Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Crowdsourcing Objective Answers to Subjective Questions Online

Ravi Iyer
Ranker

Abstract

In this demonstration, we show how Ranker's algorithms use diverse sampling, measurement, and algorithmic techniques to crowdsource answers to subjective questions in a real-world online environment where user behavior is difficult to control. Ranker receives approximately 8 million visitors each month, as of September 2013, and collects over 1.5 million monthly user opinions. Tradeoffs between computational complexity, projected user engagement, and accuracy are required in such an environment, and aggregating across diverse techniques allows us to mitigate the sizable errors specific to individual imperfect crowdsourcing methods. We will specifically show how relatively unstructured crowdsourcing can yield surprisingly accurate predictions of movie box-office revenue, celebrity mortality, and retail pizza topping sales.

Online Crowdsourcing

Opinion aggregation is a billion dollar online business that includes both business-to-business market research and direct to consumer websites such as TripAdvisor, which aggregates opinions about hotels, and Yelp, which aggregates opinions about local businesses. Unlike many opinion aggregators, Ranker's platform is not domain specific and encompasses opinions about entertainment ("most anticipated movies"), sports ("best nba players"), books ("books that changed my life"), the future ("celebrity death pool"), and many other domains. As interest in aggregated opinions grows (Iyer, 2013a), traffic to Ranker's lists have grown as well, giving Ranker a steady stream of consumer opinions about a variety of topics.

The Wisdom of Ranker Crowds

Ranker's algorithm to aggregate user opinions is based on the Wisdom of Crowds principle of aggregating as many diverse sources of "signal" as possible (Larrick, Mannes, & Soll, 2010), in the hopes that the error specific to any individual source of variance cancels out when aggregated. Visitors to Ranker provide their opinions in several different ways. The most common interaction is for visitors to vote on particular items on a list, ostensibly based on whether the item belongs on the list (e.g. is The Godfather one of the best movies of all time?), but our analyses have shown that items do get downvoted more as they move higher on lists, indicating that votes are based both on whether an item belongs on a list and on whether an item belongs at the specific position on a list. Since a registration is not required for voting, there are relatively low barriers to engagement, which enables Ranker to achieve 10-15% engagement on votable lists with voters voting on 10-15 items. These opinions tend to be "shallow" compared to other inputs, and reflect mass opinion as opposed to the opinions of particularly knowledgeable individuals.

The other input that Ranker collects from users comes in the form of "reranked" user lists where a user makes their own version of a particular list. These lists may be created from scratch or may be based on existing lists, with the order changed. The barrier to entry is higher, and so these lists tend to be made by individuals who have greater knowledge of the domain. As such, while we receive far fewer user lists, compared to user votes, we weigh user lists heavily in our algorithms in an attempt to reach a balance between expert and non-expert opinions, both of which have some amount of uncorrelated error. Further, we consider both the list position of an item, which indicates the level of passion that a user has for an item, and the number of times an item appears on user lists, which is indicative of the popularity that an item has among knowledgeable individuals. In keeping with the principle that diversity of measurement leads to better results, we feel that the combination of "shallow" voting behavior, "expert" popularity, and "expert" passion yields better algorithmic results than any particular single measure of user opinions.

External Validation

In order to validate Ranker's algorithms and methodologies, we have sought to compare our results to those of other organizations. For example, our list of "Top Dream Colleges" matches Gallup's list of best colleges quite closely, at a fraction of the cost, and with greater depth (Iyer, 2012). Our list of the "Best Movies of All Time" has fewer older films, compared to the American Film Institute's Best Movies list, but has more critically acclaimed films than Rotten Tomatoes Best Movies list, indicating that we are succeeding in bridging the divide between "expert" and "mass" opinions.

In some cases, our results can also be compared to real world metrics. For example, a common prediction market task is to attempt to predict the box office success of movies based on twitter activity, Wikipedia activity, or prediction market activity. A similar analysis of Ranker data showed that votes on our "most anticipated movies of 2012" list predicted 82% of the variance in opening weekend box office success (Iyer, 2013b), which was comparable in accuracy to other techniques, but potentially more pragmatically useful given that Ranker data was generally collected 6-12 months before the actual release date of the movie.

While the topic is rather grim, analyses of "celebrity death pool" lists which attempt to predict which celebrities will pass away in the next year, indicate that our aggregated results perform better than all but one of the 27 individual prediction lists created, in predicting which celebrities actually pass away (Lee, 2013a). Similarly, an aggregation of our list of the "tastiest pizza toppings" conformed to reported pizzeria topping sales by a New York pizza chain better than all but one of the 29 user lists made on the Ranker platform (Lee, 2013b). In summary, aggregated Ranker lists have proven predictive of real world outcomes.

Future Directions

Ranker's ultimate success will likely be determined by the ability to crowdsource superior answers to questions, as compared to experts who are constantly attempting to answer similar questions in blog and news articles. As such, we are constantly attempting to improve our algorithms by capturing new diverse sources of data. We are currently using IP address and referrer information to help identify systematic bias introduced by fans of particular items on lists and correct accordingly. We are currently partnering with outside organizations that are willing to collect data on their websites through a Ranker widget, which increases engagement on their site, and hope that data from widget partners will eventually provide a useful check on sampling bias inherent in collecting data solely on Ranker.com. We are also currently partnering with behavioral targeting companies to segment our audience and determine list items that are popular among a diverse group of visitors. Highly ranked items should theoretically be ranked highly by both male and female, young and old, American and international visitors. By layering in third party data, we can verify the robustness of user opinions and our rank order.

In Summary

Crowdsourcing in an online environment is necessarily prone to error and bias, as online opinion aggregators rarely have control over sampling. By leveraging diversity and aggregation, Ranker algorithms are able to create accurate rankings that are less prone to single source bias. By continually adding new sources of diverse data, with each source contributing different kinds of error, we are hopeful that the quality and credibility of our rankings will continue to grow.

References

Iyer, R. (2012). Validating Ranker's Aggregated Data vs. a Gallup Poll of Best Colleges. *Data.Ranker.com*. Retrieved September 4, 2013 from http://data.ranker.com/ranker-gallup-best-colleges-wisdom-of-crowds/.

Iyer, R. (2013a). Rankings are the Future of Mobile Search. *Data.Ranker.com*. Retrieved September 4, 2013 from http://data.ranker.com/rankings-are-the-future-of-mobile-search/.

Iyer, R. (2013b). Predicting Box Office Success a Year in Advance from Ranker Data. *Data.Ranker.com*. Retrieved September 4, 2013 from http://data.ranker.com/predicting-box-office-success-ranker-data/.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2011). The social psychology of the wisdom of crowds. *Frontiers of social psychology: Social psychology and decision making*, 227-242.

Lee, M. (2013ba). Recent Celebrity Deaths as Predicted by the Wisdom of Ranker Crowds. *Data.Ranker.com*. Retrieved September 4, 2013 from http://data.ranker.com/recent-celebrity-deaths-predicted-wisdom-of-crowds/.

Lee, M. (2013b). Combining Preferences for Pizza Toppings to Predict Sales. *Data.Ranker.com*. Retrieved September 4, 2013 from http://data.ranker.com/combining-preferences-for-pizza-toppings-to-predict-sales/.