HITS on Question Answer Portals:
Exploration of Link Analysis for Author Ranking

Pawel Jurczyk
Department of Mathematics and Computer Science,
Emory University

pjurczy@emory.edu

Eugene Agichtein
Department of Mathematics and Computer Science,
Emory University

eugene@mathcs.emory.edu

ABSTRACT

Question-Answer portals such as Naver and Yahoo! Answers are growing in popularity. However, despite the increased popularity, the quality of answers is uneven, and while some users usually provide good answers, many others often provide bad answers. Hence, estimating the *authority*, or the *expected quality* of users, is a crucial task for this emerging domain, with potential applications to answer ranking and to incentive mechanism design. We adapt a powerful link analysis methodology from the web domain as a first step towards estimating authority in Question Answer portals. Our experimental results over more than 3 million answers from Yahoo! Answers are promising, and warrant further exploration along the lines outlined in this poster.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval

General Terms: Algorithms, Documentation, Experimentation.

Keywords: Question-answer portals, authority estimation, link analysis.

1. INTRODUCTION

Portals allowing users to answer questions posted by others (henceforth QA portals) are rapidly growing in popularity. The reason is that people can share their knowledge with others easily and can find answers for both common and unique questions. Some popular QA portals include Naver (http://www.naver.com) and Yahoo! Answers (http://answers.yahoo.com). Yahoo! Answers already reports millions of users and over 60 million answers to over 5 million distinct questions and is growing rapidly. Unfortunately, with the increase of popularity, the *quality* of answers is uneven. While user feedback features such as voting for the best answers provide valuable information *when available*, the user ratings of answers are relatively sparse, with fewer than 35% of closed questions having a user rating for any of the answers. Hence it is becoming increasingly important to automatically *estimate* the authority of users that post the answers on such QA portals without relying exclusively on user feedback.

While authority of *pages* on the web has been an active area of research, estimating authority of *authors* in collaborative portals such as Yahoo! Answers is an open question. Previously, Jeon et al. [1] evaluated various features such as author's activity, number of clicks on answers and average length of posts for finding the *best*

answers for a given question. In contrast, we focus on estimating the *authority* of *users* that can be potentially used for ranking answers, finding "experts", incentive mechanism design (to reward authoritative users), and spam detection. As an initial approach to user authority estimation, we explore the HITS link analysis algorithm [2]. We then describe the evaluation of our method over a large crawl of Yahoo! Answers data and discuss preliminary results.

2. ADAPTING HITS FOR QA PORTALS

The HITS algorithm [2] was developed to predict importance of web-pages by assigning each page a *hub* and *authority* value. A page is considered a good hub if it links to authoritative pages, and authoritative pages are in turn linked by good hubs. This idea has an intuitive parallel for QA portals. Specifically, we can consider *question authors* as hubs, and *answer authors* as authorities. This idea is illustrated in Figure 1. In the bipartite graph in Figure 1, QA portal users Q1, Q2, and Q3 have posted questions answered by users A1, A2, A3, and A4. For each user we calculate both the hub and authority value. Intuitively, our approach gives high hub value to users posting good questions – if a question is good, it will be answered by experts in particular area. On the other hand, poor and not well formulated queries will not be answered, and users posting them will have low hub value.
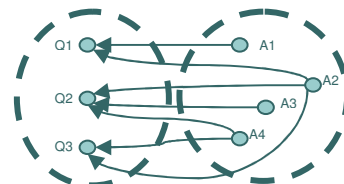

Figure 1: Graph representation of question and answer users.

Specifically, our algorithm calculates the hub and authority value of all users using Kleinberg's HITS algorithm formulation [2]:

$$H(i) = \sum_{j=0..K} A(j) \qquad A(j) = \sum_{i=0..M} H(i)$$

where $H(i)$ is the hub value of each user i from set of K users posting questions and $A(j)$ is the authority value of each user j from set of M users posting answers. The vectors H and A are initialized to all 0s and 1s respectively, and are updated iteratively using the equation above. After each iteration the values in the H and A vectors are normalized, so that the highest hub and the highest authority values are 1. This algorithm iterates until convergence, defined as when the total sum of changes in both hub and authority values becomes less than some small ε (0.001 in our experiments).

2.1 Implementation and Experimental Setup

To obtain the data for experiments we crawled a large portion of the Yahoo! Answers QA portal, obtaining the total of 495,099 questions and corresponding 3,252,345 answers in three general categories: *Science*, *Sports*, and *Arts & Entertainment*. The dataset statistics are

reported in Table 1. The crawler was based on the WebSphinx framework[1]. For each category, we recursively traversed the subcategories, with up to 1,000 question pages retrieved (Yahoo! Answers does not expose additional question pages to external requests). The crawler then parses the question pages to detect answers and stores both questions and answers in a relational database for easy access and subsequent analysis.

Category	Questions	Answers	Users	Answers per Question
Science	225,750	1,469,207	197,773	6.5
Sports	136,824	1,046,411	142,349	7.6
Arts & Entertainment	132,525	736,727	117,608	5.6
Total	**495,099**	**3,252,345**	**457,730**	**6.6**

Table 1: Yahoo! Answers dataset statistics

3. EXPERIMENTAL EVALUATION

We now present the experimental evaluation of our authority estimation method over the Yahoo! Answers datasets described in Table 1. We compare our method (**HITS**) with the number of posts per user. Frequent posters tend to have significant interest in the topic, and number of posts was shown to correlate with answer quality [1]. Hence, we will use the number of posts (**Frequency**) as a baseline for authority estimation.

3.1 Evaluation Metrics

To evaluate the accuracy of our methods, we use the feedback of users. Yahoo! Answers, as do other QA portals, provide a mechanism for users to provide feedback for each answer. We observed that high quality (authoritative) users tend to post answers that are popular (via the "thumbs up" and "thumbs down" user voting mechanism) or, alternatively, obtain high ratings from the original question posters (via the "stars" rating for the best answer). Following these observations, we define two possible "gold standard" quality scores for each author as follows:

- **Votes:** number of positive votes minus negative votes combined with historical rank that an author received from *other users*, averaged over all answers attempted.
- **Stars:** the average number of stars an author obtains when their answer is selected as the "best answer" by the original question posters over all answers attempted.

To evaluate the authority scores computed by our methods, we rank the authors in decreasing order by their scores, and compare our ranking with the ranking of users ordered by their *Votes* and *Stars* values. Specifically, we use the Pearson correlation coefficient**:**

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

where the *x* values are the ranks of users according to our authority estimation method, and *y* are the ranks of users according to the *Votes* or the *Stars* user feedback scoring.

3.2 Experimental Results

We first present the results for all question categories (Figure 2). Figure 2(a) reports the Pearson's correlation for the top *K* users ranked by the *HITS* and *Frequency* algorithms respectively, compared with their ranks ordered by the *Stars* values. Figure 2(b) reports the correlation for the top *K* users ranked by the *HITS* and *Frequency* algorithms with their *Votes* values. *HITS* correlates more

strongly with the *Stars* than with the *Votes*, but in both cases the top authorities are indicated by the *HITS* algorithm more accurately than by using the *Frequency* scores.
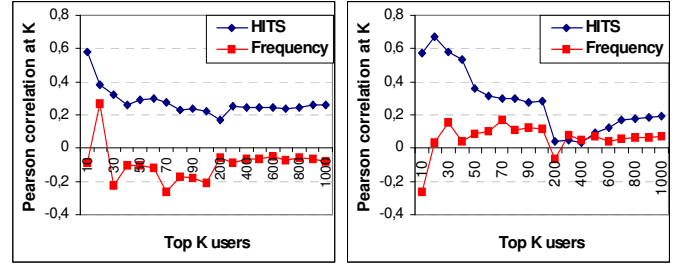

Figure 2: Pearson correlation at *K* for *HITS* and *Frequency* vs. *Stars* (a) and vs. *Votes* (b) for all question categories.

We now consider estimating user authority in a particular category, for example for answering *Science* questions. We report the results in Figure 3. We hypothesized that authority is easier to estimate within a particular domain. Interestingly, the estimated authority for users across *all* categories correlates about as well with both *Stars* and *Votes* scores as authority in the *Science* category alone.
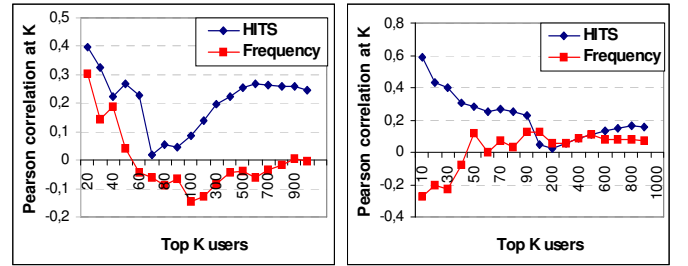

Figure 3: Pearson correlation at *K* for *HITS* and *Frequency* vs. *Stars* (a) and vs. *Votes* (b) for the *Science* question category.

4. CONCLUSIONS AND FUTURE WORK

We presented a first step towards link-based authority estimation in collaborative sources such as question answer portals. We evaluated our methods by comparing our estimated scores with measures obtained derived from explicit user feedback. In fact, our *Votes* and *Stars* scores are instances of general feedback mechanism design, where the former allows "popularity" feedback from all users, whereas the latter supports "quality" feedback from the user posing the original question. Understanding the drawbacks and advantages of each feedback mechanism is the subject of future work. While our results are promising and provide better ranking of user authority than the simple frequency baseline, our methods are not effective for the users ranked below the top few "experts". In the future, we plan to combine content features with link analysis, and to perform more extensive experiments on understanding the performance of link analysis algorithms on more fine-grained domain categories. In summary, our exploration of link analysis for estimating the authority of users in Question Answer portals suggests a promising direction for continued future research.

5. REFERENCES

[1] J. Jeon, W.B. Croft, J.H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. *In Proc. of* SIGIR 2006

[2] J. Kleinberg, Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999

---

[1] http://www.cs.cmu.edu/~rcm/websphinx/