Abstract: Different Knowledge Management Software (KMS) suites approach information retrieval in different ways. Some of them have in-built knowledge structures, and some of those structures may be called a "thesaurus". But they are not thesauri as defined by the standard ISO 2788. This paper describes the types of knowledge structure which are provided for in the Verity search engine, namely "Thesaurus" and "Topics". An examination of how they are constructed shows strengths and weaknesses, and their potential to be applied to automatic categorization of search results. Unfortunately it is difficult to quantify the performance benefits of any of these knowledge structures, or even compare them with the benefits of using the traditional techniques of vocabulary control. However, the problems of retrieving information on many intranets has led some practitioners to apply simple vocabulary control to complement the functionality of their chosen KMS.

1. Introduction

What are we to make of the astonishing proliferation of different Knowledge Management Software (KMS) products all being promoted as the ultimate solution for information retrieval on a corporate intranet? How do they compare with our traditional controlled vocabulary techniques for organizing information? And how to choose between them, while we have no consensus on quantitative, practicable techniques for comparing information retrieval performance? This paper seeks to explore one tiny piece of those large imponderables, by looking at the role of knowledge structures within some different products. We will take brief stock of the range of approaches on offer, then examine the offerings of one KMS vendor in more detail, before comparing these with the traditional thesaurus technique.

2. Different Approaches to the Use of Knowledge Structures

KMS products have one thing in common: they all offer a way of retrieving information automatically, without need of human intervention to index and/or classify individual incoming documents. Other than that, the range of approaches is vast. At one extreme are the artificial intelligence solutions, in which the computer does it all without reference to established knowledge structures. At the other we have software suites relying heavily on hierarchical structures just as elaborate as the traditional thesaurus.

Take Autonomy, for example. Its Knowledge Server product has no built-in vocabulary, but relies on pattern matching technology. According to company brochures, "It understands the real meaning of words in their correct context." A user keys in his question in his own language; the "Dynamic Reasoning Engine" analyses the query and compares it with the stored "concept maps" of documents in the system. Those which give a sufficiently close match are returned. There is no reference to any kind of stored knowledge base.

In contrast, Excalibur's RetrievalWare relies on an extensive "Semantic network". A dictionary database of concepts is backed up by syntactic analysis capability, linguistic morphology rules and a database of irregular words. The standard dictionary supplied with the product is built up from published sources including Roget's Electronic Thesaurus and the American Heritage Dictionary. And it can be supplemented with a custom dictionary incorporating the customer's own terminology. Each entry in the dictionary corresponds to a
concept, as described by a dictionary definition. The concept entries specify part of speech, synonyms and near-synonyms. Synonyms are tagged with a weighting factor, linked to the degree of synonymy, which controls how queries will be expanded. To summarize, RetrievalWare depends on a sophisticated, customizable knowledge database incorporating semantic links, coupled with syntactic and morphological processing capability.

Another KMS vendor is Dialog. Its recent acquisition of Muscat has allowed it to apply at least two different approaches within its KMS suite “k-working”. In the first step “Linguistic Inference” uses probabilistic techniques to develop “Concept Maps” of document sets and compare these with users’ interest profiles. The relative frequencies and co-occurrences of words within the document set are important, but the matching process does not depend on any manually prepared dictionary, thesaurus or topic tree. Next, however, the textual output of the process may be passed to the “InfoSort” component for categorization. InfoSort classifies documents automatically using rules for assigning documents to categories pre-established by experts. Rule-bases may be developed manually, or generated automatically by analysing a provided set of correctly classified documents. A fully developed InfoSort rule-base certainly qualifies as a knowledge structure. It bears quite a strong resemblance to the topic trees used by Verity, which will be described in more detail below.

Rather than catalogue all the ways that knowledge structures are used in the many different KMS products, we will now examine in more detail the two (or arguably three) types of knowledge structure supported by one of the leading vendors, Verity.

3. “Thesaurus” as Understood by Verity

Verity’s search interface expects a user to express his query using keywords, optionally linked with operators such as Boolean operators. Its standard products “Information Server” and “Developer’s Kit” incorporate an in-built thesaurus, comprised mostly of synonym sets. If a user of the search engine chooses to apply the thesaurus operator, his search terms are expanded by looking up the thesaurus and “ORing” the corresponding synonyms on to the original search term(s). Thus if a search for “flow” retrieves 60 documents, a search for “<thesaurus>flow” would be expanded to look for “well OR course OR flow OR rush OR stream OR pour OR surge OR gush” and probably retrieve far more documents. In other words, the thesaurus facility is intended to enhance Recall.

Enhancing Recall can sometimes be very useful. But the loss of Precision can be unwelcome in the common case where a search retrieves too many unwanted documents. The problem is shown up in Figure 1, giving a brief extract from the “circular lists” of the in-built thesaurus.

| list: "repel,parry,repulse,keep off,beat off,ward off,fend off" |
| list: "stray,depart,err,swerve,deviate,diverge,digress" |
| list: "finish,kill,murder,destroy,slay,liquidate,put away" |
| list: "well,course,flow,rush,stream,pour,surge,gush" |

Figure 1: Extract from circular lists of Verity’s in-built thesaurus

From Figure 1 it is easy to see that this thesaurus is very different from the typical information retrieval thesaurus built in conformity with ISO 2788 or ANSI Z39.19. There are no BT, NT or RT relationships. Instead of USE/UF relationships, there are synonym sets or rings, in which every term has equal status. The default thesaurus supplied by Verity includes lists of verbs and adverbs as well as nouns and noun phrases. A stemming algorithm is also applied,
so the search above will pick up "flowed", "flowing", "flows", etc., as well as "courses", "rushes", and all the other terms with their stemming variants.

The Verity thesaurus bears some resemblance to Roget's Thesaurus, in which the terms listed under one concept heading may be synonyms in certain contexts, but need not be synonymous in all contexts. If one applies a Roget-type thesaurus to information retrieval with simple OR logic, one must not be surprised to find the net is cast very wide, and the retrieval results typically include far more false drops than relevant documents.

But what if we discard the in-built thesaurus and substitute a home-made one comprised of really close synonyms? Verity offers the option of applying a "Custom Thesaurus". The author has experimented with this approach, using entries such as those in Figure 2:

```
list: "BSE,B S E,bovine spongiform encephalopathy,mad cow disease"
list: "cjd,c j d,creutzfeldt jakob disease,creutzfeldt jakob s disease,wcjd,ncjd"
list: "abattoir,abbattoir,abbatoir,slaughter house,slaughterhouse"
```

**Figure 2: Some circular list entries from a Verity-style Custom Thesaurus**

The examples in Figure 2 are based on nouns and noun phrases. They allow for acronyms/expansions, common/scientific name equivalents, British/American variants, and other spelling variants including common misspellings. With this thesaurus the user can key in a short and simple abbreviation such as BSE or CID, and retrieve all the applicable references. Gone are the problems of entering some long scientific names, and spelling them correctly.

The Custom Thesaurus also includes some entries in the format which Verity calls "asymmetric lists". Examples are shown in Figure 3.

```
list: "gin<or>whisky<or>whiskey<or>rum<or>beer<or>wine"
(keys="alcoholic beverages"
list: "cereal<or>wheat<or>oats<or>maize<or>rye<or>barley"
(keys="cereals"
list: "pea<or>cowpea<or>chickpea<or>vigna<or>vicia faba<or>broad bean"
(keys="legumes"
```

**Figure 3: Some asymmetric list entries from a Verity-style Custom Thesaurus**

These operate differently from the "circular lists" described above, in which all terms have equal status. In the asymmetric lists, only one term can be used to invoke query expansion, namely the term at the end of each entry. Thus a search for `<thesaurus>"whisky"` retrieves exactly the same as a normal search for "whisky". In contrast, a search for `<thesaurus>"alcoholic beverages"` is expanded to retrieve all the documents which mention gin, whisky, rum, etc. There is a parallel here with the BT/NT relationship of a standard thesaurus. And plainly this is a useful facility which saves the searcher keying in long strings of the narrower terms ORed together. (The examples above have been abbreviated for readability. In the live thesaurus, entries are much longer.)

To summarize, the Verity Thesaurus facility admits two types of entry, namely circular and asymmetric. The former enables query expansion to retrieve synonyms; the latter enables query expansion to retrieve narrower terms. Both have the potential to improve Recall but not to improve Precision. Both depend on the advance preparation of knowledge structures in the formats illustrated above.
4. "Topics" as Implemented by Verity

Verity provides a facility for in-built knowledge structures called "Topics", which are much more elaborate than the "Thesaurus" lists. What is a Topic? According to one of Verity's user guides, "A Topic is a grouping of information related to a concept or a subject area. Topics provide a convenient means by which you can encapsulate knowledge and make it available to end users as a shared resource." According to some recent Verity sales literature, Topics are "...customizable rules of evidence linking documents to concepts. A Verity Topic Set enables organizations to build families of stored persistent queries describing an entire knowledge domain." I like to think of a Topic Set as a nested set of complex search statements. The nesting allows you to choose the level of specificity for the search.

Figure 4: Illustration of a Verity-style Topic Set

Figure 4 shows a fictitious Topic Set for "Petcare". Our view of it has been expanded in some areas but not in others, so we can get the flavour of it but not see its full complexity. It takes the form of a tree, with the concept of petcare at its root. Each Topic in the tree has a name, such as "Pets", "Cats", "Dogs" etc, followed by a search operator in angle brackets, e.g. <And>, <Any>, <Accrue>, etc. The Topics at the next level below the root concept are "Pets", "Pet-nutrition", "Pet-behaviour", etc. "Pets" is in turn subdivided into "Cats", "Dogs", etc. Any number of levels of subdivision is possible. Optionally, the nodes and terms in the tree may be weighted.

If a searcher enters a term which is the name of a node, e.g. "Dogs", then his query will automatically be expanded to include all the subdivisions of that node, with a syntax controlled by the corresponding operator. Thus, "Dogs" will be expanded to search for any of the terms dog, dogs, hounds (or stemming variants such as hound), bitches (or bitch), puppies, puppy, pups, pup, or any of a vast number of breed names which reside under the node named
"dog-breeds". The <Accrue> operator ensures that the more of these different terms are found in a document, the higher it will be ranked among the search results. A search for "Pets" would retrieve all the cats and birds as well as the dogs. This is the nesting of search statements within broader search statements, or sub-topics within topics.

Under "nutrient-interaction-topic" we see some more sophisticated search strategies. A search for "nutrient-interaction" will retrieve the phrase, "nutrient interaction" or its plural or verbal forms. A search for "nutrient-interactions" will retrieve far more: not just the word "nutrient" within 7 words of the word "interaction", but also a similar collocation of "interaction" with any one of a vast number of proteins, carbohydrates, vitamins, minerals, etc., which we would be able to see if we had room to open up the "macronutrients" and micronutrients" boxes.

5. Verity "Thesaurus" versus "Topics"

The two facilities have a lot in common. Both can look for synonyms. Both can expand a term to look for narrower terms. But there are important differences. For the user, the search syntax is quite different. How to train users, and how to let them know what is in the knowledge structures behind the scenes, require different solutions for the two facilities. "Topics" allow much more sophisticated search strategies, addressing Precision as well as Recall. Search terms can be linked with a wide range of operators, including Boolean, proximity, and others such as <wildcard>, <case> and <stem>. Furthermore, Topics allow the weighting of search terms, affecting the ranking of documents retrieved. But Topics do not allow all members of a synonym set to act as search keys - only the one chosen as the "name" of the Topic will retrieve all the synonyms. And despite the provision within Topics for complex search statements, these still suffer from all the problems well known to searchers of full text, which may be summarized as the trade-off between Recall and Precision.

6. Verity "Topics" versus a Standard Thesaurus

By "standard thesaurus" I mean one that conforms broadly with the International Standard ISO 2788, "Guidelines for the establishment and development of monolingual thesauri" (International Organization for Standardization), or the corresponding American standard ANSI Z39.19 (National Information Standards Organization). Since the standards offer guidelines rather than a mandatory prescription, there is a great deal of variety among thesauri which follow them. But some common features are the inclusion of three types of a priori relationships between terms (namely equivalence, hierarchical and associative relationships), the presence of preferred terms with non-preferred terms to point to them, and the underlying assumption that the thesaurus will be used for purposes of vocabulary control.

Superficially, a Topic Tree is a knowledge base resembling some well-known thesauri such as MeSH (National Library of Medicine) or EMTREE (Elsevier Science), which use the hierarchical relationships between terms to present the latter in well developed tree structures. The tree structures are particularly helpful in allowing users to adjust the level of specificity of their searches, and it could be argued that Topic Trees have the same property.

After that the similarities end. A Topic shows no equivalence or associative relationships. Sibling terms may be roughly synonymous (e.g. "dogs" and "hounds") or they may correspond to an a posteriori association deemed useful for searching (e.g. "nutrient" and "interact*"). In a Topic there is no such thing as a "preferred term", because there is no intention to impose vocabulary control.

In fact the key difference is functional - the presence or absence of vocabulary control. With a standard thesaurus, the idea is that a human indexer will pick out the key concepts in a document and assign the corresponding terms from the controlled vocabulary. A searcher looking for any combination of those same concepts should find the relevant documents
because he is guided by the thesaurus to use the very same terms for searching. But with a
free text search, no-one has picked out the key concepts in documents. If a relevant Topic has
been established, it can help a searcher by looking for all the various combinations of terms an
author might have used to express a given concept. Terms will be detected whether they
represent the focus of a document or just a peripheral mention, and whether or not they have
the desired meaning. In summary, a standard thesaurus is designed to be used both on input
of documents and on searching; a Topic is designed to be used only for searching.

7. Knowledge Structures within “Knowledge Organizer”

So far we have considered Topics only for direct use in formulating searches, but there
is an additional application in Verity’s recently launched product, “Knowledge Organizer”.
This builds on earlier products, adding a facility for the automatic classification of documents
retrieved in a search. The product appears to respond to a huge market interest in “corporate
taxonomies” and a classified presentation of information, as exemplified by the Internet portal
“Yahoo!”. “Knowledge Organizer” takes a rule-based approach to automatic classification
(not unlike Dialog’s Infosort), and it supports the coexistence of more than one classification
scheme. The categories in each classification scheme must be established in advance, along
with the rules for assigning retrieved documents to the categories. The rules can be based on
file system hierarchies, URLs, document metadata, or Topics.

So the same TopicSets as before may now be applied within Knowledge Organizer. The Topic names double up as headings in a taxonomy. And the rules developed for
responding to queries are now used to assign documents to headings. Let us take an example
based on the Petcare topic illustrated above. A user searches the intranet for, say, “skin
condition”. 100 Documents are retrieved, and the results are presented under the headings we
have seen above, such as “Pets”, “Pet nutrition”, “Pet behaviour”, etc. Under the “Pets”
heading, the skin condition documents will be sub-categorized under sub-headings for cats,
dogs and birds. Some headings will not appear if they have no hits. Some documents, if they
match the criteria for two different headings, will be allocated to both headings.

Categorization of the results is often more user-friendly than an undifferentiated list of 100
documents.

Returning to our comparison of Topics with a standard thesaurus, we see that only the
former has a categorization capability. After all, categorization is traditionally done, not with
thesauri, but with classification schemes. Manual classification is time-consuming and
therefore expensive, as well as subject to human inconsistencies. Automatic categorization
cuts the cost dramatically, but by Verity’s method is still subject to the problems of free-text
retrieval. The building of Topics is more of an art than a science, involving subjectivities and
the inevitable compromise between optimizing Precision and optimizing Recall.

8. Concluding Remarks

We have looked in some detail at the nature and function of knowledge structures in
the products of one popular KMS vendor. We have compared these structures with a standard
thesaurus and seen little resemblance. We have noted that some other KMS products also
incorporate knowledge structures, usually built differently but with some common features.
The key advantage of all of these products over the standard thesaurus method, even if they
require up-front effort in building and maintaining the knowledge structures, is that they avoid
the expense and delay associated with human indexing. But how effective are they, compared
with each other and with traditional techniques?

Regrettably, we do not have an easy and reliable way of comparing the performance of
one system with that of another. We cannot even prove that the application of controlled
language indexing, confers measurable benefits over free text searching (see, for example,
Svenonius, 1986; Rowley, 1994). We do know that human indexing tends to be inconsistent (Reich & Biever, 1991; Iivonen, 1990). We also know that free text searching with keywords is fraught with difficulties (Bates, 1998; Knapp, 1993, p.xvii-xxi). But we have no evidence that the various boxes of KMS tricks, sometimes backed by sales rhetoric promising the moon, perform any better. At meetings and in the literature we do hear a frequent cry, “Help - we can’t find anything on our intranet!”

In response to this cri de coeur some practitioners are adding vocabulary control to their intranet collections, even though the task sounds dauntingly large. Doran (1999), for example, describes a project to apply terms from a controlled vocabulary to document metadata. In this project, Verity happens to be the intranet search engine, and the Verity search facilities are just as available as before, but with the additional option of controlled vocabulary retrieval.

Even before the advent of intranets and KMS products, Jessica Milstead (Milstead, 1994) remarked that, “There is some research - and many anecdotes - that shows that the best retrieval situation is a mix of human indexing using a controlled vocabulary and the ability to search full text.” Providing a mix of retrieval facilities is probably still the safest course - at least until artificial intelligence research, perhaps combined with probabilistic methods, can come up with a demonstrably superior solution.

So what is the practical choice when an organization suddenly acquires a network of vast heterogeneous resources? Initially, human indexing is out of the question and one of the automated solutions is a must. Whether then to build a knowledge structure into it is a question mark, for building and maintenance absorb significant human resources. Whether to add vocabulary control via the metadata is an even bigger question mark. But for those who can afford it, there is little doubt that a mix of techniques offers the best chance of being able to find what you want.

References


Medical Subject Headings (MeSH). (annual). Bethesda, Maryland: National Library of Medicine.


