An algorithm for labels aggregation in taxonomy-based crowd-labeling

A Ponomarev¹, T Levashova¹ and N Mustafin¹
¹ St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 39 14th Line, 199178, St. Petersburg, Russian Federation
E-mail: ponomarev.a.v@gmail.com

Abstract. Crowdsourcing provides a convenient solution for many information processing problems that are still hard or even intractable by modern AI techniques, but are relatively simple for many people. However, complete crowdsourcing solution cannot go by without a quality control mechanisms, as the results received from participants are not always reliable. The paper considers taxonomy-based crowd-labeling – a form of crowdsourcing, in which participants label objects with tags, and there exists an explicit taxonomy relation on the set of tags. We propose a method and an algorithm for label aggregation, allowing to estimate the likelihood of the true object label from a set of noisy labels received from the crowd, and to estimate the expected crowd members’ accuracy. The proposed method and algorithm can be used in a wide range of crowd-labeling applications (e.g., classification of scientific literature collections, software repositories, etc.).

1. Introduction
Crowdsourcing is usually understood as the involvement of a large group of people in solving the problems on a voluntary basis with the use of information and communication technologies. Nowadays, crowdsourcing is widely used in various fields to address problems that are relatively hard for automation, but easy for humans [1-3]. There is a variety of crowdsourcing kinds [4,5], classified by the type of tasks offered to crowd members, type of expected contributions, incentivization and rewarding schemes. One of these kinds is crowd-labeling. In crowd-labeling, the collective effort of the crowd is used to assign labels to a (usually, very large) collection of objects (i.e., images, audio or video clips, etc.).

In most crowd-labeling applications, the set of labels is either an unconstrained set of words in some natural language, or a predefined restricted set of terms relevant for the application (for example, when crowd-labeling is used to collect training datasets for machine learning, the set of labels corresponds to the set of output labels of the machine learning model). In both cases, possible semantic relationships between the labels are discarded, therefore, any two labels in the set are considered different.

This paper is dedicated to a specific kind of crowd-labeling, where there is an explicit taxonomy relation between labels in the set. One of the most important applications for taxonomy-based labeling is enabling semantic search on a collection of objects. That is, making it possible to search for objects providing explicit queries (e.g., using Semantic Web query languages) and using search engines that can make use of explicit semantic relationships between concepts describing objects.

Applying crowdsourcing to some problem is connected to potential issues with the quality of the results, obtained from the crowd [6-8]. This problem stems mainly from the general human propensity...
for errors, and from “free riding” attempts (typical for monetary crowdsourcing – when a crowd member submits low effort to get payment). Therefore, a number of quality control mechanisms has been proposed to ensure the quality of the results [8].

Most of the mechanisms are based on comparing the results obtained from different participants. In crowd-labeling applications where all labels considered to be “equally different” such comparison may be not very informative. On the other hand, taking into account information about relationships present between labels may significantly improve reliability of both aggregated labels and estimates of the individual diligence of crowd members. For example, despite the labels “Student” and “Human” are different, the label “Student” does not contradict to the label “Human” (in a common-sense taxonomy), but is a refinement of it. It is not reasonable to assume that these two labels just mismatch, in fact, they match to the point that the labeled object is a human, but one of the labels carries more information.

Nowadays, a standard way of representing taxonomies (as well as more subtle logical relationships between concepts) is to use ontologies and ontological languages, first of all, OWL 2 (the specification of which is standardized by the W3C consortium) [9]. OWL 2 can be used to precisely define concepts and semantic relationships between them. One of built-in relationships defined by OWL 2 specification is SubClassOf relationship, corresponding to taxonomy.

The paper proposes a method and an algorithm for label aggregation, allowing 1) to estimate the likelihood of the true object label from a set of noisy labels received from the crowd, and 2) to estimate the expected crowd members’ accuracy (e.g., to disable those who provide low quality labels).

2. Related work
Crowdsourcing quality control is an important problem, fettering the use of crowdsourcing and attracting the attention of many researchers in this area (see, e.g., [7]). The approach to quality control proposed in this paper falls into a large group of consensus methods (sometimes also called aggregation methods) that use the redundancy of the results of one task collected from several crowd members to resolve contradictions between them. Consensus methods are widely used partly due to their flexibility and low data requirements – they rely only on labels received from various participants. However, most of the existing consensus methods are based on a strict comparison of the results obtained from individual participants, ignoring possible relationship between labels. There are only few publications trying to adapt consensus methods to situations where there are such relationships. For example, in [10] different models for representing relationships between labels are explored (including a Bayesian network). The authors of [11] propose a probabilistic labeling model for hierarchical classification, i.e., for the situation when the labels that participants assign to objects are classes organized in a hierarchy (classification of books, goods) – very close to the goal of this paper.

However, the results obtained in these papers are not entirely applicable to the problem under consideration: the set of consensus methods proposed in [10] is based on the DS algorithm, which has high computational complexity, therefore, its applicability for labeling using large ontologies is limited. The algorithm also requires high redundancy in labels, which can become economically unfeasible if crowdsourcing is based on monetary rewarding. The method proposed in [11] is intended for the scenario of sequential tag refinement by a crowd member, which is not always reasonable and convenient.

Thus, currently, there are no methods for aggregating semantic labels for non-sequential labeling, taking into account typical size of the ontologies used in labeling.

3. The method and algorithm for label aggregation
The proposed algorithm and method are based on the following property of the taxonomy relationship (SubClassOf relationship of the OWL 2). If $C_1$ is a subclass of $C_2$ then any instance of class $C_1$ is also an instance of class $C_2$. Therefore, if some object has label $C_1$ then this object can also be described by $C_2$, such that SubClassOf($C_1$, $C_2$). Moreover, due to transitivity of the SubClassOf relation, this object can also rightfully be described by any class, reachable from $C_1$ via a chain of explicit SubClassOf
axioms. By ontology graph induced by the SubClassOf relationship in this paper we mean a graph $G_O = (V_O, D_O)$, such that the set of nodes $V_O$ is the set of concepts defined by the ontology $O$, and the set of arcs $D_O$ corresponds to the set of SubClassOf axioms in the ontology $O \ (\langle V_O(C_1), V_O(C_2) \rangle \in D_O \iff \text{there is } \text{SubClassOf}(C_1, C_2) \text{ axiom in ontology } O)$.

Crowd-labeling is connected with uncertainty. In particular, if there one of the crowd members assigned a label to some object, we cannot be sure that the object actually belongs to the class corresponding to this label. To deal with this uncertainty, the proposed model and method use the adaptation of the Buchanan model for reasoning under uncertainty. Each participant $u$ is assigned a parameter $b_u$ – a degree of belief that takes values in the range $[0; 1]$. Each statement about correspondence of a label $l$ to some object $i$ is also assigned a degree of belief $\mu(i, l)$ (also in the range $[0; 1]$).

The proposed aggregation method:

a) calculates degree of belief for class labels not reported directly by the crowd member (propagation of the degree of belief via the ontology graph induced by the SubClassOf relationship);

b) adjusts the participants’ degrees of belief based on the comparison of the sets of labels reported by different crowd members.

Propagation of the degree of belief via the ontology graph induced by the SubClassOf relationship is based on the mentioned earlier properties of the taxonomy relation. If there is evidence that object $i$ belongs to class $l$ with degree of belief $\mu(i, l)$, this degree of belief is also transferred to the fact that this object belongs to all classes (labels) located above $l$ in the graph $G_O$ (see Fig. 1).

![Figure 1](image)

**Figure 1.** Propagation of the degree of belief via taxonomy relation.

A special case is the estimation of a degree of belief that a label corresponds to an object in case there are several evidences (possibly received from different participants with similar of different degrees of belief). Let $\mu_1, ..., \mu_n$ be the degrees of belief in the statement that the object $i$ belongs to class $l$ obtained from different participants (possibly during the “propagation” of the degree of belief through the class hierarchy described earlier). The function used in combining the results of $f_m$ must satisfy the following restrictions:

4. $f_m(\mu_1, ..., \mu_n) \geq \max(\mu_1, ..., \mu_n)$. If there are several evidences that some fact is true, then the aggregated degree of belief that this fact is true should not be less that any of the evidences’ degrees of belief (evidences reinforce each other);

5. $f_m(\mu_1, ..., \mu_n) = f_m(\mu_{n1}, ..., \mu_{nn})$, where $n1, ..., nn$ give some arbitrary permutation of the indices $1, ..., n$. I.e., the value of the function should not depend on the order of the arguments.

In the proposed method, $f_m$ is defined as $1 - \prod(1 - \mu_i)$ (it is easy to see that this function satisfies the restrictions above).

The degree of belief of the labels for the object, in turn, is the basis for adjusting the degree of belief of participants. In particular, for this, various measures of semantic distance between ontology classes are used, defined on the $G_O$. The proposed algorithm for recalculation of the participant’s $u$ degree of belief is following:
Step 1. Using the previously described algorithm calculate the consensus degree of belief for all possible labels using data received from all participants except \( u \) (in particular, this means that if an object was marked up by only one user, then this does not affect his/her degree of belief).

Step 2. For the label \( c \) received from the participant \( u \) find labels (ontology classes) \( c^*, c^t, c^b \) and \( c' \) (see Fig. 2, dashed rectangle is the label provided by the user \( u \), filled rectangles – labels with non-zero degree of belief, as estimated by degree of belief propagation in Step 1). To find these labels directed acyclic graph \( G_o \) is considered. The labels are defined as follows: \( c^* \) – is the nearest node of the \( G_o \) with non-zero degree of belief reachable from \( c \), \( c^t \) – is the farthest node of the \( G_o \) reachable from \( c \), \( c^b \) – is the farthest from \( c \) node of the \( G_o \), from which \( c \) is reachable, and \( c' \) – is the farthest from \( c^* \) node with non-zero degree of belief, from which \( c^* \) is reachable.

Step 3. Participant’s degree of belief is calculated as an arithmetic average of the degree of correspondence of his/her labels and the labels of other participants for the same objects:

\[
\hat{b}_u = \frac{1}{|n_u|} \left( \sum_{i \in n_u} \left( \phi_1(c_i, l_{i,u}, l_{i,-u}) + \phi_2(c_i, l_{i,u}, l_{i,-u}) + \phi_3(c_i, l_{i,u}, l_{i,-u}) \right) \right),
\]

where \( n_u \) – the set of objects labeled by the participant \( u \), \( l_{i,u} \) – is the label of the object I provided by the participant \( u \), \( l_{i,-u} \) – aggregated labeling of the object \( i \) without participant \( u \). Functions \( \phi_j \) reflect three empirical factors, influencing the evaluation of the labeling performed by the participant \( u \). Coefficients \( \alpha_j \) give relative importance of these factors (summing up to 1). Particular values should be tuned w.r.t. the size and structure of particular ontology.

![Figure 2](image-url). Adjustment of the participant’s degree of belief.

Empirical factor evaluation functions use the labels identified in step 2 and are calculated by the following formulae:

\[
\phi_1(c, l_{-u}) = \frac{d(c^*, c^t)}{d(c^b, c^t)} \mu(c^*);
\]

\[
\phi_2(c, l_{-u}) = \max\left(\frac{d(c^b, c^t)}{d(c^b, c^t)}, \frac{d(c, c^t)}{d(c^b, c^t)} \right) \mu(c^t);
\]

\[
\phi_3(c, l_{-u}) = \frac{d(c, c^t)}{d(c^b, c^t)} \mu(c^b);
\]

Here, \( d(c_1, c_2) \) is the distance between the nodes in \( G_o \) (the number of arcs). The factors allow “transfer” some degree of belief from the aggregated labels to the degree of belief of the participant, depending on how close are his/her labels to the aggregated labels (proposed by other participants). The first factor measures how specific is class \( c^* \) (the most specific class on which the opinion of the participant agrees with the opinion of other participants) in the hierarchy, the second factor allows to "horizontally" transfer the degree of belief from the most specific aggregated label (taking into account
distance to it in the $G_0$), finally, the third factor shows how general the proposed label is (specific labels are generally better).

Recalculation of labels’ and participants’ degrees of belief occurs iteratively. Based on the initial degrees of belief of participants, degrees of belief of labels are estimated, then participants’ degrees of belief are calculated. Using new participants’ degrees of belief, again, labels degrees of belief are adjusted and so on until the moment when after a particular iteration the maximum absolute change in the value of degrees of belief will not be less than a specified threshold.

4. Conclusion
The paper considers taxonomy-based crowd-labeling – a form of crowdsourcing, in which participants label objects with tags, and there exists an explicit taxonomy relation on the set of tags. For this kind of crowd-labeling application, it proposes a method and an algorithm of quality control. The proposed method and algorithm make use of the taxonomy relationships between possible labels and allow 1) estimating the likelihood of the true object label from a set of noisy labels received from the crowd, and 2) estimating the crowd members’ expected accuracy.

The proposed method and algorithm can be used in a wide range of crowd-labeling applications (e.g., classification of scientific literature collections, software repositories, etc.).

5. Acknowledgments
The reported study was funded by RFBR, project number 19-07-01120.

References
[1] Harvey D, Kitching T D, Noah-Vanhoucke J, Hamner B, Salimans T and Pires A M 2014 Observing Dark Worlds: A crowdsourcing experiment for dark matter mapping Astron. Comput. 5 35-44
[2] Shin D, Aliaga D, Tunçer B, Arisona S M, Kim S, Zünd D and Schmitt G 2015 Urban sensing: Using smartphones for transportation mode classification Comput. Environ. Urban Syst. 53 76-86
[3] Meisel Z F, VonHoltz L a. H and Merchant R M 2015 Crowdsourcing healthcare costs: Opportunities and challenges for patient centered price transparency Healthcare pp 1-3
[4] Assis Neto F R and Santos C A S 2018 Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management Inf. Process. Manag. 54 490-506
[5] Ghezzi A, Gabelloni D, Martini A and Natalicchio A 2017 Crowdsourcing: A Review and Suggestions for Future Research Int. J. Manag. Rev. 20 343-63
[6] Allahbakhsh M, Benatallah B, Ignjatovic A, Motahari-Nezhad H R, Bertino E and Dustdar S 2013 Quality Control in Crowdsourcing Systems: Issues and Directions IEEE Internet Comput. 17 76-81
[7] Daniel F, Kucherbaev P, Cappiello C, Benatallah B and Allahbakhsh M 2018 Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques and Assurance Actions ACM Comput. Surv. 51 article 7
[8] Ponomarev A A V 2017 Quality control methods in crowd computing: literature review SPIIRAS Proc. 5 152-84
[9] Anon OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax
[10] Duan L, Oyama S, Sato H and Kurihara M 2014 Separate or joint? Estimation of multiple labels from crowdsourced annotations Expert Syst. Appl. 41 5723-32
[11] Otani N, Baba Y and Kashima H 2016 Quality control of crowdsourced classification using hierarchical class structures Expert Syst. Appl. 58 155-63