Semi-Automated Assessment of Annotation Trustworthiness

Davide Ceolin
VU University Amsterdam
Email: d.ceolin@vu.nl

Archana Nottamkandath
VU University Amsterdam
Email: a.nottamkandath@vu.nl

Wan Fokkink
VU University Amsterdam
Email: w.j.fokkink@vu.nl

***Abstract*—Cultural heritage institutions and multimedia archives often delegate the task of annotating their collections of artifacts to Web users. The use of crowdsourced annotations from the Web gives rise to trust issues. We propose an algorithm that, by making use of a combination of subjective logic, semantic relatedness measures and clustering, automates the process of evaluation for annotations represented by means of the Open Annotation ontology. The algorithm is evaluated over two different datasets coming from the cultural heritage domain.**

I. Introduction

Through the Web, institutions can reach large masses of people with intentions varying from increasing visibility (and hence, visitors), to acquiring user-generated content. Crowdsourcing has revealed to be an effective way to handle tasks which are highly demanding in terms of the amount of work needed to complete them [1], like, for instance, annotating a large number of 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 Web over time, the amount of data is too large to be accurately dealt with by experts at the disposal of these institutions within a reasonable time. 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.

Employing crowdsourcing triggers trust issues. In fact, these institutions cannot directly check user-contributed content, because checking an annotation is almost as demanding as producing it, and this would be too labour-intensive, and neither can they fully control the user behavior or intentions. Nevertheless a high quality of annotations is vital for their business. The cultural heritage and multimedia institutions need the annotations to be trustworthy in order to maintain their authoritative reputation.

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.

The goal of the work described in this paper is to show how it is possible to automate the process of evaluation of tags obtained through crowdsourcing in an optimized way. 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. On the basis of such a model, the system automatically evaluates the tags further added by the same user. 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. In order to improve the computation time, we cluster the evaluated tags to reduce the number of comparisons, and our experiments show that this preprocessing does not seriously affect the accuracy of the predictions. The proposed algorithms are evaluated on two datasets from the cultural heritage domain. In our experiments we 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%.

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. The sole 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, our experiments show that varying this parameter does not substantially affect the performance (resulting in about 1% precision variation per five new observations considered in a user reputation). Using our algorithms, we show how it is possible to avoid to ask the system administrators 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).

The rest of the paper is structured as follows: Section II describes related work; Section III provides a short definition of terms; Section IV describes the framework that we propose; Section V provides two different case studies where the system has been evaluated. Finally Section VI provides conclusions and future work description.

II. Related work

Trust has been studied extensively in computer science. We refer the reader to the work of 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 the definition of Castelfranchi and Falcone [5], reported by Sabater and Sierra, as discussed in Section III.

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[1] 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 classical wisdom of crowds approaches [9]. In our scenarios we can not make use of those approaches because the level of expertise needed to annotate cultural heritage artifacts restricts the potential set of users involved, thus making this kind of approach inapplicable or less effective. Gamification is another approach that leads to an improvement of the quality of tags gathered from crowds, as shown, for instance, in the work of von Ahn et al. [1]. The work presented here can be considered orthogonal to a gamified environment, as it allows to semi-automatically evaluate the user contributed annotations and hence to semi-automatically incentivize them. In folksonomy systems such as Steve.Museum project, traditional 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, the work of Van Damme et al. [10]) 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. Medeylan et al. [11] present algorithms to determine the quality of tags entered by users in a collaboratively created folksonomy, and apply them to the dataset CiteULike [12], 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 can not apply document-based metrics, and since we do not have at our disposal large amounts of tags per subject, we can not check for consistency among users tagging the same image. Similarly, we can not compute semantic similarity based on the available annotations (like in the work of Cattuto et al. [13]). In open collaborative sites such as Wikipedia [14], where information is contributed by Web users, automated quality evaluation mechanisms have been investigated (see, for instance, the work of De La Calzada et al. [15]). Most of these mechanisms involve computing trust from article revision history and user groups (see the works of Zeng et al. [16] and Wang et al. [17]). 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, we do not have the revision history for the tags.

Another approach to obtain trustworthy data is to find experts amongst Web users with good motivation and intentions (see the work of De Martini et al. [18]). 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 users performance. In our model, we build profiles based on users 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. Modeling of reputation and user behavior on the Web is a widely studied domain. Javanmardi et al. [19] 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. [20] build a reputation and provenance-based model for predicting the trustworthiness of Web users in Waisda? over time. Here we do not make use of provenance (although we aim at doing that in the near future), but we optimize the management reputations and the decision strategies described in that paper.

Here, we use subjective logic to represent user reputation in combination with semantic relatedness measures. This work extends the work of Ceolin et al. [20], [21] and provides a complete framework and two applications of it. Similarity measures have been combined with subjective logic in the work of Tavakolifard et al. [22], who infer new trust connections between entities (users, etc.) 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 the work of Cilibrasi et al. [23] hierarchical clustering has been used for grouping related topics, while Ushioda et al. [24] experiment on clustering words in a hierarchical manner. Begelman et al. [25] 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 [26]. 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 the user reputation, we cluster the tags contributed by the users, 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.

Different cultural heritage institutions have different values and metrics of varying scales to represent the trustworthiness of user contributed information. The accuracy of the various scales has been studied earlier. Certain cases use a binary (boolean) scale for trust values, like in the work of Golbeck et al. [27], while binomial values (i.e. the probabilities of two mutually exclusive values, which range between zero and one, that we use in our work) are used in the work of Guha et al. [28] and Kamvar et al. [29].

III. Definition of Terms

We provide a definition for the most relevant terms used in this paper.

[1] https://www.rijksmuseum.nl/

Trust: We rely on the definition of trust of Castelfranchi and Falcone [5], that is, “the decision that an agent x (trustor) takes to delegate a task to agent y (trustee) is based on a specific set of beliefs and goals, and this mental state is what we call trust.” This general definition underlies many different trust management systems. In our context, the trustor is the museum, while the trustee is a user. The set of goals comprises the need for the museum to maintain an authoritative position and to have its collection tagged, so users are trusted (by time to time) if their tags are believed to be trustworthy enough.

User: A person who provides tags via a crowsourcing application.

Reviewer: A person manually evaluating the crowdsourced tags on behalf of an institution (e.g. museum).

Tag trust value: The evaluation that a given cultural heritage institution assigns to a tag.

Reputation: A value representing the trustworthiness of a given user, based on the evaluation that a cultural heritage institution made of the tags that he (or she) contributed.

IV. System Description

A. High-level 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. Fig. 1 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 Section V, this parameter does not affect significantly the accuracy obtained. Moreover, we do not need to set an “acceptance threshold” (e.g. accept only annotations with a trust value of say at least 0.9, for trust values ranging from zero to one), despite the work of Ceolin et al. [20]. 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.

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.

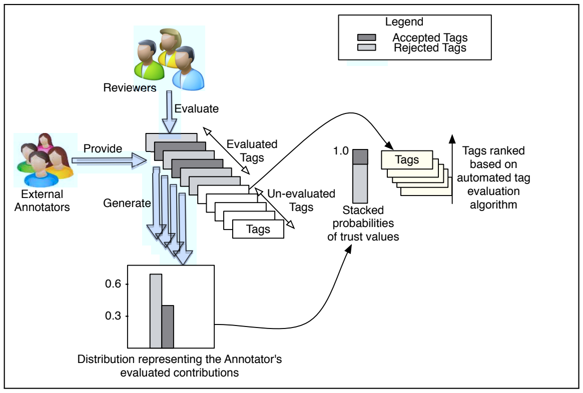
Legend
Accepted Tags
Rejected Tags
Reviewers
Evaluate
Provide
Evaluated Tags
External Annotators
Generate
Un-evaluated Tags
Tags
1.0
Stacked probabilities of trust values
Tags
Tags ranked based on automated tag evaluation algorithm
0.6
0.3
Distribution representing the Annotator's evaluated contributions

Fig. 1: High-level overview of the system.

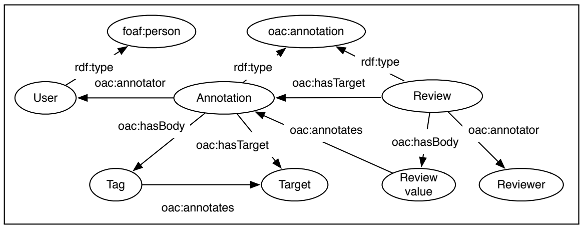

Fig. 2: We represent annotations and their reviews as annotations from the Open Annotation Model.

B. Annotation representation

We adopt the Open Annotation model [30] 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, Fig. 2 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.

```
@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 .
```

C. Trust management

We employ subjective logic [31] 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 possible to what that museum would do, based on a small sample of evaluations produced by the museum itself.

1) Subjective logic: In subjective logic, so-called "subjective opinions" (which are represented as $\omega_y^x(b,d,u,a)$) express the belief that source x owns with respect to the value of assertion y (for instance, a user reputation). When y can assume only two values (e.g. trustworthy and non-trustworthy), the opinion is called "binomial"; when y ranges over more than two values, the opinion is called "multinomial". Opinions are computed as in Equation 1, where the positive and the negative evidence counts are represented as p and n respectively, and b, d, u and a represent the belief, disbelief, uncertainty and prior probability respectively. Such an opinion is equivalent to a Beta probability distribution (see Fig. **??**), which describes the likelihood for each possible value in the $[0 \dots 1]$ interval to be the right trust value for y. An expected probability for a possible value of an opinion is computed as $E(\omega_y^x) = b + a \cdot u$.

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

2) 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. Global, since we do not relate the user reputation 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 Section IV-C4.

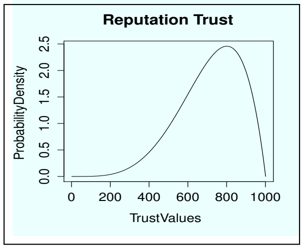

Fig. 3: Beta distribution of user trustworthiness.

We require that a fixed amount of user-contributed tags are evaluated by the museum. Based on those evaluations we compute the user reputation using subjective opinions, as in Equation 2.

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

where m and u represent the museum and the user respectively.

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 2, the algorithm makes use of the expected value of that opinion, as shown in Equation 3

$$E(\omega_u^m) = \frac{p_u^m}{p_u^m + n_u^m + 2} + \frac{1}{2} \cdot \frac{2}{p_u^m + n_u^m + 2} \tag{3}$$

So, to continue with the previous example, suppose that Alex contributed three tags: {Indian, Buddhist} where evaluated as accepted and {tulip} as rejected. His reputation is:

$$\omega_{Alex}^{museum} = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{2} \right) \quad E(\omega_{Alex}^{museum}) = 0.6 \tag{4}$$

3) Semantic relatedness measures: The target of our trust assessments are annotations, and our evidence consists of evaluated annotations. In order to increase the availability of evidence for our estimate and to make 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. We focus on deterministic semantic relatedness measures based on WordNet [32] or on its Dutch counterpart Cornetto [33]. In particular we use the Wu and Palmer [34] and the Lin [35] 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 synsets (sets of word synonyms), and each directed edge from v to w implies that w is a hypernym of v. 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, while the Lin measure considers the information content of the Lowest Common Subsumer and the two compared synsets. For more details about how to combine semantic relatedness measures and subjective logic, see the work of Ceolin et al. [36]. 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. 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.

4) Tag trust value computation and representation: Tag trust values are represented by means of subjective opinions, as in Equation 5.

$$\omega_t^m\left(\frac{p_t^m}{p_t^m+n_t^m+2},\frac{n_t^m}{p_t^m+n_t^m+2},\frac{2}{p_t^m+n_t^m+2},\frac{1}{2}\right) \quad (5)$$

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 are weighed with respect to the similarity to the tag to be evaluated. So p and n are determined as in Equation 6, where sim is a semantic relatedness measure and t is a tag to be evaluated.

$$\begin{aligned} p_t^m &= \Sigma_{t_i \in train}\, sim(t,t_i) \;\; \mathit{if\,evaluation}(t_i) = \mathit{true} \\ n_t^m &= \Sigma_{t_i \in train}\, sim(t,t_i) \;\; \mathit{if\,evaluation}(t_i) = \mathit{false} \end{aligned} \quad (6)$$

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

$$p_{Chinese}^m = sim(\text{Chinese}, \text{Indian}) + \\ + sim(\text{Chinese}, \text{Buddhist}) = 1.05$$

$$n_{Chinese}^m = sim(Chinese, \text{tulip}) = 0.1$$

$$\omega_{Chinese}^m\left(\frac{1.05}{1.05+0.1+2},\frac{0.1}{1.05+0.1+2},\frac{2}{1.05+0.1+2},\frac{1}{2}\right)$$

$$E(\omega_{Chinese}^m) = 0.95$$

5) 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 7). 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,

$$t_1 \leq t_2 \iff E(\omega_{t_1}^m) \leq E(\omega_{t_2}^m) \quad (7)$$

Recall that $E(\omega_u^m)$ is the user reputation, that is the expected percentage of correct tags created by the user. Hence, we accept the last $E(\omega_u^m) \cdot |tags|$ tags in $\{t\}_1^{|tags|}$ (see Equation 8) (as $\{t\}_1^{|tags|}$ is in increasing order).

$$evaluation(tag) = \begin{cases} rejected & \text{if } t \in \{t\}^1_{E(\omega_u^m)\cdot|tags|} \\ accepted & \text{otherwise} \end{cases} \quad (8)$$

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.*

D. Algorithm

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

Algorithm 1: Algorithm to compute trust values of tags

```
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
       $Result\_Test\_set = \{\langle tag, trust\_values\rangle\}$
for $UserID \leftarrow UserID_1$ to $UserID_n$ do
    ▷ for all tags in Training_set
    $rep[UserID] \leftarrow build\_reputation(Training\_set)$
for $UserID \leftarrow UserID_1$ to $UserID_n$ do
    ▷ for all users in Test_set
    for $Tag \leftarrow tag_1$ to $tag_n$ do
        ▷ for all tags in Test_set
        $trust\_values[Tag] = comp\_tv(Training\_set)$
    $s\_tags \leftarrow sort\_tags(trust\_values)$
    $Result \leftarrow assess(s\_tags, rep[UserID])$
return Result
```

1) build_user_reputation: Builds a reputation for each user in the training set, following Equation 2. A reputation is represented as a vector of probabilities for possible tag evaluations.

2) trust_values: Trust values are represented as vectors of probabilities of possible tag evaluations, following Equation 5.

3) comp_tv: Implements Equation 5.

4) sort_tags: The tags are sorted according to their trust value, following the ordering function in Equation 7.

5) assess: The assess function assigns an evaluation to the tag, by implementing Equation 8.

E. 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 per user, and 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 record the evidence counts. The clustering is performed on 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 [37] 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 $1 - \mathit{similarity}(\mathit{word}_1, \mathit{word}_2)$), and this is one of the algorithms that requires such kind of input; (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 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.

To continue with the previous 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}.

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

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

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

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

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.

Algorithm 2: Algorithm to compute trust values of tags, with clustering of the evaluated tags.

```
Input: A finite set of elements in
       $\mathit{Training\_set} = \{\langle \mathit{tag}, \mathit{evaluation}, \mathit{UserID} \rangle\}$
       and $\mathit{Test\_set} = \{\langle \mathit{tag}, \mathit{UserID} \rangle\}$
Output: A finite set of evaluated tags
       $\mathit{Result\_Test\_set} = \{\langle \mathit{tag}, \mathit{trust\_values} \rangle\}$
for $\mathit{UserID} \leftarrow \mathit{UserID}_1$ to $\mathit{UserID}_n$ do
    ▷ for all tags in Training_set
    $\mathit{rep}[\mathit{UserID}] \leftarrow \mathit{build\_reputation}(\mathit{training\_set})$
    $\mathit{clusters}[\mathit{UserID}] \leftarrow \mathit{build\_clust}(\mathit{training\_set})$
    $\mathit{medoids}[\mathit{UserID}] \leftarrow \mathit{get\_med}(\mathit{clusters}, \mathit{UserID})$
for $\mathit{UserID} \leftarrow \mathit{UserID}_1$ to $\mathit{UserID}_n$ do
    ▷ for all users in Test_set
    for $\mathit{Tag} \leftarrow \mathit{tag}_1$ to $\mathit{tag}_n$ do
        ▷ for all tags in Test_set
        $\mathit{trust\_values}[\mathit{Tag}] =$
        $\mathit{comp\_tv}(\mathit{medoids}[\mathit{UserID}], \mathit{rep}[\mathit{UserID}])$
    $\mathit{sort\_tags} \leftarrow \mathit{sort}(\mathit{trust\_values})$
    $\mathit{Result} \leftarrow \mathit{assess}(\mathit{sort\_tags}, \mathit{rep}[\mathit{UserID}])$
return $\mathit{Result}$
```

F. 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[2] and the Python libraries rdflib[3] and hcluster [38], and is available on the Web.[4]

V. Evaluation

We evaluate our model against two different datasets from cultural heritage crowdsourcing projects.

A. Case study 1: SEALINC Media project experiment

As part of SEALINC Media project, Rijksmuseum 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 [39]. 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. 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 I. The behavior of an annotator is classified as either correct or wrong, based on the positive and negative evidence available. The positive evidence is constituted by the tags classified as category 4 and 5, while the negative evidence comprises annotations from category 1, 2 and 3. We run the previously described algorithm for different numbers of annotations used as a basis for building user reputations, in order to analyze the impact of different sizes of training sets. The results of the experiment are reported in Table I, 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%. Clustering brings a clear reduction of the computation time without compromising accuracy (with two different values for the cut parameters, chosen to split almost evenly the $[0, 1]$ interval). The shape of the dataset and the high variance for measurements of small execution times determine a non-linear pattern in the execution

[2] http://www.swi-prolog.org/pldoc/package/semweb.html

[3] http://www.rdflib.net/

[4] Available at http://www.few.vu.nl/ dceolin/PST2013.zip

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
non-clustered results						
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
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

TABLE I: Performances on the data from the SEALINC Media project experiment.

# tags per reputation	% training set covered	accuracy	precision	recall	F-measure	time (sec.)
non-clustered results						
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
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

TABLE II: Performance on the Steve.Museum project dataset.

times. An important consideration regards the fact that some errors can be due to intrinsic limitations of the experiment rather than the imprecision of the algorithm. For instance, since training and test set are part of the same dataset, the bigger the training set is, the smaller the test set is. 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 only few annotators produced a large number of annotations. 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, and on the other hand, performance are 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.

B. Case study 2: Steve.Museum project dataset

Steve.Museum 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 musea collections, and a subset of the data collected from the crowd was evaluated for trustworthiness. 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 performed, yet. We partition this dataset into a training and a test set, as shown in Table II along with their percentage coverage of the whole dataset, together with the results obtained. This second case study focuses on a larger dataset than the first one. The average accuracy attests around 70%. This shows that our algorithm can be trained to different museum policies, because the accuracy, although lower than before, can still be considered satisfactory. The decrease of the 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 might make it harder to discriminate between trustworthy and non-trustworthy tags (as one might encounter mostly one type of observations). Different domains might lead to a different variability of the topics of the tags and this fact affects the reliability of clusters computed on 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. Moreover, one underlying assumption of the algorithm is the existence of a correlation between an artifact author and its reliability. This correlation, apparently, does not always have the same strength in all domains. However, by clustering the training set per user (in Table II we report the most significant results, with cut equal to 0.3), we almost always halve the computation time, and this gain, together with the relatively satisfactory accuracy, makes us incline to further investigate this method in the future.

VI. Conclusion and Future Work

In this paper we present a framework which helps in partially, but efficiently and accurately automating the process of annotation evaluation in crowdsourced systems. One of the major advantages of our system is that it does not require to set any particular parameter regarding decision strategies, hence the final result does not rely on our ability to choose precise values for such parameters. The only parameter we need to set is the size of the training set used to build user reputations, but we observed that it does not substantially affect our performance, thanks to the smoothing factor introduced by subjective logic: smoothing helps to compensate for the fact that small training sets might diverge substantially from the whole population they are sampled from, and this prevents a decrease of the performance. In addition, the use of semantic relatedness measures as weighing factors for the evidence allows us to make precise estimations. The use of probability distributions to represent reputations allows us to make estimates taking into account that high reputations do not necessarily imply perfect performance by the user. Clustering helps to make the computation affordable, without compromising accuracy.

As future work, we intend to optimize our model in three directions: further reduction of human interaction and of computation time, and improving prediction accuracy. The first aspect will be mainly addressed by allowing the reuse of evaluations made on tags inserted by other annotators. This was not possible at this stage as it requires further optimizations (which we will investigate) to keep the computational effort manageable. To further reduce the computation time, we will investigate other optimization and clustering techniques. Finally, the accuracy of the prediction may be improved by considering different kinds of features. Provenance is one class of information that looks suitable for such an improvement.

Acknowledgment

This research is carried out in the context of the SEALINC Media project[5] in the COMMIT research program[6] in The Netherlands. We would like to thank our colleagues Mieke Leyssen, Myriam Traub and Jacco van Ossenbruggen from CWI in Amsterdam and Valentina Maccatrozzo from VU University Amsterdam for their useful comments.

References

[1] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *CHI '04*. ACM, 2004, pp. 319–326.

[2] J. Sabater and C. Sierra, "Review on computational trust and reputation models," *Artificial Intelligence Review*, vol. 24, pp. 33–60, 2005.

[3] D. Artz and Y. Gil, "A survey of trust in computer science and the semantic web," *Web Semant.*, vol. 5, no. 2, pp. 58–71, Jun. 2007.

[4] J. Golbeck, "Trust on the World Wide Web: A Survey," *Foundations and Trends in Web Science*, vol. 1, no. 2, pp. 131–197, 2006.

[5] C. Castelfranchi and R. Falcone, "Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification," in *ICMAS '98*. IEEE Computer Society, 1998, pp. 72–79.

[6] A. Ellis, D. Gluckman, A. Cooper, and A. Greg, "Your paintings: A nation's oil paintings go online, tagged by the public," in *Museums and the Web 2012*. Online, 2012.

[7] U.S. Inst. of Museum and Library Service, "Steve social tagging project," Dec. 2012.

[8] NL. Inst. Sound and Vision, "Waisda?" http://wasida.nl, Aug. 2012.

[9] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Anchor, 2004.

[10] C. V. Damme and T. Coenen, "Quality metrics for tags of broad folksonomies," in *I-semantics'08*. J. of Univ. Comp. Sci., 2008.

[11] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *EMNLP '09*. ACL, 2009, pp. 1318–1327.

[12] CiteULike, "CiteULike," http://www.citeulike.org/, Dec. 2012.

[13] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic Grounding of Tag Relatedness in Social Bookmarking Systems," in *ISWC*. Springer, 2008.

[14] Wikimedia Fndt., "Wikipedia," http://www.wikipedia.org, Mar. 2013.

[15] G. De la Calzada and A. Dekhtyar, "On measuring the quality of Wikipedia articles," in *WICOW '10*. ACM, 2010, pp. 11–18.

[16] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness, "Computing trust from revision history," in *PST '06*. ACM, 2006, p. 8.

[17] S. Wang and M. Iwaihara, "Quality evaluation of wikipedia articles through edit history and editor groups," in *APWeb'11*. Springer-Verlag, 2011, pp. 188–199.

[18] G. Demartini, "Finding experts using wikipedia," in *FEWS 2007*. CEUR-WS.org, November 2007.

[19] S. Javanmardi, C. Lopes, and P. Baldi, "Modeling user reputation in wikis," *Stat. An. Data Min.*, vol. 3, no. 2, pp. 126–139, Apr. 2010.

[20] D. Ceolin, P. Groth, W. van Hage, A. Nottamkandath, and W. Fokkink, "Trust Evaluation through User Reputation and Provenance Analysis," *URSW 2012*, pp. 15–26, 2012.

[21] D. Ceolin, A. Nottamkandath, and W. Fokkink, "Automated Evaluation of Annotators for Museum Collections using Subjective Logic," in *IFIPTM*. Springer, May 2012, pp. 232–239.

[22] M. Tavakolifard, P. Herrmann, and S. J. Knapskog, "Inferring trust based on similarity with tillit," in *IFIPTM*, 2009, pp. 133–148.

[23] R. Cilibrasi and P. Vitanyi, "Automatic Meaning Discovery Using Google," in *Manuscript, CWI, 2004*, Mar. 2004.

[24] A. Ushioda, "Hierarchical Clustering of Words and Application to NLP Tasks," in *COLING*. ACL, Jul. 1996, pp. 28–41.

[25] G. Begelman, "Automated Tag Clustering: Improving search and exploration in the tag space," in *Coll. Web Tagging*, May 2006.

[26] Y. Hassan-Montero and V. Herrero-Solana, "Improving tag-clouds as visual information retrieval interfaces," in *INSCIT*. ACL, Oct. 2006.

[27] J. Golbeck and J. Hendler, "Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks," in *EKAW*, 2004.

[28] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *WWW '04*. ACM, 2004, pp. 403–412.

[29] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in *WWW '03*. ACM, 2003, pp. 640–651.

[30] R. Sanderson, P. Ciccarese, H. V. de Sompel, T. Clark, T. Cole, J. Hunter, and N. Fraistat, "Open annotation core data model," W3C Community, Tech. Rep., May 9 2012.

[31] A. Jøsang, "A Logic for Uncertain Probabilities," *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 3, pp. 279–212, 2001.

[32] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[33] P. Vossen, K. Hofmann, M. de Rijke, E. T. K. Sang, and K. Deschacht, "The Cornetto database: Architecture and user-scenarios," in *DIR 2007*, 2007, pp. 89–96.

[34] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *ACL '94*. ACL, 1994, pp. 133–138.

[35] D. Lin, "An information-theoretic definition of similarity," in *ICML '98*. Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[36] D. Ceolin, A. Nottamkandath, and W. Fokkink, "Subjective Logic Extensions for the Semantic Web," *URSW 2012*, pp. 27–38, 2012.

[37] Gower, J.C. and Ross, G.J.S., "Minimum Spanning Trees and Single Linkage Cluster Analysis," *Journ. of the Royal Stat. Society*, vol. 18, no. 1, pp. 54–64, 1969.

[38] D. Eads, 2008, hcluster: Hierarchical Clustering for SciPy. [Online]. Available: http://scipy-cluster.googlecode.com/

[39] M. H. R. Leyssen, M. C. Traub, J. R. van Ossenbruggen, and L. Hardman, "Is It A Bird Or Is It A Crow? The Influence Of Presented Tags On Image Tagging By Non-Expert Users," CWI, CWI Tech. Report INS-1202, December 2012.

[5] http://sealincmedia.wordpress.com

[6] http://www.commit-nl.nl