Assessing Content Value for Digital Publishing through Relevance and Provenance-based Trust

Tom De Nies

Supervised by Erik Mannens and Rik Van de Walle
Ghent University - iMinds - Multimedia Lab
Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium
tom.denies@ugent.be

Abstract. Due to the abundance of content on the Web, content authors and publishers have a pressing need for systems that select content that is valuable for them, is trustworthy and is related to their own work. Additionally, the value of their own work needs to be assessed before it is published, to guarantee high value for the consumer. In this doctoral research, we investigate how to use Semantic Web technologies to automatically assess the value of content that is – or is about to be – digitally published. To achieve this, we propose methods to assess the relevance of content to existing publications, retrieve or reconstruct its provenance, and derive a trust assessment from this provenance. We discuss our evaluation methods, and present some preliminary results.

1 Problem Statement

Nowadays, content producers and consumers in the digital publishing world are all facing the same problem: an abundance of content, available from an ever increasing number of sources. News sites, blogs, social media, and digital libraries are overflowing with content. However, human content creators and curators do not receive more time to filter through this continuous stream of content than before the existence of the Web. On the contrary, the need for immediate reporting increases, while the patience of the consumer decreases. In other words, there is a clear need for a system that assists the author or publisher by automatically assessing the value of content. Note that content value cannot be expressed as a single numerical value, as it might differ depending on the target audience. For example, in the case of news publishing, the content value of an article is primarily determined by its newsworthiness. This newsworthiness includes its relevance to a certain reader group's interests, and the trustworthiness of the information.

In this work, we analyze two aspects that we believe consolidate the essential components of content value: *relevance* and *trust*. After this decomposition, the problem becomes twofold. On the one hand, it is important to determine *for whom* the content is relevant, and *which aspects* of the content make it relevant. On the other hand, the user needs an indication whether or not the content is to be considered as trustworthy. *Provenance* – defined as information about entities, activities, and people involved in producing a piece of data or thing [22] – plays an important role in making assessments about this trustworthiness, as does the *reputation* of the entities, agents and processes involved.

2 Relevancy

The problem stated in Sect. 1 is relevant to several use cases in the digital publishing world, including production of news, eBooks, digital magazines, or more open Web content, such as blogs and microposts on social media. It is also important to note that the problem is relevant to information consumers as well as to those on the production side of information. While the areas of content filtering and recommendation systems with the consumer as end-user are widely researched, the content creator is often overlooked. This is why our approach specifically aims to assist the content creator during the production process. As the information overload on the Web has the highest impact on news organizations, supporting journalists is our main use case. Especially citizen journalists could benefit from our approach, as they cannot rely on the resources a professional journalist has access to.

3 Related Work

Most works in literature about data quality deal with the quality assessment of machine-generated data gathered by observation systems. For the assessment of human-generated content, only a limited number of solutions are proposed [21]. To the best of our knowledge, there is currently no system that integrates both relevance and trust assessments on the content level into one integrated value assessment system. However, a great deal of work is found in literature with respect to individual aspects of our approach. Therefore, we divide the related work into each of the components of our approach: *relevance assessment*, *provenance reconstruction*, and *trust assessment*.

In our work, we apply semantically aware similarity metrics when performing relevance assessments. In literature, the importance of this semantic awareness, together with the novelty of recommendations is stressed [15, 23].

As we point out in [5], existing techniques for the automated production of provenance mostly rely on disclosure by the user or capturing of all provenance information, possibly without understanding the semantics of their observations [2]. A number of domain-specific techniques used to reconstruct lost or missing provenance information exist [11, 25, 26]. However, only a limited number of approaches [5, 9, 13] have been proposed to provide a more generic method to reconstruct provenance [19].

Finally, a significant amount of work can be found on trust and quality assessment of information on the Web [1]. As we describe in [6], most approaches agree that reputation is essential to generating trust. One approach makes use of this fact, together with the trusted relationships of the user [12]. In another work, this *trust network* is automatically built, based on the similarity between users [24]. However, provenance is equally as important as reputation and trust networks, and can be used to assess data quality [14]. Some approaches use the best of both worlds, and determine a trust value by combining provenance and reputation [3, 18].

4 Research Questions

The main research question in our work is:

- *How can we automatically assess the value of content on the Web?*

To answer this question, we also answer the following questions:

- *Can we automatically assess content relevance on the Web?*
- *How can we automatically derive a trustworthiness assessment from semantic enrichments and metadata, such as provenance information, associated with content on the Web?*
- *When this provenance information is incomplete or missing, how can we retrieve or reconstruct it?*

5 Hypotheses

Our research questions have lead to the following hypotheses:

- *The two main aspects of content value are its relevance to its consumer, and its trustworthiness.*
- *Trustworthiness of content is dependent on the content's provenance and the reputation of the entities, agents and processes involved.*
- *When (partially) missing, provenance can be reconstructed based on the content's semantic similarity to other content.*
- *Semantic similarity of content can be measured using a metric based on extracted semantic features.*

6 Proposed Approach

We provide a high-level overview of our approach in Fig. 1. In the next paragraphs, we describe each of the essential components in detail.

We consider two scenarios in which our approach is applicable. In the first scenario, the *content producer* (e.g., the author or publisher) submits his or her own content, to assess its potential value for future readers. In the second scenario, the *content consumer* (e.g., a reader or research journalist) is searching a large dataset for content related to the document he or she is reading, or intends to publish. This large dataset can either be *closed* (e.g., the publisher's archive), or *open* (e.g., datasets on the Web).

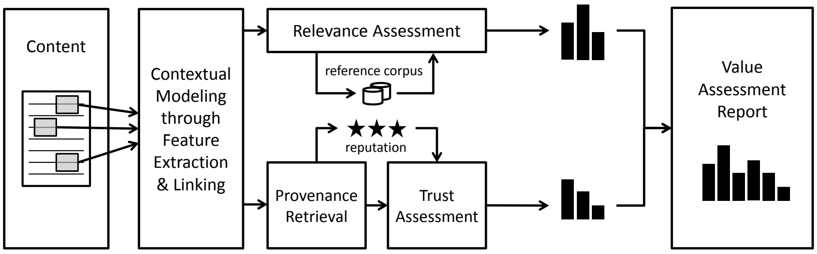

Fig. 1. High-level overview of our approach. A contextual model is used to generate the content's relevance, reconstruct its provenance, and assess its trustworthiness.

Relevance Assessment As can be seen in Fig. 1, the basis for our approach is the creation of a contextual model of the content whose the value is assessed. To achieve this, we rely on semantic feature extraction and linking methods, such as Named-Entity Recognition (NER) services. Here, it is important that the extracted Named Entities (NEs) are accurately disambiguated, and linked to a machine-understandable data source, such as DBpedia. As we describe in [5, 7], we use the properties of these entities to calculate a NE-based semantic similarity metric, to serve as an essential part of our provenance reconstruction and relevance assessment method. To determine the overall relevance of content, we calculate its semantic similarity to one or more reference corpora. These corpora can include sets of recent and/or popular news articles, micropost streams from social media, or publications categorized per topic, etc. The eventual goal is to return a list of highly relevant publications, which are then used to make a statement about the overall relevance of the input content.

Provenance Access and Reconstruction In some cases, the provenance of information on the Web can already be accessed in a standard way, as described in [16]. However, in most cases, the provenance is incomplete or missing, and we need to reconstruct it. For a detailed description of our proposed provenance reconstruction method, we refer to [5]. Essentially, we propose a method that creates clusters of semantically similar documents, using the NE-based similarity metric described above. Then, each document is assumed to have originated from the oldest document in its cluster. We also derive more fine-grained provenance, if the documents have sufficient semantic features to make accurate assumptions. A consequence of this approach is that each of our reconstructed provenance statements has a different confidence value. We model this uncertainty using our customly defined Uncertainty Provenance (UP) [4] attributes, to enable the derivation of trust statements about the provenance in a later stage.

Trust Assessment As we state in [6], we believe the key to making trust assessments about information on the Web is to reason over its provenance as well as to consider the reputation of the entities, agents and processes involved in creating it. Furthermore, as described in [17], it is important to identify *distrust events*. These are indicators that cause a user to lose confidence in the trustworthiness of a document. Instead of providing the user with a single ‘trust score’, we generate a number of statements that indicate the trustworthiness of the document. This way, the user can make a well-informed decision whether or not to trust the document. We reason over the provenance, generating statements regarding its availability, its validity and its contents (e.g., the number of sources listed). Finally, we also use crowd-based services, such as Web of Trust[1] to assess the reputation of the entities and agents mentioned in the provenance.

Determining Content Value The final step in our approach is to present the acquired information in such a way that the user is able to determine the strong and weak points of the content. Here, it is important to provide tangible, and comparable results, to allow for automatic ranking of multiple content items. However, the reasoned statements generated in the relevance and trust assessment steps should also be made available to the user. For example, in the first scenario (value assessment), the user of our system (the content producer) requires a *fine-grained* value assessment of his or her content. This means that he or she not only requires the *raw* value assessment, but also the reasoning statements, indicating *why* the content is valuable. On the other hand, in the second scenario (content selection), the user (the content consumer seeking valuable content) requires a *coarse-grained* value assessment, used to rank the results.

7 Reflections

Our approach combines relevance and trust assessment in a novel way. This allows us to make fine-grained assessments, which are useful to both content producers and consumers. Traditional recommendation systems cannot offer these fine-grained aspects, providing the user only with a ranked list of results. Our work augments this traditional approach, by presenting the user with the strong and weak points of content, together with references to other, relevant content. Additionally, the use of Semantic Web technologies allows for a generic, easily adaptable approach to content value, rather than the typically domain-specific approach of traditional recommender systems. However, and important aspect that remains to be addressed, is how to deal with the complex psychology behind relevance and trust. Indeed, relevance and trust are subjective, relative to the person making the inquiry. Therefore, we will have to consider modeling the interests and trust relations of the users of our approach as well.

8 Evaluation Plan

By definition, content value is dependent on the end-user and the use case. Together with the complex, yet generic nature of our approach, this makes it infeasible to set up a general evaluation. Therefore, we plan to evaluate the proposed approach in specific, representative use cases that provide a suitable framework

[1] `http://www.mywot.com`

to demonstrate its effectiveness. Throughout our research, our main use case has been that of online news, which is particularly suitable in an Information Retrieval (IR), provenance and trust context, and thus suits all components of our approach. Additionally, newsworthiness assessment is a suitable application of our approach as a whole. We will divide the evaluation in four parts, one for each of the three components, and one for the approach as a whole.

For the evaluation of the relevance assessment, we rely on established evaluation metrics and benchmarks from the IR community, such as TRECVID and MediaEval. Typically, these benchmarks offer an extensive textual dataset and a number of queries, of which the ground truth is known.

At the time of writing, no comprehensive benchmarks or standard datasets for provenance reconstruction are available. There has been recent work on repurposing existing datasets for provenance reconstruction [20], but none of the discussed datasets covers all of our use cases. Therefore, as a preliminary evaluation, we created a small multi-lingual dataset to serve as ground truth, consisting of 410 news stories in French and Dutch [5], of which the provenance (primary sources) was known. While the results of this small evaluation were rather positive (as can be seen in Sect. 9), a larger, more comprehensive gold standard dataset is needed in future work.

Due to its subjective nature, no ground truth datasets for trust and/or value assessments are available, to the best of our knowledge. Therefore, we will use a crowdsourcing platform such as Amazon Mechanical Turk to create our ground truth. This way, we will be able to provide a direct (subjective) comparison between the human assessments and our machine-generated assessments. Additionally, we will perform a more objective evaluation by creating a dataset of known trusted content, injected with some known untrusted content. This can be collected from real-world examples, such as spam messages, or from sources that are known to make false statements (e.g., the news satire site "The Onion").

9 Preliminary Results

In the first two years of our research, we have performed a number of preliminary experiments with the current version of the components needed for our approach. The results are summarized in Table 1, and are briefly discussed below.

Relevance Assessment [7]		**Provenance Reconstruction**		**Newsworthiness Criteria Detection**	
MRR	MAP	Precision	Recall	Precision	Recall
0.254	0.171	0.723	0.445	0.839	0.661

Table 1. Preliminary results of the relevance assessment, provenance reconstruction and value (newsworthiness) assessment approaches.

To test our relevance assessment approach, we participated in the MediaEval 2012 Search and Hyperlinking Task [10], which consisted of retrieving and linking relevant video segments from a text corpus made up by approx. 10000 automatically transcribed videos. The search task was evaluated using Mean Reciprocal

Rank (MRR), and the linking task using Mean Average Precision (MAP). As shown in Table 1, our results were promising, with room for improvement. For the full results, we refer to our working notes [7].

As discussed in Sect. 8, we evaluated our provenance reconstruction approach using a small multi-lingual dataset of 410 news stories [5]. When identifying the original sources of the news items, and reconstructing the corresponding derivations, we obtained a reasonably high precision and moderate recall [5].

To evaluate the feasibility of our value assessment approach in a practical scenario, we built a tool for content-based newsworthiness assessment [8] of news articles. The tool detects the presence of news determinants that are used in literature to classify a news article as "newsworthy". As a preliminary evaluation, we manually assessed the accuracy of the automatic detection of these determinants. The results[2] were promising, at 90% precision and 66% recall. Of course, a more thorough evaluation by domain experts is necessary before making any definitive conclusions about the approach as a whole.

To summarize, although more evaluation is needed, we have shown that our approach and its components are indeed feasible. With the right improvements, more extensive evaluation, and further integration of the approach as whole, we are confident that we will be able to test our hypotheses and provide an answer to all our research questions.

Acknowledgments. These research activities were funded by Ghent University, iMinds, the IWT Flanders, the FWO-Flanders, and the European Union.

References

[1] Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Journal of Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (2007)

[2] Braun, U., Garfinkel, S., Holland, D.A., Muniswamy-Reddy, K.K., Seltzer, M.I.: Issues in automatic provenance collection. In: Provenance and annotation of data, pp. 171–183. Springer (2006)

[3] Ceolin, D., Groth, P., van Hage, W.R., Nottamkandath, A., Fokkink, W.: Trust Evaluation through User Reputation and Provenance Analysis. In: 8th International Workshop on Uncertainty Reasoning for the Semantic Web. p. 15 (2012)

[4] De Nies, T., Coppens, S., Mannens, E., Van de Walle, R.: Modeling uncertain provenance and provenance of uncertainty in W3C PROV. In: Proceedings of the 22nd World Wide Web conference, companion. pp. 167–168 (2013)

[5] De Nies, T., Coppens, S., Van Deursen, D., Mannens, E., Van de Walle, R.: Automatic discovery of high-level provenance using semantic similarity. In: Provenance and Annotation of Data and Processes, pp. 97–110 (2012)

[6] De Nies, T., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R., Michaelides, D., Moreau, L.: Easy Access to Provenance: an Essential Step Towards Trust on the Web. In: METHOD 2013 at COMPSAC 2013 (2013)

[7] De Nies, T., Debevere, P., Van Deursen, D., De Neve, W., Mannens, E., Van de Walle, R.: Ghent University-IBBT at MediaEval 2012 Search and Hyperlinking: Semantic Similarity using Named Entities. In: MediaEval (2012)

[2] data available at `http://users.ugent.be/~tdenies/AVALON/ISWC/`

[8] De Nies, T., D'heer, E., Coppens, S., Van Deursen, D., Mannens, E., Van de Walle, R.: Bringing Newsworthiness into the 21st Century. In: Web of Linked Entities (WoLE) at ISWC 2012. pp. 106–117 (2012)

[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. pp. 1–10. USENIX Association (2009)

[10] Eskevich, M., Jones, G.J., Chen, S., Aly, R., Ordelman, R., Larson, M.: Search and hyperlinking task at MediaEval 2012. In: MediaEval 2012 Workshop, Pisa, Italy, October 4-5. CEUR-WS.org (2012)

[11] Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. Concurrency and Computation: Practice and Experience 20(5), 485–496 (2008)

[12] Golbeck, J., Mannes, A.: Using trust and provenance for content filtering on the semantic web. In: Models of Trust for the Web Workshop (2006)

[13] Groth, P., Gil, Y., Magliacane, S.: Automatic Metadata Annotation through Reconstructing Provenance. In: Semantic Web in Provenance Management, CEUR Workshop Proceedings. vol. 856 (2012)

[14] Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: Proceedings of the International Workshop on Semantic Web and Provenance Management, Washington DC, USA (2009)

[15] Iacobelli, F., Birnbaum, L., Hammond, K.J.: Tell me more, not just more of the same. In: Proceedings of the 15th international conference on Intelligent user interfaces. pp. 81–90. ACM (2010)

[16] Klyne, G., Groth , P., (Eds.) et al.: PROV-AQ: Provenance Access and Query. W3C Note (2012)

[17] Li, X., Lebo, T., McGuinness, D.L.: Provenance-based strategies to develop trust in semantic web applications. In: Provenance and Annotation of Data and Processes, pp. 182–197. Springer (2010)

[18] Lim, H.S., Moon, Y.S., Bertino, E.: Provenance-based trustworthiness assessment in sensor networks. In: Proceedings of the Seventh International Workshop on Data Management for Sensor Networks. pp. 2–7. ACM (2010)

[19] Magliacane, S.: Reconstructing provenance. In: 11th International Semantic Web Conference (ISWC), pp. 399–406. Springer (2012)

[20] Magliacane, S., Groth, P.: Repurposing Benchmark Corpora for Reconstructing Provenance. In: Semantic Publications (SePublica) Workshop (ESWC2013) (2013)

[21] Melucci, M.: Contextual Search: A Computational Framework. Foundations and Trends® in Information Retrieval 6(4-5), 257–405 (2012)

[22] Moreau, L., Missier, P., (Eds.) et al: PROV-DM: The PROV Data Model. W3C Recommendation (2013)

[23] Passant, A.: Measuring semantic distance on linking data and using it for resources recommendations. In: Proceedings of the AAAI Spring SymposiumŤ Linked Data Meets Artificial Intelligence. vol. 3 (2010)

[24] Tavakolifard, M.: Similarity-based techniques for trust management. Web Intelligence and Intelligent Agents pp. 233–250 (2010)

[25] Zhang, J., Jagadish, H.: Lost source provenance. In: 13th International Conference on Extending Database Technology. pp. 311–322. ACM (2010)

[26] Zhao, J., Gomadam, K., Prasanna, V.: Predicting Missing Provenance using Semantic Associations in Reservoir Engineering. In: 5th IEEE International Conference on Semantic Computing (ICSC). pp. 141–148. IEEE (2011)