Corroborating information from multiple sources

Minji Wu
Rutgers University

9/14/2015 at 11:30 am
Core B (Room 305)

Abstract

Information on the Internet is abundant but often inaccurate. Given a query that has a unique answer (as opposed to a Web query against a search engine), different Web sources might provide multiple conflicting answers. As a result, users are left with the burden of validating the correctness of the answer from each source. In order to tackle this problem, corroboration techniques have been proposed in order to identify the correct answer given a set of candidate answers extracted from the sources. Corroboration is the technique that evaluates the quality of the answers by considering the trustworthiness of the sources from which the answers are extracted. Intuitively, an answer extracted from a trustworthy source is more likely to be the correct answer. In return, the more correct answers it reports, the more trustworthy a source is.

Unfortunately, several challenges arise before we can successfully apply a corroboration technique to find the correct answer to a query. First of all, the prime challenge is how to evaluate the trustworthiness of the sources and henceforth derive the quality of an answer based on the sources reporting it. Secondly, in a case where all the sources agree on a single candidate answer, how to validate the correctness of the answer. Third, in an application in which each source only provides a partial answer and the final answer is a combination of partial answers from multiple sources, how to evaluate the quality of answers and how to efficiently compute the correct answer.

This thesis studies real world problems that involve information from multiple sources. We demonstrate that using information from a single source is often of low quality and in some cases insufficient. We discuss the challenges in each individual problem and present novel corroboration algorithms that efficiently compute scores for the answers by taking into consideration of the trustworthiness of the sources.