The role of automated word classification in the summarization of the contents of sets of documents

Presented at the Second International Conference on Information and Knowledge Management, CIKM '93, November 1993, Washington, DC.

Robert P. Futrelle and **Susan Gauch**[1]

Biological Knowledge Laboratory
College of Computer Science, 161CN
Northeastern University
Boston, MA 02115
futrelle@ccs.neu.edu, sgauch@tisl.ukans.edu

Abstract

In future digital libraries, even "perfect" retrieval will typically return too much material for a user to cope with. One way to deal with this problem is to produce automated summaries tailored to the user's requirements. One of the prime purposes of a summary of a collection of documents is to collapse together all of the important information elements that are common to the collection. This requires some method of discovering classes of similar items, e.g., word classes. This paper describes automated techniques for placing words in similarity classes. To do this, each target word is described by a composite vector that records the occurrence of words positioned near any occurrence of the target. Target words with similar contexts are grouped together by a clustering algorithm. We describe how such classifications can be used in information retrieval and for the summarization of biological literature.

The dilemma of "perfect" retrieval

In retrieving documents or portions of full-text documents, recall is the percentage of the desired documents that are retrieved and precision is the percentage of all retrieved documents that are of the desired type. No matter how good future systems become, even if they achieved 100% recall and precision, the amount of information that will be on line will be so large that the user will still be overwhelmed. It will rarely be the case that one returned paragraph or even one entire document will answer the user's questions. The information the user wants is typically scattered throughout the documents simply because none of the documents were written (nor could they have been written) to satisfy the interests that one particular user would have at some later time. The user could look for a review of the topic, but again, there would probably not be a review focused on the user's interests, much less one that was as up-to-date as the literature itself. Because the information desired is scattered across many documents, ranking the documents in order of relevance does not solve the problem.

One solution: Summarizing document sets

Retrieval systems could help to avoid the dilemma above if they could automatically produce a summary of the relevant documents tailored to the user's interests, particular query, level of expertise and adjusted to some particular length (from a paragraph to many pages). There has been work on extracting information from single sentences, from paragraphs (Zadrozny & Jensen, 1991), work on summarizing the arguments in whole documents (Alvarado, 1990) and work on automatic abstracting (Paice, 1990). Extensions of these techniques can be applied to summarizing the contents of sets of documents.

Manual analysis of reviews in the biological and computer science literature reveals the strategies authors use to summarize large collections of literature. One of the primary devices is to generate that describe items of a given class, citing the appropriate sources. The listings could be sets of genes or enzymes in biological articles or sorting algorithms or network protocols in computer science. Discovering the set of items in a given class in a document collection would need to be automated for this strategy to succeed. It is not appropriate to say that the system should refer to some standard listing of the items of a given class, because new terms are constantly being introduced in rapidly moving fields such as computer science or biology. Furthermore, terminology and use is often specific to a given subfield.

As an example, we would like an automated summary system to produce tables such as the following for a biological topic,

Term	**Context and source**
λ repressor	"O_L and O_R each contain a series of nonidentical binding sites for the λ repressor..." [Stryer, 1975] Pages 35-39 of Genetic Switch [Ptashne, 1992].
lac repressor	The repressor of the lactose operon [Stryer, 1975]

In a navigation (hypertext) environment the user could select any of the items in the table for expansion.

In order to select the terms that should be grouped together in tabular summaries as in the example above (λ and lac), word classification must be done. This is described in the next section.

Describing and quantifying word contexts

To discover word classes, we describe the context of a word (the target word) by the preceding two context words and the following two context words. Each context position is represented by a vector containing the joint frequencies of the 150 highest frequency words in the corpus, giving a 600-dimensional context vector. The entries in the context vectors are converted to mutual information measures, with smoothing. The similarities of the resultant context vectors for the 1,000 highest frequency words are computed from the normalized inner products of their context vectors (cosine rule). The resulting set of 500,000 similarities is used as the basis of a hierarchical clustering algorithm, a bottom-up approach producing binary trees with a similarity at each node, $-1.0 \leq \leq 1.0$. The method was inspired by (Finch & Chater, 1992) and is described in more detail in (Futrelle & Gauch, 1993). Near the leaves, the words were found to be grouped by both semantic and syntactic similarity. Further up the tree, the larger classes retained only syntactic similarity.

[1] Prof. Gauch's current address: Dept. of Computer Science, U. Kansas, Lawrence, KS.

Some examples from the biological literature

The corpus used for this analysis was the 220,000 words of text in 1,700 abstracts that completely cover the field of bacterial chemotaxis since its inception in 1965. Bacterial chemotaxis is a phenomena in which single bacteria move toward higher concentrations of chemical attractants such as sugars (and away from repellents). One of the classes of terms that is constantly being added to by biologists is genetic mutant designators. One class of these the system discovered consists of ten items:

motB, tar, tsr, cheB, cheZ, cheY, cheA, flaA, flaE, double

There are two apparent anomalies in this list, "tar" and "double", both common words in other contexts. The utility of the classification method is that it is sensitive to the particular use of these words in this specialized field. "tar" means "taxis towards aspartate" in this field and "double" is used to describe mutants which have two lesions in the same or different genes. Thus, if a table of mutants were constructed to summarize this set of papers it should include all ten items.

The following class contains compounds that are attractants used in chemotaxis studies,

aspartate, maltose, galactose, ribose, serine

These could usefully be placed in a list summarizing the major compounds of interest.

The word classes also include , which are fundamental to the understanding of living systems,

chemotaxis, taxis, sensing, motility, rotation, behavior, movement, transport, uptake

Again, a tabulation of these along with excerpts describing them or references to articles devoted to them would be useful as part of a summary.

Note that the word classes shown above are both syntactically and semantically homogeneous. The examples above contain only nouns. The homogeneity is easily seen from some other classes generated by the system,

adjectives:

higher, lower, greater, less

other, several, many

molecular, structural

nouns (physical units):

degrees, min, s, mM, microM, nm

verbs:

suggest, indicate, show, demonstrate

prepositions:

of, in, for, with, on

The semantic homogeneity exhibited by the classes is richer than simple synonymy. Words such as "higher" and "lower" are classed together because they are members of the same (graded) semantic field. They are used in similar contexts and the choice of which to use is based on the author's knowledge of the world, not on the surrounding word context. If they were determined by the word context they would be predictable and therefore information-poor. Their very unpredictability allows them to be carriers of information.

Applications in information retrieval

Our technique of classification can contribute to the solution of a number of problems in information retrieval. Its greatest utility comes from the fact that it produces . There have been two very different methods used in the past for word classification. The first is found in part-of-speech taggers, which normally operate sequentially and predict the part of the speech of a word from the part of speech categories determined for a few immediately preceding words. These methods can be categorized as methods. Semantic classification has typically been based on the co-occurrence of words in a much larger context, from 50 word neighborhoods to entire documents. This method is a one. Our method retains the actual word identities from the local context and the results above make it clear that the method is

One of the potential uses of the word classes is for automated thesaurus construction. But since the method does not distinguish words in the same semantic field, e.g., "attractant" and "repellent", we have to examine its utility carefully. Our preliminary experiments (S. Gauch, unpublished) show that expansion of search terms by words that are strongly associated by our classifier () gives a significant improvement in document retrieval. Why this should be true is most easily seen from the example of "higher". Using "higher" as a search term will find instances of "A is higher than B" but will miss the logically equivalent statement, "B is lower than A", unless "higher" in the query is expanded to include "lower". The conclusion from this example and from our experiments is that the classes found by our method are indeed useful for improving retrieval when they are used for search term expansion.

Because single words can have more than one use (multiple meanings), it is useful to distinguish these uses to gain precision in analysis and retrieval. The appearance of "tar" as a mutant designator is a clear example. By treating each individual occurrence of "tar" as a separate entity, rather than lumping all occurrences to compute the context vector, we can discover its distinct uses. Individual occurrences have extremely sparse context vectors (with at most four non-zero entries) so we have found that the separation of single word occurrences into distinct classes is significantly enhanced by expanding their context words by their simsets (Futrelle & Gauch, 1993). The next planned improvement is to iterate the method once more by distinguishing the senses of the context words, e.g., distinguishing the verb "results" from the plural noun "results". The use of disambiguated context words plus their expansion by simsets should bring our classification performance to the level of the best part-of-speech taggers.

Our method has several advantages over the standard methods: 1) It is an method that requires no tagged training corpus, so it can deal with novel text. 2) It is possible to "tune" the size of the simsets to optimize performance. For example, the physical unit class shown earlier could be retained as a separate specialized class instead of simply lumping it with . 3) It produces classes that are both syntactically semantically homogeneous.

Characterizing document sets

We will now describe how classification techniques can help to characterize document sets, including producing summaries. We use the word "document" in a general way, so that any block of text from a sentence to an entire book might be considered a document. This is appropriate, because today many of the analysis and retrieval techniques originally developed for whole documents are being successfully applied to portions of documents.

To be useful, the characterization of document sets must be tightly tied to the queries that a user presents to the system. One way to do this is simply to return a few of the most relevant passages, with relevance determined by a complex interplay of global and local context (Hearst & Plaunt, 1993; Salton, Allan & Buckley, 1993) or by trying to discover semantically relevant categories (Paice & Jones, 1993). But for the biological literature which produces over 10 million paragraph-length "documents" every year, the "most relevant passages" will either be a small and not very representative sample or too large a set to be useful.

The alternative we suggest is to return results in a more compact tabular form, possibly with further hierarchical structure. The tabular material would be synthesized from the highest ranking passages retrieved by the methods just described, using word classification to identify important sets of terms with a strong bias towards sets that contain terms from the user's query. For example, if the query involved the mutant cheZ, and the system was required to construct a short tabulation, it would favor the terms most closely associated with cheZ according to our classification analysis: cheY, cheB, tar and tsr.

One of the great opportunities afforded by Hypertext techniques is to give the user the ability to rapidly expand and collapse any collection of material on the screen to converge on the items of greatest interest. The tabular method described plus the navigation tools of Hypertext should give the user a truly effective system for information access. Methods to do this efficiently, based on structured text representations stored in object-oriented databases are under development (Futrelle & Zou, in preparation).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. DIR-8814522 and IRI-9117030. Thanks to K. Klimukhina for clustering software development, to Linda C. Smith for pointers to the automated abstracting literature and to Marti Hearst for discussions about our research program. Maurice Pescitelli compiled the entire bacterial chemotaxis corpus. Thanks to the ERC for providing a productive working environment.

References

Alvarado, S. J. (1990). . Norwell, Massachusetts: Kluwer Academic Publishers.

Finch, S., & Chater, N. (1992). Bootstrapping Syntactic Categories Using Statistical Methods. In W. Daelemans & D. Powers (Ed.), , (pp. 229-235). Tilburg U., The Netherlands.

Futrelle, R. P., & Gauch, S. (1993). Experiments in syntactic and semantic classification and disambiguation using bootstrapping. In , (pp. 117-127). Columbus, OH: ACL.

Hearst, M. A., & Plaunt, C. (1993). Subtopic structuring for full-length document access. In , (pp. 59-68). Pittsburgh, PA: ACM.

Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. , **26**(1), 171-186.

Paice, C. D., & Jones, P. A. (1993). The identification of important concepts in highly structured technical papers. In , (pp. 69-78). Pittsburgh, PA: ACM.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In , (pp. 49-58). Pittsburgh, PA: ACM.

Zadrozny, W., & Jensen, K. (1991). Semantics of Paragraphs. , **17**(2), 171-209.