Proceedings of the Ninth International AAAI Conference on Web and Social Media

Moving from Static to Dynamic Modeling of Expertise for Question Routing in CQA Sites

Reyyan Yeniterzi and **Jamie Callan**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

CQA sites are dynamic environments where new users join constantly, or the activity levels or interest of existing users change over time. Classic expertise estimation approaches which were mostly developed for static datasets, cannot effectively model changing expertise and interest levels in these sites. This paper proposes how available temporal information in CQA sites can be used to make these existing approaches more effective for expertise related applications like question routing. Adapting two widely used expert finding approaches for question routing returned consistent and statistically significant improvements over the original approaches, which shows the effectiveness of the proposed temporal modeling.

Introduction

Routing questions to users who can provide accurate and timely replies is an important application in Community Question Answering (CQA) environments. This task has been widely studied as a question specific expert identification task. Most of the prior work applied counting or profiling-based expert finding approaches to a snapshot of the environment which contains previously asked questions and their replies. In these approaches, CQA sites are assumed as static environments, and available temporal information is mostly ignored. However these sites are more dynamic in nature where new users join every day or the existing users' interests, roles and in-site activities change over time. These dynamic aspects of the environment should be taken into account for more effective question routing.

This paper identifies several reasons why current static expert finding approaches are not well suited for these environments. First, new users join to these environments everyday or inactive users may become more active over time. For instance, every month on average 12.5K users start replying to questions on the StackOverflow website. Widely used expert finding methods may not favor new users with limited reply history. Instead, they promote users who were actively replying to questions for a long time. However, in CQA sites, the only way users can show their expertise is through replying to other users posted questions. Unlike blog or microblog sites where users can post whatever they want on whenever they want; in CQA sites users' contributions are limited with posted questions related to their expertise. Furthermore, these questions should not be replied before, or replied but have not been completely resolved. Depending on the number of questions that satisfy these conditions, it requires some time for users to build reputation in these sites. Having a small reply history does not make user less of an expert when it comes to replying a question accurately. These users should also be considered in question routing.

Another dynamic aspect of CQA sites is the degree of activity change over time. For effective question routing, questions should not be only replied accurately but also within an acceptable time frame. Routing questions to inactive users can result in delays and even failures in receiving replies, therefore finding experts who can provide timely replies is also important. Cai and Chakravarthy's (2013) analysis on users' replying activities in StackOverflow over monthly intervals showed the considerable activity fluctuations over time. Additionally, we calculated the coefficient of variation (CV)[1] of replying activities for StackOverflow data over weeks, and found that around 90% of the active users [2] have $CV > 1$. This means that for most users who reply to on average n questions per week, the standard deviations are more than n, as they may reply to more than $2n$ questions in a week and may not reply any in another week.

The change in user's interest is yet another reason why users who were replying to topic relevant questions before, may not reply anymore. Cai and Chakravarthy (2013) also performed a correlation analysis on users' replies (words used) over time. Their analysis revealed possible topic drifts for some users. Changes in users' availabilities and interests are important temporal factors and should be considered for more effective question routing.

In order to overcome these problems, some prior works examined the temporal information available in these sites. Pal et al. (2012) analyzed the evaluation of experts over time and showed that estimating expertise using tempo-

[1] Coefficient of variation, which is the ratio of the standard deviation to the mean, $CV = \frac{\sigma}{\mu}$, represents the measure of variation (σ) within a distribution with respect to its mean (μ).

[2] CV is very sensitive to small changes when μ is close to 0, so only the activities of users with replying $\mu >= 1$ are used.

ral data outperforms using static snapshot of the data. Cai and Chakravarthy (2013) also used temporal features calculated between the time question and its replies are posted, to improve answer quality prediction. Some works used temporal features to estimate the availability of users for a given day (Li and King 2010; Sung, Lee, and Lee 2013; Chang and Pal 2013) or for a specific time of the day (Liu and Agichtein 2011; Chang and Pal 2013). Using users' availability with respect to certain days and hours improved the performance of expertise related tasks in CQAs. This paper extends the prior work by proposing a temporal modeling of expertise which incorporates the dynamic features of the environment to some of the existing state-of-the-art approaches in order to overcome all the mentioned problems at the same time. The proposed approach uses all prior replying activities of users without punishing the recently joined users. The temporal aspect of the approach is also useful for modeling the availability and recent interest of users.

Temporal Discounting

In this paper, we use temporal discounting for identifying experts for a given question. Temporal discounting, which is a widely studied phenomenon in economy and psychology, refers to the decrease in the subjective value of a reward as its receipt delays over time. In other words, the longer one needs to wait for a future reward, the lower its present subjective value becomes. People have a tendency to discount delayed rewards and give more value to near future rewards. This behavior can be also observed for the past. People give more value to recent events than events that occurred a long time ago. Therefore, in dynamic environments where users and their activities change over time, the system should have a tendency to give more value to recent activities and discount earlier activities especially when interacting with users in real time. CQA sites are among these systems as new users join in and existing users' activities and interests change over time constantly. Thus, more effective question routing can be possible by applying temporal discounting towards earlier posted replies and so giving relatively more value to recent replies. Such an approach will give enough credit to newly joined users while not ignoring the earlier replies of existing users. By using temporal information, this approach will also indirectly model and use the availability and interest of users.

Two forms of temporal discounting functions have been used widely, exponential and hyperbolic discounting. This paper proposes to integrate these models into existing expert finding approaches used for question routing, more specifically the Answer Count and ZScore (Zhang, Ackerman, and Adamic 2007; Bouguessa, Dumoulin, and Wang 2008) approaches. Before explaining these in detail, the time interval divisions used to group the past activities are summarized.

Constructing Time Intervals

Assume that t_1 represents the time of the first question posted to CQA site and t_q represents the time question q is posted to the site. During identifying expert candidates who can reply to question q, only questions and replies posted within the period $[t_1, t_q]$ are used. Previous approaches mostly treat replies posted within this interval equally. However, in our proposed approach, the value of a posted reply depends on its posting time, therefore, replies are initially grouped with respect to their posting times. $[t_1, t_q]$ interval is divided into specific periods of times such as days, weeks, biweeks and months. The dates of posts are used to find their corresponding time intervals. The day interval of the first post is set as 1, $d(t_1) = 1$, while the day interval of question q is equal to 1 + the number of days passed since t_1.

Exponential and Hyperbolic Discounting

In exponential (exp) discounting model, the value of replies are exponentially discounted as time goes on. Exponential discounting can be represented as $e^{-k\Delta t}$, where k represents the parameter describing the rate of decrease and Δt is the number of time intervals passed since reply was posted. The hyperbolic (hyp) discounting model is in the form $1/(1 + k\Delta t)$. The hyperbolic model shows very rapid fall initially, and then the decrease becomes more gradual as time passes, or in other words, as Δt gets higher. For a given question q and for any reply posted at time interval i, Δt_i is calculated as $d(t_q) - d(t_i)$ for days. It is always the case that $d(t_q) >= d(t_i)$ for any i, since only the earlier posted replies are used.

Integrating temporal discounting to counting-based expertise calculation algorithms is straightforward. Instead of counting all instances equally, a discounted value depending on their time of creation is used. For *Answer Count* (*AC*) approach, the static expertise estimation of user u is equal to the number of replies posted by user u. On the other hand, its temporal discounted versions are as follows:

$$AC_{exp}(u) = \sum_{i=1}^{q} R_i(u) e^{-k\Delta t_i} \quad (1)$$

$$AC_{hyp}(u) = \sum_{i=1}^{q} \frac{R_i(u)}{1 + k\Delta t_i} \quad (2)$$

where $R_i(u)$ is the number of replies posted by user u at interval i. Similarly for temporal ZScore approach, first the ZScore is calculated for each time interval and then it is discounted with respect to its temporal distance from question's interval. Its formulation is as follows:

$$ZScore_i(u) = \frac{R_i(u) - Q_i(u)}{\sqrt{R_i(u) + Q_i(u)}} \quad (3)$$

$$ZScore_{exp}(u) = \sum_{i=1}^{q} ZScore_i(u) e^{-k\Delta t_i} \quad (4)$$

$$ZScore_{hyp}(u) = \sum_{i=1}^{q} \frac{ZScore_i(u)}{1 + k\Delta t_i} \quad (5)$$

where $Q_i(u)$ is the number of questions posted by user u at interval i.

Experimental Methodology

StackOverflow[3] is a community question answering site focusing on technical topics such as programming languages,

[3] http://stackoverflow.com/

algorithms and operating systems. A public data dump of StackOverflow site from May 2014, which contains around 7.2M questions asked by 1.2M askers and 12.6M replies posted by 862K users, is used for experiments.

During evaluations, all the authors of the particular question's replies are treated as relevant while all the other retrieved users who did not reply to the question are treated as irrelevant. These other users who did not reply to this particular question may have the necessary knowledge and background to answer the particular question, but due to incomplete assessments, they were assumed irrelevant. Although this methodology is not ideal, it was used commonly in prior research, and was also used in this paper. However, in order to decrease the effects of incomplete assessments, 250 questions, each with 15 replies were selected for test set so that questions have more users assessed as relevant on average. The success of question routing task depends on one of the identified expert candidates to answer the particular question, therefore the performance is reported with Mean Reciprocal Rank (MRR). *Matching Set Count (MSC) @n* (Chang and Pal 2013), which reports the average number of the questions that were replied by any user ranked within top n identified candidates, is also used to report the performance. Two statistical significance tests, randomization (r) and sign (s), were applied in order to draw safer conclusions. Results that are significant with $p < 0.05$ are presented with r and s symbols and results which are significant with $0.05 < p < 0.1$ are presented with r' and s' symbols.

Baseline Approaches

The original AC and ZScore algorithms were used as baselines for static approaches. Their question dependent versions were used for more effective performance. For a given question, the question's tags were initially searched among other previously asked questions' tags. Then the retrieved questions' repliers were extracted, and for each user the number of retrieved replies and asked questions were used to calculate the AC and ZScore scores.

Two prior works on availability estimation are also used as temporal baselines. Sung et al. (2013) estimated availability as a sigmoid function applied recency value which is calculated as follows:

$$\frac{1}{1 + e^{-\alpha \sum_{i=1}^{|R(u)|} \frac{1}{age(r_i)+2}}} \tag{6}$$

where $|R(u)|$ is the number of replies posted by user u at any time and $age(r_i)$ is the number of days passed since reply i is posted. α is set as 0.1 based on (Sung, Lee, and Lee 2013).

Chang and Pal (2013) also built binary classifiers on previous n days of activity with different machine learning approaches. However, these classifiers did not beat the simple baselines of assuming always available or using the availability status of previous day directly. Using always available is same as the static approach, so the status of previous day is used as another temporal baseline in this paper.

The first temporal baseline uses all replies of users while the second one uses all replies of users from a certain time frame (previous day) in order to estimate availability. Our proposed approach is different from these as we only use the particular question related replies in temporal modeling. This use of topic dependent activities is useful for modeling user's interest as well. Estimating user availability is useful only if user is still replying to questions on the particular topic of question, which may not always be the case.

The following equation is used to combine the content (AC and ZScore) and availability scores:

$$finalScore = content^{\lambda} * availability^{1-\lambda} \tag{7}$$

Min-max normalization was applied to the the first availability baseline to make its range $[0, 1]$ (Sung, Lee, and Lee 2013). Similarly for the second temporal baseline, the availability will be either 0 or 1. Therefore, content scores are also normalized to have a similar range with availability scores. 10-fold cross-validation is used to find the optimum parameter setting for the interpolation.

	Answer Count		ZScore	
	MRR	MSC@10	MRR	MSC@10
static	.1545	.3000	.1438	.2880
+ Chang	$.1592_s$	$.3520^r_s$	$.1493_s$	.3240
+ Sung	$.1572_s$	$.3520^r_s$	$.1505_s$	$.3160^{r'}_{s'}$

Table 1: Experimental results of static and temporal baselines. (r/s: $p = 0.05$, r'/s': $p = 0.1$)

		Answer Count		ZScore	
		MRR	MSC @10	MRR	MSC @10
exp	day	$.2043^r$	.3800	$.1965^r$	.3760
	week	.1755	.3560	.1703	.3320
	biweek	.1691	.3720	.1619	.3720
	month	$.1879^r_s$	$.4120^{r'}_{s'}$	$.1820^r_s$	$.4040^r_s$
hyp	day	$.2170^r_s$	$.4480^r_s$	$.2050^r_s$	$.4400^r_s$
	week	$.1780_s$	$.4240^r_s$	$.1787^r_s$	$.4120^r_s$
	biweek	$.1737_{s'}$	.3960	$.1717^r_s$	$.3920^r_s$
	month	.1674	.3720	$.1664^{r'}$	$.3760^r_s$

Table 2: Experimental results of proposed temporal models.

Experiments

The results of static and temporal baselines are summarized in Table 1. Combining estimated availability with original approaches, which do not use any temporal information, provided consistent and statistically significant improvements. Using temporal information just for estimating availability is shown to be effective. In our proposed approaches, in addition to availability; the interest of users and the recently joined users' activities are also modeled. The experimental results of these are presented in Table 2.

Experiments on proposed approaches were initially performed with rate of decrease $k = 1$. The exponential and hyperbolic discounting approaches were applied to different time intervals as presented in Table 2. Results that are statistically significant to static and both of the temporal baselines are specified in the table. As seen from Table 2, the proposed dynamic modeling of expertise approaches consistently outperform the static and temporal baselines with respect to all

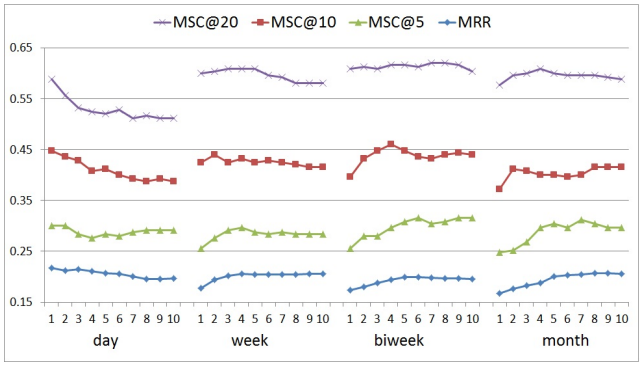

Figure 1: Experimental results of AC_{hyp} approach with respect to different k values.

experimented time intervals. Some of these differences are statistically significant over all 3 baselines.

In Table 2, different behaviors are observed for exp and hyp models, possibly due to the difference in their degree of decay over time. The discounting rate (weight difference between consecutive intervals) of two models are initially similar for small values of Δt, however as Δt increases, the drop rate exponentially increases for exp model, while the increase is linear for hyp. For instance, with $k = 1$, the weight ratios of 1st interval to the 3rd, 5th and 10th intervals are 2, 3 and 5.5 for hyp model, while these ratios are 7.4, 54.6 and 8103 respectively for exp model. This high drop rate in exp model cause recent intervals to receive relatively much more weights and dominate the overall score. Only the month interval, the longest time interval tested, returned consistent significant improvements with exp discounting; probably because activities from the most recent couple of months provide enough data to build effective user expertise and interest models. However, the same behavior doesn't apply to shorter intervals due to lack of enough information for modeling users. The day interval performs relatively better than week and biweek possibly due to its effectiveness in estimating availability of users.

On the contrary to exp model, more consistent and statistically significant improvements are observed with hyp discounting. This is mostly due to the smoother decay used in temporal modeling. Due to the smoother decrease, activities from recent intervals do not dominate the overall model. Activities from high Δt have still comparable effects on the model. In experiments with $k = 1$ (Table 2), shorter intervals perform better than the longer intervals; mainly because with shorter intervals, the availability and recent interest of users can be estimated more accurately. Therefore, day interval performs better than others. However, the relative ranking of these intervals also depend on the decay factor k. The decrease goes smoother over time when k is low. When k is high, the decrease between time intervals becomes more drastic. In order to analyze the effects of k more clearly, the performance of proposed AC_{hyp} approach with increasing k values (from 1 to 10) are presented in Figure 1 for different time intervals with additional metrics, MSC@5 and @20.

Several trends exist in Figure 1. For instance, with day intervals, the scores are highest when $k = 1$ but decrease as k gets higher values. This is because, with high values of k, the activities from small values of Δt (same day or previous day mostly) get relatively more weights in modeling expertise which negatively affects the overall ranking. On the other hand, with biweekly and monthly intervals, the performances increase as k value increases and then become more stable. This tendency towards using higher k values and giving much more value to recent biweeks and months is probably due to more effective modeling of user availability and interest in addition to expertise. Unlike days, using a couple of months activity can be enough to model users' expertise as well as their availability. Week intervals of AC_{hyp} also perform in between day, biweek and month intervals. Similar trends are also observed with $ZScore_{hyp}$.

Conclusion

This paper proposed adapting temporal discounting models to expertise estimation methods for question routing. Two widely used approaches, Answer Count and ZScore, were modified to use the available temporal information. Consistent and statistically significant improvements were observed in both approaches with hyperbolic discounting. For approaches where temporal information cannot be integrated directly, such as feature-based approaches, these proposed approaches can be used as additional features to improve the overall performance.

Acknowledgments

This research was supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

References

Bouguessa, M.; Dumoulin, B.; and Wang, S. 2008. Identifying authoritative actors in question-answering forums: The case of Yahoo! answers. In *KDD*.

Cai, Y., and Chakravarthy, S. 2013. Improving answer quality prediction in q/a social networks by leveraging temporal feature. In *WSDM*.

Chang, S., and Pal, A. 2013. Routing questions for collaborative answering in community question answering. In *ASONAM*.

Li, B., and King, I. 2010. Routing questions to appropriate answerers in community question answering services. In *CIKM*.

Liu, Q., and Agichtein, E. 2011. Modeling answerer behavior in collaborative question answering systems. In *ECIR*.

Pal, A.; Chang, S.; and Konstan, J. A. 2012. Evolution of experts in question answering communities. In *ICWSM*.

Sung, J.; Lee, J.-G.; and Lee, U. 2013. Booming up the long tails: Discovering potentially contributive users in community-based question answering services. In *ICWSM*.

Zhang, J.; Ackerman, M. S.; and Adamic, L. 2007. Expertise networks in online communities: structure and algorithms. In *WWW*.