Automated Evaluation of Crowdsourced Annotations in the Cultural Heritage Domain

Archana Nottamkandath[1], Jasper Oosterman[2], Davide Ceolin[1], and Wan Fokkink[1]

[1] VU University Amsterdam, Amsterdam, The Netherlands
{a.nottamkandath,d.ceolin,w.j.fokkink}@vu.nl
[2] Delft University of Technology, Delft, The Netherlands
j.e.g.oosterman@tudelft.nl

Abstract. Cultural heritage institutions are employing crowdsourcing techniques to enrich their collection. However, assessing the quality of crowdsourced annotations is a challenge for these institutions and manually evaluating all annotations is not feasible. We employ Support Vector Machines and feature set selectors to understand which annotator and annotation properties are relevant to the annotation quality. In addition we propose a trust model to build an annotator reputation using subjective logic and assess the relevance of both annotator and annotation properties on the reputation. We applied our models to the Steve.museum dataset and found that a subset of annotation properties can identify useful annotations with a precision of 90%. However, our studied annotator properties were less predictive.

1 Introduction

Cultural heritage institutions have large collections which can be viewed in exhibitions and often are digitised and visible online. For these institutions the metadata of these artefacts (paintings, prints, sculptures etc.) are of the utmost importance. They notably cover the physical properties of the artefact (e.g. dimensions, material), provenance properties (e.g. creator, previous owners) and the subject matter (what is depicted on the artefact). Typically, cultural heritage institutions employ professionals, mostly art historians, who mostly provide high-quality annotations about art-historical properties of artefacts, but tend to lack domain expertise for other aspects such as names of depicted items (of e.g. flowers and birds). With regard to the large scale of collections, their annotation capacity is also limited to describe the subject matter in detail.

Due to these limitations institutions are looking into the knowledge and capacity of crowds. Projects such as Steve.museum [18], Your Paintings [4] and Waisda? [8], are all examples of cultural heritage or media institutions opening up their collection to the crowd for annotation. In these projects institutions engage people from the web in different tasks with the purpose of integrating the

obtained data within their collections. However, employed professional annotators are trained and follow strict guidelines on how to correctly and qualitatively annotate artefacts, to maintain the high quality standards these institutions have. Crowdsourced annotators are not trained in such a way and their quality cannot be guaranteed in a straightforward manner.

Crowdsourced annotations thus need to be assessed, to evaluate whether they meet the institution's quality criteria. However, manually evaluating such a large amount of annotations is likely as expensive as entering the information manually. Thus there is a need to develop algorithms which can automatically or semi-automatically predict the trustworthiness of crowd annotations. The goal of this study is to understand which kinds of properties are important in deciding this trustworthiness, so that in the future suitable annotators can be recruited, or annotation tasks can be tuned in such a way to more likely obtain desired information. The results from this study will thus have implications in the fields of expert finding and task formulation in the domain of crowdsourcing cultural heritage data. In this paper we answer the following research questions:

RQ1: Which annotation properties affect the trustworthiness of crowd-provided annotations?

RQ2: Can an annotator's profile information help in the estimation of annotation and annotator trustworthiness?

In this paper we make use of the Steve.museum dataset [18] containing reviewed annotations on museum objects and information about the annotators such as *age*, *museum and annotation familiarity* and *income*. We propose a trust model for annotator reputation and make prediction models for both annotation usefulness and annotator reputation. The contributions of this paper are: 1) A trust model for reputation based on subjective logic, and 2) insights into the relevance of annotation and annotator properties on the trustworthiness of cultural heritage annotations.

The remainder of the paper is structured as follows. Section 2 compares our work to existing methods. Section 3 describes our methodology and presents the trust model and semantic model. The Steve.museum case study and semantic representation of the data are described in Section 4. Experiments and evaluations are reported in Section 5 and Section 6 provides conclusions of the paper.

2 Related Work

The problem of assessing the trustworthiness of annotations and annotators is not new. There exist several ontologies for representing trust (e.g., those of Golbeck et al. [6] and of Alnemr et al. [1]). While these put emphasis on the social aspects of trust, we are more interested in the trustworthiness of annotations and annotators. Ceolin et al. [2] employed semantic similarity measures, clustering algorithms and subjective logic for the semi-automatic evaluation of annotations in the cultural heritage domain. A probabilistic model, based on a combination of an annotators reputation and the semantic similarity with already labelled

annotations, is used to assess the usefulness of new annotations, achieving 80% correctness. In this paper we take a different approach and employ machine learning algorithms to determine the usefulness of an annotation by using features of both annotator and annotations.

Majority voting [9] is a commonly used method to assess the quality of annotations. However, for domains with a broad vocabulary, such as the cultural heritage domain, this is not optimal. Adapted annotator agreement or disagreement measures have also been studied [11,5], by considering, for example, annotator history and agreement with aggregated label. In contrast, we employ subjective logic to build a user reputation based on his/her positive and negative contributions, and focus more on identifying features about the information and the user that may help to predict his/her trustworthiness.

Task design is also important to achieve qualitative annotations. Test questions or other specialised constructions should be employed to filter out low-quality and spam workers [14] and are necessary to approximate results from experts [15].

Annotation properties have also been studies in the context of Wikipedia [19] and Twitter [17]. Annotation quality has been shown to be related to properties of the annotator. The impact of user information such as *age*, *gender*, *education* and demographics in crowdsourcing tasks have been explored in [13]. They explored the relationship between worker characteristics and their work quality and showed a strong link between them. In this paper we continue in this direction and investigate the relationship between annotation quality and a more extensive set of user properties including *income* and *internet connection speed*.

3 Methodology

In this section we describe the methodology employed in this paper. Our methodology focusses around methods to understand the importance of annotator and annotation properties and is outlined in Algorithm 1. Firstly we identify the features which are relevant for predicting the *value*, in our case the evaluation of the annotation and the reputation of the annotator. Feature identification is done through three different methods: *process analysis*, *extended analysis* and using *feature selection algorithms.* Having identified the sets of features, we perform an independent correlation analysis of each of the identified features with the *value*. We split the dataset into a test and a training set and use the feature sets to predict the *value*. The result of the feature selection methods are then compared.

In Section 3.1 we describe the trust modelling of annotator reputation and in 3.2 we describe the semantic representation of our data model.

3.1 Trust Modelling

The annotation process involves an annotator who is either a user from the crowd or an employee of a cultural heritage institution who provides information about

Algorithm 1: Algorithm to perform predictions based on relevant features

```
Input: A finite set of features F and values used for training
    Input_set ={⟨F,value⟩}
Output: A finite set of relevant features and predicted values
    Output_set ={⟨F_relevant, predicted_value⟩}
F_relevant ← Identify_relevant_features(Input_set)
for F_relevant ← F_relevant_1 to F_relevant_n do
    Compute_correlation(F_relevant, value )
Train_set ← Build_train_set(F_relevant, value)
Test_set ← Build_test_set(F_relevant)
Output_set ← Employ_machine_learning(Train_set, Test_set)
return Output_set
```

digital artefacts. A digital artefact is an image of the actual physical artefact which is published online by the cultural heritage institution. An annotation is information describing some properties of the digital artefact such as what is depicted, who is the artist, etc. A reviewer is a trusted entity, usually an employee of a cultural heritage institution who evaluates the annotation and decides if it is to be accepted or not, based on review policy of the institution.

Aggregating the annotations and their evaluations per annotator helps us understand the reputation of the annotator in the system based on the total number of *useful* and *not useful* annotations. We define reputation of an annotator as a value representing the trustworthiness of a given annotator, based on the evaluation that a cultural heritage institution made of the tags that he or she contributed.

In order to properly model and represent the user expertise and reputation based on the evidence at our disposal, we use a probabilistic logic named subjective logic [12]. It models the truth of propositions as Beta probability distributions that represent both the probability of the proposition to be true (i.e., for instance, the probability of a user to be trustworthy) and the uncertainty about this probability. In subjective logic such a probability distribution is represented by means of the "opinion" (ω) construct. An opinion that a certain *institution* holds with respect to a given *annotator* is represented as follows:

$$\omega_{annotator}^{institution}(belief, disbelief, uncertainty, apriori)$$

where

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

and

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

Here p is the amount of positive evidence (e.g., annotations evaluated as *useful*), n the amount of negative evidence (e.g., annotations evaluated as *not useful*), and *apriori* is the prior knowledge about the reputation, which is set to $\frac{1}{2}$ by

default. The actual value that we use to represent an annotator's reputation is the expected value of the corresponding Beta distribution, that is computed as:

$$E = belief + apriori \cdot uncertainty$$

Subjective logic offers a wide range of operators that allow one to reason upon the evidence at our disposal and infer the reputation based on the different features considered. But we use it merely for a representation purpose. In fact, to apply such operators we would need to know a priori the kind of relations that occur between the features that we identify and the reputation. These relations will instead be discovered by means of a machine learning approach.

We use subjective logic to model both annotator and annotation reputations by means of the expected value E. In the case of the annotators, we collect evidence about them (i.e. reviews of the tags they contributed) and we estimate their reputations by means of the subjective opinions described above. In the case of annotation reputations, we use the expected value E to model them, but their prediction is made by means of the machine learning methods.

3.2 Semantic Modelling

We adopt semantic web technologies for representing the annotations and the related metadata. This is done for two reasons. First, they provide a uniform layer that allow us interoperability and prevents us from relying on the specific structure such as relational databases. Second, they provide a means to possibly share metadata and computation results in such a manner that other institutions could benefit from them, thus promoting the sharing of possibly precious information (precious both because of their specificity and of their quality).

A (crowd) annotator performs an annotation task. The annotator's features (e.g., age, country, education) are as much as possible represented by means of the standard FOAF ontology [3], while the annotation is represented by means of the Open Annotation Model [16].

The annotation entered by the user is reviewed by an employee of the cultural heritage institution. The annotation evaluation is yet again represented by means of the Open Annotation Model, as an annotation of the first annotation. All the features we adopt in our computation that are not representable by means of standard vocabularies are represented by means of an ad-hoc construct ("ex:" prefix). An illustration of the annotation (and related metadata) representation is provided in Figure 1, where it is also indicated that we use annotator and annotation features as a basis for estimating the value of an annotation evaluation.

4 Cultural Heritage Annotations: Steve.museum

The Steve.museum [18] dataset was created by a group of art museums with the aim to explore the role that user-contributed descriptions can play in improving

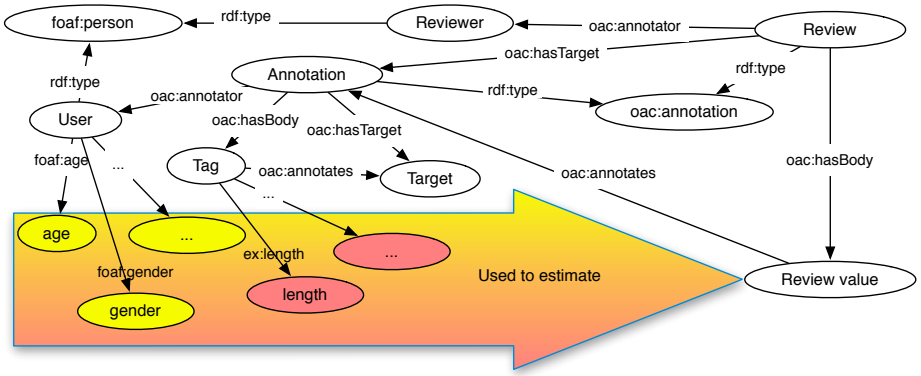

Fig. 1: Representation of an annotation and of the related metadata.

on-line access to works of art. Annotations were gathered for 1,784 artworks and the usefulness, either *useful* or *not useful*, of each annotation was evaluated by professional museum staff. The annotations including their evaluations and annotator information were published as a dataset to study[3].

We performed two pre-processing steps on the data. First, for the correct calculation of the annotator reputation we need at least five annotations per annotator and as such removed data from annotator with fewer annotations. It also occurred that multiple reviewers evaluated the same annotation. For those annotations we took the majority vote of the evaluations. In case of a tie we always chose *useful*, giving more weight to a potentially useful annotation.

The dataset contains both anonymous (730) and registered (488) annotators. Table 1 lists the annotator properties and the percentage of registered annotators who filled in each property. The distribution of the number of annotations per annotator follows a power law. The majority of the annotations (87%) were evaluated as useful. Considering other crowdsourcing initiatives this was a remarkably good crowd. Table 2 provides a summary of the complete dataset.

Table 1: Annotator properties and the percentage of registered annotators who filled in the property.

Community	**Experience**	**Education**	**Age**	**Gender**	**Household income**
431 (88%)	483 (99%)	483 (99%)	480 (98%)	447 (92%)	344 (70%)
Works in a museum	**Involvement level**	**Tagging experience**	**Internet connection**	**Internet usage**	
428 (88%)	411 (84%)	425 (87%)	406 (83%)	432 (89%)	

[3] `http://verne.steve.museum/steve-data-release.zip`

Table 2: Summary of the Steve.museum dataset.

Number of annotators / Registered	1,218 / 488 (40%)
Provided tags	45,733
Unique tags	13,949
Tags evaluated as *useful*	39,931 (87%)
Tags evaluated as *not useful*	5,802 (13%)

5 Evaluation

Annotations in the Steve.museum dataset have been assessed as either *useful* or *not useful*. Each annotator has a *reputation* score using the model described in Section 3.1. Using machine learning techniques, we aim to automatically predict the evaluation of the annotations based on features of annotators and annotations. Next to that we aim to predict the reputation of the annotator based on the annotator features. The first subsection describes the setup and tooling of our experiments. Section 5.2 contains the results of analysing the relation between annotation properties and *usefulness* of annotations and Section 5.3 between annotation properties and both annotation evaluation and user reputation.

5.1 Experimental Setup

In order to perform fair training, we randomly selected 1000 *useful* and 1000 *not useful* annotations as training set. The remainder of the dataset was used as test set. We used a Support Vector Machine (Sequential Minimal Optimisation[4], default PolyKernel[5]) on selected features to predict annotation usefulness, since that algorithm works for dichotomous variables, and is commonly used, fast and resistant against over-fitting. For prediction of the *reputation* of a user (an interval variable) we used a similar algorithm but adapted for regression. For automated selection of relevant features we used correlation-based feature subset selection [7]. This algorithm selects subsets of features that are highly correlated with the prediction class but have a low inter-correlation.

To calculate an independent correlation between different types of variables we used appropriate statistical tests; *Biserial* for interval, ordinal and nominal against dichotomous variables followed by *Wilcoxon rank sum* for ordinal and *Chi squared* for nominal; *Fisher's exact test* for two dichotomous variables; *Kendall* τ for ordinal against interval variables; and *Pearson* for both two interval variables and nominal against interval variables. Fisher's exact test signals a strong correlation above a score of 1.0.

[4] We used the implementation inside the tool WEKA `http://cs.waikato.ac.nz/ml/weka/`.

[5] There are specific kernels targeting RDF data, but these were, for simplicity reasons, not used.

5.2 Predicting Annotation Evaluation Using Annotation Features

Features Selection. We manually analysed the annotations in different evaluation categories of the Steve.museum so as to understand the evaluation policies depicted as ***F_man***. From our observations, we found out that some of the evaluations were strongly influenced by certain features of the annotation. Annotations that did not describe something actually depicted, for example sentimental annotations such as "happy", were evaluated as *not useful*. Adjectives in general were not deemed useful. Also annotations in non-English languages or misspelled words were evaluated as *not useful*. To detect these problems we created the features *is_adjective*, *is_english* and *in_wordnet*, where the latter signals a correctly spelled word. For detecting the language of a tag we used the n-gram based language detection from [10]. For detecting the adjective and spelling errors we used Wordnet,[6] where words not in Wordnet are treated as incorrectly spelled. For multi-words annotations we assessed whether either of the words matched the criteria. We explored the possibilities to extract more features which might be indicative of the evaluation of the annotation represented as ***F_all*** . We regarded the creation time (both day and hour) of the annotation, how specific the annotation was (based on the depth a word occurs at in the Wordnet tree), the length and number of words of the annotation, and the frequency with which the annotation was created for the same object.
We applied the feature selection algorithm to the features from ***F_all*** on the annotation data resulting in the feature set ***F_ml***.
F_man = *[is_adjective, is_english, in_wordnet]*
F_all = ***F_man*** + *[created_day, created_hour, Length, Specificity, nrWords, Frequency]*
F_ml = *[created_day, in_wordnet, Frequency]*

Independent correlation of annotation features. We performed an independent correlation analysis of the mentioned features with regard to the *evaluation* of the annotation. We observed a strong correlation (3.34, using Fisher's exact test) for *in_wordnet*, significant at <0.01. We observed a weak correlation for *Specificity* (-0.11), *Frequency* (0.14), *is_adjective*(0.67, Fisher) and *is_English* (0.94, Fisher, not statistically significant).

Predicting annotation evaluation. Table 3 lists the precision, recall and F-measure of the three feature sets. We observe that the precision is high, ranging from 0.90 to 0.978 in all the cases of classifying *useful* annotations. All three methods for creating the feature sets result in a model that can predict *useful* annotations very well. However, the recall is high only for the feature set ***F_man***, while the predictions using feature sets ***F_all*** and ***F_ml*** had a high number of false positives.

None of the classifiers performed well in predicting the annotations which were classified as *not useful*. There was a large number of false positives and the

[6] We used the NLTK library (`http://nltk.org/`) to query the Wordnet tree.

precision was very low in all cases, ranging from 0.13 to 0.21. Thus from our analysis we can observe that although the machine learning classifier using the three different features were comparably successful in identifying *useful* annotations, neither of them succeeded in identifying the *not useful* annotations.

Table 3: Comparison of results from SVM predictions using annotation features.

Feature set	Class	Precision	Recall	F-measure
F_man	useful	0.90	0.90	0.90
	not useful	0.21	0.20	0.20
F_all	useful	0.91	0.75	0.83
	not useful	0.18	0.42	0.25
F_ml	useful	0.98	0.20	0.34
	not useful	0.13	0.96	0.23

5.3 Predicting Annotation Evaluation And User Reputation Using Annotator Features

Feature Selection. The set ***F_ man*** is based on the annotator properties listed in Table 1. Apart from the provided features for an annotator, we also compute certain features related to the annotations they provided, which may be useful for predicting the *evaluation* of an annotation. The computed features are the total number of annotations entered by the user (*#Annotations*), the vocabulary size and diversity of the annotator, and the number of matched annotations in Wordnet (*#matched_in_wordnet*). The vocabulary size of an annotator is the number of distinct annotations after stemming has been applied. The vocabulary diversity is computed as the vocabulary size divided by the total number of annotations provided by that annotator. The definition of vocabulary diversity is reasonable in view of the fact that the number and length of annotations is relatively small in Steve.museum dataset.
Two sets are obtained when the feature selection algorithm is applied in two instances, one to identify relevant features for the annotation evaluation, represented as ***F_ ml_ a***, and in the second case to identify relevant features for annotator reputation, represented as ***F_ ml_ u***. For the prediction of the annotation evaluation, we merged the annotation data with the corresponding annotator properties and performed a prediction of annotation evaluation. We applied the feature selection algorithm to the features from ***F_ all*** on the annotation data (***F_ ml_ a***) and on the user data (***F_ ml_ u***) resulting in the following features.
F_ man = [Features in Table 1]
F_ all = [**F_ man**, *vocabulary_size*, *vocabulary_diversity*, *is_anonymous*, *#Annotations_in_wordnet*]
F_ ml_ a = [*vocabulary_ size, vocabulary_ diversity*]
F_ ml_ u = [*Language*, *Education*, *Community*, *#tags_wordnet*, *Tagging_experience*]

Independent correlation analysis of annotator features. A statistical correlation analysis was performed to determine the relationship between the annotator features with the annotation reputation and annotation evaluation as shown in Table 4. For the annotation evaluation, Experience, Education, Tagging Experience, Internet connection and Internet usage had a weak correlation that was statistically significant. However, Community had a higher correlation compared to the other features. For the annotator reputation, the computed features such as *# Annotations, vocabulary size* and *#Annotations in Wordnet* were considered significant.

Table 4: Correlation of features with annotation evaluation and annotation reputation. In brackets the statistical test (See Section 5.1). * indicates significance at $p < 0.01$. Note: Fisher signals a high correlation for values > 1.

Annotator feature	**Correlation score Annotation evaluation**	**Correlation score Annotator reputation**
Community	0.22* (C+B)	0.22 (P)
Experience	0.02* (W+B)	0.02 (K)
Education	0.02* (W+B)	0.01 (K)
Age	0.01 (B)	-0.16 (P)
Gender	1.11 (F)	-0.004 (B)
Household income	-0.14 (W+B)	-0.14 (K)
Works in a museum	0.99 (F)	-0.34 (B)
Involvement level	0.04* (W+B)	-0.10 (K)
Tagging experience	1.22* (F)	-0.08 (B)
Internet connection	0.02* (W)	0.06 (K)
Internet usage	0.02* (W)	-0.16 (K)
# Annotations	-0.06 (B)	0.27* (P)
Vocabulary size	-0.06 (B)	0.27* (P)
Vocabulary diversity	0.05 (B)	-0.03 (P)
# Annotations in Wordnet	-0.08 (B)	0.31* (P)

Predicting annotation evaluation and annotator reputation. From Table 5 we can see that the features identified from the annotator profile and those identified by the feature selection algorithm are useful in classifying *useful* annotations and have a high precision of 0.91. However, these methods also have lower values of recall, indicating a high number of false negatives. Both methods have a low precision and recall in classifying *not useful* annotations, and thus are not successful in predicting *not useful* annotations.

We used a SVM for regression to estimate the reputation of the annotator since it was hard to perform a classification for reputation. This is because the reputation is highly right skewed with 90% of the annotators having a reputation > 0.7. This makes it hard to classify data and distinguish the classes when the distribution is highly skewed. Another point is that classification of reputation is highly use case dependent. Upon performing regression on the reputation, as shown in Table 6, we can observe that all the predictions have a very high

relative absolute error and low coefficients. Another observation is that relative weights assigned to the *#Annotations in Wordnet* feature are relatively high, showing consistency with our earlier analysis.

Table 5: Comparison of results from SVM predictions using annotator features.

Feature set	**Class**	**Precision**	**Recall**	**F-measure**
F_man	useful	0.90	0.29	0.44
	not useful	0.11	0.73	0.20
F_all	useful	0.91	0.69	0.78
	not useful	0.15	0.43	0.22
F_ml_a	useful	0.91	0.55	0.68
	not useful	0.13	0.53	0.21

Table 6: Comparison of results from predicting annotator reputation using SVM regression and 10-fold cross validation.

Feature set	**corr**	**Mean abs error**	**Root mean sq error**	**Rel abs error**
F_man	−0.02	0.10	0.15	97.8%
F_all	0.22	0.09	0.13	95.1%
F_ml_u	0.29	0.09	0.13	90.4%

6 Conclusion and Future Work

In this paper we described methods which can automatically evaluate annotations. The experiment was performed on the Steve.museum dataset and investigated the effect of annotation and annotator properties in predicting trustworthiness of annotations and reputation of annotator. We also devised a model using Support Vector Machines for predicting annotation evaluation and annotator reputation. Presence of an annotation in Wordnet is shown to be indicative for the perceived usefulness of that annotation. With a small set of features we were able to predict 98% of the *useful* and 13% of the *not useful* annotations correctly. The annotator reputation was computed using a model in subjective logic. Since the reputation of annotators is highly skewed in this dataset(with more than 90% having a reputation > 0.7), we could not make successful estimations of reputation from annotator profiles.

As part of future work, we would like to repeat the experiment on other cultural heritage datasets. We would also like to build a reputation for an annotator based on topics of expertise, to obtain more accurate correlations between the semantics of the annotation and the topical reputation of the annotator. Our analysis also indicated that there is relevance in aspects related to creation time of an annotation. A more sophisticated model, such as whether an annotation was created during work or during free-time might increase the predictive power.

Acknowledgements This publication was supported by Data2Semantics and SEALINCMedia projects from the Dutch National program COMMIT.

References

1. Alnemr, R., Paschke, A., Meinel, C.: Enabling reputation interoperability through semantic technologies. In: I-SEMANTICS. pp. 1–9. ACM (2010)
2. Ceolin, D., Nottamkandath, A., Fokkink, W.: Efficient semi-automated assessment of annotation trustworthiness. Journal of Trust Management 1, 1–31 (2014)
3. Dan Brickley, L.M.: FOAF. `http://xmlns.com/foaf/spec/` (Jan 2014)
4. Ellis, A., Gluckman, D., Cooper, A., Greg, A.: Your paintings: A nation's oil paintings go online, tagged by the public. In: Museums and the Web 2012. Online (2012)
5. Georgescu, M., Zhu, X.: Aggregation of crowdsourced labels based on worker history. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). pp. 37:1–37:11. WIMS '14, ACM (2014)
6. Golbeck, J., Parsia, B., Hendler, J.A.: Trust networks on the semantic web. In: CIA. pp. 238–249. Springer (2003)
7. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1998)
8. Hildebrand, M., Brinkerink, M., Gligorov, R., van Steenbergen, M., Huijkman, J., Oomen, J.: Waisda?: Video labeling game. In: Proceedings of the 21st ACM International Conference on Multimedia. pp. 823–826. MM '13, ACM (2013)
9. Hirth, M., Hossfeld, T., Tran-Gia, P.: Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on. pp. 316–321 (June 2011)
10. Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., Feinerer, I.: The textcat package for n-gram based text categorization in R. Journal of Statistical Software 52(6), 1–17 (2013)
11. Inel, O., Aroyo, L., Welty, C., Sips, R.J.: Domain-independent quality measures for crowd truth disagreement. Journal of Detection, Representation, and Exploitation of Events in the Semantic Web pp. 2–13 (2013)
12. Jøsang, A.: A logic for uncertain probabilities. Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–212 (2001)
13. Kazai, G., Kamps, J., Milic-Frayling, N.: The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. pp. 2583–2586. CIKM '12, ACM (2012)
14. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. pp. 453–456. CHI '08, ACM (2008)
15. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. pp. 557–566. MIR '10, ACM (2010)
16. Sanderson, R., Ciccarese, P., de Sompel, H.V., Clark, T., Cole, T., Hunter, J., Fraistat, N.: Open annotation core data model. Tech. rep., W3C Community (May 9 2012)
17. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on. pp. 177–184. IEEE (Aug 2010)
18. Trant, J.: Tagging, folksonomy and art museums: Early experiments and ongoing research. J. Digit. Inf. 10(1) (2009)
19. Warncke-Wang, M., Cosley, D., Riedl, J.: Tell me more: An actionable quality model for wikipedia. In: Proceedings of the 9th International Symposium on Open Collaboration. pp. 8:1–8:10. WikiSym '13, ACM (2013)