Towards Discovering Informal Experts in Stack Overflow: A Data Mining Approach based on Winner Posts

Carlos Gómez (V00776905)[1] and Lorena Castañeda (V00776903)[2]

Department of Computer Science, University of Victoria
[1]CHiSEL Research Group, cagomezt@uvic.ca
[2]Rigi Research Group, lcastane@cs.uvic.ca

Abstract

With the massive changes in technology, the amount of software development tools and techniques had grown to respond to such constant demand. Hence, software developers face the difficulties of finding reliable information in web communities, specially over newer technologies. In fact, key users in a community represent reliable source of information but not always are explicit. Stack Overflow is a community around the theme of software development and due its massive use, stores huge amount of information. No doubt, over time it has become a reliable source for software developers to find answers over technology problems. In addition, through the qualification of winning post, the users can identify explicitly the proper answer to a question in a pool of possible answers. We argue a winner posts is strongly related with a key user, but not in a decisive way. Indeed, through a set of data mining techniques we present other variables like life time and time of response, that would help to identify key users in the community.

Contents

1 Introduction

Currently, software developers face different challenges in the scope of their work. In their daily operations is very common to search over the internet for answers to their software and development problems. However the abundance of information available make very difficult to developers to find solutions quickly, and more importantly filtering those considered a reliable source.

Social communities had become important sources of information in different fields, bringing together users with common interest around the same topic and sharing information. Different tools for social communities are forums, blogs, and question and answer websites. Hence, some had gain reputation among their users even because of their popularity of because of the reliability of the information.

In the scope of software development, Stack Overflow [1] stands out as a well known social community with high activity. Users constantly search over this website and contribute to solve problems posted by other users. In fact, the community had established a mechanism of voting to evaluate answers according their pertinence with the question providing an easy mechanism for new users to find the best answer to an already posted question. In other words, Stack Overflow is considered a reliable source of information for the software development community.

However, the community information is generated by any type of users, from developers, newcomers in technology, to professionals and experts in different fields. However, for the Stack Overflow database is very hard to truly differentiate one from another given the massive users in the community. Given that we can't tell them apart, we prefer the tern *informal expert* to qualify an user that shows well-known knowledge in a certain area, identifies key facts before everyone else and also and writes about it before anybody else. To be clear a formal expert, is well known professional, but as in the stack overflow we can't be sure of their professional credential, we can't be sure of them. Also, informal experts set is a bigger set that contains those in this group as well given that the expertise is gained by contributions and not by title.

Moreover, these informal experts represent a valuable asset in the construction of a community. They are not only experts in their field and can help newer users to solve their problems, but from the point of view of software development communities it is important to create a community of experts in order to exchange information, organize events, or even work in join software projects. From the business perspective, key users are important while testing new software, also to get ideas of improvement, and even hire them to solve specific problems.

Despite that the popularity gained trough votes is a proper mechanism to choose the best answer, we argue that identifying experts in these communities is also useful while looking for reliable sources. In social communities such as Stack Overflow, where everybody is allowed to participate, there is no mechanism to identify this type of users, thus we consider a challenge to identify those that belong to our informal experts user's set.

In Stack Overflow a *winning post* is that answer-post that has being selected by the question-post's user to be the one that truly solved the problem she or he posted. Given the importance of informal experts, and the lack of mechanisms to identify them in the Stack Overflow community, the staring point of our study is the premise that wining posts are strongly related with informal experts given the reason to actually solve a problem. However our hypothesis is that there are other variables in the Stack Overflow community that narrow the domain of potential informal experts to get a more accurate set of them.

In light of this we define the following Stack Overflow key elements that possibly affect the decision over a user to be considered informal expert:

- Reputation: shows the number of contribution to that community and is gained according the type of activities the user performs over the community (i.e., given by votes, posts, visits and answers score).
- Acceptance: The value that other users had given to his or her answers (i.e., his or her answers are highly voted, and/or are winning posts)
- Activity: measured by the lifespan of the user vs the amount of contributions he or she have had during that time (i.e., her contributions in the community are frequent),
- Winning Post Count: how many posts in the same topic are winners.
- Badges: these are status gained by the users by different means (e.g, by activity or other users' decisions), and earning badges[1] in stack overflow are very related with the qualification of expertise. We will consider experts in the level of Gold badges (i.e., Famous Question, Marshal, Copy Editor, Reversal, and Great Answer)

To demonstrate our point, we study the data from Stack Overflow. Firstly we cover a phase of *data preprocessing* in Section 3 where prepare the information in the proper structure transforming the XML files offered by Stack Overflow using our own adaptation of the *SoSlow* tool [2], to create the data set in the proper format. Secondly, we move to

[1] The complete description option of badges can be found in the official website: http://stackoverflow.com/badges

the *data mining* phase depicted in Section 4 were we use *WEKA* [3] as a tool to execute the knowledge flow, to apply data mining techniques to propose a classifier to predict winning posts and potential informal experts. Thirdly, we enter to a phase of *observations and results* in Section 4.2 where we analyse the information resulted from the winning posts and the potential informer experts, to generate a more accurate set attributes that identify an informal expert in a given domain. Fourthly, in Section 5 we present the *evaluation* of our proposal to prove our hypothesis identifying a set of 5 informal experts in the domain of C-sharp. Finally, in Section 6 we conclude our study.

2 Related Work

The study of social media is being a challenge around the software community. However, we present this project as part of the International Working Conference on Mining Software Repositories (MSR 2013) that challenges researchers to uncover interesting findings on Stack Overflow Data. Hence, we present an overview of different approaches towards this Question and Answer platform to highlight relevant aspects to our own study.

For the past years, the web 2.0 also known for user-generated content has gained popularity among users increasing the number of blogs, forums, and social networks in general. However, due the enormous amount of content, the task of filtering and ranking the useful information in these platforms is being a challenge for different communities specially Question and Answer (Q&A) websites.

Different approaches have been made over these social communities, especially in Q&A websites like Stack Overflow. Agichtein *et al.* propose a general classification framework to identify high quality content in Yahoo! Answers (a Q&A website) [4]. In their approach they study the conditions of high quality such as: semantic features (i.e., punctuation, syntax, semantics, and grammar), user relationships with the information, and usage statistics. Additionally, they study the composition of Q&A website elements to determine the important aspects while assessing the quality of the content and proposed a graph-based model with the relationships of contributors, content and usage.

In the same scope, Anderson *et al.* consider the dynamic behaviour of the community to address important factors to identify trustworthiness and accuracy of the content [5]. In their approach they study Stack Overflow and highlight two principles required to find the best answer: (1) the expertise level of the contributors, and (2) the high activity level of the question as a representation of a potential interest and good reputation of those involved.

In a related work, Bacchelli *et al.* developed a plug-in for Eclipse IDE to integrate Stack Overflow data [6]. A key element in this approach is the understanding of the structure of the Q&A posts highlighting aspects such as voting and tagging as important elements to filter the searches.

Other approaches relevant to our study include Barua *et al.* methodology to analyse the textual content of Stack Overflow whose findings are the conception of trends among the software developer community [7]. Mamykina *et al.* present and study of the time of response the users take to answer a question in Stack Overflow, arguing that is not only expertise a factor to a quick answer but also a high activity of the community [8].

In final consideration the related work exposes issues relevant for our work. Firstly, Q&A websites are recognized as reliable sources of information, specially Stack Overflow in the domain of software development. Secondly, the quality of answers is a well common challenge and voting is a useful mechanism to address such issue. Finally, users in this social community are informal also considered experts on their field. The study of the content of their answers is important to address significant information related with tools and improvements.

3 Data Preprocessing

The preprocessing process aims to clean and transform the raw data to an useful form. In order to do that, we follow the next steps:

1. **Raw data and structure analysis**: To identify the original data set and studies the elements according the domain of the information
2. **Data transformation:** To compute and transform original values into types in the domain of the study (i.e., dates as minutes, hyperlinks as URL strings)
3. **Data preparation:** To compute and calculate raw data to transform it in information, (i.e., calculating time elapsed, counting elements, averages, normalization)
4. **Training and Testing data set generation:** To generate the files required for the data mining process (i.e, ARFF files for the WEKA tool)

The details for each step are described in the next sections.

3.1 Raw data and structure analysis

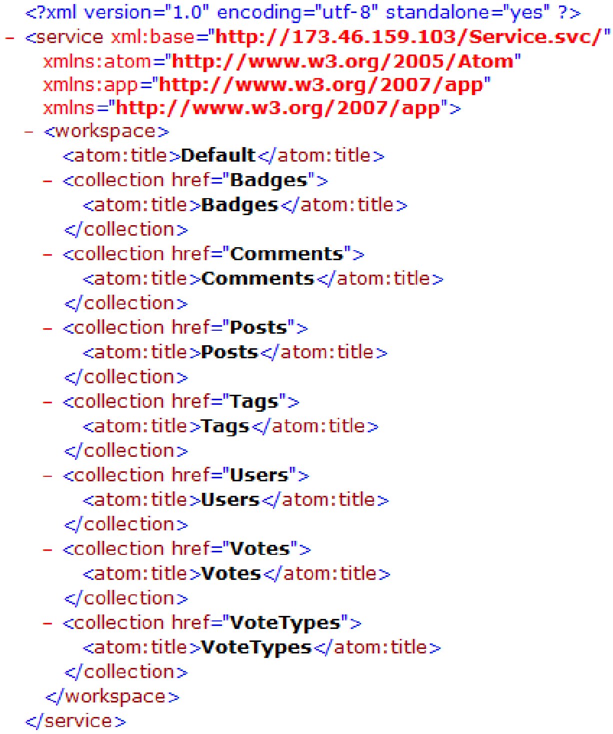
The data set used in this study corresponds to the Stack Overflow Challenge presented in the MSR conference [9] related to August 2012 provided by StackExchange[2]. This data set is presented as a set of XML files. Figure 1 shows an example of the structure of the file.

3.2 Data transformation

In order to manipulate the data in these XML files, we used the So Slow Tool [2] by Sam Saffron to import the XML data into a Microsoft SQLServer database. After identifying the elements required for our study we match them with the tables produced by the So-Slow tool, as depicted in Figure 2a : **Post, Tag, Votes, Vote Type, User, Comment** and **Badge**.

Hence, we used some modifications in the tool [10] to add two more tables as depicted in Figure 2b that are relevant to our study: **Resources, PostResources, PostTags** and **Tags**. This new tables will help to filter the data and get closer information to study our

[2] http://stackexchange.com/



```
  <?xml version="1.0" encoding="utf-8" standalone="yes" ?>
- <service xml:base="http://173.46.159.103/Service.svc/"
    xmlns:atom="http://www.w3.org/2005/Atom"
    xmlns:app="http://www.w3.org/2007/app"
    xmlns="http://www.w3.org/2007/app">
  - <workspace>
      <atom:title>Default</atom:title>
    - <collection href="Badges">
        <atom:title>Badges</atom:title>
      </collection>
    - <collection href="Comments">
        <atom:title>Comments</atom:title>
      </collection>
    - <collection href="Posts">
        <atom:title>Posts</atom:title>
      </collection>
    - <collection href="Tags">
        <atom:title>Tags</atom:title>
      </collection>
    - <collection href="Users">
        <atom:title>Users</atom:title>
      </collection>
    - <collection href="Votes">
        <atom:title>Votes</atom:title>
      </collection>
    - <collection href="VoteTypes">
        <atom:title>VoteTypes</atom:title>
      </collection>
    </workspace>
  </service>
```

Figure 1: General Structure of an XMl file in the Stack Overflow dump

expected informal experts and associate them to their fields. thus, the Tags are the fields of knowledge and the Resources the type of elements users share in the answers, such as, hyperlinks, code or plain text, among others.

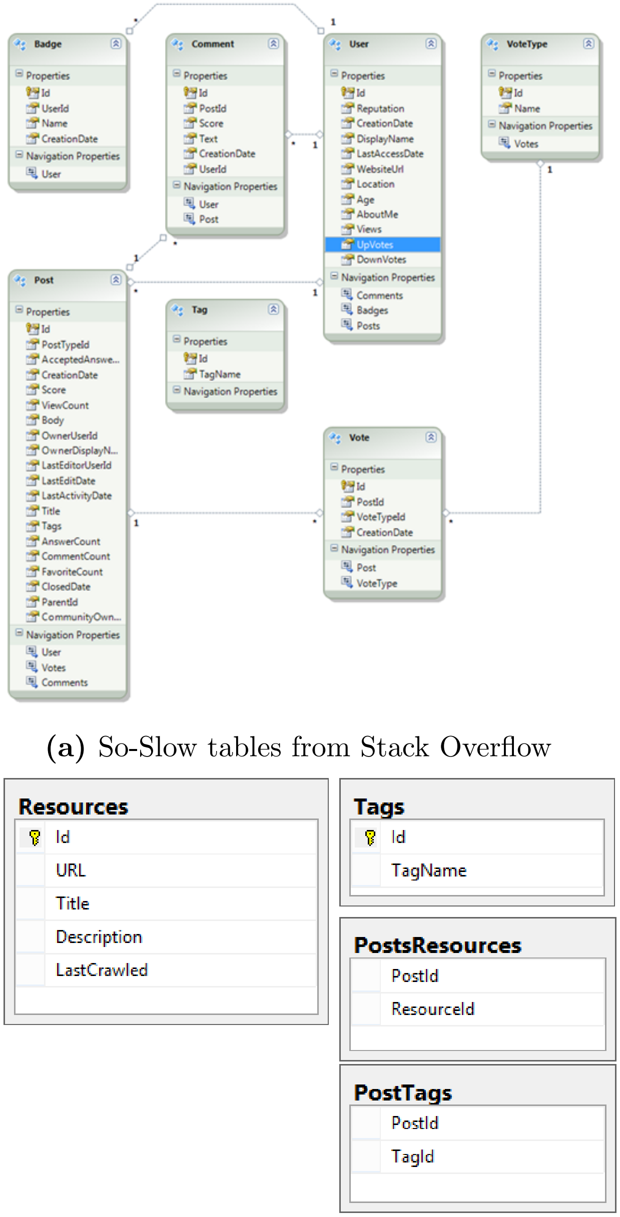
3.3 Data preparation

After importing the information into the database we analyzed both *posts* and *users* using a combination of different database queries[3], we extract the information related with posts that are questions and have one accepted answer, and users related with these posts. Using this process, we identified 2,148,455 questions post that have an accepted answer and a 2,774,809 post that try to answer these questions post.

Due to the physical capacity of our machine we reduce the size of our dataset to a smaller set. To address this, we use the top 10 tags with more answers related with answered question posts. Consequently, we extract 1,000 answered-questions post (but from the top 10 tags) in a random way. From these questions post, we extract all the related answers and the information that we need for the mining process.

Once we filtered the information in the database the following step was to tune the data into the information we required, such as transforming time into minutes by computing the

[3] These queries are not available in this document but can be requested through email



(a) So-Slow tables from Stack Overflow

(b) Additional tables after modifying So-Slow

Figure 2: Stack Overflow in the SQLServer data base representation

difference between dates as new attributes: user's life time, user-post life time, and answered time. Table 1 describes these attributes.

3.4 Training and Testing data generation

Finally, we export the information from the database as CSV files ready to be imported in WEKA.

Attribute	Description	Trying to Measure
User's Life Time (min)	Number of minutes between the user's registration and the last activity of the user	Experience
User-Post Life Time (min)	Number of minutes between the user's registration and the given answer	Experience
Views	Number of post viewed by the user	Experience
Parent Views	Number of visitors of the parent post	Experience
Votes	Number of votes given by the user	Experience
Reputation	Reputation of the user	Experience
Answer's Score Perc	Relation that shows the score of the post vs the maximum score for that parent post	Experience/ Information delivery
Answered Time (min)	Number of minutes between the question and the answer.	Knowledge
Number of Resources	Number of resources shared in the post	Knowledge
Winner Post Till Date	Number of winner post between user registration and the answer post.	Experience/ Information delivery
Winner Post	Total number of winning post of the user	Experience
Post extension	Number of character on the post	Information delivery /complexity
Winner	If the post was a winner post or not.	Class

Table 1: The list of attributes that are related with our definition of winning post

4 Data Mining

The data mining process aim to find hidden information from raw data by implementing algorithms and filters to get the better results. For this, we used WEKA to classify the attributes in order to identify those that are significant in the relationship of winning posts and key users. Figure 3 shows the data mining flow (knowledge flow) we followed to execute the classification.

The steps are as follows:

1. **Arff Loader:** This element loads the information into WEKA to be processed. In our case from a CSV file into an ARFF file, that is compatible with WEKA.

2. **Class Assigner:** This step selects the Class we want to predict. In this case, Winner that determines if the post is a winner or not.

3. **Cross Validation Fold Maker:** This step is in charge to deliver the training set and the test set to each classifier. In this case we used 6 classifiers.

4. **Classifier performance evaluator:** This step evaluates the results of the classifier and creates an output.

5. **Views:** The results of the classifier are presented in the information viewer prior analysis.

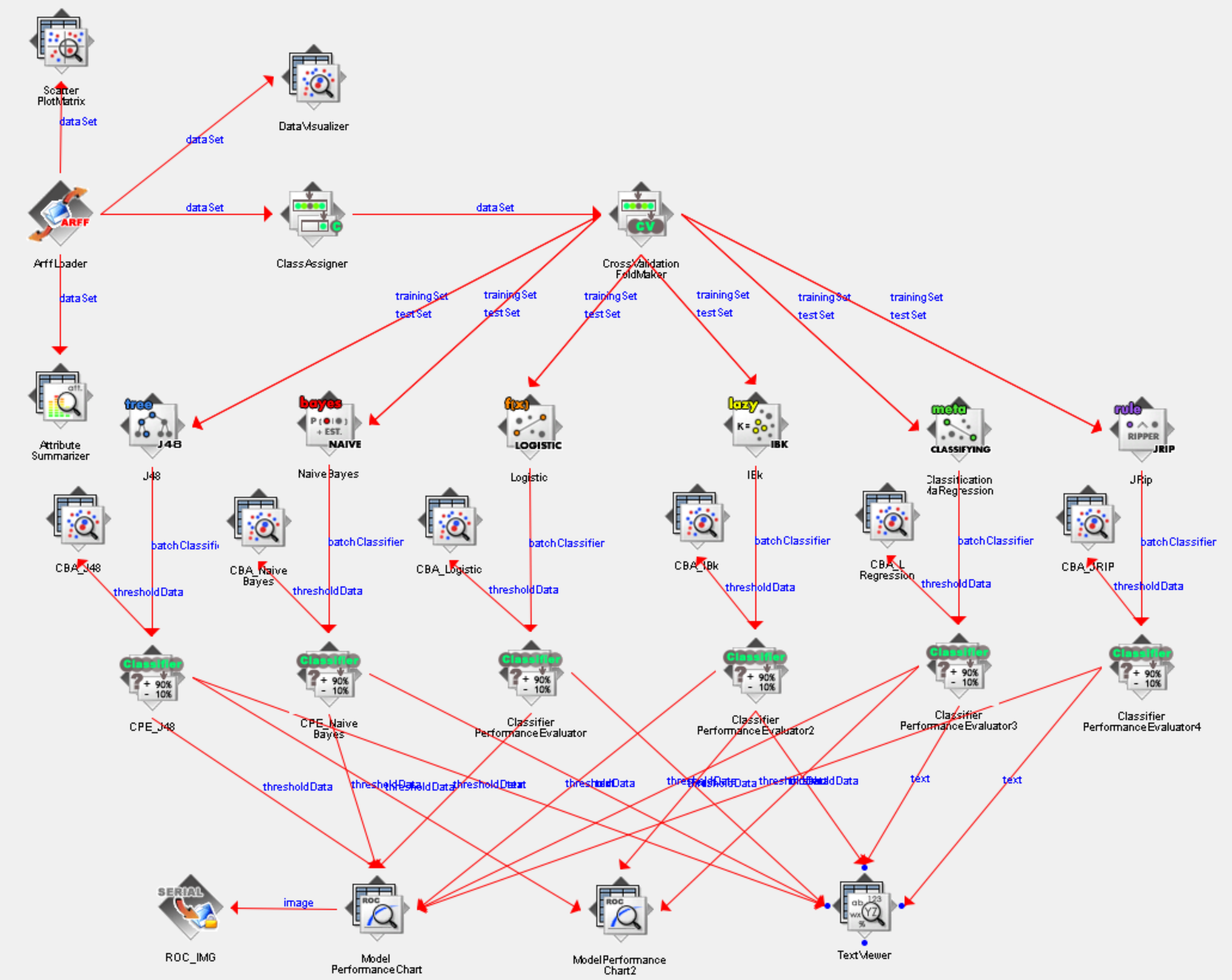


Figure 3: The data mining knowledge flow

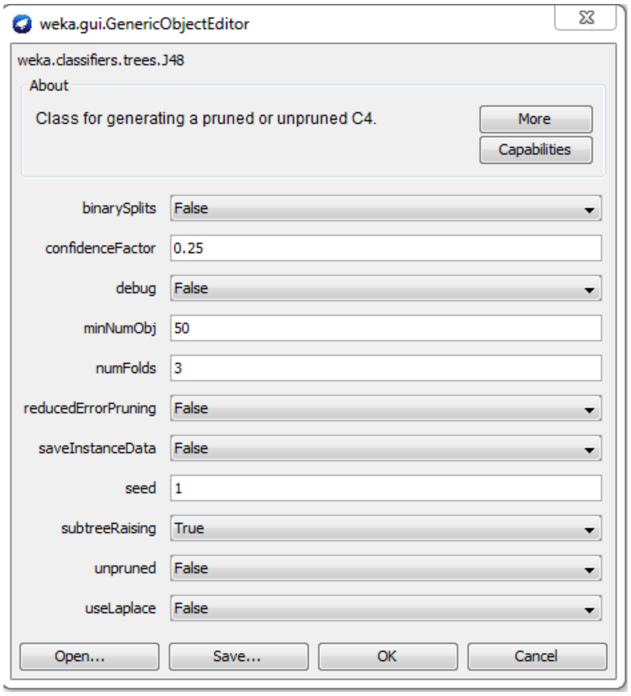
4.1 Winning posts

Following the knowledge flow, we executed 6 classifiers that are suitable for our data set: Naïve Bayes, Logistic Regression, J48, Classification via Regression, JRip and IBk. The results of the executions are described in Table 2 showing that J48 stands out over the others with 89.53% accuracy.

Given that J48 is a tree classifier it seams more comfortable to interpret while analyzing the data. Thus, we apply the M50 filter in order to reduce the size of the tree. Figure 4 presents the configuration for the classifier, and Figures 5 and 6 show the complete output of the classifier in WEKA. Finally, Figure 7 shows the tree resulting of the classifier. In the following sections we address our observations and analysis separately based on this tree.

Classifier	Correct %
Naïve Bayes	84.352%
Logistic	87.828%
J48	89.5351%
Classification Via Regression	89.476%
JRip	89.444%
IBk	83.8396%

Table 2: Results of the execution of the classifiers available in WEKA


weka.gui.GenericObjectEditor

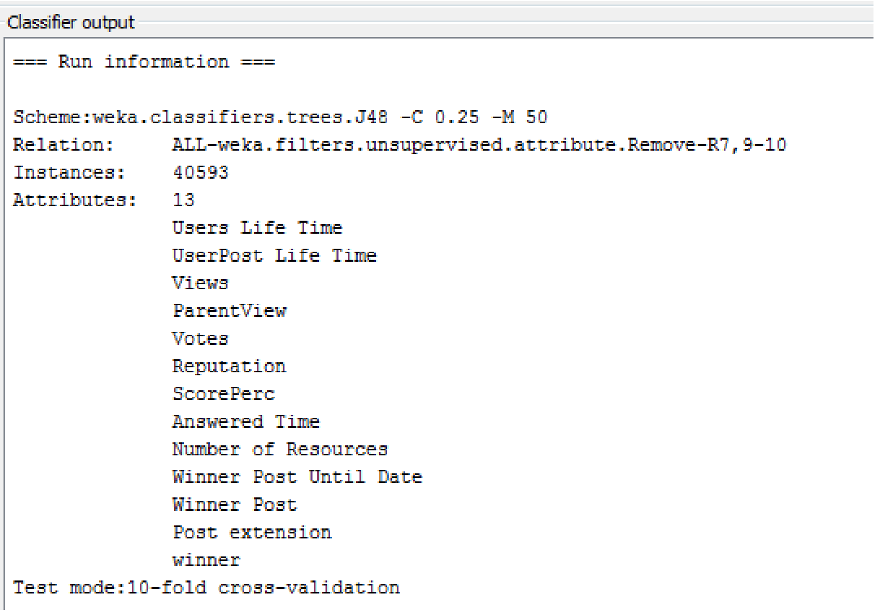
weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

binarySplits	False
confidenceFactor	0.25
debug	False
minNumObj	50
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False

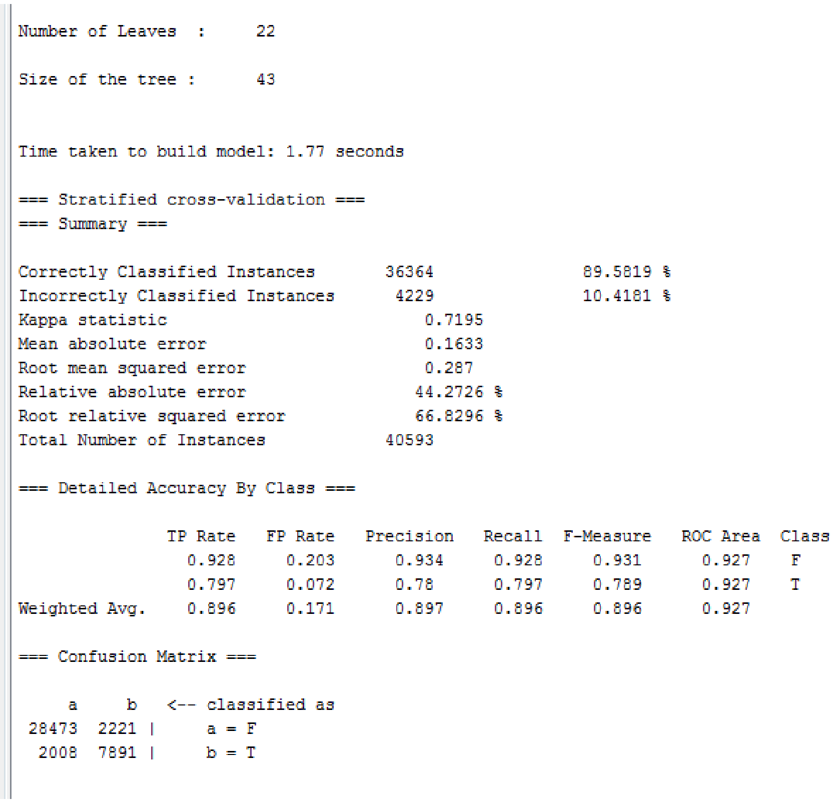
Figure 4: J48 Classifier configuration with M50 filter



```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 50
Relation:     ALL-weka.filters.unsupervised.attribute.Remove-R7,9-10
Instances:    40593
Attributes:   13
              Users Life Time
              UserPost Life Time
              Views
              ParentView
              Votes
              Reputation
              ScorePerc
              Answered Time
              Number of Resources
              Winner Post Until Date
              Winner Post
              Post extension
              winner
Test mode:10-fold cross-validation
```

Figure 5: J48 Classifier output for winning posts (Part 1)



```
Number of Leaves  :     22

Size of the tree :      43


Time taken to build model: 1.77 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       36364               89.5819 %
Incorrectly Classified Instances      4229               10.4181 %
Kappa statistic                          0.7195
Mean absolute error                      0.1633
Root mean squared error                  0.287
Relative absolute error                 44.2726 %
Root relative squared error             66.8296 %
Total Number of Instances            40593

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.928     0.203      0.934     0.928     0.931      0.927    F
                 0.797     0.072      0.78      0.797     0.789      0.927    T
Weighted Avg.    0.896     0.171      0.897     0.896     0.896      0.927

=== Confusion Matrix ===

     a     b   <-- classified as
 28473  2221 |     a = F
  2008  7891 |     b = T
```

Figure 6: J48 Classifier output for winning posts (Part 2)

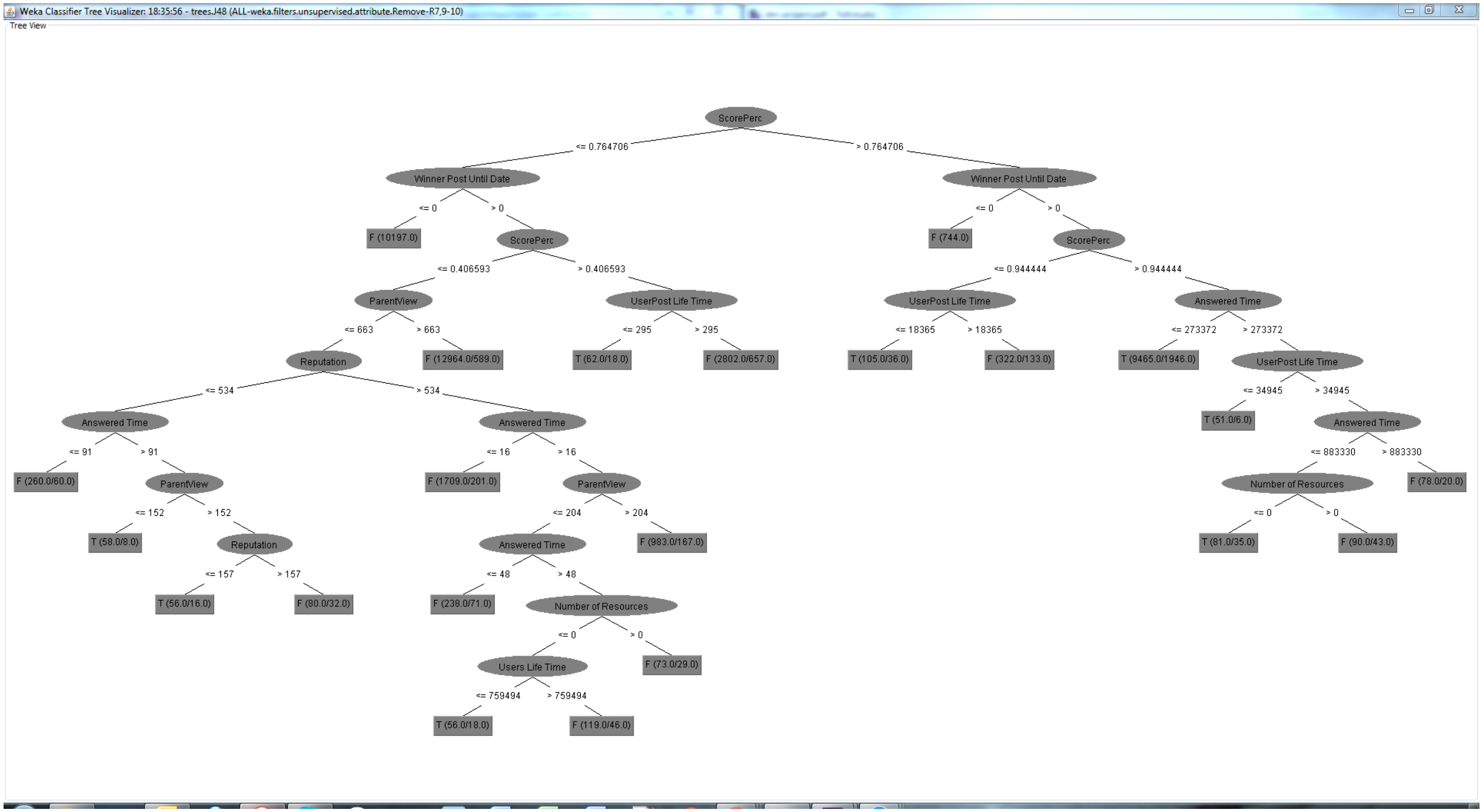


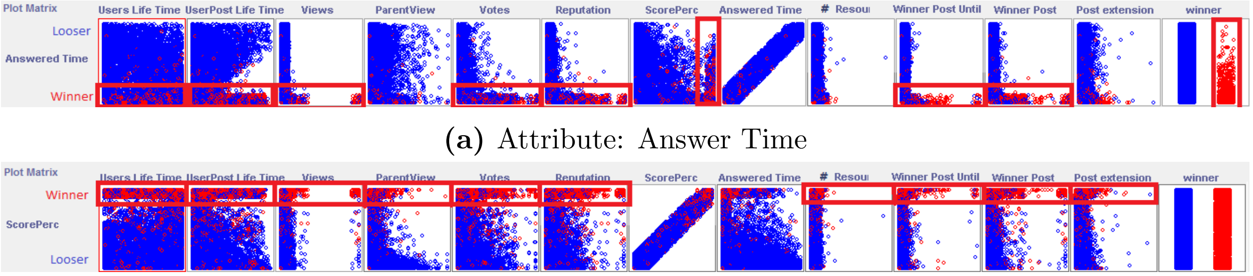
Figure 7: J48 Tree

4.2 Observations (1st set)

After selecting and executing the classifiers we proceed to analyze the results and findings.

4.2.1 OBS 1: Winning posts attributes

Figure 8 exhibits the values distribution according every attribute. As noted in Figure 8a *Answered Time* is an important attribute to classify the information due that the density of the winner posts is in the first part of the plot. This means that winner posts tend to have a very short time of answer compared with no winner posts. Finally, Figure 8b is the attribute *Score Percentage* that reflect the value that the community gave to qualify a best answer. Hence, as seen in the figure, the winner posts are distinguishable yet not dispersed. This means, that this attribute is significant to find winning posts.



(a) Attribute: Answer Time

(b) Attribute: Score Percentage

Figure 8: Classifier attributes plot analysis

Table 3 presents the winner attributes after running the evaluation of WEKA. Given the position of each attribute in the evaluation we can infer that:

- *ScorePerc* is the top 1 of most important attributes for all the evaluators. This supports the previous statement in which we argue that the community is the most appropriate to evaluate a winning post and its content.

- *Answered time* is placed in the top 5 for all the evaluators, and reflects the expertise of the user, given the fact that time of response is shorter if the user has knowledge of the topic.

- *Winner Post Until Date* is in the top 3 for 3 our of 4 evaluators, this can signify their importance, however, the 3rd evaluator has this attribute in place 7th which diminishes it at some degree. This attribute reflects both experience and activity on the community, yet does not imply a winner post.

- *Post Extension* does not present information to a conclusive position. For 2 of the 4 evaluators this attribute is in the lower 5 and for the pother two in the upper 5. Therefore, the extension of the answer can not determine for sure if the post is a winner.

- *Reputation* is an attribute that instinctively means that the user is recognized by the community and it should be important in the classification. But beyond the expectations does not appear in all the evaluators, only 3 out of. Hence on those where appears in the lower of the top 5 which conduct us to believe that it is not as important as our intuition suggested.

- *Winner Post* is on the top 6 in general, despite the fact that only participates in 3 out of 4 evaluators. This attribute still remains important as a reflection of user's expertise and activity in the community.

- *Parent View* is the last attribute to be on the top 5 in general, despite the fact that only participates in 3 out of 4 evaluators. This attribute still remains important as a reflection popularity of the question on the community.

The remaining attributes are still important for the classification and we are not discard it them, but we consider that due the place in the score are not as critical as those mentioned before.

Evaluator	**Search**	**Attribute**
CfsSubsetEval	BestFirst	1 - ScorePerc
		2 - Winner Post Until Date
		3 - Post extension
CorrelationAttributeEval	Ranker	1 - ScorePer
		2 - Answered Time
		3 - Reputation
		4 - Post extension
		5 - Parent View
		6 - Winner Post
		7 - UserPost Life Time
		8 - Winner Post Until Date
		9 - Votes
		10 - Number of Resources
		11 - Views
		12 - Users Life Time
GainRatioAttributeEval	Ranker	1 - ScorePerc
		2 - Winner Post Until Date
		3 - Answered Time
		4 - Winner Post
		5 - ParentView
		6 - Reputation
		7 - UserPost Life Time
		8 - Votes
		9 - Users Life Time
		10 - Post extension

Evaluator	**Search**	**Attribute**
		11 - Number of Resources
		12 - Views
InfoGainAttributeEval	Ranker	1 - ScorePerc
		2 - Winner Post Until Date
		3 - Winner Post
		4 - Answered Time
		5 - ParentView
		6 - Reputation
		7 - Votes
		8 - UserPost Life Time
		9 - Post extension
		10 - Users Life Time
		11 - Views
		12 - Number of Resources

Table 3: Evaluation of attributes

4.2.2 OBS 2: The resulting winning posts tree

Figure 7 shows the resulting tree of the J48 classifier. After analysis we can conclude:

- By analyzing the root of the tree and counting the winning posts on both branches of the root, we can conclude that having a high score (given by users) is a tendency for winner posts. In this case, the score percentage threshold is approximately 76%, where when is lower, 232 post were classified as winners, and when is greater, 9702 posts were classified as winner posts.

- Hence, in the left subtree it is very hard to find experts, given that other variables like reputation and parent view are too low to conclude anything.

- A formal expert, someone who is a professional on the field, is hard to recognized in the community while is a new user. In the right subtree of the root of the three having no winning post automatically discards the user as an expert not having winning posts. This supports our theory that an user's professional credential can not be found in the stack overflow data to place it as an expert.

- Finally the attributes that influence the most to have winning posts, classifying 9465 of the wining posts are: High users' acceptance (94%), the user most have had at least one winner post to the date, and the answer time must be in the range of the first 189 days. This 189 days might seem excessive, however, we have found that difficult questions take time to be answered, hence we can inferred that these type of questions indeed are answered by experts.

Finally, we conclude that the most representative branch to predict a winning post are those that follow the 9465 winning posts. Then, we select those users owners of those 9465 posts and we marked them as **potential experts**. Then we filter the original data for those attributes of the user, and perform a new data mining process to finally get the attributes that influence the most while selecting a informal expert, stating from a very representative set of winning post supporting our original hypothesis.

4.3 Potential experts data mining

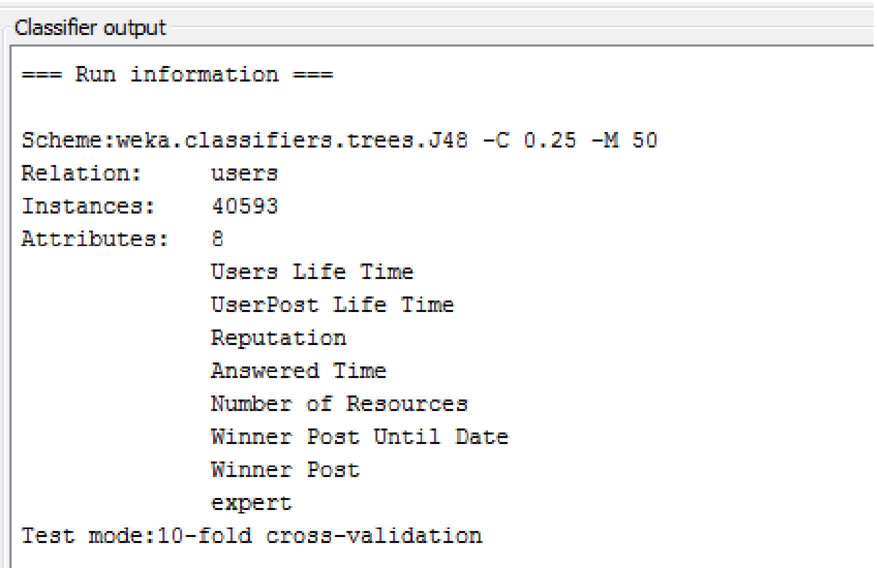
Having the data results from the winning post, we apply our hypothesis where those users owners of the winning posts are considered experts, and then we filter the set of users of those within the highest range of winning posts.

Then, we apply data mining techniques to discover the variables that influence the most while determining these experts. Similarly, we executed 6 classifiers as well. The results of the executions are described in Table 4 showing that Classification via Regression is slightly better than J48. However, due that the tree visualization is easier to analyze, we choose to user J48 classifier.

Classifier	**Correct %**
Naïve Bayes	75.092%
Logistic	76.04%
J48	77.36%
Classification Via Regression	77.48%
JRip	76.9%
IBk	73.52%

Table 4: Results of the execution of the classifiers available in WEKA in te users dataset

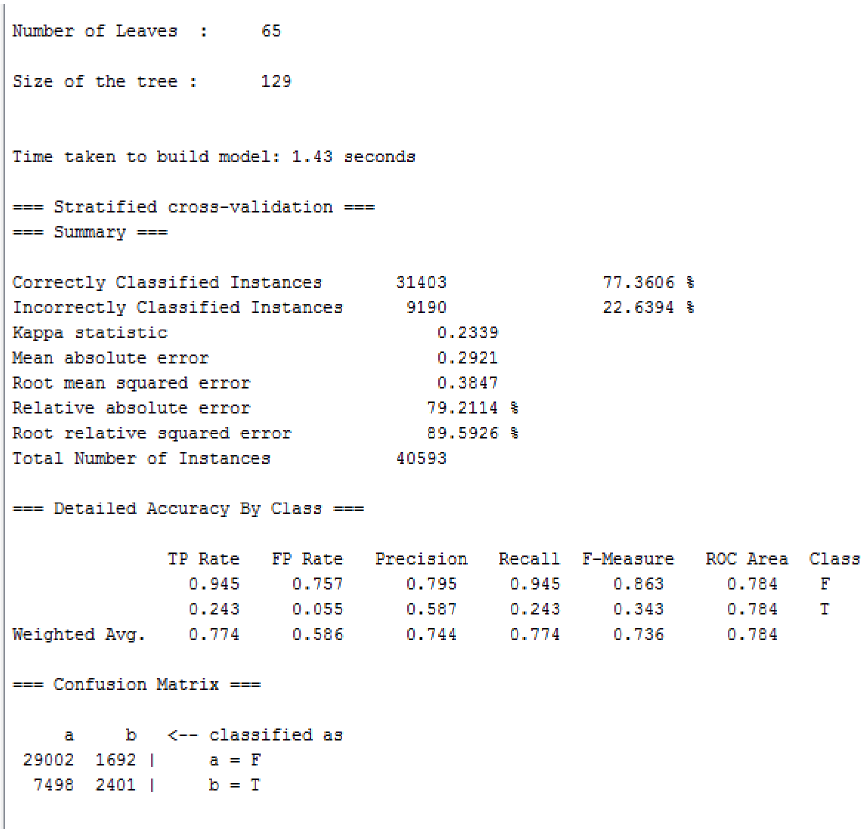
Figures 9 and 10 show the complete output of the classifier in WEKA. Finally, Figure 11 shows the tree resulting of the classifier. In the following sections we address our observations and analysis separately based on this tree.



```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 50
Relation:     users
Instances:    40593
Attributes:   8
              Users Life Time
              UserPost Life Time
              Reputation
              Answered Time
              Number of Resources
              Winner Post Until Date
              Winner Post
              expert
Test mode:10-fold cross-validation
```

Figure 9: J48 Classifier output for winning posts (Part 1)



```
Number of Leaves  :     65

Size of the tree :      129


Time taken to build model: 1.43 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       31403               77.3606 %
Incorrectly Classified Instances      9190               22.6394 %
Kappa statistic                          0.2339
Mean absolute error                      0.2921
Root mean squared error                  0.3847
Relative absolute error                 79.2114 %
Root relative squared error             89.5926 %
Total Number of Instances            40593

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.945     0.757     0.795     0.945     0.863      0.784    F
                 0.243     0.055     0.587     0.243     0.343      0.784    T
Weighted Avg.    0.774     0.586     0.744     0.774     0.736      0.784

=== Confusion Matrix ===

     a     b   <-- classified as
 29002  1692 |     a = F
  7498  2401 |     b = T
```

Figure 10: J48 Classifier output for winning posts (Part 2)

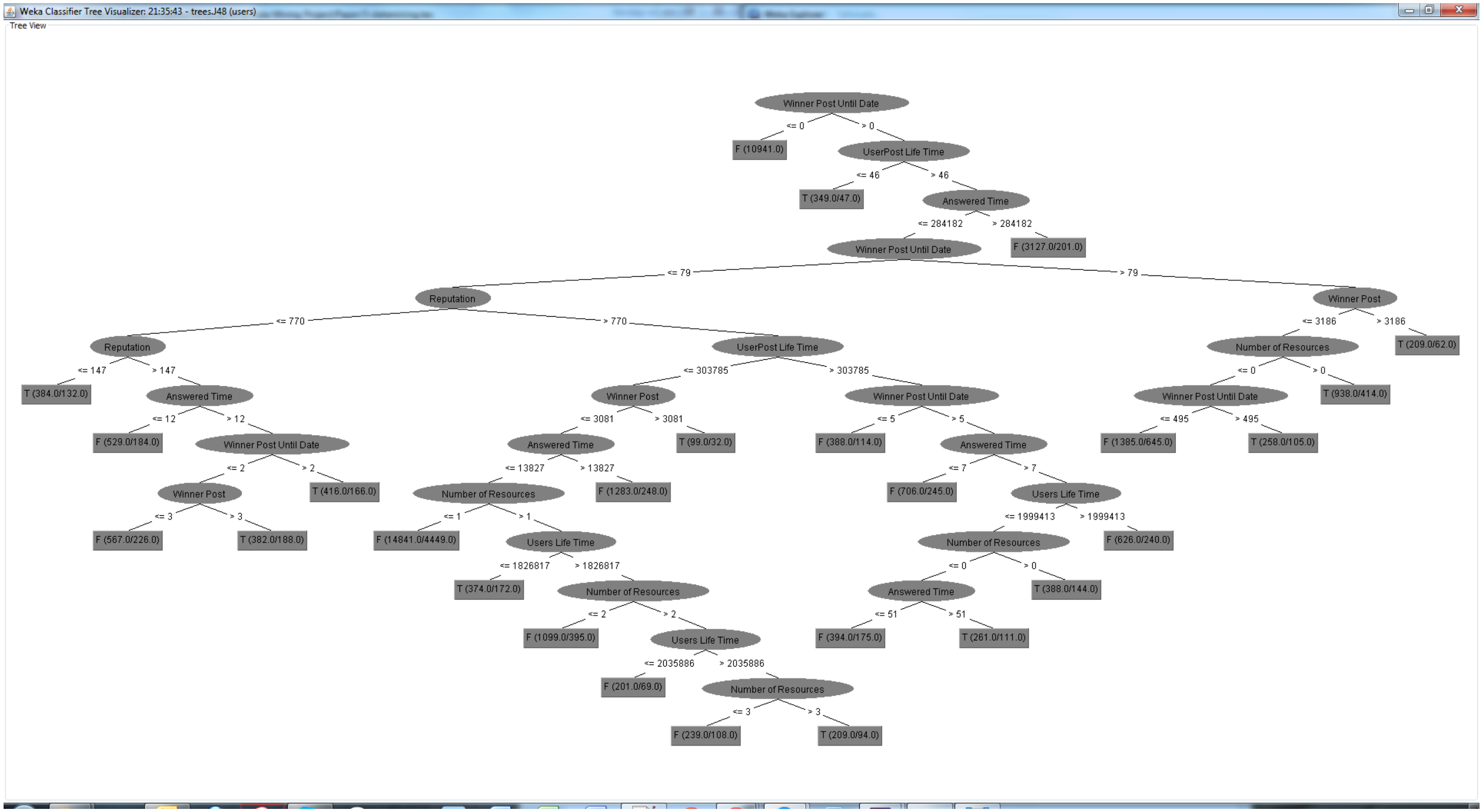


Figure 11: J48 Tree

4.4 Observations (2nd set)

4.4.1 OBS 3: Informal experts attributes

Table 5 presents the winner attributes after running the evaluation of WEKA. Given the position of each attribute in the evaluation we can infer that:

- *Answer Time and Winner posts until date* are te most representative attributes to categorize and expert.
- *Reputation and Number of Resources* even though are important are no definitely. Hence, Reputation is gain by different means so we can't yet discard it.

Evaluator	**Search**	**Attribute**
CfsSubsetEval	BestFirst	1 - Answered Time
		2 - Number of Resources
		3 - Winner Post Until Date
GainRatioAttributeEval	Ranker	1 - Winner Post Until Date
		2 - Answered Time
		3 - Winner Post
		4 - Reputation
		5 - UserPost Life Time
		6 - Users Life Time
		7 - Number of Resources
InfoGainAttributeEval	Ranker	1 - Winner Post Until Date
		2 - Winner Post
		3 - Answered Time
		4 - Reputation
		5 - UserPost Life Time
		6 - Number of Resources

Table 5: Evaluation of attributes

4.4.2 OBS 4: The resulting informal experts tree

Figure 11 shows the resulting tree of the J48 classifier. After analysis we can conclude:

- we can place a big groups of experts in the right subtree of the fourth node, that stated that users with more than 79 winner posts until their last wining post can be considered potential experts if their answer time is less than 197 days and
- Reputation and number of resources split the set in several branches, hence we cannot use it to provide a strong conclusion.

5 Evaluation

As presented in Section 4.2 our proposal is able to identify informal experts from the data of stack overflow. To prove the feasibility of our proposal we extracted 5 users in the classification and studied their profile sin the Stack Overflow database

As mentioned in Section 1, our definition for an informal expert is based in a set of condition. Additionally, and based on the results obtained in Section 4.2 we define in Table 6 the following thresholds:

Attribute	**Measurement**	**Threshold**
Reputation (R)	Standard numerical value	Not So Important
Answered Time (AT)	Standard numerical value	197 days
Winning Post (WP)	Standard numerical value	79
Badges (B)	14 types of golden badges [4]	minimum 3 types

Table 6: Thresholds to consider an informal expert

According to this, we extracted 5 random users from the classified set, and find their profiles in stack overflow. Table 7 shows this information, and the values in their profiles (extracted using Stack Exchange [11], an online tool to perform queries) according measurements, to present a comparison. The last column determines whether is (+) or not (-) an informal expert.

username	**profile website url**	R	AT	WP	B	**Result**
alex-ford	X/498624/alex-ford	9306	22	251	3	+
lukeh	X/55847/lukeh	81604	148	669	2	+
jaredpar	X/23283/jaredpar	254412	67	2334	8	+
ian-boyd	X/12597/ian-boyd	32915	64	127	6	+
dema80	X/863564/dema80	1937	2	17	0	-

Table 7: Predicted informal users and data extracted from their profiles to compare with the predictions ((X) corresponds to the common url http://stackoverflow.com/users/)

Finally, the prediction of our classifier was accurate to place this 5 random experts in their prediction. As we evaluate them, 4 of them had perfect record to be classified according the threshold. Hence, we can discard that non informal expert, and we can see that his reputation is consistent with the badges category. Therefore, these two variables can be substitute and the user could be consider a potential expert given that has good response time, wining posts, but low reputation. these are characteristics of a newcomer.

[4] The mechanism of badges acquisition is determined by Stack Overflow rules

6 Conclusion

As mentioned in Section 1, for software developers is very important to count with reliable sources of information while browsing the internet for answer to their development problems. Social communities present themselves as very resourceful site to retrieve valuable information to solve a problem. However, due the massive user, it can be very noisy to find a proper answer. However, Stack Overflow provide a mechanism based on votes and selection, to determine the best answer and the winning answer-post.

Despite that finding answers is good, finding expert users is also desired, not only to browse different approaches, but also to follow knowledge and projects. We argue that winning posts are valuable sources to identify expert users. Though we consider that having a winning post can't be the enough to qualify an user as expert. As a consequence, we propose to mine the available information to find those additional attributes that predict an expert.

We mined the data from the Stack Overflow 2012 dump, available int eh MSR challenge website [9]. Using a modified version of the So-Slow tool we imported the raw data into a SQLServer database and preformed some cleaning and transformation procedures to get the must suitable data for our approach. Later, we query the data to crate the proper CSV files to imported in the data mining tool WEKA.

Furthermore, we implemented different classifiers over the data and finally we chose J48 given the acceptable accuracy over de 80%. As a result the J48 tree exposed interesting findings such as the attributes that influence the most in the winning posts, such as: Score, Answered Time and Winner Post until date. Consequently, for the users the nos relevant attributes where: Reputation, Lifespan, and Winning posts.

Finally, we extracted 5 random users predicted by the classifier and search their profiles on Stack Overflow to compare results according the viability of them as informal experts. The results were compelling, even though the data set was from 2012 and this users profiles are in 2013, the differences in the predictions still place them in the threshold to be considered informal experts.

Acknowledgments

This paper is part of the Spring 2013 Datamining Course and it was presented as part of the final project of the course. We thank Dr. Alex Thomo and Dr. Maryam Shoaran for their guidance during the course.

References

[1] "Stack overflow." [Online]. Available: http://stackoverflow.com/

[2] S. Saffrom, "So-slow: Stack overflow creative commons database importer." [Online]. Available: https://github.com/SamSaffron/So-Slow

[3] Machine Learning Group at the University of Waikato, "WEKA: Waikato Environment for Knowledge Analysis." [Online]. Available: http://www.cs.waikato.ac.nz/ml/index.html

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: ACM, 2008, pp. 183–194.

[5] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 850–858.

[6] A. Bacchelli, L. Ponzanelli, and M. Lanza, "Harnessing stack overflow for the ide," in *Recommendation Systems for Software Engineering (RSSE), 2012 Third International Workshop on*, 2012, pp. 26–30.

[7] A. Barua, S. Thomas, and A. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical Software Engineering*, pp. 1–36, 2012.

[8] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&#38;a site in the west," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 2857–2866.

[9] Alberto Bacchelli, "Mining challenge 2013: Stack Overow. In The 10th Working Conference on Mining Software Repositories." [Online]. Available: http://2013.msrconf.org/challenge.php

[10] L. S. Carlos Gómez, Brendan Cleary, "A study of innovation diffusion through link sharing on stack overow," 2013 (To appear).

[11] "Stackexhange daata explorer." [Online]. Available: ttp://data.stackexchange.com/stackoverflow/queries