**Data mining** is about *explaining* the past and *predicting* the future by analyzing the data. It is a multi-disciplinary field which combines statistics, artificial intelligence, machine learning and database technology. Data mining predicts the future by the means of modelling. Predictive modelling is a process where in which an outcome is predicted by creating a model. If the outcome is categorical, it is called *classification* and if the outcome is numerical, it is called *regression*. Classification is a data mining task of predicting the values of a categorical variable by building a model based on one or more categorical variables. Association rules find interesting associations among observations.

**The following are the groups of classification algorithms.**

Frequency table, covariance matrix, similarity functions and others.

K Nearest Neighbors algorithm falls under the *similarity* *functions* category.

Association rules find all sets of items that have a support greater than the minimum support and using large sets to get desired rules that have confidence greater than the minimum confidence.

**The following are the algorithms which I have used in the project.**

Bootstrap algorithm

K Nearest Neighbors algorithm

AIS algorithm for association rules

**Weakly supervised method or Comparator Mining:**

A sequential pattern is defined as a sequence S (S1S2S3…. Si…. Sn), where Si can be a word, symbol or a POS tag denoting a comparator (C$) or the beginning (#start) or the end of a question (#end). A sequential pattern is called an Indicative Extraction Pattern (IEP) if it can be used to identify comparative questions and extract comparators.

**Mining Indicative Extraction Patterns**

Weakly supervised method or comparator mining is based on the following assumptions

1. If a sequential pattern can be used to extract comparator pairs, then it is very likely to be an IEP.
2. The pair is capable to compare if a comparator pair can be extracted by an IEP

Based on the above assumption, boot strapping algorithm is designed. The two main keys steps involved are

1. Pattern generation
2. Pattern evaluation

**Pattern generation:**

The three kinds sequential patterns generated from sequences of questions are

**Lexical patterns:** These patterns indicate sequential patterns consisting of only words and symbols ($C, #start, and #end).

Example: Which is better, Omaha or Lincoln?

**Generalized patterns:** A lexical pattern is too specific for matching. Lexical patterns are generalized by replacing one or more words with their POS tags.

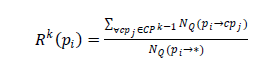
Example: Which city is better, Omaha or Lincoln?

**Specialized patterns:** Pattern specialization by adding POS tags to all comparator slots.

For example, from the lexical pattern “<$C or $C>” and the question “Omaha or Lincoln?”, “<$C=NN or $C=NN?>” will be produced as a specialized pattern.

Example: Omaha or Lincoln?

**Pattern Evaluation**



Reliability score for a candidate pattern Pi at iteration k is calculated as stated above.

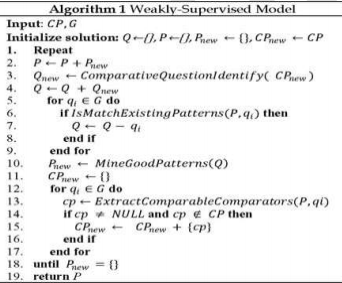
All the candidate patterns are evaluated and the pattern whose reliability score is greater than a threshold value is stored as IEP in the database.

**Comparator Extraction:**

Using IEP’s, comparator questions are easily identified and comparators are extracted from comparative questions. The following is the process for comparator extraction

1. Generate sequence for the comparator question
2. If the IEP is a pattern without generalization, then we need to tokenize the question.
3. If the IEP is a specialized pattern, then POS tags should follow the conditions specified by the pattern.

**Algorithm**



1. Bootstrapping starts with a single IEP
2. Extract initial seed comparator pairs from that single IEP
3. For each comparator pair all questions that contain the pair are considered as the comparative questions
4. From comparative questions, all the possible comparators are extracted.
5. Reliable patterns are added to the IEP repository.

**K Nearest Neighbors algorithm**

It is one of the top 10 data mining algorithms

Nearest Neighbor algorithm is a useful data mining technique to use the past data instances, with known output values, to predict an unknown output value. It is a simple algorithm where in which the training data set is stored, so that a classification for a new unclassified record may be found simply by comparing it to most similar records in the training set. Commonly used distance metric is Euclidean distance.

I have made use of this algorithm to find the comparators which are close to one another, this helps in analyzing the outcome in a more meaningful way to understand the relation between the nearest comparators.

**AIS algorithm for association rules**

Association rules find interesting associations and correlations amongst a larger set of data items.

Rule = X --> Y

Support = frq(X,Y)/N

Confidence = frq(X,Y)/frq(X)

The support of each individual item is counted in the first pass, in each subsequent pass, the large item sets determined in the previous pass is used to generate new item sets, and these are called candidate item sets. The support of each candidate item set is counted and the large ones are determined.

Candidate item sets are generated and counted on the fly as the database is scanned. The counts of the corresponding entries are increased if they were created by an earlier transaction.

Nearest neighbor algorithm and Association rules for mining add value to the comparators extracted from the bootstrapping algorithm in making a meaningful analysis.

**Note:** The implementation part I have already sent you couple of weeks ago, I have used the data from yahoo answers site called “Yahoo Research Alliance Web scope program”. This is a huge file, here I have selected 6000 questions from it.