# *Association Analysis of Wikidata Properties*

Abimanyu Yuda Dewa
*Faculty of Computer Science*
*Universitas Indonesia*
Depok, Indonesia
abimanyu.yuda@ui.ac.id

Muhammad Zahran Agung Dewantoro
*Faculty of Computer Science*
*Universitas Indonesia*
Depok, Indonesia
muhammad.zahran@ui.ac.id

Abraham Rudolf Brahmana
*Faculty of Computer Science*
*Universitas Indonesia*
Depok, Indonesia
abraham.rudolf@ui.ac.id

Seto Adhi Prasetyo
*Faculty of Computer Science*
*Universitas Indonesia*
Depok, Indonesia
seto.adhi@ui.ac.id

Al Taaj Kautsar Supangkat
*Faculty of Computer Science*
*Universitas Indonesia*
Depok, Indonesia
al.taaj@ui.ac.id

***Abstract*—Wikidata is a large-scale knowledge graph describing over 90 million real-world entities. The entity description relies on properties, linking entities to other entities as well as literals. This gives rise to the question: Is there any pattern in how properties are used to describe entities in Wikidata? This paper presents association analysis over Wikidata properties, giving hints as to which and how properties often occur together in Wikidata. What we aim to achieve is to find properties that are common between entities of different types in Wikidata using association rule mining algorithms. For instance, every entity of type human has a date of birth property or every song entity has a writer property. If entities of the same type have something in common, perhaps the same case could apply to different types of entities. First of all, all of the related definitions and association rule formulas are introduced. Secondly, the data and algorithm we used in this paper are discussed. Finally, the results of our experiment are finalized and summarized.**

***Keywords*—Association Analysis, Association Rules, Frequent Itemsets, Properties, Wikidata**

## I. Introduction

(Rajak, 2008) mentioned that association analysis has been used in many applications, such as biological databases, market basket analysis of library circulation data, to study protein composition, to study population and economic census, etc. Association analysis is used to identify patterns that occur in the data. The pattern shows combinations of items that often appear together. Such a pattern can be helpful in market basket analysis where it is used to identify items that are often bought together in transactions.

There are two types of association that can be mined from data using machine learning, Frequent Itemsets and Association Rules. An item set is a group of items. If any itemset has k-items it is called a k-itemset. An itemset is a set that has items as its element. It can be empty or filled with one or more elements. An itemset that occurs frequently is called a frequent itemset. A set of items is called frequent if it satisfies a minimum threshold value for support and confidence. (Labade & Kini, 2013) explained that frequent pattern mining is the method of mining data in a set of items or some patterns from big databases, of which support is equal to or greater than the minimum threshold. Support shows how popular an itemset is in the transaction database. Confidence shows how likely an item is going to be purchased when another is purchased. It is a collection of items that frequently occur together that are usually used on marketing systems.

An association rule is an if/then statement that helps uncover relationships between seemingly unrelated data. It is usually used to find relationships between attributes in large databases. Association rule mining consists of 2 steps: find all the frequent itemsets and generate association rules from the above frequent itemsets.

Based on the concept of strong rules, (Agrawal et al., 1993) introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale systems in supermarkets. It uses association rule learning as a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

The goal of this paper is to find interesting associations between properties of Wikidata entities (Vrandečić & Krötzsch, 2014). We use the FP-growth algorithm to analyze the properties of entities in Wikidata, we don't use apriori because it is slower on large datasets, and most of all, our computing power is not really high, so despite its limitations, FP-growth still gives result, while apriori ends in a memory error. The data we used in this paper is taken from Wikidata.

The rest of the paper is structured as follows. Section II provides the preliminaries of our paper. Section III describes the methodology. In Section IV, we report our analysis results. Finally, Section V concludes our paper.

## II. Preliminaries

### *A. FP-Growth Algorithm*

The FP-growth algorithm is the most widespread algorithm for frequent itemset mining according to (Labade & Kini, 2013). FP-growth algorithm aims to

remove the bottlenecks of the apriori algorithm. According to (Borgelt, 2010), It is based on a prefix tree representation of the given database of transactions (called an FP-tree), which can save considerable amounts of memory for storing the transactions, so all transactions are stored in the tree-like data structure.

Advantages Of FP-growth algorithm:

1. This algorithm needs to scan the database only twice when compared to apriori which scans the transactions for each iteration.
2. The pairing of items is not done in this algorithm and this makes it faster.
3. The database is stored in a compact version in memory.
4. It is efficient and scalable for mining both long and short frequent patterns.

Disadvantages Of FP-growth algorithm

1. FP-tree is more cumbersome and difficult to build than apriori.
2. It may be expensive.
3. When the database is large, the algorithm may not fit in the shared memory.

### B. *Wikidata*

Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, the other Wikimedia projects, and well beyond that which is stated by (Vrandečić, 2013). Wikidata allows humans and machines to read and edit the data. To store structured data beyond text labels and language links, Wikidata uses a simple data model (Vrandečić & Krötzsch, 2014). Data is basically described through property-value pairs.

### C. *Support*

(Alfiqra, 2018) said that support is an indication that shows how often an itemset appears in the database. Suppose we have X and Y which are items, then we calculate the support of item X and item Y with this formula.

$$Support\{X,Y\} = P(X \cap Y) = \frac{\sum transactions\ that\ contain\ X\ and\ Y}{total\ number\ of\ transactions}$$

### D. *Confidence*

According to (Zhan, Zhu, Zhang, & Wang, 2019), given a transaction database D and items X and Y, confidence is the percentage of Y in the case that the transaction in D already contains X, that is, the conditional probability.

$$Confidence(X \rightarrow Y) = P(Y/X) = \frac{PP(X\ X \cap Y\ Y)}{\sum transactions\ that\ contain\ }$$

### E. *Lift*

(Shankar & Bargadiya, 2013) mentioned that Lift is nothing but the ratio of confidence to expected confidence. Lift is a value that gives us information about the increase in the probability of the "then" (consequent) given the "if" (antecedent) part. Assume we have X and Y which are items, at that point, we figure the lift of item X and item Y with this equation.

$$Lift(X \rightarrow Y) = \frac{PConfidence(X \rightarrow Y)}{PSupport\{Y\}}$$

## III. METHODOLOGY

This section describes the methodology of the research: problem statement, dataset, experiment setup

### A. *Problem Statement*

We want to find frequent patterns of properties from entities of certain categories. We do it by using association analysis.

For instance, we have the following data and we want to know the frequent patterns with minimum support of 0.6.

| No | Items |
|---|---|
| 1 | Apple, Bread, Milk |
| 2 | Melon, Milk, Cereal, Coke |
| 3 | Apple, Melon, Cereal |

When we use FP Growth, the resulting frequent patterns are in the following table

| | support | itemsets |
|---|---|---|
| 0 | 0.666667 | (milk) |
| 1 | 0.666667 | (apple) |
| 2 | 0.666667 | (melon) |
| 3 | 0.666667 | (cereal) |
| 4 | 0.666667 | (cereal, melon) |

We are going to use this method on the properties of Wikidata entities.

### B. *Dataset*

The dataset in this paper is acquired from Wikidata using its query service. It consists of entities along with their direct properties. We have queried 10,000 data of five categories: humans, cities, books, songs, and films, so in total, we are using 50,000 data. The reason we chose those 5 categories is that we are aiming to get a balanced amount of data between them and all of them have a wide range of possible entities and properties.

### C. *Experiment Setup*

The first step that we did was preprocessing the queried data. We downloaded the JSON format of the dataset from https://query.wikidata.org/ and converted it to NTriples format. Next, we store the aforementioned NTriples

formatted data in an Apache Jena Fuseki server to make querying faster, as it is stored in a local server.

The second step was using the data on the algorithms from the Python mlxtend library for association rule mining. After that, the result given would be used to gain insights in association between the properties.

Wikidata Query Service  Examples  Help

```
SELECT ?s ?p ?o
WHERE {
    ?s wdt:P31 wd:Q5.
    ?s ?p ?o .
    FILTER(CONTAINS(STR(?o),"wikidata.org/entity"))
}
limit 10000
```

Figure 3.1

Figure 3.1 shows the example query command that we use on the Wikidata query service to get the dataset. The s is subject, p is predicate, and o is object. The wdt queries the entities with direct properties "instance of". The Q5 in the figure queries the class human, and it can be changed to query books, songs, cities, and humans. Filter command ensures the entity is from Wikidata. If the query is not filtered, then there may be data from DBpedia. The limit command is for the query to return 10,000 data.

## IV. Result and Discussion

### A. Book

| | support | itemsets |
|---|---|---|
| 1 | 0.540204 | (title) |
| 2 | 0.502831 | (author) |
| 3 | 0.388448 | (publication date) |
| 4 | 0.288788 | (language of work or name) |
| 5 | 0.224236 | (collection) |

Figure 4.1

| | support | itemsets |
|---|---|---|
| 14 | 0.318233 | (publication date, author) |
| 13 | 0.262741 | (title, author) |
| 15 | 0.261608 | (publication date, title) |
| 17 | 0.236693 | (author, language of work or name) |
| 18 | 0.225368 | (publication date, language of work or name) |

Figure 4.2

| | support | itemsets |
|---|---|---|
| 16 | 0.219706 | (publication date, title, author) |
| 57 | 0.214043 | (described at URL, title, inventory number) |
| 59 | 0.214043 | (collection, described at URL, inventory number) |
| 58 | 0.214043 | (collection, title, described at URL) |
| 75 | 0.214043 | (collection, title, inventory number) |

Figure 4.3

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 0 | (author) | (title) | 0.502831 | 0.540204 | 0.262741 | 0.522523 | 0.967269 |
| 1 | (title) | (author) | 0.540204 | 0.502831 | 0.262741 | 0.486373 | 0.967269 |
| 279 | (publisher) | (title) | 0.171008 | 0.540204 | 0.112118 | 0.655629 | 1.213670 |
| 4 | (publication date) | (title) | 0.388448 | 0.540204 | 0.261608 | 0.673469 | 1.246695 |
| 5 | (title) | (publication date) | 0.540204 | 0.388448 | 0.261608 | 0.484277 | 1.246695 |
| 16 | (language of work or name) | (title) | 0.288788 | 0.540204 | 0.195923 | 0.678431 | 1.255880 |
| 17 | (title) | (language of work or name) | 0.540204 | 0.288788 | 0.195923 | 0.362683 | 1.255880 |
| 11 | (title) | (publication date, author) | 0.540204 | 0.318233 | 0.219706 | 0.406709 | 1.278020 |
| 6 | (publication date, author) | (title) | 0.318233 | 0.540204 | 0.219706 | 0.690391 | 1.278020 |
| 290 | (publication date, publisher) | (title) | 0.152888 | 0.540204 | 0.107588 | 0.703704 | 1.302663 |

Figure 4.4

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 132 | (copyright status, title, inventory number) | (collection, described at URL) | 0.142695 | 0.214043 | 0.142695 | 1.000000 | 4.671958 |
| 257 | (title, inventory number) | (collection, described at URL) | 0.214043 | 0.214043 | 0.214043 | 1.000000 | 4.671958 |
| 131 | (copyright status, collection, described at URL) | (title, inventory number) | 0.142695 | 0.214043 | 0.142695 | 1.000000 | 4.671958 |
| 139 | (collection, described at URL) | (copyright status, title, inventory number) | 0.214043 | 0.142695 | 0.142695 | 0.666667 | 4.671958 |
| 252 | (collection, described at URL) | (title, inventory number) | 0.214043 | 0.214043 | 0.214043 | 1.000000 | 4.671958 |
| 140 | (title, inventory number) | (copyright status, collection, described at URL) | 0.214043 | 0.142695 | 0.142695 | 0.666667 | 4.671958 |
| 135 | (copyright status, described at URL, inventory... | (collection, title) | 0.142695 | 0.215176 | 0.142695 | 1.000000 | 4.647368 |
| 254 | (described at URL, inventory number) | (collection, title) | 0.214043 | 0.215176 | 0.214043 | 1.000000 | 4.647368 |
| 255 | (collection, title) | (described at URL, inventory number) | 0.215176 | 0.214043 | 0.214043 | 0.994737 | 4.647368 |
| 136 | (collection, title) | (copyright status, described at URL, inventory... | 0.215176 | 0.142695 | 0.142695 | 0.663158 | 4.647368 |

Figure 4.5

In figure 4.1, we find the support of 0.54 for book "title" is interesting because unlike "authors", which can be anonymous, we figured that a book should have a title so that it can be identified.

In figure 4.2, where each itemset consists of two items, we find that "publication date" and "author" pairing to have the most support to be interesting, because they don't seem to have much connection between them compared to "title" and "author". We believe this is because books that have no authors tend to be old books thus making their publication date also unknown.

In figure 4.3, where the item "language of work or name" appears with high support in figure 4.1 and 4.2, it doesn't seem to appear in this figure. Instead, "described at URL" and "inventory number" got boosted up.

What we found interesting is that author and title are negatively correlated (lift < 1.0) where it's supposed to rely on each other, and *vice versa*. Another odd thing in the book class is that there is no genre item in the itemsets of the

Wikidata even though in the example of the book in Wikidata there is a bible genre.

*B. City*

| | support | itemsets |
|---|---|---|
| 0 | 0.955882 | (country) |
| 1 | 0.911765 | (coordinate location) |
| 11 | 0.911765 | (located in the administrative territorial ent... |
| 2 | 0.897059 | (Commons category) |
| 3 | 0.897059 | (area) |

Figure 4.6

| | support | itemsets |
|---|---|---|
| 266 | 0.911765 | (country, located in the administrative territ... |
| 13 | 0.897059 | (coordinate location, Commons category) |
| 67 | 0.897059 | (population, coordinate location) |
| 66 | 0.897059 | (population, Commons category) |
| 35 | 0.897059 | (country, image) |

Figure 4.7

| | support | itemsets |
|---|---|---|
| 72 | 0.897059 | (population, coordinate location, Commons cate... |
| 65 | 0.882353 | (country, located in the administrative territ... |
| 24 | 0.867647 | (area, coordinate location, Commons category) |
| 74 | 0.867647 | (area, population, Commons category) |
| 73 | 0.867647 | (area, population, coordinate location) |

Figure 4.8

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 3008 | (topic's main category, population, coordinate... | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 2946 | (topic's main category, population) | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 3129 | (Commons category, topic's main category, popu... | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 2940 | (Commons category, topic's main category) | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 2934 | (topic's main category, coordinate location) | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 2994 | (Commons category, topic's main category, coor... | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 3022 | (Commons category, topic's main category, popu... | (country) | 0.852941 | 0.955882 | 0.808824 | 0.948276 | 0.992042 |
| 2951 | (country) | (topic's main category, population) | 0.955882 | 0.852941 | 0.808824 | 0.846154 | 0.992042 |
| 3007 | (country) | (Commons category, topic's main category, coor... | 0.955882 | 0.852941 | 0.808824 | 0.846154 | 0.992042 |
| 2939 | (country) | (topic's main category, coordinate location) | 0.955882 | 0.852941 | 0.808824 | 0.846154 | 0.992042 |

Figure 4.9

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 1709 | (population, coordinate location, image) | (area, located in the administrative territori... | 0.838235 | 0.808824 | 0.808824 | 0.964912 | 1.192982 |
| 1513 | (country, image, Commons category) | (coordinate location, population, located in t... | 0.838235 | 0.823529 | 0.823529 | 0.982456 | 1.192982 |
| 1538 | (image, Commons category) | (country, coordinate location, population, loc... | 0.838235 | 0.823529 | 0.823529 | 0.982456 | 1.192982 |
| 1537 | (population, image) | (country, coordinate location, located in the ... | 0.838235 | 0.823529 | 0.823529 | 0.982456 | 1.192982 |
| 1535 | (located in the administrative territorial ent... | (country, population, coordinate location, image) | 0.823529 | 0.838235 | 0.823529 | 1.000000 | 1.192982 |
| 1534 | (population, located in the administrative ter... | (country, coordinate location, image, Commons ... | 0.823529 | 0.838235 | 0.823529 | 1.000000 | 1.192982 |
| 1524 | (coordinate location, image, Commons category) | (country, population, located in the administr... | 0.838235 | 0.823529 | 0.823529 | 0.982456 | 1.192982 |
| 1840 | (country, population, coordinate location, image) | (area, located in the administrative territori... | 0.838235 | 0.808824 | 0.808824 | 0.964912 | 1.192982 |
| 1523 | (population, coordinate location, image) | (country, located in the administrative territ... | 0.838235 | 0.823529 | 0.823529 | 0.982456 | 1.192982 |
| 1521 | (coordinate location, located in the administr... | (country, population, image) | 0.823529 | 0.838235 | 0.823529 | 1.000000 | 1.192982 |

Figure 4.10

In figure 4.6, we find the support of the country is the highest. Where in figure 4.9, All of the itemsets depend on the country and the ones with a lift < 1 is related to the "topic's main category" property.

*C. Film*

| | support | itemsets |
|---|---|---|
| 0 | 0.9750 | (country of origin) |
| 1 | 0.9625 | (color) |
| 2 | 0.9625 | (original language of film or TV show) |
| 3 | 0.9625 | (genre) |
| 4 | 0.9500 | (director) |

Figure 4.11

| | support | itemsets |
|---|---|---|
| 10 | 0.9625 | (country of origin, color) |
| 11 | 0.9625 | (country of origin, original language of film ... |
| 21 | 0.9500 | (director, genre) |
| 36 | 0.9500 | (cast member, genre) |
| 14 | 0.9500 | (country of origin, genre) |

Figure 4.12

| | support | itemsets |
|---|---|---|
| 13 | 0.9500 | (country of origin, color, original language o... |
| 41 | 0.9375 | (director, cast member, genre) |
| 18 | 0.9375 | (country of origin, genre, color) |
| 17 | 0.9375 | (country of origin, genre, original language o... |
| 25 | 0.9375 | (director, country of origin, genre) |

Figure 4.13

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 5492 | (cast member, genre, director) | (country of origin, original language of film ... | 0.9375 | 0.8625 | 0.8125 | 0.866667 | 1.004831 |
| 5469 | (country of origin, original language of film ... | (cast member, genre, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |
| 5078 | (original language of film or TV show, color, ... | (cast member, genre, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |
| 5071 | (cast member, genre, director) | (original language of film or TV show, color, ... | 0.9375 | 0.8625 | 0.8125 | 0.866667 | 1.004831 |
| 5249 | (publication date, original language of film o... | (cast member, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |
| 5585 | (publication date, color, title, original lang... | (cast member, genre, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |
| 5628 | (cast member, genre, director) | (publication date, color, title, original lang... | 0.9375 | 0.8625 | 0.8125 | 0.866667 | 1.004831 |
| 4906 | (cast member, director) | (original language of film or TV show, color, ... | 0.9375 | 0.8625 | 0.8125 | 0.866667 | 1.004831 |
| 4903 | (original language of film or TV show, color, ... | (cast member, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |
| 5186 | (country of origin, original language of film ... | (cast member, director) | 0.8625 | 0.9375 | 0.8125 | 0.942029 | 1.004831 |

Figure 4.14

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 8954 | (director, duration) | (country of origin, genre, color, publication ... | 0.8000 | 0.8875 | 0.8 | 1.000000 | 1.126761 |
| 8639 | (cast member, country of origin, duration) | (genre, color, publication date) | 0.8000 | 0.8875 | 0.8 | 1.000000 | 1.126761 |
| 8573 | (cast member, duration) | (genre, color, publication date) | 0.8000 | 0.8875 | 0.8 | 1.000000 | 1.126761 |
| 8660 | (cast member, duration) | (country of origin, genre, color, publication ... | 0.8000 | 0.8875 | 0.8 | 1.000000 | 1.126761 |
| 8637 | (country of origin, genre, color, publication ... | (cast member, duration) | 0.8875 | 0.8000 | 0.8 | 0.901408 | 1.126761 |
| 8572 | (genre, color, publication date) | (cast member, duration) | 0.8875 | 0.8000 | 0.8 | 0.901408 | 1.126761 |
| 8658 | (genre, color, publication date) | (cast member, country of origin, duration) | 0.8875 | 0.8000 | 0.8 | 0.901408 | 1.126761 |
| 8863 | (director, duration) | (genre, color, publication date) | 0.8000 | 0.8875 | 0.8 | 1.000000 | 1.126761 |
| 8862 | (genre, color, publication date) | (director, duration) | 0.8875 | 0.8000 | 0.8 | 0.901408 | 1.126761 |
| 8948 | (genre, color, publication date) | (director, country of origin, duration) | 0.8875 | 0.8000 | 0.8 | 0.901408 | 1.126761 |

Figure 4.15

In figure 4.11, we find that country of origin has the highest support. Oddly the title and the publication date is not on the top 5 of the support, where nowadays title and publication date is more important than color because people will search for what is the title and when the film is published, not whether the film is colored or not, and almost every color in the film is not black and white.

The interesting thing from figure 4.14 is that the minimum lift is 1.004831 (lift > 1.0) means that all of them are positively correlated.

*D. Human*

| | support | itemsets |
|---|---|---|
| 0 | 0.965517 | (sex or gender) |
| 1 | 0.965517 | (occupation) |
| 2 | 0.948276 | (country of citizenship) |
| 3 | 0.948276 | (place of birth) |
| 4 | 0.931034 | (date of birth) |

Figure 4.16

| | support | itemsets |
|---|---|---|
| 7 | 0.965517 | (sex or gender, occupation) |
| 8 | 0.948276 | (country of citizenship, occupation) |
| 9 | 0.948276 | (country of citizenship, sex or gender) |
| 11 | 0.948276 | (place of birth, occupation) |
| 12 | 0.948276 | (sex or gender, place of birth) |

Figure 4.17

| | support | itemsets |
|---|---|---|
| 10 | 0.948276 | (country of citizenship, sex or gender, occupa... |
| 14 | 0.948276 | (sex or gender, place of birth, occupation) |
| 38 | 0.931034 | (place of birth, image, occupation) |
| 40 | 0.931034 | (sex or gender, place of birth, image) |
| 39 | 0.931034 | (sex or gender, image, occupation) |

Figure 4.18

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 597 | (date of birth) | (place of birth, country of citizenship, image... | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 506 | (date of birth) | (image, place of birth, sex or gender, country... | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 475 | (date of birth) | (occupation, image, country of citizenship, se... | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 536 | (date of birth) | (occupation, image, place of birth, country of... | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 402 | (date of birth) | (occupation, image, country of citizenship) | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 374 | (date of birth) | (image, country of citizenship, sex or gender) | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 333 | (date of birth) | (image, country of citizenship) | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 416 | (date of birth) | (image, country of citizenship, place of birth) | 0.931034 | 0.913793 | 0.844828 | 0.907407 | 0.993012 |
| 544 | (place of birth, country of citizenship, image... | (date of birth) | 0.913793 | 0.931034 | 0.844828 | 0.924528 | 0.993012 |
| 409 | (image, country of citizenship, place of birth) | (date of birth) | 0.913793 | 0.931034 | 0.844828 | 0.924528 | 0.993012 |

Figure 4.19

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 1666 | (Commons category, sex or gender) | (date of birth, country of citizenship, image,... | 0.862069 | 0.844828 | 0.844828 | 0.98 | 1.16 |
| 1828 | (sex or gender, Commons category, country of c... | (date of birth, image) | 0.844828 | 0.862069 | 0.844828 | 1.00 | 1.16 |
| 1053 | (date of birth, country of citizenship, image) | (Commons category, place of birth) | 0.844828 | 0.862069 | 0.844828 | 1.00 | 1.16 |
| 1036 | (Commons category, place of birth) | (date of birth, sex or gender, image) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1033 | (Commons category, sex or gender) | (date of birth, place of birth, image) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1032 | (date of birth, image) | (Commons category, sex or gender, place of birth) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1025 | (Commons category, sex or gender, place of birth) | (date of birth, image) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1024 | (date of birth, place of birth, image) | (Commons category, sex or gender) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1021 | (date of birth, sex or gender, image) | (Commons category, place of birth) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |
| 1004 | (Commons category, place of birth) | (date of birth, image, occupation) | 0.862069 | 0.862069 | 0.862069 | 1.00 | 1.16 |

Figure 4.20

In figure 4.17 and 4.18, we find that the top itemsets of 2-itemset and 3-itemset consist of the combination between the top itemsets in 1-itemset. From figure 4.19, we can see that the top 8 association with negative correlation has “date

of birth" property as its antecedent. This means that having the property "date of birth" makes it less likely for some itemsets to appear.

*E. Song*

| | support | itemsets |
|---|---|---|
| 0 | 0.861190 | (performer) |
| 1 | 0.798867 | (genre) |
| 5 | 0.781870 | (publication date) |
| 2 | 0.640227 | (part of) |
| 3 | 0.592068 | (lyrics by) |

Figure 4.21

| | support | itemsets |
|---|---|---|
| 8 | 0.767705 | (performer, genre) |
| 20 | 0.733711 | (performer, publication date) |
| 21 | 0.679887 | (genre, publication date) |
| 9 | 0.628895 | (part of, performer) |
| 10 | 0.603399 | (part of, genre) |

Figure 4.22

| | support | itemsets |
|---|---|---|
| 22 | 0.665722 | (performer, genre, publication date) |
| 12 | 0.597734 | (part of, genre, performer) |
| 26 | 0.572238 | (performer, genre, record label) |
| 13 | 0.563739 | (part of, publication date, performer) |
| 14 | 0.538244 | (part of, genre, publication date) |

Figure 4.23

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 394 | (performer, record label) | (language of work or name) | 0.589235 | 0.526912 | 0.305949 | 0.519231 | 0.985422 |
| 395 | (language of work or name) | (performer, record label) | 0.526912 | 0.589235 | 0.305949 | 0.580645 | 0.985422 |
| 182 | (language of work or name) | (record label) | 0.526912 | 0.592068 | 0.308782 | 0.586022 | 0.989788 |
| 183 | (record label) | (language of work or name) | 0.592068 | 0.526912 | 0.308782 | 0.521531 | 0.989788 |
| 401 | (language of work or name) | (record label, genre) | 0.526912 | 0.572238 | 0.300283 | 0.569892 | 0.995901 |
| 406 | (record label, performer, genre) | (language of work or name) | 0.572238 | 0.526912 | 0.300283 | 0.524752 | 0.995901 |
| 415 | (language of work or name) | (record label, performer, genre) | 0.526912 | 0.572238 | 0.300283 | 0.569892 | 0.995901 |
| 400 | (record label, genre) | (language of work or name) | 0.572238 | 0.526912 | 0.300283 | 0.524752 | 0.995901 |
| 419 | (performer) | (composer) | 0.861190 | 0.478754 | 0.419263 | 0.486842 | 1.016895 |
| 418 | (composer) | (performer) | 0.478754 | 0.861190 | 0.419263 | 0.875740 | 1.016895 |

Figure 4.24

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 86 | (performer, genre, publication date) | (record label) | 0.665722 | 0.592068 | 0.518414 | 0.778723 | 1.315260 |
| 99 | (record label) | (performer, genre, publication date) | 0.592068 | 0.665722 | 0.518414 | 0.875598 | 1.315260 |
| 93 | (genre, publication date) | (performer, record label) | 0.679887 | 0.589235 | 0.518414 | 0.762500 | 1.294050 |
| 92 | (performer, record label) | (genre, publication date) | 0.589235 | 0.679887 | 0.518414 | 0.879808 | 1.294050 |
| 85 | (record label) | (genre, publication date) | 0.592068 | 0.679887 | 0.518414 | 0.875598 | 1.287859 |
| 80 | (genre, publication date) | (record label) | 0.679887 | 0.592068 | 0.518414 | 0.762500 | 1.287859 |
| 68 | (performer, genre) | (record label) | 0.767705 | 0.592068 | 0.572238 | 0.745387 | 1.258956 |
| 73 | (record label) | (performer, genre) | 0.592068 | 0.767705 | 0.572238 | 0.966507 | 1.258956 |
| 29 | (performer, genre, publication date) | (part of) | 0.665722 | 0.640227 | 0.535411 | 0.804255 | 1.256204 |
| 36 | (part of) | (performer, genre, publication date) | 0.640227 | 0.665722 | 0.535411 | 0.836283 | 1.256204 |

Figure 4.25

Another interesting thing that we found in figure 4.21 is that there is no title for the song where most people search for the song based on the title. In figure 4.22 and 4.23, we find that the top itemsets of 2-itemset and 3-itemset consist of the combination between the top itemsets in 1-itemset. In figure 4.24, rows with lift < 1 are related to the "language of work or name" property.

*F. Merge*

| | support | itemsets |
|---|---|---|
| 7 | 0.480583 | (publication date) |
| 8 | 0.455617 | (title) |
| 5 | 0.350902 | (genre) |
| 19 | 0.312760 | (author) |
| 16 | 0.305825 | (language of work or name) |

Figure 4.26

| | support | itemsets |
|---|---|---|
| 46 | 0.295423 | (publication date, genre) |
| 66 | 0.264910 | (publication date, title) |
| 109 | 0.244799 | (language of work or name, publication date) |
| 130 | 0.198336 | (author, publication date) |
| 67 | 0.194175 | (performer, genre) |

Figure 4.27

| | support | itemsets |
|---|---|---|
| 77 | 0.169209 | (publication date, country of origin, genre) |
| 70 | 0.169209 | (publication date, performer, genre) |
| 111 | 0.163662 | (language of work or name, publication date, g... |
| 48 | 0.155340 | (publication date, title, genre) |
| 85 | 0.150485 | (publication date, genre, producer) |

Figure 4.28

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 114 | (genre) | (title) | 0.350902 | 0.455617 | 0.178918 | 0.509881 | 1.119100 |
| 113 | (title) | (genre) | 0.455617 | 0.350902 | 0.178918 | 0.392694 | 1.119100 |
| 119 | (title) | (publication date, genre) | 0.455617 | 0.295423 | 0.155340 | 0.340944 | 1.154086 |
| 116 | (publication date, genre) | (title) | 0.295423 | 0.455617 | 0.155340 | 0.525822 | 1.154086 |
| 527 | (author) | (title) | 0.312760 | 0.455617 | 0.165049 | 0.527716 | 1.158245 |
| 528 | (title) | (author) | 0.455617 | 0.312760 | 0.165049 | 0.362253 | 1.158245 |
| 251 | (publication date) | (title) | 0.480583 | 0.455617 | 0.264910 | 0.551227 | 1.209846 |
| 252 | (title) | (publication date) | 0.455617 | 0.480583 | 0.264910 | 0.581431 | 1.209846 |
| 422 | (title) | (language of work or name) | 0.455617 | 0.305825 | 0.181692 | 0.398782 | 1.303955 |
| 421 | (language of work or name) | (title) | 0.305825 | 0.455617 | 0.181692 | 0.594104 | 1.303955 |

Figure 4.29

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 562 | (collection, described at URL) | (inventory number, title) | 0.131068 | 0.131068 | 0.131068 | 1.000000 | 7.629630 |
| 561 | (inventory number, title) | (collection, described at URL) | 0.131068 | 0.131068 | 0.131068 | 1.000000 | 7.629630 |
| 565 | (inventory number) | (collection, described at URL, title) | 0.133842 | 0.131068 | 0.131068 | 0.979275 | 7.471503 |
| 542 | (inventory number) | (collection, described at URL) | 0.133842 | 0.131068 | 0.131068 | 0.979275 | 7.471503 |
| 541 | (collection, described at URL) | (inventory number) | 0.131068 | 0.133842 | 0.131068 | 1.000000 | 7.471503 |
| 558 | (collection, described at URL, title) | (inventory number) | 0.131068 | 0.133842 | 0.131068 | 1.000000 | 7.471503 |
| 564 | (collection, title) | (inventory number, described at URL) | 0.134535 | 0.131068 | 0.131068 | 0.974227 | 7.432990 |
| 559 | (inventory number, described at URL) | (collection, title) | 0.131068 | 0.134535 | 0.131068 | 1.000000 | 7.432990 |
| 566 | (described at URL) | (inventory number, collection, title) | 0.136616 | 0.131068 | 0.131068 | 0.959391 | 7.319797 |
| 554 | (described at URL) | (inventory number, title) | 0.136616 | 0.131068 | 0.131068 | 0.959391 | 7.319797 |

Figure 4.30

In figure 4.26, it is depicted that the highest support is less than 0.5, meaning not much of the properties are common across all entity types. In figure 4.29, all of the itemsets depend on title, the minimum lift is > 1 making them positively correlated.

### *G. Inter Class*

From our observation, book, film, and song have common properties of publication date. “Publication date” property support is higher in song compared to book. We think that songs are older than books, but because songs from the prehistoric era tend not to be documented and there is no physical evidence to prove the existence of the song unlike songs in the modern time where they would be better documented because of technologies like recordings. Compared to books that already exist back from the BC era, they have physical evidence which can be documented but the publication dates are hard to keep track of. Oddly, the support for “publication date” in film class is not as high compared to book or song class even though it is a modern media.

For the title property, books have the highest support compared to film. Interestingly, title properties in films are not in the top five of the highest support even though when searching for a film, people tend to search with the film title.

## V. Conclusion

Based on our experiment, for some of the entities, we find that the top itemsets of 2-itemset and 3-itemset consist of the combination between the top itemsets in 1-itemset.

Properties that people tend to use to be able to find certain songs and books are non-existent. Based on our result, there is no itemset that contains song title or book genre. Also, In the film class, the title and publication date property does not exist in the top 5 k-itemsets with the highest support.

In view of our test on book, song, and film have some common properties but those three don’t have anything in common with human class or city class..


## Acknowledgements

This research paper was part of the Introduction to Artificial Intelligence and Data Sciences 2021 course from Universitas Indonesia. We are grateful for the support from the teaching team for assisting us on completing this paper. Thank you to Fariz Darari, our lecturer, for giving us feedback on our paper. We are also grateful for Millenio Ramadizsa, our teaching assistant, for guiding us through the process of making this paper.


## References


[1] Rajak, Akash.(2008).Association rule mining- Applications in various areas

[2] Yabing, Jiao.(2013).Research of an Improved Apriori Algorithm in Data Mining Association Rules

[3] Agrawal, Rakesh. Imielinski, Tomasz. & Swami, Arun.(1993). Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD Record,* 22(2), 207-216. doi:10.1145/170035.170072

[4] Labade, Sheetal, Srinivas N. Kini.(2013).A Survey Paper on Frequent Itemset Mining Methods and Techniques

[5] Borgelt, Christian.(2010).An Implementation of the FP-growth Algorithm

[6] Alfiqra, Faiza Yogi Alfizi.(2018).Penerapan Market Basket Analysis Menggunakan Proses KDD (Knowledge Discovery In Database) Sebagai Strategi Penjualan Produk Swalayan

[7] Zhan, Foxiao., Xiaolan Zhu, Lei Zhang, Xuexi Wang. .(2019).Summary of Association Rules

[8] Shankar, Girja., Latita Bargadiya.(2013).A New Improved Apriori Algorithm For Association Rules Mining

[9] Vrandečić, Denny.(2013). The Rise of Wikidata.*IEEE Intelligent Systems,28(4),90-95.*

[10] Vrandečić, Denny. Krötzsch, Markus.(2014). Wikidata: a Free Collaborative Knowledgebase.*Communications of the ACM, 57(10), 78–85.*