

Learning for Question Answering and Text Classification: Integrating Knowledge-Based and Statistical Techniques

Jay Budzik and Kristian J. Hammond

Intelligent Information Laboratory
University of Chicago
1100 E. 58th Street
Chicago, IL 60637
{budzik, hammond}@cs.uchicago.edu

Abstract

It is a time consuming and difficult task for an individual, a group, or an organization to classify large collections of documents under a content-driven taxonomy. In this paper, we outline an approach for building a system which makes the classification process the responsibility of the author of the document, thus allowing the author to explain classifications and verify (or correct) automated techniques. We present our preliminary work on such a system, Q&A, which enables the distribution of the task of semantic classification and knowledge acquisition by semiautomatically learning taxonomic categorizations and document indices as it captures interactions between experts and question-asking users.

Introduction

The availability of vast quantities of online information has provided a new context for considering an old problem: given a need for information and a collection of possible information sources, how can we most optimally obtain the information we seek? Even if the information is available, it may not be easily accessible, especially if we are faced with considering a large collection of information sources spanning diverse domains. Thus, we need to minimize the amount of time and effort it takes to obtain the specific information we need, even if the number of sources is huge.

A natural solution to this problem is to organize information sources conceptually, clustering semantically related documents. Semantic organization not only helps reduce the complexity of retrieval, it also improves accuracy. Historically, there have been two distinct ways of doing this:

1. Use statistical measures of similarity between information sources to discover semantic relations among them.
2. Use a symbolic representation of information sources and a causal, semantic model of a specific domain to measure similarity and build taxonomic semantic classifications.

Each approach has distinct advantages and drawbacks. Knowledge-intensive systems are usually quite

accurate because they can reason about similarity using a causal domain theory. Yet developing a representation for concepts, creating a concept hierarchy and enumerating the relationships between concepts is a difficult phase in the development of any knowledge-intensive system. As a result, such systems also tend to be domain specific and difficult to construct. On the other hand, systems which exploit purely statistical models usually span a wider breadth of domains. In exchange, they sacrifice accuracy. This dichotomous relationship motivates the creation of a hybrid system which serves to capture both the content and the conceptual organization of various information sources relative to one another. Such a system would learn necessary knowledge along the way, allowing it to subsequently harness this knowledge for indexing and retrieval. Our goal is to produce a system which achieves both the breadth of statistics and the power of semantics by using both methods.

Document Classification in the Large

The need for semantic organization is nowhere more prevalent than on the World-Wide Web. While the Web has provided an easy way for users to make information available to a wide audience, it does not, with only few exceptions, provide a means for the wide-scale semantic organization of this available information. Thus we have a suitable solution to the first part of the above problem, while the second part, that of semantically organizing this information for retrieval, remains to be solved. It has been argued elsewhere (Etzioni 1996) that the Web is sufficiently structured to allow us to mine semantics after the documents are made available. While this is true, it does not imply that all attempts at providing a means for Web users to organize their documents should be abandoned. Efforts to classify Web documents such as Yahoo! (Yahoo! 1995) are a good first step, but document classification in the large simply cannot rely on having a central organization which performs classification by hand, without a long delay between the time a document is published and the time it is discovered and finally classified.

Instead, the document classification project should rely on the collective effort of those publishing the doc-

uments in the first place. A preliminary solution to this problem is offered by allowing a user to suggest a category for a document, as it is done at Yahoo! Yet under this paradigm, site categorizations must still be validated by hand to ensure documents are classified under the appropriate topic. A solution would be to make the task of classifying a document a user task. However, every Web user who can create an HTML document is not a knowledge engineer, and should not have to know the details of a representation or all of the ramifications of a given semantic classification.

Instead, we propose developing a system which allows the user to accomplish this task without having to interact with the details of the underlying knowledge structures the system exploits. Such a system should be able to suggest document classifications with some degree of accuracy, learn document indices relevant to the retrieval process, and learn salient features and semantic relations used in the classification process from the user, as a byproduct of user interaction with the system.

In this paper, we present a proposal for such a system, Q&A, which acquires and uses semantic organization and document indices for text classification and retrieval. In the following sections, we describe our preliminary work on Q&A. In section 3, we describe Q&A as it stands, and in section 4 consider the role of document classification within this system. In section 5, we discuss our agenda for future work. We conclude with section 6.

Q&A: a system for capturing, accessing and organizing memory

Q&A is a natural language question-answering and referral system. Q&A mediates interactions between an expert and a question-asking user. It uses its experience referring questions to expert users to answer new questions by retrieving previously answered ones. If a user's question is not found within the collection of previously answered questions, Q&A suggests a set of experts who are most likely to be able to answer the question. The system then gives the user the option of passing a question along to one or more of these experts. When an expert answers a user's question¹, the resulting question-answer pair is captured and indexed under a topic of the expert's choice for later use, and notification is sent to the user. By retrieving previously indexed questions from its knowledge base automatically, Q&A is able to reduce the amount of work typically associated with answering questions, while providing a natural way for users to access expertise. The tasks of retrieving a document or referring an expert are both strategies for achieving the expressed goal of a user as it is stated in a question. As a result, Q&A's learning is directed by the need to service this goal. In summary, a user can:

- Ask questions which will be answered by matching previously answered questions in Q&A's memory of question-answer pairs.
- Ask questions of experts directly.
- Browse an expert's topic hierarchy.

An expert can:

- Respond to a user's question and index that question under a topic.
- Add indices in the form of questions to previously indexed answers.
- Modify the contents of their topic hierarchy.
- Forward a user's question to another expert.

All of these interactions are opportunities for Q&A to learn new information about experts, their expertise and a particular question-asking user.

Q&A is a case-based system. We view question answering as a memory retrieval task. As in *FAQFINDER* (Burke *et al.* 1997), a question is treated as an index for the knowledge contained in the answer. Thus, when Q&A captures a question-answer pair, it is simultaneously acquiring codified expertise (an answer to a question) and an index to that expertise that can be used for retrieval and classification (the question itself). Similarly, when experts classify a question-answer pairs under topics, Q&A captures a forest of semantic hierarchies. In Q&A, topics are considered labels for a defining semantic similarity among the question-answer pairs beneath them. Topics not only provide a semantic organization. Because of their role in the browsing process, the text in topic labels is highly predictive of the knowledge contained beneath that topic. As a result of both the predictive nature of topic labels and the functional role they play in the retrieval process, the text in topic labels is a good candidate for indexing vocabulary (Kolodner 1993) and as such is used as an element of a predictive index (Kulyukin, Hammond, & Burke 1998).

Answering a question in Q&A is a two-stage process. First, the system attempts to classify the question under a topic. Then it retrieves the most similar questions under that topic as in (Burke *et al.* 1997), using shallow semantics combined with statistics (cosine in the vector-space (Salton & McGill 1983) model) to achieve a greater degree of recall. The complexity of retrieval is significantly reduced in this paradigm because we need only consider questions in a relevant topic. Classifying a new question-answer pair and retrieving an old one are essentially the same process, using the same structures in memory. If a similar question cannot be found in the collection of question-answer pairs, the system refers the user to experts associated with the topic of the question. Thus, text categorization, learning semantic hierarchies through interaction with experts, and combining these hierarchies are central problems.

¹ Answers can be in the form of URLs or typed responses.

Work in Progress: Text Categorization for Q&A

We view the task of categorizing a text under a topic in Q&A as similar to the task of exemplar-based classification described in (Bareiss, Porter, & Holte 1990). A topic for a question-answer pair (in the case of suggesting a topic to the expert) or a question (in the case of retrieval) is chosen based on its similarity to other question-answer pairs indexed under that topic. Our proposed method for classification in Q&A is hierarchical. At each node in the topic hierarchy, Q&A will consider the most representative terms of each subtopic, determined by the question-answer pairs indexed under that topic. To limit the number of terms to consider, our preliminary algorithms use a stop list to eliminate the most common words, and morphology analysis backed by WordNet (Miller 1995), to transform words to their base forms² as in (Kulyukin, Hammond, & Burke 1998). For example, “ran” gets converted to “run” and “jogging” to “jog”. Our algorithm then visits the subtopic that has the most representative terms in common with the document it is attempting to classify and drives down. It has found a classification for a text when it is unable to discriminate any further, or when it has reached a leaf. We are currently working on an algorithm which statistically learns these distinguishing surface-level features of topics at each level in a hierarchy, building what amounts to a discrimination network (Feigenbaum 1963) that supports both the retrieval and classification processes. The central problem for us is thus finding these representative (and predictive) document features and discerning under what contexts they apply.

In order to empirically evaluate classification algorithms, we needed a large online corpus of documents classified in a topic hierarchy. We chose to mine the Yahoo! (Yahoo! 1995) subtree of documents on Computer Science for this task. We gathered some 35241 documents under 2433 topics. Our measure of an algorithm’s success at the task of classification is whether or not it correctly classifies documents under this subtree of the Yahoo! hierarchy, given varying percentages of previously classified documents. Preliminary evaluations of both previous hierarchical classification algorithms (such as (Koller & Sahami 1997)) and work on novel classification algorithms for this purpose show that the organization reflected in Yahoo! is not nearly as well-behaved as it is in more typically used collections such as MedLine (Hersh *et al.* 1994). Moreover, we predict that algorithms which consider only correlational relationships between terms will perform disastrously on all but the most well-behaved document collections. In Yahoo!, for example, while topics are indeed labels for common semantic features of the documents beneath them, they in no way indicate the *kind* of similar semantic feature they label. The category “Bibliographies”, for instance, denotes both a struc-

tural and a functional semantic similarity among the documents below it. It does not, however, restrict the subject of the bibliographies (which is reflected in term frequencies), except that they are all related to computer science, because they are located in the computer science subtree. In contrast, on the same level in the hierarchy, the topic “Artificial Intelligence” denotes semantic similarity based on the concepts covered within those documents below it, while the function and structure of the documents underneath is not expressed or limited. This underscores the need for a richer representation of documents and semantic categories in terms of their content, function, and structure in order to reason about classification and document similarity in an intelligent way.

Discussion and Future Work

Our preliminary work on document classification has outlined a clear direction for future work: developing a representation for documents and topics as well as a domain theory for reasoning about document classification and document similarity in a general context. Several open questions then need to be addressed.

First, how can we attempt to make the process of representing a document semiautomatic? Web documents *are* already structured to a certain extent. Using heuristic approaches to reason about document features with respect to the HTML mark-up language has previously been shown to be a good indicator of both the structure of a document and the importance of a given term with respect to that document (Etzioni 1996). In a sense, we are already doing this. For example, in Q&A, terms which occur in questions are treated as more predictive for retrieval purposes. Our experiments with statistical classification algorithms show, however, that we need to do much more in order to produce a viable classification system.

Next, how can we best use the experts as a resource for learning and verification? Because Q&A is necessarily semiautomated (as a result of its knowledge acquisition process) we can use the question-answering experts as a greater resource than we currently do. While we believe that a statistical analysis of the co-occurrence of words among subtopics will play a role in the large-scale classification of documents, we also believe that we can do much better if we use experts as a source for classification explanation and document annotation. Having experts explain classifications, correct faulty automatic classifications (when the system suggests a topic to the expert for a new question-answer pair), and correct faulty referrals (Kulyukin, Hammond, & Burke 1998) will complete what (Kolodner 1993) has called the *learning cycle*. As in (Bareiss, Porter, & Holte 1990), we can expect that such additions to the system will improve both the classification and retrieval process dramatically.

Finally, we must consider how to provide the user a way to express their goals as an information seeker in a more detailed and directed way. Combining

²WordNet is a trademark of Princeton University

statistical measures of document similarity and these more knowledge-based approaches will be an interesting problem to address.

Conclusion

In summary, as a classification and retrieval system, Q&A is different in the following ways.

1. It employs a hybrid of statistics and semantic, knowledge-based techniques.
2. It views question answering as a memory retrieval task. As a result, the semantic organization of its memory of question-answer pairs is extremely important to both the efficiency and the accuracy of retrieval.
3. It is semiautomatic. Because we have access to experts as question answerers, we can use them as a resource for verification, explanation, and knowledge acquisition.
4. It learns.
 - (a) It captures new knowledge as questions are answered and organized by experts.
 - (b) It learns new indices when experts create new topics and associate old answers with new questions.
 - (c) It updates its memory of concept representations when new documents are classified and new topics are created.

The amount of knowledge it acquires increases as a function of its use.

In this paper, we have identified several substantial problems with the state of the art in text classification systems. Namely, that building a strictly knowledge-based classification system for documents in general is infeasible, but that systems based on correlations among term frequencies alone simply do not work very well on unengineered document collections. Finally, we have outlined a new approach which attempts to overcome these problems by integrating insights from case-based reasoning with both of the previous kinds of systems, and by using the diversity and ubiquity of the Web and its users.

References

- Bareiss, R.; Porter, B.; and Holte, R. 1990. Concept Learning and Heuristic Classification in Weak-Theory Domains. *Artificial Intelligence Journal* 67(2):287–328.
- Burke, R.; Hammond, K.; Kulyukin, V.; Lytinen, S.; Tomuro, N.; and Schoenberg, S. 1997. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. Technical Report TR-97-05, University of Chicago, Department of Computer Science.
- Etzioni, O. 1996. The World Wide Web: quagmire or gold mine? *Communications of the ACM* 39(11).

Feigenbaum, E. 1963. The simulation of natural learning behavior. In Feigenbaum, E., and Feldman, J., eds., *Computers and Thought*. New York: McGraw-Hill.

Hersh, W. R.; Buckley, C.; Leone, T. J.; and Hickam, D. H. 1994. OHSUMED: An interactive retrieval evaluation and new large text collection for research. In *Proc. SIGIR-94*, 192–201.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *ICML-97: Proceedings of the Fourteenth International Conference on Machine Learning*, 170–178. San Francisco, CA: Morgan Kaufmann.

Kolodner, J. 1993. *Case-Based Reasoning*. San Francisco, CA: Morgan Kaufmann.

Kulyukin, V. A.; Hammond, K. J.; and Burke, R. D. 1998. Answering questions for an organization online. In *Proc. AAAI '98*.

Miller, G. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11).

Salton, G., and McGill, M. 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill.

Yahoo! 1995. <http://www.yahoo.com/>.