Utilizing Non-QA Data to Improve Questions Routing for Users with Low QA Activity in CQA

Ivan Srba, Marek Grznar, Maria Bielikova
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
ivan.srba@stuba.sk, xgrznarm1@stuba.sk, maria.bielikova@stuba.sk

***Abstract*—Community Question Answering (CQA) systems, such as Yahoo! Answers and Stack Overflow, represent a well-known example of collective intelligence. The existing CQA systems, despite their overall successfulness and popularity, fail to answer a significant amount of questions in required time. One option for scaffolding collaboration in CQA systems is a recommendation of new questions to users who are suitable candidates for providing correct answers (so called question routing). Various methods have been proposed so far to find appropriate answerers, but almost all approaches heavily depend on previous users' activities in a particular CQA system (i.e. QA-data). In our work, we attempt to involve a whole community including users with no or minimal previous activity (e.g. newcomers or lurkers). We proposed a question routing method which analyses users' non-QA data from a CQA system itself as well as from external services and platforms, such as blogs, micro-blogs or social networking sites, in order to estimate users' interests and expertise early and more precisely. Consequently, we can recommend new questions to a wider part of a community as well as more accurately. Evaluation on a dataset from Stack Exchange platform showed that considering non-QA data leads not only to better recognition of users with low activity as suitable answerers, but also to higher overall precision of the recommendations. It implies that non-QA data can supplement QA data during expertise estimation in question routing and thus also improve a success rate of a questions answering process.**

Keywords—community question answering; question routing; question recommendation; non-qa data; expertise estimation

I. INTRODUCTION

The current web uses a variety of search engines to provide people with an ability how to effectively identify and obtain valuable information. Despite the fact that search engines in the last decade have significantly improved their successfulness and effectiveness, there are still some information needs that current search engines cannot meet effectively. This is due to several reasons [1], e.g. required information can be too complex, fragmented among several sources or even subjective, such as a recommendation. In addition in many cases, it can be difficult to describe required information as a search query just by keywords.

Web 2.0 offers an opportunity that helps to solve the mentioned problems. This option is to obtain necessary information by asking an online community in various knowledge sharing systems [1], such as web forums or mailing lists. Among them, Community Questions Answering (CQA) come more and more to the fore in the recent years. Popular CQA systems include Stack Overflow, Yahoo! Answers or Quora, to name a few. The typical process of community question answering consists of several steps. Any user can post a new question and other users can share their knowledge by providing their answer-candidates. Consequently, the community can vote on these answers to highlight the most useful ones. The process of question answering is finished as soon as the asker select one of answers as the best one.

In comparison with conventional information retrieval systems, CQA is based on two concepts: collective intelligence and wisdom of the crowds. Employment of these concepts allows CQA systems to provide satisfying answers on an enormous amount of questions each day. In spite of that, one of the most serious problems is that CQA systems quite often fail to answer questions in the required time. In 2010, only 17.6% of newly posted questions in Yahoo! Answers received a satisfactory answer within 48 hours [2]. Two years later, analyses on the same system showed that 11.95% of the questions were answered in one day and only 19.95% of the questions were answered in two days [3]. To improve performance of CQA systems and to guarantee fast and accurate answers on as many questions as possible, many collaboration support methods have been proposed so far.

In this paper, we focus on question routing which refers to a recommendation of new questions to potential answerers. Almost all of the existing approaches for question routing rely on previous users' activities in a CQA system (so called QA-data that consist mainly of asking questions and providing answers) and thus they are applicable only for users with high level of activity. Besides these users also newcomers and lurkers (i.e. users who are a part of community but do not actively participate on question answering) could be good candidates to answer questions and the system can motivate them to become more involved in the community. However, as they have no or only minimal interaction with the system, the QA-based question routing methods do not have sufficient information about their expertise and therefore it is not possible to route any questions to them. In order to address this problem, several question routing approaches based on non-QA data acquired from CQA system itself or external services (e.g. blogs, micro-blogs, and social networking sites) have been proposed very lately. These methods, however, focus primarily on estimation of users' social attributes (such as activeness, influence and connectivity) that can predict users' willingness to provide an answer on a question but not their level of expertise or interest which is even more important.

We suppose that non-QA data are suitable also to estimate users' expertise and interest. Therefore, our main contribution is a proposal of a new approach which utilizes besides QA data also non-QA data in order to estimate users' expertise. In the proposed method, we employ a probability model based on latent topics identified by Latent Dirichlet Allocation (LDA). To the best of our knowledge, our method is the first one which combines state-of-the-art latent topic modeling for expertise estimation employed in QA-based approaches with non-QA sources of data in order to estimate users' knowledge early and more accurately for users with low level of QA activity.

The rest of this paper is organized as follows. Section 2 introduces state-of-the-art question routing approaches based on QA as well as non-QA data. In Section 3, details on the proposed method are given. Experimental evaluation on a dataset from Stack Exchange platform is described in Section 4. Finally, conclusions are proposed in Section 5.

II. Related Work

Currently, CQA systems represent an interesting subject of research in the domain of social networks, information retrieval and knowledge management systems. From the previous research studies, especially approaches aimed to adaptively support users' collaboration has a significant impact on question answering successfulness and effectiveness.

The majority of adaptive support approaches applied in CQA systems fall into two groups according to the source of knowledge which is employed to answer new questions: 1) *question retrieval* tries to identify the required information in the already existing question answer pairs; 2) *question routing* refers to a recommendation of new questions to users who would be able (and possibly willing) to provide answers.

From the above classification, we focus on question routing, because this group of approaches has the best chance to improve efficiency of collaboration in CQA systems and it still provides many open research problems.

A. *Question Routing*

Question routing can be characterized as a recommendation task which aims to identify users (experts) who are suitable candidates for providing an answer for a given question. These potential answerers must have necessary expertise in the question topic. Some approaches take into consideration also additional user characteristics, such as user authority, overall activity [4] or availability [2].

The problem of question routing can be formalized as follows: given a newly posted question q we need to create an ordered list of top k users $u_1, u_2, \ldots, u_k$ who are the most suitable to answer question q. This list is usually ordered by a probability that user u would answer given question q. To obtain the list of suitable answerers, it is necessary to solve three sub-problems [5]: 1) construction of a question profile, which represents question's topics; 2) construction of a user profile, which represents user expertise/interest and optionally also additional characteristics (e.g. authority); 3) matching between profile of a new question and all relevant user profiles.

In an extensive literature survey of adaptive methods applied in CQA systems, we identified 32 papers that tackle with the question routing problem (published from 2005 to 2015). We divided all these approaches into three groups according to various models they apply to create question/user profiles or to find matching between them.

Language-Model-based Question Routing. The first group of approaches is based on language models. Traditional language model approaches (e.g. [2]) represent both question and user profiles as a bag of words (the user profile is created from all questions the corresponding user previously answered or asked). Afterwards, user profiles are ranked according to Query Likelihood Language Model (QLLM) which calculates a probability that user profiles will generate terms of the routed question. In these traditional language models, data sparseness can lead to word mismatch between the routed question and user profiles which can be caused by co-occurrence of random words in user profiles or questions [6]. This problem is solved by translation models (e.g. [6]) which employ statistical machine translation to overcome data sparseness and which is able to differentiate between exact matched words and translated semantically related ones.

Topic-Model-based Question Routing. Language models are based on exact word matching and thus they are not able to capture more advanced semantics and solve the problem of lexical gap between the posted question and user profiles [4]. As a result of this limitation, latent topic models, such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA), are employed to consider not only syntactic but also semantic similarities.

Classification and Ranking-based Question Routing. The third group of approaches (e.g. [3], [7]) tackles question routing problem in comparison with previous language and topic models as a classification or ranking problem. Question and user profiles are represented as a set of features that are consequently used in classification (e.g. SVM, random forest) or ranking (e.g. SVM rank).

While all approaches in the prior works achieved interesting results, in general the experimental evaluations confirmed that various widely used topic-based models significantly outperformed language models (e.g. [4], [8]). In addition, topic models have been already successfully used also in combination with classification-based question routing as a feature containing a text similarity calculated by LDA topic model achieved the best performance in [7].

B. *Involvement of a Whole Community*

There are, however, still large gaps and drawbacks in the existing state-of-the-art topic-model-based approaches for question routing. All these approaches significantly rely only on data from CQA system (provided answers, asked questions etc.). This high dependency is obvious also during experiments as these approaches consider only those users who previously provided more than 5 [9], 10 [5] or even 20 answers [10]. Moreover, some approaches take only users with a significant number of best answers into consideration (with more than 10 [4] or even 20 best answers [8]).

However, distribution of activities among users in CQA systems follows a typical long-tail distribution as 1% of the most active users create more than 25% of all content. Authors in [6] report that only 15.67% of all users in Yahoo! Answers answered more than four questions. It means that all these approaches involve only a small proportion of highly active users and are not able to recommend new questions to the rest of the community, which includes also newcomers (due to well-known cold start problem) [11] or lurkers, who just remain a part of community but do not actively participate on a question answering process. Nevertheless, to preserve a long-term sustainability of CQA ecosystem, it is necessary to satisfy expectations of all types of users [12]. If a question routing method would be able to involve also users with no or minimal previous activity, these users can become motivated to take more intensive participation in a community or even develop to highly active experts.

We also confirmed the necessity of involvement of a whole community in our previous case study [13]. As we pointed out, we can witness new emerging problems in the most popular CQA systems (e.g. in Stack Overflow): an increasing failure rate (i.e. a proportion of unanswered questions) and churn rate (i.e. a proportion of users who leave the community). The results of our study showed that these emerging problems are highly related to the growing amount of undesired groups of users (i.e. help vampires, noobs and reputation collectors) that produce a great amount of low quality content. As current question routing methods involve only highly active users, who are often also experts in particular topics, these users are becoming more and more overloaded with low-quality and uninteresting content. If we will be able to route questions to a whole community, a total load will be distributed among more users and thus it will improve not only an overall success rate but it will also contribute to system's long-term sustainability.

To achieve this shift and to involve all relevant users in question routing, it is possible to utilize users' various publicly available sources of information (so called non-QA data). In the current Web 2.0 era, it is possible to utilize social media tools and services [14], such as messages on Twitter, status updates and friendships on Facebook or published blogs. Possibilities are even richer in domain specific CQA systems such as Stack Overflow, which is a CQA system dedicated to programming-related questions, where we can take advantage of public source code repositories (e.g. Github). The most of CQA systems allows users to create user community profiles where users can explicitly specify links to these social media tools and also directly describe their interests or knowledge.

C. Utilization of Non-QA Data in Question Routing

In the very recent time, several studies have already investigated a potential of non-QA data that are publicly available about users for purpose of question routing in CQA systems. Pan et al. [14] conducted an exploratory study to verify feasibility of leveraging non-QA social activities in organizational enterprise settings. Users' activity in various tools provided by IBM Connections software (e.g. forums, blogs) were used to estimate users' overall activeness, influence and connectivity. Consequently correlations between these characteristics and answering behavior in a CQA system were calculated. The obtained significant correlations implied an interesting potential of non-QA data. On the basis of achieved results, the authors extended the previous study and proposed a question routing method [15], which derives from non-QA data not only users' social attributes but also expertise, nevertheless the method used only the simple bag-of-words representation to create question and user profiles.

In study [16], answerers' non-QA data was successfully used as features in a classification task whether an answerer will provide an answer on a particular question or not. Non-QA data described users' social attributes (activeness, influence and connectivity) but not their level of expertise.

Finally, authors in [17] applied non-QA data also in a question routing task. The proposed method allows to consider social following (e.g. in Twitter) and social friendship (e.g. in Facebook) when ranking answerers.

Based on the successful results achieved in all four studies, it is possible to confirm a potential of non-QA data to improve question routing and to overcome question routing issues (i.e. the cold start problem and the sparse data for users with the low level of activity). However, all these approaches used non-QA data mainly to estimate users' social attributes (i.e. activeness, influence and connectivity). Only the study [15] attempted to derive also users' expertise, nevertheless just term vectors were used for this purpose. At the same time, the state-of-the-art approaches applied at QA-data confirmed that topic models can significantly outperform bag-of-words language models.

III. METHOD FOR QUESTION ROUTING USING NON-QA DATA

To fill the identified gap between state-of-the-art question routing approaches and their application with non-QA data, we propose a novel question routing method which combines verified topic-model-based approaches with non-QA sources of data. We suppose that non-QA data can be used as a supplement for QA activities in expertise estimation and thus they will improve a prediction whether low-activity users are suitable candidates to provide an answer.

Similarly as the previous methods, we also base our method on three main building blocks: question profiles, user profiles, and a matching procedure. In the prior topic-based question routing methods (e.g. [4], [9], [5]), question profiles are created only for purpose of the matching procedure while user profiles are created separately from previously asked/answered questions by concatenation of their content. Consequently users' expertise (captured by user profiles) was represented by a topic distribution inferred from latent topic modelling. This solution has, however, a significant drawback. As soon as any user provides an answer, it is necessary to perform re-profiling of all user profiles since the topic distribution may be changed [11]. Therefore, these methods are not suitable in online situations as well as in our case since we want to evaluate how non-QA data can improve question routing after a user post his/her first answers (and recalculation of all user profiles after each new answer will be time-consuming also in offline settings).

A solution for this drawback was proposed in [12] where latent topics are inferred for questions instead of whole users' answering history and consequently the user profile is created by an aggregation of question profiles. It means that question profiles are used not only for the matching procedure but also to derive user profiles. Afterwards, when a user provides a new answer, it is necessary to update only his/her own user profile by incorporating a topic distribution from the corresponding question. In the proposal of our method, we follow this paradigm. The overall framework of our proposed question routing method is illustrated in Fig. 1.

At first, a question profile is created for a new question as well as for all previously answered questions (Step 1). The second step is specific for our approach as extraction of non-QA data for all users in the system is performed in order to create non-QA data profiles (Step 2). Then we create user profiles separately from question profiles (corresponding to users' previously answered questions) and non-QA data profiles (Step 3). It means that each potential answerer in the system is represented by two user profiles (non-QA and QA). Finally when the new question profile and all user profiles are created, we are able obtain a list of recommended answerers for the new question by common matching of these profiles (Step 4).

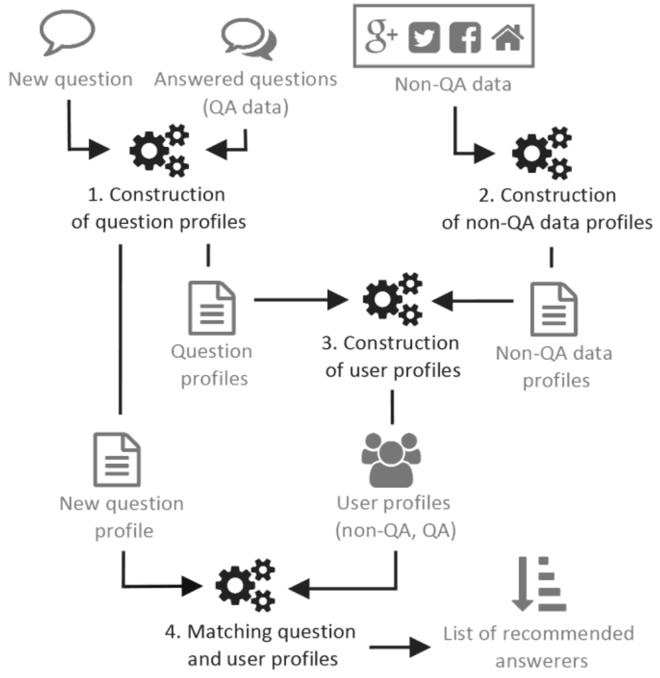

Fig. 1. Framework of the proposed question routing method.

A. *Construction of Question Profiles*

We represent each question by a question profile, which describes its topics (information need). When a question is posted, an amount of available information is limited. Thus, when creating the question profile for a new question, we only consider information available at creation time: question's tags and textual content (i.e. title and body of the question). We ignore any additional question data that may be added later (answers, votes, etc.). Therefore, the question profile is built only once at the time when the question is added to the system. The main motivation for utilizing only these parts of the question is that we do not need to perform any updates on the question profile later. This solution presents clear benefits in terms of scalability as it was declared also in [12]. We are aware that incorporating answers and votes (such as in [9]) may lead to a better estimation of question's topic and user expertise (as users with a high level of activity do not have to be necessary experts [18]). Nevertheless, this improvement can be achieved only at the expense of performance and scalability and thus we let considering these additional data as a possible extension in a future work.

Before building the question profile, we concatenate question title, body and assigned tags. Secondly, we apply basic text preprocessing methods (i.e. tokenization, stop-word removal and lemmatization). Once the question is preprocessed, we build its profile which consists of two models: 1) a unigram bag-of-words model, which is used later in the matching procedure; and 2) a Latent Dirichlet Allocation (LDA) model, which is used later to create the user QA profile.

Bag-of-words model. In order to describe the question at lexical level, we employed a unigram bag of words. In this model, the weight of each word corresponds to the frequency of its occurrence in the question.

Latent Dirichlet Allocation (LDA) model. To describe the question topic on semantic level, we employ the widely used smoothed Latent Dirichlet Allocation (LDA) model [19], which has been widely used in information retrieval and which is represented as a probabilistic graphical model in Fig. 2.

In LDA, the topic mixture is drawn from conjugate Dirichlet prior that remains the same for all questions. The process of generating question profile θ_q for a specific question q is as follows: 1) choose a multinomial distribution ϕ_z for each topic z from a Dirichlet distribution with parameter β, ϕ_z describes words distribution within topic z; 2) pick a multinomial distribution θ_q for each question profile from Dirichlet distribution with parameter α; 3) for each word token w in the question profile $\theta_{q,}$ select a topic $z \in \{1,\dots,K\}$ from the multinomial distribution $\theta_{q;}$ 4) pick word w from the multinomial distribution ϕ_z. Repeat this procedure for N_q (number of words in θ_q) times, then the question profile is generated. Finally, the above procedure is repeated N times for all questions.

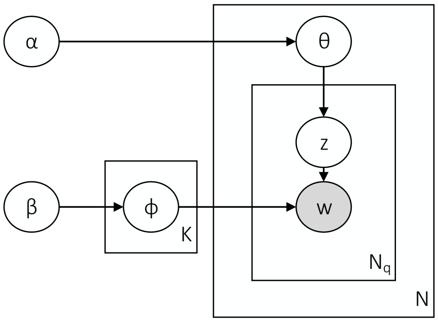

Fig. 2. Plate notation for Latent Dirichlet Allocation (LDA) model [19].

B. *Construction of Non-QA Data Profiles*

In general, non-QA data can be divided into two groups: 1) *internal non-QA data* come from a CQA system, but are not related to a question answering process itself; 2) *external non-QA data* are created by public information about a user that are spread through external systems and social networking sites.

Internal non-QA data are captured mainly by user community profiles. The current CQA systems commonly require a registration to do all interactions within a system, during which users are asked to enter a few details about themselves, such as a personal description or categories/tags a particular user is interested in. Preliminary analyses on dataset from Stack Overflow revealed that about 400K users provided personal descriptions (i.e. About me). Out of these users, more than 169K did not answer any question; and more than 270K did not answer more than 4 questions. It means that internal non-QA data are available for a significant number of users with a low level activity and it can be even improved if users will be motivated to provide these data in order to receive better recommendations.

Moreover, non-QA data does not have to be provided by a user directly in a community profile. Many systems allow users to specify a number of various links to external social tools and services that can be utilized to estimate users' expertise and interest, such as Facebook, Twitter or LinkedIn. For example, if we consider public information from the social networking site LinkedIn, we can learn which topics users are interested in, which schools they attended or where they work. Another increasingly common phenomenon are home pages and blogs, where users write articles from their professional life.

Preprocessing of non-QA data consists of three phases (for brevity, we omit the details of these phases here): (1) obtaining a required content from internal/external sources using a web crawler or API; (2) content identification (e.g. when extracting a content from a blog, it is necessary to identify relevant text paragraphs); (3) and finally, preprocessing, which applies the same text preprocessing methods as for questions (i.e. tokenization, stop-word removal and lemmatization). Finally for the preprocessed content obtained from each non-QA source, we build the non-QA data profile by inferring topic distribution $\theta_{non\text{-}qa}$ from the LDA model which we previously used for questions (so non-QA profiles share the same topics and their word distributions ϕ_z as question profiles).

C. Construction of User Profiles

The fact that we have two types of information about a user is reflected in construction of two user profiles. The first one is user QA profile which captures user expertise from QA data, while the second one does the same with non-QA data.

User QA Profile. As we introduced earlier, in contrast to the majority of prior topic-based question routing methods, we derive user QA profiles rather indirectly from question profiles. Some approaches (e.g. [5]) mix answered and asked questions during user profiling. However, as authors emphasized in [20], each user plays two different roles in CQA system simultaneously: an asker and an answerer. While answering a question can be perceived as an expression of expertise on question topics; asking a question, on the other side, can be perceived as a luck of expertise. For this reason, we decided to derive the user QA profile only from questions a corresponding user previously posted an answer on.

In order to aggregate profiles of answered questions, we use a similar approach as it was previously proposed in [12] and [9]. User QA profiles is represented by a topical distribution $\theta_{u\text{-}qa}$ which is computed as an average of topical distributions θ_q from question profiles which correspond to all questions answered before the current timestamp. Authors in [12] decided to use also a decaying factor to suppress questions answered in remote history and thus enable users to shift their answering behavior more rapidly. However, as we focus on users who have just provided their first answers, this decaying factor is not necessary.

User Non-QA Profile. We have already introduced how to extract and preprocess non-QA data from various sources and how to build non-QA data profiles. Now, we propose to build user non-QA profile $\theta_{u\text{-}non\text{-}qa}$ in a similar way as we previously proposed for user QA profiles. We aggregate all users' non-QA data profiles by averaging their topical distributions $\theta_{non\text{-}qa}$. Alternatively, we can take a diversity of non-QA data sources into consideration and use a weighted arithmetic mean.

D. Matching Question and User Profiles

The core idea behind question routing is to obtain a list of users ranked by a probability how likely they will provide a suitable answer for a routed question. In general, there are two main options how to obtain this ranking: by calculation of a similarity between question and user profiles (e.g. with dot-product measure [12]) or by employing a probabilistic model (e.g. [4], [10], [21]).

Due to the design of our question and user profiles (they are represented as a distribution over the same set of topics described by word distribution ϕ_z), it would be possible to use both options. Although we decided for the probabilistic model which directly allows us to consider also additional information about each user. Formally, given a new question q, the probability that a user u will provide a suitable answer is:

$$P(u|q) = \frac{P(u)P(q|u)}{P(q)} \qquad (1)$$

where $P(u)$ is a prior probability (includes activity, authority etc.) of user u, $P(q|u)$ is a probability that question q is generated from the user profile u (it models the degree of expertise of user u on question q). Due to the fact that $P(q)$ is a probability of generating question q, which is the same for all users, we can omit it during the following calculations. From the obtained probabilities, the ranked answerer list is created where the first record represents the most probable user to give an answer to the routed question.

Probability of Generating Question from User Profile. We compute probability $P(q|u)$ as a linear combination of two aspects: a probability derived from user QA profile $\theta_{u\text{-}qa}$ and user non-QA profile $\theta_{u\text{-}non\text{-}qa}$.

$$P(q|u) = \alpha\, P\left(q\middle|\theta_{u-QA}\right) + (1-\alpha)\, P\left(q\middle|\theta_{u-non-QA}\right) \quad (2)$$

If one of user profiles is missing (when user u has not answered any question yet or when user u does not have any non-QA data), the probability $P(q|u)$ equals to the probability derived from the second existing user profile. The individual importance of each user profile is determined by a weighting coefficient α, which depends on a number of user's previous QA activities $|QA|$. The underlying idea, why we employ the dynamic coefficient α, is that the user non-QA profile should play higher importance for users with a smaller number of previous QA activities. As the user QA profile will aggregate more question profiles, also its influence will grow (note that user profiles will be taken into consideration with the same weight when a user has 5 previously answered questions).

$$\alpha = \frac{1}{1+e^{-0.25(|QA|-5)}} \tag{3}$$

Each word in the question is expected to be generated from both user profiles independently (in this calculation, we utilize the bag-of-words model from question profile θ_q). Therefore probability $P(q|\theta_u)$ of generating question q from both user profiles is calculated as:

$$P(q|\theta_u) = \prod_{w \in \theta_q} P(w|\theta_u)^{n(w,\theta_q)} \tag{4}$$

where $P(w|\theta_u)$ is a probability of generating word w from user profile θ_u and $n(w,\theta_u)$ means how many times word w occurs in question q. Finally, the probability of generating word w from user profile θ_u can be obtained as:

$$P\left(w\middle|\hat{\theta},\hat{\phi},\theta_u\right) = \sum_{z=1}^{K} P\left(w\middle|z,\hat{\phi}\right)P\left(z\middle|\hat{\theta},\theta_u\right) \tag{5}$$

where $\hat{\theta}$ and $\hat{\phi}$ are the posterior estimates of θ and ϕ.

Prior Information of User. In the proposed probabilistic model, we decided to consider users' level of activity in the system as prior probability of user $P(u)$. Some users could be highly active for a certain time and consequently become completely inactive. Therefore, users who are active recently should be preferred. In calculation of user prior probability, we adapt an approach which was utilized also in the previous studies [4], [9]:

$$P(u) = exp^{-(t_q - t_u)} \tag{6}$$

where t_q refers to the question posting time and t_u is the most recent time when user u participated in the question answering process (i.e. posted a question, an answer or a comment).

IV. Experimental Evaluation

A. Experimental Setup

In order to evaluate the proposed method, we conducted an offline experiment in which we used a dataset collected from CQA system Android Enthusiasts, which is a part of Stack Exchange platform. It contains data from May 2009 to December 2014. During this period, about 26,000 questions and 33,000 answers were posted concerning with various topics related to Android operating system. From the total number of questions, about 18,900 questions contain at least one answer.

On the basis of the structure of the experimental dataset, we recognized three possible sources of non-QA data, particularly a personal description (a.k.a. About me), Homepage and Twitter. After further analysis of quantity of individual non-QA data sources (see Table I), we decided to use one internal non-QA source (About me) and one external non-QA source (Homepage). Approximately 14,000 out of more than 21,000 homepages can be used as a suitable non-QA data source because some users provided just general webpages (e.g. www.google.com or www.stackoverflow.com). Consequently, we implemented a web crawler, which downloaded HTML source code from each of these homepages and preprocessed it in order to create non-QA data profiles.

TABLE I. Quantity of non-QA data sources in the dataset.

Source of non-QA data	Number of users
About me (internal)	21,541
Homepage (external)	21,703
Twitter (external)	1,028
About me + Homepage	10,073
About me + Twitter	275
Homepage + Twitter	726
About me + Homepage + Twitter	10,668

For all questions included in the dataset, we created question profiles. Consequently, we selected those questions in which at least one user with non-QA data provided an answer. Finally, we calculated ranked lists of recommended answerers for these questions by matching user and question profiles. Some previous studies (e.g. [20]) involved in question routing only those users who actually provided answers on a particular question and try to identify the best answerer among them. In contrast to this approach, we included in the ranked list all relevant users from the community. This approach corresponds to a real situation in CQA systems when we want to recommend the routed question to top-k users.

To obtain a ground truth of appropriate answerers for a particular question, we followed the majority of previous studies (e.g. [4], [5], [21]) and utilized a list of users who actually provided an answer on this question. However, we are aware that this approach does not completely correspond to the real interest from users. Especially, as soon as a question receive at least one high-quality answer, other suitable candidates can express their expertise just by providing a positive vote or simply by skipping the routed question and attempting to answer another one (unfortunately, voting and question views are anonymous and thus they are anonymized also in all Stack Exchange datasets). In spite of this drawback, the list of actual answerers can be still considered as a fair precise ground truth for question routing.

Implementation and evaluation of the proposed method is based on an experimental infrastructure developed as a part of educational and organizational CQA system Askalot [22]. This experimental infrastructure can simulate events (e.g. question and answer creation) ordered by time when they actually occurred. During question and non-QA data profiling, Stanford CoreNLP tool was utilized to perform text preprocessing (i.e. tokenization, stop-word removal and lemmatization). Consequently to identify latent topics, we employed LDA implementation by Blei, Ng, and Jordan and calculate topic distributions for all documents at once (the number of LDA topics was empirically set to 20). In online settings, it would be possible to use an online implementation of LDA (e.g. [23]) in order to calculate latent topics incrementally without necessity to recalculate all previous question profiles.

B. Evaluation Metrics

To measure performance of the proposed method, we used two ranking metrics: Mean Reciprocal Rank and Precision@n.

Mean Reciprocal Rank (MRR) is an average from all routed questions Q of the rank $rank_i$ at which the first relevant (actual) answerer was returned, or 0 if the recommended list of answerers does not contain any relevant answerers (e.g. due to the absence of their both user profiles).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{7}$$

Precision@n (P@n) reports the fraction of recommended answerers U_r at top-n positions that are labelled as relevant. In comparison with MRR, which measures actual ranking of recommended users, P@n measures overall potential success of receiving a correct answer if we recommend the routed question to top-*n* answerers. Based on the prior works, we use values 5 and 10 as *n*.

$$P@n = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{|U_r|}{n} \tag{8}$$

C. Methods Compared

In order verify our assumption that non-QA data can supplement QA data during question routing, we compared the performance of three possible variants of the proposed question routing method. In the first variant, which represents a baseline, we considered only QA data and thus the probability $P(q|u)$ was calculated only with user QA profiles $\theta_{u\text{-}qa}$. The second variant was in a similar way limited to non-QA data. Finally, the last full variant takes advantage of combination of QA and non-QA data. Please, note that our contribution is in exploring a potential of non-QA data and thus we did not compare the achieved performance with other question routings methods (e.g. language models).

D. Experimental Results

The experimental results are organized into three parts, each of them focuses on evaluation of a particular aspect of non-QA data utilization in question routing.

Consistence between Non-QA and QA Data. At first, we evaluated how well user non-QA profiles correspond to QA profiles. For all users with non-QA data and with at least three provided answers, we calculated user QA profiles from all previously answered questions. Consequently, we measured a similarity between user QA and non-QA profiles by means of cosine similarity. The obtained similarity value 0.119 indicates that topics included in non-QA data at least partially correspond to those that can be derived from QA activities. In addition, the non-QA profiles of more active users correspond to QA activities better (e.g. cosine similarity for users with at least 10 answers is 0.166). This similarity is an important preliminary indicator of the potential of non-QA data to estimate users' expertise. We found out that highest similarity was achieved when non-QA data sources (i.e. about me and homepage) are combined together with the same weight and thus in the following evaluation, we used ordinary arithmetic average to create non-QA data profiles.

Earlier Estimation of User Expertise. Secondly, we investigated how consideration of non-QA data can improve a prediction whether a user with a small number of QA activities will be a relevant candidate to provide an answer. From all routed questions, we selected answerers with non-QA data and grouped them according to various amounts of their previous answers (i.e. amounts of question profiles that were aggregated in their user QA profiles). Consequently, we evaluated in how many cases the non-QA and full variant of our method predicted for an answerer a higher probability $P(u|q)$ and a higher rank in comparison with the QA variant (see Table II).

The results revealed a quite unexpected finding that both variants of the method can improve ranking not only for users who have a low amount of activity (one answer), as we originally hypothesized, but also for users with a great amount of activity (5 or 10 answers). This result can be explained by the fact that highly active users are not actually so consistent in topics they provide answers on what finally undesirable affects and hamper QA-based question routing methods. Therefore, non-QA data, which are naturally more stable, are able to improve answerers' positions even in more than 20% of cases.

TABLE II. COMPARISON OF IMPROVEMENT IN ANSWERERS' PROBABILITY AND RANKING WITH VARIOUS AMOUNTS OF ANSWERS |*A*|.

|A|	**Non-QA**		**QA + Non-QA**	
	P(u|q) % imp	***Rank % imp***	***P(u|q) % imp***	***Rank % imp***
1	24.36	30.34	23.50	11.54
2	10.45	19.40	9.70	4.48
3	10.16	14.06	10.94	5.47
4	6.19	18.56	7.22	2.06
5	7.27	25.45	15.45	8.18
10	4.29	21.43	10.01	22.86

[a] *"P(u|q) % imp" and "Rank % imp" denotes the proportion of answerers for whose consideration of non-QA data was able to improve probability P(u|q) and the recommended rank respectively.*

Higher Overall Precision of Question Routing. The finding from the previous part indicates that non-QA data can improve ranking for users with low as well as high level of activity. Now we evaluate influence of non-QA data on the overall precision of question routing. For all obtained ranked lists, we measured the positions of actual answerers (see Table III).

TABLE III. COMPARISONS OF THE POSITIONS OF ACTUAL ANSWERERS IN THE RANKED LISTS.

Metric	**QA**	**Non-QA**	**QA + Non-QA**
MRR	0.0269	**0.0320**	0.0242
P@5	0.0095	**0.0318**	**0.0104**
P@10	0.0292	**0.0494**	0.0224

[a] *Numbers highlighted in bold indicate that particular variant overcomes baseline (QA variant).*

From the results, we can derive several interesting findings. In general, the performance of QA variant according P@n is very similar or even slightly better in comparison with the prior works (e.g. [7]), while performance according to MRR metric is slightly lower. MRR metric in our approach is, however, influenced by a significantly larger number of recommended users as the previous studies focus only on a small fraction of community (with more than 10, 15 or even 20 answers as we stated in Section II). It is really surprising that the variant, which considers only non-QA data, was able to overcome the QA variant in all three metrics. It means that just the presence of non-QA data and their topic distribution provides enough information to overcome QA data, which are very sparse and inaccurate for many users. MRR value of 0.032 means that on average each question will get answered if we will route it to the top 30 users. We discovered that also the full variant was able to outperform QA variant in P@5 although it achieved slightly worse results for other two metrics. Nevertheless, we can finally conclude that consideration of non-QA data leads to the overall improvement in the precision of question routing.

V. Conclusion

In this paper, we proposed a question routing method specifically designed to utilize non-QA data as a supplement to QA activities (asking or answering questions) in estimation of users' expertise. This approach differs from the prior works, which focus primarily on QA data, in a scope of users who can be considered as possible candidates to provide an answer. The underlying idea is that different sources of non-QA data (social networking sites, blogs, etc.) may help to engage also users who are new or passive in the CQA system, what will finally contribute to the more successful question answering process as well as to the better sustainability of the CQA ecosystem.

We introduced a probabilistic model that calculates how likely a particular user will be a suitable candidate to provide an answer on the routed question. In our approach, questions and non-QA data are represented by their profiles based on the topical distribution derived from the LDA latent topic model. In contrast to the most existing topic-based question routing methods, we derived user profiles incrementally from question profiles for which corresponding users provided answers on. This solution allows us to update a user profile without re-profiling all other user profiles in the model.

We implemented the proposed method within the experimental infrastructure in CQA system Askalot created as a part of our previous work. We conducted an offline experiment to verify a potential of non-QA data to supplement QA data. The obtained experimental results revealed an improvement in the ranking of users with low as well as high amount of previous activity in the system and also in the overall precision. These results confirm our assumption that incorporating non-QA data in question routing will lead to a better recognition of low-activity user as suitable answerers.

Acknowledgment

This work was partially supported by grants No. VG 1/0752/14, VG 1/0646/15 and KEGA 009STU-4/2014.

References

[1] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor, "When web search fails, searchers become askers," in *Proc. of the 35th int. ACM SIGIR Conf. on Research and Development in Information Retrieval - SIGIR '12*, 2012, pp. 801–810.

[2] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proc. of the 19th ACM Int. Conf. on Inf. and Knowl. Management - CIKM '10*, 2010, pp. 1585–1588.

[3] T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in *Proc. of the 21st Int. Conf. Comp. on World Wide Web - WWW '12*, 2012, pp. 783–790.

[4] M. Liu, Y. Liu, and Q. Yang, "Predicting Best Answerers for New Questions in Community Question Answering," in *Proc. of the 11th Int. Conf. on Web-age Inf. Management*, 2010, pp. 127–138.

[5] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in *Proc. of the 17th Int. Conf. on Inf. and Knowl. Management - CIKM '08*, 2008, pp. 921–930.

[6] G. Zhou, K. Liu, and J. Zhao, "Joint relevance and answer quality learning for question routing in community QA," in *Proc. of the 21st ACM Int. Conf. on Inf. and Knowl. Management - CIKM '12*, 2012, pp. 1492–1496.

[7] Z. Ji and B. Wang, "Learning to rank for question routing in community question answering," in *Proc. of the 22nd ACM Int. Conf. on Inf. and Knowl. Management - CIKM '13*, 2013, pp. 2363–2368.

[8] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proc. of the 21st Int. Conf. Comp. on World Wide Web - WWW '12*, 2012, no. i, pp. 791–798.

[9] Y. Tian, P. S. Kochhar, E.-P. Lim, F. Zhu, and D. Lo, "Predicting Best Answerers for New Questions: An Approach Leveraging Topic Modeling and Collaborative Voting," in *Proc. of Social Informatics 2013 Int. Workshops*, 2013, pp. 55–68.

[10] L. Fang, M. Huang, and X. Zhu, "Question routing in community based QA: Incorporating Answer Quality and Answer Content," in *Proc. of the ACM SIGKDD WS on Mining Data Semantics - MDS'12*, 2012, pp. 1–8.

[11] B. Li, "A Computational Framework for Question Processing in Community Question Answering Services," Doctoral Thesis, The Chinese University of Hong Kong, 2014.

[12] I. Szpektor, Y. Maarek, and D. Pelleg, "When Relevance is not Enough : Promoting Diversity and Freshness in Personalized Question Recommendation," in *Proc. of the 22nd Int. Conf. on World Wide Web - WWW'13*, 2013, pp. 1249–1259.

[13] I. Srba and M. Bielikova, "Why Stack Overflow Fails? Preservation of Sustainability in Community Question Answering," *IEEE Software*, Submitted, 2015.

[14] Y. Pan, L. Luo, C. Chi, and Q. Liao, "To answer or not: What non-QA Social Activities Can Tell," in *Proc. of the 2013 Conf. on Computer Supported Cooperative Work - CSCW'13*, 2013, pp. 1253–1263.

[15] L. Luo, F. Wang, M. X. Zhou, Y. Pan, and H. Chen, "Who have got answers?: growing the pool of answerers in a smart enterprise social QA system," in *Proc. of the 19th Int. Conf. on Intelligent User Interfaces - IUI '14*, 2014, pp. 7–16.

[16] Z. Liu and B. J. Jansen, "Predicting potential responders in social Q&A based on non-QA features," in *Proc. of the 32nd Annual ACM Conf. on Human Factors in Comp. Systems - CHI EA '14*, 2014, pp. 2131–2136.

[17] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert Finding for Question Answering via Graph Regularized Matrix Completion," *IEEE Trans. on Knowledge Data Engineering*, vol. 27, no. 4, pp. 993–1004, 2015.

[18] J. Yang, K. Tao, A. Bozzon, and G.-J. Houben, "Sparrows and Owls: Characterisation of Expert Behaviour in StackOverflow," in *User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science*, 2014, vol. 8538, pp. 37–48.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, pp. 993–1022, 2012.

[20] F. Xu, Z. Ji, and B. Wang, "Dual role model for question recommendation in community question answering," in *Proc. of the 35th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval - SIGIR '12*, 2012, pp. 771–779.

[21] H. Zhu, E. Chen, and H. Cao, "Finding experts in tag based knowledge sharing communities," in *Proc. of 5th Int. Conf. KSEM 2011*, 2011, vol. 7091 LNAI, pp. 183–195.

[22] I. Srba and M. Bielikova, "Askalot: Community Question Answering as a Means for Knowledge Sharing in an Educational Organization," in *Proc. of the 18th ACM Conf. Companion on Computer Supported Cooperative Work & Social Computing - CSCW'15*, 2015, pp. 179–182.

[23] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola, "Scalable distributed inference of dynamic user interests for behavioral targeting," in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge discovery and Data Mining - KDD '11*, 2011, pp. 114–122.