Towards Reconstructing the Provenance of Clinical Guidelines

Sara Magliacane and Paul Groth

s.magliacane@vu.nl, p.t.groth@vu.nl
Department of Computer Science
VU University Amsterdam

Abstract. Understanding the provenance of clinical guidelines is important for both practitioners and researchers as it allows for deeper understanding of the provided recommendations and could potentially provide a basis for updating guidelines. Often such provenance is incomplete or unavailable. We describe a prototype of a multi-signal pipeline for reconstructing provenance and show preliminary results of reconstructing dependencies between documents in the context of clinical guidelines and associated documents.

1 Prototype description

Broadly, we target the problem of reconstructing provenance of files in a shared folder setting, in which several authors can create or edit files at different moments, and only standard filesystem metadata is available. In a previous work [3] we proposed a content-based approach that is able to reconstruct provenance automatically, leveraging several similarity measures and edit distance algorithms, which are then adapted and integrated them into a multi-signal pipeline.

Here, we present an improved version of this prototype applied to a clinical guideline and associated biomedical documents. The architecture of our prototype is shown in Fig. 1.

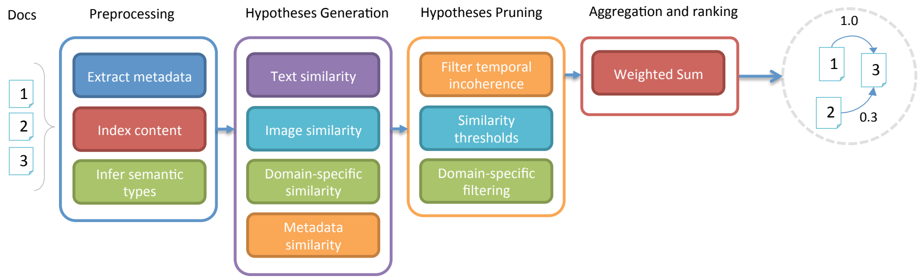

Fig. 1. System architecture

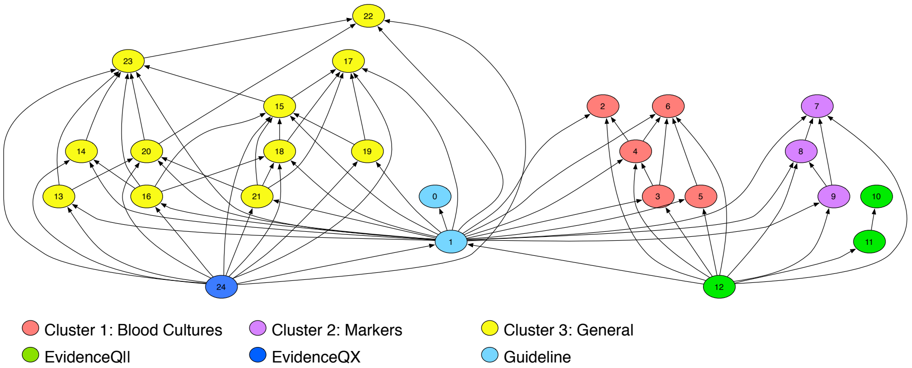

Fig. 2. Manual annotated dependencies between documents

The prototype combines several multi-modal similarity measures, in particular text, image and metadata similarity, and aggregates them into a single similarity score. The prototype performs the following tasks:

- Gather all available versions and metadata of the files (e.g. authors, revisions, timestamps) using the Dropbox Java API [1]
- Extract content (both text and images) and metadata using Apache Tika[2].
- Index the content of the files using Apache Lucene[3] and LIRE [2].
- Create a graph, in which the nodes represent the files and the edges represent the relationships between the files, using different text, metadata, and image similarity metrics.
- Prune similarity edges using temporal constrains known from the provenance literature [1], e.g. pruning the edges that indicate that a file depends from another file that was created later in the timeline.
- Aggregate the similarity measures for each couple of files into a single score.
- Output a PROV [4] graph using the Prov-toolbox[4]. PROV is the forthcoming recommendation from the W3C on representing provenance.

2 Experimental setting

The experimental setting consisted of a Dropbox folder containing the clinical guideline for febrile neutropenia, a set of publications referred to by the guideline and two Excel sheets that describe the references from the guideline for two research questions. The provenance of the files in the folder was manually annotated in PROV-DM, as shown in Fig. 2, in which each node represents a file and each edge a dependency of the origin file from the destination file.

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

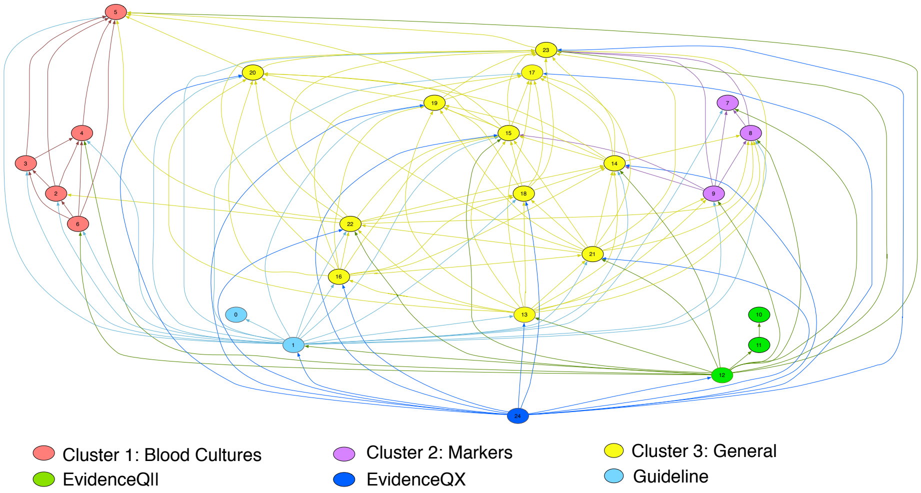

Fig. 3. An example of the predicted dependencies between documents

The main document is the guideline, which has two versions in the Dropbox folder (light blue nodes in Fig. 2). All the publications are cited by the guideline and there are two Excel sheets that contain copy-paste text from the guideline. Each of these sheets details the references needed to answer a research question. In particular EvidenceQII (green nodes in Fig. 2) details the references for Question A, while EvidenceQX (blue node) focuses on Question B.

In the manual annotation, we considered citations as dependencies. Then we divided the publications in 3 clusters based on the citation network: 1) publications on blood cultures (red nodes); 2) publications on markers (purple nodes); 3) other publications (yellow nodes).

3 Results and evaluation

We ran our prototype in the previously described experimental setting with different sets of similarity measures and obtained several predictions of dependency graphs. One example can be seen in Fig. 3, in which we show the predicted dependency graph using all the implemented similarity metrics.

In order to evaluate the results we obtained, we compared the edges of the original dependency graph and each dependency graph, predicted with our method. The results are shown in Table 1, where the rows represent the evaluation using different similarity measures. The first row represents our baseline, i.e. the approach described in [1].

We compared the different systems to see if there was a statistically significant difference in the results. Using the T-test provided in the R statistical package, we obtained a small difference between the baseline and system1 (p-value is 0.4216), while system2 and system3 are very different from the baseline (both have p-value 2.388e-06).

Similarity measures	Precision	Recall	F1-score
baseline: text	0.638	0.403	0.494
system1: text, metadata	0.621	0.415	0.498
system2: text, metadata, inverse lucene	0.696	0.717	0.706
system3: text, metadata, inverse lucene, images	0.692	0.717	0.704

Table 1. Comparison of results using different similarity measures

As we can see from Table 1, much of the structure of the original dependency graph is well-predicted. The Excel sheets depend on the guideline and the guideline is connected to all of the publications. The clusters of citations are quite recognizable.

Among the errors, some can be easily explained. For example, some papers are connected even when there is no citation, e.g. the newer clinical guideline is connected to its older version, but does not cite it. The two Excel sheets are connected because they have the same author and creation data (metadata similarity). There are some difficulties in finding the appropriate temporal order, since some documents have a very different creation and publication date. Due to the temporal pruning that we perform, this means that several dependencies were discarded because of temporal inconsistency.

4 Future Work & Conclusion

The issues with temporal ordering can be partially solved by retrieving bibliographic information on the publications. There are also other domain-specific improvements than can be made, e.g. using the knowledge of citations. Moreover, there is the need for a better aggregation algorithm. Up to now we targeted high recall, in our future work we aim at refining the predictions in terms of precision. Finally, we want to apply the technique on a much larger corpus of the biomedical papers and guidelines. Overall, we have shown that multimodal similarity combined with knowledge of the structure of provenance graphs is a good start towards reconstructing the provenance of clinical guidelines.

Acknowledgements This work was funded by Data2Semantics project in the Dutch national program COMMIT.

References

1. 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)
2. 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)
3. Magliacane, S.: Reconstructing provenance. In: Procs. ISWC 2012 (2012)
4. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model, http://www.w3.org/TR/prov-dm/