Reconstructing Provenance

Sara Magliacane
Advisors: Paul Groth, Frank van Harmelen

Department of Computer Science
VU University Amsterdam
s.magliacane@vu.nl

Abstract. Provenance is an increasingly important aspect of data management that is often underestimated and neglected by practitioners. In our work, we target the problem of reconstructing provenance of files in a shared folder setting, assuming that only standard filesystem metadata are available. We propose a content-based approach that is able to reconstruct provenance automatically, leveraging several similarity measures and edit distance algorithms, adapting and integrating them into a multi-signal pipeline. We discuss our research methodology and show some promising preliminary results.

1 Problem statement

The provenance of a data item is the metadata describing how, when and by whom the data item was produced. Provenance information is crucial for many applications, from data quality and aggregation to trust, and it has been researched from several perspectives (see surveys [9,12,20]).

In science, provenance helps scientists reproduce and repeat experiments. In business, understanding who made a decision, produced a document, or designed a product allows for effective accountability. However, since tracking provenance requires effort, it is often not done in real-world settings, resulting in collections of files with only basic metadata, e.g. timestamps. Thus, addressing these use cases becomes difficult or impossible.

In our work, we target the problem of reconstructing provenance of data in a shared folder setting, in which several authors can create or edit the data at different moments, and only standard filesystem metadata are available. Some of the data in the folder have been created by a sequence of operations on other data. The research questions we wish to answer are the following:

> How can one automatically, accurately and efficiently reconstruct the provenance of data in a shared folder, intended as the sequences of operations connecting the data?

A desirable solution should reconstruct provenance across multiple data types. It should be applicable also without domain-specific knowledge, while improving its accuracy in case this knowledge is available. Efficiency is intended both in terms of run-time performance and scalability. An additional desirable property

would be to produce the results with an anytime strategy, i.e. returning an approximated output at any time of the computation, in which the accuracy increases the longer we wait.

2 Related work

As we pointed out in [16], recently the issue of missing or incomplete provenance has attracted the attention of the provenance community and lead to few initial attempts to address this problem. On the other side, there are also several other fields that face similar problems and propose approaches that could be adapted to reconstructing provenance.

Related Work	**Reconstruction**	**Entities**	**Operations**	**Required knowledge**
Provenance in reservoir engineering [26]	Generating Process	Instances of concepts	Processes in the system	Previous executions
Provenance in network setting [15,2]	Dependency	Nodes	Sending information	Network structure
Provenance in stream processing [18]	Sequence	Tuples in data streams	Processes in the system	Coarse-grained provenance
Monitoring at OS-level [17,13]	Sequence	Application data	Application	OS-level reads and writes
Provenance as data mining [10]	Dependency	Text	Any on text	None
Provenance discovery using semantic similarity [22]	Sequence	Named Entities in Documents	Replacement, Generalization, Specialization, Addition, Omission	None
Text-reuse [7,4]	Dependency	Text	Any on text	None
Image Mining for Historical Manuscripts [6]	Dependency (same manuscript)	Images of historical manuscripts	Distortions on images	Library of known images
Edit distance [5,14]	Sequence	Strings, trees and graphs	Few and simple	None
Change detection [8]	Sequence	Hierarchically structured data	Few and simple	None
Ontology change detection [23]	Sequence	Ontologies	Low-level operations are similar to graphs	Rules for inferring high level changes
Web Service Composition [3,24]	Sequence based on user requirements	Inputs and Outputs	Web Services	Formal description of Web Services
Learn data transformations [25]	Sequence	Instances of semantic types	Any defined by grammar	Grammars of operations, More examples
Workflow Mining [1]	Sequence	Inputs and Outputs	Workflow components	Execution Traces

Table 1. Classification matrix of the related work

In Table 1, we take a broad view of reconstructing provenance and present a classification of the related work, listing in some cases only few representative examples of a field. The type of provenance that is reconstructed (column Reconstruction in Table 1) between the entities (column Entities in Table 1) can be:

– Dependency - the dependency relationship between two entities;

– Sequence - the sequence of operations that connect two entities;
– Generating Process - the process which created the entity;

The type of entities involved in the reconstruction ranges from text to data structures like graphs and ontologies. The type of operations involved in the reconstruction varies accordingly from simple operations, like inserting a node in a graph, to an arbitrary complex operation as a web service. Finally, we have classified also the required knowledge in the case of each related work.

As can be seen from Table 1, most of the approaches in the provenance literature [26,15,2,18,17,13] require a lot of knowledge, leveraging the network structure or execution environment. There are two exceptions: Deolalikar et al. [10] who reconstruct dependency chains of documents using a basic text similarity metric, and Nies et al. [22] who reconstruct sequences of a limited set of operations on Named Entities in documents using semantic similarity, i.e. the cosine similarity of vectors of Named Entities contained in each document. Both of these approaches offer a partial solution to the problem of reconstructing provenance, since they consider only one type of entities (text) and few operations.

More refined similarity measures are used in the context of text-reuse (e.g. [7,4]) in order to detect content reuse between documents, which can be seen as a type of dependency relationship. There are also approaches that use image similarity to reconstruct dependencies between documents, e.g. Hu et al. [6], who consider several electronic versions of historical manuscripts. There exists extensive research on reconstructing sequences of operations based on input and output data, but either the entities and operations involved are very simple [5,14,8] or they are tailored to a specific situation [23]. Other approaches require a lot of knowledge, either a formal description of the operations and user requirements on the composite operation (e.g. [3,24]), grammars of edit operations and a number of examples [25], or execution traces for several executions [1].

While there is extensive related work, this specific problem is only beginning to be addressed in the provenance community (see [10,22]), thus there are still a wide variety of open issues and improvements to be made.

3 Proposed approach

We propose a content-driven approach that reconstructs provenance using the contents of the files. Inspired by the DeepQA approach of IBM Watson [11], we aim at developing a multi-signal pipeline, which will combine several signals representing evidence on the relationships between files and propose a ranked list of plausible reconstructions. Our multi-signal pipeline consists of four stages, each containing several components that can be executed in parallel:

1. **Preprocessing phase**: contains the components that extract all available metadata from the files, infer the semantic types of the data, preprocess the content and index it in order to speed up the following phases.
2. **Hypotheses generation phase**: contains several *Signal Detectors*, which gather evidence of possible relationships between all the documents, generating several hypothesis graphs, that are expressed in the PROV-DM model

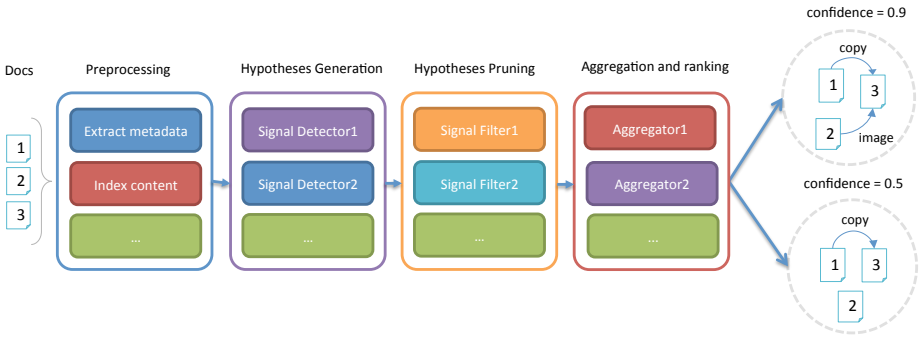

Fig. 1. Multi-signal pipeline for reconstructing provenance

[21]. Signal Detectors can be implemented using a number of existing techniques, from change detection algorithms to various types of similarity measures for different types of entities, e.g. text-reuse measures [4], image similarity [19] or semantic similarity [22].

If domain-specific knowledge is available, we can integrate it into one or more Signal Detectors. Moreover, if there is a library of possible operations, an AI planning technique similar to [16] can be employed, parametrized with the appropriate domain-specific heuristics.

The semantic type of data from the previous phase helps in deciding which Signal Detector to use. In order to speed up the evaluation, the cheapest Signal Detectors are executed first.

3. **Hypotheses pruning phase**: contains several *Signal Filters* that prune inconsistent or non-relevant hypotheses. One example is the Signal Filter that prunes temporally inconsistent edges in the hypotheses graphs or Signal Filters that enforce domain-specific rules triggered by the semantic type of the data. For example, if we are comparing two patient records, there could be a domain-specific rule that defines that two records can refer to the same patient only if they have the same identifier. For each hypotheses graph, the system executes all relevant Signal Filters in a cascade, but being independent one from the other, their order is not important. Therefore, we can devise a scheduling algorithm to parallelize their execution.
4. **Aggregate and ranking of hypotheses phase**: contains several *Aggregators* that aggregate the hypotheses, each with a confidence value that is based on the semantic type of data, e.g. a domain-specific aggregator has a greater confidence than a general one.

There are several challenges in this approach. The first major challenge involves finding the appropriate components for each of the phases in order to have results that are accurate enough for a broad range of domains and types of entities. We address this challenge by researching existing approaches in literature and integrating them in our pipeline. Moreover, we plan to integrate some simple domain-specific components (e.g. for dealing with bio-medical publications).

Another important challenge is computational efficiency, due to the large number of components, which are possibly already computationally expensive. We propose to address this issue by parallelizing the execution of components as much as possible, and schedule the cheapest components in each phase first. Moreover, all the components should feed their outputs, i.e. the hypotheses graphs, to the next phase as soon as they are ready. The Aggregator components, which need to aggregate several hypotheses graphs, should implement an anytime strategy that is able to give an approximation of the results based on its inputs, and gets more and more refined as there are more inputs.

A possible approach that we are considering involves using some of the computationally cheaper Signal Detectors as an approximation of the dependencies between entities, in order to suggest which pairs of entities are more promising to be compared.

4 Planned research methodology

To address the reconstructing provenance problem, we will follow an iterative process and we will incrementally build a framework for reconstructing provenance. In particular, we will use an empirical approach, in which each iteration will be guided by the results of the evaluation of the previous iteration. Each iteration of the process will consist of three phases.

The first phase will be focused on analyzing the state of the art approaches in literature, that could be compatible and complementary to the ones already present in our framework.

In the second phase, we will adapt and integrate one of these approaches into a framework, possibly reusing existing open-source software.

In the third phase, we will evaluate the performance of the system, both in terms of correctness of predictions and computational efficiency, on benchmark corpora. In case there are no corpora available, we will construct one, either automatically or manually, depending on the case. In the evaluation and testing phase, we will follow and adapt the standard IR and NLP approaches.

5 Preliminary results

The first approach for reconstructing provenance we devised was inspired by AI planning techniques and change detection algorithms, as described in [16]. The goal of this work was to reconstruct the sequence of transformations between entities by using the A* algorithm combined with a heuristic function based on the edit distance. In this case, there are three limitations :

- we need to define the library of possible operations;
- we need to define the heuristics;
- for each entity, we have to compute the edit distance - an expensive algorithm- for all entities, not only the more promising entities.

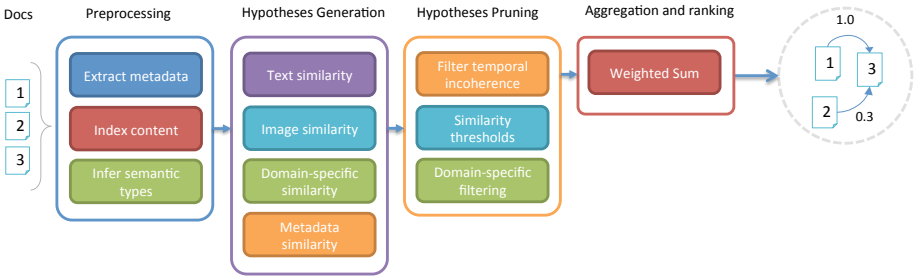

Fig. 2. Current system architecture

Therefore, we developed a complementary approach based on our multi-signal pipeline. As a first step, we considered the simpler problem of reconstructing provenance intended as dependencies between entities. There has been some prior work addressing reconstructing provenance as dependencies using text similarity (e.g., [10]), which we expanded by considering several multi-modal similarity measures. The setting we considered is reconstructing dependencies among a set of documents of different types (including images, Latex files, PDFs, MS Office documents, etc.) in a shared Dropbox folder.

We implemented a first prototype of the multi-signal pipeline by taking advantage of existing libraries and frameworks. Currently, all execution is sequential and we have not yet developed the anytime behavior.

In the Preprocessing stage, the system gets all available versions and metadata using the Dropbox API [1], extracts content (both text and images) and other metadata using Apache Tika[2]; and indexes the text using Apache Lucene[3] and images using LIRE [19].

We implemented four Signal Detectors: 1) text similarity using Lucene; 2) metadata similarity using SimMetrics[4]; 3) image similarity using LIRE [19]; 4) a simple domain-specific similarity, e.g. the exact match of the document name in the content.

We developed two Signal Filters: 1) filter dependencies using temporal information, e.g. a document in the past cannot depend on a document in the future; 2) filter dependencies with a score lower than a specified threshold;

The Aggregator we implemented is a weighted average of all the scores from the Signal Detectors and output a PROV-DM [21] graph using the Prov-toolbox[5].

We evaluated the prototype in a preliminary experiment on a Dropbox folder containing all data for a workshop paper, where the provenance of the files in the folder was manually annotated. With respect to our baseline, i.e. the approach described in [10], which uses only text similarity, our approach that combines

[1] https://www.dropbox.com/developers/reference/sdk
[2] http://tika.apache.org/
[3] http://lucene.apache.org/
[4] http://sourceforge.net/projects/simmetrics/
[5] https://github.com/lucmoreau/ProvToolbox

multi-modal similarities is able to increase the precision from 0.57 to 0.63 and the recall from 0.65 to 0.80, showing that even a simple approach can lead to significant improvements. More details on our experiment can be found in the Technical Report[6].

6 Conclusions and Future Work

In this paper, we have described the problem of reconstructing provenance, introducing a possible approach to address it using a multi-signal pipeline. The results we had obtained on the small test pilot are encouraging and we are currently creating a corpus for a more extensive evaluation.

The next step we will take is to implement the parallel and anytime behavior suggested in the proposed approach. Then we will extend the prototype with additional Signal Detectors, like text-reuse similarity measures (e.g. [4]), semantic similarity [22], normalized compression distance or other domain-specific Detectors as citation analysis techniques. We also plan to add domain-specific Signal Filters and to implement more Aggregators, possibly by using a supervised learning to assign the weights to the different Signal Detector scores. Moreover, we will integrate and adapt the approach presented in [16].

Acknowledgements This publication was supported by the Data2Semantics project in the Dutch national program COMMIT.

References

1. van der Aalst, W., van Dongen, B.F.and Herbst, J., Maruster, L., Schimm, G., Weijters, A.: Workflow mining: A survey of issues and approaches. Data & Knowledge Engineering 47(2), 237–267 (Nov 2003)
2. Barbier, G., Liu, H.: Information provenance in social media. Social Computing, Behavioral-Cultural Modeling and Prediction pp. 276–283 (2011)
3. Baryannis, G., Plexousakis, D.: Automated Web Service Composition: State of the Art and Research Challenges. Tech. Rep. October, Tech. Rep. 409, ICS-FORTH (October 2010) (2010)
4. Bendersky, M., Croft, W.B.: Finding text reuse on the web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining (2009)
5. Bille, P.: A survey on tree edit distance and related problems. Theoretical Computer Science 337(1-3), 217–239 (Jun 2005)
6. Broder, A.Z.: On the resemblance and containment of documents. In: In Compression and Complexity of Sequences (SEQUENCES'97 (1997)
7. Chawathe, S., Garcia-Molina, H.: Meaningful change detection in structured data. In: ACM SIGMOD Record. pp. 26–37 (1997)
8. Cheney, J., Chiticariu, L., Tan, W.C.: Provenance in databases: Why, how, and where. Found. Trends databases 1, 379–474 (April 2009)
9. Deolalikar, V., Laffitte, H.: Provenance as data mining: combining file system metadata with content analysis. In: First workshop on on Theory and practice of provenance. p. 10. USENIX Association (2009)

[6] http://www.few.vu.nl/ sme340/papers/techreport.pdf

10. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An overview of the DeepQA project. AI Magazine 31(3), 59–79 (2010)
11. Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for computational tasks: A survey. Computing in Science and Engg. 10, 11–21 (May 2008)
12. Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. Concurrency and Computation: Practice and Experience 20(5), 485–496 (2008)
13. Gao, X., Xiao, B., Tao, D., Li, X.: A survey of graph edit distance. Pattern Analysis and Applications 13(1), 113–129 (Jan 2009)
14. Govindan, K., Wang, X., Khan, M., Dogan, G., Zeng, K., Davis, C.: PRONET : Network Trust Assessment Based on Incomplete Provenance. IEEE The Premier International Conference for Military Communications (2011)
15. Groth, P., Gil, Y., Magliacane, S.: Automatic Metadata Annotation through Reconstructing Provenance. ESWC (2012)
16. Holland, D.A., Seltzer, M.I., Braun, U., Muniswamy-Reddy, K.K.: Passing the provenance challenge. Concurrency and Computation: Practice and Experience 20(5), 531–540 (2008)
17. Hu, B., Rakthanmanon, T., Campana, B., Mueen, A., Keogh, E.: Image mining of historical manuscripts to establish provenance. In: SIAM Conference on Data Mining (SDM) (2012)
18. Huq, M., Wombacher, A.: Inferring fine-grained data provenance in stream data processing: reduced storage cost, high accuracy. Database and Expert Systems pp. 118–127 (2011)
19. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: Proceedings of the 16th ACM international conference on Multimedia. pp. 1085–1088 (2008)
20. Moreau, L.: The foundations for provenance on the web. Found. Trends Web Sci. 2, 99–241 (February 2010)
21. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model, http://www.w3.org/TR/prov-dm/
22. Nies, T.D., Coppens, S., Deursen, D.V., Mannens, E., Walle, R.V.D.: Automatic Discovery of High-Level Provenance using Semantic Similarity. In: IPAW 2012
23. Noy, N., Kunnatur, S., Klein, M., Musen, M.: Tracking changes during ontology evolution. The Semantic Web–ISWC 2004 pp. 259–273 (2004)
24. Rao, J., Su, X.: A survey of automated web service composition methods. Semantic Web Services and Web Process Composition pp. 43–54 (2005)
25. Wu, B., Szekely, P., Knoblock, C.A.: Learning data transformation rules through examples: Preliminary results. In: Ninth International Workshop on Information Integration on the Web (IIWeb 2012) (2012)
26. Zhao, J., Gomadam, K., Prasanna, V.: Predicting Missing Provenance using Semantic Associations in Reservoir Engineering. In: Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. pp. 141–148. IEEE (2011)