Deriving concept hierarchies from text

Mark Sanderson
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK
+44 114 22 22648
m.sanderson@sheffield.ac.uk

Bruce Croft
CIIR, Computer Science Department
University of Massachusetts
Amherst, MA, 01007, USA
+1 413 545 0463
croft@cs.umass.edu

ABSTRACT

This paper presents a means of automatically deriving a hierarchical organization of concepts from a set of documents without use of training data or standard clustering techniques. Instead, salient words and phrases extracted from the documents are organized hierarchically using a type of co-occurrence known as subsumption. The resulting structure is displayed as a series of hierarchical menus. When generated from a set of retrieved documents, a user browsing the menus is provided with a detailed overview of their content in a manner distinct from existing overview and summarization techniques. The methods used to build the structure are simple, but appear to be effective: a small-scale user study reveals that the generated hierarchy possesses properties expected of such a structure in that general terms are placed at the top levels leading to related and more specific terms below. The formation and presentation of the hierarchy is described along with the user study and some other informal evaluations.

Keywords

Concept hierarchy, subsumption, term co-occurrence, multi-document summary.

1. INTRODUCTION

The organization of a set of documents into a concept hierarchy derived automatically from the set itself is undoubtedly one goal of information retrieval. Were this goal to be achieved, the documents would be organized into a form somewhat like existing manually constructed subject hierarchies, such as the Library of Congress categories, or the Dewey Decimal system. The only difference being that the categories would be customized to the set of documents itself. For example, from a collection of media related articles, the category "Entertainment" might appear near the top level; below it, (amongst others) one might find the category "Movies", a type of entertainment; and below that, there could be the category "Actors & Actresses", an aspect of movies. As can be seen, the arrangement of the categories provides an overview of the topic structure of those articles.

The classic automated method of associating documents with each other is based on so-called *polythetic clustering [van Rijsbergen 79]* where each cluster is defined by a set of words and phrases (referred to here as terms). A document's membership of a cluster is based on its possession of a sufficient fraction of the terms that define the cluster. An example of this technique is the Scatter/Gather *[Hearst 96]* system which was applied, with some success, to whole document collections as well as to the documents retrieved in response to a query. In both cases Scatter/Gather produced an initial set of clusters each of which were re-clustered to produce a second level of more specific clusters, which themselves were reduced through a recursive process to yet more specific clusters until only individual documents remained.

Such a hierarchy of polythetic clusters has quite different properties from manually generated hierarchies. Returning to the hypothetical media example, one can see that each of the three categories is defined by a single feature (i.e. entertainment, movies, etc) which a document must possess for it to be included. These categories are in fact *monothetic clusters*: clusters where membership is based on only one feature. This alternative form of clustering has two advantages over the polythetic variety. The first is the relative ease with which one can understand the topic covered by each cluster. This can be more difficult for polythetic clusters. Their presentation to users typically takes one of two forms. Clusters can be presented through some visual metaphor, typically points in a two or three-dimensional space, where users are required to spot clusters of points and investigate the documents "behind" each point to determine the cluster topic. Such a layout, though visually arresting, has never proved to be useful. An alternative presentation involves showing a list of the terms within a cluster and a small number of key passages extracted from the documents that are its members. This presentation style also has problems. To illustrate, term lists describing four polythetic clusters are shown below. These are taken from a Scatter/Gather paper *[Hearst 96]*, and the clusters are from a set of documents retrieved in response to the query "auto car vehicle electric". Titles of cluster documents were also displayed in the paper, but are not reproduced here. Undoubtedly, one can deduce the topic of each cluster, but it is hardly an ideal form of description.

- control drive accident program office design front-wheel inventory ap track generate recall
- battery california technology mile state recharge impact official cost hour government
- import j. rate honda toyota trk light veh drop mazda percentage domestic

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Deriving concept hierarchies from text	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Advanced Research projects Agency,3701 North Fairfax Drive,Arlington,VA,22203-1714**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

- export international unit japan trade manufacturer citation german output trd news south

The second advantage of monothetic clusters is that one can guarantee that a document within a cluster will be about that cluster's topic (at least in the opinion of the person or process that did the categorizing). Such a guarantee is not possible with documents in a polythetic cluster. Although those closest to its centroid are highly likely to be about the cluster's general topic, that likelihood is less for more peripheral documents. For example, a document that is a member of the second polythetic cluster (listed above) by virtue of possessing the terms "battery, technology, recharge" is likely to be about electric cars, but for a document possessing the terms "california, mile, state, impact official cost hour", it is less clear.

Currently the commonest forms of monothetic concept hierarchies are the well-known categorization schemes such as Yahoo *[Yahoo]* and those mentioned at the beginning of this introduction. Little work has been conducted on automatically constructing a concept hierarchy educed from a set of documents. Therefore, this paper describes such an attempt. It describes the method used to build the concepts and organize them into a hierarchy. The section also contains a review of previous work in this area. Following on from this, a set of examples illustrating the structure and the general appearance of the concept hierarchy is presented. The examples are then compared to other clustering methods to highlight their differences. Next, a preliminary user experiment designed to test the properties of the structure is outlined, its results are described, and finally, conclusions are drawn.

2. BUILDING A CONCEPT HIERARCHY

The objective of this project was to automatically build a hierarchical organization of terms from a set of documents that would provide an overview of those documents. This was translated into five basic principles of design:

- terms for the hierarchy were to be extracted from the documents and had to best reflect the topics covered within them;
- their organization would be such that a parent term would refer to a more general concept than its child, in other words, the parent's concept *subsumes* the child's;
- the child would cover a related sub topic of the parent;
- forming a strict hierarchy, where every child had only one parent, was not considered to be important, therefore, the structure could be more like a directed acyclic graph;
- and finally, ambiguous terms would be expected to have separate entries in the hierarchy, one for each sense appearing in the documents.

It might be expected that a parent-child relation might also hold transitively for all the descendants of the parent, however, as pointed out by Woods *[Woods 97]*, some types of relationship between a general concept and its related more specific descendants are intransitive. Using an example from Woods, a "ship's captain" is a "profession" and "Captain Ahab" is a "ship's captain", but the relationship between "Captain Ahab" and the concept "profession" is less clear. In practice many parts of a created concept hierarchy may show transitivity, but it is unreasonable to make it a requirement.

With these principles in mind, the practicalities of building a hierarchy were addressed, starting with the means of relating terms to each other.

2.1 Discovering term relationships

Much work has been conducted in the field of locating and typing term relationships derived from text. An overview of these relationships is first presented followed by a description of the method chosen to build the hierarchy.

2.1.1 Previous work

The planned concept hierarchy was in some ways like the WordNet thesaurus *[Miller 95]*: an organization of terms with synonym, antonym, hyponym/hypernym (is-a/is-a-type-of), and meronym/holonym (has-part/is-part-of) relations. There has been some past work on automatically deriving thesaural relationships from texts. Grefenstette *[Grefenstette 94]* measured the similarity of a term's context as a means of locating synonyms. The contexts were first parsed to help normalize them and then a form of the Jaccard similarity measure was used to relate contexts to each other. Using a number of evaluation schemes, Grefenstette found the success of his method varied depending on the frequency of occurrence of the words he was analyzing. He attempted to use his derived thesaurus to aid automatic query expansion with mixed success.

Hearst *[Hearst 98]* found that certain key phrases could be an indicator of a hyponym/hypernym relation. Three of the phrases she found were

- "such as", e.g. "...popular forms of entertainment such as movies...";
- "and other", e.g. "...Robert De Niro and other actors...";
- "especially", e.g. "...most horror films, especially Psycho and The Exorcist.".

Sentences that contained these phrases were parsed to identify the noun phrases being related. Hearst discovered around ten such phrases that were accurate identifiers of the "type-of" relation. However, manual intervention was required for their discovery and the scope of the noun phrase pairs identified was limited. Hearst suggested using the key phrases to help thesaurus lexicographers search for new relations.

In later work, Grefenstette *[Grefenstette 97]* described another form of classification, where, through the use of simple syntactic analysis, he was able to place noun and verb phrases into one of nine classifications. He illustrated his ideas by examining all possible phrases containing the word "research". For example depending on whether "research" was the head or the modifier of a noun phrase, Grefenstette was able to classify types of research (e.g. market research, recent research, scientific research, etc) from research things (e.g. research project, research program, research center, etc). No tested application of this classification scheme was reported.

Woods also used phrase analysis in addition to a large knowledge base to organize terms into a concept hierarchy *[Woods 97]*. By locating the head and modifier of noun and verb phrases, Woods was able to make choices on how to classify phrases. For example in the phrase "car washing", Woods' system would identify "car" as the modifier and "washing" as the head of the phrase. This would inform the system to classify the phrase "car

washing" under "washing" and not "car". The success of the technique relied on a large morphological knowledge base of information to help identify phrase components. Woods used the concept hierarchy to automatically expand non-matching terms of a query. In a set of retrieval experiments, Woods reported that use of the expansion method significantly improved the effectiveness of his retrieval system.

Simpler methods, such as term co-occurrence, have also been used to produce structures or maps of related terms [*Doyle 61*, *Thompson 89*]. To the best of our knowledge, however, most work in this area used term relations that were symmetric. Our interest was in producing a concept structure with an ordering from general terms to more specific. Such a production was performed in work by Forsyth and Rada *[Forsyth 86]*. They used the cohesion statistic to measure the degree of association between terms. The generality and specificity of terms was determined by their document frequency (*DF*), the more documents a term occurred in, the more general it was assumed to be[1]. The authors reported building a small multilevel graph like structure of terms. Although no testing of its properties were reported, the hierarchy of terms appeared promising. Despite the apparent success of the more sophisticated methods cited above, it was decided to start with Forsyth and Rada's much simpler ideas and explore what could be achieved using them, leaving open the possibility of adopting the more sophisticated methods for future work.

2.1.2 Method used

Although it was used to create a concept hierarchy, Forsyth and Rada's term association method (cohesion) was not originally designed to find the types of association found in concept hierarchies: where, as was stated at the start of this section, a parent node subsumes the topics of its children. Therefore, it was decided to drop cohesion in favor of a test based on the notion of subsumption. It is defined as follows, for two terms, *x* and *y*, *x* is said to subsume *y* if the following two conditions hold,

(1) $P(x|y) = 1$, $P(y|x) < 1$.

In other words *x* subsumes *y* if the documents which *y* occurs in are a subset of the documents which *x* occurs in. Because *x* subsumes *y* and because it is more frequent, in the hierarchy, *x* is the parent of *y*. Although a good number of term pairs were found that adhered to the two subsumption conditions (1), it was noticed that many were just failing to be included because a few occurrences of the subsumed term, *y*, did not co-occur with *x*. Subsequently, the first condition was relaxed and subsumption was redefined as

(2) $P(x|y) \geq 0.8$, $P(y|x) < 1$.

The value of 0.8 was chosen through informal analysis of subsumption term pairs.

Subsumption satisfied three of the design principles outlined at the start of this section. As a form of co-occurrence, subsumption provided a means of associating related terms. It did not prevent children from having more than one parent. Also, the *DF* of terms provided an ordering from general to more specific. The next principle to be tackled was the issue of the senses of ambiguous terms.

2.2 Ambiguous terms

As the terms of the hierarchy were to be extracted from documents, it was necessary to know the senses of the terms in those documents. Though a great deal of work has been expended on performing automatic word sense disambiguation *[Yarowsky 95, Ng 96]*, the low accuracy and general lack of availability of such systems effectively precluded the possibility of disambiguating a collection of text. However, one could ignore the issue of ambiguity by choosing to only derive concept hierarchies from sets of documents where ambiguous terms were used in only one sense. This was achieved by using top ranked documents retrieved in response to a query. Such documents would have some degree of commonality between them, meaning that the terms within them would most likely be used unambiguously. Working on such a set of retrieved documents also has practical importance, as building a concept hierarchy from such a set would provide a overview of those documents, and should prove useful to users wishing to quickly discover the topic structure of the retrieved set. The collection of documents and queries chosen for this work was TREC. For each of the queries (known as topics in TREC), the 500 top ranked documents were chosen as the set to process.

With the issues of ambiguity and the documents to process resolved, only the final design principle remained to be addressed: how to extract good terms from the documents.

2.3 Term selection

The initial source of terms came from the query which retrieved the documents in the first place. Before doing this however, certain query improvements were applied. As can be seen in the TREC conferences *[TREC]*, much research has successfully addressed the issue of improving queries through means of automatic expansion. This works in the following manner, an initial set of documents is retrieved in response to the original query and the best matching passages of the top ranked documents are examined to find words and phrases that commonly co-occur with each other across many of the passages. The best of these terms are then added to a query and a new, hopefully better, retrieval is performed. Local Context Analysis (LCA) *[Xu 96]* is regarded as one of the better performing expansion methods available. Therefore, before extracting terms from the queries, they were automatically expanded with around 70 additional LCA terms. As good as these terms were, it was felt that the resulting concept hierarchies would be rather small. Therefore, an additional source of terms was needed.

The second means of selection was a simpler process using a comparison of a term's frequency of occurrence in the retrieved documents with its occurrence in the collection. An empirically derived threshold was set to decide which terms would be selected; it was defined as follows,

(3) $x_r/x_c \geq 0.1$: x_r is the frequency of occurrence of *x* in the retrieved set, x_c is its occurrence in the collection.

[1] Use of a term's *DF* to determine specificity is not unusual, *IDF* weighting presumes that less frequent query terms provide a more specific description of an information need *[Sparck Jones 72]*.

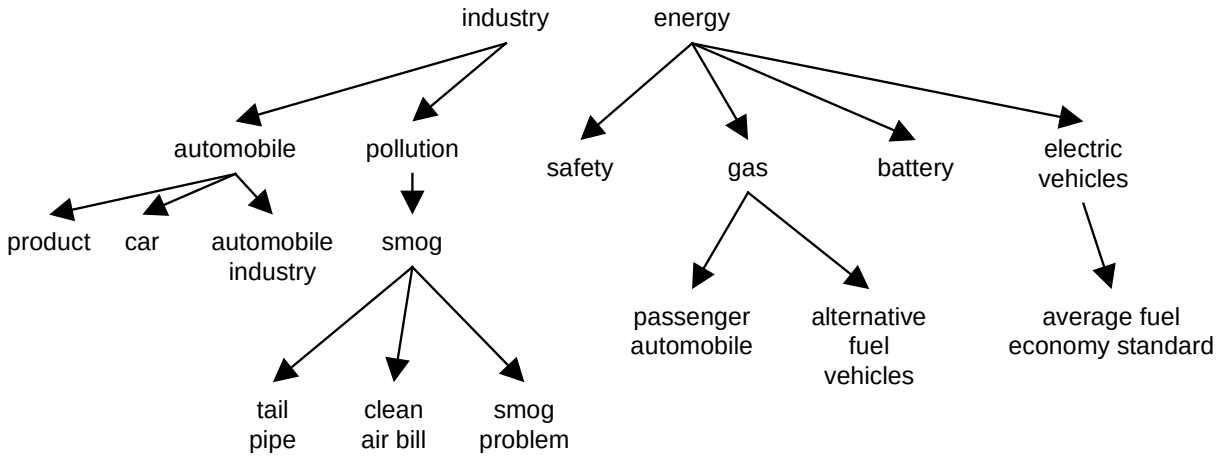

Figure 1 Fragment of concept hierarchy from TREC topic 230.

Not all the words and phrases found in the retrieved set of documents were extracted, only those found in the best passage of each document (i.e. the passage of the document most similar to the query) were subjected to this test. On average for each query, 2,000 words and 350 phrases were extracted from the 500 documents.

With this set of terms selected, the process to create a concept hierarchy could now take place.

2.4 Creating a hierarchy and contrasting it with other methods

Given a query and a collection of documents, LCA was used to expand the query with additional terms. Then retrieval was performed using the expanded query. The 500 top ranked documents were selected and the two forms of term selection took place, yielding, on average, 2,420 terms. Next, every selected term was compared to every other term to test for subsumption relationships. Around 200 subsumption pairs were identified. They were then organized into a concept hierarchy, which involved the removal of very infrequent terms. Figure 1 shows a fragment (~10%) of one of these structures resulting from TREC topic 230: "Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?".

As can be seen, much of the concept organization is promising, especially under "pollution". Other term pairs - "average fuel economy standard" and "electric vehicles" or "safety" and "energy" - seem less sensible, although examination of the underlying documents may reveal some unanticipated link. One other encouraging sign is that the hierarchy displays the desired property of general terms at the top leading to more specific terms below.

According to Hearst *[Hearst 96]*, topic 230 is "reminiscent" of the topic used to illustrate Scatter/Gather's display of clusters shown at the start of this paper (Section 1). As can be seen, there is no similarity between that display and Figure 1 even though retrieval was performed on the same TREC collections for roughly the same query. This should not be surprising, however, as polythetic document clustering works quite differently from the

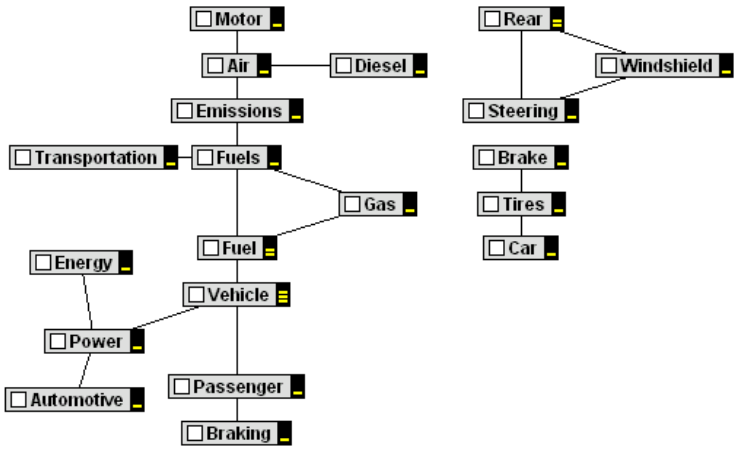

Figure 2 Clustered term structure from Refine.

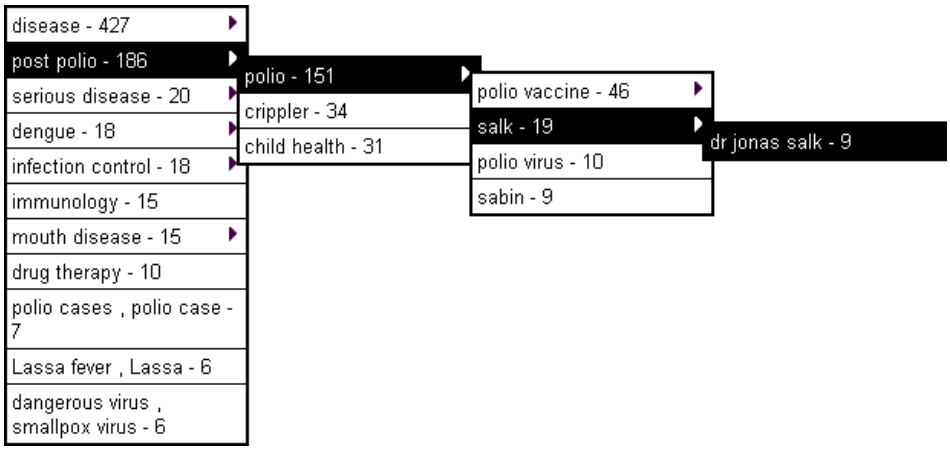

Figure 3: A fragment of concept hierarchy from topic 302

monothetic clustering used here. Document clustering is based on finding document wide similarities to form clusters. The organization of terms used in the concept hierarchies is much more akin to term clustering techniques where similarity is based on smaller passages of text. One such example is the Refine system (previously known as LiveTopics and Cow9) at the AltaVista search site *[AltaVista, Bourdoncle 97]*. Documentation of this system is somewhat limited, but it would appear that sets of words found to commonly co-occur with each other are grouped together. Groups with a degree of similarity to each other are linked up to form an undirected graph structure. The word groupings are presented to users for possible query expansion. Figure 2 shows the output of Refine after entering the query "auto car vehicle electric" (Hearst's query reminiscent of topic 230; use of the full TREC topic produced poor output). Each node is a word grouping, which is expanded via a pop-up menu. Remembering that Refine is working from a different document collection (i.e. the web), there is some degree of similarity between its output and the presentation in Figure 1. The main difference is in the organization of terms. The layout of the Refine groups has no significance. The method presented in this paper, like Refine, groups commonly co-occurring terms. However, the hierarchy lays out the related terms in order of specificity down the branches of its tree, and if possible joins groups of terms together via more general terms higher up. For example, the words "pollution", "smog", and "tail pipe" all co-occur with each other and would be possible candidates for a cluster of terms in the Refine system. It is perhaps less likely that "industry" would be included in a Refine-like cluster as it is too frequent and co-occurs with too many terms.

From these informal comparisons, it was concluded that the concept hierarchies were producing what appeared to be sensible organizations of terms in a manner that was distinct from existing techniques. The remaining issue was to determine a means of presenting the structures to users.

3. PRESENTING A CONCEPT HIERARCHY

As seen in Figure 2, it is possible to lay out a small graph structure on screen, however, the concept hierarchies being generated were much bigger: the fragment Figure 1 showed only one tenth of the whole concept hierarchy. Laying it all out on screen may not be possible[2]. Therefore, some alternative means of displaying the structure was examined. Although it is visually pleasing to see the entire arrangement of concepts laid out, it is not entirely necessary. A more minimal presentation may suffice: showing only the current layer the user is interested in and the path used to get there. A hierarchical menu provides such a means of presentation. Since it is a standard feature of operating systems, it was felt that menus would be familiar and easy to manipulate by users. Therefore, they were chosen as the means of presentation.

In order to display the hierarchies on any computer, a menu system was located that worked on web browsers *[DHTMLAB]* (using DHTML and JavaScript). Most menu systems are designed to allow a user to get to a known item in a sub-menu as fast as possible without making a mistake. This is generally achieved using delays related to mouse movement, which temporarily prevent the closing of the currently open sub-menu. Such a provision was not helpful for the task required here as the user was to be encouraged to browse around the entire structure as fast as possible. Luckily, the menu system obtained did not have such delays and so was well suited to the browsing task.

Figure 3, Figure 4 and Figure 5 shows three parts of another concept hierarchy, this time generated from TREC topic 302: "Poliomyelitis and Post-Polio: Is the disease of Poliomyelitis (polio) under control in the world?". The number next to each term is the *DF* of that term. As can been seen, from the three figures and the structure in Figure 1, the concept hierarchy provides a form of overview of the content of the retrieved documents regardless of whether they are relevant or not. In keeping with the forth design principles, notice that "Salk" (inventor of a polio vaccine) appears both in the "polio" and the "disease->vaccine" sections of the hierarchy; both sensible locations for this term. The structure while satisfying could be improved: in Figure 4 for example, "Fauci", the surname of an AIDS researcher, might have been better categorized under

[2] Although we have recently been made aware of a number of publicly available graph drawing packages that we plan to try.

"AIDS" instead of "virus". Nevertheless, as an initial step, the structure appears to be promising.

With a means of showing the concept hierarchy to users in place, it was now necessary to perform an evaluation of the structures.

4. EXPERIMENT

Evaluating the concept hierarchies presented a challenge, their intended purpose was to provide users with an overview of the topical structure of the documents retrieved in response to a query. Measuring how well something provides an overview was not going to be counted by some objectively derived value. In a paper on user evaluation of Scatter/Gather, Pirolli et al *[Pirolli 96]* reported using a method aimed at testing how well users understood the topical structure of documents after seeing Scatter/Gather clusters. Unfortunately, the test involved asking users to draw a concept hierarchy, something that would inevitably be influenced after seeing the structures generated here.

Clearly, it is possible to design a user study of the hierarchy's over viewing capabilities. However, it was felt that before expending time on such an effort, some of the basic properties of the structure should be examined first. Therefore, an experiment was created that addressed the second and third design principles outlined at the start of Section 2: testing the relatedness of a child to its parent; and examining the type of relationship between the two. The design of the experiment was as follows. Users were presented with a child term, it's parent and, if they existed, its grand and great grand parents. They were asked to make judgments about the child and parent, the other two terms were shown to provide the contextual path that led to the parent. The visual presentation of the terms was in a form very similar to the hierarchical menus discussed in Section 3. First, users were asked if they thought the relationship between the child and parent was interesting, uninteresting or they did not know. The word "interesting" was used as opposed to related or unrelated as it was felt that judging the relatedness of terms was not possible unless one examined the document texts. Asking a user if a relation was interesting would indicate if they would be willing to explore the child and the terms underneath it. If users did think a relationship was interesting, they were then asked to decide on the type of relationship between child and parent. Four of the organizing relations in WordNet were presented to users to choose from along with the ubiquitous "don't know". The names of the relations were changed to try to make them easier to understand by the users. They were asked if the child was either

- an aspect of the parent (a holonymic relation), e.g. an actor is an aspect of a movie;
- a type of the parent (hypernymic), e.g. Psycho is a type of movie;
- the same as the parent (synonym);
- the opposite of the parent (antonym);
- or they did not know or they had some other relation.

The first two relation types indicated that a child was more specific than its parent.

Fifty concept hierarchies were constructed from TREC topics 301-350 and a group of eight users (6 graduate students, and 2 authorial relatives) were asked to pass judgement on parent-child pairs. In order for the numbers provided by this experiment to have some context, users were also asked to judge a set of hierarchies formed by a random process. It was formed in the same manner as the concept hierarchies (as described in Section 2.4) except that when all terms were compared to all other terms, random selection was used to form parent-child pairs instead of subsumption. Note the ordering of terms based on frequency of

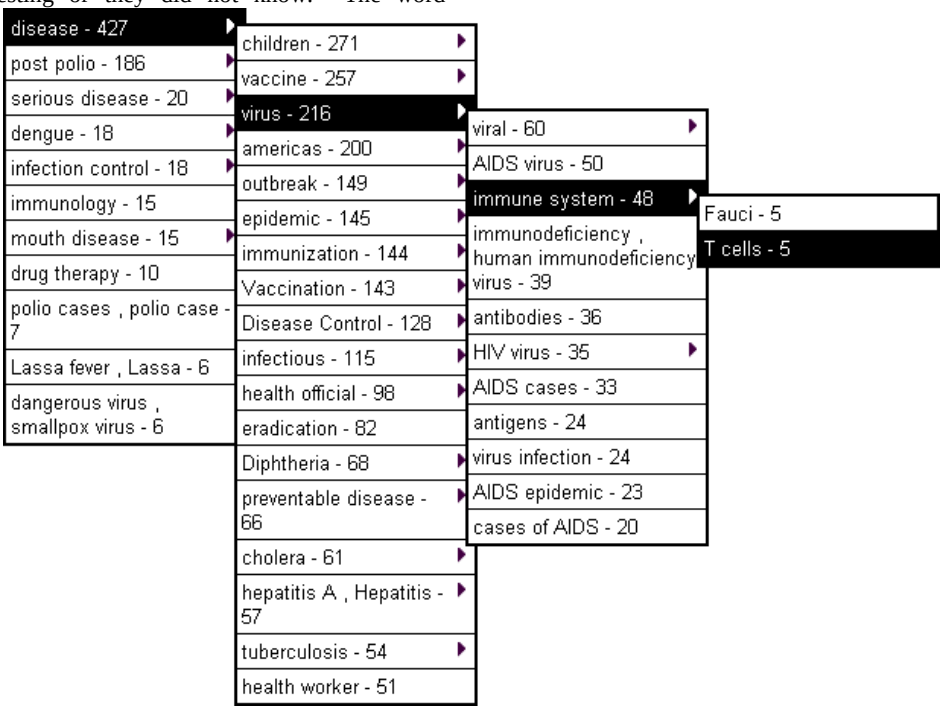

Figure 4: Second fragment of concept hierarchy from TREC topic 302

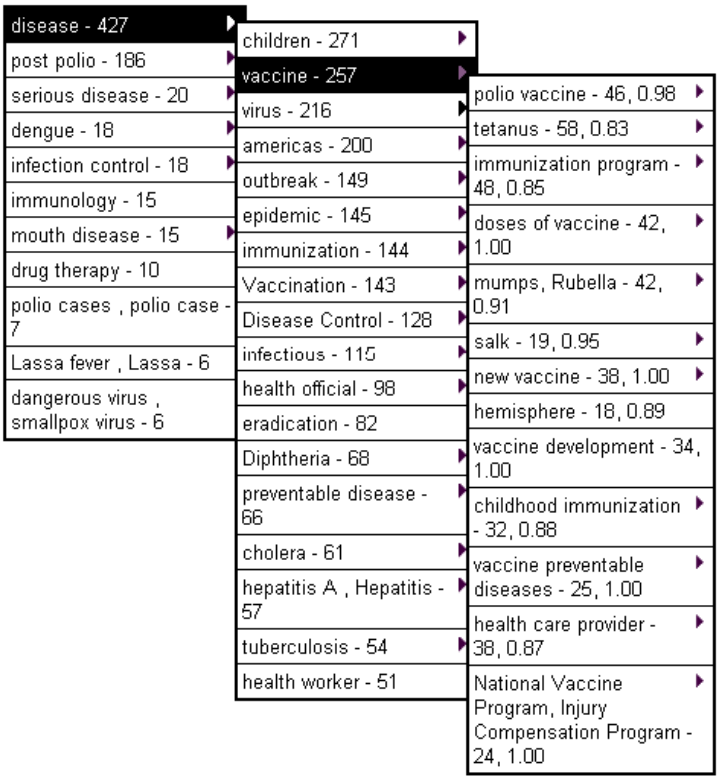

Figure 5: Third fragment of concept hierarchy from TREC topic 302

occurrence was still present in this structure. Users were not aware that the term pairs they were judging were formed from the two different processes.

5. Results

The results for the initial question, asking if term pairs were interesting, were as follows: 51% of randomly associated terms were judged interesting; this contrasted with 67% interesting term pairs from the concept hierarchy. Therefore, more term pairs were judged interesting in the hierarchies formed by subsumption than by those created randomly. This result was confirmed as significant after performing a one-sided paired t-test which produced the result $p<0.002$. The high number of random pairs thought to be interesting was not a surprising result. The terms in the hierarchies were all from similar documents retrieved by the same query, therefore, any pairing of the terms was likely to produce interesting associations.

	Aspect of	Type of	Same	Opposite	Don't Know
Random	47%	8%	1%	0%	43%
Subsumption	49%	23%	8%	1%	19%

Table 1: Results of user classification of term pair type if pair was judged as being interesting

The result of the second question, on the type of term pair relationships, is shown in Table 1. Here, it can be seen that although half of the random term pairs were thought interesting, many of them (43%) had a relationship that was not one of the standard WordNet relations. By contrast, only 19% of the subsumption formed pairs, were judged to be this unclassifiable type. This difference in the unclassifiable relationships was found (via the same t-test as above) to be significant: $p<0.01$. In terms of the generality of parent terms and the specificity of their child, 72% (49% + 23%) of the subsumption pairs had the "aspect of" or "type of" relationships, an encouraging result. Something unexpected was the high level of "aspect of" types found in the random pairs (47%). Remembering that the randomly formed hierarchies still used a term's frequency of occurrence to judge generality or specificity, this result would seem to indicate that this simple statistic is capable of indicating this quality of terms with a relatively high degree of accuracy.

Overall, the percentage of term pairs judged interesting and having an "aspect of" or "type of" relationship for the hierarchy formed through subsumption was 48% (67% * (49%+23%)). This compared to 28% (51% * (47%+8%)) for the randomly generated hierarchy. Although there was room for improvement, the experimental results indicated that the generated structures did in fact possess the desired qualities of a concept hierarchy.

6. FUTURE WORK AND CONCLUSIONS

In the future, it is intended to conduct experiments to examine the underlying documents to discover the extent and accuracy of topical coverage provided by the structure, particularly to examine how good the hierarchy is as a multi-document summary. In addition, it is planned to explore the utility of the hierarchy building system when applied to small document collections, for

example, a person's email. One other application of the concept hierarchies might be as a means of presenting possible query expansion terms as part of a retrieval system. Possible improvements to the quality of the hierarchical structure will also be examined: methods such as statistical co-variance of terms, examination of thesauri and use of Information Extraction techniques will be explored.

Through use of a simple term association technique, a method for building concept hierarchies has been presented. The hierarchies were informally compared to other methods that derive structure from collections of documents. From this comparison, it was shown that a hierarchical organization of monothetic clusters is quite different from both document and term clustering. Finally, through a small-scale user study, it has been shown that the generated concept hierarchies emulate some of the properties of manually generated subject hierarchies.

7. ACKOWLEDGEMENTS

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and also supported in part by United States Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235.

Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor(s).

8. REFERENCES

Alta Vista, www.altavista.com

Bourdoncle F. (1997) LiveTopics: recherche visuelle d'information sur l'Internet (LiveTopics: visual search for information on the Internet) in the Proceedings of RIAO: 651-654.

DHTMLAB, www.dhtmlab.com

Hearst M.A., Pedersen J.O. (1996): *Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results*, in the Proceedings of the ACM SIGIR, 19.

Doyle L.B. (1961): *Semantic Road Maps for Literature Searchers* in the Journal of the Association of Computing Machinery (ACM), 8(4): 553-578.

Forsyth R., Rada R. (1986): *Adding an edge* in Machine Learning: applications in expert systems and information retrieval, Ellis Horwood Ltd: 198-212.

Greffenstette G. (1994): *Explorations in automatic thesaurus discovery*, Kluwer Academic Publishers

Grefenstette G. (1997): *SQLET:Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text*, in the Proceedings of RIAO: 500-509.

Hearst M.A. (1998): *Automated Discovery of WordNet Relations*, in WordNet: an electronic lexical database, Christiane Fellbaum (Ed.), MIT Press.

Miller G.A. (1995): *WordNet: A lexical database for english*, in the Communications of the ACM, 38(11): 39-41.

Ng H.T., Lee H.B. (1996): *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*, in the Proceedings of the ACL, 34: 40-47.

Pirolli P., Schank P., Hearst M.A., Diehl C. (1996): *Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection*, in the Proceedings of ACM CHI Conference on Human Factors in Computing Systems: 213-220.

Spark Jones K. (1972): *A statistical interpretation of term specificity and its application in retrieval*, in the Journal of Documentation, 28(1): 11-21.

Thompson R.H., Croft W.B. (1989): *Support for browsing in an intelligent text retrieval system*, in the International Journal of Man Machine Studies, 30: 639-668.

Van Rijsbergen C.J. (1979): *Information retrieval* (second edition), Chapter 3, Butterworths, London.

TREC 97, *The Sixth Text REtrieval Conference (TREC-6)*, Editors: E. M. Voorhees and D. K. Harman, Department of Commerce, National Institute of Standards and Technology

Woods W.A. (1997): *Conceptual Indexing: A Better Way to Organize Knowledge*, a Sun Labs Technical Report: TR-97-61. Editor, Technical Reports, 901 San Antonio Road, Palo Alto, California 94303, USA.

Xu J., Croft W.B (1996): *Query Expansion Using Local and Global Document Analysis*, in the Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR 96) ,Zurich, Switzerland, pp. 4-11.

Yahoo, www.yahoo.com

Yarowsky D. (1995): *Unsupervised word sense disambiguation rivaling supervised methods*, in the Proceedings of the ACL, 33.