Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

Computer-Aided Knowledge Engineering for Corporate Information Retrieval

Marlene Rockmore

Digital Equipment Corporation, Maynard, MA 01754, USA

ABSTRACT

In 1987, Digital Equipment Corporation's internal Market Information Services Group / Information Access Services (IAS) decided to build a single thesaurus system to support production and retrieval of multiple applications. This system TIMS (Thesaurus / Indexing Management System) had to be dynamic and allow for easy modification and merging of volatile business terminology. A faceted approach was used for knowledge-base building and semantic representation. The system allowed the knowledge engineer to determine a classification structure and to develop relation types suited to a specific application's requirements.

1. INTRODUCTION

The value of classification has been an open question. Croft has suggested that classification is not cost-effective when compared to other techniques, [1] while other studies such as comparison of PLEXUS, a knowledge-base driven application, to INSTRUCT, a statistically based system, also showed that the knowledge based approach, using thesauri as a knowledge base structure, performed consistently better in recall and precision than the statistical approach. [2]

One of the assumptions applied in the design and development of the system is that the goal of an information retrieval (IR) system is to help users find appropriate information to solve a problem. Although, ultimately, the IR system will transform a query into its most atomic unit — terms — the IR system needs to understand some of the intentions behind the terms. The architecture of an IR system, then, includes components to represent and store documents and components that represent or store problem descriptions or queries. Attempts to build systems that contain comprehensive descriptions of problems/queries as well as documents include the ASK system, as described by Belkin. Belkin argues, "Representation of the user's query was as critical to the system as the representation of the text." [3] Indexing and classification have typically provided the link between the query representation and document representation.

Automated approaches to query analysis such as latent semantic analysis techniques require that data be available in electronic format. These approaches, while showing some ability to extract noun phrases for constrained domains, have not been able to analyze other functions of language such as ambiguities or semantic categorization. [4] Semantic and pragmatic analysis typically depends on subjective knowledge of the user, specific conditions of the user's environment and domain, and general world and pragmatic knowledge. Semantic knowledge is not easily captured by automated techniques or traditional thesaurus construction techniques that depend on analysis of texts within constrained domains.

The ability to capture and analyze the intention of the user's terms has proved to be very important to maintaining a high degree of reliable retrieval. Data collected from Digital Equipment Corporation's internal Competitive Information System have shown that using

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

thesauri can improve reliability of the system to return documents that match the user's search terms. This system provides current information on volatile business subjects. Following implementation of the machine-assisted thesaurus/indexing component, TIMS, document throughput rose 300% with no increase in staff. Yet, as the size of the database grew, there was no degradation in index performance, which averaged 78-80%.

We attributed this consistency in performance to features of the system which allowed for continual design and maintenance of the vocabulary. These features are:

- Machine-Assisted Thesaurus Construction — This component allows the structuring of thesaurus relation types (hierarchies, synonyms, related terms) within facets. The thesaurus update mechanism supported automated thesaurus update processes such as reciprocations. Relations types within a facet were dynamically defined when the facet was defined.
- Machine-Aided Indexing Interface — This module connected the faceted thesaurus to indexing providing interactive access to the vocabulary. Indexers were encouraged to update the vocabulary through a semiautomatic queue file for managing candidate terminology.[5]
- Candidate Term Management — Indexers were encouraged to add vocabulary as they were automatically notified when a term did not exist in the vocabulary. They could then choose whether to add a term to a candidate term queue file or whether to search for another term that existed in the vocabulary. When indexers chose to add a term, they were also asked to suggest a semantic category (facet) and to flag any ambiguity about the term by using the term relation structures.

These tools assisted in the analytical processes used to build vocabulary and manage indexing to a support high rate of retrieval. We believed tools that allowed dynamic modification to the vocabulary were critical to system performance. However, successful use of the tools depended on application of methodology that allowed good description of the particular business environment, its needs, and its concepts. This methodology is described below.

2. METHODOLOGY

Methodology to build and maintain a thesaurus-based application then consisted of two phases:

- An initial analysis and design was guided by models of the business environment and the decision-making process.
- Maintenance process that allowed dynamic modification to the knowledge components above.

The initial design has three objectives:

- To understand the business environment and the user's role and needs within that environment.

ISSN: 2324-9773

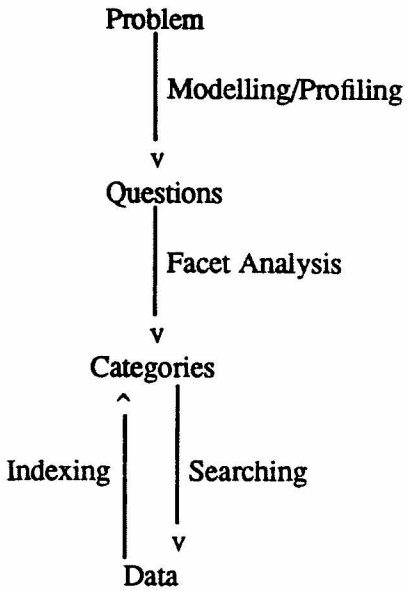
Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

- To understand specific categories or facets by which users look for information and specific sources that mapped to these categories.
- To extract specific vocabulary.

2.1 FROM DATA TO DECISIONS

IR system users in the corporate environment seek information to make decisions. These decisions address questions that can range from "Where do we locate a new plant?" to "What software tools should I use to do X?" These decisions are made to solve business problems, and how these decisions are made depends on factors including corporate vision and goals as well as accessibility to the right information at the right time in the right format.*

A business problem calling for a decision can be broken into a set of questions. Each question represents a smaller set of information that needs to be collected and analyzed. From this analytical process, broad topics (categories) for which information will be needed are identified. These topics are then matched to data or documents. Matching of data to topics is done through the indexing process. The facet analysis process is used to create the broad topics or categorization scheme for classifying the data. The steps towards building the categorization scheme are included in the steps in the process of developing the IR system, enumerated later on. The part of the decision model from decision need (problem) to data may be diagramed as follows:



2.2 STEPS IN DEVELOPING THE SYSTEM

Thus, the immediate problem to solve in building a faceted classification scheme is the creation of information categories. To do this, one needs to gather and analyze information about

* I would like to acknowledge Gil Press of Digital Equipment Corporation, who has done extensive work on modelling the process of data to decisions in the business environment. We have had many conversations on how the TIMS system fits into this framework.

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

the organization's business, information needs, and requirements. This process comprises phase 1 of our application as mentioned earlier, and consists of the following steps:

1. Understand the business environment.
2. Capture specific queries.
3. Analyze queries into semantic categories (begin facet design).
4. Map the relations between terms (continue facet design).
5. Map relations between facets and develop cross-connections (complete facet design).
6. Identify sources for additional terminology.
7. Determine an editorial policy for standardized format of terms.
8. Build and implement the faceted thesaurus system by evaluating the application.

3. APPLYING THE METHODOLOGY: EXAMPLE

This section uses a scenario to illustrate and further explain the steps in our methodology.

3.1 UNDERSTANDING NEEDS AND CAPTURING QUERIES

Information categories created later on will depend on the type of business or application. for this scenario, as a result of background research (step 1) we identified the software engineering need. As a method to capture queries (step 2), we then conducted group interview sessions with software engineers where we encouraged dynamic interplays between participants. We selected the following query resulting from these sessions: "What software tools can I use to do X?" Obviously, more information is needed in order for the software engineer to make a judicious selection of tools. Therefore, further queries might be: "What existing applications do X?" and "What existing technologies were used to build X?" Additionally, it might be helpful to know about new technologies and associated learning curves. It would also be important to know hardware or operating system environment requirements as well as personal goals such as skills one would like to develop.

3.2 SEMANTIC CATEGORIZATION

Using the above queries, step 3 was applied, creating a semantic categorization scheme resembling the needs of software engineers and functioning as access points to information in the ir system. This facet analysis resulted in the following set of categories (facets):

- Software Products and Tools
- Hardware

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

- Operating Systems
- Applications
- New Technologies
- Personal Skills Inventory

3.3 MAPPING RELATIONS WITHIN FACETS

To continue the thesaurus design, we would map relations between terms within facets (step 4). Each of these categories has a specific terminology, and associated relations between terms. In our application we use thesaurus relations (broader-than, narrower-than, associative, and equivalent) to express relations between terms in a category. In one facet, Software Products and Tools, all thesaurus relations might be used. In another facet, Personal Skills, only equivalent terms (e.g., synonyms) might be required. There are other conditions that might be set. For example, the Software Products and Tools facet may be a preexisting public vocabulary while the Personal Skills inventory might be considered private and only available to a specific user.

3.3.1 DISAMBIGUATING TERMINOLOGY

The problem of ambiguity is complex. In typical thesaurus construction, ambiguities are flagged with parenthetical modifiers, for this example: *availability (system status)* versus *availability (product)*. In our system design, we handle two types of ambiguity structurally, through semantic classification and term relations, thereby eliminating the need for parenthetical qualifiers. Even though the same term might appear in different places in the classification, documents would nevertheless be described unambiguously for precise retrieval, using word order, format, and the classification to represent the user's context or possibly anomalous state of knowledge.

The first type of ambiguity we called conceptual ambiguity. A conceptual ambiguity in our system is a word or phrase that can be resolved by classifying the term in the appropriate semantic category. For example, a conceptual ambiguity might be the term *availability*. A software engineer is typically interested in system availability, that is, whether a system is up. Thus we might classify *availability* as a term associated with the software engineering function. However, in the context of our sample query, "How do I order Product X?", the term *availability* refers to product availability, which is an attribute of a software product or tool, which is a sales and marketing function. Thus *availability* would be classified as product availability and may ultimately be classified in a sales domain.

The second type of ambiguity is a homograph which occurs within the same facet. In this case, the ambiguity cannot be resolved by the appropriate classification of the term into a category. A common example of this second type of ambiguity are product names, where the product name is a number such as 3000. In this case, we created a specific relation type within the facet, which we called a USA. Integrity checks in the Thesaurus Construction Module ensured that a synonym that was a homograph to a preexisting synonym in the same facet was tagged with this relation type.

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

3.4 MAPPING RELATIONS BETWEEN FACETS TO BUILD INFERENCES

To continue our thesaurus design, we created explicit relations between categories (step 5). There are also explicit relations between categories. In the example above, there is a relation between software products and tools and the operating systems on which these products can run. This relation type was called a cross-connection and provided a mechanism for establishing inferences between categories. [6] The relation type label of the cross-connection was application-definable. To implement the cross-connection to support the example above in TIMS would require the following steps:

- Define a facet for Operating Systems.
- Define a facet for Software Products and Tools.
- Add a relation *RO* (Runs-On) to the Software Product and Tools facet. This builds the linkage between specific products and operating systems that they run on.

The RO operator then appeared in the Thesaurus Construction Module and could be updated with any appropriate values.

3.5 ADDITIONAL TERMINOLOGY AND EDITORIAL POLICY

Having provided functionally-rich structuring of facets and thesauri for domain-specific content in order to improve quality of indexing and retrieval, next we identified sources for additional terminology (step 6), such as published thesauri, tables of contents and back of book indexes, appropriate trade literature and business reports, and subject experts. [7,8] Again, group interview sessions provided a good source for terminology, including jargon.

The terminology then was classified into the appropriate category; the relation types (lead term, broad term, narrow term, related term, etc.) were structured, and the terms were entered into the thesaurus following an editorial policy (step 7). This editorial policy adapted the guidelines published by NISO (National Information Standards Organization) to the specific organization or application. It covered general syntactical rules about the format of terms such as capitilization or preferred form (noun vs. verb). While published standards provided general guidelines for constructing thesaurus relations within the facet, the knowledge engineer is responsible for understanding the impact of specific relation types on retrieval and for ensuring effectiveness of the underlying thesaurus on enduser retrieval and index performance. A method for evaluating the effectiveness of the design will be discussed later on.

3.6 IMPLEMENTATION AND MAINTENANCE

The final step in the first phase of development (step 8) is to test the thesaurus against the database by indexing a sample set of documents and retrieving these documents. Initially the thesaurus should retrieve at 65-75%, depending upon the depth to which the users' vocabulary was initially analyzed and the critical mass of content available in the database. Over time, as the content grows, and the vocabulary is updated and matures, index performance should measure 80%. This growth curve has been described by Lancaster. [9] Conversely, retrieval will degrade if the vocabulary is not maintained. This evaluation would complete the first phase of

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

development. An actual evaluation is presented in a subsequent section of this paper.

The second phase, vocabulary maintenance allowing dynamic modification, involves the following processes:

- Capture of input terms from user interface.
- Encouraging indexers to update terminology through the candidate term system.
- Ongoing analysis and evaluation of corporate information retrieval needs and user need.
- Use of flexible, easy to use software tools for defining new categories and merging specific terms and their relations.

4. AN EVALUATION OF THE THESAURUS-BASED APPLICATION

The enduser application, described earlier, offered a unique system for information retrieval experimentation because this system has two retrieval interfaces, free text and thesaurus-controlled, accessing similar types of documents.

The free text portion of the system is updated as needed with brief articles from newswire feeds. This portion of the database is searched using free text techniques. During a typical week, it contains about 500 articles. Articles are purged every two weeks. The thesaurus-controlled section of the database contains about 12,000 documents. About 100 new documents are added each week. Articles in this portion of the database were varied in length, style, and content. Interfaces to each module were similar. Usage of both modules averaged about 1000 sessions per month.

Every session was logged in a session file written by an application program, utilizing Digital's VAX VTX/VALU product. Session statistics included when the session was connected, how much of the menu structure was navigated during the session, whether the user chose to mail an article or read it online, and search terms. Each week, a report was generated that showed which terms the user entered, the field in which the user entered the term, whether that field was controlled by a thesaurus, and whether that term was found in the inverted index file. We review these reports for a two-year period. The free text module had a consistent rate of 40% index performance; that is, only two of every five searches successfully returned a match on the search term.

In the second module, which used the thesaurus-controlled search, there was a index performance of over 80%; that is, a search had a successful match 8 of every 10 times. However, if terminology was not updated quickly in the thesaurus, this rate fell to 70th percentile. We found that when new terms and relations were added, retrieval improved again to the 80th percentile. Without ongoing modification to the vocabulary, retrieval performance would degrade.

This measure gave us a way to compare the robustness of index performance between different indexing methods, as we knew whether the term had been added to the index file from the thesaurus or whether the term had been added automatically by free-text. Since the database content had been expertly selected and acquired according to the needs of users in our

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

environment, it could be assumed that all content in the database was potentially relevant, and therefore, was likely to be highly relevant to these needs in general. The measure of index performance described here may have some significant implications towards the more difficult measure of recall as an assessment of relevance. Reliability is a very important factor in a business setting. Users need to feel confident that when they use a system they will get information (reliability) and that, of course, the information will be relevant.

5. CONCLUSIONS

Past research, while concluding that adding domain-specific knowledge to a system improved retrieval, has failed to look closely at the relationship between methodology and tools. Our data show that good methodology and analysis using a thesaurus-based system will outperform comparable statistical or boolean systems.

Our current research and development efforts have focused on programs and rule bases that assist in automating enduser navigation of thesaurus structures. In order for endusers to benefit from these structures, they need to be designed and built with an understanding of complexity and depth of users' information access needs. This means that there need to be knowledge engineers who can skillfully use a new generation of tools for the computer-aided design of complex and volatile corporate information retrieval systems.

Currently, there are few commercially-available systems for computer-aided design assistance for knowledge engineering. This reflects on a more significant issue. There is a growing need to attract and retain qualified knowledge engineers who understand the business analysis components and system components of a corporate information retrieval system. The future of the field of computer-aided knowledge engineering for corporate information retrieval depends in the near term on a convergence of perceptions. There needs to be a market for these systems, and this market needs to value the expertise knowledge engineers provide; there need to be tools built to assist the knowledge engineer; and last, there need to be professional opportunities and growth that attract and encourage talented people to this field.

NOTES

1. Croft, W. Bruce. "Is Classification Necessary for Retrieval." *Proceedings of the 44th Annual Meeting of the American Society for Information Science*, Vol. 18, 1981.

2. Wade, Stephen. "A comparison of knowledge-base and statistically-based techniques for reference retrieval." *Online Review*, 1988, Vol. 12, No. 2, pp. 91-105.

3. Belkin, Nicholas J. "Anomalous States of Knowledge as a Basis for Information Retrieval." *Canadian Journal of Information Science*, 1980, Vol. 5, pp. 133-143.

4. Streeter, Lynn. "Who Knows: A System Based on Automatic Representation of Semantic Structure." *User Oriented Content-based Text and Image Handling; RIAO 88*, Vol. 1, pp. 379-388.

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

5. Anderson, James D. "Information Organization based on Textual Analysis (IOTA): Instructional Programs for Database Design." *The Library Microcomputer Environment: Management Issues* (Sheila S. Intner and Jane Anne Hannigan, eds.). Phoenix: Oryx Press, 1988.

6. Humphrey, Susanne M. "Interactive knowledge-based indexing: The MedIndEx System," *User Oriented Content-based Text and Image Handling; RIAO 88*, Vol. 2, pp. 883-898.

7. Aitchison, Jean and Alan Gilchrist. *Thesaurus Construction: a Practical Manual.* 2nd. ed. London: Aslib, 1987.

8. Batty, David. "Thesaurus Construction and Maintenance." *Database*, February 1989, Vol. 12, No. 1, pp. 13-20.

9. Lancaster, F.W. *Vocabulary Control for Information Retrieval.* 2d ed. Arlington, VA: Information Resources Press, 1986.

ISSN: 2324-9773

Rockmore, M. (1990). Computer-aided knowledge engineering for corporate information retrieval. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 139-148. doi: 10.7152/acro.v1i1.12473

ISSN: 2324-9773