Information Retrieval using Patterns and Relations

Boris Katz, Igor Kaplansky, Rebecca Schulman & Ali Ibrahim

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: We are working on designing an information retrieval system that answers questions about a large text collection by employing a pattern matching language to identify certain syntactic relations and other types of information about the text, such as people, places, times, and other simple information.

Motivation: Search engines have become the primary tool to search through large, poorly organized archives of information, such as the World Wide Web, newspaper archives, and electronic books. Search engine technology is currently limited in most cases to keyword searches. Pattern based indexing is an attempt to improve on this sort of system, by indexing not only keywords, but also syntactic relations, and occurrences in a text of items such as a person, a time, or a location.

Previous Work: The idea of text pattern indexing is an extension of previous work in information retrieval that has centered around indexing bigrams or phrases of words, in addition to individual words.

Approach: We separate the information retrieval problem into two parts. First, patterns to extract interesting information from the document are created, and this information is indexed. After that a query generator receives questions from a user and formulates queries to the pattern database. Hence the system is comprised of the following parts:

1. a simple indexer which creates a word based database of the text sources.
2. a pattern-based indexer which finds and stores relations and patterns based on lists of words and regular expressions.
3. a query engine, which takes questions from a user, generates patterns from them, and retrieves documents using the generated patterns.

We created a pattern language in order to construct complicated text patterns easily. Pattern elements may be a regular expression, a list of terms (any one of which is a match), a subroutine, or a collection of patterns. Patterns maybe collected into sets of patterns that must be matched consecutively, or may be matched alternatively. Using this language we have written patterns to find names of people, numbers, distances, and places, among other things.

Since indexed patterns are specific to a particular data source like encyclopedia or a scientific journal, the query engine can use this fact to direct the query to the source which is most likely to contain the necessary information. After determining the right source, a query engine would identify relations and patterns contained in the question and use all available indices and a scoring mechanism to retrieve the best answer.

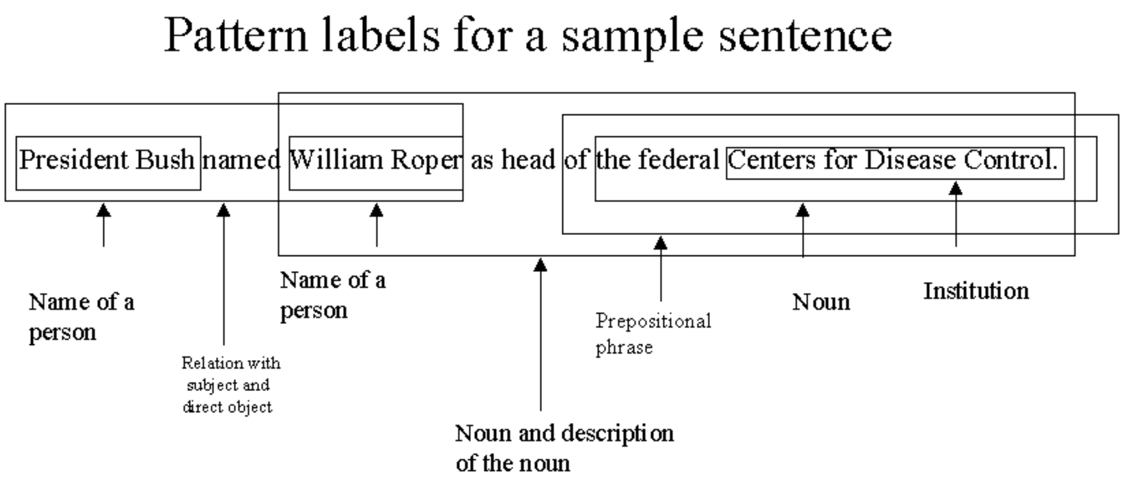
Difficulty: Correctly identifying even simple patterns is not easy because it is not always clear where to draw the boundaries. In addition, patterns may overlap or be nested. Assuming selected patterns have been successfully identified, an efficient representation is necessary to store and retrieve them.

Impact: Improving search engine technology will allow more efficient searching of large document collections. In situations where too much information is available, the ability of a program to effectively pick out important information is crucial. Indexing documents by their content rather than their textual representation will allow better searching.

Our information retrieval system can also be used by START. Currently, in responce to a question that START understands, it returns a page that contains a wealth of information. When asked a specific question, based on that information, START often fails to answer since it has no knowledge about the information contained within the page.

Using our system for indexing all of START's knowledge base will give START access to the information it already possessed, thus allowing it to answer a wider range of questions.

Future Work: We will increase the number and precision of patterns that the information retrieval system will recognize and use to answer questions. Further work on our nascent system for recognizing anaphors (words such as "he" or "the chairman" that refer to things mentioned earlier) and indexing them to the original term will significantly improve our information retrieval. To make sure that our work is scalable to cover large amounts of text, we will experiment with different indexing schemes to find the optimally efficient representation.



Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:

[1] B. Katz. A Three-step Procedure for Language Generation. A.I. Memo 599, MIT AI Laboratory, 1980.

[2] B. Katz. Using English for Indexing and Retrieving. in P. H. Winston and S. A. Shellard (eds.), *Artificial Intelligence at MIT: Expanding Frontiers,* V. 1, MIT Press, 1990.

[3] B. Katz. "Annotating the World Wide Web using Natural Language." Proceedings of RIAO'97, Montreal, Canada (1997).

[4] B. Katz, et al. Omnibase: A Universal Data Source Interface. *MIT Artificial Intelligence Laboratory Research Abstracts, September 2000* (this volume).