Authority vs Affinity: Modeling User Intent in Expert Finding

Suratna Budalakoti and K. Suzanne Barber
Laboratory for Intelligent Processes and Systems
The University of Texas at Austin
1 University Station C5000, Austin, TX 78712, USA
Email: {sbudalakoti, barber}@lips.utexas.edu

Abstract—This paper considers the problem of recommending experts for a question. The problem of expert finding has been investigated in the past, within the context of information retrieval, where the focus has been on measuring expert authority. However, we treat the problem as a recommendation problem in the context of social search, that is, expert recommendations are personalized for a questioner based on his/her history. The intuition is that questioners look not only for authoritative responders, but for a combination of authority and personal trust or affinity. The significance of these factors depends on user intent in asking the question, and varies with each question, and with questioner personality. An iterative algorithm is proposed to model user intent for a question in terms of whether they are looking for authoritative experts, or individuals in their social networks, and an expert recommendation algorithm based on it is tested on data crawled from the Yahoo! Answers website.

I. INTRODUCTION

While search engines have been the dominant mode of user interaction with the internet, question and answer(Q&A) forums have always been a significant component. Apart from specialized Q&A forums dedicated to specific areas, general-purpose Q&A forums such as Yahoo! Answers \(^1\) and Answerbag \(^2\) combined generate over 3 million questions each month. These forums enable users to seek expertise and advice on topics where ordinary search results may not meet expectations, for example in scenarios requiring personalized information, recommendations and advice/opinions. Despite the success of these forums, the method used to find responders for various questions is generally primitive. The common approach is broadcast: each question is posted on the forum for all users to see, with the expectation that interested users will answer the question.

This approach is wasteful in terms of the time investment required by the responders, as they need to read through all the questions on the forum to identify those of interest to them. It also discourages highly specialized questions from being asked or answered, as specialists in highly specific areas would have to read through an extremely large number of questions to find the few of interest to them. An alternative to broadcast is a recommendation-based approach where potential responders are recommended questions, based on their interests. Traditionally, the broadcast approach has dominated Q&A websites, although expert finding has received attention in the context of enterprise search, that is, searching data within an organization \(^9\).

More recently, there have been attempts to move away from a broadcast model towards a recommendation-based one in online forums as well: two commercial examples being Fluther\(^3\) and Aardvark\(^4\) (recently acquired by Google\(^5\)). An impetus behind this interest is the evolution of the web from a source of information to a venue for social interaction. As people use the internet more and more to connect with their friends, it is natural that they should turn to them when they need information. A recent survey \(^1\) at Microsoft found that around 50% had used the status box of their social network page to ask a question. The survey also found that users preferred information from sources they knew personally for certain categories of questions such as recommendations, personal opinions, and occasionally even factual information. This finding points to the potential of expert search in social networks, especially as this behavior is a user innovation, and the primary purpose of the status box is not to ask questions. Horowitz and Kamwar \(^3\) characterize this as the ‘village paradigm’ of information exchange, as opposed to the historical ‘library paradigm’ of internet information retrieval, and place this type of information-seeking under the ambit of social search.

Social search is a broad term that can encompass all methodologies that take into account the social relationships of users while finding information of relevance to them. Traditional search is based on the assumption that users prefer information from authoritative sources. Social search is based on a different assumption, that users prefer information not only from authoritative sources, but also from sources that they know personally and trust, or have some kind of personal affinity for. This is true not only of online communities where users personas can be mapped to identities in the real world, but also in pseudonymous social networks such as Yahoo! Answers or Flickr\(^6\). People

\(^1\)http://answers.yahoo.com
\(^2\)http://www.answerbag.com
\(^3\)http://www.fluther.com
\(^4\)http://vark.com
\(^5\)http://www.google.com
\(^6\)http://www.flickr.com
interacting on these forums also form virtual social groups and relationships, which may differ from social network relationships mainly in being pseudonymous. Despite this difference, concepts such as authority and personal affinity or trust that are applicable to social network relationships, are also applicable to Q&A forums. The main contribution of this paper is a new algorithm for exploring of the trade-off between source authority in the topic and personal affinity to the source that users make when searching for information in Q&A forums and social networks.

II. RELATED WORK

The assumption behind traditional information retrieval approaches is that users prefer information from authoritative sources. However, the concept of authority is itself defined only in context of the hyperlinked collection of documents being analyzed: it is not an objective concept in the sense that a source is authoritative because it has been independently verified as corresponding with external reality. Authority in popular link analysis approaches [4][5] is conferred, in one way or the other, by other pages via links or references: it is only the consensus arrived at with respect to the quality or trustworthiness of a page by all the participants in the system. For example, Page et. al’s random surfer model [5] defines intuitively a page’s authority as the probability a random web surfer would visit the page given that he/she starts at a random page, and selects a random outlink at each timestep. Similarly, in the HITS algorithm [4], a webpage’s authority score is equal to the sum of the ‘hub’ scores of the pages that out-link to it.

Social search attempts to combine the concept of authority with the assumption that, given a question/query, people value the responses from their friends, or people they trust, provide, or people who share their interests. We loosely term this as the social affinity between the questioner and prospective responders. This could cover the relationship between the two users across topic categories, or the expected compatibility between them based on their profile information. Horowitz and Kamvar [3] at Aardvark measure the 'connectedness' probability between two users as a function of a number of factors such as their social connection, demographic similarity, vocabulary match, verbosity match, etc. The expert recommendations by Aardvark are made by multiplying this value with the probability of success based on the responders’ expertise in the topic. However this may not be the best approach as, as studies of search engine logs [2] have shown, not all search engine queries benefit from personalization. The same intuition carries to personalization in Q&A settings: the degree of personalization required for a question depends on the nature of the question and the intent of the questioner in asking the question. This paper tries to address this problem of predicting the degree of personalization required for a question. We base our predictions on a model of the question topic and type, and questioner personality. Questions in some topics require more personalization than others. Similarly different questioners desire different levels of personalization.

A broad dichotomy commonly recognized among questions asked on Q&A forums is that between informational questions and conversational questions. Harper et al [10] define informational questions as those asked with the intention of obtaining information, and conversational questions as those asked with the intent of stimulating discussion. Morris et al [1], in their study of questions posed on social network status boxes, divide questions into eight different categories. The largest four categories: recommendation, opinion, factual knowledge, and rhetorical, account for 82% of the questions. Opinion and rhetorical can be classified as questions with conversational intent, and account for 36% of the questions, while factual knowledge, or informational questions, are 17% of the total. The largest group, recommendations (29%), involves components of both informational and conversational intents.

Intuitively, a user’s intention in asking a question is an important component in the type of responder he is seeking. In seeking responses for informational questions, he may be less concerned with the compatibility, or his social relationship with the responder, as opposed to conversational questions, where he is likely to be more concerned. For questions that involve recommendations, he may prefer either, or a combination of both, depending on the type of recommendation being sought, and his personality.

III. MODELING QUESTIONER INTENT

Consider the following scenario: there is a set of users U in a question-answer forum, who can take the role of questioner and responder, depending on the situation. Each time a user asks a question, he receives responses from a set of responders. He rates a subset of these responses as satisfactory, and closes the question, so that no more responses are possible. We also assume each question belongs to a single topic, and this topic is known. The central assumption of this paper is that a questioner asks a question with the intent of seeking one of two things: expertise in a topic, or interaction from socially proximate individuals. So all ratings by a questioner are based on one of two factors:

1) **Responder expertise**: The responder’s expertise in a topic determines the quality of the content produced by him/her.

2) **Social Affinity**: The social affinity between two users is independent of the topic at hand, and depends on the relationship the questioner and responder share across all topics, or even in the real world outside the social network.

The trade-off between these two factors depends on a variable called **objectivity** associated with each questioner. A more objective questioner seeks experts in the topic, while
a less objective one is more interested in what socially proximate responders have to say.

A reasonable definition of objectivity would be the degree to which a questioner’s preferences can be explained by assuming he is seeking experts in the topic. To apply this definition, we need to define expertise, which we define for a responder in a topic as the consensus among the objective questioners about the quality of responses from him in the topic. As these definitions are recursive, we derive an iterative algorithm to arrive at expertise and objectivity values for responders and questioners respectively.

In the next section we provide a formal definition of responder expertise. Following that, we provide algorithms to estimate the degree to which a questioner in a topic prefers expertise to socially close responders, or his/her objectivity.

IV. RESPONDER EXPERTISE

A. Expertise Strength: Definition

A straightforward measure of the expertise level, or the strength, of a responder in a topic is the probability that a question in the topic will be provided a satisfactory response by the responder [11]. However, this definition has the drawback of treating all questions as equally difficult. In practise, some questions are at a higher difficulty level than others, and a good measure of expertise should be able to take this fact into account.

Suppose we had a measure for the level of difficulty of each question in a topic. Then we can define the topic strength of a responder as the fraction of questions the responder has answered in the topic, each multiplied by the difficulty level of the question.

We still need a measure of the difficulty level of each question: we define this for a question as the strength of the questioner within that topic. This seems reasonable as a user would ask questions that are at least harder than his/her level of knowledge, but not so hard as to not be interesting to him/her.

Then, let the strength (or authority) of user i in a topic be written as $a_i$, and the number of times user i answered a question by user j satisfactorily be written as $r_{ji}$. Let $N_Q$ be the total number of questions in the topic. Then $a_i$ can be written as follows:

$$a_i = \sum_{j=1}^{N} \frac{a_j}{N_Q} r_{ji}$$  \hspace{1cm} (1)

where $N$ is the number of users. Now, let the number of questions asked by user j be written as $q_j$. Then, after normalizing with the total strength of all the questions asked, we can rewrite $a_i$ as follows:

$$a_i = \frac{\sum_{j=1}^{N} q_j \cdot a_j \cdot p_{ji}}{\sum_{j=1}^{N} q_j \cdot a_j}$$  \hspace{1cm} (2)

where $p_{ji}$ is the fraction of questions by j answered by i. Dividing both numerator and denominator by $N_Q$, we get:

$$a_i = \frac{\sum_{j=1}^{N} \frac{q_j}{N_Q} \cdot a_j \cdot p_{ji}}{\sum_{j=1}^{N} \frac{q_j}{N_Q} \cdot a_j}$$  \hspace{1cm} (3)

Interpreting $\frac{q_j}{N_Q}$ as the probability that user j will ask a question, written as $P_q^j$, we get:

$$a_i = \frac{\sum_{j=1}^{N} P_q^j \cdot a_j \cdot p_{ji}}{\sum_{j=1}^{N} P_q^j \cdot a_j}$$  \hspace{1cm} (4)

B. Absorbing Random Walk Interpretation of Expert Strength

We can write this in matrix form: let $Q$ be a diagonal matrix, where $Q(i,i) = P_q^i$, let $P$ be a matrix such that $P(i,j) = p_{ij}$, and let $a$ be a vector corresponding to $a_{i...N}$ above. Then the above equation can be written as:

$$(QP)^T a = a$$  \hspace{1cm} (5)

We can add a small prior vector to P, which signifies a small probability that any user can answer another user’s question, even if there is no evidence in the data for it. To ensure a probabilistic interpretation, we also update $Q$, by adding 1 to the denominators of $Q(i,i)$. The prior vector $a$ can be written as $er^T$, where $e_i = 1$ for all i, and r sums to 1. We also restrict $a$ to vectors that sum to 1. Then we can rewrite the above equation as:

$$(QP + er^T)^T a = a$$  \hspace{1cm} (6)

$$\Rightarrow (QP)^T a + re^T a = a$$  \hspace{1cm} (7)

$$\Rightarrow a = (I - (QP)^T)^{-1} r$$  \hspace{1cm} (8)

$$\Rightarrow a = ((I - QP)^{-1})^T r$$  \hspace{1cm} (9)

$$\Rightarrow a = (I - T)^{-1} r$$  \hspace{1cm} (10)

where $T = QP$.

As all rows of matrix T sum to less than 1, we can interpret T as the transition matrix for a reducible Markov chain [14] with N+1 states as follows: we create an extra absorbing state Z, which is the exit state. Z is a recurrent state: the system, if in state Z, transitions to Z with probability 1 at each timestep. On the other hand, at a timestep, if the system is currently in state i, it transitions to the state Z with a probability $1 - Q(i,i)$, and to another state j with probability $Q(i,i) \times P_{ij}$. Interpreting T as an absorbing Markov chain, it is easy to see that $A = (I - T)^{-1}$ is the definition of fundamental matrix of an absorbing Markov chain [14], that is $A = I + \sum_{i=1}^{\infty} T^i$.

Then, if a random walk is executed across the absorbing chain, $a_{ij}$ is the expected number of visits to state j before exit, if the walk started in state i, if the transition probability
from i to j \( P^q_{ij} \), and the exit probability is \( 1 - P^q_{ij} \). Then, as \( r \) is a probability vector, \( a = A^T r \) gives the expected time spent in each state, if the initialization probability of the walk at vertex \( i \) is given by \( r \). We refer to \( A \) as the expected visits matrix, and \( a \) as the expected visits vector, or the authority vector, for a given initialization vector \( r \).

It is possible to replace the diagonal values of \( Q \) with values different from the number of questions asked \( (P^q_{ii}) \) in the above equations. The drawback of the current definition of \( Q \) is that a questioner’s influence in determining responder strength is directly proportional to the number of questions asked by him. This could lead to an overemphasis on the number of questions asked, at the expense of the quality of the questions, and it would be possible for questioner to ‘crowd out’ the influence of other questioners, simply by asking a lot of questions, irrespective of the quality of the questions. For this reason, we use a sigmoid function on the number of questions to define the \( Q \) matrix, so that questioner influence levels of after he has asked a certain number of questions. This is discussed in further detail in the next section.

An option for setting the values of \( Q \) would be to fix all \( Q_{ii} \) to some constant \( c \), and modify \( r \), the initialization vector. This would mean that we take \( q_i \) into account only when initializing the random walk, and ignore it after that. In the next section, we show that the vector \( r \) can be interpreted as a generalization of the pagerank vector equation: specifically, we show that that for \( Q(i,i) = c \) for all \( i \), where \( c \) is a constant, the value of vector \( a \) above corresponds to the pagerank vector [5].

An intuitive interpretation of the expected visits vector \( r \) is as a voting process, as follows: everytime a questioner \( i \) asks a question, we ask him/her to recommend a responder who he/she believes is an expert in the question topic. Assume the questioner picks a responder at random from his preference distribution \( P_i \). So, the user recommends a responder \( j \) with probability \( P_{ij} \). We accept this recommendation with a certain probability \( Q(i,i) \). If we accept the recommendation, we now ask user \( j \) to recommend a responder, a recommendation we accept with probability \( Q(j,j) \), and so on. If we do not accept a recommendation, the process exits. Then the vector \( s \) represents the number of votes we expect each user to get, using this process.

C. Comparing \( a \) to the Pagerank Vector

Pagerank [5] is a popular algorithm for link analysis over a collection of hyperlinked documents. It treats the collection of documents as a graph, with each document as a vertex. A hyperlink from document \( i \) pointing to document \( j \) in the collection is represented by a directed edge. The weight of the edge is usually defined as \( \frac{1}{h_i} \), where \( h_i \) is the total number of outgoing links from document \( i \). This can be written as a transition matrix \( M' \), where \( M'_{ij} \) is the weight of the link from \( i \) to \( j \). To ensure that the matrix is stochastic, the transition matrix is usually updated to \( M = d M' + (1-d)E \). Here, \( E = e v^T \), where \( v \) is a probability vector, and \( e \) is an \( n \)-vector with \( e_i = 1 \).

The pagerank vector is then calculated by solving \( x = M^T x \). The common approach to solving this is via an iterative method. However, solving \( x = M^T x \) algebraically, as also described in [7] gives:

\[
\begin{align*}
x &= d M^T x + (1 - d)v \\
\Rightarrow (I - d M^T)x &= (1 - d)v \\
\Rightarrow x &= (1 - d)(I - d M^T)^{-1} v
\end{align*}
\]

The above equation can be rewritten as:

\[
x = (1 - d)((I - d M^T)^{-1})^T v
\]

In the above equation, setting and \( P \equiv M' \), \( v=\bar{r} \), in accordance with equation (9), we get:

\[
x = (1 - d)((I - d P)^{-1})^T r
\]

Comparing equation (15) to equation (9), it can be seen that \( a \) differs from the pagerank vector \( x \) only by a constant \( 1 - d \) in the special case when \( Q_{ii} = d \). Hence the expected visits vector \( a \) can be seen as a generalization of the pagerank vector: \( s=x \) when all questioners are weighed equally, irrespective of the number of questions asked. This seems reasonable for webpages, but for Q&A forums, questioners who ask a lot of questions in a topic are more likely to be seriously interested in the topic, and likely to be better judges of response quality. It would be useful to have this effect...
level off at some point, so that users cannot increase their
influence as questioners simply by asking a lot of questions.
For this reason we use a sigmoid function to set $Q$. We set
$Q(i, i) = \frac{1}{1 + e^{-\text{number} \cdot \alpha}}$. This means that for questioners who
have asked 100 or more questions $Q(i, i)$ is effectively equal
to 1.

D. Defining Objectivity

Given the responder strength/authority vector, we need
an estimate of the objectivity of a particular selection by
a questioner? For this purpose, we define objectivity as
follows: the probability that a selection by a questioner $q$
is objective, is equal to the probability that any other
questioner, say $q'$, (drawn from a distribution) would have rated
positively the same responder. The distribution (called $\alpha$) the
user $q'$ is drawn from defines the probability of a user being
drawn/chosen as proportional to the estimated expertise
of the user $q'$ (in his role as responder). This is based on
our initial assumption that responders rated positively by
stronger experts are more likely to be experts themselves.
Hence, the greater the extent to which a responder agrees
with expert users, the more objective he is estimated to be.
Notice that this definition of objectivity does not consider
only the lack of subjectivity of the questioner (in which case
he should be expected to make selections close to those made
by users at his expertise level). The definition, instead, is
closer to the concept of ‘hub score’ in the HITS algorithm
[4], that is, it takes into account whether the questioner asks
questions that interest high quality responders, and whether
he is able to identify and select them from the responses he
receives.

We also weigh the distribution $\alpha$ by the weight attached
to each user, based on number of questions asked, or $Q$. Based
on this, the distribution $\alpha$ can be written as $\alpha = (QP)^T \alpha$.
However, as defined in equation (5), $a = (QP)^T a$. Hence
$\alpha = a$, normalized. Based on this, we call $\alpha$ the authority
distribution, and an objective responder is one whose selec-
tions match draws from this distribution.

The next section discusses the iterative algorithm that
alternatively estimates questioner objectivity and responder
authority.

V. EXPERTISE ESTIMATION ALGORITHM

A. Background

Assuming we have a dataset of Q&A documents $D$
consisting of $N$ documents $D = \{d_1, d_2, \ldots, d_N\}$. We
assume that we already know the topic for each question: this
may be true in cases where users supply tags for questions,
or if questions are grouped into topics in advance (as is
the case at a very broad level in Yahoo Answers). If not, a
standard clustering algorithm such as the k-means algorithm
[15] can be used to cluster documents into topics. The
following algorithm is run separately for each topic, so we
do not need notation to refer to each topic separately.

The dataset is generated by users $U = u_1, u_2, \ldots, u_{N_u}$
taking on the role of questioners and responders, where
$N_u$ is the total number of users. Each document $d_i$ has a
questioner, referred to as $q_i$, and a responder vector $r_i$, where
$r_{ij} = 1$ if responder $j$ was rated positively for question $i$.
We also construct a matrix $K$, so that $K(i, j) = 1/N_u^q$
for all $j$, where $N_u^q$ is the total number of questions asked
by user $u_i$. Each time a questioner rates an answer by a
responder as satisfactory, we call it a selection. A questioner
may positively rate more than one response for a questioner,
so the total number of selections in a topic can be more than
the number of questions.

We also assume that each user $q$ has a social affinity based
distribution $\omega_q$, which is known. There is also a user-specific
probability $\omega_q$, with which the user picks an expert to rate
positively, everytime he has to rate responses to a question.
When calculating the expertise distribution $\alpha$ for a topic,
we would like to attach more importance to questioners
with a higher value of $\omega_q$. This is based on the assumption
that questioners who select authoritative responders more
frequently are more likely to be interested in the topic, and
better judges of the quality of responses in the topic.

Another way to understand this is: ideally, we would have
preferred to assign a single random variable, say $o_q$, to each
questioner, so that $o_q = 1$ if the questioner is objective, and
$o_q = 0$, if he is not. Then, assuming we knew the values
of $o_q$ for all questioners, if we set $O = \text{diag}(o_1, \ldots, o_N)$,
and $T = OQP$ in equation (10), we would only consider
objective responders ($o_q = 1$) when calculating the expertise
strength vector $s$ (which is the same as $\alpha$). However, in
practise, for cases where questioners by some users can run
into hundreds or even thousands, a single random variable
could be too restrictive. So, instead, we assume a probability
of being objective $\omega_q$ associated with each questioner, and
set $O = \text{diag}(\omega_1, \ldots, \omega_N)$. 

Figure 2. Generative Model for Question Answer Recommender
B. Algorithm Outline

Following initialization, the algorithm iteratively executes two main steps, the objectivity estimation step, and the authority estimation step. A likelihood step is executed between these two steps, which calculates the log likelihood of the data based on the objectivity estimation step. The algorithm terminates when the likelihood stops increasing. The algorithm is given in Figure (1).

During the objectivity estimation step, an objectivity score is assigned to the selection made for each question $d_i$, based on an estimate of their likelihood of being drawn from the two candidate distributions: the current questioner $q_i$’s social affinity based distribution $\sigma_{q_i}$, and the current estimate of the authority distribution $\alpha$. In the expertise estimation step, the probability $\omega_{q_i}$, of each questioner being objective for a given question, is estimated for each questioner. Based on these estimates, the expertise probability vector $\omega$ is recalculated, reweighing each row of the transition probability matrix $P$ by the objectivity of the corresponding questioner.

The algorithm can be understood in terms of the Random Surfer model as follows: in Page et al.’s [5], there is a constant probability with which a surfer teleports to another random page (or vertex in a web graph) is constant. In our model, the probability of randomly teleporting from a given vertex $q$ is $\omega_q$. In other words, the higher the estimated objectivity of the current vertex, the greater the probability that an outgoing edge from the current vertex will be followed. The lower the estimated objectivity, the higher the probability of teleportation to a random vertex. Hence, ratings by questioners estimated as making more selections based on a personal distribution will be less influential than selections made by questioners estimated as making few or no selections based on a personal distribution.

C. Personalized Recommendation

For a user(questioner) $q$, given the estimated objectivity score ($\omega_q$), the overall expertise vector ($\alpha$), and the social affinity distribution $\sigma_q$, the probability of selecting a user $u$ as a satisfactory responder is given by:

$$P(u|q) = \omega_q P(u|\alpha) + (1 - \omega_q) P(u|\sigma_q) \quad (16)$$

In the case of a new user $q'$ for whom we do not have an objectivity estimate in a topic, we set $\omega_{q'}$ to the marginalized objectivity value, calculated as:

$$\omega_{q'} = \sum_{i=1}^{N_u} \frac{N_u^{q_i}}{N} \omega_u \quad (17)$$

VI. Experimental Results

We extracted data from the Yahoo! Answers website for three categories of Q&A documents, spread over two months, between January and March 2010. The three categories we Astronomy and Space (A&S), Books and Authors (B&A), and Wrestling(Wr). The data consisted of between 6000-10,000 documents in each category. The categories were chosen to take into account the range of interactions that occur on Yahoo! Answers: with Space and Astronomy, Books and Authors, and Wrestling expected to place decreasing importance on authority/expertise, and increasing importance on social affinity.

The text data for the Q&A documents were processed as follows: the hypertext information was removed, and stopping and stemming was performed. The user interaction information was processed so that only users who has asked or answered at least 30 questions in total were considered. This corresponds roughly to an average of one interaction every two days. Based on this criteria, 295 users remained in the A&S category, 349 in B&A, and 403 in Wrestling.

As an initial experiment, the expertise estimation algorithm was run over all user interactions for the categories of Astronomy and Space and Books and Authors. As the social affinity distributions $\sigma$ for each user, in the absence of information, we use a uniform multinomial distribution across all responders. Based on this, we constructed a cumulative probability graph of the expertise distribution $\alpha$. The responders were sorted by decreasing order of the expertise/authority assigned to them by the algorithm. As can be seen from the Graph in Figure 2, the top responders in the A&S category have a much larger share of overall authority ascribed to them. In comparison, the B&A graph is much smoother. This is because the algorithm identifies a small subset of responders as highly authoritative in the A&S dataset, while expertise in B&A is much more diffuse. This suggests that the A&S category is more information-oriented, as there are a few responders that everyone seems to agree upon as authoritative. In comparison, there is no consensus on expertise in the B&A category, which suggests it is more conversation oriented.

Following this, we ran an experiment to identify if our approach is capable of improving expert recommendation
in the task of social search. Documents from each category were divided into two equal parts as the training dataset and the test dataset. Following this, we constructed a term frequency inverse document frequency (TFIDF) based vector [15] for each Q&A interaction document in the training set, and clustered these documents based on a k-means-cosine algorithm [15]. Each of these clusters can be treated as a sub-topic within the larger topic. Following this, we ran our expertise estimation algorithm over each sub-topic cluster, to assign expertise scores to all responders in the sub-topic, and objectivity scores to all questioners. For the social affinity distributions \( \sigma \), we used the user’s selection distribution across the other clusters. Then, given a question in the test dataset, it is assigned the sub-topic with whose centroid it has the highest cosine similarity. The authority distribution \( \alpha \) of this sub-topic/cluster can then be used to make predictions for this question. These estimates were then used to make predictions based on equations (16) and (17). Table 1 shows the results of prediction across the three categories. The hybrid algorithm refers to the approach which combines both expertise and social affinity based models based on questioner objectivity. The other two columns show results for only when one approach is used. While the recommendation accuracy is low overall as the problem is a difficult one, due to the volatility of the Yahoo! Answers dataset, in all cases the hybrid approach does better than the other two approaches.

VII. CONCLUSION

As the Web moves from a platform for information search to a forum for social interaction, the assumption that users, when searching for information, are only searching for authoritative sources, becomes less true. However, the degree to which a user requires personalization is far from straightforward. The common naive approach till date has been to personalize all queries in the same way. However, different users have different expectations based on user personality and the type of questions.

In this paper, we formalize this intuition by defining the concepts of authority and objectivity for users interacting in the context of question and answer forums. The definition of authority as defined here can be seen as a generalization of authority as defined by the pagerank algorithm [5]. We also present an algorithm for calculating the trade-off required between authority and social affinity for such forums. The algorithm was used to make recommendations for users based on question topics. The results suggest that different users want different levels of personalization in the answers to their questions, and it is important to take this into account in order to make successful recommendations. Our experiments and results are focused on data drawn from the Yahoo! Answers website. However, the methodology and algorithms we develop are general and can be applied to any social network of producers and consumers, where users search for content generated by their peers. Examples of this include, besides Q&A sites, photo sharing sites such as Flickr, book and movie review sites (where users search for quality reviews), and video sharing sites such as youtube7.

REFERENCES


7http://www.youtube.com
