A brief explanation about an application idea of association rule

Name:
Gengcong Yan, 1009903

Article Name:
A Text Mining Technique Using Association Rules Extraction

Link:
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.212.8624&rep=rep1&type=pdf

Reference:
Mahgoub, Hany, et al. "A text mining technique using association rules extraction." International journal of computational intelligence 4.1 (2008): 21-28.

Answer 1:
Base on association rule extraction, researchers developed a text mining system, generating association rules from raw data with a shorter time compared to the Apriori-based system. The proposed system consists of complete workflows like Text Preprocessing Phase, Association Rule Mining Phase and Visualization Phase. Data used in the paper are 100 webpages of news that are related to the outbreak of bird flu disease The collection contained 30000 single words. Each document contained on average 300 single words.

Answer 2:
The single words in documents are seen as sets in the system. The criteria for association patterns are support, confidence rule and their weight values TF-IDF.

Answer 3:
The workflow of the system simply introduces as follow. The raw data obtained from webpages include a large number of repetitive and meaningless words and symbols, so our first step is the text pre-processing phase, which removes a large amount of redundant and useless information and allows each document to form a compact collection of keywords. The algorithm ignores the order of the keywords in the document and focuses on their frequency distribution. Then with the TF-IDF weighting scheme, we assign a score to all keywords based on occurrence in documents. And the top N keywords are selected as the final set for the next phase. In the association rule mining phase, the system

applies Generating Association Rules based on Weighting scheme algorithm based on minimum support and confidence to obtain expected association rules. Last, the system converted all rules easy to read and interpret in the visualization phase.

The strength of the system is a reduction in execution time. Because it generates all frequent keyword sets from filtered data that satisfy the threshold weight value, whereas Apriori makes repeated scanning on the original documents. The potential risk is the ability to handle a large dataset containing millions of documents. In this case, calculating the weight information of all keywords inside all documents will become difficult and consume a lot of computing power and time.

Answer 4:

A similar approach can be applied in social network analysis. Through the information network associated between users, we can extract association rules through the public profile and tweets of users, then present and analyze the similarities between users and popular trends relationship in social networks.