# **Interesting Sequences in Online Retail Transactions**

Karri Srikanth

2016AAPS0234H

*Abstract*—Data Mining is a process to extract information for developing significant relationship with variables stored in large data warehouses. Online Retail has become a newest trend in the recent time. People preferring online shopping these days because of its flexibility, time saving and variety of options. The companies who are providing platforms for such online retail shopping has also seen a boom. As a result of this, companies has got a lots of data because of large number of people are using these platforms. The companies has levelled up their business using data to serve customers well. The data that has been stored from years of their customers previously helping them to ease their operations well and make the shopping more intuitive for customers. For example, companies uses the data and find the requirement of different products in different regions so that they can store them in respective regions which will help them to deliver to customers faster than keeping all the products in one store. Also, the products which are not in demand are stored uselessly. It also helps the companies to serve customers by recommending products when customer trying to buy one based on the patterns of several customers who shopped before. The objective of this research is to find patterns or sequences in the online retail data of firm using Generalised sequential Pattern Mining(GSP) algorithm. This algorithm helps to find interesting sequences in the online retail dataset. The sequences are the orders where most of the customer bought in such a way.

Keywords—Data Mining, Online Retail Data Mining, Retail Data, Patterns, Sequences, Sequential Mining.

I. Introduction

Data mining is a process to discover knowledge from data which the stored in data base and data warehouses responsibility. Sequential pattern mining is a topic of [data mining](https://en.wikipedia.org/wiki/Data_mining) concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus [time series](https://en.wikipedia.org/wiki/Time_series) mining is closely related, but usually considered a different activity. Sequential pattern mining is a special case of [structured data mining](https://en.wikipedia.org/wiki/Structured_data_mining). There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, sequence mining problems can be classified as string mining which is typically based on [string processing algorithms](https://en.wikipedia.org/wiki/String_(computer_science)) and itemset mining which is typically based on [association rule learning](https://en.wikipedia.org/wiki/Association_rule_learning). Local process models  extend sequential pattern mining to more complex patterns that can include (exclusive) choices, loops, and concurrency constructs in addition to the sequential ordering construct. The data we have is online retail data which will help us to find the interesting sequences of the transactions made by customers. The data has wide variety of transactions including various type of items. This will help to understand the GSP algorithm clearer and helps to find interesting sequences.

1. Online Retail

Online shopping is a form of [electronic commerce](https://en.wikipedia.org/wiki/Electronic_commerce) which allows consumers to directly buy [goods](https://en.wikipedia.org/wiki/Good_(economics)) or [services](https://en.wikipedia.org/wiki/Service_(economics)) from a seller over the [Internet](https://en.wikipedia.org/wiki/Internet) using a [web browser](https://en.wikipedia.org/wiki/Web_browser). Consumers find a product of interest by visiting the [website](https://en.wikipedia.org/wiki/Website) of the retailer directly or by searching among alternative vendors using a [shopping search engine](https://en.wikipedia.org/wiki/Shopping_search_engine), which displays the same product's availability and pricing at different e-retailers. As of 2020, customers can shop online using a range of different computers and devices, including [desktop computers](https://en.wikipedia.org/wiki/Desktop_computer), [laptops](https://en.wikipedia.org/wiki/Laptop), [tablet computers](https://en.wikipedia.org/wiki/Tablet_computer), [smartphones](https://en.wikipedia.org/wiki/Smartphone), and [smart speakers](https://en.wikipedia.org/wiki/Smart_speaker).An online shop evokes the physical analogy of buying [products](https://en.wikipedia.org/wiki/Product_(business)) or services at a regular ["bricks-and-mortar"](https://en.wikipedia.org/wiki/Brick_and_mortar_business) [retailer](https://en.wikipedia.org/wiki/Retailing) or [shopping center](https://en.wikipedia.org/wiki/Shopping_center); the process is called business-to-consumer (B2C) online shopping. When an online store is set up to enable businesses to buy from another businesses, the process is called [business-to-business](https://en.wikipedia.org/wiki/Business-to-business) (B2B) online shopping. A typical online store enables the customer to browse the firm's range of products and services, view photos or images of the products, along with information about the product specifications, features and prices.Online stores usually enable shoppers to use "search" features to find specific models, brands or items. Online customers must have access to the Internet and a valid [method of payment](https://en.wikipedia.org/wiki/Online_shopping#Payment) in order to complete a transaction, such as a [credit card](https://en.wikipedia.org/wiki/Credit_card), an [Interac](https://en.wikipedia.org/wiki/Interac)-enabled [debit card](https://en.wikipedia.org/wiki/Debit_card), or a service such as [PayPal](https://en.wikipedia.org/wiki/PayPal). For physical products (e.g., paperback books or clothes), the e-tailer ships the products to the customer; for digital products, such as [digital audio files](https://en.wikipedia.org/wiki/Digital_audio_file) of [songs](https://en.wikipedia.org/wiki/Song) or [software](https://en.wikipedia.org/wiki/Software), the e-tailer usually sends the file to the customer over the Internet. The largest of these online retailing corporations are [Alibaba](https://en.wikipedia.org/wiki/Alibaba_Group), [Amazon.com](https://en.wikipedia.org/wiki/Amazon.com), and [eBay](https://en.wikipedia.org/wiki/EBay).

***B) Data***

The data is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. This data was taken from UCI. The data contains 8 attributes and 4.2 lakhs data points. The attributes are InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation. StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product. Description: Product (item) name. Nominal.  
Quantity: The quantities of each product (item) per transaction. Numeric. InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.  
UnitPrice: Unit price. Numeric, Product price per unit in sterling. CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.  
Country: Country name. Nominal, the name of the country where each customer resides. The dataset contains wide range of attribute type and good number of data points to work with.

C) Preprocessing

Data pre-processing is crucial in any data mining process as they directly impact success rate of the project. This reduces complexity of the data under analysis as data in real world is unclean. Data is said to be unclean if it is missing attribute, attribute values, contain noise or outliers and duplicate or wrong data. Presence of any of these will degrade quality of the results.

Here are few important data pre-processing techniques that can be performed are handling missing data, dimensionality reduction, feature selection, feature creation, binarization.

***D) Visualisation***

Without the concept of visualization, mining and analysis doesn’t play any role of importance as data mining is the idea of finding inferences by analysing the data through patterns and those patterns can only be represented by different visualization techniques.

## **Uses of data visualization**

* Powerful way to explore our data with presentable results.
* Primary use is the pre-processing portion of the data mining process.
* Supports in data cleaning process by finding incorrect and missing values.
* For variable derivation and selection means to determine which variable to include and discarded in the analysis.
* Also play role in combining categories as part of the data reduction process.

## Techniques:

* Box plots
* Histograms
* Charts

## **II.Data Preprocessing**

The preprocessing techniques used here are :

1. ***Data Cleaning***

The **missing values** in each numerical attributes is replaced by **mean** of those columns. The categorial attribute is replaced by **mode**. Since the number of missing values in each column is not so significant , removing them seemed to be best option as the number is less and it won’t effect predictions anyway. This helped to increase data points so that it can be helpful to make better predictions.

1. ***Data Reduction***

The attributes which convey same meaning will not help for any data mining techniques so removing such features help in **dimensionality reduction** thus contributing towards feasible processing of data for pattern and sequence recognition. Attributes which do not contribute any significant information to our data were removed (including attributes Stock code as description and stock code convey same thing, quantity and unit price are redundant as we are interested in find the patterns in items not prices or quanity).

1. ***Data Transformation***

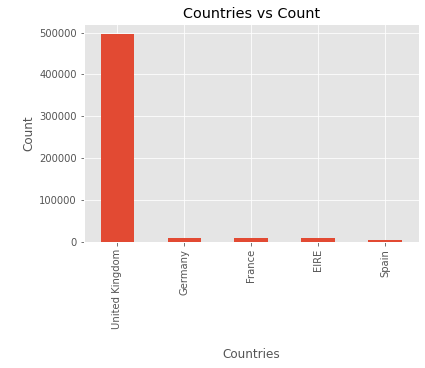
A new attribute was introduced **(feature creation)** which consists of the aggregate of multiple columns thus giving us an overview of data in each row. **Feature selection** was done thus redundant features were removed. The description having same Invoice ID are clubbed together as one transaction and sorted based on time to form sequential data

## **III.Data Visualisation**

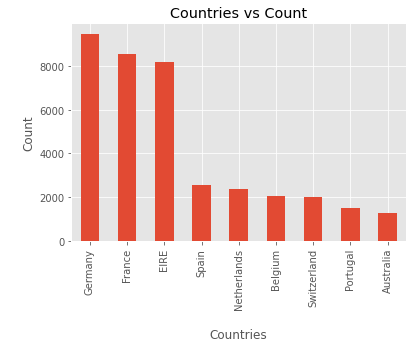
## 

Data was visualised using matplotlib, seaborn, apache superset, where pie charts, bar graphs, heat maps, plot box, pair plot, scatter matrix and word cloud were plotted between various attributes. Snippets of some of those graphs are shown below:

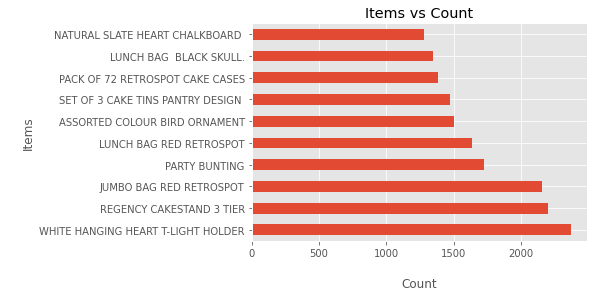
* **Top 5 countries in no. of Transactions**



* **Top 10 countries where higher transactions were made excluding UK**



#### **Top 10 products sold**



***IV. Techniques for obtaining results***

There are very good number of techniques in data mining to obtain the required results depending on the nature and characteristics of data. The technique used here is Generalised Sequential Pattern Mining (GSP).

**A. *Associate Rule Mining***

GSP algorithm (*Generalized Sequential Pattern* algorithm) is an [algorithm](https://en.wikipedia.org/wiki/Algorithm) used for [sequence mining](https://en.wikipedia.org/wiki/Sequence_mining). The algorithms for solving sequence mining problems are mostly based on the [*apriori*](https://en.wikipedia.org/wiki/Apriori_algorithm) (level-wise) algorithm. One way to use the level-wise paradigm is to first discover all the frequent items in a level-wise fashion. It simply means counting the occurrences of all singleton elements in the database. Then, the [transactions](https://en.wikipedia.org/wiki/Transaction_(database)) are filtered by removing the non-frequent items. At the end of this step, each transaction consists of only the frequent elements it originally contained. This modified database becomes an input to the GSP algorithm. This process requires one pass over the whole [database](https://en.wikipedia.org/wiki/Database).

GSP algorithm makes multiple database passes. In the first pass, all single items (1-sequences) are counted. From the frequent items, a set of candidate 2-sequences are formed, and another pass is made to identify their frequency. The frequent 2-sequences are used to generate the candidate 3-sequences, and this process is repeated until no more frequent sequences are found. There are two main steps in the algorithm.

* Candidate Generation. Given the set of frequent (k-1)-frequent sequences Fk-1, the candidates for the next pass are generated by joining F(k-1) with itself. A pruning phase eliminates any sequence, at least one of whose subsequences is not frequent.
* Support Counting. Normally, a [hash tree](https://en.wikipedia.org/wiki/Hash_tree_(persistent_data_structure))–based search is employed for efficient support counting. Finally non-maximal frequent sequences are removed.

**Algorithm:**

F1 = the set of frequent 1-sequence

k=2,

do while Fk-1 != Null;

Generate candidate sets Ck (set of candidate k-sequences);

For all input sequences s in the database D

do

Increment count of all a in Ck if s supports a

End do

Fk = {a ∈ Ck such that its frequency exceeds the threshold}

k = k+1;

End do

Result = Set of all frequent sequences is the union of all Fk's

The above algorithm looks like the [Apriori algorithm](https://en.wikipedia.org/wiki/Apriori_algorithm). One main difference is however the generation of candidate sets. Let us assume that: A → B and A → C are two frequent 2-sequences. The items involved in these sequences are (A, B) and (A,C) respectively. The candidate generation in a usual Apriori style would give (A, B, C) as a 3-itemset, but in the present context we get the following 3-sequences as a result of joining the above 2- sequences A → B → C, A → C → B and A → BC. The candidate–generation phase takes this into account. The GSP algorithm discovers frequent sequences, allowing for time constraints such as maximum gap and minimum gap among the sequence elements. Moreover, it supports the notion of a sliding window, i.e., of a time interval within which items are observed as belonging to the same event, even if they originate from different events. The run-time was high.

Applying the same on the data the results were good number. The following are few of each length:

1-length sequences:

* ['CREAM CUPID HEARTS COAT HANGER'],
* ['KNITTED UNION FLAG HOT WATER BOTTLE'],
* ['RED WOOLLY HOTTIE WHITE HEART.'],
* ['SET 7 BABUSHKA NESTING BOXES'],
* ['WHITE HANGING HEART T-LIGHT HOLDER'],
* ['WHITE METAL LANTERN'],
* ['HAND WARMER UNION JACK'],
* ['ASSORTED COLOUR BIRD ORNAMENT'],
* ['BOX OF VINTAGE ALPHABET BLOCKS']

etc

2-length sequences:

* ['WHITE HANGING HEART T-LIGHT HOLDER', 'ASSORTED COLOUR BIRD ORNAMENT'],
* ['WHITE HANGING HEART T-LIGHT HOLDER', 'WOODEN FRAME ANTIQUE WHITE '],
* ['WHITE HANGING HEART T-LIGHT HOLDER', 'WOODEN PICTURE FRAME WHITE FINISH'],
* ['WHITE HANGING HEART T-LIGHT HOLDER', 'RED HANGING HEART T-LIGHT HOLDER'],
* ['WHITE HANGING HEART T-LIGHT HOLDER', 'JUMBO STORAGE BAG SUKI'],
* ['WHITE HANGING HEART T-LIGHT HOLDER', 'LUNCH BAG RED RETROSPOT'],

Etc..

3-length sequences:

* ['ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED '],
* ['JUMBO  BAG BAROQUE BLACK WHITE', 'JUMBO BAG PINK POLKADOT', 'JUMBO BAG RED RETROSPOT'],
* ['JUMBO  BAG BAROQUE BLACK WHITE', 'JUMBO STORAGE BAG SUKI', 'JUMBO BAG RED RETROSPOT'],
* ['JUMBO  BAG BAROQUE BLACK WHITE', 'JUMBO SHOPPER VINTAGE RED PAISLEY', 'JUMBO BAG RED RETROSPOT'],
* ['JUMBO BAG PINK POLKADOT', 'JUMBO BAG PINK VINTAGE PAISLEY', 'JUMBO BAG RED RETROSPOT'],
* ['JUMBO BAG PINK POLKADOT', 'JUMBO STORAGE BAG SUKI', 'JUMBO SHOPPER VINTAGE RED PAISLEY']

Etc..

4-length sequences:

* ['ROSES REGENCY TEACUP AND SAUCER ', 'GREEN REGENCY TEACUP AND SAUCER', 'PINK REGENCY TEACUP AND SAUCER'],
* ['CHARLOTTE BAG SUKI DESIGN', 'CHARLOTTE BAG PINK POLKADOT', 'RED RETROSPOT CHARLOTTE BAG'],
* ['CHARLOTTE BAG SUKI DESIGN', 'CHARLOTTE BAG PINK POLKADOT', 'WOODLAND CHARLOTTE BAG'],
* ['CHARLOTTE BAG SUKI DESIGN', 'RED RETROSPOT CHARLOTTE BAG', 'WOODLAND CHARLOTTE BAG'],
* ['CHARLOTTE BAG PINK POLKADOT', 'RED RETROSPOT CHARLOTTE BAG', 'WOODLAND CHARLOTTE BAG']

Etc..

***IV. Conclusion***

The preprocessing of the data is done based on various techniques to make it more meaningful and useful. The techniques helped the data to become clean and more expressive.

It also helps in making better models and data mining techniques to give meaningful. Visualization is very useful , it helps in giving a vague idea of whole data. The various visualization tools helps in presenting data more pleasant. The sequences were clean and good. Since each sequence appearance in transactions is much lesser compared to data size, the support taken is 0.01. The maximum length of interesting sequences is 4. There are large number of sequences found and can be found if support is decreased. The results were satisfying at that given support.

***V. Acknowledgements***

I would to like to express my appreciation and gratitude for the lectures taught by Prof. Manik Gupta. The lessons were quite informative and helped in completing this assignment. Also we would like to thank all the Teaching Assistants for helping all the time throughout the assignment.

***V.References***

[1] The data is taken from [https://data.gov.in](https://data.gov.in/)

[2] Techniques taught in class.