Deliverable 1: Final Year Dissertation

Sahil Pattni

BSc. Computer Science (Honours)

Supervisor: Neamat El Gayar

December 2, 2020

Declaration

I, Sahil Manojkumar Pattni, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Sahil Manojkumar Pattni

Date: December 2, 2020

Abstract

In this digital age, data is being generated at an exponential rate, with data analytics being used by corporations and small businesses alike to produce insights, reduce costs, optimize operations and increase profits. Generating association rules allow us to find non-intuitive associations that bring new insights to the management, allowing them to leverage this information to maximize their profits. A prime example would be the *'Beers and Diapers'* case **beers˙diapers**, where a company looked at their point-of-sale data and found a strong association between beers and diapers being co-purchased, which seems rather unintuitive.

In this study, we will extract a minimum spanning tree (MST) from a data set using machine learning techniques. We will then use this minimum spanning tree to segment products together, and produce association rules and extract the most interesting ones. The resulting ruleset will be compared to rules generated by the established Apriori Algorithm **apriori**, which is the foundation of association rule mining. Additionally, we will study how the grouping of the MST compares to clustering algorithms, and how the structure of our MST would change before and after promotional events, leading to insights that may help the management in such firms make better informed decisions about the type of promotions they would like to run.

Contents

1 Introduction

1.1 Context

For business which deal with the sale of a heterogeneous physical assets - such as groceries, hypermarkets and select retail outlets - operations such as inventory management and product placement play an instrumental role in determining the business' financial success. These involve asking questions such as:

- Which products should be placed at the entrance of the store? Which should be placed closer to the exit?
- Which products will benefit the most by being placed at eye-level?
- Which products should be placed next to each other to maximize the purchase volume?

One way to find optimal solutions for such questions is to employ the use of Association Rule Mining (also known as Market Basket Analysis). This set of techniques assess frequent itemsets (e.g. from sales data) and generate association rules between products. Several algorithms and techniques exist for association rule mining, such as the Apriori Algorithm and FP-Growth **fp˙growth**. One problem with these algorithms is that they tend to generate an enormous amount of rules, of which many of the rules themselves are large This makes it grossly inconvenient for the end-user to retrieve any actionable information from the results.
What we will do is construct a network of products - where each vertice represents a product or product category, and an edge between two vertices represents the simultaneous occurrence of the the two products. The result is a graph (i.e. a network), whose architecture is visually informative of the products and their associations. The insights that can be gained from the MST can help the management of the mentioned businesses to maximize their profit.

1.2 Aims

The aim of this study is to study the effectiveness of a minimum spanning tree as a market basket analysis tool - from product clustering to association rule mining. Additionally, we will determine how the architecture of the MST will change during and after large events such as promotions. The model will be tested on a relatively large dataset of sales data.

1.3 Objectives

The research objectives for this project have been laid out below, in the order that they will be carried out.

1. Acquire a suitable dataset upon which the MST can be constructed.

2. Explore and evaluate MST extraction algorithms.
   (While Prim's **primsprims˙og** and Kruskal's **kruskal** algorithms are the most commonly used methods to extract the MST from a graph, research has been conducted on more efficient ways to extract the MST using machine learning techniques such as Artificial Neural Networks and K-Nearest-Neighbors, and these will be discussed in the literature review [SECTION POINTER ONCE CREATED]).

3. Generate an undirected graph $G(E, V)$ where the edges E are the correlation values between the product vertices.

4. Extract a minimum spanning tree from this graph using the technique determined in step 2.

5. Analyze this MST and use its architecture to determine product clusters and generate association rules.

6. Generate association rules using the Apriori and FP-Growth algorithm and compare them in both their time complexity and their *'interestingness'* (i.e. the unintuitive rules they generate).

7. Validate the product grouping present in the MST against a clustering algorithm (e.g. such as the K-Means algorithm).

1.4 Core Concepts

This section details the core concepts this dissertation is based upon.

1.4.1 Graphs

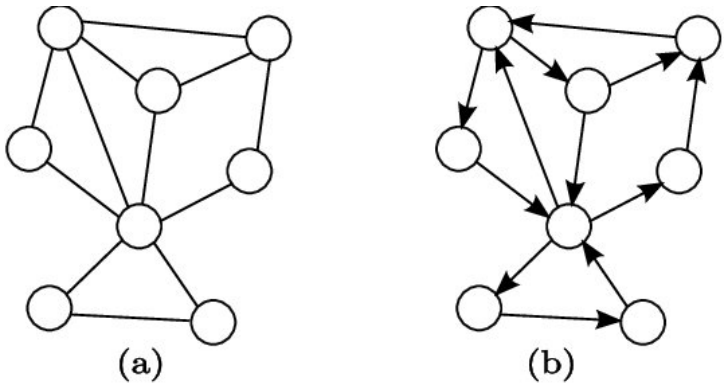

Figure 1: Undirected and Directed Graphs

In discrete mathematics and more specifically - graph theory, a graph is a data structure that contains a set of nodes (i.e. vertices) connected by lines (i.e. edges). These edges may be directed - such as in Figure 1:(a), or undirected - such as in Figure 1:(b). A graph G with a set of vertices V and a set of edges E can be represented via the notation $G = (V, E)$. For the purposes of our research, we will be focusing on undirected graphs, where the weight between two vertices is the same in both directions.

1.4.2 Minimum Spanning Trees

Given an undirected $G = (E, V)$, a *spanning tree* can be described as a subgraph that is a tree which includes all the vertices V of G with the minimum number of edges required. A *minimum spanning tree* (MST) is the spanning tree with the smallest sum of edge weights. This means that if the graph has n vertices, each spanning tree - including the minimum spanning tree - will have $n - 1$ edges.