Journal of Trust Management
a SpringerOpen Journal

RESEARCH Open Access

Efficient semi-automated assessment of annotations trustworthiness

Davide Ceolin*, Archana Nottamkandath* and Wan Fokkink

*Correspondence: d.ceolin@vu.nl; a.nottamkandath@vu.nl
The Network Institute, VU University Amsterdam, de Boelelaan, 1081a, 1081HV Amsterdam, The Netherlands

Abstract

Crowdsourcing provides a valuable means for accomplishing large amounts of work which may require a high level of expertise. We present an algorithm for computing the trustworthiness of user-contributed tags of artifacts, based on the reputation of the user, represented as a probability distribution, and on provenance of the tag. The algorithm only requires a small number of manually assessed tags, and computes two trust values for each tag, based on reputation and provenance. We moreover present a computationally cheaper adaptation of the algorithm, which clusters semantically similar tags in the training set, and builds an opinion on a new tag based on its semantic relatedness with respect to the medoids of the clusters. Also, we introduce an adaptation of the algorithm based on the use of provenance stereotypes as an alternative basis for the estimation. Two case studies from the cultural heritage domain show that the algorithms produce satisfactory results.

Keywords: Trust; Annotations; Semantic similarity; Subjective logic; Clustering; Cultural heritage; Tagging; Crowdsourcing; Provenance

Introduction

Through the Web, cultural heritage institutions can reach large masses of people, with intentions varying from increasing visibility (and hence visitors) to acquiring user-generated content. Crowdsourcing is an effective way to handle tasks which are highly demanding in terms of the amount of work needed to complete and required level of expertise [1], such as annotating artifacts in large cultural heritage collections. For this reason, many cultural heritage institutions have opened up their archives to ask the masses to help them in tagging or annotating their artifacts. In earlier years it was feasible for employees at the cultural heritage institutions to manually assess the quality of the tags entered by external users, since there were relatively few contributions from Web users. However, with the growth of the Web, the amount of data has become too large to be accurately dealt with by experts at the disposal of these institutions within a reasonable time. Nevertheless a high quality of annotations is vital for their business. The cultural heritage institutions need the annotations to be trustworthy in order to maintain their authoritative reputation. This calls for mechanisms to automate the annotation evaluation process in order to assist the cultural heritage institutions to obtain quality content from the Web. Annotations from external users can be either in the form of tags or free text, describing entities in the crowdsourced systems. Here, we focus on tags in

the cultural heritage domain, which describe mainly the content, context and facts about an artifact by associating words to it.

The goal of this paper is to propose an algorithm for computing the trustworthiness of annotations in a fast and reliable manner. We focus on three main evaluation aspects for our algorithm. First, the trust values produced by our algorithm are meant as indicators of the trustworthiness of annotations, and therefore they must be accurate enough to warrant their usefulness. Accuracy of trust values is achieved by carefully handling the information at our disposal and by utilizing the existence of a relationship between the features considered (e.g., the annotation creator) and the trust values themselves. If the information is handled correctly and the relationship holds, then the trust values are accurate enough and form a basis to automatically decide whether or not to use the annotations. We evaluate this first research question by applying our algorithm on two different datasets, one from a SEALINC Media project experiment and the other from the Steve.Museum dataset. In both cases we divide the dataset into two parts, training set and test set, so as to build a model based on subjective logic and semantic similarity in the training set, and then evaluate the accuracy of such a model on the test set.

The goal of the work described in this paper is to automate the process of evaluation of tags obtained through crowdsourcing in an effective way, by means of an algorithm. In fact, crowsourcing provides massive amounts of annotations, but these are not always trustworthy enough. So we aim at automating the process of deciding whether these are satisfactorily correct and of high quality, i.e. of evaluating them. This is done by first collecting manual evaluations about the quality of a small part of the tags contributed by a user, and then learning a statistical model from them. Throughout the paper we refer to this set as "training set". On the basis of such a model, that assumes the existence of a relation between the user reputation and his overall performance, the system automatically evaluates the tags further added by the same user or user stereotype (i.e. set of users behaving similarly, e.g., users that always provide their tags on Sunday morning). We refer to this set of "new" tags as "test set". Suppose that a user, Alex, provides annotations to the Fictitious National Museum. We propose a method that automatically evaluates these annotations, based on a small set of annotations that the museum previously evaluated from which we derive Alex' reputation, or the reputation of the users whose annotation behaviour is similar to Alex'. We will return on this example more in detail in the following sections.

We employ Semantic Web technologies to represent and store the annotations and the corresponding reviews. We use subjective logic to build a reputation for users that contribute to the system, and moreover semantic similarity measures to generate assessments on the tags entered by the same users at a later point in time. By reputation, we mean a value indicating the estimated probability that the annotations provided by a given author (or, later in the paper, by a given user stereotype) are positively evaluated. In subjective logic, we use the expected value of an opinion about an author as the value of the reputation. The opinion is based on a set of previously evaluated tags. In order to reduce the computation time, we cluster evaluated tags to reduce the number of comparisons. Our experiments show that this preprocessing does not seriously affect the accuracy of the predictions, while significantly reducing the computation time. The proposed algorithms are evaluated on two datasets from the cultural heritage domain. These case studies show that it is possible to semi-automatically evaluate the tags entered by users

in crowdsourcing systems into binomial categories (good, bad) with an accuracy above 80%.

Apart from using subjective logic and semantic similarity, we also use provenance mechanisms to evaluate the quality of user-contributed tags. Provenance is represented by means of a data record per tag, containing information on its creation such as time of day, day of the week, typing speed, etc., obtained by tracking user behavior. We use provenance information to group annotations according to the "stereotype" or "behavior" (or "provenance group") that produced them. In other words, we group them depending on whether they are produced by, for instance, early-morning or late-night users. Once the annotations have been grouped per stereotype, we compute a reputation for each stereotype, based on a sample of evaluations provided by an authority: we learn the policy adopted by the authority in evaluating the annotations and we apply the learnt model on further annotations. The fact that we take into account provenance and not rely solely on user identities makes our approach suitable for situations where users are anonymous and only their behavior is tracked. By "policy" we mean a set of rules that the institution adopts to evaluate tags. The fact that we do not have an explicit definition of such rules (nor we could ask for it) determines the need to learn a probabilistic model that aims at mimicking their evaluation strategy. For instance, we do not know a priori if one of two conflicting tags about the same image is wrong, both because we do not know the image (these could refer to two distinct image parts) and the museum policies (which may prohibit their coexistence, in principle). So, we rely only on the museum evaluations and we do not consider the possible impact of conflicting tags.

We propose three algorithms for estimating the trustworthiness of tags. The first learns a reputation for each user based on a set of evaluated tags. Then it predicts the evaluations of the rest of the tags provided by the same user by ranking them according to the user performance in each specific domain (using semantic similarity measures) and then accepting a number of tags proportional to the user reputation. The second algorithm reduces the computational complexity by clustering the tags in the training set, and the third algorithm computes the same prediction on the basis of the user stereotype rather than on the basis of each single user identity. The novelty of this research lies in the automation of tag evaluations on crowdsourcing systems by coupling subjective logic opinions with measures of semantic similarity along with provenance metrics. The only variable parameter that we require is the size of the set of manual evaluations that are needed to build a useful and reliable reputation. Moreover, in the experiments that we performed, varying this parameter did not substantially affect the performance (resulting in about 1% precision variation per five new observations considered in a user reputation). We will further investigate in the future about the impact of this parameter. Using our algorithms, we show how it is possible to avoid asking the curators responsible for the quality of the annotations to set a threshold in order to make assessments about a tag trustworthiness (e.g., accept only tags which have a trust value above a given threshold).

Background and literature review

Trust has been studied extensively in Computer Science. We refer the reader to Sabater and Sierra [2], Gil and Artz [3] and Golbeck [4] for a comprehensive review of trust in computer science, Semantic Web, and Web respectively. The work presented in this paper

focuses on trust in crowdsourced information from the Web, using an adapted version of the definition of Castelfranchi and Falcone [5], reported by Sabater and Sierra [2], so we decide to trust or not trust tags based on a set of beliefs and assumptions about who produced the tags and how these were produced. We quantify these beliefs, for instance, though reputations.

Crowdsourcing techniques are widely used by cultural heritage and multimedia institutions for enhancing the available information about their collections. Examples include the Tag Your Paintings project [6], the Steve.Museum project [7] and the Waisda? video tagging platform [8]. The Socially Enriched Access to Linked Cultural (SEALINC) Media project investigates also in this direction. In this project, Rijksmuseum [9] in Amsterdam is using crowdsourcing on a Web platform selecting experts of various domains to enrich information about their collection. One of the case studies analyzed in this paper is provided by the SEALINC Media project.

Trust management in crowdsourced systems often employs wisdom of crowds approaches [10]. In our scenarios we cannot make use of those approaches because the level of expertise needed to annotate cultural heritage artifacts restricts the potential set of users, thus making this kind of approach inapplicable or less effective. Gamification, that consists of using game mechanisms for involving users in non-game tasks, is another approach that leads to an improvement of the quality of tags gathered from crowds, as shown, for instance, in von Ahn et al. [1]. The work presented here is orthogonal to a gamified environment, as it allows us to semi-automatically evaluate the user-contributed annotations and hence to semi-automatically incentivize them. By combining the two, museums could increase the user incentiviazion (showing his reputation may be enough to incentivize a user) while curating the quality of annotations. Users that participated in the experiments that provided the datasets for our analyses did not receive monetary incentives, so leveraging incentives related to gamification and personal satisfaction (by means of reputation tracking) may reveal to be an important factor in increasing the accuracy of the tags collected. In folksonomy systems such as the Steve.Museum project, tag evaluation techniques such as comparing the presence of the tags in standard vocabularies and thesauri, determining their frequency and their popularity or agreement with other tags (see, for instance, Van Damme et al. [11]) have been employed to determine the quality of tags entered by users. Such mechanisms focus mainly on the contributed content with little or no reference to the user who authored it. Also, in folksonomy systems the crowd often manages the tags, while in our scenarios, the crowd only provides the tags, that are managed by museums or other institutions, according to specific policies. Medeylan et al. [12] present algorithms to determine the quality of tags entered by users in a collaboratively created folksonomy, and apply them to the dataset CiteULike [13], which consists of text documents. They evaluate the relevance of user-provided tags by means of text document-based metrics. In our work, since we evaluate tags, we cannot apply document-based metrics, and since we do not have at our disposal large amounts of tags per subject, we cannot check for consistency among users tagging the same image. Similarly, we cannot compute semantic similarity based on the available annotations (like in Cattuto et al. [14]). In fact, since we do not have at our disposal image analysis software nor explicit museum policies, we can not know if possible conflicts between tags regarding the same image are due to the fact that some are correct and some not, or to the fact that they refer to different aspects (or parts) of a complex picture. Therefore, instead of

assuming one of the two cases a priori, we determine the trustworthiness of the tags on the basis of the reputation of their user or provenance stereotype. As future direction, we plan to consider also inputs from image recognition software that will help us dealing with conflicting or dubious tags. In open collaborative sites such as Wikipedia [15], where information is contributed by Web users, automated quality evaluation mechanisms have been investigated (see, for instance, De La Calzada et al. [16]). Most of these mechanisms involve computing trust from article revision histories and user groups (see Zeng et al. [17] and Wang et al. [18]). These algorithms track the changes that a particular article or piece of text has undergone over time, along with details of the users performing the changes. In our case study, tags do not have a revision history.

Another approach to obtain trustworthy data is to find experts amongst Web users with a good intention (see De Martini et al. [19]). This mechanism assumes that users who are experts tend to provide more trustworthy annotations. It aims at identifying such experts, by analyzing the profiles built by tracking user performance. In our model, we build profiles based on user performance in the system. So the profile is only behavior-based, and rather than looking for expert and trustworthy users, we build a model which helps in evaluating the tag quality based on the estimated reputation of the tag author. However there is a clear relation between highly reputed users and experts, although these two classes do not always overlap. Modeling of reputation and user behavior on the Web is a widely studied domain. Javanmardi et al. [20] propose three computational models for user reputation by extracting detailed user edit patterns and statistics which are particularly tailored for wikis, while we focus on the annotations domain. Ceolin et al. [21] build a reputation- and provenance-based model for predicting the trustworthiness of Web users in Waisda? over time. We optimize the reputation management and the decision strategies described in that paper.

We use subjective logic to represent user reputations, in combination with semantic relatedness measures. This work extends Ceolin et al. [21,22]. Similarity measures have been combined with subjective logic in Tavakolifard et al. [23], who infer new trust connections between entities (e.g., users) given a set of trust connections known a priori. In our paper, we also start from a graphical representation of relations between the various participating entities (annotators, tags, reviewers, etc.), but: (1) trust relationships are learnt from a sample of museum evaluations, and (2) new trust connections are inferred based on the relative position of the tags in another graph, WordNet. We also use semantic similarity measures to cluster related tags to optimize the computations. In Cilibrasi et al. [24], hierarchical clustering is used for grouping related topics, while Ushioda et al. [25] experiment on clustering words in a hierarchical manner. Begelman et al. [26] present an algorithm for the automated clustering of tags on the basis of tag co-occurrences in order to facilitate more effective retrieval. A similar approach is used by Hassan-Montero and Herrero-Solana [27]. They compute tag similarities using the Jaccard similarity coefficient and then cluster the tags hierarchically using the k-means algorithm. In our work, to build user reputations, we cluster the tags along with their respective evaluations (e.g., accept or reject). Each cluster is represented by a medoid (that is, the element of the cluster which is the closest to its center), and in order to evaluate a newly entered tag by the same user, we consider clusters which are most semantically relevant to the new tag. This helps in selectively weighing only the relevant evidence about a user for evaluating a new tag.

In general, different cultural heritage institutions employ different values and metrics of varying scales to represent the trustworthiness of user-contributed information. The accuracy of various scales has been studied earlier. Certain cases use a binary (boolean) scale for trust values, as in Golbeck et al. [28], while binomial values (i.e., the probabilities of two mutually exclusive values zero and one, that we use in our work) are used in Guha et al. [29] and Kamvar et al. [30].

Relevant for the work presented in this paper is the link between provenance and trust. Bizer and Cyganiak [31], Hartig and Zhao [32] and Zaihrayeu et al. [33] use provenance and background information expressed as annotated or named graphs to produce trust values. We use the same class of information to make our estimates, but we do not use named graphs to represent provenance information. We represent provenance by means of the W3C recommendation PROV-O, the PROV Ontology [34,35]. Provenance is employed for determining trustworthiness of user-contributed information in crowdsourced environments in Ceolin et al. [21], where provenance information is used in combination with user reputation to make binomial assessments of annotations. Also, they employ support vector machines for making the provenance-based estimates, while we employ a subjective logic-based approach. Provenance is used for data verification in crowdsourced environments by Ebden et al. [36]. In their work, they introduced provenance tracking into their online CollabMap application (used to crowdsource evacuation maps), and in this way they collect approximately 5,000 provenance graphs, generated using the Open Provenance Model [37] (which has now been superseded by the PROV Data Model and Ontology). In their work they have at their disposal large provenance graphs and can learn useful features about the artifact trustworthiness from the graphs topologies. Here, the graphs at our disposal are much more limited, so we cannot rely on the graph topology, but we can easily group graphs in stereotypes. Provenance mechanisms have also been used to understand and study workflows in collaborative environments as discussed in Altintas et al. [38]. We share the same context with that work, but we do not focus on the workflow of artifact creation.

The rest of the paper is structured as follows: first we describe the research questions tackled and we connect the datasets used with them. Then we explain our research work in detail and present and discuss the obtained results. Lastly, we provide some final conclusions.

Research design and methodology

The goal of this paper is to propose an algorithm for computing the trustworthiness of annotations in a fast and reliable manner. We focus on three main evaluation aspects for our algorithm. First, the trust values produced by our algorithm are meant as indicators of the trustworthiness of annotations, and therefore they must be accurate enough to warrant their usefulness. Accuracy of trust values is achieved by carefully handling the information at our disposal and by utilizing the existence of a relationship between the features considered (e.g., the annotation creator) and the trust values themselves. If the information is handled correctly and the relationship holds, then the trust values are accurate enough and form a basis to automatically decide whether or not to use the annotations. We evaluate this first research question by applying our algorithm on two different datasets, one from a SEALINC Media project experiment and the other from the Steve.Museum dataset. In both cases we divide the dataset into two parts, training set

and test set, so as to build a model based on subjective logic and semantic similarity in the training set, and then evaluate the accuracy of such a model on the test set.

The second evaluation we make regards the possibility to perform trust estimations in a relatively fast manner, by properly clustering the training set on a semantic similarity basis. Here the goal of the contribution is to reduce the computational overhead due to avoidable comparisons between evaluated annotations and new annotations. We evaluate this contribution by applying clustering mechanisms in the training set data of the aforementioned datasets and by running our algorithm for computing trust values on the clustered training sets. The evaluation will check whether clustering reduces the computation time (and in case it does, up to which magnitude) and whether it affects the accuracy of the predictions.

Finally, we show that the algorithm we propose is dependable and not solely dependent on the availability of information about the author of an annotation. Our assumption is that when the identity of the author is not known or when a reliable reputation about the author is not available, we can base our estimates on provenance information, that is, on a range of information about how the tag has been created (e.g., the timestamp of the annotation). By properly gathering and grouping such information to make it utilizable, we can use it as a "stereotypical description" of a user's behavior. Users are often constrained in their behavior by the environment and other factors; for instance, they produce tags within certain periodic intervals, such as the time of the day or day of the week. Being able to recognize such stereotypes, we can compute a reputation per stereotype rather than per user. While on the one hand this approach guarantees the availability of evidence, as typically multiple users belong to the same stereotype, on the other hand this approach compensates on the lack of evidence about specific users. We evaluate our hypothesis over the two datasets mentioned before by splitting them into two parts, one to build a provenance-based model and the other to evaluate it.

Methods

Here we describe the methods adopted and implemented in our algorithm. We start describing the tools that were already available and that we incorporated in our framework, and then we continue with the framework description.

Preliminaries

The system that we propose aims at estimating the trustworthiness of new annotations based on a set of evaluated ones (per user or per provenance stereotype, as we will see). In order to make the estimates, we need to make use of a probabilistic logic that allows us to model and reason about the evidence at our disposal while accounting for uncertainty due to the limited information available. For this reason we employ subjective logic, which fits our needs. Moreover, since the evidence at our disposal consists of textual annotations, we use semantic similarity measures to understand the relevance of each piece of evidence when analyzing each different annotation.

Subjective logic

Subjective logic is a probabilistic logic that we extensively use in our system in order to reason about the trustworthiness of the annotations and the reputations of their authors, based on limited samples. In subjective logic, so-called "subjective opinions" (represented

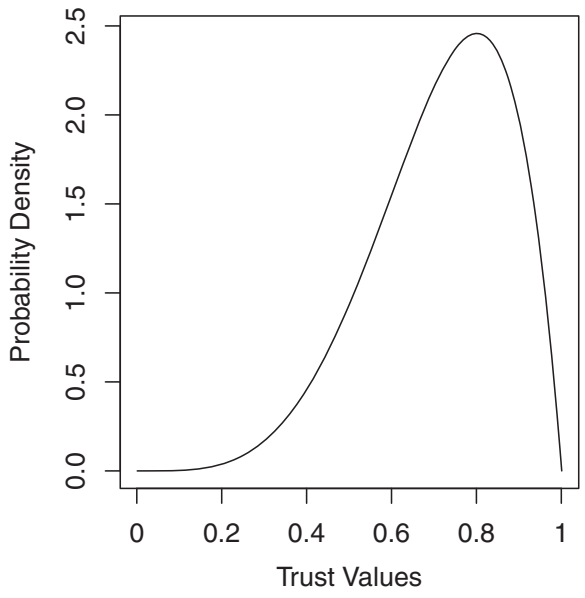

Figure 1 Beta distribution of the user trustworthiness. We use a Beta distribution to describe the probability for each value in the [0 . . . 1] interval to be the right reputation for a given user or the right trust value for a given tag.

as ω) express the belief that source x owns with respect to the value of assertion y (for instance, a user's reputation). When y can assume only two values (e.g., true or false), the opinion is called "binomial"; when y ranges over more than two values, the opinion is called "multinomial". Opinions are computed as follows, where the positive and negative evidence are represented as p and n, respectively, and b, d, u and a represent the belief, disbelief, uncertainty and prior probability, respectively. A binomial opinion is represented as:

$$\omega_y^x(b, d, u) \tag{1}$$

where:

$$b = \frac{p}{p+n+2} \quad d = \frac{n}{p+n+2} \quad u = \frac{2}{p+n+2} \quad a = \frac{1}{2} \tag{2}$$

Such an opinion is equivalent to a Beta probability distribution (see Figure 1), which describes the likelihood for each possible trust value to be the right trust value for a given subject. An expected probability for a possible value of an opinion is computed as:

$$E = b + a \cdot u \tag{3}$$

The belief b of the opinion represents how much we believe that the y statement is true (where y is, for instance, the fact that an annotation is correct), based on the evidence at our disposal. However, the evidence at our disposal is limited (and since the evidence is obtained by asking an authority, such as a museum, to evaluate an annotation, we aim at keeping it limited in order to avoid overloading the museum with such requests), so we must take into account the fact that other observations about the same fact might, in principle, disagree with those currently at our disposal. This leads to uncertainty, which is represented by the corresponding element of the opinion, u. The disbelief d represents, instead, the disbelief that we have about the statement y based on actual negative evidence

at our disposal. Disbelief is the counterpart of belief, and the sum of belief, disbelief and uncertainty is always one ($b + d + u = 1$).

Whereas belief, disbelief and uncertainty are based on the actual evidence observed, the base rate a encodes the prior knowledge about the truth of y according to x (before observing any evidence). The final opinion about the correctness of y is then obtained by aggregating a and b in a weighted manner, so as to make b count more if it is based on more evidence than a (and thus its uncertainty is lower). This explains the meaning of E, which represents exactly such an aggregation. One last consideration regards the fact that E represents both the expected value of the opinion and of the Beta distribution equivalent to it. u is tightly connected to the variance of the Beta distribution, as they both represent the uncertainty about the belief to be significant.

These elements make subjective opinions ideal tools for representing the fact that the probabilities that we estimate about the trustworthiness of annotations, authors and provenance stereotypes (that are defined later in this section) are based on evidence provided by a museum (that consist of evaluations based on the museum's policy), and that the estimates we make carry some uncertainty due to the fact that they are based on a limited set of observations.

Semantic similarity

The target of our trust assessments are annotations (that is, words associated to images, in order to describe them), and our evidence consists of evaluated annotations. In our system we collect evidence about an author or a provenance stereotype (which is defined later in this section) that consist of annotations evaluated by the museum, and we compare each new annotation that needs to be evaluated against the evidence at our disposal. If, for instance, we based our estimates only on the ratio between positive and negative evidence of a given author, then all his annotations would be rated equally. If, instead, we compared every new annotation only against the evidence at our disposal of matching words, we would have a more tailored, but more limited estimate, as we cannot expect that the same author always uses the same words (and so that there is evidence for every word he uses in an annotation).

In order to increase the availability of evidence for our estimate and to let the more relevant evidence have a higher impact on those calculations, we employ semantic relatedness measures as a weighing factor. These measures quantify the likeness between the meaning of two given terms. Whenever we evaluate a tag, we take the evidence at our disposal, and tags that are more semantically similar to the one we focus on are weighed more heavily. There exist many techniques for measuring semantic relatedness, which can be divided into two groups. First, we have so-called "topological" semantic similarity measures, which are deterministic measures based on the graph distance between the two words examined, based on a word graphs (e.g. WordNet [39]). Second, there is the family of statistical semantic similarity measures, which include, for instance, the Normalized Google Distance [40], that measures statistically the similarity between two words on the basis of the number of times that these occur and co-occur in documents indexed by Google. These measures are characterized by the fact that the similarity of two words is estimated on a statistical basis from their occurrence and co-occurrence in large sets of documents.

We focus on deterministic semantic relatedness measures based on WordNet or its Dutch counterpart Cornetto [41]. In particular, we use the Wu and Palmer [42] and the Lin [43] measure for computing semantic relatedness between tags, because both provide us with values in the range $[0, 1]$, but other measures are possible as well. WordNet is a directed and acyclic graph where each vertex v, w is an integer that represents a synset (set of word synonyms), and each directed edge from v to w implies that w is a hypernym of v. In other words w shares a "type-of" relation with v. For instance, if v is the word "winter" (hyponym), w can be the word "season" (hypernym). The Wu and Palmer measure calculates semantic relatedness between two words by considering the depths between two synsets in WordNet, along with the depth of the Least Common Subsumer, as follows:

$$score(s1, s2) = 2 * \frac{depth(lcs(s1, s2))}{depth(s1) + depth(s2)}$$

where $s1$ is a synset of the first word and $s2$ of the second. WordNet is an acyclic graph where nodes are represented by synsets and edges represent hypernym/hyponym relations. If a synset is a generalization of another one, we can measure the depth, that is the distance between the two. The first ancestor shared by two nodes is the Least Common Subsumer. We compute the similarity of all synsets combinations and pick the maximum value, as we adopt the upper bound of the similarity between the two words. The Lin measure considers the information content of the Least Common Subsumer and the two compared synsets, as follows:

$$2 * \frac{IC(lcs(s1, s2))}{IC(s1) + IC(s2)}$$

where IC is the information context, that is the probability of finding the concept in a given corpus, and is defined as:

$$IC(s) = -log\left(\frac{freq(s)}{freq(root)}\right)$$

and $freq$ is the frequency of the synset. So the Wu and Palmer measure derives the similarity of two concepts from their distance from a common ancestor, while the Lin similarity derives it from the information content of the two concepts and their lowest ancestor. The Wu and Palmer similarity measure is more recent and has shown to be effectively combinable with subjective logic in Ceolin et al. [44], so when we deal with datasets of tags in English (Steve.Museum dataset), we use its implementation provided by the python nltk library [45]. Instead, when we work on datasets composed of Dutch tags (e.g., the dataset from the SEALINC Media project experiment), we rely on pyCornetto [46], an interface to Cornetto, the Dutch WordNet. pyCornetto does not provide a means to compute the Wu and Palmer similarity measure, but it provides the Lin similarity measure, and given the relatedness of the two measures, in this case we adopt the Lin measure.

For more details about how to combine semantic relatedness measures and subjective logic, see the work of Ceolin et al. [44]. By choosing to use these measures we limit ourself in the possibility to evaluate only single-word tags and only common words, because these are the kinds of words that are present in WordNet. However, we choose these measures because almost all the tags we evaluate fall into the mentioned categories and because the use of these similarity measures together with subjective logic has already been theoretically validated. Moreover, almost all the words used in the annotations that form the dataset we used in our evaluations are single-word tags and common words,

hence this limitation does not affect our evaluation significantly. The algorithm proposed is designed so that any other relatedness measure could be used in place of the chosen ones, without the need of any additional intervention. The choice of the semantic similarity and how the semantic similarity is used both affect the uncertainty of the expected results of the algorithms that we propose. In fact, these algorithms use semantic similarity to weigh the importance of evidence when evaluating tags, i.e. words associated with cultural heritage artifacts. We use a deterministic semantic similarity measure which, although it constitutes a heuristics, is based on a trustworthy data source (WordNet) and this implies that the measure is less uncertain than a probabilistic measure based on a limited document corpus. Still, the semantic similarity measure represents an approximation and part of the uncertainty these imply is due to their use: semantic similarity measures represent the similarity between synsets, but we have at our disposal only words without an indication of their intended synset (a word may have more meanings, represented by synsets). Since we are situated in a well-defined domain (cultural heritage), and since words are all used to tag cultural heritage artifacts, we assume that words are semantically related, and hence, when computing the semantic similarity between two words, we make use of the maximum of the similarity between all their synsets. Despite the fact that this introduces an approximation, we will show in Section Results that the method is effective.

Datasets adopted

We validate the algorithms we propose over two datasets of annotations of images. The annotations contained in these datasets consist of content descriptions and the datasets contain also the evaluations from the institutions that collected them. For each annotation, the datasets contain information about its author and a timestamp. Since each institution adopts a different policy for evaluating annotations, we try to learn such a policy from a sample of annotations per dataset, and find a relationship between the identity of the author or other information about the annotation and its evaluation.

SEALINC media project experiment

As part of SEALINC Media project, the Rijksmuseum in Amsterdam [9] is crowdsourcing annotations of artifacts in its collection using Web users. An initial experiment was conducted to study the effect of presenting pre-set tags on the quality of annotations on crowdsourced data [47]. In the experiment, the external annotators were presented with pictures from the Web and prints from the Rijksmuseum collection along with a pre-set annotations about the picture or print, and they were asked to insert new annotations, or remove the pre-set ones which they did not agree with (the pre-set tags are either correct or not). A total of 2,650 annotations resulted from the experiment, and these were manually evaluated by trusted personnel for their quality and relevance using the following scale:

- 1 : Irrelevant
- 2 : Incorrect
- 3 : Subjective
- 4 : Correct and possibly relevant
- 5 : Correct and highly relevant
- typo : Spelling mistake

These tags, along with their evaluations, were used to validate our model. For each tag, the SEALINC Media dataset presents the following elements: author identifier, artifact identifier, timestamp, evaluation. We do not focus on the goals of the experiment from which this dataset is obtained, that is, we do not analyze the relation between the kind of tag that was proposed to the user, and the tag that the user provided. We focus on the tag that the user actually proposes and its evaluation and we try to predict the evaluation of the tags provided by each user, given a small training set of sample evaluations about each of them.

We neglect the tags evaluated as "Typo" because our focus is on the semantic correctness of the tags, so we assume that such a category of mistakes would be properly avoided or treated (e.g., by using autocompletion and checking the presence of the tags in dictionaries) before the tags reach our evaluation framework. We build our training set using a fixed amount of evaluated annotations for each of the users, and form the test set using the remaining annotations. The number of annotations used to build the reputation and the percentage of the dataset covered is presented in Table 1: in the first column "# annotation per reputation" we report the number of evaluated annotations we use to build each reputation, while in the second column, "% training set covered" we report the percentage of annotation used as training set compared to the whole dataset.

Steve.Museum project dataset

Steve.Museum [7] is a project involving several museum professionals in the cultural heritage domain. Part of the project focuses on understanding the various effects of crowdsourcing cultural heritage artifact annotations. Their experiments involved external annotators annotating museum collections, and a subset of the data collected from the crowd was evaluated for trustworthiness. In total, 4,588 users tagged the 89,671 artifacts using 480,617 tags from 21 participating museums. Part of these annotations consisting of 45,860 tags were manually evaluated by professionals at these museums and were used as a basis for our second case study. In this project, the annotations were classified in a more refined way, compared to the previous case study, namely as: {Todo, Judgement-negative, Judgement-positive, Problematic-foreign, Problematic-huh, Problematic-misperception, Problematic-misspelling, Problematic-no_consensus, Problematic-personal, Usefulness-not_useful, Usefulness-useful}. There are three main categories: judgement (a personal judgement by the annotator about the picture), problematic (for several, different reasons) and usefulness (stating whether the annotation is useful or not). We consider only "usefulness-useful" as a positive judgement, all the others are considered as negative evaluations. The tags classified as "todo" are discarded, since their evaluation has not been

Table 1 Results of the evaluation of Algorithm 1 over the SEALINC Media dataset

# Tags per reputation	% Training set covered	Accuracy	Precision	Recall	F-measure	Time (sec.)
5	8%	0.73	0.88	0.81	0.84	87
10	19%	0.76	0.87	0.84	0.86	139
15	31%	0.76	0.86	0.86	0.86	221
20	41%	0.84	0.87	0.96	0.86	225

Results of the evaluation of Algorithm 1 over the SEALINC Media dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

performed yet. The Steve.Museum dataset is provided as a MySQL database and consists of several tables. Those most important for us are: "steve_term", that contains information like the identifiers for the artifact annotated and the words associated with them (tags); "steve_session" that reports information about when the tags are provided and by whom, and "steve_term_review" that contain information about the tag evaluations. We join these tables and we select the information that is relevant for us: the tags, their authors, their timestamps (i.e. date and time of creation) and their evaluation. We partition this dataset thus obtained into a training and a test set, as shown in Table 2, along with their percentage coverage of the whole dataset and the obtained results (in the second column, "% of training set covered"). We use the training set to learn a model for evaluating user provided tags, and we consider the tags in the test set as tags newly introduced by the authors for which we build a model.

System description

After having introduced the methods that we make use of, and the datasets that we analyze, we provide here a description of the system that we propose.

High-level system overview

The system that we propose aims at relieving the institution personnel (reviewers in particular) from the burden of controlling and evaluating all the annotations inserted by users. The system asks for some interaction with the reviewers, but tries to minimize it. Figure 2 shows a high-level view of the model.

For each user, the system asks the reviewers to review a fixed number of annotations, and on the basis of these reviews it builds user reputations. A reputation is meant to express a global measure of trustworthiness and accountability of the corresponding user. The reviews are also used to assess the trustworthiness of each tag inserted afterwards by a user: given a tag, the system evaluates it by looking at the evaluations already available. The evaluations of the tags semantically closer to the one that we evaluate have a higher impact. So we have two distinct phases: a first training step where we collect samples of manual reviews, and a second step where we make automatic assessments of tags trustworthiness (possibly after having clustered the evaluated tags, to improve the computation time). The more reviews there are, the more reliable the reputation is, but this number depends also on the workforce at the disposal of the institution. On the other hand, as we will see in the following section, this parameter does not affect significantly the accuracy obtained. Moreover, we do not need to set an "acceptance threshold"

Table 2 Results of the evaluation of Algorithm 1 over the Steve.Museum dataset

# Tags per reputation	% Training set covered	Accuracy	Precision	Recall	F-measure	Time (sec.)
5	18%	0.68	0.79	0.80	0.80	1254
10	27%	0.70	0.79	0.83	0.81	1957
15	33%	0.71	0.80	0.84	0.82	2659
20	39%	0.70	0.79	0.84	0.81	2986
25	43%	0.71	0.79	0.85	0.82	3350
30	47%	0.72	0.81	0.85	0.83	7598

Results of the evaluation of Algorithm 1 over the Steve.Museum dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

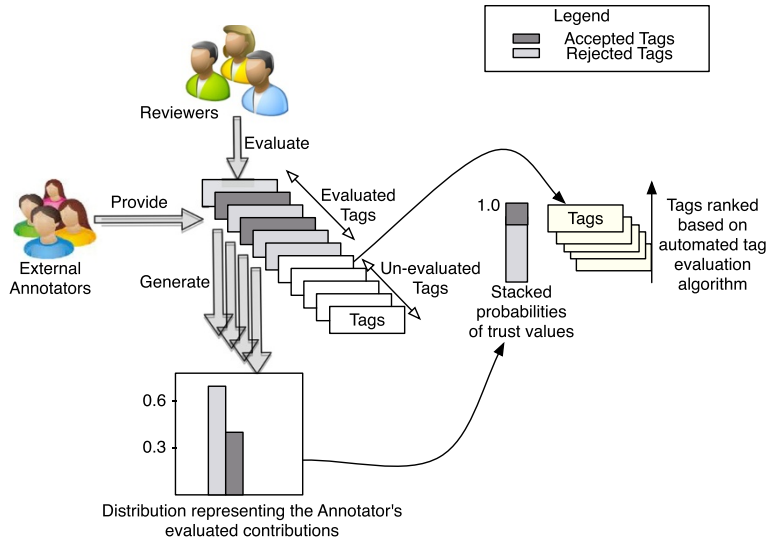

Figure 2 High-level overview of the system. The figure shows the high-level overview of the system. We evaluate a subset of user-contributed tags and use it to build the annotator reputation model. We evaluate the remaining set of the tags based on semantic similarity measures and subjective logic. The unevaluated tags are then ranked based on the annotator reputation model and the model serves as a basis to accept or reject the unevaluated tags from the users.

(e.g., accept only annotations with a trust value of say at least 0.9, for trust values ranging from zero to one), in contrast to the work of Ceolin et al. [21]. This is important since such a threshold is arbitrary, and it is not trivial to find a balance between the risk to accept wrong annotations and to reject good ones.

We introduce here a running example that accompanies the description of the system in the rest of this section. Suppose that a user, Alex (whose profile already contains three tags which were evaluated by the museum), newly contributes to the collection of the Fictitious National Museum by tagging five artifacts. Alex tags one artifact with "Chinese". If the museum immediately uses the tag for classifying the artifact, it might be risky because the tag might be wrong (maliciously or not). On the other hand, had the museum enough internal employees to check the external contributed tag, then it would not have needed to crowdsource it. The system that we propose here relies on few evaluations of Alex's tags by the Museum. Based on these evaluations, the system: (1) computes Alex's reputation; (2) computes a trust value for the new tag; and (3) decides whether to accept it or not. We describe the system implementation in the following sections.

Annotation representation

We adopt the Open Annotation model [48] as a standard model for describing annotations, together with the most relevant related metadata (like the author and the time of creation). The Open Annotation model allows to reify the annotation itself, and by treating it as an object, we can easily link to it properties like the annotator URI or the time of creation. Moreover, the review of an annotation can be represented as an annotation which target is an annotation and which body contains a value of the review about the annotation.

To continue with our example, Figure 3 and Listing 1 show an example of an annotation and a corresponding review, both represented as "annotations" from the Open Annotation model.

Listing 1 Example of an annotation and respective evaluation. The annotation is represented using the Annotation class from the Open Annotation model. The evaluation is represented as an annotation of the annotation.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix oac: <http://www.w3.org/ns/openannotation/core/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:user_1 oac:annotator Annotation; foaf:givenName "Alex" .
ex:annotation_1 oac:hasBody tag:Chinese;
                oac:annotator ex:user_1;
                oac:hasTarget ex:img_231;
                rdf:type oac:annotation .
ex:review oac:hasBody ex:ann_accepted;
          oac:annotator ex:reviewer_1;
          oac:hasTarget ex:annotation_1;
          rdf:type oac:annotation .
ex:annotation_accepted oac:annotates ex:annotation_1 .
```

Trust management

We employ subjective logic [49] for representing, computing and reasoning on trust assessments. There are several reasons why we use this logic. First, it allows to quantify the truth of statements regarding different subjects (e.g. user reputation and tag trust value) by aggregating the evidence at our disposal in a simple and clear way that accounts both for the distribution of the observed evidence and the size of it, hence quantifying the uncertainty of our assessment. Second, each statement in subjective logic is equivalent to a Beta or Dirichlet probability distribution, and hence we can tackle the problem from a statistical point of view without the need to change our data representation. Third, the logic offers several operators to combine the assessments made over the statements of our interest. We made a limited use of operators so far, but we aim at expanding this in the near future. Lastly, we use subjective logic because it allows us to represent formally the fact that the evidence we collect is linked to a given subject (user, tag), and is based on a specific point of view (reviewers for a museum) that is the source of the evaluations.

Trust is context-dependent, since different users or tags (or, more in general, agents and artifacts) might receive different trust evaluations, depending on the context from which they situate, and the reviewer. In our scenarios we do not have at our disposal an explicit description of trust policies by the museums. Also, we do not aim at determining a generic tag (or user) trust level. Our goal is to learn a model that evaluates tags as closely as

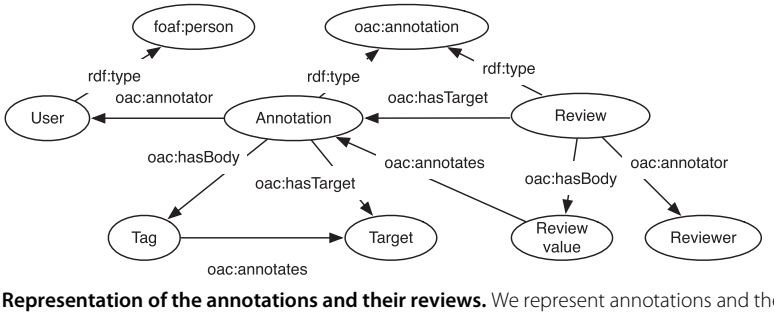

Figure 3 Representation of the annotations and their reviews. We represent annotations and their reviews as annotations from the open annotation model.

possible to what that museum would do, based on a small sample of evaluations produced by the museum itself.

User reputation computation and representation We define a user reputation as a global value representing the user's ability to tag according to the museum policy. With global we mean that the user reputation is not related to a specific context, because this value should represent an overall trust level about the user production: a highly reputed user is believed to have the ability to produce high-quality tags and to choose tags/artifacts related to his/her domain of expertise. Also, the possible number of topics is so high that defining the reputation to be topic-dependent would bring manageability issues. Expertise will be considered when evaluating a single tag, as we will see in the next paragraph.

We require that a fixed number of user-contributed tags are evaluated by the museum. Based on those evaluations we compute the user reputation using subjective opinions, as in Equation 4.

$$\omega_u^m \left(\frac{p_u^m}{p_u^m + n_u^m + 2}, \frac{n_u^m}{p_u^m + n_u^m + 2}, \frac{2}{p_u^m + n_u^m + 2}, \frac{1}{2} \right) \tag{4}$$

where m and u represent the museum and the user respectively and p and n the count of positive and negative pieces of evidence respectively. So, for instance, p_u^m is the count of positive pieces of evidence that the museum m collected about user u, and n_u^m the negative ones.

The algorithm that we will describe makes use of a single value representing the user reputation, so in place of the values computed as in Equation 4, the algorithm makes use of the expected value of that opinion, as shown in Equation 5.

$$E(\omega_u^m) = \frac{p_u^m}{p_u^m + n_u^m + 2} + \frac{1}{2} \cdot \frac{2}{p_u^m + n_u^m + 2} \tag{5}$$

To continue with the running example, suppose that Alex contributed three tags: {Indian, Buddhist} were evaluated as accepted and {tulip} as rejected. His reputation is:

$$\omega_{Alex}^{museum} = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{2} \right) \quad E(\omega_{Alex}^{museum}) = 0.6 \tag{6}$$

Tag trust value computation and representation Tag trust values are represented by means of subjective opinions, as in Equation 7.

$$\omega_t^m \left(\frac{p_t^m}{p_t^m + n_t^m + 2}, \frac{n_t^m}{p_t^m + n_t^m + 2}, \frac{2}{p_t^m + n_t^m + 2}, \frac{1}{2} \right) \tag{7}$$

Here, we still use the tags created by the user and the corresponding evaluations to compute the trust value, but despite the computation of the user reputation, evidence is weighed with respect to the similarity to the tag to be evaluated. This means, that we do not consider each piece of evidence as equally contributing to the computation of the reputation, i.e. evidence is weighed according to the semantic similarity with respect to the tag that we are evaluating. So p and n are determined as in Equation 8, where *sim* is a semantic relatedness measure and t is a tag to be evaluated and so, despite Equation 4 where each piece of evidence counted as one, here each piece of evidence

counts as a real number between zero and one corresponding to the value of the semantic similarity.

$$p_t^m = \Sigma_{t_i \in train} sim(t, t_i) \text{ if } evaluation(t_i) = true$$
$$n_t^m = \Sigma_{t_i \in train} sim(t, t_i) \text{ if } evaluation(t_i) = false \quad (8)$$

The tag "Chinese" inserted by Alex is evaluated as:

$$p^m_{Chinese} = sim(\text{Chinese, Indian}) + sim(\text{Chinese, Buddhist}) = 1.05$$
$$n^m_{Chinese} = sim(\text{Chinese, tulip}) = 0.1$$
$$\omega^m_{Chinese}\left(\frac{1.05}{1.05+0.1+2}, \frac{0.1}{1.05+0.1+2}, \frac{2}{1.05+0.1+2}, \frac{1}{2}\right)$$
$$E(\omega^m_{Chinese}) = 0.95$$

Tag evaluation In order to evaluate tags (i.e. decide to accept or reject them), we define an ordering function on the set of tags based on their trust values (see Equation 9). The ordered set of tags is represented as $\{t\}_1^{|tags|}$, where $|tags|$ is the cardinality of the set of tags. For tags t_1 and t_2,

$$t_1 \leq t_2 \iff E(\omega^m_{t_1}) \leq E(\omega^m_{t_2}) \quad (9)$$

Recall that $E(\omega^m_u)$ is the user reputation, being the expected percentage of correct tags created by the user. Hence, we accept the last $E(\omega^m_u) \cdot |tags|$ tags in $\{t\}_1^{|tags|}$ (see Equation 10) as $\{t\}_1^{|tags|}$ is in ascending order, so we accept the tags having higher trust value.

$$evaluation(tag) = \begin{cases} rejected & \text{if } t \in \{t\}^1_{E(\omega^m_u)\cdot|tags|} \\ accepted & \text{otherwise} \end{cases} \quad (10)$$

We saw how the reputation of Alex was 0.6. He inserted five new tags, so $0.6 \cdot 5 = 3$ will be accepted. The tag "Chinese" had a trust value of 0.95, which ranks it as first in the ordered list of tags. Therefore the tag "Chinese" is *accepted*.

Algorithm

We provide here a pseudocode representation of the algorithm that implements the tag evaluation procedures, and we explain it in detail.

Input The algorithm takes as input two vectors. The first vector, i.e. the training set, is composed of tuples formed by tags, their evaluation (e.g., "useful") and the user identifier (which consists of a URI, since we use the Semantic Web representation described above). The second vector (test set) is composed of tuples formed by tags and the identifier of the user that provided them.

Output The intended output consists of a vector of tuples formed by the tags in the test set and their estimated evaluation.

build_user_reputation Builds a reputation for each user in the training set, following Equation 4. A reputation is represented as a vector of probabilities for possible tag evaluations.

trust_values Trust values are represented as vectors of probabilities of possible tag evaluations, following Equation 7.

Algorithm 1: Algorithm to compute trust values of tags base on user reputation.

Input: A finite set of elements in $Training\_set = \{\langle tag, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle tag, UserID\rangle\}$

Output: A finite set of evaluated tags $Result\_Test\_set = \{\langle tag, trust\_values\rangle\}$

```
1  for UserID ← UserID_1 to UserID_n do
2      ▷ for all tags in Training_set
3      rep[ UserID] ← build_reputation(Training_set)
4  for UserID ← UserID_1 to UserID_n do
5      ▷ for all users in Test_set
6      for Tag ← tag_1 to tag_n do
7          ▷ for all tags in Test_set
8          trust_values[ Tag] = comp_tv(Training_set)
9      s_tags ← sort(tags(trust_values)
10     Result ← assess(s_tags,rep[UserID])
11 return Result
```

comp_tv Implements Equation 7 using Equation 8. The value actually stored is the expected value of the opinion, that is $E(\omega_t^m) = \frac{p_t^m}{p_t^m+n_t^m+2} + \frac{1}{2} \cdot \frac{2}{p_t^m+n_t^m+2}$.

sort_tags The tags are sorted according to their trust value, following the ordering function in Equation 9.

assess The assess function assigns an evaluation to the tag, by implementing Equation 10.

Clustering semantically related tags

Reputations built using large training sets are likely to be more accurate than those built using smaller ones. On the other hand, the larger the set of tags used for building the reputation, the higher the number of comparisons we will have to make to evaluate a new tag. In order to reduce this tension, we cluster the tags in the training set of a user based on semantic similarity, for each resulting cluster we compute the medoid (that is, the element of the cluster which is, on average, the closest to the other elements), and we record the evidence counts. Clustering is performed on a semantic basis, that is, tags are clustered in order to create subsets of tags having similar meanings. After having clustered the tags, we adapt the algorithm so that we compute a subjective opinion per cluster, but we weigh it only on the semantic distance between the new tag and the cluster medoid. In this way we reduce the number of comparisons (we do not measure the distance between the new tag and each element of the cluster), but we still account for the size of the training set, as we record the evidence counts of it. We use hierarchical clustering [50] for semantically clustering the words, although it is computationally expensive, because: (1) we know only the relative distances between words, and not their position in a simplex (the semantic distance is computed as $1 - similarity(word_1, word_2)$), and this is one of the algorithms that requires such kind of input; and (2) it requires only one input argument, a real number "cut", that determines the number of clusters of the input set S of words. If cut $= 0$, then there is only one cluster; if cut $= 1$, then there are n clusters, where n is the cardinality of S. Clustering is performed offline, before any tag is evaluated, and here we

focus on the improvement of the performance of the newly introduced tags. Algorithm 2 incorporates these optimizations. As Algorithm 1, Algorithm 2 takes as input the training set (composed of tuples formed by a tag, its evaluation and its author identifier) and a test set (composed of tuples formed by tags and their author identifier) and outputs a set of tuples formed by the tags in the test set and their estimated evaluations.

To continue with the running example, the museum can cluster the tags inserted by Alex before making any estimate. We have only three tags in the training set, which result in two clusters, {Indian, Buddhist} and {tulip}.

$$p^m_{Chinese} = sim(\text{Chinese, Indian}) \cdot 2 = 1.75$$

$$n^m_{Chinese} = sim(\text{Chinese, tulip}) = 0.1$$

$$\omega^m_{Chinese}\left(\frac{1.75}{1.75+0.1+2}, \frac{0.1}{1.75+0.1+2}, \frac{2}{1.75+0.1+2}, \frac{1}{2}\right)$$

$$E(\omega^m_{Chinese}) = 0.72$$

This result is different from the previous trust value computed in a non-clustered manner (0.95). However, this variation affects all the computed trust values, and the overall performance of the algorithm even benefits from it, as a consequence of a better distribution of the evidence weights.

Provenance-based trust values

The algorithms described so far are based on the fact that there exists a relationship between the identity of an author and the trustworthiness of his annotations, or that the user reputation is a meaningful estimate. However, there might be cases when the user reputation is not available, for instance if there is not enough evidence about his trustworthiness or in case his identity is not known. We show that the algorithm is not firmly

Algorithm 2: Algorithm to compute trust values of tags based on user reputation, with clustering of the evaluated tags in the training set.

Input: A finite set of elements in $Training\_set = \{\langle tag, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle tag, UserID\rangle\}$

Output: A finite set of evaluated tags $Result\_Test\_set = \{\langle tag, trust\_values\rangle\}$

```
1  for UserID ← UserID_1 to UserID_n do
2      ▷ for all tags in Training_set
3      rep[UserID] ← build_reputation(training_set)
4      clusters[UserID] ← build_clust(training_set)
5      medoids[UserID] ← get_med(clusters,UserID)
6  for UserID ← UserID_1 to UserID_n do
7      ▷ for all users in Test_set
8      for Tag ← tag_1 to tag_n do
9          ▷ for all tags in Test_set
10         trust_values[Tag] = comp_tv(medoids[UserID],rep[UserID])
11     sort_tags ← sort(trust_values)
12     Result ← assess(sort_tags,rep[UserID])
13 return Result
```

dependent on the user reputation and, in case this is not available, other classes of information can be used as well. This class of information is so-called provenance information about how an artifact (in this case, an annotation) has been produced, and represents, therefore, an extension of the information about the sole author of the annotation.

We follow a reasoning similar to a previous work of Ceolin et al. [21], as we use "provenance stereotypes" to group annotations. By stereotype we mean a class of provenance traces classified according to the user behavior they hint at. For instance, we could have "Monday early morning users" or "Saturday night users". We suppose that a given behavior should be associated with a particular reputation and hence with a given degree of trustworthiness of the annotations created in that manner, for two reasons:

- The trustworthiness of a given annotation might be affected by when it is created. For instance, late at night, users may on average be more tired and hence less precise than on other moments of the day.
- Users tend to follow a regular pattern in their behavior, because, for instance, their availability for annotating is constrained by their working time. Therefore, by considering their behavior, we implicitly consider their identity as well, even when they act as anonymous users, as shown in Ceolin et al. [21].

In order to apply this kind of reasoning, we need to refer to the provenance information at our disposal about the annotations. In particular, these include only the day of the week and the time of creation for the dataset considered, but other information, when available, might be used as well (e.g., the typing duration for a given annotation). Since annotations are hardly created at the same time, in general do not coincide, we need to group them in order to be able to identify patterns in the data that allow us to link specific provenance information to the trustworthiness of the tags. In fact, the creation time of a tag may be recorded as a timestamp, but since tags are probably created at different times, we need to increase the granularity of this piece of information and analyze the part of the day or the day of the week when the tag was created, rather than the exact moment (tracked by the timestamp). Of course, this grouping introduces some uncertainty in the calculations because it introduces an approximation and because, in principle there are several possible groupings that we can apply, with different granularity and semantics (e.g., the days can be distinguished in weekdays and weekends, or simply be kept as single days of the week). In the next section, we report the results we obtained and we provide a possible explanation of why the grouping we propose allowed us to obtain the results we achieved, in the case studies we analyzed. Lastly, from the modeling point of view, each group or stereotype can be thought of as a **prov:bundle** from the PROV Ontology [35], that is a "named set of provenance descriptions", where each set groups provenance traces according to the day of the week and the part of the day they belong to.

Despite the mentioned previous work, we do not apply support vector machines to learn the trustworthiness of the annotations created with a given stereotype. Rather, we collect a predefined amount of evidence (i.e. of evaluated annotations) per group, and we evaluate the remaining annotations of the same group based on the reputation estimated using the evidence collected, so as to exploit the provenance semantics instead of using it only as a statistical feature.

For representing provenance information we adopt the W3C Recommendation PROV-O Ontology [35], which provides founding types and relations for representing this

specific kind of information, like entities and activities, which coincide with tags and tag creation processes respectively.

Computing the reputation of a provenance stereotype Once we have decided how to group the provenance traces, we start collecting evidence per group. We fix a limit to the amount of evidence needed to create the opinion representing the stereotype's reputation. (In the experiment described in the next section we vary this limit to evaluate the impact it has on the accuracy of the reputation itself.). The reputation is computed as in the **build_reputation()** procedure described in Algorithm 3. First we determine which stereotype the annotation belongs to. Then we increment the evidence count for the evaluation of the current tag until we reach the limit per stereotype. Lastly, we convert the list of evidence counts in subjective opinions.

Algorithm 3: Algorithm to compute trust values of tags using provenance stereotypes. First we present the procedure for computing the reputation of the provenance stereotypes and then we predict the trustworthiness of tags based on their provenance group.

```
1  procedure build_reputation()
       Input: A finite set of elements in Training_set = {⟨tag,evaluation, ProvenanceID⟩}
       Output: A set of provenance group reputations
                   Result_Test_set = {⟨ProvenanceID, reputation_values⟩}
2      for tag in training_set_tags do
3          i ← tag.get_stereotype_id()
4          if length(trainingset[stereotypes[i])]) <n then
5              trainingset[ length(trainingset[stereotypes[i]]) + 1] ←
               get_eval( tag)
6          else
7              testset[ length(testset[stereotypes[i]]) + 1] ← get_eval(tag)
8      for s in stereotypes do
9          rep[ s] ← compute_reputations(s)
10     return s
   Input: A finite set of elements in Training_set = {⟨tag,evaluation, ProvenanceID⟩}
          and Test_set = {⟨tag, ProvenanceID⟩}
   Output: A finite set of evaluated tags Result_Test_set = {⟨tag,trust_values⟩}
1  for s in trainingset[stereotypes] do
2      rep[ s] ← build_reputation(Training_set)
3  for s in testset[stereotypes] do
4      for Tag ← tag_1 to tag_n do
5          trust_values[ Tag] ← compute_tv(Training_set)
6      s_tags ← sort_tags(trust_values)
7      Result ← assess(s_tags,rep[s])
8  return Result
```

In a previous work, Ceolin et al. [51] estimated the trustworthiness of the tags in a test set by weighing them based on semantic similarity with all the tags by the same user from

the training set. Another work, Ceolin et al. [52], demonstrated how the user expertise in a specific topic can be estimated from evidence from semantically close areas. Here we follow a similar approach, but we differ in that the works mentioned evaluate the annotations on a user basis, while we use provenance stereotypes instead.

Once the training set has been built, we evaluate the trustworthiness of the annotations in the test set for each group. We compare each annotation to be evaluated against each piece of evidence in the training set, and we use the semantic similarity emerging from that comparison to weigh the evidence and compute an opinion per annotation.

Once we have obtained one trust value per tag, we have to decide whether or not to accept the tag itself. To be more precise, for each tag we compute an entire opinion, representing the probabilities for each tag to be correctly evaluated with one of the possible evaluations. Now we must decide which evaluation to assign to the annotation. One strategy would use, for each annotation, the evaluation having the higher probability. We do not adopt this strategy because by doing so we will most likely tend to evaluate all tags of a given stereotype with the same dominant evaluation. For instance, if 95% of the training set annotations of one stereotype are useful, we will most likely evaluate all its annotations in the test set as useful. In turn, this implies that we do not take into account that we estimated that 5% of the annotations are not useful.

So we use an approach that combines the stereotype reputation with the trust values of the annotations, because we want to take fully into account the probabilities that are estimated by means of the reputation, and trust values estimate the trustworthiness of annotations.

Algorithm 3 presents the algorithm for annotation evaluation. First, it provides a procedure for computing the reputation of provenance stereotypes that takes as input a training set composed of tuples formed by tags, their evaluation and the identifier of the provenance stereotype they belong to. This procedure returns a set of pairs consisting of provenance stereotype identifiers and their reputation. Then the algorithm evaluates the new annotations, i.e. the annotations in the test set. This second procedure takes as input the training set (formed by tuples composed of tags, their evaluations and the identifier of the provenance stereotype they belong to) and the test set (formed by tags and their provenance stereotype identifier) and outputs a series of pairs consisting of the list of tags in the test set and the corresponding predicted evaluations.

To continue with the running example, suppose that Alex created his tag ("Chinese") on Monday at 13.00. Suppose, further, that in the cluster Monday-afternoon already the tags {Japanese, Christian} have been evaluated as useful, while {rose} has been evaluated as not useful. Now the trust value of the tag Chinese is evaluated as before, with as only difference that the evaluation is made on the basis of the provenance group it belongs to, and not of the author:

$$p^{m}_{Chinese} = sim(\text{Chinese, Japanese}) + sim(\text{Chinese, Christian}) = 0.9 + 0.63 = 1.53$$

$$n^{m}_{Chinese} = sim(\text{Chinese, rose}) = 0.57$$

$$\omega^{m}_{Chinese}\left(\frac{1.53}{1.53+0.57+2}, \frac{0.9}{1.53+0.57+2}, \frac{2}{1.53+0.57+2}, \frac{1}{2}\right)$$

$$E(\omega^{m}_{Chinese}) = 0.62$$

The reputation of the cluster is:

$$\omega^{m}_{Cluster}\left(\frac{2}{5},\frac{1}{5},\frac{2}{5},\frac{1}{2}\right)$$

$$E(\omega^{m}_{Cluster}) = 0.6$$

So the tag inserted by Alex will be accepted only if it is one of the 60% best tags belonging to that cluster.

Implementation

The code for the representation and assessment of the annotations with the Open Annotation model has been developed using SWI-Prolog Semantic Web Library [53] and the Python libraries rdflib [54] and hcluster [55], and is available on the Web [56].

Results and discussion

We evaluated the algorithms that we proposed by running them on Steve.Museum and SEALINC Media experiment datasets. As described before, we split each dataset into a training and a test set, learn a model based on the training set, and evaluate it on the test set. There is a tradeoff between complexity and performance. On the one hand, a larger training set in general produces a more accurate model. On the other hand, an increased size of the training set induces a larger number of comparisons for each estimate, and hence an increased computation cost. To determine an optimal size for the training set in each case study, we have run the algorithm with different training set sizes, expressed in terms of annotations per user reputation, and tracked their performance.

Some errors can be due to intrinsic limitations of the experiment rather than imprecision of the algorithms. For instance, since training and test set are part of the same dataset, a larger training set means a smaller test set, and vice versa. Since our prediction is probabilistic, a small training set forces us to discretize our predictions, and this increases our error rate. Also, while an increase of the number of annotations used for building a reputation produces an increase of the reliability of the reputation itself, such an increase has the downside to reduce our test set size, since often only few annotators produce a large number of annotations. Nonetheless, we are bound to this limitation because we can only rely on learning reputations and trust values from museum evaluations since we do not have any possibility to decide if the internal inconsistency of the tags regarding a given image implies low trustworthiness of one or more of them.

Both the Steve.Museum dataset and the SEALINC Media dataset present an unbalanced distribution of tags, since about 76% of tags is evaluated as “useful” in the first, and 74% of the tags is evaluated as “4” or “5” in the second. In other words, in each of the datasets, about three quarters of the tags are positively evaluated. So, in principle, if we predict that all the tags are correct, then our accuracy would be 74% and 76% respectively, but that can hardly happen. In fact, our algorithm is made in such a manner that, even if an annotator has a very high reputation (e.g., 95%), still we do not accept all his tags, rather we accept only the 95% of them. New tags are all classified as trustworthy only if the user reputation is 100% or if it is very high (e.g., 99%) and because of discretization, the amount of untrustworthy tags is so small (e.g., 2%) that it is neglected. So, it may happen that all the tags provided by a given user are predicted to be trustworthy, but since users are treated as “silos”, i.e. they are evaluated independently of each other in our system, then this means that there are other users in the dataset for which some tags are predicted

to be untrustworthy, so to justify an overall percentage of trustworthy tags of 76% or 74%. Another important fact is that we cannot evaluate our system on a test set that is artificially balanced in terms of amount of positive and negative evidence. Indeed the basic assumption of our system is that the annotator reputation is representative enough of his performance. So, if a user has 80% reputation, our system will accept about 80% of his new tags. If we build the test set so that it is balanced, then our system will not be able to properly classify all the tags. Instead, we prefer to work with real data, so to be able to test if the annotator reputation is really representative of his performance. Since all the users in our system have high reputation, then necessarily our test set is unbalanced. Lastly, we must add that, since our system hardly evaluates all the tags as trustworthy, if the system was not able to predict at least some of the real trustworthy tags as trustworthy and some untrustworthy tags as untrustworthy, then the accuracy of the system would be higher than 74% or 76%. The fact that this is not the case, as we will see in the remainder of this section, testifies the effectiveness of the algorithms proposed.

Estimation of annotation trustworthiness based on user reputation - Algorithm 1

First, we evaluated the performance of algorithm 1. The results of SEALINC Media experiment are reported in Table 1, where correct tags are considered as a target to be retrieved, so that we can compute metrics such as precision, recall and F-measure. This first case study provided us interesting insights about the model that we propose. The evaluation shows positive results, with an accuracy higher than 80% and a recall higher than 85%.

Then, we applied the same evaluation over the Steve.Museum dataset and we reported the results obtained in Table 2, using the same metrics as before (that is, precision, recall, accuracy and F-measure). Here the performance is less favorable than for the first case study (accuracy around 70% and precision around 80%). This is possibly due to the different size of the Steve.Museum dataset, which may make it more varied than the SEALINC Media dataset. Moreover, the basic assumption of our algorithm is the existence of a correlation between the user identity and his trustworthiness. This might not always be the case, or the correlation might not have always the same strength (e.g. a good user in some situations might not annotate accurately). Also, we aim at learning the museum policies for trusting annotations, but these are not always easy to learn. Lastly, the decrease of accuracy with respect to the previous case is possibly due to the different tag distribution (of positives and negatives) of the dataset and different domains. Different distributions can make it harder to discriminate between trustworthy and untrustworthy tags (as one may encounter mostly one type of observations). Different domains can lead to a different variability of the topics of the tags and this fact affects the reliability of clusters computed on a semantic basis (since clusters will tend to contain less uniform tags, and medoids will be, on average, less representative of their corresponding clusters), and consequently affects the accuracy of the algorithm.

It is important to stress that, on the one hand, the increase of the size of the training set brings an improvement of the performance, while on the other hand, performance is already satisfactory with a small training set (five observations per user). Also, this improvement is small. This is important because: (1) the sole parameter that we did not set (i.e. size of the training set) does not seriously affect our results; and (2) when the size of the training set is small, the performance is relatively high, so the need of manual evaluation is reduced. The results are satisfactory even with a small training set, also thanks

to the smoothing factor of subjective logic that allows us to compensate for the possibly limited representativity (with respect to the population) of a distribution estimated from a small sample.

Improving computational efficiency of the estimation of annotation trustworthiness - Algorithm 2

We evaluated the performance of Algorithm 2 on both datasets. Table 3 and Table 4 report the results for the SEALINC Media and the Steve.Museum datasets, respectively. Algorithm 2 is a variant of Algorithm 1 as it attempts to improve the computational efficiency of the first, while trying not to compromise its performance. We ran our evaluation with the same setting as before, with the same training set sizes. Moreover, in one case (Table 3) we also ran the algorithm with two different values for the "cut" parameter, to check its influence on the overall performance.

By comparing Table 3 with Table 1 we can see how the performance of Algorithm 1 is kept, and in some cases even improved, while the execution time is significantly reduced. The same holds for the Steve.Museum case, as we can see by comparing Table 4 and Table 2. Here, in a few limited cases the performance degrades, but in a negligible manner, and the computational time saving is even more evident than in the SEALINC Media case. The "cut" parameter, apparently, does not affect the performance much.

These considerations make us conclude that, at least in these case studies, it is worth clustering the training set on a semantic similarity basis, as this leads to a better computational efficiency, without compromising the performance in terms of precision, accuracy and recall.

Estimation of annotation trustworthiness based on provenance stereotypes - Algorithm 3

We evaluated the performance of Algorithm 3 on both datasets. Table 5 and Table 6 present the results for the SEALINC Media and the Steve.Museum datasets. We ran this evaluation with the same setting as before. Since we were interested only in checking whether the trustworthiness estimations based on provenance stereotypes perform as well as those based on user reputations in terms of precision and recall, we do not report the execution time of the algorithm.

Table 3 Results of the evaluation of Algorithm 2 over the SEALINC Media dataset

# Tags per reputation	% Training set covered	Accuracy	Precision	Recall	F-measure	Time (sec.)
clustered results (cut = 0.6)						
5	8%	0.73	0.88	0.81	0.84	43
10	19%	0.82	0.87	0.93	0.90	24
15	31%	0.83	0.87	0.95	0.91	14
20	41%	0.84	0.87	0.96	0.91	18
clustered results (cut = 0.3)						
5	8%	0.78	0.88	0.88	0.88	43
10	19%	0.82	0.87	0.93	0.90	14
15	31%	0.84	0.87	0.95	0.91	16
20	41%	0.84	0.87	0.96	0.92	21

Results of the evaluation of Algorithm 2 over the SEALINC Media dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

Table 4 Results of the evaluation of Algorithm 2 over the Steve.Museum dataset

# Tags per reputation	% Training set covered	Accuracy	Precision	recall	F-measure	Time (sec.)
clustered results (cut = 0.3)						
5	18%	0.71	0.80	0.84	0.82	707
10	27%	0.70	0.79	0.83	0.81	1004
15	33%	0.70	0.79	0.84	0.82	1197
20	39%	0.70	0.79	0.84	0.82	1286
25	43%	0.71	0.79	0.85	0.82	3080
30	47%	0.72	0.79	0.86	0.82	3660

Results of the evaluation of Algorithm 2 over the Steve.Museum dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

By looking at the results we see that the performance is very satisfactory, and that the results achieved with this algorithm outperform those reported in the tables before, obtained with Algorithm 1 and Algorithm 2. In Table 5 precision is about 88% and recall ranges between 73% and 88%. The decrease in accuracy for the training set built with 20 annotations per reputation is plausibly due to the fact that many provenance stereotypes do not have 20 or more annotations available, so these cluster cannot contribute to the overall accuracy measurement, while they did with 5, 10 or 15 annotations per reputation.

Moreover, the amount of evidence needed to make these assessments is low, as demonstrated by the percentage covered by the training set over the dataset. In Table 6 the performance is even higher than in Table 5. First, this is due to the existence of a correlation between the provenance group an annotation belongs to and its trustworthiness. Second, the fact that the provenance stereotypes that we considered for this experiment are 21, which is much less than the number of users, together with the unbalance between useful and non-useful annotations in the Steve.Museum dataset (the first are much more plentiful than the latter) compensates a collateral effect of smoothing. In fact, smoothing helps in allocating some probability to unseen events (for instance, possible future mistakes of good users). So, because of smoothing, we predicted the existence of non-useful annotations for users who actually did not produce them (the dataset contains only relatively few non-useful annotations). Since there are many more users than provenance stereotypes, this error is higher with user-based estimates, where there are many more smoothed probability distributions (one per author), which causes many more annotations to be wrongly evaluated as non-useful. On the other hand, with provenance stereotypes, this error was much more limited, because the corresponding smoothed reputations introduced fewer wrong non-useful evaluations. Still, we will continue employing

Table 5 Results of the evaluation of Algorithm 3 over the SEALINC Media dataset

Annotations in each reputation	Accuracy	% of Dataset Covered by the Training set	Precision	Recall	F-measure
5	0.68	1.69%	0.88	0.73	0.80
10	0.71	3.35%	0.87	0.80	0.83
15	0.78	4.97%	0.88	0.88	0.88
20	0.72	6.45%	0.87	0.80	0.83

Results of the evaluation of Algorithm 3 over the SEALINC Media dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

Table 6 Results of the evaluation of Algorithm 3 over the Steve.Museum dataset

Annotations in each reputation	Accuracy	% of Dataset Covered by the Training set	Precision	Recall	F-measure
5	0.84	0.25%	0.84	0.99	0.90
10	0.84	0.45%	0.84	0.99	0.90
15	0.84	0.66%	0.84	0.99	0.90
20	0.84	0.86%	0.84	0.99	0.90
25	0.84	1.04 %	0.84	0.99	0.90
30	0.84	1.22 %	0.84	0.99	0.90

Results of the evaluation of Algorithm 1 over the Steve.Museum dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset actually covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

smoothing, as these are posterior considerations based on the availability of privileged information about the test set (i.e. its evaluation), and smoothing allows to compensate the lack of this information. On the other hand, the specific Steve.Museum dataset possibly shows a limitation of smoothing.

In the previous section, we hypothesized that the time of creation of an annotation may implicitly affect its trustworthiness and that the users follow approximatively regular patterns in their behaviors. To support these statements, we made the following analyses:

- we computed the average of the user reputations per provenance group. The averages vary from 0.73 to 0.84 in the Steve.Museum case study and from 0.75 to 0.91 in the SEALINC Media case study. Each user that took part in the SEALINC Media experiment, participated only once. Moreover, their contributions are concentrated in the mid part of the weekdays, so we could not make additional checks. In the Steve.Museum dataset, instead, we also run a series of Wilcoxon signed-ranked tests at 95% confidence level (since the data distribution is not always normally distributed, as shown by a Shapiro-Wilk test at 95% confidence level, we prefer not to use a t-student test), and we discovered that:
  - there is no significant difference within user reputations in the morning, afternoon, and night slots respectively across the week. For instance, we took the reputations in the morning slots for Monday, Tuesday, etc. and the Wilcoxon signed-rank test showed no significant difference. The same holds for the afternoon and the night ones;
  - there is a significant difference between the morning and the afternoon slots and the afternoon and night slots. Here we compared the series of reputations per slot across the week;
  - if we compare the averages of the reputations with respect to the days (for instance, considering the three slots of Monday versus the three slots of Tuesday, etc.) we see no significant difference;
  - there is no significant difference between weekends and weekdays.

  The first two points support our hypothesis because they show that actually there are some relevant differences between groups and actually these depend on the time of creation of an annotation. The third and the fourth point show that, at least in this case study, it is not useful to keep track of the day of the week when the annotation was created. On the other hand, the fact that we recorded the day of the week

allowed us to check if there is any difference both among days and between weekend and weekdays, while if we started directly with this latter distinction, we could not have decreased the granularity.

- as we stated in the previous item, the average number of provenance groups a user contribution belongs to is 1 in the SEALINC Media dataset. In the Steve.Museum dataset, instead, the average number of groups a user contributions belongs to is 1.17, variance 0.56. This means that most of the users' contributions belong to one group. So we can say that, approximatively, there exists a one-to-many relation that links the groups with the users: given a group, we can identify a group of users that provide annotations mostly in that group. This means that, when we analyze the annotations that belong to a given group, then we implicitly analyze the annotations produced by a group of users that annotate mostly in that time interval. So the provenance group acts as a proxy to this group of users, and hence, in practice, we analyze the annotations in that group based on the reputations of the users linked to that group. In principle, there may be a high variance among the users belonging to a given provenance group. However, in the case studies analyzed in this paper, this does not happen to be the case, since the variance of the users reputation belonging to a given group is low.
- in the Steve.Museum case study, the variance of the user reputations ranges between 0.12 and 0.15. This shows that, even if the averages of user reputations per group range between 0.73 and 0.84, the reputations are not sparsely distributed. Rather, within provenance groups users tend to be rather homogeneous in terms of reputation. The same holds for the SEALINC Media case study, where the variance of user reputation per provenance group ranges between 0.004 and 0.01;
- the time that we used in our computation is the server time and the fact that, in principle, the annotations are collected worldwide, this might imply that our calculations are misleading. However, since: (1) as shown before, there is a consistent distinction between morning, afternoon and night reputations (which is determined by user performance, and users tend to contribute at fixed times), (2) the amount of tags annotated as "problematic-foreign" is very small (about 1.9%) and (3) the artifact annotated in the case study belong mainly to U.S. cultural heritage institutions, we assume that the annotations are approximatively provided by users in the same time zone or in the neighboring ones.

When grouping the tags based on time, the choice between coarser and finer granularity is not trivial and, in general, affects the uncertainty of the final result. Grouping the tags at a coarser granularity allows easily collecting evidence for a given group and finding a semantic justification for the differences between groups. If we find a difference between morning and afternoon tags, we can easily suppose (and possibly test) that this is due to the influence that different parts of the day have on the user conditions (tired, sleepy, etc.). If we find a difference between tags made at 8.00 a.m. and at 9.00 a.m., we may need additional information to justify semantically the reasons of such differences. On the other hand, a finer granularity may reveal to be useful to avoid to group together heterogeneous tags. All these are generic considerations, and the choice of the best granularity depends on the peculiarities of the single use case evaluated. In our cases, as is evident from the considerations above, we chose a coarser granularity for the hours of

the day and a finer one for the days of the week, because this combination was the most significant and gave us the highest accuracy. Future work will investigate the possibility to automatically determine the best granularity level for this grouping.

Conclusions

We presented an algorithm for automatically evaluating the trustworthiness of user-contributed annotations by using subjective logic and semantic similarity to learn a model from a limited set of annotations evaluated by an institution. Moreover, we introduce two extensions of this algorithm. The first extension makes use of semantic similarity to cluster the set of evaluated annotations at our disposal (training set) and hence improve the computational efficiency of the algorithm. The second extension regards the possibility to adapt the algorithm to use provenance information instead of the user reputation as a basis for the trustworthiness estimations.

We evaluated each algorithm on two different datasets of annotations from the cultural heritage domain. The algorithm based on user reputation satisfactorily allows us to estimate the annotation trustworthiness with an accuracy of about 80% in one case and 70% in the other one. Clustering effectively helps in increasing the efficiency of the first extension, and the use of provenance information actually allow us to compute accurate estimates of annotations trustworthiness.

With the growth of information on the Web and with active contributions from online users, it becomes necessary to devise algorithms to automate the evaluation of the quality of the contributed information. Our methods are been proven to evaluate user contributed tags in cultural heritage domain with relatively high accuracy. We will aim, on the one hand, at reducing even further the need for evaluated annotations to bootstrap our system so to reduce the burden of cultural heritage institutions in this process, and on the other hand, we will investigate methods for further increasing the accuracy of our algorithms and for making effective use of more complex provenance information. This can be vital for the cultural heritage institutions which do not have many resources in terms of labour or finances at their disposal and decide to rely on crowdsourcing platform, as well as for many other institutions in similar situations.

Lastly, one of the research directions we intend to pursue regards the possibility to add analyses about the content of the annotated artifacts. For instance, the output of visual analysis tools in the case of cultural heritage artifacts could help to improve the quality of our estimates by combining the evidential reasoning we adopt with knowledge about the artifacts themselves. Such a direction opens up for applications of our approach in other contexts. In fact, we could apply our algorithm in combination with natural language processing methods in order to obtain tools for automatically reviewing, for instance, wiki articles or restaurant reviews.

Abbreviation

SEALINC Media: Socially enriched access to linked cultural media.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DC contributed by designing the research questions, collecting one dataset, by designing the algorithms and implementing them, and by writing the article. AN contributed by collecting the datasets, by assisting the algorithm design and implementation and by writing the article. WF supervised the research carried, helping in defining the research questions, and proofread the article. All authors read and approved the final manuscript.

Received: 30 October 2013 Accepted: 5 February 2014
Published: 20 May 2014

References

1. von Ahn L, Dabbish L (2004) Labeling images with a computer game In: Proceedings of the 2004 Conference on Human Factors in Computing Systems, CHI '04, Association for Computing Machinery, pp 319–326
2. Sabater J, Sierra C (2005) Review on computational trust and reputation models. Artif Intell Rev 24: 33–60
3. Artz D, Gil Y (2007) A survey of trust in computer science and the Semantic Web. J Web Semantic 5(2): 58–71
4. Golbeck J (2006) Trust on the World Wide Web: a survey. Foundations Trends Web Sci 1(2): 131–197
5. Castelfranchi C, Falcone R (1998) Proceedings of the 4th International Conference on Multi-Agent Systems, ICMAS '98. IEEE Computer Society, pp 72–79
6. Ellis A, Gluckman D, Cooper A, Greg A (2012) Your paintings: a nation's oil paintings go online, tagged by the public In: Proceedings of Museums and the Web 2012, Online
7. US Institute of Museum and Library Service (2013) Steve Social Tagging Project. [Accessed 9 January 2013]
8. Netherlands Institute for Sound and Vision (2012) Waisda? http://waisda.nl. [Accessed 14 August 2012]
9. Rijksmuseum (2013). https://www.rijksmuseum.nl/. [Accessed 23 September 2013]
10. Surowiecki J (2004) The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Anchor
11. Damme CV, Coenen T (2008) Quality metrics for tags of broad Folksonomies In: Proceedings of the 2008 International Conference on Semantic Systems, I-semantics'08. Journal of University Computer Science
12. Medelyan O, Frank E, Witten IH (2009) Human-competitive tagging using automatic keyphrase extraction In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP '09. Association for Computational Linguistics, pp 1318–1327
13. CiteULike (2012) CiteULike. http://www.citeulike.org/ [Accessed 8 December 2012]
14. Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic Grounding of Tag Relatedness in Social Bookmarking Systems In: Proceedings of the 7th International Semantic Web Conference, ISWC2008. Springer
15. Wikimedia Foundation (2013) Wikipedia. http://www.wikipedia.org. [Accessed 4 March 2013]
16. De la Calzada G, Dekhtyar A (2010) On measuring the quality of Wikipedia articles In: Proceedings of the 4th Workshop on Information Credibility, WICOW '10. Association for Computing Machinery, pp 11–18
17. Zeng H, Alhossaini MA, Ding L, Fikes R, McGuinness DL (2006) Computing trust from revision history In: Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, PST2006. Association for Computing Machinery, p 8
18. Wang S, Iwaihara M (2011) Quality evaluation of wikipedia articles through edit history and editor groups In: Proceedings of the 13th Asia-Pacific Web Conference, APWeb'11. Springer-Verlag, pp 188–199
19. Demartini G (2007) Finding experts using Wikipedia In: Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics, FEWS 2007. CEUR-WS.org
20. Javanmardi S, Lopes C, Baldi P (2010) Modeling user reputation in wikis. Stat Anal Data Mining 3(2): 126–139
21. Ceolin D, Groth P, van Hage WR, Nottamkandath A, Fokkink W (2012) Trust evaluation through user reputation and provenance analysis In: Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web, URSW 2012, pp 15–26. CEUR-WS.org
22. Ceolin D, Nottamkandath A, Fokkink W (2012) Automated evaluation of annotators for museum collections using subjective logic In: Proceedings of the 6th IFIP WG 11.11 International Conference on Trust Management, IFIPTM. Springer, pp 232–239
23. Tavakolifard M, H P, Knapskog SJ (2009) Inferring trust based on similarity with TILLIT In: Proceedings of the 3rd IFIP WG 11.11 International Conference on Trust Management, IFIPTM, pp 133–148
24. Cilibrasi R, Vitányi PMB (2006) Automatic meaning discovery using Google In: Kolmogorov Complexity and Applications. Dagstuhl Seminar Proceedings
25. Ushioda A (1996) Hierarchical clustering of words and application to NLP tasks In: Proceedings of the 16th International Conference on Computational Linguistics, COLING. Association for Computational Linguistics, pp 28–41
26. G Begelman PK, Smadja F (2006) Automated tag clustering: improving search and exploration in the tag space In: Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006
27. Hassan-Montero Y, Herrero-Solana V (2006) Improving tag-clouds as visual information retrieval interfaces In: Proceedings of Multidisciplinary Information Sciences and Technologies Conference, INSCIT 2006. Association for Computational Linguistics
28. Golbeck J, Hendler J (2004) Accuracy of metrics for inferring trust and reputation in semantic web-based social networks In: Proceedings of the 14th International Conference Engineering Knowledge in the Age of the Semantic Web, EKAW
29. Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust In: Proceedings of the 13th International World Wide Web Conference, WWW2004. Association for Computing Machinery, pp 403–412
30. Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The Eigentrust algorithm for reputation management in P2P networks In: Proceedings of the 12th International World Wide Web Conference, WWW2003. Association for Computing Machinery, pp 640–651
31. Bizer C, Cyganiak R (2009) Quality-driven information filtering using the WIQA policy framework. J Web Semantics 7: 1–10
32. Hartig O, Zhao J (2009) Using web data provenance for quality assessment In: Proceedings of the 1st International Workshop on the role of Semantic Web in Provenance Management, SWPM 2009. CEUR-WS.org
33. Zaihrayeu I, da Silva PP, McGuinness DL (2005) IWTrust: Improving user trust in answers from the Web In: Proceedings of the 3th International Conference on Trust Management, vol 3477 of *iTrust*. Springer, pp 384–392
34. W3C (2013) PROV-DM: The PROV Data Model. http://www.w3.org/TR/2012/CR-prov-dm-20121211/. [Accessed 16 July 2013]
35. W3C (2013) PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/. [Accessed 16 July 2013]

36. Ebden M, Huynh TD, Moreau L, Ramchurn S, Roberts S (2012) Network analysis on provenance graphs from a crowdsourcing application In: Proceedings of the 4th International Conference on Provenance and Annotation of Data and Processes, IPAW'12. Springer-Verlag, pp 168–182
37. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, Groth P, Kwasnikowska N, Miles S, Missier P, Myers J, Plale B, Simmhan Y, Stephan E, den Bussche JV (2011) The open provenance model core specification (v1.1). Future Generations Comput Syst 27(6): 743–756
38. Altintas I, Anand MK, Crawl D, Bowers S, Belloum A, Missier P, Ludäscher B, Goble CA, Sloot PMA (2010) Understanding collaborative studies through interoperable workflow provenance In: Proceedings of the 2nd International Conference on Provenance and Annotation of Data and Processes, IPAW'10. Springer, pp 42–58
39. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11): 39–41
40. Cilibrasi RL, Vitanyi PMB (2007) The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3): 370–383. http://dx.doi.org/10.1109/TKDE.2007.48.
41. Vossen P, Hofmann K, de Rijke M, Sang ETK, Deschacht K (2007) The Cornetto database: architecture and user-scenarios In: Proceedings of 7th Dutch-Belgian Information Retrieval Workshop, DIR 2007, pp 89–96
42. Wu Z, Palmer M (1994) Verbs semantics and lexical selection In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ACL '94. Association for Computational Linguistics, pp 133–138
43. Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, ICML '98. Morgan Kaufmann Publishers Inc., pp 296–304
44. Ceolin D, Nottamkandath A, Fokkink W (2012) Subjective logic extensions for the semantic web. In: Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web, URSW, pp 27–38. CEUR-WS.org
45. Loper E, Bird S (2002) NLTK: The Natural Language Toolkit. In: ETMTNLP '02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 63–70
46. Marsi E (2013) pyCornetto. https://github.com/emsrc/pycornetto.
47. Leyssen MHR, Traub MC, van Ossenbruggen JR, Hardman L (2012) Is it a bird or is it a crow? The influence of presented tags on image tagging by non- Expert users. CWI Tech. Report INS-1202, CWI
48. Sanderson R, Ciccarese P, de Sompel HV, Clark T, Cole T, Hunter J, Fraistat N (2012) Open annotation core data Model. Tech. rep., W3C Community
49. Jøsang A (2001) A logic for uncertain probabilities. Int J Uncertainty Fuzziness Knowledge-Based Syst 9(3): 279–212
50. Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. J R Stat Soc 18: 54–64
51. Ceolin D, Nottamkandath A, Fokkink W (2013) Semi-automated assessment of annotation trustworthiness In: Proceedings of the 11th Annual Conference on Privacy, Security and Trust, PST2013. IEEE Computer Society
52. Ceolin D, Nottamkandath A, Fokkink W (2012) Automated evaluation of annotators for museum collections using subjective logic In: Proceedings of the 6th IFIP WG 11.11 International Conference on Trust Management, IFIPTM 2012. Springer, pp 232–239
53. SWI-Prolog Semantic WebLibrary (2013). http://www.swi-prolog.org/pldoc/package/semweb.html. [Accessed 10 April 2013]
54. Python libraries rdflib (2013). http://www.rdflib.net/. [Accessed 10 April 2013]
55. Eads D (2008). http://scipy-cluster.googlecode.com/. [Hcluster: Hierarchical Clustering for SciPy]
56. Code published online (2013). http://trustingwebdata.org/JTM2013. [Accessed 23 September 2013]

doi:10.1186/2196-064X-1-3
Cite this article as: Ceolin *et al.*: **Efficient semi-automated assessment of annotations trustworthiness.** *Journal of Trust Management* 2014 **1**:3.