Reconstructing Human-Generated Provenance Through Similarity-Based Clustering

Tom De Nies, Erik Mannens, and Rik Van de Walle

In: Provenance and Annotation of Data and Processes, 9672, 191-194, 2016.

http://link.springer.com/chapter/10.1007/978-3-319-40593-3_19

Reconstructing Human-Generated Provenance Through Similarity-Based Clustering

Tom De Nies(✉), Erik Mannens, and Rik Van de Walle

Ghent University – iMinds – Data Science Lab, Ghent, Belgium
{tom.denies,erik.mannens,rik.vandewalle}@ugent.be

Abstract. In this paper, we revisit our method for reconstructing the primary sources of documents, which make up an important part of their provenance. Our method is based on the assumption that if two documents are semantically similar, there is a high chance that they also share a common source. We previously evaluated this assumption on an excerpt from a news archive, achieving 68.2 % precision and 73 % recall when reconstructing the primary sources of all articles. However, since we could not release this dataset to the public, it made our results hard to compare to others. In this work, we extend the flexibility of our method by adding a new parameter, and re-evaluate it on the human-generated dataset created for the 2014 Provenance Reconstruction Challenge. The extended method achieves up to 86 % precision and 59 % recall, and is now directly comparable to any approach that uses the same dataset.

1 Introduction

Even with the recommendation of the PROV model by W3C in 2013, there is still a plethora of data on the Web that lacks associated provenance. Research that works towards reconstructing this provenance is still very new in the community, and datasets suitable for evaluation are rare. Thus, together with VU Amsterdam, we initiated the 2014 Provenance Reconstruction Challenge[1]. The aim of this challenge was to help spur research into the reconstruction of provenance by providing a common task and datasets for experimentation. In this paper, we present our own evaluation results on this dataset.

2 The Dataset

Challenge participants received an open data set and the corresponding provenance graphs (in W3C PROV format). They could then work with the data trying to reconstruct the provenance graphs from the open data set. The data consists of two distinct sets: one machine-collected, and one human-generated. This way, we are able to evaluate the reconstruction accuracy for provenance that was automatically collected based on observations, and provenance that

[1] http://www.data2semantics.org/prov-reconstruction-challenge/.

M. Mattoso and B. Glavic (Eds.): IPAW 2016, LNCS 9672, pp. 191–194, 2016.
DOI: 10.1007/978-3-319-40593-3_19

was generated based on information provided by humans, which could not be captured automatically.

The machine-collected dataset can be downloaded at: http://git2prov.org/reconstruction/machine-generated-dev.zip, and the human-generated set at: http://git2prov.org/reconstruction/human-generated-dev.zip.

The ground truth (*groundtruth.ttl*) for the machine-generated dataset was generated from a number of Github repositories using the Git2PROV tool [3]. As raw data, it includes every version of each file that was ever present in the repository (including deleted files). However, the filenames are randomized, to simulate a scenario where all provenance was lost. Due to these randomized filenames, the timing metadata associated with the files may differ from the original. The correct timings can be found in the ground truth provenance. The main goal here is to reconstruct the derivation graph of the original files, serialized as PROV-O. Evaluations should report at a minimum the precision/recall of the detected PROV relations (`prov:wasDerivedFrom`, `prov:wasGeneratedBy`, etc.).

The ground truth for the human-generated dataset was created using the sources mentioned in news articles from *WikiNews*. The link between news articles and their sources is modeled using the `prov:hadPrimarySource` relation. The raw data consists of the entire HTML of the WikiNews articles, without the sources, and a list of URIs (*human_sources.txt*). In other words, the goal of this task is to match the source URIs from this list to the correct WikiNews article. Approaches may use any information embedded in the files or external information, save from the ground truth or WikiNews, for obvious reasons. Evaluations should report at a minimum the results of precision/recall of the `prov:hadPrimarySource` relations.

3 Our Approach

We applied our method as described in [2], applying the assumption *"if a set of documents is highly similar, there is a high chance they also share a common source"*. This method clusters all documents in the dataset using a lower bound on similarity, expressed as the threshold T_s. Then, for each cluster, the oldest document is selected, and asserted as the (indirect) primary source of all others in that cluster. Note that clusters can overlap, so multiple primary sources can be asserted for one document. The level of uncertainty is annotated using the similarity measured between the two documents to help end-users make a decision on which assertion to trust, if there is a conflict. As parameters, we used the **cosine similarity with TF-IDF weighting**, **10 different the similarity thresholds** T_s, and **no cluster-size threshold** (so no re-clustering). Additionally, the following considerations were made during the implementation:

– For a number of articles which do not include a date, the original WikiNews articles were consulted, and the date reported there was used. In certain cases, this is the date of access by the writers of the article. Because a number of sources provide a datetime, while others only provide the day of publishing, *only the day of publishing* was used for all articles.

– We re-formatted the dataset to be usable with our software. To do this, the text and date had to be extracted from each HTML document, without advertisements, images, videos, etc. To obtain results that reflect the performance of our approach, not influenced by automatic text extraction methods, we performed this extraction manually, thereby assuming an *'ideal' text extractor*.

4 Evaluation

We evaluated our approach only on the human-generated dataset, for which it was primarily designed, and which is harder to capture in an automatic way. The results are shown in Table 1. At first glance, our method only achieved a rather disappointing maximum precision of 27 % and recall of 16 %. However, these results can be explained by looking deeper into how the human-generated dataset was constructed, and how our method tries to reconstruct it.

Table 1. Results of our method as described in [2] on the human-generated dataset

T_s	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.30	0.14	0.20	0.21	**0.27**	0.25	0	0	0	0
Recall	0.12	0.13	**0.16**	0.15	0.12	0.066	0	0	0	0

In our method as described in [2], we assume the *oldest document* in a cluster to be the (indirect) source of *multiple documents* – i.e., all others in the cluster. However, the ground truth dataset was constructed in exactly the opposite way: the *newest document* is derived from *multiple sources*. This means that with a very minor adjustment to our method, we might be able to achieve much better results. Therefore, we extended our method for this benchmark, by including a **new parameter** that allows the algorithm to select the *newest document* in every cluster instead of the oldest, and making all other documents in the cluster primary sources of the former. When we ran our reconstruction algorithm with this parameter enabled, it confirmed our suspicions, and we achieved much better results, as shown in Table 2. Now, our method achieves 86 % precision and 59 % recall with $T_s = 0.4$.

Table 2. Results of our slightly adjusted method on the human-generated dataset

T_s	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.52	0.54	0.70	**0.86**	0.77	0.69	0.2	0	0	0
Recall	0.26	0.51	0.57	**0.59**	0.33	0.18	0.016	0	0	0

5 Comparison to Related Work and Conclusion

While a number of domain-specific techniques used to reconstruct provenance exist, these techniques all predate the PROV standard and do not offer a comparable evaluation. For example, Zhao et al. [7] predict missing provenance based on semantic associations in the domain of reservoir engineering. Zhang et al. [6] exploit the logging capabilities of existing relational database management systems to retrieve lost source provenance traces. The work of [4,5] focuses on tracing news and quotes (referred to as *memes*) on the Web over time.

More recently, Aierken et al. [1] presented their multi-funneling approach to provenance reconstruction. They apply three techniques: one based on *IR techniques and the Vector Space Model (VSM)* similar to our approach, one based on the *machine learning and topic modeling*, and one based on *dynamic programming and matching the longest common subsequence*. They report a precision and recall of 77 % and 47 % for human-generated provenance, and 78 % and 68 % for machine-generated provenance, respectively. However, since their method relies heavily on training data, they used the human-generated challenge dataset as a training set for their method, and created a new WikiNews dataset using the same procedure for their evaluation. This means that while at first glance, our reported results seem to outperform theirs, they are not entirely comparable. However, their results together with ours – and the results we measured on our news dataset in [2] (68.2 % precision and 73 % recall) – can at least be interpreted as an indication of the level of accuracy that is achievable with the current state of the art in this field. While not perfect, these methods can certainly help a human-user reconstruct lost provenance, as opposed to doing it all manually.

References

1. Aierken, A., Davis, D.B., Zhang, Q., Gupta, K., Wong, A., Asuncion, H.U.: A multi-level funneling approach to data provenance reconstruction. In: IEEE 10th International Conference on e-Science, vol. 2, pp. 71–74. IEEE (2014)
2. De Nies, T., Coppens, S., Van Deursen, D., Mannens, E., Van de Walle, R.: Automatic discovery of high-level provenance using semantic similarity. In: Groth, P., Frew, J. (eds.) IPAW 2012. LNCS, vol. 7525, pp. 97–110. Springer, Heidelberg (2012)
3. De Nies, T., Magliacane, S., Verborgh, R., Coppens, S., Groth, P., Mannens, E., Van de Walle, R.: Git2PROV: exposing version control system content as W3C PROV. In: ISWC Posters & Demos, pp. 125–128 (2013)
4. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506. ACM (2009)
5. Simmons, M.P., Adamic, L.A., Adar, E.: Memes online: extracted, subtracted, injected, and recollected. In: ICWSM 2011, pp. 17–21 (2011)
6. Zhang, J., Jagadish, H.V.: Lost source provenance. In: 13th International Conference on Extending Database Technology, pp. 311–322. ACM (2010)
7. Zhao, J., Gomadam, K., Prasanna, V.: Predicting missing provenance using semantic associations in reservoir engineering. In: Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 141–148. IEEE (2011)