Discovering Authorities in Question Answer Communities by Using Link Analysis

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 quickly becoming rich sources of knowledge on many topics which are not well served by general web search engines. Unfortunately, the quality of the submitted answers is uneven, ranging from excellent detailed answers to snappy and insulting remarks or even advertisements for commercial content. Furthermore, user feedback for many topics is sparse, and can be insufficient to reliably identify good answers from the bad ones. Hence, estimating the authority of users is a crucial task for this emerging domain, with potential applications to answer ranking, spam detection, and incentive mechanism design. We present an analysis of the link structure of a general-purpose question answering community to discover authoritative users, and promising experimental results over a dataset of more than 3 million answers from a popular community QA site. We also describe structural differences between question topics that correlate with the success of link analysis for authority discovery.

Categories and Subject Descriptors:

H.3.3 Information Search and Retrieval

General Terms: Algorithms, Design, Experimentation.

Keywords: Question-answer portals, 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, and can find answers for both common and unique questions. Some of these information needs are too specific to formulate as web search queries, or the content simply does not exist on the web. Other users seek opinions of the community, or are not adept at searching the web and would prefer other people to help them find the relevant information. Some popular QA portals include *Naver* and *Yahoo! Answers*. All non-abusive answers later become available for search and retrieval. Since going live relatively recently, *Yahoo! Answers* attracted millions of users and over 100 million answers for more than 20 million questions.

Unfortunately, the quality of the submitted answers is uneven, ranging from excellent detailed answers to snappy and insulting remarks or even advertisements for commercial content. Therefore, it is increasingly important to better understand the issues of authority and trust in such communities, which differ drastically from previously studied online communities both in types of interactions that are available to users, and the content of the sites. QA portals provide many mechanisms for community feedback. When a question author chooses a best answer, he or she can provide a “quality” rating. Another measure of quality of answer are the “thumbs up” and “thumbs down” votes. Such community feedback is extremely valuable, but requires some time to accumulate, and often remains sparse for obscure or unpopular topics. In a large sample of the Yahoo! Answers portal that we analyzed, fewer than 35% of all questions had any user votes cast for any of the answers (as of Jan 2007). Therefore, it becomes important to estimate the authority of users without exclusively relying on user feedback. In particular, we attempt to discover *authoritative users* for specific question categories by analyzing the link structure of the community.

We present a large-scale study of the link structure of community question answering portal for discovering authorities in topical categories. In particular, we formulate a graph structure for the QA domain, and adapt a web link analysis algorithm for topical authority estimation (Section 3). We describe an experimental evaluation over a dataset of more than 3 million answers (Section 4), demonstrating the viability of our approach (Section 5). We summarize our findings in Section 6, which concludes the paper.

2. RELATED WORK

Link analysis has played an important role in bringing order to the web (e.g., [1, 2, 5]). The most common link analysis algorithms are PageRank and *HITS*. The *HITS* algorithm is based on the observation that there are two types of pages: (1) hubs which group links to authoritative pages and (2) authorities which are source of information on given topic. *HITS* assigns each page two scores, hub and authority value. A hub value represents the quality of outgoing links from the page while authority represents the quality of information located on that page. The PageRank algorithm does not distinguish between hub and authority pages, and instead estimates the likelihood that a random walk following links (and occasional random jumps) will visit a page. The algorithm has been extended to bias the jump probability for particular topics [2] and for many others static web ranking tasks.

We adapt link analysis techniques to community question answering. There has been much research focused on modeling communities and finding good users. Reputation in a community [8] and other questions have been addressed with link analysis

CIKM’07, November 6--8, 2007, Lisboa, Portugal.

methods. McCallum et al. [7] describe techniques for finding experts for particular topics using social network structure. However, the nature of question answer portals centered around *questions* results in structures that are different from the static web analyzed by the previous link analysis algorithms.

Closest to our work, Zhang et al. [10] evaluated link algorithms such as PageRank or *HITS* for expert finding in a closed domain. We focus on a web-scale general purpose question answer portal with millions of users, and hundreds of millions of postings and interactions, which introduces a variety of novel challenges. Also related to our work, Jeon et al. [3] and Liu et al. [6] evaluated answer features such as author's activity, number of clicks, 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 could be exploited for ranking, incentive mechanism design, and spam detection.

3. AUTHORITY IN QA PORTALS

To derive the link structure of the question answering community, we begin by representing the structure of a single question. As shown in Figure 1 (a), a particular question has a number of answers associated with it, represented by an edge from the question to each of the answers. We also include vertices representing authors of questions or answers. An edge from a user to a question means that the user asked the question, and an edge from an answer to a user means that the answer was posted by this user. In our example, User 2 has posted questions 2 and 3, but has never answered any questions, while User 5 has answered Question 1 and Question 2, but has never asked a question.

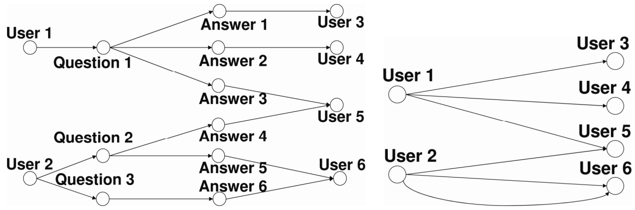

Figure 1. Representing users, questions, and answers (a) and summary multigraph representing user relationships (b).

We summarize the relationships between users in a *multigraph* shown in Figure 1(b). This graph contains vertices representing the users and omits the actual questions and answers that connect the users. Therefore, there may be multiple edges connecting the same users. For example, there are two edges from User 2 to User 6, as User 6 has answered two questions posted by User 2.

Note that we are only considering the links between user nodes and not past explicit feedback history for each user (e.g., total positive votes or best answer fraction received by the user in the past). Such history can be naturally modeled in our graph formulation as weights on the edges in Figure 1(b). We plan to explore this issue in more depth in future work. Also note that unlike in the original definition of HITS, in our model the authors create edges themselves by answering questions. In this case, the system might be susceptible to spamming. However, this problem is reduced with extensive user feedback and abuse reporting which are present in most of QA portals.

Authority Estimation: Consider the user relationship graph shown in Figure 1(b). Poorly formulated or nonsense questions will have few or no answers (resulting in low outdegree for the question user nodes), while good questions will tend to have many answers, often posted by experts in the subject area (resulting in high outdegree). In turn, users answering many questions from "good" users will have high indegree. This immediately suggests that nodes representing questions authors act as "hubs" while nodes representing answer authors correspond to "authorities." By considering *question authors* as hubs, and *answer authors* as authorities, our graph representation of user interactions can be used as input to the *HITS* algorithm [5] that computes the hub and authority values for web pages connected via hyperlinks. Intuitively, we may view the question category (e.g., *Science*) as a query topic, and calculate the hub and authority values of users:

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

where *H(i)* is the hub value of each user *i* from set of users *0..K* posting questions, and *A(j)* is the authority value of each user *j* from set of users *0..M* posting answers. The vectors *H* and *A* are initialized to all 0 and 1 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.

4. EXPERIMENTAL SETUP

To evaluate the accuracy of our methods we use the explicit feedback of users, when available. We observe that 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 three possible "gold standard" quality scores:

- ***Votes*:** number of positive votes minus negative votes combined with total percent of positive votes an author received from *other users*, averaged over all answers attempted.
- ***%Best*:** the fraction of best answers awarded to an author by asker over all answers attempted.
- ***Ratings*:** the number of stars an author obtains when their answer is selected as the "best answer" by the asker, averaged over all answers attempted.

To evaluate the authority estimation methods, we rank the users in decreasing order by their authority scores, and compare with the ranking of users by their *Votes*, *%Best*, and *Ratings* scores. Specifically, we use the Pearson correlation coefficient**:**

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

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 scores derived from the user feedback.

Methods compared: We compare two link-based methods:

• ***HITS***: Our method, described in Section 3.

• ***Degree*** (Baseline)**:** Frequent posters tend to have significant interest invested in the topic, and, as shown by Jeon et al.[3] and Zhan et al.[10],degree of a node correlates with answer quality.

Datasets: We crawled a large portion of the Yahoo! Answers QA portal, retrieving 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.

Table 1: Yahoo! Answers dataset statistics

Category	*Questions*	*Answers*	*Users*	*Avg. Answers*
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**

5. EXPERIMENTAL RESULTS

We first present the results for all question categories (Figures 2(a), 3(a) and 4(a)). Figure 2(a) reports the Pearson's correlation for the top *K* users ranked by the *HITS* and *Degree* algorithms, correlated with the *Ratings* metric defined above. Figure 3(a) reports the correlation for the top *K* users ranked by the *HITS* and *Degree* algorithms with the *Votes* metric. *HITS* correlates more strongly with *Ratings* scores than with the *Votes* scores, but in both cases up to top 30 authorities are indicated by *HITS* more accurately than by *Degree*. Figure 4(a) reports the correlation with the *%Best* scores for each user. As before, *HITS* correlates strongly with the *%Best* metric for up to top 40 authorities.

We now consider the case of discovering authority within a particular category, for example, the *Science* category. We report the results in Figures 2(b), 3(b) and 4(b). We hypothesized that authority is *easier* to estimate within a particular domain. Indeed, the correlation is significantly higher for the *Science* domain than overall, with up to 30 "experts" predicted by *HITS* significantly better than by *Degree*. These results are encouraging, as it is unlikely that a category would have more than 20 or 30 authorities, and so our method would be sufficient in practice.

We also compared two variants of our algorithm (single vs. multigraph, as described in Section 3). The idea behind this approach is to make it harder for group of friends to boost their rankings by constantly answering each others questions and rating their friends' answers as best. We discovered that collapsing duplicate edges results in the degradation of *HITS* performance. Therefore, for our experiments, we use the multigraph formulation for reporting results.

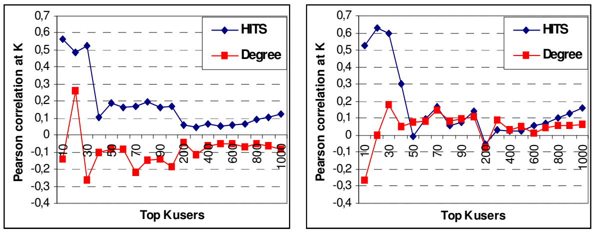

Figure 2: Pearson correlation at *K* for *HITS* and *Degree* vs. *Ratings* for (a) all categories and (b) *Science* category.

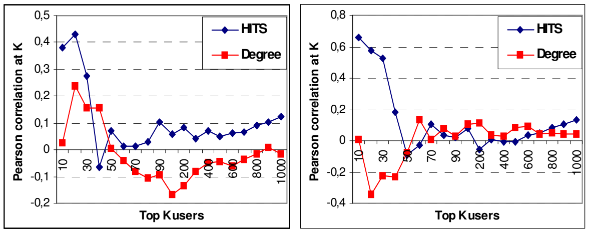

Figure 3: Pearson correlation at *K* for *HITS* and *Degree* vs. Votes for (a) all categories and (b) *Science* category.

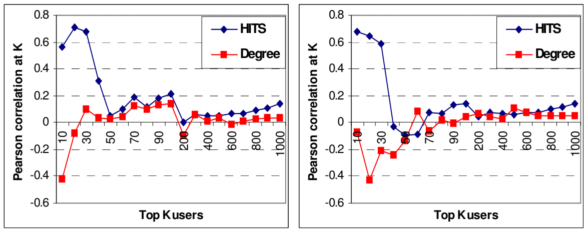

Figure 4: correlation at *K* for *HITS* and *Degree* vs. *%Best* scores for (a) all categories and (b) *Science* category.

Table 2: Percentage of answers chosen as Best for the top 10 HITS authorities (computed 5 months after HITS calculation)

HITS rank	*Sports Cycling*	*Sports Tennis*	*Science Engineering*
1	12%	16%	73%
2	31%	50%	27%
3	7%	22%	11% (Top Contributor)
4	11%	5%	17% (Top contributor)
5	18%	67%	51%
Average *%Best* scores for top 10 Authorities	16%	***32%**	***36%**
Average *%Best* scores for *all* users	14%	11%	19%

We now consider how well *HITS* (computed in January 2007) predicts the quality of the user 5 months into the future (May 2007). Table 2 presents the percentage of best answers for users ranked by *HITS* at positions 1 – 5 in Cycling, Tennis and Engineering categories. The percentage of best answers is computed as of May 2007, or 5 months *after our crawl was completed.* In many cases, the average percent of best answers for users with high authority values is often significantly higher than the average for all users in the category.

Table 3: Correlation of the top 10 Authorities with the Ratings and %Best metrics. * indicates moderate or strong correlation.

Category	*%Best*		*Ratings*		*Number of Posts*	*Average Degree*	*Power law fit*
	HITS	*Degree*	*HITS*	*Degree*			
Arts	***0.71****	***0.22***	***-0.28***	***-0.34***	***736,727***	***6.26***	
Sports	***0.58****	***-0.17***	***-0.06***	***0.23***	***1,046,411***	***7.35***	
Tennis	**0.69***	-0.10	**0.59***	-0.48	32,969	2.69	
Cycling	**0.76***	-0.073	**0.96***	0.23	20,158	2.46	
Science	***0.68****	***-0.07***	***0.38****	***0.03***	***1,469,207***	***7.43***	
Mathematics	-0.41	**0.31**	0.02	-0.26	49,825	3.47	0.88
Engineering	**0.77***	0.13	**0.49***	-0.32	36,198	2.45	**0.72**
Government	-0.11	-0.09	**0.48***	-0.21	75,279	3.17	0.85
Civic participation	**0.54***	0.21	-0.05	-0.60	59,743	2.57	**0.82**
Politics	0.15	0.21	-0.51	-0.25	109,861	6.86	**0.77**
Women studies	-0.11	0.00	-0.62	-0.11	79,542	2.94	0.85

We have observed that for some question categories *HITS* performs much better than for others. Therefore, we now take a closer look at different categories to determine what makes a given category amenable to the *HITS* authority estimation. Table 3 reports the correlation for *HITS* and *Degree* rankings for some interesting categories which we investigated further. As we can see, there is no correlation between the size of a category (i.e., number of posts) and accuracy of *HITS*. There is also no correlation between the average degree in a given category and its *HITS* accuracy values. Furthermore, *HITS* is more robust than *Degree*, and there is only one category where *Degree* performs better than *HITS* (namely, for the Science/Mathematics category which we investigate further). These findings suggest that there are some underlying structural differences in the corresponding graphs which are responsible for the differences in the authority estimation accuracy.

Even though the average degree of the graphs did not correlate with *HITS* effectiveness, we found that the goodness of fit (R^2) of the power law trendline correlates *inversely* with *HITS* performance on predicting the *%Best* scores for each authority, i.e. the more a category graph distribution deviates form the power law, the better *HITS* authority scores correlate with user feedback. This suggests exploring *local* properties of the graphs. We have visualized the graphs, focusing on the neighborhoods of top authority users for the Government and Engineering (Figure 5) and Mathematics and Civic Participation (Figure 6). The red nodes represent the top 10 authorities according to *HITS*, and the blue nodes represent all other users connected to the authorities.

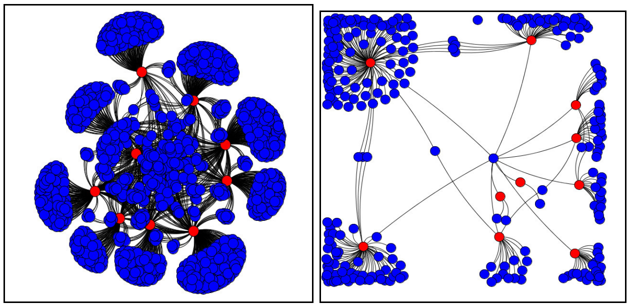

Figure 5: Neighborhood graphs of top 10 *HITS* authorities for (a) *Science/Government*, and (b) *Science/Engineering*.

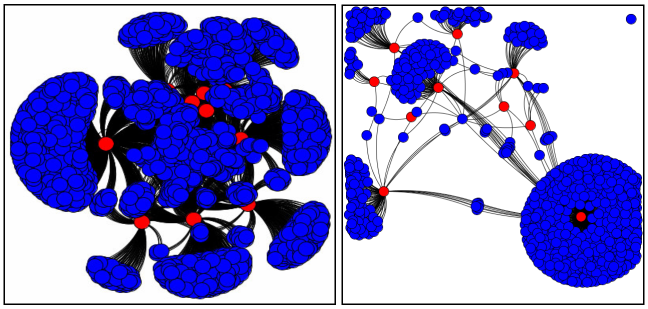

Figure 6: Neighborhood graphs of top 10 *HITS* authorities for (a) *Mathematics* and (b) *Civic Participation*.

The graphs for categories where *HITS* worked well (Government and Engineering) appear qualitatively different than for categories where *HITS* correlation was low. In the case of Government and Engineering, we can distinguish smaller groups (communities) which appear around subtopics. In this case, *HITS* can successfully find expert users. In contrast, Figure 5 show the neighborhood of top authority users in the categories Mathematics and Civic Participation (which exhibit poor correlation of *HITS* with *Ratings* and other metrics). Here a small number of users answer thousands of questions, and there is not much balance across virtual "topics" or around local authorities.

6. CONCLUSIONS

QA portals are a rapidly emerging alternative to web search that exhibit dynamics and structures different from the traditional static Web. This domain requires rethinking link analysis algorithms previously developed for different settings and interaction modes. In this paper we presented an adaptation of the *HITS* algorithm for predicting experts in QA portals such as Yahoo! Answers. We performed a large scale empirical evaluation of this method, demonstrating its effectiveness for discovering authorities in topical categories. We have also performed an extensive analysis of our results to shed some light on *why* link analysis performs well for some categories but not for others. A significant factor appears to be the deviation from the power law degree distribution, which indicates local structures in the graph that *HITS* is able to exploit. In the future we plan to extend this analysis to provide automatic prediction of the expected success or failure of *HITS*-like link analysis methods for a given user relationship graph.

In summary, we presented a first step towards analyzing the structure of the general web-scale question answer portals. These portals are emerging as a valuable alternative to web search, and are rapidly generating knowledge that rivals more established collaborative sources such as Wikipedia. The challenging questions of identifying authorities, managing trust, and estimating quality of this dynamic and rapidly changing content present an exciting new area of research in information retrieval and knowledge management.

ACKNOWLEDGEMENTS: We thank Ravi Kumar for valuable discussions.

REFERENCES

[1] A. Borodin, G.O. Roberts, J.S. Rosenthal and P. Tsaparas, Link Analysis Ranking Algorithms Theory And Experiments. *ACM Transactions on Internet Technology*, 2005.

[2] T. H. Haveliwala, Topic-sensitive PageRank. *WWW*, 2002

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

[4] P. Jurczyk and E. Agichtein. HITS on Question Answer Portals: an Exploration of Link Analysis for Author Ranking. *SIGIR*, 2007

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

[6] X. Liu, W. B. Croft and M. Koll, Finding experts in community-based question-answering services, *CIKM*, 2005.

[7] A. McCallum, A. Corrada-Emmanuel and X. Wang, Topic and Role Discovery in Social Networks. *IJCAI,* 2005

[8] L. Nie, B. D. Davison and B. Wu, From Whence Does Your Authority Come? Utilizing Community Relevance in Ranking. *AAAI*, 2007

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1998

[10] J. Zhang, M. S. Ackerman and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. *WWW,* 2007