Acquisition of Learning Object Metadata Using Crowdsourcing

Marek LÁNI*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
mareklani@gmail.com

In past years the Web began to be used largely for education purposes. There are many TEL or Question Answer portals and web sites which are being used to gain knowledge and information. Many times these systems are designed not only to provide benefits for users but also to benefit from their users. We can say, that it is a win-win relationship. It is because the content of these systems is often crowdsourced, it means generated by users themselves, so these systems build up or enlarge their knowledge base thanks to crowdsourcing approach. Typical example of such a system is a QA community portal StackOverflow[1].

But as it is not guaranteed, that people who create content within these systems are experts in a specified area, it is necessary to evaluate quality of their contributions. This quality validation should be automated for reason of its time consumption. But the question is: How?

Some approaches are trying to analyze semantics of questions and corresponding crowdsourced answers, in order to determine whether the answer is relevant or not.

We think that there exists better method to determine quality of user generated content and it is usage of "second level" crowdsourcing, where not only the content is being generated by users, but also the act of evaluation is outsourced to the users. The evaluation should be done in form of ratings. However it is not as simple as it seems to be. Many research works defined problem, that non filtered rating evaluation is not giving us satisfying results [1][2]. The obtained raw rating data also suffer from certain degree of imperfection based on non-expert contributing. The usage of next iteration of crowdsourcing would be meaningless so we have to analyze, filter and weight obtained rating data.

In our work we are focused on QA component of TEL system. In this system we have set of question-answer pairs which were collected during the examinations during the school year at subject Principles of Software Engineering. Every QA object has

---

* Supervisor: Jakub Šimko, Institute of Informatics and Software Engineering

[1] www.stackoverflow.com

Spring 2013 PeWe Workshop, April 5, 2013, pp. 77–78.

assigned teacher's evaluation. This QA component allows students to study before examinations in form of reviewing the QA pairs and assigning the rating to answers. We want to take this data and analyze them in order to determine if the crowd is able to agree with teacher on assigned rating.

There are several approaches how to filter and weight user ratings. Aim of our work is to take, combine and modify some of them to achieve satisfactory results in evaluation of answers. We plan to use some of the following approaches:

- **Spam filtering**

  Agichtein [1] defined two main reasons why there is spam within the rating data. One reason is insufficient knowledge and the second is malicious activity. In our case, within the collected ratings, there can be outbound ratings obtained from users, who have only a little knowledge or no knowledge about the questioned subject at all. As for the every rating the user is awarded by activity points, there can be produced also malicious rating activity in order to gain points. We want to detect some general patterns of named behaviors and ignore the ratings which were produced by behavior like this.

- **Weighting of rates and expert determination**

  According to Chen et al.[2] a vote calibration can bring large improvement in answer relevance evaluation. We aim to analyze users' profiles in order to determine their expertise level. This will help us to estimate probability of correctness of user's ratings. We will also take in account the frequency and count of user's votes when designing the weighting respectively calibration mechanism.

We want to experiment with these weighting parameters and filtering in order to achieve the best results in crowdsourced answer relevance evaluation and to determinate whether the crowd is capable of self evaluation or not.

To enlarge our dataset and mainly the count of ratings, we want to create interface of our QA component for mobile devices. We also think about usage of probabilistic approach of inferring the missing ratings via Probabilistic Matrix Factorization.

We plan to evaluate our methods of filtering and weighting of ratings produced by users, by comparison to the ratings assigned to the answers by experts - teachers.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Agichtein, E.: A Few Bad Votes Too Many ? Towards Robust Ranking in Social Media. In.: AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web, ACM, 2008, p.p. 53–60.

[2] Chen, B.-C., Dasgupta, A., Wang, X., & Yang, J.: Vote calibration in community question-answering systems. In: SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012, p.p. 781-790.