Corroborating Information from Web Sources

Amélie Marian, Minji Wu

Department of Computer Science, Rutgers University, New Brunswick, NJ, USA

{amelie,minji-wu}@cs.rutgers.edu

Abstract

Information available on the Internet is abundant but often inaccurate. Web sources have different degrees of trustworthiness based on their accuracy, freshness, origin or bias. Web users are then left with the daunting task of assessing the correctness of possibly conflicting answers to their queries. In this paper, we present techniques for corroborating information from different web sources. We discuss techniques that estimates the truthfulness of answers and the trustworthiness of the sources based on an underlying probabilistic model. We show how to apply data corroboration to a web setting where data sources can have multiple forms, all with various quality issues: individual web sites, search engine query results, user reviews, map and street view data, and social tags.

1 Introduction

Traditional query processing techniques assume correctness of the information provided by an underlying data source (e.g., a database, a document collection). For many data needs, accessing a single trusted source of information is no longer possible; the information needs to be gathered from several, often completely independent sources. This is particularly true when accessing Web sources, which provide a variety of information and viewpoints with different degree of trustworthiness based on the sources' origin or bias. When different sources provide conflicting or incomplete information, guaranteeing good query answer quality can be challenging. 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 Such conflicting answers can arise from disagreement, outdated information, or simple errors.

Using corroborative evidence, in the form of similar answers from several sources, is an intuitive way to increase trust in the returned answers. The use of data corroboration can significantly improve query answer quality in a wide variety of scenarios [14, 4, 7, 13, 8].

In this paper we discuss different approaches to corroborate data (Section 2). We show how data corroboration can improve data quality in a web source scenario, where data sources can have multiple forms, all with various quality issues: individual web sites, search engine query results, user reviews, map and street view data, and social tags (Section 3). Finally, we discuss open issues in data corroboration (Section 4).

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Background

The problem of identifying the best outcome from multiple sources of information is not a new one. Voting theory and multi-criteria decision making have been subjects of research for centuries. Data corroboration is a natural way of dealing with conflicting information, and has been used in various aspects of society: justice, journalism, gossip. The more (independent) sources corroborate a fact, the more likely the fact is true, assuming the sources are trustworthy. Unfortunately, many web sources are not trustworthy, because of erroneous, misleading, biased, or outdated information.

In this section we describe some of the recent advances in data corroboration and comment on some related work.

Web Sources. In [13], we developed corroboration scoring techniques for web data in order to save users the hassle of manually checking query-related web sites for corroborated answers. The existence of several sources providing the same information is then viewed as corroborating evidence, increasing the quality of the corresponding information. We score each answer by taking into account the number, relevance, and originality of the sources reporting the answer as well as the prominence of the answer within the sources. Our results show that corroboration-based methods provide significant improvements in answer quality compared to simple frequency-based approaches.

TRUTHFINDER [14] assigns confidence to web sources by recursively computing confidence in the sources based on the expected truth of the facts they report. Presence of similar facts in different sources is then seen as positive reinforcement of the expected truth of the fact.

Probabilistic Model. A probabilistic data model for corroboration was introduced in [7] to take into account the uncertainty associated with facts reported by the sources, the possibility of conflicting information coming from separate sources, as well as the limited coverage of the sources. The model sets the bases for a systematic study of trust-based corroboration, where trust values are assigned to sources of information and used to infer truth of facts reported by the source. In addition, the probabilistic model was used as the basis for corroboration algorithms based on fixpoint computation techniques that derive estimates of the truth value of facts reported by a set of sources, as well as estimates of the quality of the sources. The resulting corroboration techniques consistently outperformed existing voting-based strategies.

Several theoretical works have focused on estimating the probability of an event in the presence of conflicting information. Osherson and Vardi [11] study the problem of inconsistent outcomes when aggregating logic statements from multiple sources. Their goal is to produce a logically coherent result. Work in subjective logic and trust management [9] consider the issue of trust propagation from one source to another, in a model where the sources are not independent.

Statistics. The expectation-maximization (EM) algorithm [3] was presented in 1977 to find the maximum likelihood estimates of parameters in a model containing latent variables. It was implemented [2] to take into account observer (source) errors to assess performance of individual observers and derive an agreement in classification tasks. More recently, it was adapted to estimate worker quality in Mechanical Turk experiments [8], and the derived quality used to generate agreements. A limitation of the EM algorithm is scalability, which makes it difficult to apply to a web source scenario.

Question Answering. Question answering systems, such as [1, 10, 6] consider the frequency of an extracted answer as a measure of answer quality. However, these techniques rely mostly on redundancy of information and do not consider the trust associated with each extraction source to score extracted answers.

3 Case Study: Identifying Good Business Listings

We now rely on the web to get details about the location and contact information of local businesses. While the old-fashioned paper-based yellow pages did have inaccuracies: places gone out business, new stores not yet included, several names, preferably starting with the letter "A," —an early spamming technique— for the same business; overall the information available was of high quality as businesses had to pay for their listing to appear and a human would typically check the listing before it went to print. In contrast it is easy and cheap to add information to the web through a web page, or by registering listings directly on services such as Google Places. This gives the opportunity to make more up-to-date and comprehensive listings available to the users, but the sheer number of businesses makes it impossible to verify the accuracy of each of these listings manually.

We consider the use of data corroboration techniques in a restaurant listings scenario, in which our goal is to assess whether a reported listing ((restaurant,address) pair) is true (correct) or false (incorrect). We consider a wide variety of data sources, and harness the power of the web 2.0 to improve data quality. Data corroboration allows us to (1) assess the accuracy of web business listings, and (2) assess the quality of the sources from which the listings are derived to identify sources of spam and determine the usefulness of each source's contribution.

We corroborate information from traditional web sources, user reviews and web 2.0 data (e.g., user check-ins, restaurant reservation portals, mechanical turkers assessments of the visibility of the business at the listing's address on Google Streetview). By themselves, none of these sources offer definite information as to whether the listing is correct or not. Similarly, the fact that some of these sources may not have information for a business does not mean that the listing is inaccurate. By corroborating information from different sources and considering the number of sources sharing similar information we can provide an estimated truth for each of the listings in the sample.

3.1 Data Model

We consider a probabilistic data model similar to that described in [7]. Sources of data can be seen as views over the real world W. Views report beliefs that are positive or negative statements. Based on these beliefs, our task is to "guess" what the correct values for the real world W are.

Let $\mathcal{F}$ be a set $\{f_1 \dots f_n\}$ of *facts*. A *view* (over $\mathcal{F}$) is a (partial) mapping from $\mathcal{F}$ to the set $\{T, F\}$ (T stands for *true*, and F for *false*). We consider a set of views $\mathcal{V} = \{V_1 \dots V_m\}$ from which we try to estimate the real world W, defined as a total mapping from $\mathcal{F}$ to the set $\{T, F\}$. From a mathematical viewpoint, based on some probabilistic model, we want to estimate the most likely W given the views. (Note that views can also report no belief for a given fact, in which case we consider the fact unknown in the source ($V_i(f_j) = ?$).

We implemented the COSINE data corroboration strategy [7].COSINE is a heuristic approach, based on the classic cosine similarity measure that is popular in Information Retrieval, that estimates the truth values of facts and the trustworthiness of views with values between -1 and 1, where -1 means false facts, systematically wrong views, 0 means indeterminate facts, views with random statements, and 1 means true facts, perfect views. The idea is then to compute, for each view V_i, given a set of truth values for facts, the similarity between the statements of V_i, viewed as a set of ± 1 statements on facts, and the predicted real world.

3.2 Experimental Settings

We have extracted a sample of New York City restaurant listings from three web sources *YellowPages, CitySearch* and *MenuPages*, a total of 1,000 distinct listings. We then cross checked each of the three sources to identify whether listings extracted from the other two sources appeared in them. For each 1,000 listing, if it exists in a source, then the source reports it as true; if it does not, then it is considered unknown by the source.

We also considered four sources of data:

Source coverage	Yelp	Foursquare	OpenTable	M. Turk	YellowPages	CitySearch	MenuPages
	0.52	0.42	0.05	0.91	0.65	0.70	0.41
Source overlap	Yelp	Foursquare	OpenTable	M. Turk	YellowPages	CitySearch	MenuPages
Yelp	1	0.58	0.09	0.49	0.43	0.44	0.55
Foursquare	0.58	1	0.11	0.40	0.35	0.40	0.51
OpenTable	0.09	0.11	1	0.05	0.05	0.07	0.10
M. Turk	0.49	0.40	0.05	1	0.61	0.64	0.40
YellowPages	0.43	0.35	0.05	0.61	1	0.46	0.26
CitySearch	0.44	0.40	0.07	0.64	0.46	1	0.37
MenuPages	0.55	0.51	0.10	0.40	0.26	0.37	1

Table 1: Source coverage and overlap

	Yelp	Foursquare	OpenTable	M. Turk	YellowPages	CitySearch	MenuPages
Source Trust	0.98	0.99	0.99	0.86	0.87	0.89	0.81

Table 2: Corroborated trust in the sources

- *Yelp*, a user reviewing web site. If a listing has a review in *Yelp*, then the source reports it as a true fact. We did a quick analysis of the text of the reviews to identify closed restaurants and report those as false. Listings for which there are no *Yelp* entry are considered unknown.

- *Foursquare*, a user check-in service. If a restaurant has associated entries, then it is reported true by *Foursquare*, otherwise it is reported unknown.

- *OpenTable*, a restaurant reservation service. If a restaurant is available for reservations in OpenTable, then it is considered true, if not it is unknown.

- *Mechanical Turk*, we considered a crowdsourcing approach where mechanical turkers were asked to manually check Google Streetview to see if the restaurant was located at the address. Unfortunately, it is well documented that *Mechanical Turk* data tend to be of low quality [8]. To prevent skew in the result due to spamming done by MT workers, we asked 5 workers to evaluate each listing and use a majority vote to assign true, false or unknown values.[1]

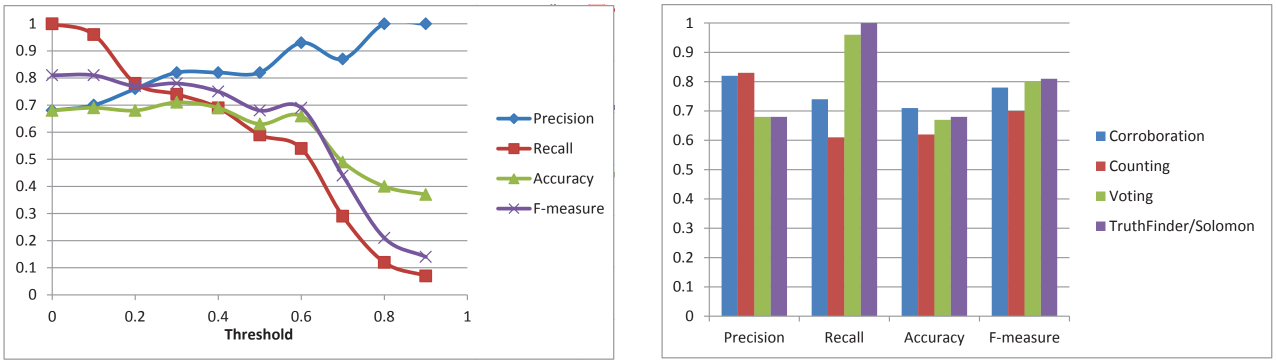
Table 1 reports the source coverage (the fraction of the sample that is contained in each source), as well as the source overlap (a measure of how much of the sample two sources have in common). All sources contain only a fraction of the domain. Considering several sources is then critical to have a large picture view of the data. *Mechanical Turk* has the largest coverage as it was specifically created by asking MT workers about each specific restaurant in the sample, nevertheless, 9% of the sample is not covered by this source as MT workers could not assess the presence of the restaurant at the specified address (missing Streetview, obstructed view, etc.). On the opposite, OpenTable only contains 5% of the sample as it is a commercial website that connects to real restaurant reservation systems.

3.3 Experimental Results

We corroborated our listing data using the COSINE algorithm presented in [7]. Table 2 shows the quality values for each source, as computed by the COSINE algorithm (trustworthiness value). Unsurprisingly, *OpenTable*,

[1] We integrated the EM-based algorithm from [8] to corroborate MT workers answers. Unfortunately, it did not perform as well as expected as our data set is highly skewed towards true listings, and our experiments had a majority of workers with few answers.

which connects to real restaurant reservation systems is of very high quality, but with low coverage (Table 1). The dynamic user-based sources *Foursquare* and *Yelp* also show high quality, whereas the directory sources *CitySearch*, *YellowPages* and *MenuPages* have lower quality, possibly because of stale results. *Mechanical Turk*'s quality is reasonable but not as good as hoped despite our postprocessing of its data through voting due to the low quality of individual workers.



(a) Precision, Recall, Accuracy and F-measure values when varying the corroboration threshold

(b) Comparison of the quality of the corroborated answers with that of standard voting techniques

Figure 1: Evaluation of the corroborated answers

We evaluate the quality of our corroborated answers using a "gold standard" test set of 100 restaurant for which we have identified the correct answer using a combination of *Mechanical Turk* experiments (with 10 workers) and manual evaluation. For each of the restaurant in the test set we compare its correct value with that reported by the corroborated approach.

We report on the Precision, Recall, Accuracy, computed as the fraction of correctly classified listings, and F-measure of the true facts (confirmed listings) varying the threshold truth value at which we classify listing as being true in Figure 1(a). Note that negative truth values are associated with negative facts (i.e., confidence that the restaurant is not at the listed address). Our sources mostly report positive facts, except for *Mechanical Turk* and *Yelp*, which reports a few restaurants as closed. As a result only eight listings have a final truth value lower than 0, and the Precision, Recall and Accuracy values for threshold values between -1 and 0 are equal to that reported at $threshold = 0$. As expected, the higher the threshold, the better the quality of the listings identified as true is (higher Precision), but more correct listings are classified as false (lower Recall). The highest Accuracy and F-measure values are found around a threshold of 0.3. (High F-measure values at threshold 0 and 0.1 are due to high Recall as almost all listing are classified as true.) In practice, this means that the results are more accurate when at least two or three sources report the listing as true with no dissenting view, as expected.

In Figure 1(b), we compare the corroborated results (at $threshold = 0.3$) with standard voting techniques: COUNTING , which assigns a listing as true if at least half the sources report it true, and VOTING , which assigns a true value to any listings that has more sources reporting it true than false. As expected VOTING has high Recall but low Precision, since any listing that is reported true by only one source is likely to be classified true. In contrast COUNTING has high Precision but low Recall since at least half of the sources have to report on a listing for it to be identified as true. We also compare our results with that of techniques that consider conflicting sources information to make truth assessments: `TRUTHFINDER` [14] and `SOLOMON` [4]. On our restaurant listing data set, both techniques report the same results. Interestingly, they return all listings as true; this mechanically results in perfect Recall and high F-measure values. We believe this is due to the fact that our restaurant listing sources report contain very few false values (only two sources, *Mechanical Turk* and *Yelp*, have false facts). `TRUTHFINDER` and `SOLOMON` will then not detect much conflicting information and will assign high trust levels to all sources, resulting in facts being considered true even if they are reported by only one source. In contrast, COSINE will assign lower scores to facts that are reported by only a few sources. By varying COSINE 's threshold value, we can tune our corroboration to require more source evidence to report a listing as true. As a result, the corroborated COSINE approach has the best Accuracy (percentage of correctly classified listings) over the five techniques.

4 Open Issues

We showed that data corroboration can effectively be used to identify answers in the presence of multiple incomplete and possibly conflicting sources. This work opens the road to many interesting research challenges:

- **Multiple Answers:** So far, most of the work on data corroboration [7, 13, 14] assume that there is only one valid answer for each fact (e.g., a listing is correct or not, a historical figure has only one valid birth date). In many cases, facts can have multiple answers (e.g., a person can have multiple phone numbers). The presence of multiple answers raise several interesting challenges, such as how many answers to consider based on the expected distribution of answers? How to adapt trust in the sources when all possible answers may be correct? How to rank, and possibly aggregate similar answers?

- **Functional Dependencies:** Many query scenarios have underlying functional dependencies. For instance if an email address can only be associated with one person, any source reporting a mapping (john, js@gmail.com) is stating implicitly that all other mappings ($person, js@gmail.com) are false. A simple way to model this would be to add a false statement for the mapping to js@gmail.com for every $person. This is not practical and creates a blowup of the number of facts. In addition, more complex functional dependencies are not simple to address, e.g., each person can only have one father and one mother. Integrating complex functional dependencies in a data corroboration model such as the probabilistic model of [7] raises interesting theoretical modeling issues.

- **Uncertain Data:** In addition to the trustworthiness of the source, we can also include another source of approximation in the model: uncertainty of the source. Sources providing results of ranking queries, belief databases, probabilistic databases are example of sources that report fact with an associated degree of confidence. How to integrate the uncertainty of the data, as reported by the source is also an interesting avenue to extend the corroboration model.

- **Domains:** In many cases, sources are specialized and the quality of a source should be assessed with respect to specific domains. This could be used for source selection during corroboration focusing on sources that have high quality rather than considering all possible sources [12] .

- **Time:** One of the main reasons for errors in web sources is due to outdated information (e.g., a restaurant has closed). We could leverage information from sources with timestamp data (e.g., user checkins, user reviews) in the corroboration, giving more weight to the "freshest" sources [5].

- **Social Network Trust:** Users originating queries may have a personal bias that they wish the corroboration to take into account. This could be preference towards sources sharing the same political views, or more trust in our friends beliefs. Including bias in the truth assessment can be done in several ways: by using biased facts as a training set of a recursive corroboration algorithm, or by heavily weighting trusted sources. A study of the tuning and impact of such bias may lead to interesting insights on how information propagates.

- **Source Dependence:** Recent work has focused on finding interdependencies, due to copying, between the sources [5, 4]. Ignoring such dependencies in corroboration can result in assigning more weight to sources that were heavily copied, regardless of their quality. Combining copying detection and corroboration is a promising direction for improving data quality.

References

[1] E. Brill, S. Dumais, and M. Banko. An analysis of the AskMSR question-answering system. In *Proc. of the ACL conference on Empirical methods in natural language processing (EMNLP'02)*, 2002.

[2] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] X. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proc. VLDB*, Lyon, France, 2009.

[5] X. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. In *Proc. VLDB*, Lyon, France, 2009.

[6] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *Proc. of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, 2005.

[7] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. of the Third ACM International Conference on Web Search and Data Mining (WSDM'10)*, 2010.

[8] P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proc. of the Second Human Computation Workshop (KDD-HCOMP 2010)*, 2010.

[9] A. Jøsang, S. Marsh, and S. Pope. Exploring different types of trust propagation. In *Proc. Trust Management*, Pisa, Italy, May 2006.

[10] C. C. T. Kwok, O. Etzioni, and D. S. Weld. Scaling question answering to the web. In *Proc. of the 10th International Conference on the World Wide Web (WWW'01)*, 2001.

[11] D. Osherson and M. Y. Vardi. Aggregating disparate estimates of chance. *Games and Economic Behavior*, 56(1):148–173, July 2006.

[12] A. D. Sarma, X. L. Dong, and A. Halevy. Data integration with dependent sources. In *Procȯf the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, 2011.

[13] M. Wu and A. Marian. A framework for corroborating answers from multiple web sources. *Information Systems*, 36(2), 2011.

[14] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of the 13th ACM International Conference on Knowledge Discovery and Data Mining (KDD'07)*, 2007.