Automated Evaluation of Annotators for Museum Collections using Subjective Logic

Davide Ceolin, Archana Nottamkandath, and Wan Fokkink
{d.ceolin,a.nottamkandath,w.j.fokkink}@vu.nl

VU University Amsterdam
De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

Abstract. Museums are rapidly digitizing their collections, and face a huge challenge to annotate every digitized artifact in store. Therefore they are opening up their archives for receiving annotations from experts world-wide. This paper presents an architecture for choosing the most eligible set of annotators for a given artifact, based on semantic relatedness measures between the subject matter of the artifact and topics of expertise of the annotators. We also employ mechanisms for evaluating the quality of provided annotations, and constantly manage and update the trust, reputation and expertise information of registered annotators.[1]

1 Introduction

Cultural and heritage preserving organizations such as museums are rapidly digitizing their collections, and at the same time migrating digitized collections to the Web. Thus, there is a growing need for seeking experts world-wide for providing high quality annotations for digitized artifacts. This paper presents an architecture for finding such experts.

Unlike online content collaboration sites such as Wikipedia, museums cannot risk anyone say anything about a particular topic. Annotations should be provided only by trusted sources, and should be validated by museum experts or peers who have sufficient proven expertise in the same topic. We assume that a generic initial classification of the artifact is already available in the form of specific set of tags or keywords by digital curators, (e.g. indicating the period of production or the type of artifact). Most museums use a standard thesaurus (such as Iconclass [11]), which serves as a basis for deriving relations between the various artifacts, and forms a controlled vocabulary for annotations. The selection of experts who can provide an annotation for a certain topic is based on a proper average of an expert's reputation and the "semantic similarity" between the requested topic and the recorded expertise areas of the expert. Thus we select experts who can be trusted to provide quality annotations.

Moreover, we employ mechanisms for evaluating the quality of provided annotations. We constantly manage and update the trust, reputation and expertise information of registered annotators by employing trust algorithms based

[1] This research is carried out as part of a Dutch FES COMMIT project entitled Socially Enriched Access To Linked Cultural Media (SEALINC).

on subjective logic [14], which is a probabilistic logic that takes into account uncertainty and belief ownership, to model and analyze situations that involve incomplete knowledge. This work extends previous work on determining the quality of annotations using (Semantic) Web sources [4] by combining subjective logic with measures of semantic relatedness, thereby providing an extensive model for managing annotations.

2 Related work

Cultural heritage organizations are opening up their archives to external user contributions mainly classified as: (1) *Social tagging*, where users link artifacts with "tags", i.e, words generically related to them (see for instance the "Steve Social Tagging Project" [13] and the "Uncovering Nation's Art Collection" project [1] from BBC); (2) *Collaborative authoring*, mainly in case of encyclopedias (see [8]); (3) *Annotations*, where the requested "tags" specifically describe one aspect of the artifact [4].

Recommender systems are widely employed in media-related systems to provide valuable suggestions to the users (e.g. [10,20]). In collaborative communities such as Wikipedia, the correct allocation of tasks is done by intelligent task routing systems [7]. User participation increased up to four-fold when online tasks were mapped to user interests (see [5]). In cultural heritage organizations, the quality and trustworthiness of contributions play a vital role. There is a considerable amount of research on finding experts (as trustworhty contributors) in online communities such as Wikipedia [2] and online forums [22].

Semantic relatedness is a concept where sets of terms are assigned a metric based on the likeness of their meaning/semantic content. User interests are recorded and various thesauri are employed for deriving semantically related interests in content-based recommender systems for museums (e.g. [19]). Recent standardization efforts, such as SKOS [17], have lowered the technical boundaries to publish thesauri on the Web.

3 Adopted methods and technologies

Semantic Relatedness This is a measure that indicates how closely two concepts relate in a taxonomy, given all existing relations between them. We use a WordNet [15] based similarity relatedness measure. WordNet is a large lexical database of English, often used by museums and similar to the models that they use to categorize and describe their artifacts.

Semantic Web technologies They include a wide range of formats and technologies aimed at enhancing the Semantic Web vision (which may be summarized with the slogan "moving from a Web of documents to a Web of data"). We use some of them, in particular:

- URIs: Uniform Resource Identifiers offer unique references to any possible entity (e.g.: annotators, artifacts, concepts).
- RDF: the Resource Definition Framework is basically a language for representing graphs. RDF statements are "triples" (Subject, Predicate, Object), where each of these elements can be either a URI or a literal value (with some restrictions).
- Ontologies: defined using RDFS/OWL language, ontologies define types, properties, etc., of URIs in particular contexts. For example, they allow to distinguish URIs referring to sets of users from those representing concepts. We use the following ontologies:
  - Friend Of A Friend (**foaf**) [3]: for representing people and connections among them.
  - Simple Knowledge Organization System (**skos**) [17]: for representing "concepts" and semantic relations among them.
  - Hoonoh (**hoonoh**) [9]: for representing expertise.
  - RDF Data Cube (**qb**) [6]: for representing multi-dimensional data.
  - Dublin Core Terms (**dcterms**) [12]: for representing meta-data.
  - PROV (**prov**) [18]: for representing provenance information.

Subjective logic Evidence about the expertise and reliability of annotators is handled by means of a probabilistic logic named subjective logic which represents the estimated truth value of propositions by means of subjective opinions. An opinion $\omega^{object}_{subject}(belief, disbelief, uncertainty, apriori)$ is defined by (1) and (2).

$$belief + disbelief + uncertainty = 1, \qquad apriori \in [0...1] \tag{1}$$

$$belief = \frac{p}{p+n+2} \quad disbelief = \frac{n}{p+n+2} \quad uncertainty = \frac{2}{p+n+2} \tag{2}$$

$$E = belief + apriori \cdot uncertainty \tag{3}$$

p and n are the amount of positive and negative evidence respectively. *apriori* is the prior knowledge owned about the expertise, which does not change over time; its influence on the trust value computation lowers as we collect new evidence. In case of a lack of prior knowledge, the default value for *apriori* is 0.5, which is equally far from zero (*false*) and one (*true*).

The expected value of an opinion (3), which corresponds to the trust value we want to represent, is an expected value in the statistical sense, since it is the expected value of the Dirichlet distribution equivalent to the opinion. The distribution describes the probability of each value between zero and one to be the right trust value and its shape depends on the value of the opinion.

4 Model

Our model aims at obtaining trustworthy annotations through crowdsourcing. It is composed of two parts, strongly interlinked: data representation and algorithm. These two parts are connected by subjective opinions: the first part

provides a representation for the expertise, i.e., the "object" of our opinions, whereas the algorithm computes the trust levels and outputs the most trustworthy annotations.

4.1 Data representation

The expertise of each annotator is recorded, through the hoonoh ontology, by linking the URI representing the user to the one representing the concept of expertise. In RDF statements, it is represented as follows:

```
eg:T1 a hoonoh:Topic, skos:Concept.
eg:user a foaf:Person .
eg:E1 a hoonoh:ExpertiseRelationship;
    hoonoh:from eg:user;
    hoonoh:toTopic eg:T1 .
```

We define a data structure representing a subjective opinion, we link it to the corresponding hoonoh:ExpertiseRelationship and then populate it with observations, i.e., opinion instances:

```
eg:Opinion a qb:DataStructureDefinition;
    qb:component
        [ qb:measure eg:belief; ],
        [ qb:measure eg:disbelief; ],
        [ qb:measure eg:uncertainty; ],
        [ qb:measure eg:apriori; ] .

eg:dataset a qb:DataSet;
    qb:structure eg:Opinion;
    dcterms:subject eg:E1 .

eg:obs1a a qb:Observation, prov:Entity;
    qb:dataSet eg:dataset;
    prov:wasAttributedTo eg:Museum;
    eg:belief 0.4;
    eg:disbelief 0.2;
    eg:uncertainty 0.4;
    eg:apriori 0.5.
```

Museum artifacts are annotated objects of type skos:Concept. E.g.:

```
eg:item1 dcterms:subject eg:T1.
```

4.2 Trust (expertise) management

We are interested in determining the user expertise about a given topic, so, if **eg**:E1 is of type **hoonoh**:ExpertiseRelationship, an opinion is:

$$expertise(user, T1) = \omega^{\text{eg:E1 hoonoh:from eg:user}}_{\text{eg:E1 hoonoh:toTopic eg:T1}}(b, d, u, a) \qquad (4)$$

We assume that users are evaluated (e.g. through a questionnaire) when registered. This evaluation is represented by the *apriori* component, which provides an initial indication of the user expertise. As the user provides candidate values for annotations and these are evaluated, the weight of the *apriori* on the trust value will decrease. When evaluating the expertise of the user about a topic T1, the opinion is computed as in (2) but, before summming them, each piece of evidence is weighed on its semantic similarity with T1.

4.3 Algorithm

We introduce a pseudo-code algorithm that computes the trust levels and outputs the most trustworthy annotations, and we provide a qualitative description of it.

for all *request* **do**
  users ← *select_users*(*request*)
  for all *users* **do**
    result ← *append_value*(*user*, *request*)
  end for
  output ← *evaluate_results*(*result*)
  update_expertise(*users*)
  return *output*
end for

select_users This function selects a set of annotators to whom we forward an input *request*. A *request* should contain:

- A reference to the artifact to be annotated.
- A first, high-level classification of the item, that facilitates the annotators selection (e.g., the century when it was made)
- The requested "facet", necessary to obtain comparable candidate values (e.g., the "what" facet, i.e. the artifact content).

The selection procedure depends on internal policies of the museum deploying the system, so we do not make it explicit. Some examples:

- Select the n highest ranked experts about the requested topic.
- Consider all the experts. Weigh their reputation with regards to the distance from the request. Order and select them.
- Consider also the belief and uncertainty (and impose some conditions on them) when selecting annotators.

append_value Collects the contributions obtained from the selected annotators. *result* is a list of couples like (*value*, *annotators_opinions*).

evaluate_results Aggregates results and takes a decision about them. Subjective logic's cumulative fusion operator is a possible aggregation function. A possible decision strategy is to choose the highest-rated value. A decision strategy has to select a candidate value, while reducing the risk of taking a wrong decision and solving possible controversies, such as when multiple candidate values all share the highest rank.

update_expertise After having evaluated the candidate values for the annotation, annotators will be "rewarded" (if their candidate was selected) or "penalized" (otherwise). In principle, this means adding a positive evidence to the first ones and a negative evidence to the last ones, but once again, this may depend on museum policies.

Output The annotation selected can be directly accepted by the museum, or ranked qualitatively according to its trust level (e.g. "accept" when trust level is higher than 0.9, "review" otherwise), so that appropriate actions are taken.

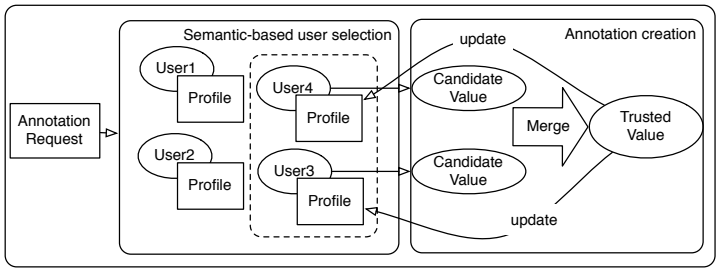

Fig. 1. Algorithm workflow.

5 Evaluation

This section describes some analyses performed on the "Steve Social Tagging Project" [13] dataset, for validating our proposed approach. For this experiment, we computed the semantic relatedness by using the Wu & Palmer measure [21] on WordNet using an online service [16]. This gave us a $measure \in]0..1]$.

The "Steve Social Tagging Project" is a collaboration of museum professionals and others aimed at enhancing social tagging. We used the small portion of the data available as part of a 2006 project entitled "Researching social tagging and folksonomy in the ArtMuseum". This dataset comprises 1784 images from the museums which are open for tagging by the users. Each image is tagged by a single user with multiple tag words. There are 15,167 distinct tags. The 45,859 tag reviews of only 11 users are available as open source. Tag evaluations from the museum (e.g. "useful", "non-useful") are used as evidence in the training part and as a *gold standard* in the prediction evaluation.

A first empirical overview of the dataset hinted at the presence of possible semantic clusters. We then manually selected the candidate set of single words and proved that the semantic relatedness among those words is high. An example of clusters found is available in Fig. 2. After having shown the existence of these semantic clusters, we compared the expertise of people using words from those

clusters and noticed that people having a high amount of positive (or negative) evidence regarding one word in a particular cluster also had a high amount of positive (or negative) evidence about the other words in the same cluster. Positive and negative evidence is derived from the evaluation by the museum: tags evaluated as useful are counted as positive evidence, non-useful as negative. This manual and empirical analysis gave us a first concrete evidence about the relatedness between reputation based on evidence and semantic similarity.

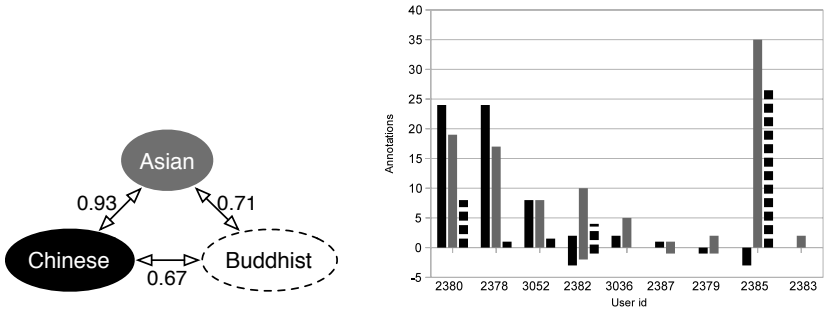

Fig. 2. Cluster and corresponding positive/negative evidence per user.

We also built each user's reputation using a subset of the evaluations made by the museum and, based on this, we predicted the usefulness of future tags inserted by each user[2]. The prediction is performed as described in Section 4.2. Tags having a trust level of at least 0.7 are labelled as "useful". As a side effect of weighing, *uncertainty* of reputations rises, since weighing reduces the amount of evidence considered. However, often, this consequence did not worsen our results, especially when the reputation was already quite high (e.g., the reputation of an annotator reduced to 0.92 from 0.97). On the contrary, our approach allowed us to be prudent in our evaluations, so we could avoid accepting as useful tags with high *uncertainty*. Weighing improved the performance of subjective logic in a statistically significant manner, as proven by applying the sign test with a confidence interval of 95% on the compared errors.

6 Conclusion and Future Work

We demonstrate the potentials of combining subjective logic, semantic relatedness measures and Semantic Web technologies for handling users expertise and annoations trustworthiness.This work is an initial step in a promising direction that will be further explored. For instance, we plan to extensively deploy the described architecture and to extend the range of semantic relatedness measures and vocabularies adopted.

[2] The complete set of analyses is available at: `http://bit.ly/y3uzlP`

References

1. BBC. Uncovering Nation's art collection. http://www.bbc.co.uk/yourpaintings, January 2012.
2. J. G. Breslin, U. Bojars, B. Aleman-meza, H. Boley, L. J. Nixon, A. Polleres, and A. V. Zhdanova. Finding experts using internet-based discussions in online communities and associated social networks. In *FEWS 2007*. CEUR-WS.org, 2007.
3. D. Brickley and L. Miller. Foaf vocabulary specification 0.98. http://xmlns.com/foaf/spec/, January 2012.
4. D. Ceolin, W. Van Hage, and W. Fokkink. A trust model to estimate the quality of annotations using the Web. In *WebSci10*, 2010.
5. D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *IUI '07*, pages 32–41. ACM, 2007.
6. R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube vocabulary. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html, February 2012.
7. G. Demartini. Finding experts using wikipedia. In *FEWS 2007*, pages 33–41. CEUR-WS.org, 2007.
8. W. Emigh and S. C. Herring. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *HICSS'05*, page 99a. IEEE, 2005.
9. T. Heath and E. Motta. The hoonoh ontology for describing trust relationships in information seeking. In *PICKME 2008*. CEUR-WS.org, 2008.
10. L. Hollink, G. Schreiber, and B. Wielinga. Patterns of semantic relations to improve image content search. *Journal Web Semantics*, 5(3):195–203, 2007.
11. IconClass. A multilingual classification system for cultural content. http://www.iconclass.org/, January 2012.
12. Dublin Core Metadata Initiative. Dcmi metadata terms. http://dublincore.org/documents/dcmi-terms, February 2012.
13. U.S institute of Museum and Library Services. Steve Social Tagging Project, January 2012.
14. A. Jøsang. A logic for uncertain probabilities. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.
15. Princeton University. A lexical database for English. http://wordnet.princeton.edu/, January 2012.
16. Princeton University. Wordnet::Similarity. http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi, February 2012.
17. W3C. SKOS Simple Knowledge Organization System Primer. http://www.w3.org/TR/skos-primer/, January 2012.
18. W3C. The PROV Ontology: Model and Formal Semantics. http://http://www.w3.org/TR/prov-o, February 2012.
19. Y. Wang, N. Stash, L. Aroyo, P. Gorgels, L. Rutledge, G. Schreiber, and Rijksmuseum Amsterdam. Recommendations based on semantically-enriched museum collections. *Journal Web Semantics*, 6(4):283–290, 2008.
20. Y. Wang, N. Stash, L. Aroyo, and L. Hollink. Using semantic relations for content-based recommender systems in cultural heritage. In *WOP 2009*, pages 16–28. CEUR-WS.org, 2009.
21. W. Zhibiao and M. Palmer. Verbs semantics and lexical selection. In *ACL '94*, pages 133–138. ACL, 1994.
22. Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. Routing Questions to the Right Users in Online Communities. In *ICDE '09*. IEEE, 2009.