Cases as Structured Indexes for Full-Length Documents

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
and
Xerox Palo Alto Research Center
marti@cs.berkeley.edu

Abstract

Two long, full-length texts are not likely to discuss all, or almost all, of the same subtopics or subpoints. Even if the documents contain many of the same terms, the ways the terms are grouped to form subtopical discussions still might be quite different. A solution is to create a description of a document which lists all of its subtopical discussions as well as its main topics. An index that indicates this structure is an abstract representation of the document, and we can think of this index as a case in the Case-Based Reasoning (CBR) sense. This paper proposes the use of cases to represent the high-level structure of full-length documents for the purpose of information retrieval. The cases are to be used both for assessing document similarity and for helping the user construct viable queries. The case can be transformed in various ways in order to make it more similar to the descriptions of other documents; these transformations include generalizing, substituting, and emphasizing subtopic descriptions. An advantage of this approach is that the cases that represent the document are automatically generable.

Introduction

This paper proposes the use of CBR-like cases to represent the high-level structure of full-length documents. The cases are to be used both for assessing document similarity and for helping the user construct viable queries. Before describing the nature of these cases, some background on retrieval from full-length documents is warranted.

Much of information retrieval is concerned with determining which documents from a large collection are most similar to one another, or to a query. What does it mean for two documents to be similar? Most IR methods assume that term frequency counts provide a good measure of similarity, and this assumption is reasonable for many document genres.[1]

For example, if two technical abstracts share many terms and if many of these shared terms are uncommon with respect to the rest of the corpus, there's a good chance the two abstracts are similar in content. This applies equally well to short texts like newswire reports.

However, when we begin to consider a full length text instead of the abstract that describes it, or a 5-10 page science article instead of a 5 paragraph newswire, we may want to adjust our criteria for similarity. It isn't realistic to expect two long documents to discuss all, or almost all, of the same subtopics or subpoints. And if two documents only have a small fraction of their terms in common, can we really assume they are similar? Even if they contain many of the same terms, the ways the terms are grouped to form subtopical discussions still might be quite different, thus indicating the documents are dissimilar at least along this dimension.

This issue has not yet been addressed by many researchers, most likely because large volumes of full-length text have only recently become available online. The most comprehensive work to date on retrieval issues pertaining to full-length text is that of

[1]This is a deliberate simplification. Many sophisticated probabilistic models have been devised for how to combine term counts and work has been done on how to incorporate information external to term counts as well. See, e.g., (Croft & Turtle 1992). However, the basic assumption of most vector-space and probabilistic methods is that a calculation of some sort is done based on overall frequency counts (Salton 1988).

Salton and Buckley (Salton & Buckley 1991b),(Salton & Buckley 1991a), who take pains to normalize the lengths of the documents that they compare. They have compared paragraphs within a large document (e.g., Salton's book), articles within an online encyclopedia, and electronic mail messages (inquiries and their replies). According to Salton and Buckley, a good way to ensure that two larger segments, such as two paragraphs, are similar is to make sure they are similar both globally and locally (i.e., sentence-by-sentence). For two sections to be similar, they must be similar both globally, at the paragraph level, and at the sentence level. Salton and Buckley's results show that their procedure is quite effective in many cases.

The OFFICER system of (Croft *et al.* 1990) provides a retrieval interface to complex full-length documents, where the documents are represented according to their orthographical markings: title, author, sections, paragraphs, figures, and so on. Users are able to create queries that are sensitive to these structuring elements. However, the kinds of texts under consideration in this paper are those that don't have much orthographically specified information; rather they are continuous, unbroken text. In future the ideas presented here should be applied to texts that are structured by the author, and in that case it might be useful to integrate these ideas with a system like OFFICER.

A Proposal

Normalizing documents to allow more justifiable term-based comparison retains the existing assumptions about how documents should be compared. We need to revise these assumptions to better account for the structure of full-length text. Consider expository text, specifically scientific reporting as is found in, say, a *Discover* magazine article. This kind of text often consists of a main topic and a series of subtopic discussions that are related in some way to the main topic. Often these subtopics are of interest in their own right.

The Salton and Buckley algorithm allows comparisons of subtopics of documents, provided that a subtopic's extent aligns exactly with a predetermined boundary unit, e.g., a section or paragraph. However, a subtopic discussion may have an irregular length, perhaps one third of a section. It would be more useful in some circumstances to identify the subtopical discussions in their complete extent in advance, and make comparisons based on these.

Furthermore, recall that Salton and Buckley combine global and local comparisons in order to make a similarity judgement. Instead, we should keep this kind of information distinct. (Hearst & Plaunt 1993) describe a retrieval paradigm in which a user can specify not only the subtopic to retrieve on, but also which main topic the subtopic should appear in the context of. In other words, we advocate allowing a user to make a distinction between, for example, retrieving a discussion of *volcanic activity* in the context of *planetary exploration* and a discussion of *volcanic activity* in the context of *Roman history*. This is a new information access paradigm especially tailored toward full-length documents.

Here I propose extending this idea in the following way: Create a description of a document which lists all of its subtopical discussions as well as its main topics. By determining the subtopical structure we are in effect imposing some high-level structure on the document. An index that indicates this structure is an abstract representation of the document. We can think of this index as a case in the CBR sense.

(Cutting *et al.* 1992) mention the benefits of browsing a document via the table of contents (TOC) at the front of it, as opposed to the index at the back (which is the standard procedure in IR if we follow this analogy). In the framework proposed here, cases encode information that is similar in form to a TOC. However, in order to be effective, the terms involved in the cases will be richer and more descriptive than what is found in a typical TOC; how these terms are acquired will be described momentarily. Bear in mind that the term TOC is being used loosely here; the idea is that we recognize the subtopic structure of the document whether it has a real table of contents or not.

Given a large collection of documents, we shouldn't expect many pairs of documents to have well-matched subtopic structure. In other words, if one document discusses subtopics $S_1, S_2, ... S_k$ in the context of main topic M, it's not likely that we'll find another document with exactly the same structure (see Figure 1). In order to allow users to query for documents similar in structure to other documents, the cases representing the structure of the documents can be grouped according to similarity along some axes.

We can also envision a new relevance feedback paradigm that would allow users to construct queries by obtaining a case corresponding to an existing document and transforming it appropriately.

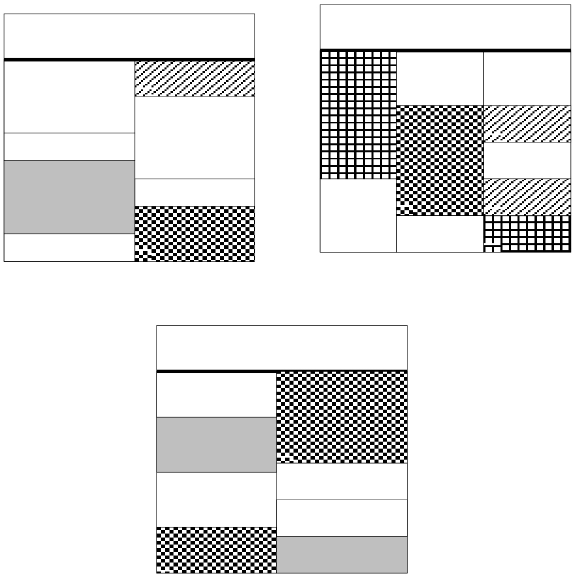

Figure 1: A sketch of the subtopic structure of three texts that share the same main topics. Instead of comparing the the texts in their entirety we construct cases that reflect the subtopic structures. Then the case representations can be adjusted to indicate which portions of the texts should be considered for comparison.

The transformed case is then compared to the other cases in the document collection and those with the most similar structure are retrieved. This setup also allows retrieval based on similarity to some subpart of a document.

To be more specific about how this might work, consider an environmental impact report written by the Department of Water Resources. The document in question is a study of the conditions necessary for growing a particular kind of sycamore tree. A human judge classified the document as having the following structure (with accompanying paragraph numbers):

1-7 *environmental impact*
8-15 *planting site and project specifications*
16-21 *irrigation*
22-23 *wind effects*
24 *weed effects*
25-27 *browse effects*
28-40 *assessment of study*

Given this template, a user who wanted to find a document that is similar along some dimensions could perform any or all of the following operations:

- Generalization:
  convert *wind effects* to *climate effects*
- Substitution:
  change *irrigation* to *pest elimination*
- Elimination:
  remove *site* section
- Remove ordering dependencies:
  say the effects can occur in any order
- De-emphasize:
  assign a low weight to some sections
- Emphasize:
  assign a high weight to some sections
- Background Change

The last item on this list indicates that a document should be found whose subtopic structure is similar to the current one's but whose overall topic is quite different. Because the information about the document is represented in a structured manner, it contains information about which terms strongly indicate the main topic(s). Therefore, change in background could be accomplished by removing from consideration, when performing similarity comparison, all terms that have been found to indicate the main topic(s).

This scheme allows the user to be selective about what parts of the example document are relevant, what parts should be matched on, and what parts should be substituted for others. Note that in a standard retrieval framework in which a query is a block of text it is not possible to substitute one topic for another.

Instead of having users adjust the structures of cases, we can consider a setup more congruous with standard CBR. We can imagine that documents are processed one by one, and assigned their TOC-like case. Each document's case is then added to a case network or lattice, positioned according to which existing cases it is most similar to. The system must specify which components of the case can be allowed to vary in order for two cases to be considered to be related to one another. This is a tricky part of the implementation for most CBR systems. (In HYPO (Ashley 1987), similarity is determined according to *dimensions*, in CYRUS (Kolodner 1983), according to values of *features*.) Often the similarities between features is determined by relationships in domain knowledge.

For the scheme proposed here, there are several ways to organize the similarity determination. One way would be to associate a vector with each entry in the TOC-like structure, indicating the terms that are most representative of the corresponding text segment. Thus each case is represented by k vectors, where k is the number of entries in the TOC. After first organizing cases according to their main topic(s), documents can then be grouped according to which and how many of their cases' vectors are similar to one another, using a standard similarity measure, such as cosine (Salton 1988). Documents which have the same number of vectors, all of which are similar, are considered to be closest, documents which have one differing vector are slightly less close, and so on. Alternatively, documents that have simply one or two vectors in common might be considered closely linked, or documents that have a set of contiguous vectors that are similar. It is straightforward to see how substitution of a new section or removal of an existing one can be accommodated by such a setup.

(Smith *et al.* 1989) report a study which monitored the suggestions made by professional searchers who were helping information seekers find relevant abstracts from the *Chemical Abstracts*. The suggestions were classified into a set of tactics that users can try in order to transform their query into something that would better match the contents of the corpus. The four most common tactics observed were: i) add more terms to the query, ii) remove terms from the query, iii) replace a term with a more specific instance of the term, and iv) replace a term with a more general instance of the term. They also suggested broadening the query with synonyms and restricting the search field (i.e., looking at the title only vs. the title and the abstract). That these tactics are similar to several of the operations suggested above, and that they were observed in the behavior of professional searchers, confirms their feasibility. The information retrieval system developed by (Smith *et al.* 1989) is one in which the abstracts are represented by hand-built semantic frames. Fittingly, they suggest that the adjustment tactics can be considered equivalent to adding or deleting slots or slot fillers from semantic frames. This kind of representation has the disadvantage of requiring a huge knowledge engineering investment. As discussed below, the approach suggested here does not require hand-encoded representations.

Case Generation

A problem that plagues many CBR systems is a lack of an automated mechanism for converting some form of data into its representative case. This is especially problematic for CBR systems in the legal domain, since programs that can successfully convert legal text into cases are still distant research goals. In order to realistically integrate IR with CBR, however, case construction must be automated.

In this proposal, cases can be built up automatically, their structure being based on the data rather than on a predefined framework. (Hearst 1993) describes TextTiling, a method for partitioning full-length expository texts into multiparagraph discourse units. Each 'tile', or segment, is intended to represent a dense discussion of a subtopic. Thus if a term is only mentioned in passing it will not be identified as a true subtopic of the document. Similarly, if a term occurs many times but is scattered approximately uniformly throughout the body of the document, it again is discounted as evidence for discussion of a subtopic. This procedure thus helps distinguish real discussions involving a term from false alarms. Furthermore, TextTiling will be able to create an informative description of each tile, or section, which can subsequently be used for similarity comparisons. It will also create a description of what terms conspired to indicate the main topic(s) of the document. These descriptions are thus more informative that what would typically be found in an entry in a table of contents.

Once the tiles have been identified, labels must be generated to provide the user with an encapsulated description of each segment. One way to do this is to pick relevant terms from those used to identify the tile. Another is to classify the segments' relevant terms according to an ontology or thesaurus. Experiments with employing a disambiguation algorithm proposed by (Yarowsky 1992) to this problem, have promising initial results. This procedure in effect incorporates domain knowledge since the classifications are based on statistics gathered by use of a lexical thesaurus. If a domain specific thesaurus is used, a domain specific classification can be made.

This ability to classify tiles might be useful for allowing users to specify generalizations and substitutions to the case they are modifying. The system can store previously seen examples of subtopics of each entry in the ontology and use these instead of data from a specific document for the substitution in the similarity comparison procedure.

Is This CBR?

The setup sketched in this paper is not a standard CBR system in several respects. Perhaps the most glaring omission is that it does not make extensive use of domain knowledge in order to decide how to make one case better match another. In other words, it is missing the 'R' from CBR.

(Bareiss 1989) lists six characteristics for comparing case-based reasoning systems:

(1) How cases are indexed for efficient retrieval.
(2) How the similarity between a new problem and a retrieved case is assessed.
(3) How cases are selected for retention.
(4) How indexing information is learned.
(5) How any additional domain knowledge required for the assessment of similarity is acquired.
(6) How generalization (if any) occurs during learning.

Although the system proposed here does specify how to index cases for efficient retrieval (criterion (1)), the mechanism for doing so is not really based on a domain model, but rather on a statistical assessment of term count similarity between subparts of the case. This applies equally well to criterion (2). Criteria (3-5) are not relevant for my framework but criterion (6) is relevant with respect to allowing the user to specify changes to a given case in order to formulate a query, as discussed above.

Perhaps it is incorrect to assert that this proposal incorporates CBR into an IR framework since several aspects of standard CBR, notably the dependence on reasoning from domain knowledge, are absent. (Although domain knowledge is to be used for classifying subtopics.) The TextTiling procedure is intended to make a good approximation of the meaningful contents of a full-length document, but it cannot discover, say, causal relationships or detailed feature assignments. It may be the case that in order to build a system that can scale to the demands of an IR system for unrestricted text, this element must be missing, since automated methods for converting full-length texts to complex, detailed representations are still not feasible.[2]

[2]Recently researchers have become more successful at converting short, domain-specific texts into template-like representations. (Liddy 1991) reports work on converting empirical abstracts into knowledge structures, and several of the researchers participating in the MUC competition (Lehnert & Sundheim 1991) are showing promising results at classifying the contents of newswire articles.

The proposal presented here is intended to allow users access to a partially structured representation of full-length documents, in a framework that can be implemented automatically and relatively efficiently; these are crucial criteria from the IR viewpoint. By thinking of these representations as CBR cases, we are opened up to the ideas of structuring, adjusting, and comparing queries that reflect the actual structure of the document. This is one suggestion for how ideas from CBR can help influence the next generation of information retrieval systems.

Acknowledgements

I would like to thank Jan Pedersen, Narciso Jaramillo, David Lewis, and the three anonymous reviewers for helpful suggestions to improve an earlier draft of this paper. This research was sponsored in part by the University of California and Digital Equipment Corporation under Digital's flagship research project Sequoia 2000: Large Capacity Object Servers to Support Global Change Research, and in part by Xerox Palo Alto Research Center.

References

Ashley, K. D. (1987). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. PhD thesis, Univeristy of Massachusetts, Amherst.

Bareiss, R. (1989). *Exemplar-Based Knowledge Acquisition*. Perspectives in Artificial Intelligence. Academic Press, Inc.

Croft, W. B., R. Krovetz, & H. Turtle (1990). Interactive retrieval of complex documents. *Information Processing and Management*, 26(5):593–616.

Croft, W. B. & H. R. Turtle (1992). Text retrieval and inference. In P. S. Jacobs, editor, *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pages 127–156. Lawrence Erlbaum Associates.

Cutting, D. R., J. O. Pedersen, D. Karger, & J. W. Tukey (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR*, pages 318–329, Copenhagen, Denmark.

Hearst, M. A. (1993). TextTiling: A quantitative approach to discourse segmentation. submitted.

Hearst, M. A. & C. Plaunt (1993). Subtopic structuring for full-length document access. submitted.

Kolodner, J. L. (1983). Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7(4):243–280.

Lehnert, W. & B. Sundheim (1991). A performance evaluation of text-analysis technologies. *AI Magazine*, 12(3):81–94.

Liddy, E. (1991). The discourse level structure of empirical abstracts – an exploratory study. *Information Processing and Management*, 27(1):55–81.

Salton, G. (1988). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, MA.

Salton, G. & C. Buckley (1991a). Automatic text structuring and retrieval: Experiments in automatic encyclopedia searching. In *Proceedings of SIGIR*, pages 21–31.

Salton, G. & C. Buckley (1991b). Global text matching for information retrieval. *Science*, 253:1012–1015.

Smith, P. J., S. J. Shute, D. Galdes, & M. H. Chignell (1989). Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems*, 7(3):246–270.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 454–460, Nantes, France.