

Corroborating Information from Disagreeing Views*

Alban Galland
INRIA Saclay – Île-de-France
LSV ENS Cachan
alban.galland@inria.fr

Amélie Marian
Rutgers University
amelie@cs.rutgers.edu

Serge Abiteboul
INRIA Saclay – Île-de-France
LSV ENS Cachan
serge.abiteboul@inria.fr

Pierre Senellart
Institut Télécom; Télécom
ParisTech; CNRS LTCI
pierre.senellart@telecom-paristech.fr

ABSTRACT

We consider a set of views stating possibly conflicting facts. Negative facts in the views may come, e.g., from functional dependencies in the underlying database schema. We want to predict the truth values of the facts. Beyond simple methods such as voting (typically rather accurate), we explore techniques based on “corroboration”, i.e., taking into account trust in the views. We introduce three fixpoint algorithms corresponding to different levels of complexity of an underlying probabilistic model. They all estimate both truth values of facts and trust in the views. We present experimental studies on synthetic and real-world data. This analysis illustrates how and in which context these methods improve corroboration results over baseline methods. We believe that corroboration can serve in a wide range of applications such as source selection in the semantic Web, data quality assessment or semantic annotation cleaning in social networks. This work sets the bases for a wide range of techniques for solving these more complex problems.

Categories and Subject Descriptors

H2.5 [Database Management]: Heterogeneous Databases;
H2.8 [Database Management]: Database Applications—
data mining

General Terms

Algorithms, Experimentation

Keywords

Corroboration, view, confidence, probabilistic model, fixpoint, contradiction

*This work has been partially funded by the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) / ERC grant Webdam, agreement 226513. <http://webdam.inria.fr/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM’10 February 4–6, 2010, New York City, New York, USA
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...\$10.00.

1. INTRODUCTION

The Web provides an interface to access a wide variety of information and viewpoints from individual Web sources that have different degree of trustworthiness based on their origin or bias. The most daunting problem when trying to answer a question seems not to be *where* to find an answer, but *which* answer to trust among the ones reported by different Web sources. This happens not only when no true answer exists, because of some opinion or context differences, but also when one or more true answers are expected. Such conflicting answers can arise from disagreement, outdated information, or simple errors.

Simple questions often yield disagreeing answers from different sources. As an example, the birth date of Napoleon Bonaparte, a contentious topic of importance to historians as it determines whether Napoleon was born French or Italian, is reported as August 15, 1769 or as January 7, 1768 depending on the sources. A more familiar everyday example is a simple professional contact information search: contact information is time-dependent; yet because of the nature of Web sources, many sources will continue to list outdated information if a person has switched jobs. For instance, as of the writing of this paper, a Google search for “Mor Naaman” lists three possible affiliations in the first ten results: Stanford University, Yahoo! Research Berkeley, and SCILS, Rutgers University. The correct current affiliation, SCILS, does not appear in first position. In addition, sources may identify the object incorrectly; in the case of a contact search this can happen in the presence of homonyms (the first page of Google results for “Mor Naaman Facebook” returns two separate Facebook profiles), misspellings or name changes.

We consider each Web source as a separate view over the data. To accurately answer a question in the presence of conflicting information, a natural approach is to simply count the number of occurrences of each answer, i.e., the number of views reporting each answer. This simple voting strategy performs well in many scenarios but is easily misguided in a Web environment where many sources can either malignantly collude to propagate false information, or naively replicate outdated or wrong data. The quality of the views should then be taken into account when corroborating answers to identify the best answer to a query. Without *a priori* knowledge on the quality, or trustworthiness, of views, or on the correctness of answers, we are left with a recursive definition: a correct answer is returned by many trusted views and a trustworthy view returns many correct answers.

