Predicting Potential Responders in Twitter: A Query Routing Algorithm

Cleyton Caetano de Souza[1], Jonathas José de Magalhães[1], Evandro Barros de Costa[2], and Joseana Macêdo Fechine[1]

[1] Department of Systems and Computing
Federal University of Campina Grande
Campina Grande-PB, Brazil
[2] Computing Institute
Federal University of Alagoas
Maceió-AL, Brazil

Useful Points:
-Twitter API
-WPM

Abstract. A phenomenon not so recent is the substantial increase in popularity and use of online social networks. With that has emerged a new way to find information online: the social query, which consists of posting a question in a social network and wait for responses from close friends. Usually, a question is posted to be visible to everyone, but we believe that this is not the best way: there will be the possibility of receiving several responses (including wrong), keep receiving answers where there is no need, do not receive answers, etc. The query router problem consists of finding the most able individual in the personal social network of the questioner. This work presents an algorithm to Routing Questions in Twitter. The model was validated through its predict capacity and the results shows that its recommendations match in half cases only when combined with a technique to enrich the information present in the question.

Keywords: Social Query, Routing Algorithm, Social Network, Twitter.

1 Introduction

A phenomenon not so recent, but perceived with greater intensity in last two years, is the substantial increase in popularity and use of online social networks. The Facebook[1], for instance, is the most visited website actually, surpassing even Google[2] [23]. On these virtual environment, people with common characteristics meet each other and discuss topics of mutual interest. The social networks was designed initially only to allow interaction between people. They have evolved and today are used for different purposes, e.g., for the distribution of games or the announcement of products between users.

Another interest phenomenon is a new way to find information online which born in Community Question & Answer Sites (Q&A Sites) and was extended to

[1] http://facebook.com
[2] http://google.com.br

B. Murgante et al. (Eds.): ICCSA 2012, Part III, LNCS 7335, pp. 714–729, 2012.

usual social networks: the social query, which consists of posting a question in a social network and wait for responses from close friends [15].

In social networks like Twitter[3], for example, the social query, appears as a valid alternative, in some cases, to find answers online [19].Usually, a question is posted to be visible to everyone. However, we believe that this is not the best way. After posting a question that will be visible to everyone there are some possible scenarios, e.g., receiving several responses, including wrong or continue to receive responses when no longer needed. Moreover, there is the possibility of potential responders may never see the question, thus they will never answer it.

One solution to this problem is the routing of questions, which consists of identifying what is the user connected to the questioner is more able to provide the correct answer and direct the question just for him (query router) [3] Such recommendation is characterized by the intimacy implicit between the questioner and the holder of knowledge; because they are direct connect in the social network [19].

However, decide to whom to direct the question is not a trivial task. When you choose the wrong individual can spend a long time to get a response, you can get one wrong answer or the chosen one can simply ignore the question. Thus, the technical problem of this research consists in identify the individual from the personal social network of the questioner who is better able to respond correctly and timely the question and direct it only for him/her. In this respect, this work proposes an algorithm for Routing Questions, which can be understood as a technique of recommendation that identifies the better able individual using various criteria (knowledge, social aspects, availability, etc.). We use the Weight Product Model (WPM), a strategy for making decisions with multiple criteria, to qualify the aptitude of all candidates [14].

The Routing algorithm proposed was designed to work in the Twitter, but it can be easily adapted to any context of other social networks. Previous work applies Routing algorithm in small social network developed only for Q&A and to offer support to the algorithms. The differential of our work is (1) we start proposing a model in a pre-existent and very popular social network and (2) we lead with the problem as a problem of decision making with multiple criteria, instead as a probabilistic problem like some most part of previous work.

Another pertinent question is how evaluating a Query Routing algorithm? There is no an evaluation technique common to the entire academic community and the majority of researches that deals with personalized recommendation present a qualitative evaluation, which difficult the comparison among the different algorithms [9]. However, assuming that the initial goal of a recommendation technique is to predict something that will interest someone [12], we decide to evaluate the proposal on the perspective of their predictive ability. Does the proposed Routing algorithm make recommendations that will reflect the events of the real world? Does the proposed Routing algorithm can predict who will answer questions posted openly on Twitter? With this purpose was validated the following hypotheses:

[3] `http://twitter.com`

- $H_{0,1}$: The proposed Routing algorithm cannot predict the events of the real world at least 50% of trials;
- $H_{a,1}$: The proposed Routing algorithm can predict the events of the real world at least 50% of trials.

To evaluate the validity of these hypotheses, we designed an experiment where some posted questions on Twitter were monitored and recorded who answered them. Then, the necessary information was passed to the algorithm and a list of recommendations with the most able was requested. The list of recommendation was compared with the users who really answer the question. Next, we calculate the hit rate (recall) obtained, aiming to study the validity of the cited research hypotheses. Importantly, the objective of the proposed algorithm is not predict who will respond, but indicate who respondents are more likely. We have assumed that who answer the question automatically must be considered one of the fittest and therefore possibly the Routing algorithm should indicate him/her as a recommendation.

In Twitter, the users can post message with until 140 characters, i.e., the questions that users have posted are subject to this limitation. Manipulate short questions is a hard task, because there is a need for terms that characterize well the topic of the question [6]. For this reason, the same experiment was performed using a synonymy expansion in question before it be passed to the algorithm. Thus, it is expected to obtain a recall rate equal to or greater than the recall rate of the first method, since the expansion terms have the ability to prioritize the most relevant results, according with Ramalho and Robin [17].

To evaluate the effectiveness of the proposed Routing algorithm combined with the synonymy expansion, we consider the following hypotheses:

- $H_{0,2}$: The proposed Routing algorithm combined with the synonymy expansion in question cannot predict the events of the real world at least 50% of trials;
- $H_{a,2}$: The proposed Routing algorithm combined with the synonymy expansion in question can predict the events of the real world at least 50% of trials.

Finally, to evaluate if the combination of Routing algorithm proposed with synonymy expansion produces a recall rate higher than the simple Routing algorithm, we consider the following hypotheses:

- $H_{0,3}$: The combination of the Routing algorithm with the synonymy expansion do not produces a recall rate higher than the same technique without expansion;
- $H_{a,3}$: The combination of the Routing algorithm with the synonymy expansion produces a recall rate higher than the same technique without expansion.

The study was conducted with nine persons who publish twenty nine questions and involved the processing of a graph composed for 1201 users, 131.962 messages

and 2.047.305 links between users. The analysis over the recall rate indicated that the Routing algorithm combined with the synonymy expansion reached the level expected (50%), but the simple technique did not reach, i.e., the hypotheses $H_{0,1}$ and $H_{a,2}$ and was accepted. Furthermore, the recall rate of both techniques were compared and the obtained conclusion is that the technique with synonymy expansion present results statically better than the simple technique (without expansion), confirming the hypothesis $H_{a,3}$. In our opinion, these results make clear the need for methods that improve the analysis of the content of the question.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work concerning previous studies that deal with social networks focused on finding resources on the Web. Section 3 presents the social network Twitter and discusses about the reasons that led us to choose it as context of this research. Section 4 describes the model, the characteristics that should be considered when we discuss about the capability of a candidate and how we measure these. Section 5 presents the Weight Product Model (WPM), as a model adopted in our approach for finding the best candidates. Section 6 describes the experiment and results used for validating our proposal; and, finally, Section 7 offers conclusions and discusses some future work.

2 Related Work

We decided to split this section into four parts. The first Subsection (2.1) presents researches that deal with the act of publishing questions on the web. The second Subsection (2.2) presents researches about Expertise Finding using multiple criteria. The third Subsection (2.3) presents studies that specifically address the query router problem either in conventional social networks (like Twitter or Facebook) or CQA Sites. Finally, the Subsection (2.4) presents the main differences between the previous work and our proposal.

2.1 About the Act of Publish Question in the Web

The fact is that search engines are not always the best way to find information on Web. To some needs of information are better solved by people, for instance, personal questions, recommendations, opinions, advices and high contextualized questions [6]. An alternative to this kind of questions are the CQA Sites as AnswerBag [4] and the Yahoo! Answers [5], which consists in virtual communities where users post and answer questions voluntarily. After posting a question, the user waits for answers of others users, who usually are unknown to him. But, people prefer ask questions to their close friends in social networks than to unknown persons in CQA Sites [18].

Regarding to the explicit action of publishing a question on social networks, Morris, Teevan and Panovich [15] present important statistics that confirm this

[4] `http://www.answerbag.com`
[5] `http://answers.yahoo.com`

as a viable strategy to get answers online. In their study case, 93.5% of users received answer to their questions after post them and these responses, in 90.1% of cases, were provided within one day. Paul, Hong and Chi [16] conduct a similar study using only Twitter, they conclude that, in this specific context, only a few part of questions posted receive answers (18.7%) and that the fact of receive or do not is intrinsically connected to the amount of followers of the questioner. However, questions posted in Twitter normally be answered quickly, in their study 67% of the responses come within the range of 30 minutes and 95% within the range of ten hours. We believe these findings are result mainly of the features of Twitter: when a user post questions to all followers, only a portion of his/her followers will view and a smaller portion will respond (as will be detailed in Section 3). Thus, users with more followers are more likely to get answers, because there is a larger viewing of their messages. And, with respect to agility in respond receiving, this is mainly due to the nature of Twitter as a real-time social network.

However, we believe that these results could be improved applying the routing questions: identifying an expert on the subject of the question, the answer could come faster and with higher quality. Horowitz and Kamvar [6] established a correlation between social query and the village paradigm: when an individual in a village looking for information, before consult the libraries, he turns first to the most intelligent people he knows.

2.2 About Expertise Finding

In fact, the problem that this paper aims to address can be understood as an Expertise Finding Problem. However, usually the Expertise Finding involves a context much large of candidates. The work proposed in this paper deals with the detection of specialists in much smaller subset of the entire social network (the set of friends/followers of the questioner) and thus the conditions under which a friend is marked as a specialist in each case differ because of the context that is analyzed ,i.e., the Expertise Finding addressed here is personalized. Moreover, usually Expertise Finding involves only the discovery of who owns knowledge about a given topic, while (as will be presented in Section 4) the algorithm proposed here involves multiple criteria.

Sarda et al. [18] deals with the identification of experts in Orkut[6] using two criteria: knowledge and confidence. However, this work has focused on the mining of expert opinion about products and not necessarily on the resolution of questions. Smirnova and Balog [20] propose a Recommendation Technique to identify experts using two criteria: knowledge gain (calculate using Bayes Theorem) and contact time (the calculation will vary with the type of social network, but is distance, though some metric, between the questioner and the expert). However, we believe that the model for contact time (which would be the time needed to receive answers) proposed in this work is not consistent with reality, because it not measures the availability of the other user to provide the answer.

[6] `http://orkut.com`

2.3 About Routing Questions

The query router consists of an algorithm of recommendation (or a technique) that objectives find an expert present in a group and direct a question to him/her [21]. In [2], Banerjee and Basu presented a probabilistic and decentralized model for the routing questions problem. This means that there is not an entity that makes all decisions and the Routing algorithm works based on the probability of actions taken in past be repeated. Other work is the [5] that presents a centralized model: the iLink is a global entity that decides who will receive the questions and, in some cases, is also able to offer answers.

Some examples of systems that implement a model of query router are the Aardvark [6], a social network that belongs to Google, Q-Sabe [1], an academic tool for exchange of information focused on education. Both systems consist of CQA Sites where users publish questions (questioners) that are directed to other users (respondents) and these can choose to answer or ignore the question, and AskWho [13], a plugin to Facebook which aims present friends of the questioner as possible answers to a question.

2.4 About the Differential of Our Research

The researches of Andrade et al. [1], Davitz et al. [5] and Horowitz and Kamvar [6] propose query router techniques and developed environments where they works. Our research follows the reverse path. We propose a Query Router algorithm that works in a pre-existent and popular social network: the Twitter, one of most popular social networks currently and which, apparently, will benefit of our technique. In [13] is presented a plugin to Facebook, but AskWho do not use any special technique to match friends, consisting only in a search engine which compare the content of the question with the friends. The differential our research in relation to the works cited is that we take as our starting point a specific and popular social network and we build our model to fit in the context of this network. Moreover, our model inherits characteristics of other related work, like trust model and friendship. We presented these concepts of a way that they can be easily adaptable in other contexts. We believe that the query router problem can be treated as a multi-criteria decision making problem, instead a probabilistic problem considered by previous work. For this reason, another differential of our research is the solution of the model using WPM, a strategy for making decisions with multiple criteria, considered adequate for the amount of variables involved [22] and that also allows a dynamic evolution of our technique trough the addition of new criteria.

In [21], we present our Formal Model do the Query Router Problem to the context of Twitter. The work that will be presented below shows our model in an algorithm form, easily compressive and which exposes new features and adaptations of the technique.

3 Twitter

Twitter is a kind of microblog, a variant of the blogs that have some type of limitation of the content, where users can tweet (post a message) on any topic using 140 characters [4]. In less than three years, Twitter gained such popularity that became the microblog with the largest number of users [4] and awakening the interest of the scientific community about it [8,11]. In January 2010 the microblog counted more than 73 million users and in March, month in which it completed four years, reached the mark of 10 billion posted messages [4].

Via Twitter users can follow other users and can be followed by other users. In the context of Twitter, to follow a user means exposing publicly interest in the content posted by him. Among the reasons that lead a user to follow another one are admiration, friendship and reciprocity, for example. In addition, the user may want to follow other user by considering that content posted by him is relevant. The tweets (posts) may or may not be publics and any user is allowed to refer to others within a tweet. Because of these features many users use the microblog as a public chat [7].

When a user publish a public question in the Twitter, if it will not be answered quickly the chances of be visualized and answered in future are lowest because the question will down in timeline of followers of him. After publish a question, probably, not all users who follow the questioner will see the question. Among the users that visualize, only a few will provide an answer and there is no guarantee that any of these answers will satisfy the information needs of the questioner. Some users will not provide an answer because, as the question was posted for all followers, they do not feel an obligation to help. And as time passes, the chances of that question be viewed and consequently answered in the future are lower, because it will fall positions on the timeline of the followers of the questioner.

We believe that when a tweet (question) is directed previously to someone the probability of it be visualized are much larger, because the user mentioned can filtrate the messages which mention him/her. When a user mention other user, the user mentioned, immediately, receive an email inform about the message. There is the possibility of the mentioned user disable these notifications, but any user can filter their mentions, as already commented. Given these facts, we believe that direct the question to someone, practically, guarantees that the message will be visualized, but there is no guarantee that the message will be answered, either on the quality of the response.

Appears evident that to direct a question increases the probability of it be visualized, while the probability of it be good answered depends to who it will be directed. A Formal Model to Query Router in Social Networks consists in a recommendation algorithm (technique) that analyses the information available on social network to infer who the user most able to respond the question. In Figure 1 is illustrated the query router process working.

The question is formulated without a mention. The Query Router algorithm (or Routing algorithm) analyzes the information about the followers of the questioners and ranks them according to aptitude to answer the question. The algorithm adds a mention in the question. In Figure 1 the question is being directed

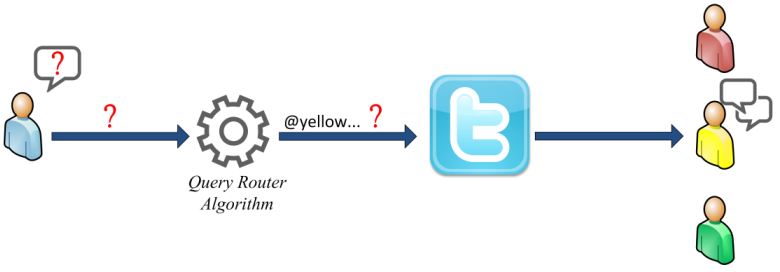

Fig. 1. Using the Query Router algorithm to Tweet a Question

to only one follower (the yellow user), but, as the algorithm ranks all user, will be possible send it to the n followers in bests positions.

In the next section, will be presented the Formal Model to Query Router, which is the proposal of this work.

4 Formal Model

Informally, the problem that the model proposes to solve is given a question posted by a user (questioner), find among his followers that user with the characteristics:

1. Knows the answer: if we direct the question to a follower who has no knowledge on the subject of some question, the quality of the response will be low and probably will not satisfy the information needs of the questioner;
2. Has the trust of the questioner: if we direct the question to a follower who has no confidence of the questioner, regardless of the answer, probably, the questioner could not believe in the responder and prefer continue to receive answers from other followers;
3. Provides the answer quickly: if we direct the question to a follower who does not access the social network often, a long time can spend until the questioner get your answer;

Thus, we need find a way to decide who user has the best combination of these three characteristics. Next, we will detail our Formal Model to Routing Questions in the Context of the Social Network Twitter.

4.1 The Social Network

The Twitter is a social network defined by the tuple $T = (U, R)$, where $U = \{u_1, u_2, u_3, ..., u_{|U|}\}$ is a set of social network users and R corresponds to the set of relationships as $r_{i,j}$ between two users i and j, where $i \in U$ and $j \in U$. In the specific context of Twitter the relationships are not necessarily reciprocal, so $r_{i,j} \neq r_{j,i}$. The existence of relationship $r_{i,j}$ means that the user i is a follower of the user j.

An user u has these sets as attributes: $Followers_u$, $Following_u$ and M_u. The set $Followers_u \subset U-u$ contains all user x with whom u maintains a relationship of form $r_{x,u}$. The set $Following_u \subset U-u$ contains all user x with whom u maintains a relationship of form $r_{u,x}$. The set $M_u = \{m_{u,1}, m_{u,2}, m_{u,3}, ..., m_{u,|M_u|}$ consists of all messages (tweets) posted by the user u. Each message m_u posted by u has the following attributes: d_{mu} (corresponds the date that the message m_u was posted by the user u) and s_{m_u} (corresponds the string that represents the message content).

4.2 Problem Definition

The problem is to find a user $f_u \in Followers_u$ that has a higher probability of answering a question q_u that requires the knowledge $k_{f_u,q}$ and was published by user u in form of a message m_u. The calculation of this probability, called p_{q_u,f_u}, takes into account three abstract concepts:

1. $k_{f_u,q}$: knowledge of f_u in relation to the question q_u;
2. t_{u,f_u}: trust of u on the user f_u;
3. a_{fu}: activity of f_u in social network.

So the problem can be summarized in find a user $f_u \in Followers_u$ whose tuple ($k_{f_u,q}$,t_{u,f_u},a_{fu}) maximizes the probability p_{q_u,f_u}, i.e., the user that has the best tuple ($k_{f_u,q}$,t_{u,f_u},a_{fu}).

The details of how we believe that these concepts are quantified are described with in [21]. Bellow, we present our Model in an algorithm form, to make easier understand it.

5 The Routing Algorithm

Figure 2 illustrates the pseudo code of the Routing algorithm used to qualify candidates.

First, it is verified the need of make or not make the synonymy expansion (line 1). The variable *DEFAULT_AMOUNT* (line 2) limits the max amount of synonyms that should be added to each word.

After synonymy expansion part, we calculate the attributes of each follower of the questioner: knowledge (line 6), trust (line 7) and activity (line 8). We define knowledge as an expertise that some user has about a topic. In social networks the knowledge of a user u is direct related with the content posted by him: M_u [5,6]. Trust is a measure that quantifies the faith that a user has over the information posted by another user (credibility) [19] and in our model is calculated based on friendship and similarity. The activity level corresponds to the frequency with the user post new tweets [21]. Calculate the tuple ($k_{f_u,q}$,t_{u,f_u},a_{fu}) for any user $f_u \in Followers_u$ is a simple task, but compare the tuples of two different users and decide which user is the best qualified to answer the question is not always a trivial task.

```
1  if (expansion == TRUE)
2     q = synonyms (q,DEFAULT_AMOUNT)
3
4  for each f_u in Followers_u do
5     Begin
6        k_{f_u,q} = knowledge (f_u,q)
7        t_{u,f_u} = similarity (f_u,u) * friendship (f_u,q)
8        a_{f_u}   = activity (f_u)
9        f_u.victories = 0
10    End
11
12 /*Weight Product Model*/
13 for each f1_u in Followers_u do
14    for each f2_u in Followers_u do
15       Begin
16          if (f1_u == f2_u)
17             continue;
18
19          k = k_{f1_u,q} ÷ k_{f2_u,q}
20          t = t_{u,f1_u} ÷ t_{u,f2_u}
21          a = a_{f1} ÷ a_{f2}
22
23          comparison = (k^RELEVANCE_KNOWLEDGE_LEVEL) *
24                       (t^RELEVANCE_TRUST_LEVEL) *
25                       (a^RELEVANCE_ACTIVITY_LEVEL )
26          if (comparison~ = 1)
27             Begin
28                f1_u.victories ++
29                f2_u.victories ++
30             End
31          else if (comparison~ > 1)
32             f1_u.victories ++
33          else if (comparison < ~1)
34             f2_u.victories ++
35       End
36
37 sort_according_victories (Followers_u)
38
39 return Followers_u
```

Fig. 2. Routing Algorithm

This way, as we already have commented, we consider that as a problem of decision making with multiple criteria (or multi-criteria decision making) and trying to find the best tuple among users $f_u \in Followers_u$, after calculate the attributes of each follower, we used the Weight Product Model (WPM) as a method for making decisions because it is the most appropriate for the conditions and context presented (dependence up to three variables) [22] (start in line 13).

The values *RELEVANCE_KNOWLEDGE_LEVEL*(line 23), *RELEVANCE_TRUST_LEVEL* (line 24) and *RELEVANCE_ACTIVITY_LEVEL* (line 25) are called factors of importance, must be set according to user need and their sum must results in 1, i.e., $(A+B+C+... = 1)$ [14]. When the user wants to prioritize the speed of response (to get answers quickly) he must establish a high value for the factor *RELEVANCE_ACTIVITY_LEVEL*, in case the user wishes to prioritize the answers from friends (because it requires a very personal response) he must establish a high value for the factor *RELEVANCE_TRUST_LEVEL*

and, finally, when the user wants to prioritize the knowledge that his friends have about the domain of the question (to find good answers) it must establish a high value for the factor *RELEVANCE_KNOWLEDGE_LEVEL*.

Still about WPM we use a variable called *comparison* to compare users in pairs. If *comparison* > 1, then f_1 is superior to f_2 and we put 1 in position (f_1, f_2) of the matrix and 0 in position (f_2, f_1) (line 32). If *comparison* < 1, then f_2 is superior to f_1 and we put 1 in position (f_2, f_1) of the matrix and 0 in position (f_1, f_2) (line 34). If *comparison* $= 1$, then f_1 is equivalent to f_2 and we put 1 in position (f_1, f_2) of the matrix and 1 in position (f_2, f_1) (lines 28 and 29). But in our approach we use a little confidence interval to decide. For this reason, we use in pseudo code the symbols to express: $\sim=$ (near), $\sim>$ (a little larger) and $\sim<$ (a little smaller).

We summarize the amount of victories of each user (start in line 26) and after compare all user we order them according with the number of victories (line 37). At the end, this ordered list also represents the relevance order of each user and it is returned as a solution of the algorithm (line 39).

We decided to address the problem as a multi-criteria decision making because it makes the algorithm easily expandable. The addition of new criteria, for example reciprocity [10] or the latency time (time between the last message and the current instant), requires only: (1) the addition of its calculation to each follower, (2) its usage in the ratios that calculates the value of the *comparison* (start in line 23) and (3) the calibration of the factors of relevance of each criteria.

6 Evaluation and Results

This section describes the details of evaluation process and then discusses how validation of the Routing algorithm proposed in this paper was performed. Additionally, the obtained results were reported.

6.1 Methodological Aspects

To validate the Routing algorithm proposed was draw an experiment whose objective was to ascertain its ability to reflect, trough recommendations, what happened in real world. For the research, nine people posted on Twitter 29 questions which were answered by 44 users. These questions are designed by the participants themselves, were visible to all followers and anyone that wanted to could answer it. Each question was considered a trial, being the input variables: the question (q), the user information (u) and the list of users who responded on Twitter (*real responders*). For each question was requested to Routing algorithm a list of recommendation (*recommended responders*). The list of recommended users was then compared with the list of users who answered the question in the real world and we analyze the recall rate obtained. As already commented, in order to analyze the benefits that the synonymy expansion would bring in the Routing algorithm, the same experiment was performed only passing in different an extended version of the question.

The study was conducted with nine persons who publish 29 questions and involved the processing of a graph composed for 1201 users, 131.962 messages and 2.047.305 links between users. All information used in research may be available by contacting any of the authors. To conduct the study, the proposed Routing algorithm was implemented in Java, for extracting data from the social network we used Twitter Streaming API[7], for application of Natural Language Processing (NLP) techniques was used BrazilianAnalyser[8], that belongs to the Lucene API, the thesaurus used for the synonym expansion was from the website "Thesaurus da Língua de Portuguesa do Brasil"[9].

6.2 Results and Discussion

In Figure 3 is showed the amount of true positive obtained by the Routing algorithm proposed in that the size of list of recommendations grows for the two versions compared (with and without synonym expansion). The horizontal axis represents the size of the list of recommendations and the vertical axis the total number of true positive within the 44 possible.

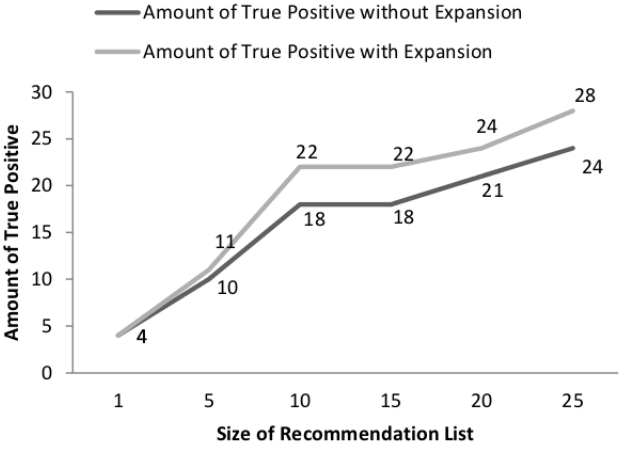

Fig. 3. Amount of True Positive of the Routing Algorithm.

Looking at Figure 3, apparently the combination with the expansion technique achieved better performance than a simple Routing algorithm (without synonymy expansion). Moreover, in both situations, even with a list of twenty five recommendations for each question, we do not match all 44 users. Even so,

[7] https://dev.twitter.com/docs/streaming-api
[8] http://lucene.apache.org/core/old_versioned_docs/versions/3_0_0/api/contrib-analyzers/org/apache/lucene/analysis/br/BrazilianAnalyzer.html
[9] http://alcor.concordia.ca/~vjorge/Thesaurus/

the true positive score using 10 recommendations (P@10) was consided positive, being respectively 40% (18/44) and 50% (22/44) to a simple Routing algorithm and for the combination with the synonymy expansion. For this reason we choose this size to continue the investigation.

Using ten recommendations (P@10), the recall rate of the Routing algorithm without synonymy expansion for each one of the twenty nine trial was, in order: 0%; 30%; 100%; 100%; 0%; 0%; 0%; 100%; 100%; 50%; 0%; 100%; 0%; 0%; 100%; 0%; 0%; 50%; 0%; 100%; 0%; 100%; 50%; 0%; 0%; 100%; 50%; 100%; 100%. We can see many of the recall values are most extreme; this happen because most of the questions posted (59%) was answered by only one person. In this situation, is only possible to obtain a recall 0 (0%) or 1 (100%). To evaluate if the Routing algorithm hit at least 50% of cases, each trial where happen at least one true positive ($recall > 0$) was considered as one and the cases where no match ($recall = 0$) will be considered zero. Thus, the percentage of success obtained was 55% (16/29). To analyze whether this value is statistically significant was performed a one-tailed binomial test where he obtained a $p\text{-}value$ of 0.3555 for $\alpha = 0.05$. This means that there is no significance to accept $H_{a,1}$. Thus, $H_{0,1}$ was the first hypothesis accepted for this work, i.e., the Routing algorithm proposed, statistically, did not get successes by more than half of the trials.

Using ten recommendations (P@10), the recall rate of the Routing algorithm with synonymy expansion for each one of the twenty nine was, in order: 50%; 30%; 100%; 100%; 100%; 100%; 100%; 100%; 100%; 50%; 0%; 100%; 0%; 0%; 100%; 100%; 0%; 50%; 0%; 100%; 0%; 100%; 50%; 0%; 0%; 100%; 0%; 100%; 100%. Thus, using the same adaptation before, the percentage of success obtained by the Routing algorithm combined with the synonymy expansion was 69% (20/29). To analyze whether this value is statistically significant was performed a one-tailed binomial test where was obtained a $p\text{-}value$ of 0.03071 for $\alpha = 0.05$. This means that the hypothesis $H_{a,2}$ was accepted, i.e., the Routing algorithm combined with the synonymy expansion obtained successes in more than half of the trials.

Finally, the successes rates of each trial were compared to analyze if the technique with synonymy expansion is superior. Initially, a study was conducted aiming to verify the normality of the recall rates of both distributions using the Shapiro-Wilk Test which resulted in a $p\text{-}value$ 6.582e-06 for combination technique and 4.73e-06 for the technique without the expansion. This means that both distributions are not normal. Then, using the Wilcoxon Signed Rank Test, we check if the distribution of recall rates by the Routing algorithm combined with synonymy expansion is superior to technique without expansion. The result was a $p\text{-}value$ of 0.03299 to a α=0.05, what means that the hypothesis $H_{a,3}$ also was accepted, i.e., the Routing algorithm proposed combined with a synonymy expansion obtained recall rates better than the same Routing algorithm without expansion.

The results showed that the combination of the Routing algorithm with the synonymy expansion got successes rates in more than half of the trials; however it is considered that such results are no so expressive. To examine whether the

recall rate for each trial (P@10) was superior to 50% was carried out again the Wilcoxon Signed Rank Test. This time, the *p-value* was 0.5981 for the recall distribution without synonymy expansion and 0.4007 for the Routing algorithm with synonymy expansion. These *p-values* indicate that the average value of successes per trial is not 50%. This negative result, in particular, is due to the high variance in success rates (many extreme values) and the statistical test did not guarantee a reliable range for the average of successes rate per trial. This raises the following question: *why have so many successes rates equal to 0%*? During the study, it was noted that the proposed task was naturally difficult. When posting a question that is visible to all, the questioner is held hostage of various random factors such as, for example, anyone who see the question might want (or not) answer it, if the question is not viewed by anyone quickly chances of someone to answer it in the future are smaller, people we never expect an answer can answer, or, e.g., if the perfect individual according to the Routing algorithm never see the question because of other random reasons (e.g., tired, blackout, did not pay the Internet bill and the service was suspended) he/she never answer it. However, it the study results was considered very positive, because being the main propose of this work to infer who are the best candidates to answer a particular question, the fact that the recommendation match with what happens in the real world consists of a predictive validity of the conceptual model, but little refers to the quality of the recommendation. Thus, we credit the negative trials (where the successes rate was 0%) mainly to the random factors mentioned above.

7 Conclusion

This paper presented a proposal for a Query Routing algorithm to Twitter. The purpose of this Routing algorithm is to find, among the followers of a given user (questioner), the individual most able to provide the answer. The differential of this research compared to the work here described is due to the fact that we take as our starting point a social network that does not fit into the category of Q&A Site, but where the users usually publish questions. Furthermore, the Routing algorithm was developed with facilities to be easily adaptable to the context of other social networks and to be also easily expansible. Finally we presented the model solution by using the WPM. To validate the proposed Routing algorithm a case study was performed where it was evaluated its ability to predict who would answer a question in the real world.

This study indicated that only when combined with a technique of expanding the synonymic, the Routing algorithm gets hit by more than half of the trials. This combination, when compared with the same technique without expansion, was more efficient even with respect to the rates of correct answers for each test. These results demonstrate that the proposed routing policy can still be improved.

An immediate future work includes a qualitative evaluation of the recommendations by the own questioner, besides we want increase our predict rate to 75%

of trials. Another interesting further research is a study on which factor is most important on the recommendation of experts: knowledge ($k_{f_u,q}$), trust (t_{u,f_u}) or activity (a_{fu}); and if its importance depends on the type/topic [15] of question or another particular factor. We also proceed an investigation to establish whether the direction of questions to a user (or a small number of users) is more effective than post the question to all followers. Furthermore, aiming to improve the results obtained by Routing algorithm we can apply semantic web techniques or analyze the insertion of Bayes Theorem to calculate the probability based on the same criteria or (when the algorithm is available to Twitter community) to implement the idea presented in [20] and [13], which shows to questioner a list of recommendations and him/her takes autonomy to decide to who will direct the question.

References

1. Andrade, J.C., Nardi, J.C., Pessoa, J.M., de Menezes, C.S.: Qsabe-um ambiente inteligente para endereçamento de perguntas em uma comunidade virtual de esclarecimento. In: LA-WEB 2003 (2003)
2. Banerjee, A., Basu, S.: A social query model for decentralized search. In: Proceedings of the 2nd Workshop on Social Network Mining and Analysiss, vol. 124. ACM, New York (2000)
3. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Parreira, J.X., Weikum, G.: Peer-to-peer information search: Semantic, social, or spiritual. IEEE Data Eng. Bull. 30(2), 51–60 (2007)
4. da Silva, M.L.H.: O Twitter dentro do Universo da Cibercultura Uma Abordagem Teórica da Ferramenta. intercom.org.br, 1–15 (2010)
5. Davitz, J., Yu, J., Basu, S., Gutelius, D., Harris, A.: iLink: search and routing in social networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 931–940. ACM (2007)
6. Horowitz, D., Kamvar, S.D.: The anatomy of a large-scale social search engine. In: Proceedings of the 19th International Conference on World Wide Web, pp. 431–440. ACM (2010)
7. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1), 8 (2009)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
9. Konstan, J.A.: Introduction to recommender systems: Algorithms and evaluation. ACM Trans. Inf. Syst. 22, 1–4 (2004)
10. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, WOSP 2008, p. 19 (2008)
11. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors. In: Most, pp. 591–600 (2010)
12. Lima, W.T., Branco, C.F.C., Barbosa, P.: Sistemas de recomendação de notícias nas mídias sociais buscam substituir o gatekeeping dos meios de comunicação de massa. Comunicação & Inovação, 36–45 (2009)
13. Liu, C.: AskWho: Finding Potential Answerers for Status Message Questions in Social Networks. agora.cs.illinois.edu, 1–5 (2010)

14. Miller, D.W., Starr, M.K.: Executive Decisions and Operations Research. Prentice-Hall, NJ (1969)
15. Morris, M.R., Teevan, J., Panovich, K.: What do people ask their social networks, and why? A Survey Study of Status Message Q&A Behavior. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 1739–1748. ACM Press, New York (2010)
16. Paul, S.A., Hong, L., Chi, E.H.: Is twitter a good place for asking questions? a characterization study. In: ICWSM (2011)
17. Ramalho, F., Robin, J.: Avaliação empírica da expansão de consultas baseada em um thesaurus: aplicação em um engenho de busca na web. RITA 10(2), 9–28 (2004)
18. Sarda, K., Gupta, P., Mukherjee, D., Padhy, S., Saran, H.: A Distributed Trust-based Recommendation System on Social Networks. In: 2nd IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb 2008). IEEE (December 2008)
19. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Weikum, G.: Efficient top-k querying over social-tagging networks. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, p. 523 (2008)
20. Smirnova, E., Balog, K.: A User-Oriented Model for Expert Finding. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 580–592. Springer, Heidelberg (2011)
21. De Souza, C.C., de Magalhães, J.J., De Barros Costa, E.: A Formal Model To The Routing Questions Problem In The Ccontext Of Twitter. In: IADIS Internaitional Conference WWW/Internet, ICWI 2011 (2011)
22. Triantaphyllou, E., Mann, S.H.: An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. Decision Support Systems 5(3), 303–312 (1989)
23. Ylan, Q.: Mui and Peter Whoriskey. Facebook passes Google as most popular site on the Internet, two measures show. The Washington Post (2010)