Automatic Discovery of High-Level Provenance Using Semantic Similarity

Tom De Nies, Sam Coppens, Davy Van Deursen, Erik Mannens, and Rik Van de Walle

In: Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science, 7525, 97-110, 2012.

http://link.springer.com/chapter/10.1007%2F978-3-642-34222-6_8

Automatic Discovery of High-Level Provenance using Semantic Similarity

Tom De Nies, Sam Coppens, Davy Van Deursen,
Erik Mannens, and Rik Van de Walle

Ghent University - IBBT
Department of Electronics and Information Systems, Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium
{tom.denies,sam.coppens,davy.vandeursen,
erik.mannens,rik.vandewalle}@ugent.be

Abstract. As interest in provenance grows among the Semantic Web community, it is recognized as a useful tool across many domains. However, existing automatic provenance collection techniques are not universally applicable. Most existing methods either rely on (low-level) observed provenance, or require that the user discloses formal workflows. In this paper, we propose a new approach for automatic discovery of provenance, at multiple levels of granularity. To accomplish this, we detect entity derivations, relying on clustering algorithms, linked data and semantic similarity. The resulting derivations are structured in compliance with the Provenance Data Model (PROV-DM). While the proposed approach is purposely kept general, allowing adaptation in many use cases, we provide an implementation for one of these use cases, namely discovering the sources of news articles. With this implementation, we were able to detect 73% of the original sources of 410 news stories, at 68% precision. Lastly, we discuss possible improvements and future work.

Keywords: Provenance, Data Model, Semantic Web, Linked Data, Similarity, News

1 Introduction

Nowadays, as interest in provenance grows among the Semantic Web community [1], media content authors are faced with a dilemma. While they clearly see the advantages of providing provenance information with their data, the process of manual annotation is labor intensive and dull work, especially for those without a technical background [2]. Clearly, there is a need for automated ways to add provenance to produced content.

Most existing automatic provenance collection techniques in literature either observe all activity on the target resources (so called *observed* provenance), or require that the users specify formal workflows which are used to create and modify the resources (*disclosed* provenance) [3]. The first approach often results

in a low-level view of the provenance associated with a resource, which is not always suitable (e.g., in the use case described in this paper). The latter approach requires significant effort from the user, and is not always applicable, since many creative processes are difficult, if not impossible, to formally describe.

In this paper, we propose a new approach for automatic discovery of provenance from limited information, at multiple levels of granularity. Whereas low-level provenance denotes the exact change at the finest granularity (e.g., at the character level), higher-level provenance denotes changes at a coarser granularity (e.g., at the document level). To achieve this, we detect inter-document derivations, using clustering methods based on semantic similarity, resulting in provenance complementary to the observed and disclosed kind. We apply the approach to a specific use case, originated from the news sector. We will attempt to reconstruct missing provenance, solely based on the content and timing information, allowing us to track down the original source of an article.

The paper is structured as follows: first, we explain our interpretation of high-level provenance, and how this fits into the ongoing standardization efforts of the W3C Provenance Working Group[1]. Next, we provide an in-depth explanation of the proposed approach and describe our use case implementation, which we then use to evaluate our approach. Before concluding, we discuss the results, followed by the related and future work.

2 Terminology & Key Concepts

Before describing our proposed approach, we explain our view on high-level provenance. We also provide a summary of the relevant features of the Provenance Data Model (PROV-DM), currently under development by the W3C Provenance Working Group.

2.1 High-Level Provenance

In our research, we make the distinction between low-level and high-level provenance. Low-level provenance is the sort of provenance expected from capturing systems and versioning systems. A typical example is that of a programmer's versioning system, where the provenance of each document is stored as a list of characters that where changed, together with their position in the document. An example of high-level provenance, at the *document level* might be: "Document A is a revision of document B".

While these types of provenance are certainly important in many cases, for our research, we aim for a more conceptualized form of provenance, and propose an intermediary approach. For example: "Document A is a derivation of document B, with concept 'Magistrate' in document A narrowed down to 'Prosecutor' in document B". We will label this as provenance at the *semantic level*, providing more details than at the document level, but remaining high-level, at a coarser granularity than low-level systems.

[1] http://www.w3.org/2011/prov/

In this paper, we will investigate ways to generate high-level provenance, both at the document level and the semantic level.

2.2 PROV-DM: The Provenance Data Model

Currently, the W3C Provenance Working Group is composing a standard data model for provenance. In our research, we aim to comply with the latest working draft of PROV-DM, at the time of writing (WD6, [4]). For a full description of the data model, we refer to [4]. Below, we provide a brief overview of the concepts needed for our research.

PROV-DM provides us with 3 essential (core) elements: *entities*, *activities* and *agents*. Entities can be related to each other, and to activities acting upon them. For our research, the most important entity-entity relations are **derivation**, **alternate** and **specialization**. Entity-activity relations are limited to usage and generation. Throughout this paper, in all figures and examples, the standard notation specified in [4] is used to specify these relations.

According to PROV-DM, a **derivation** is anything that transforms an entity into another, that constructs an entity from another, or that updates an entity, resulting in a new one. However, the underpinning activities and their associated details are not always known. Therefore, we will make the distinction between *precise* and *imprecise* derivations. When two entities are linked by a **precise-1** derivation, it means they are connected by a single, known activity, which uses (consumes) one of the entities and generates the other. When the activity connecting two entities is unknown, but it is certain that they are connected by a single activity, we obtain an **imprecise-1** derivation. For an **imprecise-n** derivation, the number of activities interconnecting the two entities is unknown. Note that while the formal distinction between imprecise and precise provenance was removed from PROV-DM since the fifth working draft, the informal distinction is still relevant to the work in this paper, and remains supported by PROV-DM (all parameters of derivation regarding the involved activity are optional).

The **alternate** relation connects two entities that refer to the same thing in the world, in different environments. For example, 'fbase:Magistrate'[2] is an alternate entity of 'dbpedia:Magistrate'[3]. The **specialization** relation connects two entities that refer to the same thing in the world, at different levels of abstraction. For example, 'dbpedia:Prosecutor'[4] is a specialization of 'dbpedia:Magistrate'.

In additional to these relations, PROV-DM allows to provide provenance of provenance. Concretely, this means that all provenance entities, activities, agents and relations can be organized in *bundles*. A bundle holds the provenance of a resource, and can have, in turn, its own provenance. This way, it becomes possible to provide provenance of the provenance, explaining how it was obtained.

[2] http://rdf.freebase.com/ns/Magistrate
[3] http://dbpedia.org/resource/Magistrate
[4] http://dbpedia.org/resource/Prosecutor

A final method that we use to provide organization among entities, is the *collection* entity. According to PROV-DM, this is an entity that provides a structure to some constituents, which are themselves entities, and connected to the collection by the **memberOf** relation.

3 Proposed Approach

In this section, we provide an in-depth description of how we aim to discover provenance derivations, using semantic similarity. While we want to keep our approach as general as possible, it is necessary to make some assumptions about the data we will be providing provenance for.

We will assume that the data essentially consists of two types of entities. We define a *document* as an entity that is characterized by multiple other entities, which we will refer to as *semantic properties*. Both documents and semantic properties can be modeled as a prov:Entity[5], and thus can be connected through activities and/or entity-entity relations. In our news use case, an example of a document would be a news article, whereas examples of semantic properties would be the descriptive metadata annotations of this article. We also assume that timing information (i.e., date of creation) is available for all documents.

The general goal of our research is to analyze documents to automatically discover provenance information about them. Since this is very general, we will narrow it down to 3 subgoals. Starting from a set of documents S, we aim to:

1. Discover high-level **imprecise-n** and **imprecise-1** derivations at a *coarse granularity*.
2. Convert these imprecise derivations to high-level **precise-1** derivations.
3. Discover additional **precise-1** derivations at a *finer granularity*.

Below, we describe how we achieve these goals.

3.1 Discovering Imprecise Derivations

To discover provenance at the coarsest granularity, we rely on the semantic similarity of documents. Since it is safe to assume that revisions of the same document are semantically similar to each other, we can assume that in many cases (unfortunately, not always), the inverse also holds: if documents are very similar to each other, it is likely that they are also a revision of the same document.

First, we group (or *cluster*) all semantically similar documents into clusters S_i, so that for all documents $doc_a \in S_i$:

$$doc_a \in S_i \Leftrightarrow \forall doc_b \in S_i : sim_D(doc_a, doc_b) > T_s \tag{1}$$

with T_s an empirically determined *similarity threshold*, and $i \in \{1, 2, ..., N\}$ with N the number of clusters[6]. Sim_D is a similarity metric, which enables semantic

[5] http://www.w3.org/ns/prov-dm/Entity

[6] Note that overlap between clusters is possible.

comparison of documents. Note that this similarity metric is interchangeable, and a more accurate similarity metric will result in better clustering (in our implementation, semantic similarity of documents is based on the comparison of their semantic properties). To avoid clusters becoming too large, resulting in poor derivations, all clusters larger than a *clustering threshold* T_c, are re-clustered with a higher similarity threshold T_s.

Next, we order all documents in each cluster according to their date of creation. For each cluster, we assume that the document doc_1 that was created first is the original source of all other documents in the cluster. This means that we can now connect each document of the cluster to doc_1 by an **imprecise-n** derivation, as illustrated by Fig. 1(a).

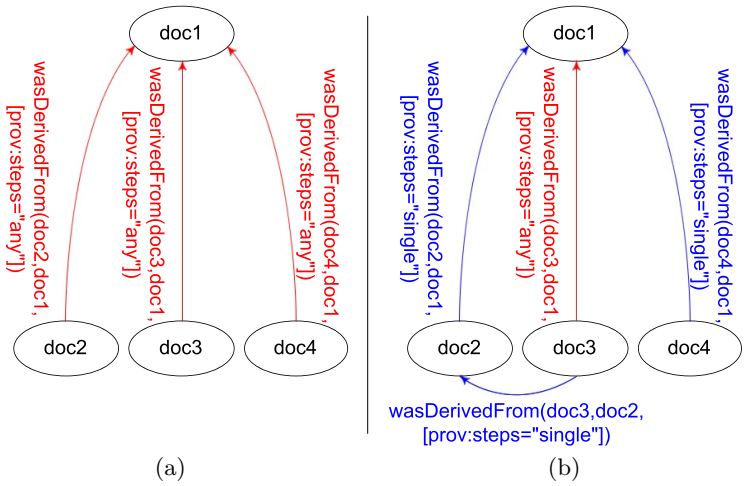

Fig. 1. Example of how documents doc_2, doc_3 and doc_4 within one cluster are related (a) to the original source doc_1 by imprecise-n derivations, and (b) to each other by imprecise-1 derivations. We assume that $time(doc_i) < time(doc_j) \Leftrightarrow i < j$. Here, doc_2 is most similar to doc_1, doc_3 most similar to doc_2 and doc_4 most similar to doc_1. Even though doc_4 was created after doc_3, it was directly derived from doc_1.

In order to create **imprecise-1** derivations, we take both the inter-document similarity and timing information into account[7]. In each set S_i, for each document $doc_a \in S_i$ (with $a \neq 1$), we find the semantically most similar document doc_b, and connect them by an imprecise-1 derivation, following Formula 2.

[7] Note that simply considering the timing and connecting successive documents with imprecise-1 derivations is not a correct approach, since multiple revisions can be based on a single document, regardless of timing.

$$\begin{aligned} \exists doc_b \in S_i : (\forall k \neq a : sim_D(doc_a, doc_b) \geq sim_D(doc_a, doc_k)) \\ \wedge \\ time(doc_b) < time(doc_a) \\ \Rightarrow wasDerivedFrom(doc_a, doc_b, [prov : steps = \text{“}single\text{”}]) \end{aligned} \quad (2)$$

The direction of this derivation depends on which document was created first. In Fig. 1(b), we apply this method to the example from Fig. 1(a).

3.2 High-Level Precise Derivations

Precise-1 derivations need to specify an *activity*, responsible for using the original entity, and generating the derived entity. Converting the imprecise-1 derivations from Sect. 3.1 to precise-1 is done by defining a *revision* activity for each imprecise-1 derivation, as illustrated by Fig. 2.

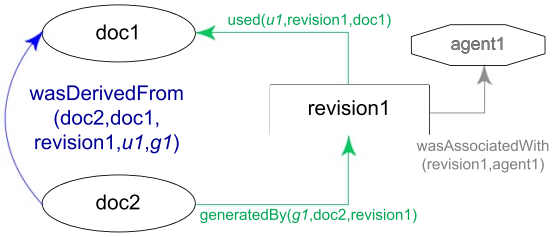

Fig. 2. The imprecise-1 derivation of doc_2 from doc_1 is converted into a precise-1 derivation by specifying an activity $revision_1$, which uses doc_1 and generates doc_2, and is associated with an agent $agent_1$

Specifying this activity enables us to vary the granularity of the obtained provenance (see Sect. 3.3), and to model *responsibility* for the revision, by specifying an agent, if available. In the best case scenario, this agent is found in the document's metadata, as the annotated author or editor. In the worst case, when no agent can be found, the provenance of the revision can still be asserted, without an agent. In other cases it might be possible to find the correct agent by querying other data sources and finding a matching document, with author information available. However, for this paper, reconstructing this missing author information would lead us too far.

3.3 Precise Derivations at Finer Granularity

To obtain provenance at a finer granularity, we will use the semantic properties characterizing the documents. As a document is revised, some of its semantic

properties will change, and others will remain the same. Changes might imply *replacements*, *generalizations* or *specializations*. Some properties might be *omitted* from the document, whereas new ones may be *added*. All of these changes can be modeled with the PROV-DM model. We start from the coarse-grained provenance bundle associated with a set of related documents, as generated in the previous steps, and create a new, fine-grained bundle, enclosing it.

How the semantic properties of a document are identified is dependent on the type of data, and may vary for each use case. In our use case (as can be seen in Sect. 4), this is achieved by applying a named entity extraction technique to the documents. Once the properties are identified, we define a *usage* activity for each of them, linking the properties to the document they are used by.

Next, the properties of each document pair related by a precise-1 derivation are semantically compared. Once again, this comparison is dependent of the type of data and use case. However, it is important that the comparison can model **replacements**, **generalizations** and **specializations**. Additionally, we will model **additions** and **omissions**.

In PROV-DM, **replacements** or synonyms are modeled by the *alternateOf* relation. The replaced property p_i is *used* by the revision activity, which *generates* the new property p_j. **Specialization** is modeled by the PROV-DM *specializationOf* relation. The more general property p_i is *used* by the revision activity, which *generates* the specialized property p_j. **Generalization** is modeled as an inverse specialization. **Addition** is modeled by a revision activity that *generates* a property p_i, but does not use a replaced, specialized or generalized property. Similarly, **omission** is modeled by a revision activity that *uses* a property p_i, but does not generate a replacing, specializing or generalizing property.

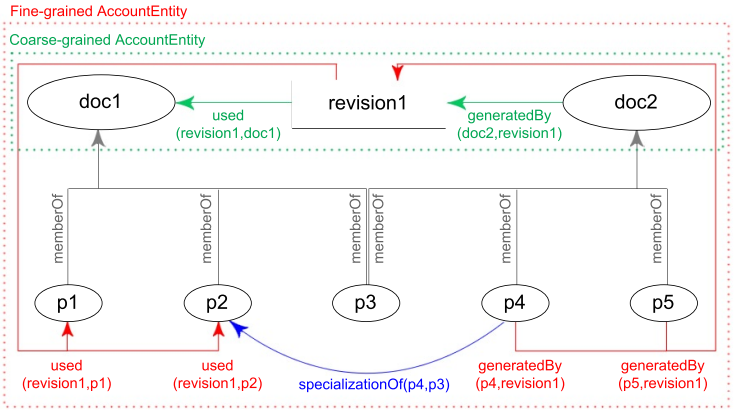

Fig. 3. Finer-Grained Precise Derivations (some usage and generation arguments omitted for clarity)

As an example, we consider the coarse-grained bundle associated with two documents doc_1 and doc_2, as illustrated by Fig. 2. Suppose we were able to identify three properties p_1, p_2, p_3 of doc_1 and three properties p_3, p_4, p_5 of doc_2. Figure 3 shows the usage activities linking these properties with doc_1 and doc_2. When comparing the properties, it was discovered that p_4 is a specialized concept of p_2. This is modeled by the usage of p_2 and generation of p_4 by $revision_1$, and the specialization relation between p_4 and p_2. p_1 was omitted from the revised document, which is modeled by the usage of p_1 by $revision_1$ and the lack of a generation of a related property. p_5 was added to the revised document, which is modeled by the generation of p_5 by $revision_1$ and the lack of a usage of a related property. Storing these assertions into a new, fine-grained bundle, encompassing the original, coarse-grained bundle, provides us with a multi-level view of the provenance of doc_1 and doc_2

4 Use Case: News Versioning

We kept our description of the proposed approach as general as possible, since it is applicable in many use cases. However, for clarification and evaluation purposes, we will describe a particular use case, originating from the news sector. In today's news industry, specification and justification of sources are key factors for producing high quality journalism. Unfortunately, due to the strong time constraints inherent to news production, provenance information is often incomplete or omitted. The consumers' need for near-immediate reporting also results in an abundance of very similar publications by all leading news organizations, often slightly modified versions of the same article, with limited to no possibility to determine the original source, or to determine which modifications were made to the content. This is exactly where our approach fills the gap. By detecting the derivation of one revision into another, our approach makes it possible to find the original source of an article, as well as the intermediary revisions. In this section, we describe how our approach is implemented for this use case.

4.1 Documents & Properties

For the implementation of our approach, we need to identify "documents" and "properties", as described in Sect. 3. As documents, we use *news stories*, provided in different *revisions*. A news story starts as an *alert*, which is then expanded into a *short story*, a *brief article*, and finally a *full article* (in some cases one or more of these stages are skipped). The articles are available in several languages, so multiple brief articles can be derived from one short story, etc.

As semantic properties, we use *Named Entities* (NEs) associated with the news stories. These can be manually added, or automatically extracted from the content. In either case, the NEs are enriched, linking them to unique resources in the Linked Open Data (LOD) Cloud[8]. For the implementation of our approach,

[8] http://linkeddata.org

the named entities are also modeled as entities in PROV-DM, with each news article linked to the entities corresponding to the metadata by a *usage* activity.

4.2 Extracting Properties through Named Entity Recognition

When news articles are not annotated with sufficient descriptive metadata, as is often the case in real-world scenarios, we need to automatically generate this metadata ourselves. The availability of accurate metadata associated with the documents will be beneficial to the resulting provenance.

To achieve this, we use publicly available Named Entity Recognition(NER) services. These services accept regular text as input, and output a list of linked NEs, detected in the text. The NERD [5] comparison tools allow us to evaluate the services and select the most fitting one for our work. For our implementation, we choose to use OpenCalais[9], a well-established, thoroughly tested [6] and freely available NER service. Note that as OpenCalais does not support Dutch, nor French at the time of writing, an automatic translation step is performed before sending the data, using the Microsoft Bing API[10].

4.3 Similarity Measure

Traditionally, document similarity is calculated using the Vector Space Model (VSM), also known as the "bag of words" model. When using this method, documents are viewed as vectors of *Term Frequency - Inverse Document Frequency (TF-IDF)* weights, signifying the importance of each term in the document. We adapt this approach to work with Named Entities (NEs) instead of words. This will allow two documents containing similar concepts, but of significantly varying length, to receive a high similarity score, whereas the classic TF-IDF approach would yield a lower score, due to the difference in text length.

The similarity measure is calculated as follows. When comparing two documents A and B, we create two vectors representations a and b of their NEs, where a_i is the weight of NE i in document A (analogous for B), as determined during the NER step. The similarity between the documents is then calculated as the *cosine similarity* of the vectors, given by Formula 3.

$$Sim_{VSM}(A, B) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}} \tag{3}$$

When no NEs were detected, we revert to the classic "bag of words" approach, using TF-IDF weights for every word in the text. Note that the semantic awareness of this similarity metric can be improved (see discussion, Sect. 7).

[9] http://www.opencalais.com
[10] http://www.microsofttranslator.com/dev/

4.4 Coarse-Grained Provenance through Clustering

As described in Sect. 3.1 and Sect. 3.2, we obtain the first, coarse-grained provenance by clustering sufficiently similar documents together. Using the similarity-measure in Sect. 4.3, we cluster the total set of news articles into sets of closely related articles. As shown in [7], clustering with a lower bound on similarity is an NP-Hard optimization problem. Fortunately, the authors of [7] also provide a greedy heuristic, SimClus, which we choose to use to cluster our dataset.

The applied algorithm is summarized as follows. The set of possible cluster centers S_{pc} initially contains all elements (with at least three NE's, to ensure accuracy of the similarity measure) of S. We compute the complete similarity matrix of the dataset S, which is then used to determine a *cover-set* S_u for each item $u \in S$. S_u contains all elements of S *covered* by u, which means their similarity to u is above an empirically determined threshold T_s. We now choose the cluster centers as follows:

1. Choose the item $u \in S_{pc}$ with the largest cover-set S_u as the next cluster center (if multiple items are tied, choose the one with the most properties; if there is still a tie, choose arbitrarily).
2. Remove all elements of S_u from S_{pc}.
3. Repeat step 1.

The algorithm terminates when there are no items left to choose as cluster center. The dataset is now divided into (possibly overlapping) clusters, corresponding to the cover-sets of each cluster center. As an optimization, clusters with more items than a predetermined upper bound T_c are clustered again with a higher similarity threshold T_s. In our implementation, we choose $T_c = 10$, since news items rarely have more than ten revisions. For each cluster, we now add the imprecise-n and imprecise-1 derivations according to the method described in Sect. 3.1. Next, we construct the activities as in Sect. 3.2, resulting in precise-1 derivations.

4.5 Finer-Grained Provenance

Starting from the coarse-grained provenance bundle from Sect. 4.4, we can create a finer-grained bundle in the manner described in Sect. 3.3. Note that the semantic properties are already identified in the NER step (see Sect. 4.2). Since these properties are linked to the LOD Cloud, information regarding synonyms, specializations and generalizations is available by following (or dereferencing) these links to popular datasets such as DBPedia, WordNet, Freebase, etc. Synonym relationships include *owl:sameAs* and *skos:exactMatch*, whereas examples of links specifying generalization and specialization are (respectively) *skos:broader* and *skos:narrower*. Using the methods in Sect. 3.3, we create the correct derivations, usages and generations linked to the revision activities from the coarse-grained provenance, and create a new, finer-grained provenance bundle, encompassing the original. In Fig. 4, this is illustrated for one news item.

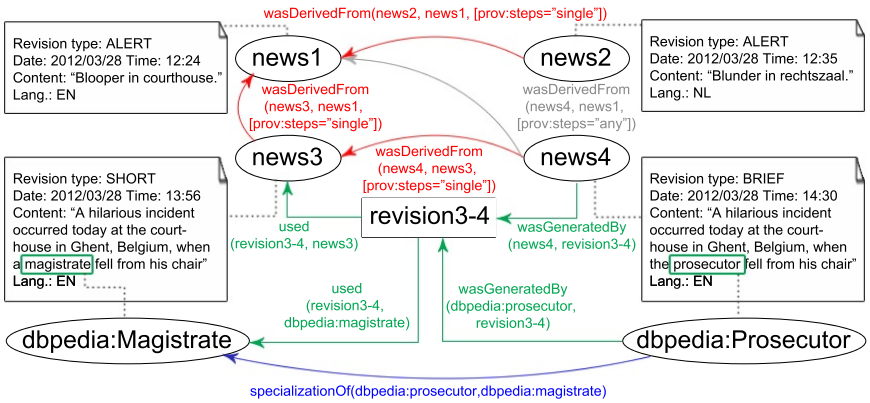

Fig. 4. Example of discovered provenance in the news use case. The news item starts as an English alert $news_1$, which is then translated into a Dutch alert $news_2$. Soon after that, a short story $news_3$ is written based on the English alert. Finally, the short story is revised to a brief story $news_4$, replacing the word "magistrate" with "prosecutor".

5 Evaluation

Our evaluation data consists of a set of 410 news stories, corresponding to 100 news items, in up to two different languages (Dutch and French), acquired from Belga[11], a professional Belgian news agency, over the course of one week.

The originally available provenance for the news stories, as specified by the content provider, is limited to the *revision types*, *original sources* and *imprecise-n derivations*. The source of a news item is always the earliest news story associated with that news item (usually an alert or short story). All following stories about that news item are (directly or indirectly) derived from its source (as an imprecise-n derivation).

Since there is no formal workflow to describe the creative process of news production, indisputably correct imprecise-1 derivations are nearly impossible to determine, even for the content providers (which is why our approach is so useful to them). Therefore, we restrict the evaluation to imprecise-n derivations.

We constructed coarse-grained provenance using the approach described in Sect. 4, based only on the (enriched) content and timing information of the news stories in our dataset. We can now compare the detected clusters, sources and imprecise-n derivations to the original information provided by the news agency. In Table 1, the results are shown for different initial similarity thresholds T_s. In the optimal case, with $T_s = 0.5$, we were able to detect 73% of the original news sources, with 68.2% precision. The imprecise-n derivations constructed from these sources have a precision of 72.3% and a recall of 44.5%.

An explanation for these figures is found when examining the clustered news stories. In Table 2, it is shown that for nearly all clusters (96% with $T_s = 0.5$), the

[11] http://www.belga.be

Table 1. Accuracy of provenance discovery with similarity threshold $T_s \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and cluster threshold $T_c = 10$. p_{source} and r_{source} represent the precision and recall of the detected news sources, compared to original derivations from the dataset.

	$T_s = 0.2$	$T_s = 0.3$	$T_s = 0.4$	$T_s = 0.5$	$T_s = 0.6$	$T_s = 0.7$	$T_s = 0.8$
p_{source}	68.0%	67.3%	**69.2**%	68.2%	64.3%	59.3%	57.7%
r_{source}	70.0%	68.0%	72.0%	**73.0**%	72.0%	70.0%	71.0%
$p_{imprecise-n}$	56.3%	61.6%	67.6%	**72.3**%	71.5%	57.1%	57.9%
$r_{imprecise-n}$	45.8%	**48.1**%	45.2%	44.5%	41.3%	28.4%	26.1%

news stories in the cluster all belong to the same original news item. However, $r_{newsitem}$ shows that many of the original news items are spread across more than one cluster, which creates more than one cluster per news item, resulting in lower overall accuracy of the detected provenance.

Table 2. Percentage $p_{cluster}$ of clusters of which all news stories originally belong to the same news item and percentage $r_{newsitem}$ of original news items that were cataloged into a single cluster.

	$T_s = 0.2$	$T_s = 0.3$	$T_s = 0.4$	$T_s = 0.5$	$T_s = 0.6$	$T_s = 0.7$	$T_s = 0.8$
$p_{cluster}$	83.8%	86.0%	93.3%	96.0%	97.7%	98.0%	100%
$r_{newsitem}$	30.0%	37.0%	32.0%	31.0%	26.0%	11.0%	8.0%

The accuracy of the fine-grained provenance depends strongly on the correctness of the detected named entities, and the quality of their links to ontologies that describe alternates, specializations and generalizations. When processing the 410 news stories, OpenCalais extracted 722 distinct named entities. Upon manual evaluation of these NEs we labeled 20 of them as incorrectly detected, resulting in 97.2% precision. Criteria for labeling a property as incorrectly detected were *non-existence* (no such concept exists) and *incorrect disambiguation* (linked to the wrong resource). These results are consistent with those of a larger performance analysis of OpenCalais, described in [6]. Of the 722 Named Entities, 47 were automatically linked to a resource in the LOD Cloud by OpenCalais.

6 Related Work

When it comes to automated production of provenance information, several methods exist. These techniques mostly focus on either *observed provenance*, or *disclosed provenance*[8]. In [3], it is noted that these systems need to capture all activities, since they do not necessarily understand the semantics of their observations. Although domain-specific techniques used to reconstruct lost or missing provenance information do exist, such as in [9] and [10], no generic solution is available to date.

As shown in [2], provenance produced by these methods is often low-level and/or too complex for a domain expert (e.g., a journalist) with limited knowledge of computer science. In [11], the need for high level provenance is motivated, and a conceptualization is proposed, namely as a combination of interconnected elements including *"what", "when", "where", "how", "who", "which", and "why"*. However, a recent survey, [12], shows that high-level knowledge provenance is still a sparsely researched topic.

7 Discussion & Future Work

The results of our evaluation clearly show that our approach to discover provenance of resources, solely based on their (enriched) content and timing information, is feasible and provides the foundations for future work. A better, more semantically aware similarity measure, such as the one described in [13], is likely to have a significant impact on the overall accuracy. To accommodate such a metric, the extracted semantic properties need to be accurately linked to the Semantic Web. To achieve this in future implementations, additional disambiguation and enrichment techniques are being developed to combine with the available NER services. Finally, even though it would make the approach less general, it might prove worthwhile considering domain specific information, as it may significantly improve accuracy and levels of granularity of the discovered provenance.

In this paper, we illustrated our approach with one specific use case: news versioning. However, thanks to the general nature of the proposed provenance discovery method, several other use cases are feasible. Examples of possible applications include plagiarism detection, provenance of code snippets and the tracing of information sources used for quotes in online content, such as blogs. Implementation of one or more of these use cases will allow us to further evaluate the approach, and provide more meaningful fine-grained provenance assertions.

8 Conclusions

We developed an approach that succeeds in creating provenance derivations for a large dataset, discovered from a limited amount of information (content and timing information). Our approach is general enough for adaptation in several domains, and is compliant with the current standard, the Provenance Data Model (PROV-DM). When adapted to the use case of news versioning, our approach detected the original source of a news item with 68% precision and 73% recall. These results are promising, considering that there are several potential improvements to be made to the current implementation. Implementing these improvements is the key to future research in this field, in which additional links to the Semantic Web and a more semantically aware similarity measure will further improve the accuracy of the discovered provenance.

Acknowledgments. The research activities in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT, a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union, in the context of the IBBT project Smarter Media in Flanders (SMIF). Companies involved are Belga, Concentra, VRT and Roularta, with project support of IWT.

References

1. Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., Da Silva, P.P.: Provenance XG final report. Final Incubator Group Report (2010)
2. Gómez-Pérez, J.M., Corcho, O.: Problem-solving methods for understanding process executions. Computing in Science & Engineering 10, 47–52. IEEE (2008)
3. Braun, U., Garfinkel, S., Holland, D., Muniswamy-Reddy, K.K., Seltzer, M.: Issues in automatic provenance collection. Provenance and annotation of data, 171–183. Springer (2006)
4. PROV-DM Part 1: The Provenance Data Model, W3C Editor's Draft 29 May 2012 `http://dvcs.w3.org/hg/prov/raw-file/default/model/prov-dm.html`
5. Rizzo, G., Troncy, R.: NERD: Evaluating Named Entity Recognition Tools in the Web of Data. Workshop on Web Scale Knowledge Extraction (WEKEX'11) (2011)
6. Iacobelli, F., Nichols, N., Birnbaum, L., Hammond, K.: Finding new information via robust entity detection. Proactive Assistant Agents AAAI Fall Symposium (2010)
7. Hasan, M., Salem, S., Pupacdi, B., Zaki, M.: Clustering with lower bound on similarity. Advances in Knowledge Discovery and Data Mining, 122–133 (2009)
8. Zhao, J., Sahoo, S.S., Missier, P., Sheth, A., Goble, C.: Extending semantic provenance into the web of data. Internet Computing, 40–48. IEEE (2011)
9. Zhao, J., Gomadam, K., Prasanna, V.: Predicting Missing Provenance using Semantic Associations in Reservoir Engineering. 2011 Fifth IEEE International Conference on Semantic Computing (ICSC)(2011)
10. Zhang, J., Jagadish, HV.: Lost source provenance. Proceedings of the 13th International Conference on Extending Database Technology, ACM (2010)
11. Ram, S., Liu, J.: A new perspective on Semantics of Data Provenance. First International Workshop on the role of Semantic Web in Provenance Management (SWPM) (2009)
12. Moreau, L.: The foundations for provenance on the web. Now Publishers (2010)
13. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., Milios, E.: Information retrieval by semantic similarity. International Journal on Semantic Web and Information Systems (IJSWIS), 55–73 (2006)