Project proposal (draft)

In this project, our aim is to analyze the attributes of online articles that influence their popularities. In specific, we summarize a heterogeneous set of features about articles published on Mashable in a period of two years. These features cover a large range of attributes, including content, keywords, publication time and so on. The goal then is to predict the popularity of that article reflecting in the number of shares in social networks. (Q1)

The project will be done in three main phases. First, we fetch data from the target website and create proper dataset for next steps. Second, we analyze data by selecting meaningful combination of metrics and calculating statistics (e.g correlation between attribute and popularity) of the given datasets. Besides analysis of existing data, our project will have a prediction module where we apply machine learning algorithms (linear regression, random forest) to predict the potential popularity of a given article and provide the optimal suggestion of rules that will make the article reach its best popularity. Finally, we will visualize the most significant findings using a set of tools and language(Q3)

We believe that this project, if properly conducted, would benefit a large number of people. First, for countless people who publish articles online every day, they would like to know those little tricks that make their article more popular. They may find, for example, that the number of shares tends to drop when the word count exceeds 2000 and then decide to write in a more concise way. Second, for editors from millions of websites all over the world, they can gain an insightful understanding of how to organize and release articles on internet. When they got first-hand materials from authors, they are responsible to find the best way to publish them. They may decide to add some pictures and videos in between lines to improve the reading experience. Moreover, the timing of publication can be a crucial factor as the popularity of different types of articles varies dramatically from different releasing time. Celebrity gossips, for example, usually spread faster during weekends. Our project will cover important features like those mentioned above and media and websites can benefit the most from our result. (Q4)

The achievement of this project can be measured in various ways. The most straightforward succeed may be reflected in the rise of visits and shares. Specifically, we apply rules indicated by analysis that can promote the popularities of online articles and compare the result with that of the same kind of articles without following the rules we suggest. As mentioned above, given considerable datasets and reasonable analysis, the result of our project will have a positive influence on the spreading of numerous online articles. Moreover, many other fields like advertising can more or less learn from this project and decide to make some adjustments of their marketing strategies, since the exposure of products is somehow similar to the popularity of articles in a sense that the more people get access to the content, the better.

There are, however a few risks we need to pay attention to. The first, and the most important one, which probably also the problem faced by all similar research groups, is the noise in data analysis. For example, articles from popular authors and famous journalists can easily get a huge number of shares even though some features of their article is against what we suggest. Another example of noise is the topic and content of an article (consider a poorly phrased and organized article which talks about a breaking news and has a fancy topic, people may easily choose to share it as well). These noises may jeopardize the accuracy, validity and availability of our project.

Leo Breiman proves that random forests are an effective tool in prediction since they avoid overfitting using the Law of Large Numbers. He further argues that random forests give results competitive with boosting and adaptive bagging, and their accuracy indicates that they act to reduce bias.

<https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>

Rutvija and Jayati Pandya briefly talks about how decision tree works, which divides a dataset into smaller subsets. Leaf node represents a decision. Based on feature values of instances, the decision trees classify the instances. Each node represents a feature in an instance in a decision tree which is to be classified, and each branch represents a value. The author also compares different classification algorithms including Bayesian classification, rule-based classification, K-nearest neighbor and so on. However, one of the main problems of using decision tree is overfitting, when the model mistakenly describes random error or noise instead of the underlying relationship

COMPARISION OF CLASSIFICATION ALGORITHMS

<http://research.ijcaonline.org/volume117/number16/pxc3903318.pdf>