CMPT741 Course Project Report

YIJI WANG
301286922
Course-based

Introduction

In this project, DocumentWords.txt is given for essays with id from 4 mini-assignments after data preprocessing with words represented by different integers. 5 or more than 5 words are removed from each essay. The goal is to predict the 5 removed words with highest possibility.

Problem Definition

The nature of this project is to tell which words are most likely to appear with the existence of some words, which is similar to the recommendation system.

For each word ***w*** to be predicted has following features:

1. ***w*** has never appeared in one essay, but has to appeared i the whole DocumentWords.txt
2. ***w*** will only be predicted once for one essay (appear once in the 5 predicted words)
3. ***w*** will be a word with at least frequency 1 in the document
4. each ***w*** recommended in one essay will have a confidence by which the recommendation list will be sorted

Approach

Here *Predictive Association Rule Mining Method* is used to achieve the goal. We can predict the removed words by finding the rules like "If A, B and C, then D"

Data Preprocessing

The preprocessing of the data can be divided into following steps.

1. Redundancy removal-remove all the words with frequency 1 in the essay.
2. Data formatting-prepare the data for different use (e.g. association rule mining, clustering, matching)

Design

The design of approach and experiments can be divided into following operations.

1. Predict by frequency only
2. Predict by association rules with no sorting
3. Predict by association rules sorted with priority support>confidence>lift
4. Predict by association rules sorted with priority confidence>support>lift
5. Predict by association rules sorted with priority lift>support>confidence
6. Predict after clustering the essays into 4 groups with association rules sorted with priority support>confidence>lift
7. Predict after clustering the essays into 4 groups with association rules sorted with priority confidence>support>lift
8. Predict after clustering the essays into 4 groups with association rules sorted with priority lift>support>confidence

Experiment

The results returned can be shown in the tables below(the experiment number corresponds to the above operations in design part)

Experiment	#1	#2	#3	#4	#5	#6	#7	#8
average MAP@5	19.21%	21.46%	30.19%	29.17%	17.23%	34.37%	31.29%	16.82%

Association Rules VS NO-Association Rules

Experiment	With Association Rules	No Association Rules
average MAP@5	25.79%	19.21%
Max MAP@5	34.37%	19.21%

Clustering VS NO-Clustering

Experiment	With Clustering	Without Clustering
average MAP@5	24.55%	24.92%
Max MAP@5	34.37%	30.19%

Priority of support, confidence and lift of Association Rules

Experiment	Support First	Confidence First	Lift First
average MAP@5	32.28%	30.23%	17.03%
Max MAP@5	34.37%	31.29%	17.23%

Here the clustered dataset is 95.2% accuracy testing with CMPT459 Validation test samples

Conclusion

After several experiments with different designs, the design that can achieve the highest MAP@5 score so far is: Cluster the essays then use association rules sorted by support>confidence>lift. And for the current provided 20% test dataset, the method could achieve 34.37% accuracy.

It is interesting to find that the dominance of an association rule is different from the usual cases, which is lift>confidence>support. It is because in the recommendation system, the support plays a more important in the data we recommend. Association rules and items will be more featured with higher support.

Future Improvement

To further improve the accuracy of the project result, we can use boosting method to boost the association rule mining procedure. Because if seen as a classifier, association rule classifier in this project is really a weka one. And considering the order of the recommended words will matter al lot, using boosting can further improve the accuracy by evaluating the order of the words. Clustering first, association rule mining combined with boosting for the recommended words will be greatly improve the result.
Because of time limit, I cannot merge my new idea into this project. (Actually I came up this idea on the last day of submission…) But related work will continue when I have time. I believe the result could be raised up to 50%.

Reference

1. A Recommendation Algorithms, https://docs.oracle.com/cd/B14099_19/bi.1012/b14052/appendix.htm
2. Jian Pei, Jiawei Han, Wenmin Li, CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules.
3. W. Li. Classification based on multiple association rules.