Questionnaire generation for better opinion polling

submitted for blind review

No Institute Given

Abstract. This paper studies a novel problem. Given a topic, how to automatically generate an opinion poll from a question bank collected from online debate forums, so that the questionnaire can best predict public opinion on the topic? This is not a trivial task, as even opinion polls composed by human experts are not always reliable. We propose a wording quality aware framework which selects a set of questions that cover all aspects of the topic and achieve maximal wording quality. Two challenges arise. (1) How to measure the quality of wording so that question alternatives can be compared? (2) How to match question alternatives? We further present statistical technologies to address these challenges. Both quantitive and qualitative experimental results demonstrate the competency of our methods.

1 Introduction

Opinion poll is not uncommon in modern society. The first known example of an opinion poll was conducted nearly 200 years ago[1]. It successfully predicted the outcome of United States presidency in 1824. Since then, it has been a convention to organize an opinion poll to measure public opinions and document the experiences of the public on a range of subjects, from your favorite video games in market to important health and political issues. Polling results provide information for academics, researchers and government officials and help policymakers and other decision makers.

In an opinion poll, a series of questions are handed out to sample voters from all population. The goal of an opinion poll is to extrapolate generalities based on the answers to the questions by the sample voters. Opinion polling is conventionally conducted by professionals with expertise. There are polling organizations in virtually every country with elections. Furthermore, all the major television networks, independently or working in conjunction with newspapers or magazines, operate opinion polling.

Recently, online social media has attracted researchers to study public opinions [2,1]. Many social media sites encourage the creation and exchange of ideas and opinions. In particular, a debate forum is a bank of simple agree/disagree scale questions, with many users voting and commenting on one option. Thus debate forums are valuable platforms to facilitate public opinion mining.

[1] https://en.wikipedia.org/wiki/Opinion_poll

In this paper, we explore the possibility of generating an opinion poll automatically from a question bank for any given topic. An illustrative example of such a desired system is shown in Fig. 1. Naturally, the accuracy of an opinion poll is highly affected by the design of questionnaire. It is well established that "asking enough questions to allow all aspects of an issue to be covered" is critical in composing the opinion poll [2]. For example, the output in Fig. 1 covers all relating issues of the given topic "drug use", including "prescription drug", "marijuana" and so on. Also, the wording of each question must be carefully selected and organized so that the true feelings of respondents are revealed. Small changes of words in a question can lead to big changes in the answer. For example, in the toy example, it is difficult for teens to ?obtain? marijuana, but it is easy for teens to "buy" marijuana[3].

However, it is not a trivial problem to get the wording of a questions right. Even opinion polls generated by experts are not always reliable in predicting public opinions on an unseen topic. For instance, Donald Trump's election in 2016 signals the biggest and complete poll failure ever.

Input: question bank	Output: opinion poll
1. Is it difficult for someone your age to buy marijuana? – yes: 39, no: 61 2. Is it easy for someone your age to obtain marijuana? – yes: 39% no: 61 3. Is it easy for someone your age to find marijuana? – yes: 40% no: 60 4. Have you abused prescription drugs? – yes: 50% no:50%	– Is it easy for someone your age to obtain marijuana? – Have you abused prescription drugs?

Fig. 1. An illustrative example of opinion poll generation

There are two challenges raised in the procedure of picking the appropriate question. (1) How to match question alternatives? (2) How to measure the quality of wording so that question alternatives can be compared?

As debate forums consist of user generated content(UGC), there could be numerous questions regarding the same issue, but with different wording and options. The first challenge is to match question alternatives. It involves two steps: (1) identify question alternatives that are similar (2) pair up options.

[2] `https://en.wikipedia.org/wiki/Opinion_poll`

[3] The toy example is a modified version of 2003 survey of teens and drug uses

For example, the first, second and third questions in the left part of Fig. 1 are equivalent as they are all about the ”obtain” of marijuana. The ”yes” option in question 1 should be paired up with the ”no” option in question 2 and 3, and vice versa.

Suppose we have a set of question alternatives, which one is the best to predict public opinion? In the literature of social research, there are general guidelines [3] for asking effective questions. However, these rules are objective and indirect, while a practical measure should be subjective and direct. Intuitively, if a problem is worded nicely to reflect the true public opinion, the resulting output must accord with most predictions of other question alternatives. For example, in the right part of Fig. 1, we choose question 2 because it agrees with most predictions in the first issue (obtaining marijuana).

Our contributions in this paper are three fold. (1) We study the problem of opinion poll generation for a given topic from a question bank collected from online debate forums. To the best of our knowledge, this is a novel task. We present a wording quality aware framework to generate a set of questions that reflect all related aspects of the topic while maximize the wording quality. (2) We define a score function to measure the quality of wording by calculating how much the prediction of a question accord with other predictions on the same aspect. (3) We address the matching of question alternatives by applying a combination of NLP techniques. We conduct quantitive and qualitative experiments on real data sets. The experimental results demonstrate that our proposed method is competent.

This paper is structured as follows. We briefly introduce related work in Sec 2. We describe details of our framework and algorithms in Sec 3 and Sec 4. The experimental results are presented and analyzed in Sec 5. We give our introduction and a outlook for future work in Sec. 6.

2 Related Work

2.1 Mining Public Opinion

http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842
A user-centric model of voting intention from Social Media

2.2

3 Framework

It is well regarded that the selection, ordering and wording of questions are critical []. For simplicity, we don’t consider the It is well established that asking enough questions to allow all aspects of an issue to be covered.

4 Evaluate Wording Quality

5 Experimental Analysis

5.1 Experimental Setup

The data set used in the experiment is obtained from xxx. We crawl xxx from xxx. The statistics is shown in Tab. It contains

Table 1. Statistics of data set

Number of queries	6777	Avg. length of query	12.22
Number of comments	28911	Avg. length of comment	16739
Number of users	16739	Avg. votes per user	1.73

5.2 Performance of Matching Question Alternatives

5.3 Evaluating the wording quality

[3]

5.4 Case Study

6 Conclusion

References

1. V. Lampos, D. Preotiuc-Pietro, and T. Cohn. A user-centric model of voting intention from social media. In *Proceedings of ACL*, pages 993–1003, 2013.
2. B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.
3. W. M. Trochim. *The Research Methods Knowledge Base.* Atomic Dog Publishing, Cincinnati, OH., second edition edition, 2000.