph. D. Dissertation, Yale University.

ISSN: 2324-9773

Hampton, J. A. 1988. Over Extension of Conjunctive Concepts: Evidence for a Unitary Model of Concept Typicality and Class Inclusion. *Journal of Experimental Psychology*, 14:1, 12-32.

He, S. 1993. Set as the Basis for Conceptual Representation of Noun-noun Compounds in English and Chinese. Paper presented at the ASIS Miniconference on Student Research in Information Science, April 24, 1993, Indiana University.

Huchins, W. J. 1975. *Languages of Indexing and Classification.* Sterenage: Peter Peregrinus.

Jones, K. P. 1981. Problems Associated with the Use of Compound Words in Thesauri, with Special Reference to BS 5723:1979. *Journal of Documentation*, 37:2, 53-68.

Jones, K. P. 1971. Compound Words: a Problem in Post-Coordinate Retrieval Systems. *Journal of the American Society for Information Science*, 22, 242-50.

Lebowitz, M. 1984. Using Memory in Text Understanding. *Proceedings of the 1984 European Conference on Artificial Intelligence*, Pisa, Italy.

Lees, R. B. 1970. Problems in the Grammatical Analysis of English Nominal Compounds. In Bierwisch, M. & K. E. Heidolph (eds.). *Progress in Linguistics: A Collection of Papers.* Mouton & Co. The Hugue, 1970.

Lehnert, W. G. 1988. The Analysis of Nominal Compounds. In Eco, U., M. Santambrogio, and P.

Leonard, R. 1984. *The Interpretation of English Noun Sequences on the Computer.* Amsterdam: North-Holland.

Levi, J. N. 1979. *The Syntax and Semantics of Complex Nominals.* New York, NY: Academic Press.

Li, C. N. and S. A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar.* Berkeley, CA: University of California Press.

Liberman, M. and R. Sproat. 1992. The Stress and Structure of Modified Noun Phrase in English. In Sag, I. A. and A. Szabolcsi. (eds.). *Lexical Matters.* Stanford, CA: Stanford University Press.

McDonald, D. & F. Hayes-Roth. 1978. Inferential Searches of Knowledge Networks as the Approach to Extential Language-Understanding Systems. In Waterman, D. A. & F. Hayes-Roth (eds.). *Pattern-Directed Inference Systems.* New York, NY: Academic Press. 1978.

Murphy, G. L. 1990. Noun Phrase Interpretation and Conceptual Combination. *Journal of Memory and Language*, 29, 259-288.

Ritchie, G. D., G. J. Russell, A. W. Black, and S. G. Pulman. 1992. (eds.). *Computational Morphology: Practical Mechanisms for the English Lexicon.* Cambridge, MA: The MIT Press.

Sager, J. C. 1990. *A Practical Course in Terminology.* John Berjamins Publishing Company.

Sowa, J. F. 1991. Toward the Expressive Power of Natural Language. In Sowa, J. F. (ed.). *Principles of Semantic Networks: Explorations in the Representation of Knowledge.* San Mates, CA: Morgan Kaufmann.

Sowa, J. F. 1984. *Conceptual Structures: Information Processing in Mind and Machine.* Menlo Park, CA: Addison-Wesley.

Sproat, R. 1992. *Morphology and Computation.* Cambridge, MA: The MIT Press.

Vanderwende, L. 1993. SENS: The System for Evaluating Noun Sequences. In Jensen, K., G. H. Heidorn and S. Richardson (eds.). *Natural Language Processing: the PLNLP Approach.* Boston. MA: Kluwer Academic Publishers.

Warren, B. 1978. *Semantic Patterns of Noun-noun Compounds.* Gotevorg: ACTA Universitatis Gothoburgensis.

ISSN: 2324-9773

Hierarchical Classification as an Aid to Browsing

J. Royce Rose
Caroline M. Eastman
Department of Computer Science
The University of South Carolina
Columbia, South Carolina 26208
rose@cs.scarolina.edu
eastman@cs.scarolina.edu

A navigational aid for databases that relies on unsupervised hierarchical classification is being developed and tested on databases of chemical reactions (Rose and Gasteiger 1994). The approach to hierarchical classification, based on both semantic and topological features, supports the creation of deep hierarchies in which succeeding levels represent increasing degrees of abstraction. This allows the user to quickly evaluate the result of a query and to locate interesting items and classes of items by performing a tree traversal rather than a sequential perusal of a hit list or a series of *ad hoc* query refinements. In very large databases where classical querying methods are increasingly inadequate such as chemical reaction databases, such a browsing method is required in order to manage the flood of information the user is confronted with.

In order to derive substantial benefit from giving hierarchical order to data, the resulting classification trees should strike a balance between depth and breadth. For this reason, one important goal in the design of the classification algorithm was to produce classification hierarchies expressing a large range of abstraction. This requirement motivated us to design a classification algorithm combining both phases of semantic and topological classification. A classification based on semantic features makes it possible to recognize similarity between objects that may be topologically dissimilar. On the other end of the spectrum, consideration of topological features makes it possible to refine a classification by extending it in the direction of greater specificity. The algorithm is shown in figure 1.

This method starts by calculating the semantics of the objects being classified. A conceptual clustering approach (Michalski and Stepp 1983, Fisher and Langley 1985, Fisher, Xu and Zard 1992) is taken in order to limit the kinds of classes that may be produced by domain-specific considerations. A hallmark of our approach is the alternation between phases of classification and generalization and the way in which semantic and topological classification is combined. An initial classification based on semantic features is performed. The semantic features in the case of chemical reactions consist of descriptions of chemical structure in terms of electronic and energy parameters. These describe the *meaning* of the structure and make it possible to create chemically valid equivalence classes of reactions. The semantic classification is followed by alternating phases of topological classification and generalization of both semantic and topological descriptions. After the topological classification stabilizes a final generalization based on the initial semantic classification is performed.

ISSN: 2324-9773

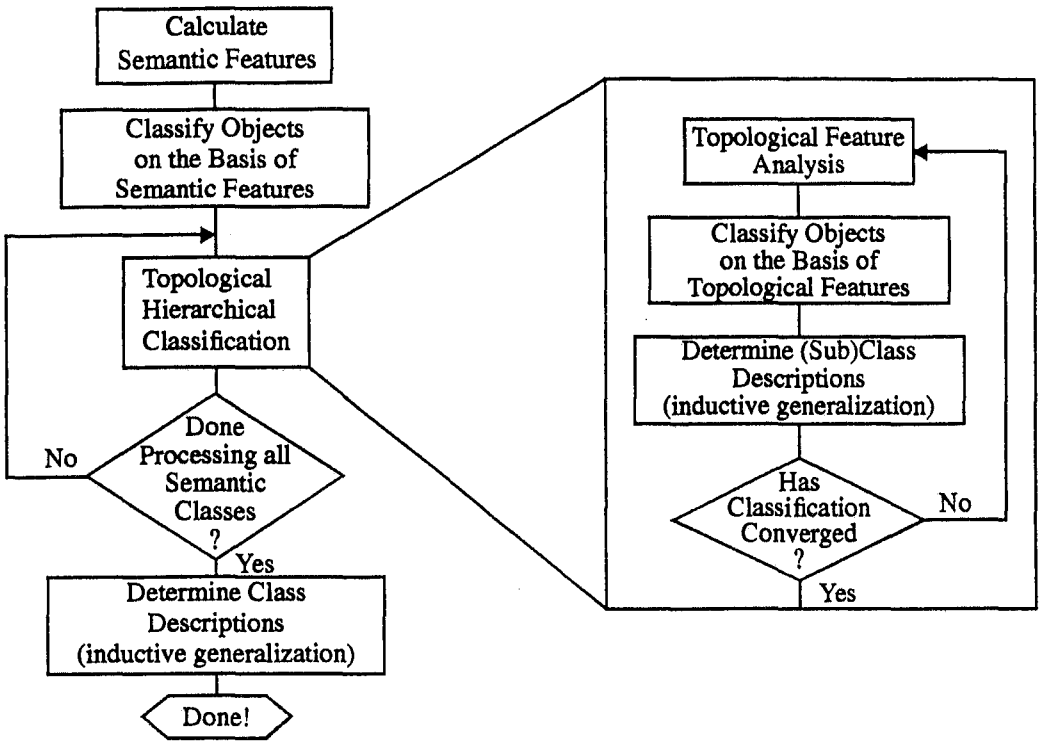

Figure 1. Hierarchical classification algorithm combining semantic and topological metrics.

As can be seen in figure 2, the topological hierarchical classification actually expands the classification tree in between the semantic classification and the final semantic generalization level. In a given hierarchy, each level represents a different degree of abstraction. The original objects being classified are at the lowest level. These items are then classified on the basis of similarity. The next layer consists of generalizations of the classes formed by classification of these items.

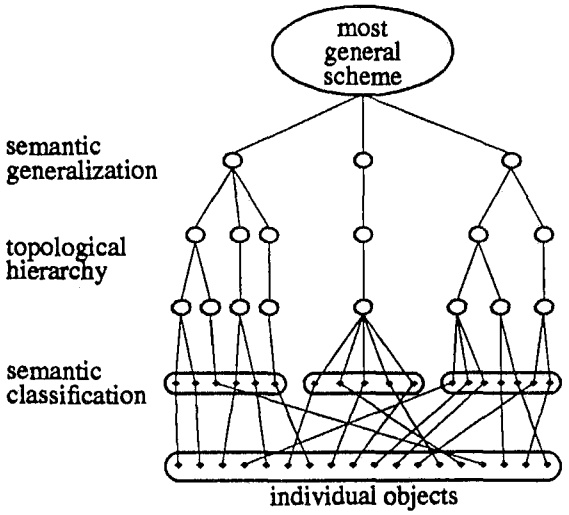

Figure 2. Stylized classification tree.

ISSN: 2324-9773

Each level in the hierarchy is an abstraction of the level below it.The goal is to provide class summaries which are stored at the next highest level of abstraction in the hierarchy. The topmost item in the hierarchy summarizes all of the objects in the tree and is therefore the most general description.

The creation of a hierarchical classification on a hit list allow the user to examine the hit list by performing a tree-traversal. This makes it possible to rapidly evaluate the contents of the hit list and quickly locate those items or sets of items the user is searching for or to determine that they are not present. Browsing by traversing such a hierarchy is equivalent to being able to query by similarity.

REFERENCES

Rose, J.R., Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci.*, 1994, 34, 74-90.

Michalski, R.S., Stepp, R. Learning from Observation: Conceptual Clustering. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.), *Machine Learning: An artificial intelligence approach.* Palo Alto, CA: Tioga Publishing Company, 1983.

Fisher, D., Langley, P. Approaches to Conceptual Clustering. *Proc. of the 9th Intl. Joint Conf. on Artificial Intelligence.* Los Angeles, CA: Morgan Kaufman, 1985.

Fisher, D., Xu, L., & Zard, N. Ordering Effects in Clustering. In *Proc. of 9th Intl. Conf. on Machine Learning*, Aberdeen, Scotland, (July), 163-168, 1992.

ISSN: 2324-9773

ISSN: 2324-9773

OPACs, Integrated Thesauri, and User Language

David L. Weisbrod
School of Information and Library Science
Pratt Institute
Brooklyn, NY 11205
weisbrod@zodiac.rutgers.edu

The author's purposes are (1) to develop quantitative measures that can be used both (a) to describe one particularly critical aspect of the performance of an OPAC or other IR system, *viz.*, the system's ability to accommodate and facilitate the end user's task of gaining initial access to the database when performing a subject search, and (b) to describe the usefulness of an integrated thesaurus or other syndetic apparatus as one particular genre of mechanism for improving the ability of the system to facilitate the user's ability to accomplish this task; (2) to demonstrate the viability of these measures by actually employing them to describe a body of observed test data based on "real world" use of OPACs.

Subject Areas: Information retrieval (IR), online public access catalog (OPAC), thesaurus, user behavior, performance measurement and evaluation, failure analysis.

RESEARCH QUESTION

Two very practical challenges to the design and implementation of IR systems (OPACs in particular) are the challenge of *subject search failure* and the subsequent challenge of *nothing retrieved; search abandoned.* The failure of certain end-users of online public catalogs (OPACS) to gain initial entry to the database when performing a subject search has been observed by a number of researchers. If relevant materials, appropriately cataloged, are actually present in the library's collection, but the user nonetheless fails to locate in the catalog any record representing even a single one of these materials (i.e., if the user fails to gain initial entry to the bibliographic database), then this *subject search failure* (SSF) is a problem. The user's SSF may often be followed by the user's abandoning the search altogether, rather than by the user's persevering by exploring some alternative avenue of approach. If relevant materials, appropriately cataloged, are actually present in the library's collection, this subsequent *nothing retrieved; search abandoned* (NRSA) behavior is a second problem, more severe than the first.

These challenges, which have been observed by many writers, are addressed in the present writer's research by taking as given (1) the view that IR is, in some fundamental sense, a matter of natural language and of its use, (2) the enormous variability in the use made of natural language by humans; and by recognizing (3) the disparate sources of the language involved in IR systems and processes, as well as (4) a number of exacerbating conditions that bear negatively on the performance of the contemporary OPAC as a tool for IR. The idea that end-user searching of an OPAC can be expedited through the addition of an integrated thesaurus is not a new one, but the question of devising and applying quantitative measures of the effectiveness of such a thesaurus in facilitating the critical user task of gaining initial access to the targeted database has not been aggressively pursued. This introduces the cluster of research questions that this research addresses:

ISSN: 2324-9773

Do syndetics help? Can syndetics help? Can we measure quantitatively the effectiveness of such help?

RESEARCH GOALS AND DESIGN

To investigate the effectiveness of the entry vocabulary component of a thesaurus, integrated into an OPAC, in facilitating the initiation of subject searches by end-users, this research employs automatically recorded session logs that were generated to monitor the behavior of real users consulting the OPAC in three different libraries as they performed their own searches on their own topics. Each of these OPACs possessed little or nothing in the way of syndetic structure. This research operationalizes the concept of success in search initiation by identifying its complement, failure in search initiation, as embodied in co-occurrences of *subject search failure* and its sequela, *nothing retrieved; search abandoned.* It operationalizes the concept of facilitating subject search initiation with an integrated thesaurus by taking failed searches and processing them against a (stand-alone, non-integrated) thesaurus that is available to the researcher but which was unavailable to the OPAC user. This research seeks to measure the need for an integrated thesaurus in the OPAC (i.e, for integrated syndetics) by calculating, inter alia, the incidence of SSF and NRSA outcomes. It seeks to measure the potential effectiveness of an integrated thesaurus by calculating the percent of "rescuable" searches as a fraction of the total number of failed subject searches (a "rescuable" search is a failed subject search for which a bridge or link can be found in the thesaurus between one or more of the subject terms appearing in the end-user's original search argument and the vocabulary of subject headings employed in the catalog).

PROGRESS TO DATE

The researcher has obtained a substantial body of unique data from Professors Tefko Saracevic and Nicholas Belkin, collected for an earlier research project. These data include OPAC session transaction logs with full capture both of user query and of system response, online questionnaires completed by users, direct observation of patron behavior by research assistants, and interviews with patrons. The present researcher has designed and has had written specially enhanced software to replay the OPAC session transaction logs in a manner that very closely simulates the original session. The researcher has identified one machine-readable database (LCSH on CD-ROM — not strictly speaking a thesaurus, but still a usable syndetic apparatus) as well as a second, back-up tool (the WordNet thesaurus, at Princeton University) to be used for estimating the potential effectiveness of an integrated thesaurus. The summer of 1994 will be dedicated to analysis and evaluation of these materials. The researcher expects to have reportable results by the Fall, in time for presentation at the SIG/CR Classification Research Workshop.

ACKNOWLEDGMENTS

This research is being pursued as the dissertation in partial fulfillment of the requirements for the degree of Ph.D. in Communication, Information and Library Studies at Rutgers University, New Brunswick, NJ. The assistance of the author's committee is acknowledged with thanks: Drs. James D. Anderson, chair; Nicholas J. Belkin; Kathleen M. Burnett; and Jessica L. Milstead. The generous programming assistance of Shuyuan Allan Zhao is also recognized.

ISSN: 2324-9773

Classification of Management Information: A Perspective from Recognition and Their Semantics

Deh-Min Wu
Department of Information Science
Technical University of Berlin
Institute of Open Communications Systems
Hardenbergplatz 2, D-10623
Berlin, FR. Germany
wu@fokus.gmd.de

This paper proposes an approach to classify network management information so that the awareness of the network states and events can be maintained in an open environment. This classification concept is different from methods like neural network using numerical weight- functions, and applies situation classification capability as found in KL-ONE and Description Logic in order to recognize the situation description for network management. This approach is applied to semantically mediate information from networks.

1. CLASSIFICATION OF MANAGEMENT INFORMATION

Today integrated networks cause management information overhead to increase and hence a global common understanding of the state of affairs of the network is more difficult to achieve than before. One example of this sort of problem is to semantically filter and mediate the network state and events. Our classification approach applies term classification facilities [3,4] as those found in the knowledge representation language KL-ONE and its successors, such as description logic. *Classification of information* is applied to determine whether collected informations possesses particular characteristics so that mediation of information corresponding to these properties can be carried out.

For distributed applications like classification of management information in a network, situation theory based on cognitive science [1,2] is applied to model a description of a network state and event. A situation is applied as the discrete information, which is modeled as a set of *infons*. An infon is a unit information item represented as a tuple $f_i = <r_i, o_{i1}, o_{i2},\ldots, o_{in}, p_i>$, where r_i is a relationship, o_{ij} are concepts and p_i is the polarity of the infon. If the polarity is *1*, then the infon is a positive state of affairs with relation to network management. The value *0* identifies it as negative. A situation is represented as a set of infons $\{f_1, f_2,\ldots, f_m\}$.

Situations are exchanged by autonomous agents to describe the network states and events so that decisions for recovering a network can be made. Any relations and concepts used in a situation are inductively defined information to ensure a consistent interpretation. In the conventional entity-relationship model [5], the relationship is not represented as an entity. We have to introduce relationship classes with the following ASN.1 definition:

ISSN: 2324-9773

<relationship-class-label> RELATIONSHIP CLASS
[DERIVED FROM <relationship-class-label>
[,<relationship-class-label>];]*
[BEHAVIOUR <behavior-definition-label>
[<behavior-definition-label>];]*
[ROLE <role-label> roleproperties
*[REGISTERED AS object-identifier];] **
REGISTERED AS object-identifier;

In this proposal an n-ary relationship is represented by many roles which are defined as binary relationships. The integrated classification of relationships and concepts simplifies the subsumption propositions. Such an approach is suitable to define new specific concepts by more generic primitive concepts and roles like those in [3,4]. In a comparison of term-classification facilities defined in [3], the situation classification allows the following two subsumption (denoted as $\leq$) rules.

Proposition 1: For infons f and f', $f' \leq f$ iff $f' = \langle a_1, a_2, \ldots, f, \ldots, a_n \rangle$

Proposition 2: For infons $f=\langle a_1,a_2,\ldots,a_n,a_{n+1},\ldots,a_m \rangle$ and $f'=\langle a_1',a_2',\ldots,a_{n'},s \rangle$, and a situation $s=\{f_1, \ldots, f_k\}$, $f' \leq f$ iff both of the following are satisfied:
—$\forall\ a_i, a_i' \leq a_i$, where $1 / i / n$
—$\forall a_i, n+1 \leq i \leq m, \exists f_j \in s, f_j = \langle \ldots, b_l, \ldots \rangle, b_l \leq a_i$, where $j = 1,\ldots,k$

The above propositions indicate that a more specific situation f' is subsumed by a more general one f if f is a descriptive part within the situation f'. A subsumption test checks recursively with these propositions in order to verify whether a partial situation is defined within embedded situational information. As an example to illustrate the application of these propositions, suppose situations A, B and C are defined as:

Situation A: *"A line has cable failure at 20:00"* $\Leftrightarrow$ *<has, line, cable failure,* l_a, *20:00, 1>*

Situation B: *Situation A and <observes,* C_i*::Controller, situation A,* l_c*, 20:00, 1>*

Situation C: *<has, connection, <atleast, 1, failures>,* l_b*, night, 1> and <observes, *::Controller, *::failure,* l_c*, 20:00, 1>*

We can prove that all infons in C subsumes at least one infon in B with these preconditions: *line* $\leq$ *connection, cable failure* $\leq$ *failure and 20:00* $\leq$ *night.* Applying proposition 2, situation C subsumes situation B.

Our classification is based on information structure and semantics of concepts and relationships. The structure of information defined to represent a network situation is similar to the relational model in term of representation power. In such a model the relationships between network elements are specified by the participants associated in a relationship. Our approach differs from

ISSN: 2324-9773

the relational model in the representation of a situation. We model it as a description. A situation description may contain an individual or a variable within it.

2. CONCLUSION

The objective of the proposal is to mediate information semantically, based on embedded situational information inductively defined by more general information. The classification facilities determine the situation types and their corresponding semantics. Each element in a situation is translated to its most generic definition and then these components are compared. This characteristic allows the information exchange to be achieved with a common understanding on most generic, primitive definitions, but with no claim on the specific classes. The classification of the situation description enables us to obtain the desired awareness of a network situation.

REFERENCES

[1] Jon Barwise, "The Situation in Logic", *CSLI, Lecture Notes* 17, 1989

[2] Devlin, Keith J., *Logic and Information,* Cambridge University Press, 1991

[3] Bernhard Nebel, "Reasoning and Revision in Hybrid Representation Systems", *Lecture Notes in Artificial Intelligence* 422, Edited by J. Siekmann, Springer-Verlag, pp. 73-102, 1990

[4] Thomas Hoppe, Carsten Kindermann, J. Joachim Quantz, Albrecht Schmiedel and Martin Fischer, *BACK V5: Tutorial & Manual,* Technical University of Berlin, Project KIT-BACK, March 1993

[5] Peter Pin-Shan Chen, "The Entity-Relationship Model - Toward a Unified View of Data", *ACM Transactions on Database Systems,* Vol. 1, No. 1, March 1976, pp. 9-36

ISSN: 2324-9773

ISSN: 2324-9773