Managing Quality of Crowdsourced Data

Nguyen Thanh Tam
LSIR, I&C, EPFL

***Abstract*—The Web is the central medium for discovering knowledge via various sources such as blogs, social media, and wikis. It facilitates access to contents provided by a large number of users, regardless of their geographical locations or cultural backgrounds. Such user-generated content is often referred to as *crowdsourced data*, which provides informational benefit in terms of variety and scale. Yet, the quality of the crowdsourced data is hard to manage, due to the inherent uncertainty and heterogeneity of the Web. In this proposal, we summarize prior work on crowdsourced data that studies quality dimensions and techniques to assess data quality. However, they often lack mechanisms to collect data with high quality guarantee and to improve data quality. To overcome such limitations, we propose a research direction that emphasises on (1) guaranteeing the data quality at collection time, and (2) using expert knowledge to improve data quality for the cases where data is already collected.**

***Index Terms*—crowdsourced data, quality management**

I. Introduction

THE Web emerged as the repository of *crowdsourced data*. The term refers to wide-ranging types of contents that are generated by a (generally large) number of people [1]. One source to collect crowdsourced data is from people who voluntarily produce content by reporting scientific studies, uploading their comments, writing product reviews, and sharing knowledge via various Web platforms, such as blogs (e.g. Tumblr, Wordpress), social media (e.g. Twitter, Facebook), and wikis (e.g. Wikipedia, Wikirate). Another source of crowdsourced data is employing "human workers" via online services (e.g. Amazon Mechanical Turk, CrowdFlower). Examples of crowdsourced data and their statistics can be found in Table I. In general, the "wisdom of the crowd" in crowdsourced data is widespread, as the users may participate at any time and location convenient for them [2], [3].

TABLE I
EXAMPLES OF CROWDSOURCED DATA

Data sources	Size	#Users	Content
Twitter [4]	$\sim$ 0.5B tweets/day	$\sim$ 0.3B active users	opinions
Tumblr [5]	$\sim$ 100B posts	$\sim$ 0.25B blogs	arguments
Wikipedia [6]	$>$ 35M articles	$>$ 70K active contributors	facts
CrowdFlower [7]	$\sim$ 1B tasks	$\sim$ 5M workers	annotations

The usage of crowdsourced data brings many benefits. First, the ubiquitous nature of crowdsourced data enables its collection in a *timely* manner. For instance, a large number of people distributed throughout different locations report crisis events of a particular area such as earthquake or violence via mobile devices in a matter of seconds [8], [9]. Moreover, the collection of crowdsourced data is often *cost-efficient* for application providers, as the Web gives access to human knowledge at virtually zero cost. Another example, social media users share their opinions on news events; and thus, these crowdsourced opinions can be used as *augmentation* to the information from mainstream media (e.g. BBC, CNN).

Managing quality of crowdsourced data is important, especially when the accessibility and scalability of data increases. This is because in practice, one have limited control over the selection of crowd participants and little insights into the level of expertise and reliability of the users who provide data. Applications built upon the crowdsourced data will be untrustworthy and of little value without any policies for filtering and repairing errors or omissions in the data. As a result, quality management becomes a paramount task to ensure that the most consistent and reliable data are made available and delivered to the users. Applications that benefit by a rich data quality management include systems for query answering, knowledge bases, decision-support, and recommendations, to name a few.

To understand the crowdsourced data and the state-of-the-art of quality management techniques, I have studied the following works as the foundation of my research proposal.

- *"Verifying crowdsourced social media reports for live crisis mapping: An introduction to information foren-*

Proposal submitted to committee: October 15th, 2015; Candidacy exam date: October 22nd, 2015; Candidacy exam committee: Prof. Boi Faltings, Prof. Karl Aberer, Prof. Matthias Grossglauser.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(name and signature)

Thesis director: _____
(name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(B. Falsafi) (signature)

sics" [8]: introduces the theoretical background and the quality dimensions for crowdsourced data. It also discusses manual techniques to verify data quality.
- *"People on drugs: credibility of user statements in health communities" [10]:* studies the accuracy dimension of quality. It models the accuracy as credibility and develops an automatic technique to classify the credibility of data.
- *"Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose" [11]:* studies the representativeness dimension of quality when a sample of data is provided. It concerns how much data is collected versus the coverage of original information. Automatic techniques are shown to measure the representativeness of sampled data and practical insights are given.

In this proposal, my research goes beyond the state-of-the-art by considering two novel directions:

- *Guiding data acquisition with quality constraints:* Different applications might have different requirements for data quality. To satisfy these requirements, it is often cost effective to control the quality during the time data is collected. We go beyond existing work by enabling users to define quality requirements and guiding the data acquisition process to meet such requirements.
- *Leveraging expert knowledge to improve data quality:* In practice, there are cases that data is already collected and its quality needs to be improved. We propose leveraging expert knowledge to improve data quality, since there is no generic improvement heuristic and user-generated content is easily validated by human. However, the availability of experts is often limited and they incur high costs. We develop techniques to minimize expert effort.

The rest of the writeup is organized as follows. Section II discusses the work of Meier [8] about the data quality dimensions and verification strategies. Section III and IV describe the work of Mukherjee et al. [10] and Morstatter et. al. [11] about automatic techniques to evaluate the accuracy and representativeness dimensions of data quality, respectively. Section V details the research directions. And finally, Section VI provides a summary and future plan to realize the proposal.

II. Crowdsourced Data: Quality Dimensions and Manual Verification

In this section, we discuss the work of Meier [8], which provides important insights about the quality dimensions of crowdsourced data and the strategies to assess data quality.

A. Quality Dimensions

Analyzing and developing clear and measurable metrics for crowdsourced data is important. This is because the usages of data vary from applications to applications and are subjective to end-user interests. Without pre-defined data requirements, quality management is an inexact science in terms of controls and evaluations. To define quality requirements that are general, one often considers the following quality dimensions.

Accuracy: The degree of correctness and precision with which the data is represented. As provided by humans, crowdsourced data might contain incorrect or partially correct information due to several reasons. The openness of the Web allows people, with wide-ranging levels of expertise, to be free to contribute without any proper control of quality. The identity and trustworthiness of the users who provide the data are often unknown before-hand. As a result, there is a need of identifying the dirty data such as human errors and spams for a better accuracy assessment.

Representativeness: The degree to which the data is representative of overall population [12]. Crowdsourced data is a form of non-probability sampling as the characteristics of each user and the relationships between users are unknown [8]. For example, collecting data about 'climate change' from social media might select only users who have Internet access and use social media. The collected information could be biased since the selected users are not necessarily representative to the overall population (e.g. China does not allow Twitter).

Relevancy: The degree to which the data is relevant for the application domain [13]. For example, tweets about Arab holidays would be irrelevant for a journalist looking to collect information about the 2010 Arab Spring event. Evaluating the relevancy of crowdsourced data is error-prone due to its unstructured and multiple formats (e.g. texts, images, and videos). For instance, Twitter messages do not follow any formal specifications and often lack proper syntax or spelling.

Redundancy: The degree to which the data is duplicated or similar to each other. Redundancy comes from the fact that content in one data source is often the results of copying and adapting existing sources [14]. Even worse, the copied data might be published, shared, or reused without provenance. Redundancy increases the cost and quality in general as the same content is processed many times. For example, if we collect data about side-effects of vaccination from different websites, we can end-up with similar arguments.

Timeliness: The degree to which the data is up-to-date. In live applications such as monitoring social events, natural disasters, or political conflicts, the access to timely data is important. Further, given the uncertain nature of crowdsourced data, not only the data itself has to be updated frequently, but also its quality has to be verified in a timely manner to preserve the usefulness of such applications.

Consistency: The degree to which the data satisfies some pre-specified constraints. Practical applications often use constraints or rules to ensure the consistency of information [15], [16], [17]. The consistent data implies correctness. For example, during the 2010 Parliamentary Elections in Egypt, the U-Shahid team verified user reports from social media. They marked a user report to be of high quality if the user identity was clear and the report itself contained picture or video as evidence.

B. Verification Strategies

The verification strategies used to assess the crowdsourced data quality can be categorized as follows (Figure 1).

Content-based. The idea is to analyze the features of the data itself. There are three main data types of crowdsourced data:

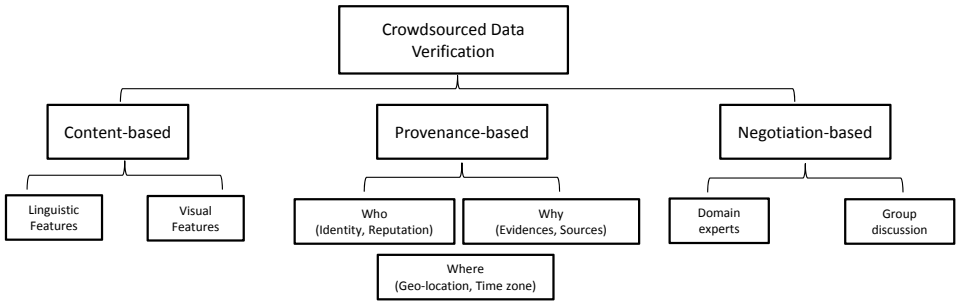

Fig. 1. Verification Strategies for Data Quality

texts, pictures, and videos with different characteristics to be verified. For example, a tweet adopting the language of breaking news such as "urgent", "breaking", and "exclusive" might indicate an important event that needs to be react quickly. However, the overuse of these *linguistic features* might have a negative effect, e.g., the user is trying to get attentions rather than reporting a true event. Videos and pictures can be validated by investigating the visual features. For example, one can analyze the background (weather, lights, shadows) of a picture to check whether the conditions shown fit with the claimed date and time of the reported event. Moreover, one can also analyze the audio features such as vocabulary, slang, accents inside the video to check whether the location is matched with the reported event.

Provenance-based. Other than content characteristics, one can use the provenance of data for verification. The provenance information describes the origins and history of data, including:

- *Who-provenance:* traces the *identity* of authors or data providers such as name, picture, bio, social media account, etc. If a user provides sufficient identity details, he is likely to be confident at the data he provides. We can even contact him via his email/phone for confirmation. Moreover, the quality of data is also affected by user *reputation*. For example, a tweet is given by domain experts is often more accurate than by normal users. One can analyze the social media activities (e.g. #posts, #followers, #thanks) for profiling the reputation of a user.
- *Why-provenance:* concerns the reason that the data is generated in the first place. A crowdsourced data is more reliable if it provides (correct and relevant) evidences. For example, a reporter tweeting about a violence incident during the 2010 Egypt Elections often has a picture as his evidence. But if this picture is validated as actually taken in Tunisia, then the tweet is regarded as incorrect.
- *Where-provenance:* concerns the original sources of data. especially when data on the Web is copied and shared via multiple intermediate channels. For example, a study quoting from a scientific publication is more trustworthy than from news. Moreover, we can also check the geo-location (via IP address) and timezone of the provided data to validate if it is matched with the event it reports.

Negotiation-based. There are extreme cases that crowdsourced data cannot be properly validated via content-based and provenance-based techniques. For example, the linguistic features are not discriminative enough or the reputation value is in the middle. This calls for negotiation-based techniques, that involve a trusted group of validators who discuss with each other to reach an agreement on the final quality. Each participant is responsible for collecting the evidences of data, with or without using the above verification techniques in his own, or evaluating the quality based on his own experiences. The consensus is reached when their verification results are matched with each other or via voting in case of conflicts.

The validator group can be formed by different ways. One can hire some domain experts on the topic that the crowdsourced data is about. If the experts are not available or insufficient, one can create a small social network by allowing the validators to recommend their friends/colleagues who they trust about the given topic. This kind of network is likely to provide good results as there is a certain degree of trust between the participants. Moreover, one can make use of supporting technologies to handle group discussions such as Skype chat group with 2000 members limit.

III. Accuracy of Crowdsourced Data

Accuracy is an important quality dimension to ensure that the high-quality and reliable data are made available and delivered to end-users. The work of Mukherjee et al. [10] proposes the notion of credibility to measure the accuracy. While the first paper studies manual strategies, the second paper proposes an automatic approach for accuracy measurement.

Credibility is a perceived concept to assess whether the data can be trusted or distrusted. Credible data implies a high chance of accurate data. This aligns with the ultimate goal of quality management which is to achieve consistent data for sharing and reusing. Applications built on top of credible data already provide practical usecases. For example, question-answering systems like Quora.com, Yahoo! Answers or Ask.com become popular since they have mechanisms to evaluate the trustworthiness of users who answers the questions. Or a medical system could provide useful warnings about the usage of drugs if they have a reliable crowdsourced data about the side-effects of drugs.

To compute the credibility of crowdsourced data, the paper follows a machine learning approach. The output is to classify for each piece of data whether it is credible or not. To achieve this goal, we need to identify the features of data and to design a classification model. The work of Mukherjee et al. [10]

proposes such a learning approach for the healthcare domain. In what follows, we will summarize their approach.

A. Setting

In this work, crowdsourced data is user-generated textual content in an online health portal, namely healthboards.com. In particular, the information about side-effects of drugs are extracted from the online posts of portal members; i.e. drug users. The quality concern is that the members of this online source are not experts, who do not take responsibility of their statements about side-effects of drugs. As a result, there is a need of assessing the credibility of the statements they provide.

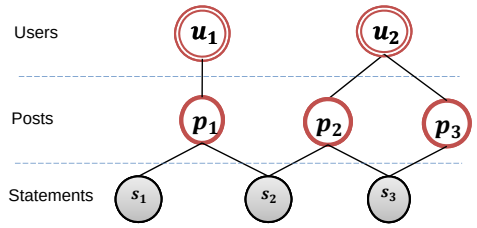

Fig. 2. Correlations in crowdsourced data

Figure 2 illustrates the elements of online health data. There are correlations between user, post, and statement. Each user could have one or many posts but a post is authored by a single user. Each post could have one or multiple statements. For example, a statement about a side-effect of drug is *"Depo-Provera reduces bone density"*, which is extracted from the following user post: *"...Depo is very dangerous as a birth control and has too many long term side-effects like reducing bone density ..."*. Statement extraction from natural-language text is performed by Subject-Predicate-Object extraction methods [18]. The same statement can be mentioned by different posts of the same or other users.

Goal of the paper. The credibility assessment for online health data is formulated as a classification problem: assigning the label for a statement as 'credible' or 'non-credible'. More precisely, given a set of users $U = \{u_1, \ldots, u_{n_U}\}$, a set of posts $P = \{p_1, \ldots, p_{n_P}\}$, and a set of statements $S = \{s_1, \ldots, s_{n_S}\}$, the goal is to infer the labels for every unlabeled statement in S. $s_i = 1$ means the statement s_i is credible and $s_i = 0$ means the statement s_i is non-credible.

B. Classification Features

The classification problem is basically about predicting true values for the output variables from the observed variables. This section identifies the features of such observed variables, including users and posts. The features are indicators of user trustworthiness and language quality that have a strong influence on the credibility of a statement. Intuitively, a statement is likely credible if it is posted by a trustworthy user using confident and objective language.

User Features. The trustworthiness of a user is captured by his profile and activities in the health portal. User profile contains private information such as age and gender. His activity logs include the number of posts and thanks received, among others. For example, a senior member having thousands of posts and receiving hundreds of thanks has good incentives to post credible statements; hence, he should be regarded as trustworthy. Every user is associated with the same number of features, but the feature values are particular to each user.

Post Features. The linguistic characteristics of an online post can reveal the author intention towards providing high-quality information. Intuitively, a post written in confident and objective language often provides credible statements. Technically, every post is associated with the same number of linguistic features. Each feature is computed as the frequency of the words belonging to that feature over the length of the post. To indicate whether a post is objective, two kinds of features are studied:

- *Stylistic Features:* capture the usage of modals, named entities, inferential conjunction, etc. The frequent appearances of strong modals such as "can, may, would" depicts a high degree of uncertainty, implying that the user is not confident in his post. The usage of inferential conjunctions such as "hence, therefore, thus" indicates that the user is giving an argument rather than based on his own subjective opinion.
- *Affective Features:* reflect the attitudes and emotions of an author. For instance, a user writing a post with a high level of hate and negativity might bias the information; and thus, it is doubtful that his statements are credible.

C. Classification model

Towards classifying the credibility of all statements, we have computed the user and post features. However, such information is still not sufficient due to the joint reinforcing relations between users, posts, and statements:

- *Causal relation:* Each statement is provided in a post by a user. Thus, statement credibility jointly depends on both user trustworthiness and language quality; e.g. a trustworthy user could provide a subjective post as his attitude and intentions might change over different contexts.
- *Mutual relation:* Causal relations are further complicated by the overlaps between them: different posts (by different or same users) can mention the same statement. For example, if one user disagrees with a statement but all other users accepts that statement; then he should be regarded as untrustworthy.

To capture these causal/mutal relations along with the observed features, the paper uses a probabilistic model, namely Conditional Random Field (CRF), which is excel at capturing the joint inference between random variables. Specifically, the causal relationship is captured by a clique in CRF and the mutual relationship is captured by the overlaps between different cliques, leading to the factorization of different cliques to compute the probability distributions.

Modeling by Conditional Random Field. A CRF is an undirected graph of random variables U, P, S as vertices. There are three kinds of edges between vertices: (i) edge

between a user and a post, (ii) edge between a post and a statement, and (iii) edge between a user and a statement. We define a clique in CRF as a triangle of a user, a post, and a statement. Since the same statement can be provided in different posts by the same or other users, any random variable can be "repeated" in different cliques. This is the reason why the CRF is constructed on top of the data itself.

Among random variables, U and P are observed variables and S is output variables. Therefore, the classification problem can be expressed in the form of conditional distribution:

$$Pr(S|P,U) = \frac{1}{Z(P,U)} \prod_{c=\{s_i,p_j,u_k\}\in C} \phi(s_i,p_j,u_k) \quad (1)$$

where C is the set of all possible cliques. s_i, p_j, u_k are statement , post, and user of a clique c respectively. $Z(P,U) = \sum_S \prod_c \phi(s_i,p_k,u_k)$ is the normalization constant to ensure that the sum of probabilities over all configurations of S is equal to 1. As aforementioned, each post $p_j = (p_{j,1},\ldots,p_{j,m_P})$ is a feature vector of m_P post features and each user $u_k = (u_{k,1},\ldots,u_{k,m_U})$ is a feature vector of m_U user features. The problem then becomes computing the conditional distribution of S given the user features and post features. This is implemented by the log-linear model (a.k.a. logistic regression) that expresses the log of potential function as a linear combination of associated features:

$$\log \phi(s_i,p_j,u_k) = \sum_{b=0}^{1} I_b(s_i) \times (\sum_{t=1}^{m_P} w_{b,t} \times p_{j,t} + \sum_{t=1}^{m_U} w_{b,t} \times u_{k,t}) \quad (2)$$

where $I_b(s_i)$ is the indicator function of the current configuration of s_i; i.e. $I_b(s_i)$ equals to 1 if $b = s_i$ and 0 otherwise. Hence, we have different weights for each configuration of S. This is straightforward in terms of classification because the features are discriminative indicators for credibility labels of the statements. For example, if a statement is posted in objective language by a confident user, then the probability of its being credible is higher than the probability of its being non-credible given the same set of features. Moreover, we also have different weights for different features. This is because the relative importance between these features varies from applications to applications. As a result, there is a need of learning these feature weights via labeled data.

Learning model parameters. We denote the set of all above weight parameters as W. In other words, the probabilistic model is characterized by W. Learning the model parameters W is usually done by maximum log-likelihood criterion on top of labeled data $\text{argmax}_W \log Pr(S^L|P,U;W)$, where S^L is a set of labeled statements whose ground truth is derived from an expert database. The log-likelihood optimization is convex since the logarithm is a monotonically increasing function and the probability distribution is in exponential form.

However in practice, one cannot always has all data labeled, especially for large-scale data. Hence, the inference problem has to be solved for unlabeled data, which is represented by the conditional distribution:

$$Pr(S^U|S^L,P,U) = \frac{1}{Z(S^L,P,U)} \prod_{c=\{s_i,p_j,u_k\}\in C} \phi(s_i,p_j,u_k) \quad (3)$$

where S^L is a set of labeled statements and S^U is a set of unlabeled statements. For parameter learning, we need to maximize the marginal log-likelihood which is now iterated over unlabeled data:

$$\underset{W}{\text{argmax}} \log Pr(S^L|P,U;W) = \underset{W}{\text{argmax}} \log \sum_{S^U} Pr(S^L,S^U|P,U;W) \quad (4)$$

The problem becomes intractable due to the exponential configurations of S^U. One possible solution is using the Expectation-Maximization approach (EM), that infers the labels of the variables S^U and learn the weight parameters W mutually. The EM algorithm is an iterative method (quick convergent and computationally efficient) that is particularly useful when the likelihood is an exponential family (i.e. maximizing log-likelihood becomes maximizing a linear function). Each EM iteration consists of two steps: (E) *expectation step* which estimates the labels from the current parameter values, and (M) *maximization step* that computes the new parameter values by maximizing the expectation of log-likelihoods as a weighted average of the probability distribution of current label estimates. The process converges when the difference between two consecutive estimates of parameters is insignificant; i.e. the parameters are fit for the labeled data and the likelihood of whole data.

In the end of probabilistic inference, we obtain the credibility labels for all statements S^U and S^L, as an output for the classification problem. At the same time, we also obtain the feature weights W, which can be reused for other datasets which share the same set of features.

IV. Representativeness of Crowdsourced Data

Representativeness is an important dimension of data quality to ensure that we have a valid and unbiased representation of underlying information. The needs for evaluating the representativeness of crowdsourced data become more intense, especially when as more and more data sources and users are integrated in the Web. Since the large-scale of whole data is too cumbersome to use directly, we often extract a sample of data as a representative collection of original contents. The representativeness concern becomes whether the sample data covers similar characteristics of the whole data.

To study the representativeness of crowdsourced data, we discuss the work of Morstatter et. al. [11] on Twitter data, which contains hundreds of millions of users and tweets per day. Twitter publishes different samples of its data: 1%, 10%, etc. of tweets that users mark as 'public'. Since Twitter does not reveal how it samples the data, the concern becomes whether the provided samples are representative of all (public) tweets. To this end, the authors compare the smallest and largest versions of Twitter data, namely: (i) *sample data* –

the 1% of public tweets, provided by Twitter Streaming API, and (ii) *full data* – the 100% (theoretically) of public tweets, provided by Twitter Firehose. The study is limited to the politic domain: Twitter data about Syrian Revolution in 2011 and 2012 is queried for 28 days. To evaluate the representativeness of Twitter data, the paper studies three important data characteristics: *content*, *user network*, and *geolocation*. Via these characteristics, the aim is to provide practical guidelines on how to sample the data effectively.

A. Content-based Measures

Crowdsourced data reflects a wide range of interests by user community. To capture this diversity property, the content-based measures model the data contents as instantiations of different topics. The more topics are included, the higher completeness of sample data. In the context of Twitter data, there are two types of topical measures: (i) *explicit* – each tweet can be assigned manually with a set of hashtags that allows user to mark their common topics explicitly, and (ii) *implicit* – the topics are implicitly mentioned in the text via topical words and thus they need to be discovered automatically.

Hashtag measure. A tweet associated with a set of hashtags explicitly indicates that the tweet belongs to one or many topics correspondingly. As a result, each hashtag can be ranked by its frequency over all given tweets. The ranking of top-k hashtags reflects the most popular topics in the data. Intuitively, a sample data with high representativeness should have the list of popular topics in a similar order to the full data.

More precisely, a ranking of hashtags is an ordered list which defines a total ordering between any two hashtags. To compare two rankings, we take into account the following cases of a given pair of hashtags appearing in both lists:

- Concordant pair: a pair of hash tags $\#A$ and $\#B$ is concordant if both lists rank $\#A$ higher than $\#B$. A high number of concordant pairs indicate a strong similarity in ranking.
- Discordant pair: a pair of hash tags $\#A$ and $\#B$ is discordant if $\#A$ is ranked higher than $\#B$ in one list and ranked lowered than $\#B$ in the other. A high number of discordant pairs indicate a strong dissimilarity in ranking.
- Ties: if both lists rank $\#A$ and $\#B$ equally, it is a two-side tie. If only one of two lists ranks $\#A$ and $\#B$ equally (regardless of the other), it is a one-side tie. Two lists are similar if the ties are two-side and dissimilar otherwise.

Such cases are captured by the Kendall's τ_β coefficient [19], whose value ranges from -1 (two rankings are inverse of each other) to 1 (two rankings are the same). Two rankings are more similar if there are more concordant pairs, less discordant pairs, and less one-side ties.

Guidelines for hashtag measure: The ranking of hashtags indicates how well the sample data preserves the common topics across all tweets. An important finding is that the hashtag measure is not always better when we get more data. For example, if only less than top 100 hashtags are compared, the sampled data performs poorly ($\tau_\beta \approx 0.5$), indicating that the most popular topics from full data is not well-preserved. Another key finding is that the sample data returned by Twitter API has a worse ranking than a random sample, which means that a simple and uniform sampling technique already has a high representativeness in terms of hashtag.

Topic measure. Since hashtag is an ad-hoc annotation by users, it could be too specific or too short to understand. In some cases, it is interesting to have a more coarse-grained view of topics, where a large number of tweets can be grouped into general topics. For example, tweets about BBC and CNN can be grouped into a topic called 'news'. Moreover, some users might not use hashtag; and thus, topics become implicit and hidden in the text.

As a result, it might be useful to automatically detect possible topics. This can be done with Latent Dirichlet Allocation (LDA) [20], a well-known technique that models each tweet as a mixture of topics and each topic as a mixture of words. In other words, LDA classifies each tweet into one or many topics and gives the percentage of each topic assigned to the tweet (e.g. a tweet could be 30% about 'violence' and 70% about 'death'). At the same time, LDA represents each topic as a set of words and derives the proportions that each of these words constitutes the topic (e.g. the 'violence' topic could be constituted by 20% of the word 'injure', 30% of 'harm', and 50% of 'damage').

Topic distribution is a representative characteristic of a given data. Each dataset is characterized by an unordered list of topics, each of which is represented by a distribution of words. To compare the topic distributions between sampled data and full data, we consider two types of information:

- Syntactic similarity: Since each topic is not a concrete label, there is a need of comparing the topics textually. For example, a topic of $\{$injure, harm$\}$ is similar to the topic of $\{$injure, harm, damage$\}$ but not to the topic of $\{$disease, suicide, accident$\}$. One way to compute such syntactic similarity is counting the overlapping words between two topics. As produced by LDA, each dataset has the same pre-defined number of topics. To find all pairs of similar topics between two datasets, we use the maximum weight matching algorithm for bi-partite graph where the weight is the similarity between a topic in the first dataset and another topic in the second dataset.
- Semantic similarity: For a pair of similar topics, we need to also compare the proportions of words that constitute each topic. This is because these proportions represent the co-occurrence of the words over the given data. A good sample data should include topics that are not only textually similar but also semantically similar to the original data. To compare the two proportions, we use the Kullback-Liebler (KL) divergence [21] that measures the information lost when using a probability distribution to approximate another probability distribution. The Kullback-Liebler (KL) divergence is suitable in this case because the sample data is an approximation of full data.

In brief, the topical comparison between two datasets is a histogram of KL divergence values between two lists of topics.

The more small divergence values in the histogram indicates that the two datasets are more similar in terms of topic.

Guidelines for topic measure: Topic measure shares similar findings with the hashtag measure. First, the more data we get, a better distribution of topics is not always achieved. Second, random sampling is better than the Twitter API, which means the true distribution of topics is not well-captured.

B. User network-based Measures

Since crowdsourced data is generated by users, it forms an ad-hoc social network of users where there are different groups of users sharing a wide range of particular interests. For example, one can construct a user-by-user retweet network on top of Twitter data, in which a directed link $u_1 \rightarrow u_2$ indicates that u_1 retweets u_2. Such a network has many connected components, e.g. people interested in 'cat' often retweet each other but not those interested in 'dog'.

As a result, it is necessary to understand the user network characteristics of data, in addition to the content characteristics. The network-related information allows to trace the provenance of data as well as the common group of interests of users. To this end, there are two types of user-based measures studied: *centrality* – identifies the most influential users, and *connectivity* – the degree of correlations between users.

Centrality. This indicator concerns the identification of influential users who are the 'key-players' of making crowdsourced data widespread. There are three possible centrality indicators:

- Degree centrality: counts the number of retweets from a particular user. A network having a high number of nodes with high degree centrality means the data is rich of original information. Oppositely, it means the data is poor and rather redundant because of duplication.
- Betweenness centrality: identifies the brokerage nodes that are bridges of connected components. Basically, it identifies the intermediate users that connect different communities with each other. A high betweenness centrality means that the data is mixed of different sources of information, thus contain unbiased information.
- Potential reach: counts the number of nodes that can reach many other nodes in the network, weighted by the reach distances. In other words, the potential reach measures the degree of spreading a particular information over users.

Guidelines for centrality measure: A key observation is that the centrality measures are more accurate when the sample is larger. Especially, the Twitter API is already able to identify more than 50% of key-players in Twitter.

Connectivity. This measure studies how well the users are connected with each other (*density*) and the distribution of user characteristics (*centralization*) over the network.

- Density: describes the structure of the network as sparse or dense. This is measure by clustering coefficient [22], which reflects the degree of connectivity (i.e. the number of edges) in each connected component. The larger clustering coefficient is, the more dense the network is.
- Centralization: capture the distribution of the above centrality measures over all nodes in the network. It is measured as the ratio of the centrality difference between any two nodes over the maximum possible difference [23]. A high centralization indicates a long-tail network: there are few users with very high influence and there are many users with low influence. Otherwise, we have a uniform distribution of influence.

Guidelines for connectivity measure: A key finding is that all of the connectivity measures are similar to full data when the sample size is substantially large. However, it is difficult to determine exactly how much data is enough as there is no formal comparison between two given measures.

C. Geolocation-based Measures

While content-based measures and user-based measures capture the diversity in data contents and users, geolocation-based measures reflect the diversity of crowdsourced data by geolocation information. Geolocation is an important facet of crowdsourced data to study as e.g., the cultural characteristics and common interests might differ across countries. In Twitter, users can turn on the "location services" that allow to geotag the tweet with their current location. Although there is only a small portion of geotagged tweets ($< 5\%$), it already covers different continents such as Asia, Europe, America, etc.

Intuitively, a sample data with high representativeness should show similar geographic information as the full data. To compute such similarity, two geolocation-based measures are studied: (i) the *percentage* of geotagged tweets from the full data is included in the sample data, and (ii) the *distribution* of geo tags in the data. More precisely, to compute these measures, geotagged tweets are first filtered from both sample data and full data. Then, we count the number of geotagged tweets in full data that are preserved in sample data to compute the *percentage*. Next, the locations of these tweets are grouped by continent and we show the *distribution* of different continents over each dataset.

Guidelines for geolocation measure: An important finding is that the sample data of Twitter API covers more than 90% of geotagged tweets in the full data. And it also retains a similar distribution of geo tags. For example, if 'Asia' and 'American' are most popular geotags in the full data, it is also the most popular geotags in the sample data. It seems that geolocation is the only information (studied so far) that the Twitter API preserves accurately despite of the sample size.

V. Research Proposal

We propose two directions toward managing crowdsourced data quality, including data acquisition and data enrichment.

A. Guiding data acquisition with quality constraints

Data applications have different requirements on quality. For example, a healthcare application often needs a higher accuracy than a social one. Existing work on crowdsourced data often follows a post-processing approach, where the quality is evaluated statically after all data is acquired [8], [10]. To

reach the required quality, it is reasonable (and cost-efficient) to consider incremental data acquisition, where the quality is dynamically controlled. Specifically, we go beyond the state-of-the-art by enabling users to define quality requirements and collecting data intelligently to meet such requirements.

To achieve this goal, we will investigate practical applications for possible quality requirements. This involves identifying key quality dimensions, which in turn enables decomposing a complex requirement into particular dimension(s). This is the basis to interpret and formalize human requirements as a guide for the acquisition of (new) data and guarantee a certain degree of resulting quality. Moreover, we also consider the extensibility of adding new quality dimension.

Another aspect is to understand the trade-offs between quality dimensions (e.g. accuracy, representativeness) and acquisition cost (e.g. time, effort, money). For example, when we try to obtain data as much as possible (high recall), we also get redundant and erroneous data (low precision). Or collecting only data from a highly popular source might introduce low diversity or biases [24]. Satisfying these quality constraints in data acquisition is a multi-criteria optimization problem [25]. We will leverage the advance of techniques on this research topic to find the best strategy for data acquisition.

We envision a general framework of end-to-end data acquisition process, consisting of the following routines. (1) Collect an initial subset of data. (2) Evaluate data quality and apply inference techniques to learn the characteristics of data sources. For instance, which source updates data frequently? which source has biases? and which source has low cost? (3) Driven by this understanding, we will develop a prediction model to estimate the expense of different acquisition strategies and select the best one. This is an adaptive process as the arrival of more data would give better understanding; and thus, change the acquisition strategy accordingly.

B. Leveraging expert knowledge to improve data quality

While the above proposal focuses on data acquisition, this proposal considers the cases where data is already collected. Our goal is to improve the quality of data. Here we go beyond existing work by leveraging expert knowledge to validate data, since there is no generic validation heuristic and user-generated content is often better understood by human. As the availability of experts is limited and they incur high costs, there is a need of minimizing expert effort [26], [27], [28], [29], [30].

To achieve this goal, we need to generate all possible candidates and then guide the expert validation to the most beneficial one. The beneficiary of each candidate can be measured by its potential to identify problematic data and low-quality data sources [28]. Removing problematic data from low-quality sources would enhance the overall quality. Besides, we also minimize the interaction time by asking a group of top-k candidates for validation. This is because human involvement incurs a high latency between two consecutive validations. However, defining the beneficiary for a set of candidates is challenging as they might not be independent.

Another aspect to consider is propagating expert input effectively for learning and reasoning. The history of expert validation can be used as adaptive information to improve heuristics for generating and ranking problematic data. Further, it can be used to train an automatic model that gradually takes over the expert work to infer the correctness of non-validated data. As a result, the cost of acquiring expert input is even more reduced during the improvement.

VI. Conclusions and Future Work

This writeup demonstrates the importance of managing crowdsourced data. On the one hand, crowdsourced data brings the wisdom of the crowd from online sources. On the other hand, the openness of the Web makes crowdsourced data inherently uncertain, calling for the need to manage its quality. Several techniques have been proposed to assess the quality of crowdsourced data in different dimensions. However, there are still open problems regarding quality constraints and mechanisms to improve data quality. In this proposal, we go beyond the state-of-the-art by collecting the data with a guaranteed quality degree and involving an expert to repair invalid data for a better quality. We will realize our approach on scientific data [31] (scientists share experimental results), in which there is no similar work to our knowledge.

References

[1] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu, "Maximizing benefits from crowdsourced data," *CMOT*, pp. 257–279, 2012.

[2] Q. V. H. Nguyen, T. T. Nguyen, N. T. Lam, and K. Aberer, "Batc: a benchmark for aggregation techniques in crowdsourcing," in *SIGIR*, 2013, pp. 1079–1080.

[3] Q. V. H. Nguyen, T. Nguyen Thanh, T. Lam Ngoc, and K. Aberer, "An evaluation of aggregation techniques in crowdsourcing," in *WISE*, 2013, pp. 1–15.

[4] (2015, August) Twitter. [Online]. Available: http://www.twitter.com

[5] (2015, August) Tumblr. [Online]. Available: http://www.tumblr.com

[6] (2015, August) Wikipedia. [Online]. Available: http://wikipedia.org

[7] (2015, August) Crowdflower. [Online]. Available: http:/crowdflower.com

[8] P. Meier, "Verifying crowdsourced social media reports for live crisis mapping: An introduction to information forensics," iRevolution.net, Tech. Rep., 2011.

[9] N. Q. V. Hung, S. Sathe, D. C. Thang, and K. Aberer, "Towards enabling probabilistic databases for participatory sensing," in *CollaborateCom*, 2014, pp. 114–123.

[10] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, "People on drugs: credibility of user statements in health communities," in *KDD*, 2014, pp. 65–74.

[11] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *ICWSM*, 2013, pp. 400–408.

[12] N. Q. V. Hung, H. Jeung, and K. Aberer, "An evaluation of model-based approaches to sensor data compression," *TKDE*, pp. 2434–2447, 2013.

[13] T. T. Nguyen, Q. V. H. Nguyen, M. Weidlich, and K. Aberer, "Result selection and summarization for web table search," in *ICDE*, 2015, pp. 231–242.

[14] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," in *VLDB*, 2009, pp. 562–573.

[15] Q. V. H. Nguyen, S. T. Do, T. Nguyen Thanh, and K. Aberer, "Privacy-preserving schema reuse," in *DASFAA*, 2014, pp. 234–250.

[16] A. Gal, M. Katz, T. Sagi, M. Weidlich, K. Aberer, H. Q. V. Nguyen, Z. Miklós, E. Levy, and V. Shafran, "Completeness and ambiguity of schema cover," in *CoopIS*, 2013, pp. 241–258.

[17] A. Gal, T. Sagi, M. Weidlich, E. Levy, V. Shafran, Z. Miklós, and N. Q. V. Hung, "Making sense of top-k matchings: A unified match graph for schema matching," 2012, p. 6.

[18] P. Ernst, C. Meng, A. Siu, and G. Weikum, "Knowlife: a knowledge graph for health and life sciences," in *ICDE*, 2014, pp. 1254–1257.

[19] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, pp. 993–1022, 2003.
[21] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[22] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *Nature*, pp. 440–442, 1998.
[23] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, pp. 215–239, 1979.
[24] N. T. Tam, D. C. Thang, N. Q. V. Hung, and K. Aberer, "An evaluation of diversification techniques," in *DASFAA*, 2015, pp. 215–231.
[25] D. T. Anh, V. H. Tam, and N. Q. V. Hung, "Generating complete university course timetables by using local search methods." in *RIVF*, 2006, pp. 67–74.
[26] N. Q. V. Hung, N. T. Tam, C. V. Tuan, T. K. Wijaya, Z. Miklos, K. Aberer, A. Gal, and M. Weidlich, "Smart: A tool for analyzing and reconciling schema matching networks," in *ICDE*, 2015, pp. 1488–1491.
[27] N. Q. V. Hung, D. C. Thang, M. Weidlich, and K. Aberer, "Erica: Expert guidance in validating crowd answers," in *SIGIR*, 2015, pp. 1037–1038.
[28] N. Q. V. Hung, D. C. Thang, M. Weidlich, and K. Aberer, "Minimizing efforts in validating crowd answers," in *SIGMOD*, 2015, pp. 999–1014.
[29] H. Q. V. Nguyen, T. K. Wijaya, Z. Miklós, K. Aberer, E. Levy, V. Shafran, A. Gal, and M. Weidlich, "Minimizing human effort in reconciling match networks," in *ER*, 2013, pp. 212–226.
[30] Q. V. H. Nguyen, X. Luong, Z. Miklos, T. Quan, and K. Aberer, "Collaborative schema matching reconciliation," in *CoopIS*, 2013, pp. 222–240.
[31] K. Aberer, A. Boyarsky, P. Cudré-Mauroux, G. Demartini, and O. Ruchayskiy, "Sciencewise: a web-based interactive semantic platform for scientific collaboration," in *ISWC*, 2011.