Snowball: A Prototype System for Extracting Relations from Large Text Collections

Eugene Agichtein, Luis Gravano, Jeff Pavel, Viktoriga Sokolova, Aleksandr Voskoboynik

Computer Science Department
Columbia University
{eugene,gravano,jeff,vicky,av69}@cs.columbia.edu

Text documents often hide valuable structured data. For example, a collection of newspaper articles might contain information on the location of the headquarters of a number of organizations. If we need to find the location of the headquarters of, say, Microsoft, we could try and use traditional information-retrieval techniques for finding documents that contain the answer to our query. Alternatively, we could answer such a query more precisely if we somehow had available a table listing all the organization-location pairs that are mentioned in our document collection. One could view the extraction process as automatically building a materialized view over the unstructured text data. In this demo we present an interactive prototype of our Snowball system for extracting relations from collections of plain-text documents with minimal human participation. Our method builds on the DIPRE idea introduced by Brin [3]. Our system and techniques were presented in detail in [2] and [1].

Figure 1: The main components of Snowball.

The basic architecture of Snowball is shown in Figure 1. Initially, we provide Snowball with a handful of instances of the tuples in the desired relation. Our system searches for occurrences of the example tuples in the documents, identifying text contexts where entities of the appropriate type (e.g., organization and location names) appear together. The system learns extraction patterns from these example contexts. The patterns are then used to scan through the collection, which results in new tuples being discovered. The new tuples are evaluated, the most reliable ones are used as the new seed tuples, and the process repeats.

A crucial step in the extraction process is the generation of patterns, which is accomplished by grouping the occurrences of known tuples in documents that occur in similar contexts. More precisely, Snowball generates a term vector for each text context where a seed tuple occurs, and then clusters these vectors using a simple single-pass bucket clustering algorithm. Patterns are represented as cluster centroids.

Using these patterns, Snowball scans the document collection to discover new tuples. The system first identifies text segments that include the entities of the appropriate type. Then, for each text segment, a most similar pattern is found, and if the similarity is higher than a threshold, a candidate tuple is generated. For each candidate tuple, we store the set of patterns that generated it. Snowball assigns a weight to extraction patterns based on their selectivity (estimated during our scan of the corpus to discover new tuples), and trusts the tuples that they generate accordingly. From these, the most reliable tuples are selected as seed for the next iteration of the system.

Demo Interface and Operation. In our demo, we present a prototype Snowball system that operates over a local collection of documents. The prototype includes a graphical user interface written in Java that allows users to specify a relation to be extracted, to examine the statistics on patterns and tuples as they are generated, and to explore different extraction parameters. Initially, the user selects the types of entities (e.g., Organization, Location, or Person’s Name) in a desired relation to be extracted (e.g., Located-in, President-of, Employee-of, Competitor-of), and a set of seed tuples for the relation. The system examines the contexts in which the seed tuples appear to generate extraction patterns. These patterns are displayed as they are learned, and the statistics for each pattern (including the distribution of supporting tuples) are graphically displayed. Using the extraction patterns, Snowball scans the collection to extract new candidate tuples, displaying the most reliable tuples found in any point in time. The output of the system is a table of extracted tuples of the desired relation.

Please refer to http://snowball.cs.columbia.edu for more information about the Snowball system and this demo.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grants No. IIS-97-33880 and IIS-98-17434.

