Determining the Success and Failure of Online Communities by the Presence or Absence of Web 2.0 Tools and Tags

Stephanie Lukin
slukin@soe.ucsc.edu

December 5, 2011

Table of Contents

Abstract 1
1 Introduction 1
2 Related Work 2
3 Methodology 3
  3.1 Software . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
  3.2 Proposed Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
    3.2.1 Success Experiments with K-Means . . . . . . . . . . . . . . . . 4
    3.2.2 Web 2.0 tools Experiments with EM . . . . . . . . . . . . . . . 5
    3.2.3 Tag Experiments with JRip . . . . . . . . . . . . . . . . . . . . . 5
4 Experimental Results 6
  4.1 Success Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  4.2 Web 2.0 Tools Experiments . . . . . . . . . . . . . . . . . . . . . . . 7
  4.3 Tag Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
5 Discussion 9
  5.1 What Was Accomplished and Not: Future Challenges . . . . . . . . . . . 9
  5.2 Self Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
6 Acknowledgements 10
References 10
Abstract

Using data collected from approximately 1,500 IBM Lotus Communities, we intend to explore characteristics of Web 2.0 usage. We hypothesis that there will be some combination of tools and usage that lead to a successful Community. We also believe that some Communities may succeed based upon their topic and with what tags they represent themselves. This is a novel approach, so we will use forms of unsupervised machine learning in order to cluster together Communities with similar tool usage patterns. We also intend to utilize rule based learners for finding similar trends in Community tags. At this first stage in our research, we are unable to support either hypothesis, but have found interesting tool usage patterns that lead us to believe that there is a relationship; we just have not been able to quantify it yet.

1 Introduction

The advent of the internet and online forms of communication allow us to share our ideas and visions freely with a global audience. Seeing this as a chance for easy collaboration and knowledge sharing, International Business Machines (IBM) has created a resource for its employees and partners to connect across the globe. IBM Lotus Communities is an online resource that facilitates this sharing of a common interest through a variety of Web 2.0 tools.

This project will look at websites, provided through IBM’s Lotus Communities service, that use a variety of Web 2.0 social and collaborative tools. Some sites are more successful than others, so we want to understand what those sites are doing right. We first define a success assumption. Then we use an unsupervised, machine learning approach to cluster together sites that have similar trends in their Web 2.0 tool usage (e.g. does a Community use wikis, file sharing, etc. and if so, to what extent?). We look within each cluster to see if they exhibit our success assumptions. Finally, given a new website with certain Web 2.0 tools, we can predict if it will be successful or not. We also have information about the tags with which each Community self identifies. We can use a different machine learning approach in order to analyze the sets of tags and determine if a Community topic is correlated to our success assumptions as well.

Lotus Communities is a commercial service provided by IBM that gives its members the opportunity to create, browse, or search (using tags) through a variety of topics via a central website. These are called Communities and can be business or leisure oriented. Communities can be used as a homepage for members with various interests, project teams, client relationship teams, and committees. They can have the following Web 2.0 style social and collaborative tools active: forums, blogs, bookmarks, files, wikis, and news feeds [20].
We want to look at how successful Communities are using Web 2.0 tools. We hope to gain insight into what tools are working very well so that we can say “in order to be successful, do like these Communities have done”. By understanding the tools and Community success, we can create more comprehensible and more attractive sites in the future, as well as establishing a good interface for sharing of resources and knowledge.

We also believe that the success of a Community may be related to its content. By grouping together similar Communities based on their tags, we can see if success varies across different topics or even within the same topic. If the latter is true, perhaps certain Communities are better at utilizing Web 2.0 tools than their shared topic competitors.

Defining the success of a Community is not trivial task. Based upon the given dataset and its features, we predict that success can be correlated to the total amount of text (higher text could indicate more contributions), a high amount of activities (a single activity is a forum, wiki, etc), and rate of contributions (the higher this ratio, the more active participants in the Community). We will use these three features as our success assumptions.

Because we do not understand the nature of how the Communities use Web 2.0 tools, we will perform clustering algorithms over the dataset. More specifically, K-Means will be used to examine how our success features are observed throughout the Communities. Then, the Expectation-Maximization (EM) algorithm will be performed over the data in order to see how the Communities are using these social and collaborative tools and what similar trends emerge. Finally, we will compare the success metric within each cluster the EM algorithms give us. In order to analyze the tags, we will use a propositional rule learner over our dataset. JRip is a supervised, classification algorithms that will use the cluster number (determined by our K-Means experiments) for their class label.

The algorithms and how we used them on our dataset will be described more in section 2.

2 Related Work

Unsupervised machine learning is good when we are "not told what the categories are: [the algorithm] determines the most useful segmentation [and is] particularly useful for clustering data where clusters are not known in advanced" [9]. This is our main problem: we have a dataset of Communities with no indicator of their success. There are two unsupervised learning algorithms we believe will be applicable with our assessment–K-Means and EM [10], [15], [16]. K-Means learning makes a hard assignment. This means that a Community can only belong to one unique cluster [2]. We believe for determining our success clusters, it will be good to have distinct boundaries. The EM algorithm, on the other hand, makes a soft assignment [2]. We believe it will be better to be more lenient with these classifications, because we are uncertain as to if the success clusters will be preserved in the first place. Gaussian mixture models will be used because they are good at dealing with latent variables which can give us insight into the way the dataset operates [2].
A common evaluation method for clustering algorithms is by using reference labels. "If reference labels are available for the cluster data, external cluster validation can be used to compare the cluster partition with the set of reference labels" [1]. By performing our K-Means success experiments, we are creating our reference labels for the EM tool experiments.

Cluster analysis is becoming multi-disciplinary across many fields and applications. Such diverse examples are social psychology [1], neighborhood census data analysis [19], vector quantization[4], [11], image segmentation [6], and marketing [18]. Cluster analysis encourages exploratory analysis [1] which is what we attempt to do with our IBM Communities data. To our knowledge, this dataset is unique to our studies, and we will see if cluster analysis will be good for us. We hope to uncover insights into a better understanding of online websites and collaborations to improve upon for the future.

Our motivation for these experiments is based upon the research of many social media psychologists interested in understanding the nature of and designing online websites [14]. Even more are dedicated to trying to improve collaboration efforts [5], [13], [8] and understand the relationship between the users and the technology [7].

Finally, our rule based analysis is motivated by related work in psychology and analysis of human behavior. The analysis involves concepts and the algorithm is motivated by "the search for and listing of attributes that can be used to distinguish exemplars from non exemplars of various categories" [3]. Several research topics include looking at human categorization behavior through these rule based learners [17] as well as looking at rational analysis of humans [12]. If these psychology studies can learn to categories and deduct rules about human behavior, a task looks at classifying strings into similar clusters based on a rule set seems like a good start for understanding online concepts and behaviors.

3 Methodology

The dataset contains 1,418 Communities. For each Community, we have information about their current membership numbers and number of active contributors. The data also includes whether or not a Community supports bookmarks, news feeds, files, blogs, wikis, forums, and, if the tool is supported, how many a Community has. The Community creator can specify tags with which to identify the site, so that it will show up in the most searches with similar keywords. After our initial parsing of the tags, we found there were approximately 5,000 unique tags. But because there is no language boundary across the Communities, we had to eliminate the tags with foreign characters so our parser would run. Now, we have observed 3,775 unique tags across all Communities. The dataset also contains information about the total number of text, total number of activities, and the average distance of the community members and main contributors. The dataset we will use is only 13% of the original number of Communities. Before I received the data, 87% of all Communities were discarded because they did not have at least 50 activities (an indicator to us that there is not much going on within the Community—a poor research subject).
3.1 Software

Weka will be used for running the K-Means and EM clustering algorithms, as well as providing simple 2-Dimensional plots. Weka also supports the rule based learner, JRip. As tests are run with more features, we will move to graphing in higher dimensions using Octave, such as visualizing our clusters as a mixture of Gaussian distributions. In order to facilitate the extraction of tags from the Communities dataset, python scripts will be written.

3.2 Proposed Experiments

Since we have no indicator of success from the given data, we will use unsupervised techniques to label the data.

We make one assumption:

\( a_0 \): success is defined by the total amount of text in Community, the rate of contributions, and the number of activities

We will use K-Means clustering to determine our success clusters based on \( a_0 \). Once we have the distribution of success features over our data, we will create a new dataset and assign each community the cluster number that the 3-feature K-Means success cluster analysis produced from \( a_0 \).

We have two test hypothesis:

\( h_1 \): success is influenced by the Web 2.0 tools that a community uses
\( h_2 \): success is influenced by the tags with which a community associates itself

We hypothesis that communities with tags similar to communities deemed successful will also have similar success features. Furthermore, communities with similar Web 2.0 tool usage will also have similarly successful features. Once we determine our initial success clusters, we will run EM to observe the Community’s usage of Web 2.0 tools, hoping that the K-Means success clusters are preserved within the Web 2.0 tools clusters. Finally, we will use JRip to analyze the tags of each Community and predict the correct class label (again, based on the K-Means success clusters).

3.2.1 Success Experiments with K-Means

We explicitly define success features based on some of the given dataset information information. We want to ensure that they are significant features, indicative of a successful Community. As aforementioned, we will consider the total amount of text in a Community, the total number of activities, and the rate of contributions as our hypothesized success features, \( a_0 \).
We will run K-Means tests over the dataset using our success features in order to observe clusters with similar success values. For each Community, we will add this cluster number to the feature set for future comparisons. Weka will automatically do this for our Web 2.0 tools dataset, and a simple python script can be created to do this for our tag dataset. In our Experiments Section, we will discuss the characteristics of the Communities within each cluster assigned by the algorithm.

### 3.2.2 Web 2.0 tools Experiments with EM

We will then use EM distributions for clustering Web 2.0 tool usage with mixture of Gaussian models. Gaussian distributions assume the features are continuous. Mixtures of Gaussians can be used to fit how much the Web 2.0 tools are used (the number of feeds a Community has, the number of wiki pages, number of forum topics, etc.).

Using this EM algorithm, we will train on the Web 2.0 tool usage over all of the data. Now that we have tool based clusters from the EM algorithm, we can look in each cluster to see if the Communities inside it are of the same success cluster defined by the K-Means algorithm. We will now have distributions which predict the success of a Community based on which cluster of tool usage into which they fall. In the future, we hope to take a new Community and fit it into our distribution to determine its future success based on its tool usage patterns.

### 3.2.3 Tag Experiments with JRip

Our other experiment is looking at the tags of a Community to see if they are correlated with a Community’s success. We have created a matrix that encapsulates all observed tags and a boolean value that indicates whether or not a community self identifies with that specific tag. There are 3776 observed tags. In addition to the booleans, the matrix will also contain the clusters as determined by the K-Means analysis. We will attempt to group together similar tags and communities; we will then look at the success features in each group and determine how similar/dissimilar they are. This analysis can help us determine if certain groups of tags are more successful than others.

For this analysis, we will use the JRip algorithm. JRip is a propositional rule learner based of the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm. It cycles between growing and pruning phases in attempts to optimize the information gain produced by every possible rule. After the initial rule set is created, the algorithm attempts to optimize, and deletes rules that would increase the information gain if they were removed.
4 Experimental Results

From our tests, K-Means did not produce an equal distribution of Communities. The algorithm was good at clustering together unique Community behavior, but these made up only 6% of the total number of communities (table 1). Our unbalanced distribution did give us the opportunity to look at the overall patterns of Community behavior, proving insightful for understanding the kind of characteristics Communities in different clusters have.

In order to proceed with our experiments, we used EM on our success features to get a more balanced distribution in terms of spread as well as provide a more informative visualization of the data. For the experiments with Web 2.0 tools and tags, the success clusters determined by our new EM tests are used.

Our experiments with EM could not support $h_1$. We ran an EM algorithm over each of the six Web 2.0 tools looking for preservations of the success clusters, and did not find any. Our attempts and further examination of the data will be described more fully below.

The JRip supervised technique failed on our tag dataset. Even after tuning algorithm parameters, the same results appeared. In attempts to continue with our hypothesis, we implemented other rule based learning algorithms supported by Weka: Decision Rules and PARK. All three algorithms failed to support $h_2$ and will be discussed below.

4.1 Success Experiments

We began our K-Means experiments by seeing how each success feature permeated through the Communities by itself; not considering the other success features yet. Through this, we got a good idea of how the dataset is distributed in terms of amount of text, contributors, and activities (figure 1). We see there are distinct boundaries between the created clusters for each success feature. This encouraged us because it seemed that our success features from $a_0$ could be easily and cleanly broken up (thanks to K-Mean’s hard assessment)

We also ran the experiment clustering on two success features. In figure 2, we can see the centroids produced by both the rate and number of activities, and how each cluster’s characteristics comprise of both success features. We could now start to see the relationships across each success feature in a Community. Again, we can see semi-distinctive boundaries defining each cluster.

We then combined all three features into our 'final' but 'not good enough' distribution. The distribution was unbalanced over our success features. We did not think this was an acceptable distribution of the data points over each cluster: 74% of the data set was contained within a single cluster. Our 'final' distribution was with five clusters. At five clusters, we had a good breakup of Community characteristics (summarized below) even though the Communities in each cluster was unbalanced. Greater than five clusters, the algorithm produced even more small clusters, still with one largest cluster.
Therefore, we decided to run the EM algorithm for our three feature success analysis. Table 1 shows that the EM for success was more balanced over the dataset than K-Means with its largest cluster at 49% instead of 74% and its smallest at 3% and 7% compared against 1% and 1%. What was considered clusters 2, 3, and 4 in the K-Means experiments kept the same centroid and characteristics of its activity, text, and rate level. Two clusters were flipped in the EM experiments, but still displayed the same characteristics as previously.

The breakdown of cluster characteristics can be seen in figure 3, indicating the number of activities per community, total amount of text, and rate of contributions respectively. In summary we conclude the following about our Communities clustered over our success assumption $a_0$ (the percent is the percentage of the dataset that was in each cluster with EM):

- cluster 0 (49%) has low activity, text, and rate
- cluster 1 (7%) has medium activity, medium text, and high rate
- cluster 2 (3%) has high activity, medium text, and high rate
- cluster 3 (21%) has medium activity, low text, and low rate
- cluster 4 (20%) has medium activity, medium text, and low rate

Because the EM results seem more stable and have more evenly distributed clusters, we will use these in our future analyses.

### 4.2 Web 2.0 Tools Experiments

For this series of experiments, we ran a new EM algorithm on each of the Web 2.0 tools: forums, file sharing, blogs, wikis, news feeds, and forums. Our hope was to find that at least one group had very strong cluster similarities to the success clusters already formed. For each tool, we used EM to get an understanding of how the Communities within each new cluster used that specific tool. For example, our tests showed that cluster 4 of the forum tools have very few forum posts and topics, while cluster 1 consisted of Communities that all supported forums but also had the most number of singleton posts. These are the kinds of insightful information we could use to help create better Community websites; based upon the success or failure of Communities because of the way they utilize their tools, we can try to avoid pitfalls as such. But before these conclusions can be reached, we must quantitatively prove $h_1$ successful.

Table 2 shows the percentages of the Community distribution over each cluster for each tool (tested separately). It is important to note the difference between cluster numbers in this table and in table 1. What was considered cluster 0 in the success experiments could be correlated to cluster 3 in one of the tool experiments. The numbering is arbitrary between experiments; the consistency should be seen by observing a one-to-one relationship between success clusters and tool clusters. This relationship does not necessarily have to be linear. Our predictions should have looked like figure 4 (which has been fabricated for this explanation). Instead, we got something like figure 5 for each tool experiment. This does not support $h_1$.  

7
Even though our clusters were not preserved, we could still look at the distribution the EM tool experiments gave. Disregarding the success cluster number we found, we looked inside the tool clusters to observe characteristics. We found that there is some relationship between our success factors and the tool usage. For example, we found that Communities in cluster 4 of the EM Bookmark clusters do not utilize the bookmark tool very often. Also cluster 1 and 3 have a medium usage of bookmarks. When we look at the total text and number of activities within this distribution, we see interesting patterns. We find that cluster 3 has a high number of activities, and it and cluster 1 have a high amount of text, while cluster 4 has a low amount of text. At first glance, it does not seem like that Communities that use bookmarks often and well also have a higher amount of text and activities. Could there be a relationship here between bookmark usage and success? We are unable to concretely conclude so at this point. Further future investigation will need to be performed in order to try and extract and quantify these emerging patterns.

4.3 Tag Experiments

In order to analyze the tags and their implied content, we first had to extract the tags used to describe each Community. First, we wrote a short program in Python to do this. Then, based on the total occurrences, we created an .arff file which Weka can read. After the success clusters had been determined, we wrote another script to automatically insert these into the .arff file with the tag information. At this point, the data is set up to be run through a supervised classification algorithm.

The first attempt was with JRip. Using 10-fold cross validation, JRip produced 3 rules: [if a tag is ”brasil”, assign it to cluster 2; if the tag is ”communitiesofpractice”, assign it to cluster 4; otherwise assign the Community to cluster 0]. Even without looking at the resulting statistics, one can speculate that these are not good rules. The numbers support this. JRip classifies Communities as cluster 0 in almost every instance, seeming to use no rules at all. However, the number of overall correctly classified communities is 49% (698 communities). This would seem good since it is above the baseline of 20% for a five labeled classification problem. It is not, however, because roughly half of the data already belongs to cluster 0 (center chart in table 6 for exact classification numbers—the boxes in gray indicate correct predictions.). Even after changing the number of folds and other parameters such as seeds and number of optimizations, we could not increase the number of rules JRip used. A new algorithm was needed.

Another rule based classification algorithm supported in Weka is a Decision Table. This did better than JRip, though not by much. Cluster 0 is still being over-classified (3rd chart in table 6). But this time we see that clusters 3 and 4 are getting slight attention, through not significant. There seemed to be no way to increase the number of rules produced. In this test, the Decision Table produced 8 rules, but they seem only to be the first few tags observed in the tag file alphabetically.

Yet another rule based learner is PARK. This algorithm created 6 rules on our data that seemed to consider more tags than the Decision Table produced (i.e. the considered tags spanned the entire alphabet). The breakdown is shown in the 1st chart in table 6. This analysis produces the least number of predictions in cluster 0 of the three rule based learners, yet the other clusters are still getting no significant attention.
Our Future Work Section will discuss possibilities for breaking down the tag matrix from a 1418 x 3776 matrix to a smaller one. But at this time and with all logical rule based algorithms and parameters exhausted on our dataset, we decided to stop further experiments on the tags so we can reconsider our approach.

5 Discussion

5.1 What Was Accomplished and Not: Future Challenges

Since this project works with a dataset that has not been looked at or annotated before and with a proposed method that has not been looked at in terms of these Communities, we are in many ways starting at square one. Overall, many of our initial ideas fell through and we could not support \( h_1 \) or \( h_2 \) with our current approach.

Too much time was spent attempting to refine the K-Means algorithm parameters and trying to get good visualizations of the distributions. When we realized that K-Means was not yielding 'good enough' results (the unbalanced cluster breakdown), we proceeded to use the EM algorithm. As it turns out, both algorithms produced the same cluster characteristics involving number of activities, total text, and rate, but there were smaller variances within each cluster using EM.

One reason why these K-Means experiments failed may be because we were looking at the dataset too broadly. As aforementioned, the clusters we found describe the overall characteristics and trends in the Communities, even if these trends are not very common (some clusters included a very small percent of the data, e.x. 1% or 4%). Finding these variety of trends was good, but for dealing with the largest cluster (74% of the dataset) we may have to look closer and temporarily ignore the other smaller clusters. Another reason our clustering was not good may have been our choice of success features assumed by \( a_0 \). There may be other potential choices that may turn out to be more influential, such as the ratio of contributions per day of each Community.

After this mishap and slight recovery by using EM for success, we looked at the Web 2.0 tools. The inconsistency between the success and tool cluster experiments leads us to believe we must redefine our \( a_0 \). We support this even more so because some patterns seem to be showing when we look inside the EM tool clusters; there is information there but we do not have the right techniques to mine it.

Finally, we looked at the Community tags. The first attempt using JRip failed. The algorithm was not able to deduce any meaningful rules for our dataset. Unfortunately, the other rule based learners also failed. This may be because the tag matrix is too sparse. We considered using a Singular Value Decomposition (SVD) and Principle Component Analysis (PCA) for comparison of influential tags, but perhaps the SVD may be able to help determine the most influential tags through its decomposition.
5.2 Self Evaluation

Throughout this project, I learned how to take a virtually new and unexamined problem and make it go somewhere. I defined my own success criteria and experiments, as well as interpreted the results with nothing to compare against as a baseline. Setbacks are almost inevitable, and I learned that after a certain point, things must be let go and something new must be tried. Instead of seeing the failed experiments as a failed project, I see it as figuring out how the dataset does not work and finding different approaches that may prove more fruitful. Our initial hypothesis may still be able to be supported through further research and refining of the dataset, algorithm parameters, and success features.

6 Acknowledgements

I acknowledge and thank Steve Whittaker for providing me with the IBM dataset and the project proposal.

I also acknowledge Marilyn Walker for helping me figure out what machine learning techniques and approaches to use on the dataset.

References


[20] Steve Whittaker. Should i use files or blogs to find a document? studying interrelationships among a groups social tools.
<table>
<thead>
<tr>
<th>cluster #</th>
<th>EM%</th>
<th>K-Means %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>49</td>
<td>74</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1: K-Means and EM success cluster breakdown

<table>
<thead>
<tr>
<th></th>
<th>Cluster 0</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>forums</td>
<td>10%</td>
<td>67%</td>
<td>11%</td>
<td>10%</td>
<td>2%</td>
</tr>
<tr>
<td>bookmarks</td>
<td>3%</td>
<td>24%</td>
<td>0%</td>
<td>71%</td>
<td>2%</td>
</tr>
<tr>
<td>files</td>
<td>1%</td>
<td>12%</td>
<td>51%</td>
<td>22%</td>
<td>15%</td>
</tr>
<tr>
<td>feeds</td>
<td>23%</td>
<td>51%</td>
<td>11%</td>
<td>14%</td>
<td>0%</td>
</tr>
<tr>
<td>blog</td>
<td>6%</td>
<td>6%</td>
<td>17%</td>
<td>1%</td>
<td>69%</td>
</tr>
<tr>
<td>wiki</td>
<td>8%</td>
<td>0%</td>
<td>1%</td>
<td>64%</td>
<td>27%</td>
</tr>
</tbody>
</table>

Table 2: Distribution of clusters from each EM Tool test

Figure 1: Success Characteristics for 1 Success Feature each K-Means Trials (clusters encapsulated in boxes for readability)
Figure 2: Success Characteristics for 2 Success Features in a K-Means trial

Figure 3: Success Characteristics for 3 Success Features for EM trial (clusters encapsulated in boxes for credibility)
Figure 4: Fabricated Success and Tool Cluster Relationship

Figure 5: Observed Success and Tool Cluster Relationship

Figure 6: Tag Classification (L to R: PART, JRip, Decision Table)