International Journal of Database Theory and Application
Vol.9, No.4 (2016), pp.309-320
http://dx.doi.org/10.14257/ijdta.2016.9.4.29

Modeling and Tracing Web Content Provenance

Jing Ni [1], Jia Hao [2], Xuemei Li[1] and Tong Zhao[1]

[1]*Department of Information Management, Beijing Institute of Petrochemical Technology, Beijing, China 102617*
[2]*School of mechanical engineering, Beijing Institute of Technology, Beijing, China 100081*
[1]*nijing@bipt.edu.cn*, [2]*haojia632@bit.edu.cn*,

Abstract

In recent years, research in data provenance has attracted a lot of attention, since it helps to judge the relevance and trustworthiness of the information enclosed in the data. However, many webpages still lack provenance annotation, and this is a main obstacle of tracing the content. In this paper, we propose a model for on-line Web paper variation, based on the W3C PROV Data Model. A semantic similarity clustering method is adopted to determine the relationship within the documents derivation, and feature words variation and the responsible person can be found with the aid of PROV-O. To verify this model, a detailed case study is shown in this paper.

Keywords*: semantic web, linked data, data provenace modeling, PROV-O, web content derivation*

1. Introduction

Provenance, refers to origin of data product. It allows us to verify their quality, to discover the dependences between data items, and to decide whether they can be trusted. Data provenance is a hot topic in the Semantic Web field [1]. Provenance is an essential part of trust and value assessment of web content, as it describes everything involved in producing this content. The PROV-AQ document t[2] describes several options to access provenance: providing a link header in the HTTP response of the resource; providing a link element in its HTML representation; providing a prov:has_provenance relation in its RDF representation. However, most of existing web content does not have these information. It is therefore necessary to discover provenance information on web content in automated ways.

In regard to discovering provenance information on the web automatically, prior work of the semantic community can be grouped in two main categories. (1) Use analysis of the annotated historical datasets with complete provenance information to capture semantic associations that may imply identical provenance [3]. (2) Develop automatic collecting system runs at the operating system level [4]. In [5], authors investigated the characteristics and requirements of provenance on the Web and described how the Open Provenance Model (OPM) can be used as a foundation for the creation of W3P, a provenance model and ontology designed to meet the core requirements for the Web. Recent studies [6] have focused on high-level knowledge provenance. We agree with the author on the fact that providing information as RDF and liking to LOD cloud would make provenance metadata more transparent and interlinked with other sources.

PROV-DM (PROV Data Model) is recommendation from the W3C on representing provenance and has been applied to various use cases. For example, authors in [7] developed a simple extractor and used PROV-DM to encode the essential elements in express history of revisions in Wikipedia. In [8], authors describe a prototype of a multi-signal pipeline for reconstructing provenance and show preliminary results of

ISSN: 2005-4270 IJDTA

reconstructing dependencies between documents in the context of clinical guidelines and associated documents which model the history of clinical guidelines with PROV Such work facilitates the understanding of provided recommendations by practitioners. List of implementations reported to the PROV Working Groups [9] has been over 65 and still continue to grow.

PROV was deliberately kept as generic and extensible as possible, to allow for all possible use cases. For example, authors in [10] present the Political Roles (PRoles) ontology, an OWL2 DL ontology for the description of political relationships between persons. Building upon existing the Provenance Ontology (PROV-O) .PRoles provides a clear ontological characterization of political roles and related events, establishing a link between the description of such concepts and the documents from which this information is distilled.. A general extension to PROV-DM was proposed by [11] in order to capture the concept of uncertainty in two ways: uncertainty in provenance statements and uncertainty about the content of an entity whose provenance is assessed. This last extension is very useful when algorithms with a certain degree of uncertainty are used to assert the provenance.

Our contributions are as follows: in this paper, we introduce a number of new attribute values to extend PROV-DM [12], to govern the use of these attributes values. In more detail, we provide: (1) a structured ontology for information variation and provenance on web content; (2) extensions of entities and activities relevant for web content; (3) we apply the similarity and clustering approach for extracting the entity and properties, and (4) through the property recognition by named entity relation extraction, document properties can be built, linking to the LOD cloud.

2. Analysis Classes and Properties of PROV-O in our Research

The PROV Ontology (PROV-O) defines the OWL2 Web Ontology Language encoding of the PROV-DM. We begin by providing an overview of the PROV provenance model PROV-O[12] provides us with three essential (core) elements: entities, activities and agents.

Entities (class prov: Entity) are arbitrary things we want to describe the provenance of. They can have relations between each other. The notable relation between entities are the derivation of an entity from another (prov:wasDerivedFrom)

Activities (class prov: Activity) act upon entities. Mainly include the usage (prov:used) of an entity by an activity, the generation (prov:wasGeneratedBy) of a new entity by an activity.

Agent (class prov: Agent) has responsibility. For the existence of entities, it is expressed by the property prov:wasAttributedTo; for past activities, the responsibility of an agent is expressed by prov:wasAssociatedWith.

A prov:collection is subclasses of prov:entity which provides a structure to some constituents and are themselves entities. These constituents are said to be member of the collections described with property prov:hadMenber. In this research, we take the document as the collection of feature properties and the feature properties of the document as members of the entity.

We briefly summarize the alternation and specialization relation in PROV-O as properties: prov:alternatedOf, prov:wasGeneralizedBy and prov:specializationOf . prov:alternatedOf can be used to specify same thing in different backgrounds or the same thing from different views, which is similar to owl:seeAlso , owl:sameAs or skos:exactMatch. prov:wasGeneralizedBy and prov:specializationOf refer to the relationship between the upper and lower concepts. So prov:alternatedOf can be mapped to owl:sameAs, owl:seeAlso or skos: exactMatch; prov:wasGeneralizedBy can be mapped to skos:broader and prov:specializationOf can be mapped to skos:narrower. Shows as Table 1.

Table 1. Mapping Between General Ontology and Prov Ontology

General Ontology	Properties of PROV
owl:sameAs，owl:seeAlso skos:exactMatch,	prov：alternatedOf
skos:broader	prov：wasGeneralizedBy
skos:narrower	prov：specializationOf

The PROV Ontology (PROV-O) defines the OWL2 Web Ontology Language encoding of the PROV Data Model This ontology is a lightweight ontology that can be adopted in a wide range of applications. The classes and properties in PROV Ontology are defined such that they can not only be used directly to represent provenance information, but also can be specialized for modeling application-specific provenance details in a variety of domains. Thus, the PROV Ontology is expected to be both directly usable in applications as well as serve as a "reference model" for creating domain-specific provenance ontologies and thereby facilitates interoperable provenance modeling [12].

In next section, we introduce a number of new attribute values to extend PROV, and relevant extensions to PROV-Constraints [7] to govern the use of these attributes values.

3. Modeling Paper on-Line

The PROV-POL model refers to paper on line provenance. The purpose of PROV-POL is to offer an easily reusable model to trace the content and acquire data provenance, especially offering maximum expressiveness. For another, we borrowed the already defined concepts from PROV-DM wherever possible, and defined our own extensions for specific use cases. This way we improve clarity and we encourage reusability of the model.

3.1 Entity Extension

In order to model papers that are published by users, we propose the following extensions that are subtypes of prov:Entity:

prov-pol:Paper: denotes the general class of Papers.

Papers on the web might be original papers or revised papers. We define the following categories as subtypes of prov-pol:Paper:

prov-pol:OriginalPaper denotes an original paper that is not derived from any other paper and the user who published it is the author of a specific paper.

prov-pol:RevisedPaper denotes a paper that is produced by modifying an existing paper. This means that the user who submits such a paper may or may not share the original opinion of the original paper. It is possible that the information carried by the original paper is altered.

prov-pol:P denotes a property which can mainly express the main concept or metadata of specific on-line paper.

3.2 Activity Extension

Next we define the following activity that refers to paper revision and is a subtype of prov:Activity

prov-pol:RevisePaper denotes a generic revision of a paper. It must generate a prov-pol:Paper, and may use another prov-pol:Paper.

Note that the subtype of the generated prov-pol:Paper (original or revised) can be inferred from the usage of another prov-pol:Paper by the prov-pol:RevisePaper. If the content of the generated Paper was altered from that of the used one, it is a prov-pol: RevisedPaper.

Whereas an original Paper does not have dependencies on other Papers, revised Papers can be traced back to their original sources through derivation. PROV-O already provides most of the concepts needed to model this, in the form of prov:Revision, and prov:PrimarySource.

We observe that paper3 was indirectly derived from paper2. To model this dependency, we introduce the concept prov-pol:IndirectDerivation. This way we can model multi-step provenance and trace how Papers are being derived, without being restricted to the previous step only.

At this point, we express the following constraints:

An prov-pol:OriginalPaper cannot be derived from a prov-pol: Paper.

A revised Paper should always be derived from another Paper. Generation of a provpol: RevisedPaper and usage of a prov-pol:Paper by a prov-pol:RevisePaper implies that the first Paper was derived from the later by prov:Revision.

3.3 Agent Extension

We define the following agent is a subtype of prov:Agent

prov-pol:OriginalAuthor denotes an original author whose paper is not derived from any other paper and the user who published it is the author of a specific paper.

prov-pol:AuthorFollower denotes an author whose paper is produced by modifying an existing paper.

A prov-pol:Paper is always attributed to a user *prov:Agent* using the relationship *prov:wasAttributedTo*.

4. Web Content Provenance Approach

4.1 Analysis of Webpage Direvation

Suppose there are any two web documents a and b, if $time\ (doc_b) > time\ (doc_a)$, then the following two probabilities may occur: (1) Document a is direct sources of document b; (2) Document b is not directly changed from document a, instead, document a straightlyimpact the document c, or, from document a to document b may go through a number of steps in the process of variation(*prov-pol:IndirectDerivation*). In general, the notable features of web text provenance are that the variation details are often unknown and cannot be informed of specific processes.

In PROV-O, *prov:Derivation* is a class that means conversion from one entity to another entity, reconstruction of one entity to another or a new entity resulting from entity updating. Derivation may consist of the following process: the entity B is produced from entity A, and the activity between these two entities is a specific conversion derivation; however, changes may also be not uncertain, such as changes from entity A to entity K, which experiences several evolutions. However, specific changes between entities and the details of activities are often unknown.

4.2 Analysis of Web Content Properties

Vocabulary is the element of the document; derivations of a Web page content are represented by changes of vocabulary, especially vocabulary that reflects the subject of the topic, namely metadata. Therefore, by adopting the PROV-O, we make the following assumptions: there are two types of entities, one is the document entity, and the other is document properties. Because the document feature description consists of multiple semantic properties, semantic property is defined as a member of the document entity. The document belongs to a specific field, and the semantic property mainly refers to the metadata of the article, assuming that the creation time of all documents is available. Our

aim is to analyze the document provenance relationship through semantic properties automatically.

Based on the above analysis, we developed the following schema: (1) finding similar pages and discovering agents (Prov:agent); (2) tracing property changes and finding derivations in detail.

4.3 Automatic Discovery Processes for Document Provenance

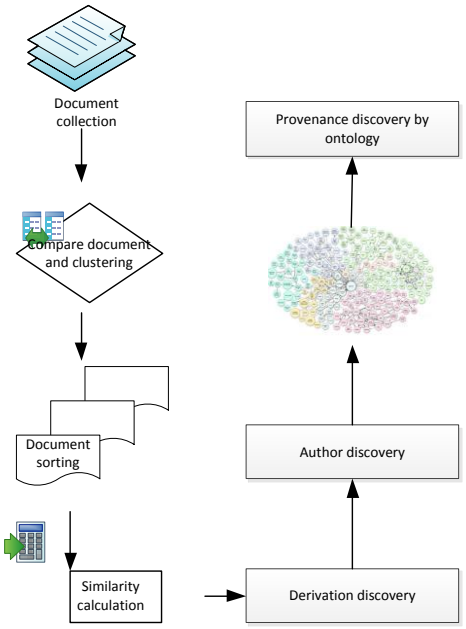

Figure 1. Automatic Discovery Processes for Document Provenance

As shown in Figure 3, documents are collected first and are then clustered by document clustering algorithms. Then, the documents are sorted according to the time for the same cluster. Second, through the calculation of text similarity, the maximum document similarity under the same clustering can be found. After determining the relationship between entities according to the PROV model, those responsible are identified. From the above steps, similar documents and the agents are determined. Finally, details of document changes are discovered by the internal property variation of the document, mainly through named entity recognition extraction. The relationship of the document and the property are established and linked to the LOD cloud by software. Then, the upper and lower properties or similar properties can be found through the ontologies in LOD. As a result, the provenance of changes can be located with fine granularity.

5. The Process Procedures

5.1 Similar Document and Agent Discovery Based on Document Clustering

Assume that the same document has semantic similarity after variations. Documents were clustered according to semantic similarity first, and *Ts* is set as similarity threshold decided by empirical test. If any two documents belong to the same cluster, given document doc_a, the condition that document similarity between doc_a and doc_b is always greater than any other document similarity with known documents doc_a, and the release date of doc_b is later than doc_a, namely, time (doc_b) > time (doc_a). We consider doc_b

originated from doc_a and modified it. From the above procedure, the relationship "*wasDerivedFrom*" between entity doc_a and entity doc_b, is determined by the PROV model, and the "*used*" or "*wasGeneratedBy*" relationship between entities and activities can also be determined.

Normally, authors or editors of the document, those who are responsible for similar documents, can be found through document metadata tagging. For example, foaf:givenName.

5.2 Tracing Property Changes in Details Within a Cluster

Because the document has been modified, some semantic property changed, and others remained the same. These changes will be represented by PROV-O with *alternative, generalization and specialization.* Some properties are ignored and some new properties are added.

In this research, we acquire the properties of the document by applying named entity extraction technology. Once the properties are identified, we define the activity property *usage* to link the property and the document that the property belongs to and to obtain agent information by semantic comparisons. However, to trace at the fine-grained level, modeling of *alternative, generalization and specialization* are described. As Figure 5 shows, according to the PROV-O model, new property Pj may be *alternateOf* or *specializationOf* or *wasGeneralizedBy* Pi. Generation is the completion of production of a new entity by an activity. This entity did not exist before generation and becomes available for usage after this generation. As Figure 4 shows,

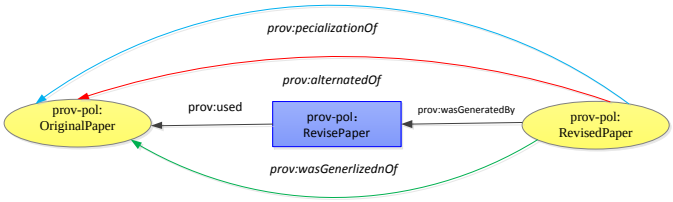

Figure 2. Illustration for Property Derivation

Because relationship between the property and the document is a type of "memberOf" relationship, the provenance of a document can be discovered through the properties of the document. For example, assuming we can identify properties of document1 as P1, P2, and P3 and properties of document2 as P3, P4, and P5, comparing document properties of document1 and document2, P3 is a common property, and P4 is a refinement of P2 (*specializationOf).* Namely, the revision activity used P2 and generated property P4. Property P1 was ignored, and P5 was added, as shown in Figure 5.

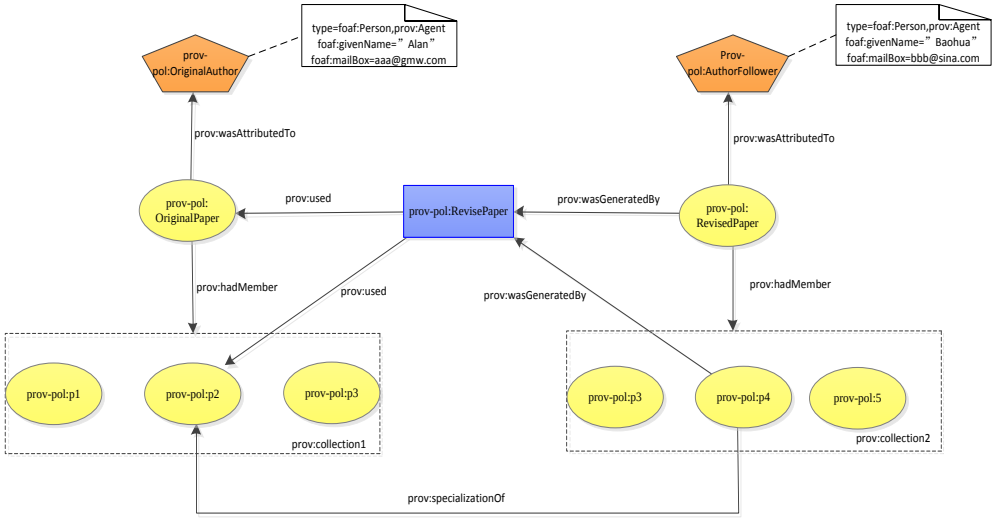

Figure 3. PROV Model for Property Derivation of Paper

Case study and verification to illustrate the problem and assess the results more clearly, we choose gm news topics as an example. References are not needed in news report, which leads to a lack of ability to determine the original source or content, especially inner changed content. The approach of this paper aims to fill this gap by detecting the variations from one release to another and finding the provenance and intermediate revision.

1) Sample selection and experimental platform. GM is becoming a hot topic in food safety. We use this topic as an example to test news provenance. The corpus of this research was derived from www.sina.com.cn and news.gmw.com, with 500 papers. We use Java as code and choose Weka as clustering engine. For segmentation engine, we use IKAnalyzer.

2) Preprocess before clustering. The preprocessing before clustering consists of the following basic processes: preprocessing, Chinese word segmentation, statistical calculations, and feature extraction.

Pre-processing. This case applied a relational database to store content with UTF-8 encoding format.

Chinese word segmentation. Take the continuous text content into Word collection (bag of words), taking into account the well-known Chinese "three noes" issue that requires a dictionary, including specialty dictionaries, as well as stop words dictionary support. The IKAnalyzer default dictionary is used while adding the GM-related vocabulary. This process generates a structured dataset shown as Figure4.

News 1	word1	word2	word3	·····	word5	word6	word7
News 2	word5	word7	word10	·····	word15	word16	word17
News 3	word16	word19	word20	·····	word27	word20	word3
				·····			
News 80	word16	word19	word20	·····	word27	word20	word3

Figure 4. Chinese Word Segmentation Structure

Statistical calculations. The statistics obtained in the previous step to obtain the weight of each word in the document are used to obtain the properties of the vector representation.

TF-IDF (Term Frequency-Inverse Document Frequency) [13] is often used to construct a vector space model in information retrieval. It evaluates the importance of a word in a paper. The importance increases proportionally with the number of times that a word

appears in a paper, compared to the inverse proportion of the same word in the whole collection of papers. Specifically, Suppose N represents the number of words in a paper; M is the total number of words in a single paper; D represents the total number of papers; Dw represents the number of papers in which the keyword appears. Then, we calculate the weight W=TF*IDF as: TF=N/M; IDF=log (D/DW). The term, which is highest in the score of TF*IDF, can best express the subject of the content.

Generally, the vocabulary is large; thus, we select important words to express the paper properties in a reduced volume. Suppose Di represents document i,Ti represents term j,and Wij represents the weight of term j in document i. The weight vectors can be got. We sort the weight and select the terms(properties) with higher W value.

3)Using K-means algorithm for text clustering. Because the characteristics of the concept set in this research belong to the no sample set, the k-means algorithm is used. Taking each sample as a vector of the space, the dimensions of the vector space equal the number of properties of the concept.

Figure 8 is the clustering results output in XML format. In which, each branch represents a cluster category, each leaf represents an article in a cluster, and each value corresponds to the news number.

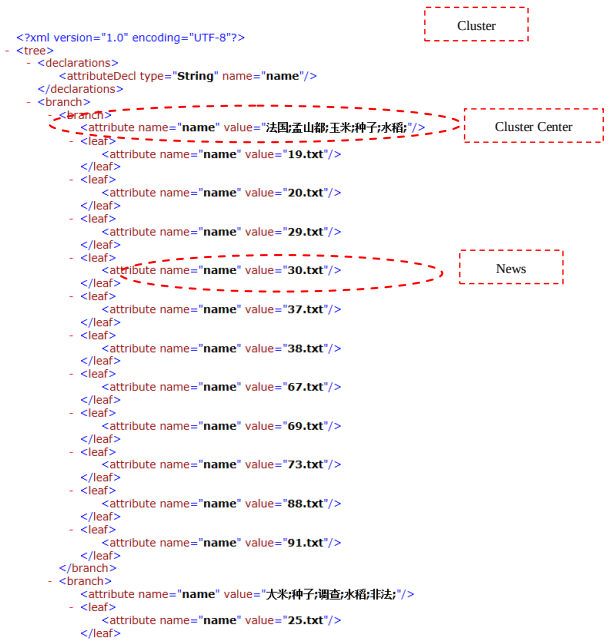

Figure 5. The Clustering Results Output in XML Format

From the steps above, similar documents are get together by clustering.

4) Semantic metadata generation. The OpenCalais Web Service [14] automatically creates rich semantic metadata for the content submitted – in well under a second. Using natural language processing (NLP), machine learning and other methods, Calais analyzes the document and finds the entities within it. However, Calais goes well beyond classic entity identification and returns the facts and events hidden within the text. Calais is now officially part of the Linking Open Data (LOD) Cloud. The Calais ecosystem is exposed via Linked Data endpoints. When Calais extracts an entity from a given text, it also returns an entity URI. This URI is dereferenceable–we can submit an HTTP request, programmatically or via a browser, and get in response useful information and links to other Linked Data and Web assets– all relevant to the entity that is described by the URI. Currently, OpenCalais supports English, French, and Spanish, but it does not support Chinese.

News articles are often not sufficiently marked with descriptive metadata and require tagging to obtain accurate metadata. OpenCalais is a well-constructed, thoroughly tested and free of NER service. However, because OpenCalais currently does not support Chinese and surveys[16] show that Google has the overall highest quality among online translation tools, we chose the Google online translation tool and call the Google API for auto-translation before sending data.

5) The maximum similarity calculation. Tversky's "Contrast Model" [15] systematizes this feature approach. A central assumption of the model is that the similarity of object a to object b is a function of the features common to a and b ("A and B"), those in a but not in b (symbolized "A-B") and those in b but not in a ("B-A"). A diagram exemplifying this is shown in Figure 8. Similarity is not just a function of common features but depends on features that are unique to each object, and the relative importance of these features varies with the parameters y and z.

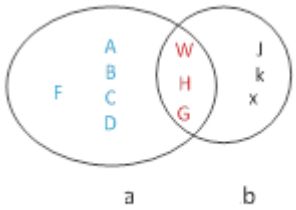

Figure 6. Tversky's "Contrast Model"

Based on this and several other assumptions, Tversky derived the following relationship:

$S(a,b) = xf(a+b) – yf(a-b) – zf(b-a)$.

Here, S is an interval scale of similarity, f is an interval scale that reflects the salience of the various features, and x, y and z are parameters that provide for differences in focus on the different components.

In our project, f(a-b) expresses some properties which are ignored by document b and f(b-a) expresses some properties are added by document b. The common properties (for example:W,H,G) mean the properties are semantically similar, including synonyms, hypernyms, and hyponyms. This is because semantic properties have been identified in the NER steps and these properties are already linked in the LOD cloud, we can follow or dereference these links and connect to a linked data set, such as DBPedia, providing synonyms, hypernyms, and hyponyms. Synonyms include owl:sameAs and skos:exactMatch; hypernyms and hyponyms are SKOS:broader and SKOS:narrower, respectively. By using the method described in Section 2, we create the proper derivation, utilization and generalization relationships.

The maximum degree of similarity was obtained by calculating the public properties according to the comparison model above and finally we get the chain of data provenance. As illustrated by Figure 7.

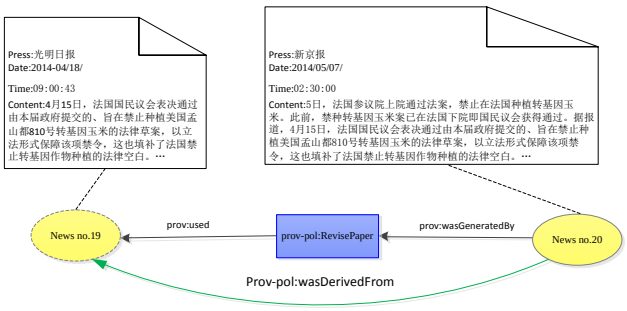

Figure 7. Example of Discovered Provenance Use Case

6. Experimental Results

We use OpenCalais to obtain the set of properties in a specific cluster and select the Tversky algorithm for similarity calculation in the same cluster. We calculate the maximum similarity while harvesting f(a-b) and f(b-a). Tracing different property changes in similar papers by f(a-b) and f(b-a),we can get the property collection. With a field ontology dictionary, the relation, such as alternateOf or specialization or generlizationOf, in the PROV model can be acquired.

Paper no.19 and paper no.20 are the most similar, and the named entity are extracted from OpenCalais.

We integrated crawler tools in the bottom layer, text clustering, similarity calculation, Google translate and the CALAIS module. Figure 8 shows the results of provenance through the interface.

Figure 8. Automatic Discovery Systems for Web Paper Provenance

We change the threshold Ts for different values and test the precision at different thresholds, Ts=0.5-0.9, the precision can be achieved between 87.6% and 90.1%.

7. Conclusions

In this paper, based on the PROV-O, we proposed a new model named PROV-POL which enables to analyze on-line paper. Data provenance and text processing method are applied to analysis the derivation of web papers. Taking GM food web pages as a case study, we apply linked data and semantic web technology to acquire the provenance of the Web. This will help to judge relevance and trustworthiness of information on Web. Overall, we have shown that text similarity combined with PROV model and linked data is a good start towards reconstructing the provenance and applied in provenance tracing.

Acknowledgement

This work is sponsored by the Education Ministry Funding Project for the Development of Liberal Arts and Social Sciences (No.12YJA87001, 15YJAZH052), Beijing Social Science Fund (No.14SHB010), National Science Fund(No.51505032), National Training Program of Innovation and Entrepreneurship for Undergraduates(No.2016J00134)

References

[1] Y. Gil, J. Cheney, P. Groth, O. Hartig, S. Miles, L. Moreau and P. P. Da Silva, "Provenance XG final report", Final Incubator Group Report, **(2010)**.
[2] G. Klyne, P. Groth, L. Moreau, O. Hartig, Y. Simmhan, J. Myers, T. Lebo, K. Belhajjame and S. Miles, "PROV-AQ: Provenance Access and Query", W3C, **(2013)**.
[3] J. Zhao, K. Gomadam and V. Prasanna, "Predicting Missing Provenance using Semantic Associations in Reservoir Engineering", Proceedings of 2011 Fifth IEEE International Conference on Semantic Computing, Palo Alto, USA, **(2011)**.
[4] U. Braun, S. Garfinkel, D. A. Holland, K. K. Muniswamy-Reddy and M. I. Seltzer, "Issues in Automatic Provenance Collection", Proceedings of First International Provenance and Annotation Workshop, Chicago, USA, **(2006)**.
[5] A. Freitas, T. Knap, S. O'Riain and E. Curry, "Building an OPM based provenance model for the Web", Future Generation Computer Systems, vol. 27, **(2011)**, pp. 766-774.
[6] D. N. Tom, C. Sam, V. D. Davy, E. Mannens and R. V. de Walle, "Automatic Discovery of High-Level Provenance Using Semantic Similarity", Proceedings of 4th International Provenance and Annotation Workshop, Santa Barbara, USA, **(2012)**.
[7] P. Missier and Z. Chen, "Extracting PROV provenance traces from wikipedia history pages", Proceedings of 16th International Conference on Extending Database Technology/ 16th International Conference on Database Theory, Genoa, Italy, **(2013)**.
[8] S. Magliacane and P. Groth, "Towards reconstructing the provenance of clinical guidelines", Proceedings of Semantic Web Applications and Tools for Life Sciences, Paris, France, **(2012)**.
[9] T. Huynh, P. Groth and S. Zednik, "(Eds.), and W3C Provenance Working Group", PROV Implementation Report. W3C Working Group Note, **(2013)**.
[10] D. Marilena, P. Silvio, T. Francesca and V. Fabio, "Political Roles Ontology (PRoles): enhancing archival authority records through Semantic Web technologies", Proceeding Computer Science, vol. 38, **(2014)**, pp. 60-67.
[11] T. D. Nies, S. Coppens, E. Mannens and R. V. de Walle, "Modeling uncertain provenance and provenance of uncertainty in w3c prov", Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, **(2013)**.
[12] T. Lebo, S. Sahoo and D. McGuinness, "PROV-O: The PROV Ontology", Retrieved, from http://www.w3.org/TR/2013/REC-prov-dm-20130430/, **(2014)**.
[13] G. Salton, A. Wong and C. S. Yang, "A vector space model for automatic indexing", Communications of the ACM, vol. 18, no. 11, **(1975)**, pp. 613-620.
[14] Calais, "Connect Everything. http://www.opencalais.com/about".
[15] Tversky, "Contrast Model", http://www.pigeon.psy.tufts.edu/avc/dblough/theory.htm
[16] X. W. Xu, "The Evaluation of Translation of Online Translation Websites", Journal of Zhengzhou Institute of Aeronautical Industry Management (Social Science Edition), vol. 31, no. 2, **(2012)**, pp. 124.

Authors

Jing Ni, is an associate professor in Department of Information Management, Beijing Institute of Petrochemical Technology. Her research interest includes data provenance, semantic web and linked data. She has published more than 30 research papers in the local and international journals. She also participated in number of international conferences and technical workshops.

Jia Hao, received the Ph.D degree in mechanical engineering from the Beijing Institute of Technology, Beijing, China in 2014. While pursuing the Ph.D degree, he spent one year (2012.10 to 2013.10) at the University of Michigan, Ann Arbor, USA. From 2014 to 2016, he was with Beijing Institute of Technology as a postdoc research. Now he is a visiting researcher at the University of Aizu, Aizu-Wakamatsu, Japan. His research

interests include awareness system, decision support systems, knowledge engineering and tacit knowledge management.

Xuemei Li, is undergraduate student in Department of Information Management, Beijing Institute of Petrochemical Technology.

Tong Zhao, is undergraduate student in Department of Information Management, Beijing Institute of Petrochemical Technology.

