A Semantic Approach for Quality Control in Crowdsourcing

lalit.mohan, Prof. Raghu Reddy

March 2016

1 Abstract

Human intelligence tasks in crowdsourcing platform include responses to questions, image recognition and language translation. The quality of responses ensure sustainability of crowdsourcing. Most of the crowdsourcing platforms use majority voting and rating for assessing the quality of responses; this may not ensure responses are complete and consistently correct. Completeness, Correctness and Consistency in Software Requirement Engineering ensure quality control. We propose an approach for quality control of answers to questions using an enriching domain ontology and identify contributing gurus on a crowdsourcing platform. The approach would be validated for Information Security domain and generalized for other domains.

2 Introduction

The ability to collaborate across the world is increasing with increasing adoption of internet, reflecting in greater than 60% growth rate of crowdsourcing. Crowdsourcing platforms like Amazon MTurks, Crowdflower, oDesk, Freelancer, StackExchange, Quora, etc are catering to variety of tasks that includes image recognition, language translation, software development, design, testing, question and answers, etc. The motivation to collaborate in crowdsourcing could be extrinsic and intrinsic, this could determine the quality of output[2]. For quick money or due to lack of complete and correct knowledge, the responses suffer in quality. This is particularly the case for knowledge-intensive tasks that demand verbal production in a coherent way, e.g. 'information gathering tasks', tasks like gathering information from different documents given some question or problem definition, putting it together and producing a coherent verbal answer. The current quality assessments use peer review/two-man rule, continual rating of responses, majority voting, pre-assessment, plagiarism,etc techniques in a manual or semi-automatic way. However, the approaches are not complete [4] and there are still quality issues in crowdsourcing. Also, seeking responses for already responded question(includes semantically same) that is semantically same is non-motivating, leads to inconsistency in responses and loss of time.

Software engineering attributes for quality control include Completeness, Correctness and Consistency. Our research attains quality of responses for a question that are complete and consistently correct. To validate our approach, we plan to build a crowdsourcing system using open source technologies. Our approach contains 2 major parts, pre-processing for building an enriching domain ontology, web credibility of contributors and processing for quality control of responses. Information Security domain is considered for demonstrating the implementation. Information Security domain is well defined owing to ISO 27000 series, COBIT 5, standards and guidelines from NIST[3], ENISA[1], etc. The researcher is involved in setting up a platform(IB-CART) for banks to share security incidents anonymously similar to FS-ISAC in US. This implementation has strengthened the motivation on Information Security and also extend the idea of crowdsourcing.

3 Why 3Cs(Consistency, Completeness and Correctness) ?

Software Engineering brings process rigor of meeting requirements. In Software Requirements Engineering, requirement is analyzed for completeness and consistency and validated for correctness. Consistency and Completeness can be defined in syntatic or semantic terms using formal methods. Consistency would mean there are no contradictions and free from undesired nondeterminism. This would mean there exists at least one possible situation when all sentences are true. If sentences are true and false, they are said to be contradictory. An example of self contradictory sentences are firewall is used for applying access controls. Therefore, database can be used for role based access control. An example of consistent sentence firewall is used for applying access controls. Therefore, it can be used for information security. Completeness would mean a response for which it is intended. An example of completeness, confidentially, integrity and availability are characterisitics of information security. The sentence is semantically incomplete, if one of the characterisitics is missing. The sentence is syntatically incomplete if there is a missing subject and/or verb, unfinished sentences with no proper punctuation and other gramatical reasons. The software solution is validated for correctness, a degree to which entity's behavior matches specification of a requirement. An example of correctness, authentication and authorization are used for implementing user security. An example of incorrrectness, authentication is used for providing add/update/delete access based on role. In our approach, "Question" on a Crowdsourcing platform is mapped to a software "requirement" from a requester. To meet this requirement, 3Cs (Consistency, Correctness and Completeness) are attributed for quality control of response.

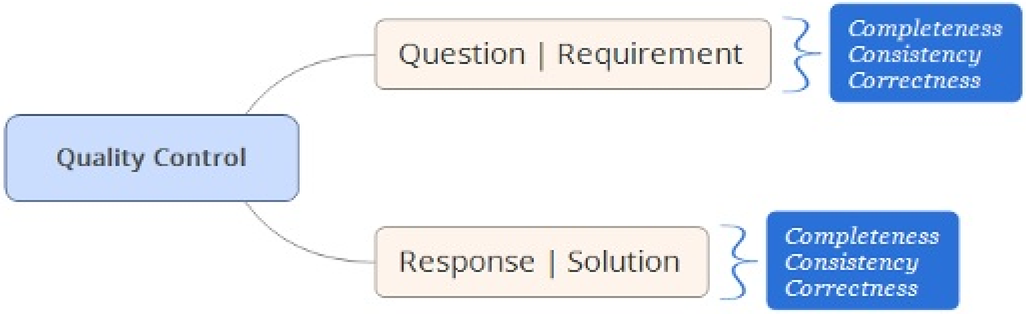

Figure 1: 3Cs of Quality Control

4 Approach

We describe the pre-processing stage and the processing stage for quality control of crowdsourcing question and answer platform. As part of the research, pre-processing is in advanced stage of completion.

Pre-processing stage : The public sites are crawled for obtaining domain specific content. The seed urls are obtained from open-content directory of World Wide Web links like DMOZ, Freebase and Wikipedia. With the obtained seed URLs and content available from OWASP site and ISO 27000 series, the classifier is trained for Information Security domain. We evaluated various java based open source crawlers and have shortlisted crawler that has better extraction rate, allows multi-threading, honors robot.txt, good community support, etc including other non-functional requirements such as maintainability, usability, documentation, extensibility, etc. The content crawling is in progress and the following activities as lemmatization, classification and the related metrics are yet to be implemented. HTML tags removal, stop words(“the”,”is”,”at”,”on”,etc) removal and lemmatization(“Information Security” as “IS”) after POS tagging would be applied to crawled content(html text, pdf and doc files) for classification. Support Vector Machines(SVM), Naive Bayes, Latent Dirichlet Allocation(LDA) , Decision Tree, etc techniques are explored for classifying crawled content. Precision, Recall, F-Score and Confusion matrix would be generated for the crawled content to identify the classifier that is most relevant for classification. This classified content can also be used for building an Information Security Search Engine (domain specific search engine)

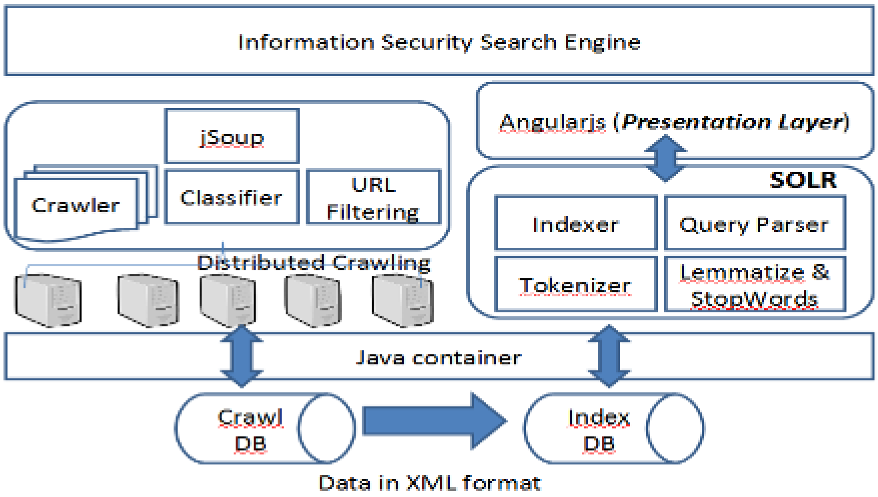

Figure 2: Information Security Search Engine

The classified content is used for building the domain ontology and identifing the crowd gurus. Ontology is the characterization of a domain, this includes formal representation of concepts and the relationship between the concepts. Ontology would be built for 114 controls in 14 groups. For faster parsing, the ontology would be represented in OWL(Web Ontology Language) format for each of the controls. An example of Ontology representation for cabling security is available in Figure 3

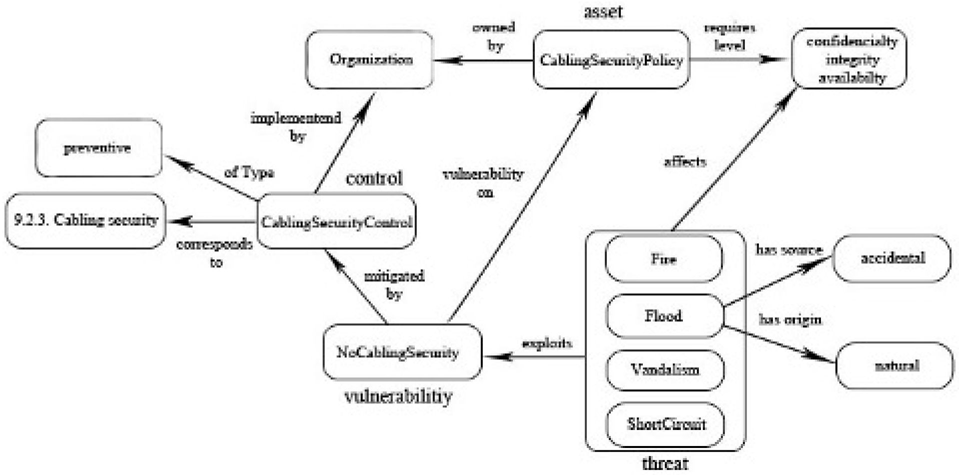

Figure 3: Domain Ontology

After POS tagging of the classified content, Named Entity Recognition using Stanford NER tool would be implemented for identifying individual names and organization names. Web credibility (Sonal and TweetCred) would be assigned to identify crowd gurus in each of the control areas of ISO 27001.

In the processing stage, when a person asks a question, the relevance of question for a domain will be validated. If the question is related to apparel, tourism, etc, an error stating that “question is not related to domain” is displayed. The validation is done based on the content in the question vis-a-vis information security domain ontology. The following question types would be validated for Quality Control – 1) Factual – responses are based on facts or awareness. E.g- When did Stuxnet attack surface ? 2) Convergent – responses are based on comprehension, material read/presented/known and basic level of cognitive thinking. E.g – What are the reasons for DDOS attack ? 3) Evaluative – responses require deeper cognitive thinking. E.g – Explain the differences between rule based access control and role based access control ? A semantic similarity of questions and answers would be validated using Similarity/Distance techniques like Latent Similarity Analysis. E.g "if there is already a response for "Provide a list of Phishing Patterns" and a new question comes up "What are the email lures because of which users are visiting a fake site and sharing account credentials".

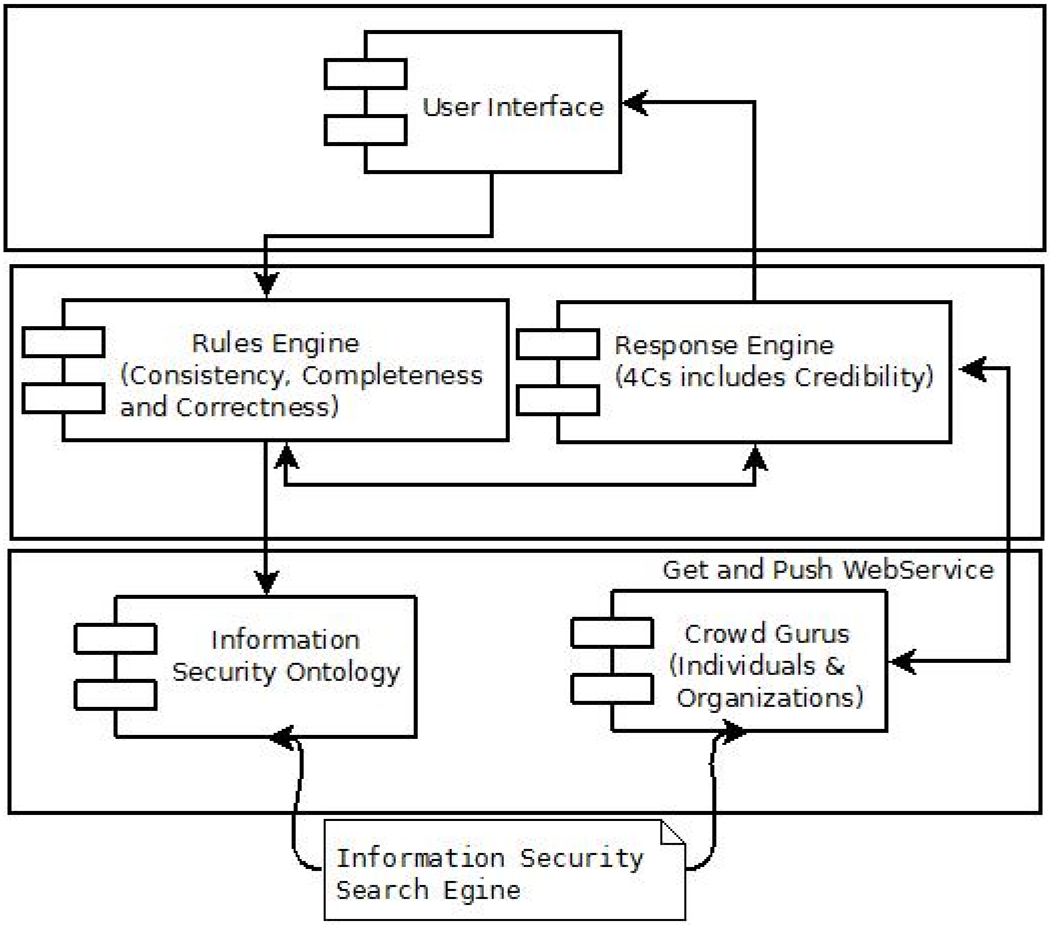

Figure 4: Quality Control - Component Diagram

If there are no semantically similar questions that already have an answer, the responses from crowd is sought. Quality control from consistency, completeness and correctness is applied to the obtained responses.

(QC) = aConsistency + bCompleteness + cCorrectness

Sentence consistency is checked against contradictions like negations in the response. Responses are also validated using Ontology to ensure that theren't multiple different answers for same question. If there aren't contridictions and no undesired undeterministic responses, the ontology depth and responder/crowd factor are considered as part of consistency calculation.

d(Consistency) = oOntologyClassDep + pPeopleContr

e(PeopleContr) = rPastContrOnCrowdPlat + sPastContrOnInternet

Completeness check for syntax/grammar is checked to ensure sentence is fully formed. Precision and Recall parameters provide the completeness of response with reference to domain ontology.

g(Completeness) = mPrec + nRecall

Correctness could be binary of yes/no. In the current research, we propose the degree to which the response meets the "question" requirement. Correctness also depends on the person responding to the question. A person with past credentials is more likely to provide a correct answer, this past credentials is already factored in Consistency check. The correctness check is obtained using Pseudo Relevance Feedback to obtain relevant responses.

e(Correctness) = oPseudoRelevanceFeedback

The weights for each of the factors shall improve based on the impact that each of the factors have on the response quality. Also, domain Ontology and Crowd gurus will change with increasing adoption and crawling for more content on information security.

5 Research Questions

Some of the research questions that require further analysis in the coming days are

1) Analysis of the quality control dimensions. A pertinent question is specific quality dimensions that can be addressed using semantic approach ?

2) Examine a) semantic repetition of the question/task and providing a response without bothering crowd workers b) possibility of giving specific feedback when certain components are missing in the answers produced. Or even c) to combine answers containing non-overlapping concepts in their answers.

3) Evaluation of improvements in quality control. How can we assess the proposed method of improvement - empirically or quantitatively?

4) Evaluate usefulness of results for other applications. Can results obtained be used in other applications such as recommender systems?

6 Approach validation

We propose two empirical approaches for validation of the research. Type 1: Same questions are posted on StackExchange, Quora and the new crowdsourcing platform. The responses are compared manually for completeness, consistency and correctness. Type 2 : Responses to the questions posted on the new crowdsourcing platform would be evaluated by information security experts of Indian Banking sector.

7 Conclusion

The research uses software engineering practice of treating question as "a requirement" and brings quality control attributes for completeness, consistency and correctness. The usage domain ontology, crowd gurus for quality control can be generalized for other domains as well. The content from social media sites may also be included for building ontology and identifying gurus and contributors of a domain. An approach for building domain specific search engines is another output of this research.

References

[1] ENISA. Network and information security guidelines, 2016.

[2] Kaufmann Nicolas, et al. More than fun and money. worker motivation in crowdsourcing – a study on mechanical turk, 2011.

[3] NIST. Computer/cyber/information security guidelines , recommendations and reference materials, 2015.

[4] Ipeirotis Panos. *A Framework for Quality Assurance in Crowdsourcing.* San Val, 1995.