Semi-Automatic Semantic Moderation of Web Annotations

Elaheh Momeni
«Supervised by: Prof. Dr. Wolfgang Klas»
University of Vienna, Faculty of Computer Science
Liebiggasse 4/3, Vienna 1010, Austria
elaheh.momeni.roochi@cs.univie.ac.at

ABSTRACT

Many social media portals are featuring annotation functionality in order to integrate the end users' knowledge with existing digital curation processes. This facilitates extending existing metadata about digital resources. However, due to various levels of annotators' expertise, the quality of annotations can vary from excellent to vague. The evaluation and moderation of annotations (be they troll, vague, or helpful) have not been sufficiently analyzed automatically. Available approaches mostly attempt to solve the problem by using distributed moderation systems, which are influenced by factors affecting accuracy (such as imbalance voting). Despite this, we hypothesize that analyzing and exploiting both content and context dimensions of annotations may assist the automatic moderation process. In this research, we focus on leveraging the context and content features of social web annotations for semi-automatic semantic moderation. This paper describes the vision of our research, proposes an approach for semi-automatic semantic moderation, introduces an ongoing effort from which we collect data that can serve as a basis for evaluating our assumption, and report on lessons learned so far.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

social web annotation, semantic moderation system, annotation system, semantic web

1. INTRODUCTION AND CHALLENGES

User-generated annotations facilitate the association of additional information with existing resources [18] and deliver valuable economic, social, and cultural information. Therefore, many social media portals, hosting large collections of digitized items, such as Facebook or Flickr.com, feature annotation functionality. With the growing availability and popularity of annotated resources on the social media portals, new opportunities and challenges arise as users can, and do, actively use information technologies to understand the opinions and discover the knowledge of others [15]. This helps to find new trends and extract knowledge of the end-users to facilitate the recommendation, retrieval, and search processes [17].

However, users, who annotate (annotators), have different levels of knowledge, different views of the world, and different intentions [4]. The quality of user-generated annotations varies dramatically from excellent to abusive and vulgar [10]. Managing and hosting these annotations can be costly and time consuming, hence their owners have a great interest in ensuring that these annotations can help to improve the curation process. Therefore, as the volume of such annotations increases, the task of semantic moderation is becoming increasingly important.

Most of the available moderation services (such as MicroSourcing.com and moderation.pro) are based on human intervention, placing annotations in a queue to be checked by a group of moderators or a "forum administrator" before they are viewable by the public. It is obvious that due to increasing volumes of annotations, using such a system without any automation process is almost impossible. Recently, several platforms (such as Slashdot.org, delicious.com) have attempted to solve the problem automatically by using distributed moderation and meta-moderation systems (sometimes referred to as reputation systems). Distributed moderation allows all users to vote and moderate the contributions of other users. Meta-moderation enables any user to judge (moderate) the evaluation (voting) of another user. However, a closer analysis by Lampe et al. [12] revealed that it often takes a long time for especially worthwhile comments to be identified. Moreover, Liu et al. [13] show, that voting is influenced by factors which affect accuracy. For example, in imbalance voting, when an annotation receives a higher rating simply because users may assume that, since it already has a higher rating, it must be a pertinent one. Consequently they vote for it.

Semi-Automatic semantic moderation of web annotations is a relatively new and complex concept that is expected to infer automatically the annotation type (such as troll, vague, or helpful) by analyzing the semantic of annotations. Consequently, definitions of annotation types will vary in different platforms, therefore, we believe, that the semantic moderation is codependent on the policies of the communities that support the annotation systems, the time, the annotation content features, and context features. Content features are textual features such as the annotation sentiment, annota-

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

tion length, etc. Context features are all features that can be extracted from social context and activities of annotators, annotated resources, and other annotations on the same resource. Furthermore, social web annotation is a relatively general term which can refer to tags, product reviews, postings in the CQA and discussion forums, comments on digital resources and so on. However, the focus of this research is on textual comments on multimedia resources such as annotations on cultural heritage resources.

More precisely, in our doctoral research we are interested in developing a semi-automatic semantic moderation system for social web annotations, which is expected to automatically infer annotation types such as helpful, troll, vague, etc. We hypothesize, that exploiting content and context features of annotations can help us to infer more accurate annotation types and furthermore to achieve the semi-automatic semantic moderation. We use the term semi-automatic because the content administrator contributes by defining type definition policies, and by taking the final decision on inferred annotations. Moreover, semantic web technology provides a pragmatic way to model and infer annotation system relations and semantics in machine-processable structures, thus facilitating the implementation of such a system.

Therefore, the general challenge we face in our research is leveraging the context and content of social web annotations for semi-automatic semantic moderation. Our general challenge manifests itself in a number of specific research challenges, such as: (1) which types of annotation can be defined in specific policy-dependent use cases, (2) which content and context features are most adequate and how to model, capture, and construct them, (3) what are the correlations between the features and types of annotations, and (4) what is the most accurate method to infer annotation types in an annotation system based on the respective community moderation policy and features and how does the method deal with incomplete feature sets.

In the following section we give a short overview of available related work. Section 3 provides an overview of our proposed methodology, moderation system framework, and evaluation method. In Section 4 we describe an experiment, that we started to collect data to serve as basis for building the moderation gold standard and a deeper automatic moderation evaluation. Furthermore we give an overview of the preliminary results achieved to date. Finally, Section 5 concludes our discussion.

2. RELATED WORK

Related problems that many researchers are confronted with are how to predict: the helpfulness of a product review (e.g., how many people have considered a particular product review helpful), the quality of a posting in a CQA platform, the helpfulness of a collaborative social tag, or the credibility of a posting in micro-blogging. Many of these approaches have demonstrated that a few relatively straightforward features and strategies can be used to predict with high accuracy whether a posting is helpful, high quality, or credible. Table 1 shows some of these features and strategies we found in related work, which are categorized into two categories: "Content" and "Context" features.

There are some areas of research on folksonomy, which are related to our research. Bischoff et al. [3] discuss the potential of different kinds of tags from collaborative tagging systems to improve search and they compare the kinds of tags with user queries posted to search engines. Halpin et al. [9] discuss a number of issues relevant to the question of whether a coherent way of organizing metadata can emerge from distributive tagging systems. Kawase et al. [11] propose an approach to generate and enrich resource profiles that exploit the multiple types of contextual information available in social tagging systems.

Table 1: Overview of some features and strategies extracted from related work

Features & Strategies	Ref	Short Description
Content Features		
Text-Structure	[1][7]	length, readability, #token, etc
Text-Sentiment	[7][6]	tone (subjective or objective), sentiment polarity, #name entities
Context Features		
Semantic Meaning	[11]	the annotation, which has been linked to external resources
Spatial & Temporal	[11]	area size or length of the selected fragment
Annotator role	[2][16]	the behavior and background of the annotator in the system
Annotator consistency	[14][6]	the similarity between the annotations of the same annotator
Trust consistency	[14]	the similarity between the annotations of two annotators, who trust each-other
Co-citation consistency	[14]	the similarity between annotations of two annotators who are trusted by the third party

Assessing the quality of user-generated information is also critical in other domains such as evaluation and propagation of trust and reputation assessment in social networks. Golbeck [8] has developed trust metrics and used ontologies to express trust and reputation information (FOAF schema is extended to include trust assertions with values ranging from 1 to 9). Bizer and Cyganiak [4] propose the WIQA Information Quality Assessment Framework, which enables users to employ different information filtering policies and generate explanations about the filtering process.

3. METHODOLOGY AND APPROACH

In this section, first we present a formal definition of moderation system elements and then we describe our proposed methodology and approach.

3.1 Definition

An annotation system consists of three finite sets, U, A, and R, whose instances are called annotators, annotations, resources. There are also some relationships between these sets: Y is a ternary relation between them, i.e., $Y \subset U \times A \times R$, that maps each annotation, a, to a unique resource and unique annotator. S is a relation, $S \subset U \times U$ that defines the social network relationships between annotators. ft_a is a function which assigns a temporal marker to each Y. Furthermore, A_r is a sub-set of A, whose instances are all annotations on a resource $r \in R$ and P is a set of all A_{r_i}, $P = \{A_{r_1},..., A_{r_N}\}$. The information about annotated resources, annotators along with the social network of the annotators, and other annotations on a same resource places the annotations within an annotation context [14]. We present a formal definition of an annotation context as follows:

Definition 1 *(Annotation Context). Given a set of annotations A, we define the annotation context of the set A as the tupel $C(A) := <U, R, Y, ft_a, S, P>$, of the set of anno-*

tators U, the set of resources R, the annotator and resource mapper relation Y, temporal marker function ft_a, the social network relation S, and the set of all annotations on a resource P.

A moderation system consists of a finite set of annotations A, a finite set of annotation types T, annotations context $C(A)$, and a function M. M is the moderator function, that, based on the given input data $\{A\ ,\ C(A)\}$ infers the annotation type $t \in T$ for an annotation a, i.e., $M : A \rightarrow T$. Therefore, we want to train the moderator function M, that, for an annotation a, infer the annotation type. An annotation a is represented as an f-dimensional vector over a feature space F constructed from information in A and $C(A)$.

3.2 Methodology

In order to develop and train the moderator function, define the elements of type set, and develop the moderation system we purpose a methodology composed of 4 stages:

(1) **Modeling the features of annotations and designing the moderation system**. To capture features and related activities of annotators we define an ontology AMOWA (Automatic Moderation Of Web Annotation), which is partly composed of and extended from available relevant vocabulary. For example we used OAC[1] for modeling the annotation characteristics, FOAF[2] for modeling the annotators, and SIOC[3] for modeling user activities and interactions. Figure 1 shows an abstract overview of this ontology. The context of an annotation changes by adding more annotations on the same resource and annotator activities. Therefore, the model stores different versions of annotation types at different points in time. This data helps to infer the types of annotations more accurately.

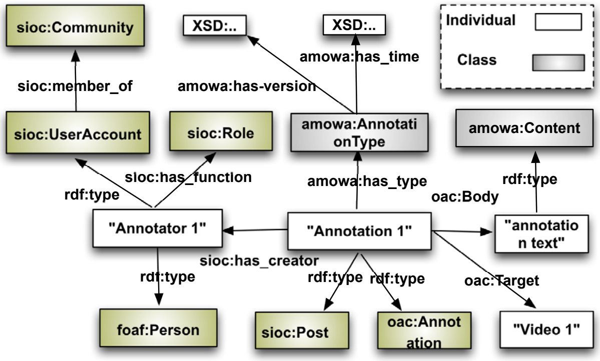

Figure 1: Abstract overview of AMOWA ontology

(2) **Setting up different experiments for gathering the training material**. The training material may be prepared from the historical data of annotation systems and can be continuously updated with new cases. During the learning phase, various moderation judgments are gathered from independent human judges in order to understand the correlation between features and annotation types. More precisely we will set up different experiments for:

- defining annotation types (define elements of the T set) regarding a policy of the community the annotation system is trying to preserve and the level of moderation the community wishes to impose on annotation. Furthermore, exploring respective user agreements. For defining the exact set of types, we, first, define general types based on different use cases made available by state-of-the-art analysis and, second using the crowd-sourcing mechanism for gathering type suggestions and complete the set.

- defining content and context feature patterns which return the most accurate results regarding annotation type inference and defining features, which emerge as most important (define elements of the F set).

- defining bins for features. In order to normalize and combine the features we need to map features-to-value, allowing the bound of the value to be adjusted, therefore through binning we set the bounds. For example, a process that discretizes continuous values of features into the certain bins (such as “low”, “medium” or “high’).

- building a “gold standard” to train our moderation system based on user studies such as human moderation judgment. For the user study we develop a questionnaire and use the crowd-sourcing mechanism in order to collect judgements from independent judges (partly are selected from experts) for each annotation in an annotation set, crawled from a real annotation system. However, creating the gold standard is a critical task and therefore, in order to avoid human judgment errors, we will initially set up small experiments based on a small set of annotations and will ask different independent trustworthy users to judge annotation types and related features. Secondly we will analyze the results and extract user agreements. Subsequently, we will set up more extensive experiments based on a bigger set of annotations by using crowd-sourcing. To validate these extensive experiments we will utilize the user agreements of the small experiments. The following section describes our primary effort in building our gold standard.

(3) **Developing and training a method to infer the annotation types** (training the moderator function M). The method will be able to infer annotation types based on the extracted features and strategies. It must also be able to infer annotation types if some features are missing or are inaccurate. We apply different algorithms (such as Decision-tree and Naive Bayes), which are mostly used in available approaches [5][16] for solving similar problems, and compare which algorithm returns the most accurate results.

(4) **Evaluation of the proposed moderation system**. We will evaluate the system based on two evaluation strategies: first we verify whether the moderation system produced by our solution confirms the people’s perceptions of moderation. Second we configure the proposed moderation system with different combinations of annotation features and then compare the results with the created gold standard from the second step of the methodology. Finally, we analyze and evaluate which feature combinations return the most accurate results. In both strategies, if the accuracy and

[1] http://www.openannotation.org/
[2] http://xmlns.com/foaf/spec/
[3] http://sioc-project.org/ontology

performance of the system is not as expected, we will repeat the second step. I.e. we will set up other experiments or repeat the previous experiment with a new setting, re-design, re-develop and improve the system.

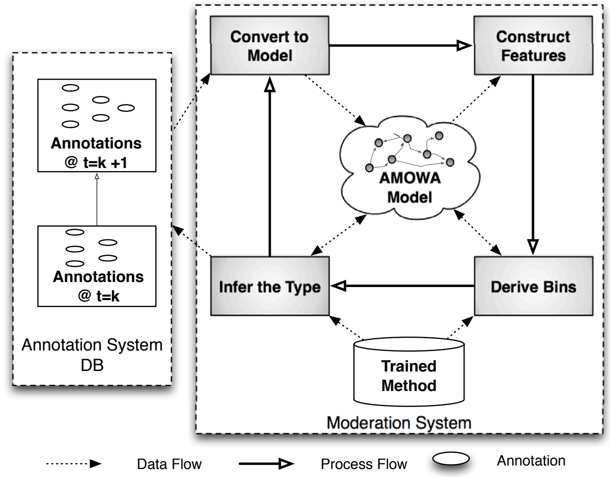

Figure 2: Overview of the proposed approach for semi-automatic semantic moderation of web annotation

3.3 Proposed Approach

Our proposed approach[4] for semi-automatic moderation is shown in Figure 2 and is composed of four stages that function in a cyclical manner. The moderation system: (1) converts user-generated web annotation content and context and user activities to the AMOWA model described in the previous section, (2) constructs the content and context features at a given point in time, (3) derives bins for each feature, and (4) applies the trained method to infer annotation types. Finally, it escalates annotation types to the first level of moderation. This system based on the frequency of receiving the annotations repeats the assessment process and stores the annotation type of each iteration at a given point in time. We use the term semi-automatic, because the content administrator contributes by defining policies and takes the final decision on inferred annotations.

4. EXPERIMENT SET UP

In carrying out the second step of the proposed methodology (especially in order to learn about annotation types and features in the cultural heritage domain) to create a gold standard for the moderation system, develop and train the moderator function, and evaluate the proposed moderation approach, we are conducting an ongoing experiment on real world annotation data harvested from Flickr.com. In order to prepare the training material we compiled a data set from real world annotations on the Flickr-photos of the Library of Congress (LOC) and the related contextual information. Afterwards, we started a user study via a questionnaire. We use the crowd-sourcing mechanism in order to collect relevance judgements from independent judges for annotations in the crawled annotation. We will invite voluntary users to collect judgements for each annotation. Volunteers are recruited from appropriate mailing lists in the digital library and history domain. Subsequently, in order to validate judgments and to verify user agreements about questions (because as we described in previous section building the gold standard is a very critical task and depends on how standardized we designed the questionnaire) we therefore selected a random sub-set of crawled annotations, in total 1,000 annotations and collected judgements from three independent trustworthy judges for each annotation in a sub-set of the dataset.

Judges were asked to select an annotation type using (a) a multiple-choice selection of pre-defined types (such as vague, troll, helpful-informative, helpful-opinion, personal-opinion) and (b) an open question for gathering type suggestions from the judges. Moreover, judges were asked to answer different questions based on the content and context features of annotations such as "what is the tone of the annotation". For this experiment we try to examine features shown in Table 1 and we defined four types of annotations as follows: "Helpful-Informative" when a comment very straight forward provides an informative, well written, and comprehensive description of resource entities (e.g., "He is Mr X and performed Y in 1920"). "Helpful-Opinion" comment, which provides the informative and subjective opinion of an annotator about resource entities (e.g., "He was one of the best football players of 1980"). "Personal-Opinion" when comment generally describes the emotion of the author about image/image set (e.g., "I love old photos like this"). "Vague" annotation, which is highly personal and irrelevant and cannot add any value to the system (e.g., "wooooow, he looks like my father"). "Troll" annotation, which provides inflammatory, extraneous, or off topic annotations, with the primary intent of provoking readers into an emotional response (e.g., "He was one of the best killers of the decade, I appreciate him").

4.1 Results Achieved to Date

In order to examine the user agreement for annotation types we did some analysis on the results of the first phase of the experiment. The details of the user agreement analysis are given in figure 3. In both charts, J-1, J-2 and J-3 represent the 3 judges. We can observe from Chart A that the number of annotation types, selected by the judges are similar for four types: Helpful-Informative, Helpful-Opinion, Personal-Opinion, and Vague. However, they judged only a few annotations as Troll type. This shows, that there are not so many Troll annotations in the selected set. Chart B reports the level of user (inter-rater) agreement based on Cohen's Kappa. From Chart B we can see that the Kappa scores are all above 0.8, which indicates almost perfect agreements.

5. CONCLUSIONS

Analyzing the content of web annotations has recently begun to attract more attention. Manual moderation cannot handle the increasing volume of annotations. Therefore, the task of semi-automatic semantic moderation of web annotations becomes increasingly important. We believe that analyzing and exploiting both content and context dimensions of annotations may help us to achieve the development of the

[4] Based on the approach proposed by Angeletou et al. [2] for inferring the user roles in online communities.

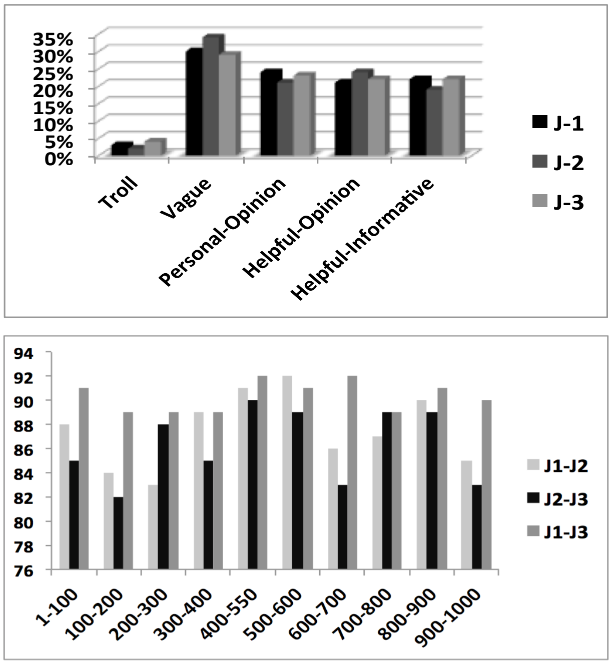

Figure 3: Overview of the user agreement results on annotation types

semi-automatic semantic moderation system. Our proposed methodology to achieve semantic moderation is composed of four stages: modeling the content and context features of annotations and designing the moderation system, setting up different experiments for gathering the training material, developing and training a method to infer the annotation types, and finally evaluation of the proposed moderation system.

Our next steps will be: understanding the correlation between the features and annotation types and train a method for semantic moderation, setting up other experiments based on other data sets (e.g., comments on Youtube.com videos) to work on Troll type, discovering that how the community policy can be represented and integrated into our inferring method, verifying whether the result of moderation method produced by our solution confirms the people's perceptions of moderation, and applying some text enrichment using external resources (such as Linked Open Data resources), and discover, how these resources can help the moderation task.

6. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of WSDM*, 2008.

[2] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *Proceedings of the 10th international Semantic Web Conference*, ISWC '11, 2011.

[3] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, 2008.

[4] C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semant.*, 7(1):1–10, 2009.

[5] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *the 20th international conference*, WWW, 2011.

[6] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 2009.

[7] A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC Proceedings of the ninth international conference on Electronic commerce*, 2007.

[8] J. A. Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland at College Park, 2005.

[9] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, 2007.

[10] N. Jindal and B. Liu. Analyzing and detecting review spam. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining.*

[11] R. Kawase, G. Papadakis, and F. Abel. Generating resource profiles by exploiting the context of social annotations. In *Proceedings of the 10th international Semantic Web Conference*, ISWC '11, 2011.

[12] C. Lampe and P. Resnick. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, 2004.

[13] J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou. Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[14] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.

[15] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Journal of Foundations and Trends in Information Retrieval*, 2(1-2), 2008.

[16] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Extended Semantic Web Conference, ESWC*, 2011.

[17] B. Haslhofer, E. Momeni, M. Gay, and R. Simon. Augmenting europeana content with linked data resources. In *Linked Data Triplification Challenge, co-located with I-Semantics 2010.*

[18] R. Simon, B. Haslhofer, W. Robitza, and E. M. Roochi. Semantically augmented annotations in digitized map collections. In *ACM?IEEE Joint Conference on Digital Libraries (JCDL)*, New York, 2011. ACM.