ANTIPLAGIAT
PLAGIARISM DETECTION SYSTEM

Text reuse report No.1

Author: Шмид Дэвид -
Checked by: ApiCorp
Organization: Национальный Исследовательский Университет "Высшая Школа Экономики"

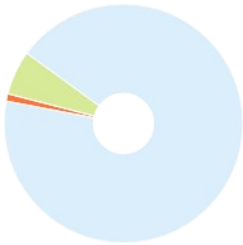
Report presented by the Antiplagiat service - http://hse.antiplagiat.ru

DOCUMENT INFORMATION

Document No.: 435511
Uploading start: 29.01.2023 16:08:20
Uploading duration: 00:00:11
Initial file name: Detecting-patterns-in-purchase-history-using-association-rule-learning-methods_latest_V2.docx
Document: Detecting-patterns-in-purchase-history-using-association-rule-learning-methods_latest_V2
Text size: 76 KB
Document type: Other
Number of characters: 77455
Number of words: 10541
Number of sentences: 625

REPORT INFORMATION

Check start: 29.01.2023 13:08:32
Check duration: 00:02:48
Comment: not specified
Search with editing factored in: Yes
Checked sections: title p. 1, main p. 2,5-47, content p. 3-4, bibliography p. 48-50
Search modules: Cross language search, Collection of the National Library of Uzbekistan, Media collection, Search module of Wiley Open Library paraphrases, Wiley Open Library Cross language, Wiley Open Library, Cross language search on eLIBRARY.RU (EnRu), Cross language search over the Internet (EnRu), Medical collection, Search module INTERNET PLUS, ELS joint collection, Search module of eLIBRARY.RU paraphrases, Search module of Internet paraphrases (EN), Search module of Internet paraphrases, Search module of common phrases, Bibliography separation module, Legal information system «Adilet» search module, RSL collection, Cross language search (RuEn), Dissertations and abstracts of National Library of Belarus, Search module of GARANT analytics paraphrases, GARANT analytics, GARANT legal documentation, Patents collection, eLIBRARY.RU collection, Citations, Institutes Unity collection, Модуль поиска "ВШЭ"



SIMILARITY	TEXT RECYCLING	CITATIONS	ORIGINALITY
0,81%	0%	5,72%	93,47%

Similarity is a share of all found overlapping text fragments, with the exception of those categorized as citations by the system, as compared to the entire document size.
Text recycling refer to the share of text fragments (relative to the whole document) in the document being checked for reuse that are completely/partially identical to the text fragments from a source which was authored/co-authored by the individual whose document is being checked.
Citations are a share of overlapping text which has not been created by the respective author, but the system considered its use correct as compared to the entire document size. These include citations complying with the GOST standard; common phrases; text fragments found in legal and regulatory documentation source collections.
Overlapping text is a text fragment of the checked document which is identical or almost identical to a source text fragment.
Source is a document indexed by the system and contained in the search module which is used for check.
Originality is a share of text fragments in the checked document that have not been found in any checked sources as compared to the entire document size.
Similarity, text recycling, citations, and originality are separate indicators. Their sum is equal to 100%, with respect to the entire text of the checked document.
Please note that the system finds overlapping texts in the checked document and text sources indexed by the system. At the same time, the system is an auxiliary tool. Correctness and adequacy of reuse or citations, as well as authorship of text fragments in the checked document must be determined by the verifier.

№	Report share	Text share	Source	Valid from	Search module	Report blocks	Text blocks	Comment
[01]	5,72%	5,72%	not specified	13 Jan 2022	Bibliography separation module	2	2	
[02]	0%	0,93%	Protecting Sensitive knowledge in assoc... https://doi.org	31 Jan 2012	Wiley Open Library	0	8	Source excluded. Reason: Low overlapping percentage.
[03]	0%	0,86%	Inter-transactional association rules for ... http://elibrary.ru	25 Aug 2014	eLIBRARY.RU collection	0	5	Source excluded. Reason: Low overlapping percentage.
[04]	0%	0,84%	A lattice-based approach for I/O efficien... http://elibrary.ru	22 Aug 2014	eLIBRARY.RU collection	0	5	Source excluded. Reason: Low overlapping percentage.
[05]	0%	0,77%	Progressive Weighted Miner: An Efficien... http://elibrary.ru	25 Aug 2014	eLIBRARY.RU collection	0	5	Source excluded. Reason: Low overlapping percentage.
[06]	0%	0,72%	180030_b-pmin41_2022_1	09 Nov 2022	Institutes Unity collection	0	5	Source excluded. Reason: Low overlapping percentage.
[07]	0%	0,7%	A survey of itemset mining https://doi.org	31 Jul 2017	Wiley Open Library	0	5	Source excluded. Reason: Low overlapping percentage.
[08]	0,48%	0,68%	Association Rules Mining: A Recent Over... http://docplayer.net	06 Jan 2018	Search module of Internet paraphrases (EN)	1	2	
[09]	0%	0,68%	Multimedia Data Mining: State of the Ar... http://comp.nus.edu.sg	07 Jan 2018	Search module of Internet paraphrases (EN)	0	2	
[10]	0%	0,67%	Frequent itemset mining: A 25 years rev... https://doi.org	30 Nov 2019	Wiley Open Library	0	4	Source excluded. Reason: Low overlapping percentage.
[11]	0%	0,66%	Beyond Market Baskets: Generalizing As... http://elibrary.ru	22 Aug 2014	eLIBRARY.RU collection	0	4	Source excluded. Reason: Low overlapping percentage.
[12]	0%	0,65%	Қарор қабул қилишга кўмаклашиши ... http://diss.natlib.uz	09 Jun 2021	Collection of the National Library of Uzbekistan	0	4	Source excluded. Reason: Low overlapping percentage.
[13]	0%	0,64%	Peer Reviewed Journal http://ijera.com	09 May 2019	Search module INTERNET PLUS	0	9	Source excluded. Reason: Low overlapping percentage.
[14]	0%	0,62%	AI-Based Modeling: Techniques, Applica... https://link.springer.com	29 Jan 2023	Search module INTERNET PLUS	0	7	Source excluded. Reason: Low overlapping percentage.
[15]	0%	0,59%	Combined association rules for dealing ... http://elibrary.ru	28 Aug 2014	eLIBRARY.RU collection	0	3	Source excluded. Reason: Low overlapping percentage.

[16]	0%	0,58%	Зубков, Сергей Александрович Разра... http://dlib.rsl.ru	27 Jun 2022	RSL collection	0	3	Source excluded. Reason: Low overlapping percentage.
[17]	0%	0,58%	http://sclab.yonsei.ac.kr/courses/09-AI/... http://sclab.yonsei.ac.kr	28 May 2021	Search module INTERNET PLUS	0	6	Source excluded. Reason: Low overlapping percentage.
[18]	0%	0,57%	https://arxiv.org/ftp/arxiv/papers/1807/... https://arxiv.org	18 Feb 2020	Search module INTERNET PLUS	0	12	Source excluded. Reason: Low overlapping percentage.
[19]	0%	0,57%	On the discovery of association rules by... https://doi.org	30 Sep 2011	Wiley Open Library	0	4	Source excluded. Reason: Low overlapping percentage.
[20]	0%	0,56%	Global Land Surface Dry/Wet Condition... https://doi.org	30 Sep 2021	Wiley Open Library	0	3	Source excluded. Reason: Low overlapping percentage.
[21]	0%	0,56%	https://sci2s.ugr.es/keel/pdf/specific/art... https://sci2s.ugr.es	29 Jan 2023	Search module INTERNET PLUS	0	6	Source excluded. Reason: Low overlapping percentage.
[22]	0%	0,53%	https://www.cse.iitk.ac.in/users/nsrivast... https://cse.iitk.ac.in	11 Sep 2022	Search module INTERNET PLUS	0	5	Source excluded. Reason: Low overlapping percentage.
[23]	0%	0,53%	https://www.cs.ubbcluj.ro/~gabis/DocDi... https://cs.ubbcluj.ro	07 Apr 2022	Search module INTERNET PLUS	0	5	Source excluded. Reason: Low overlapping percentage.
[24]	0%	0,51%	Self-Tuning Clustering: An Adaptive Clus... http://elibrary.ru	24 Aug 2014	eLIBRARY.RU collection	0	3	Source excluded. Reason: Low overlapping percentage.
[25]	0%	0,5%	Feature selection for classification of os... https://doi.org	30 Nov 2012	Wiley Open Library	0	2	Source excluded. Reason: Low overlapping percentage.
[26]	0%	0,48%	Covariate Balancing through Naturally ... https://doi.org	28 Feb 2018	Wiley Open Library	0	2	Source excluded. Reason: Low overlapping percentage.
[27]	0%	0,48%	Role of soft computing as a tool in data ... http://ijcsit.com	07 Jan 2018	Search module of Internet paraphrases (EN)	0	1	
[28]	0%	0,47%	nechaeva_p_a_optimizaciya-spiskov-kli...	19 May 2022	Модуль поиска "ВШЭ"	0	3	Source excluded. Reason: Low overlapping percentage.
[29]	0%	0,47%	Холод, Иван Иванович Модели и мет... http://dlib.rsl.ru	15 Oct 2019	RSL collection	0	3	Source excluded. Reason: Low overlapping percentage.
[30]	0%	0,47%	RFID Tag Performance: Linking the Labo... https://doi.org	31 Oct 2018	Search module of Wiley Open Library paraphrases	0	2	Source excluded. Reason: Low overlapping percentage.
[31]	0%	0,45%	Fundamentals of association rules in da... https://doi.org	31 Mar 2011	Search module of Wiley Open Library paraphrases	0	2	Source excluded. Reason: Low overlapping percentage.
[32]	0%	0,44%	Extended vertical lists for temporal patt... https://doi.org	31 Oct 2019	Wiley Open Library	0	4	Source excluded. Reason: Low overlapping percentage.
[33]	0%	0,43%	Global Land Surface Dry/Wet Condition... https://doi.org	30 Sep 2021	Search module of Wiley Open Library paraphrases	0	2	Source excluded. Reason: Low overlapping percentage.
[34]	0%	0,42%	Tatiana Pavlovna Makhalova; [Место за... http://dlib.rsl.ru	19 Aug 2020	RSL collection	0	3	Source excluded. Reason: Low overlapping percentage.
[35]	0%	0,41%	Материалы студенческой научной се... http://biblioclub.ru	21 Jan 2020	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[36]	0%	0,4%	Method and system for mining associat... http://freepatentsonline.com	09 Nov 2016	Patents collection	0	2	Source excluded. Reason: Low overlapping percentage.
[37]	0%	0,4%	i DECLARATION http://csse.monash.edu.au	07 Jan 2018	Search module of Internet paraphrases (EN)	0	2	Source excluded. Reason: Low overlapping percentage.
[38]	0%	0,38%	Интеллектуальный анализ данных https://e.lanbook.com	22 Jan 2020	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[39]	0%	0,38%	Интеллектуальный анализ данных: уч... https://e.lanbook.com	22 Jan 2020	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[40]	0%	0,38%	Интеллектуальный анализ данных: уч... http://biblioclub.ru	21 Jan 2020	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[41]	0%	0,38%	Крайнов, Александр Юрьевич диссер... http://dlib.rsl.ru	25 Dec 2015	RSL collection	0	2	Source excluded. Reason: Low overlapping percentage.
[42]	0%	0,38%	http://myweb.sabanciuniv.edu/rdehkh... http://myweb.sabanciuniv.edu	26 May 2022	Search module INTERNET PLUS	0	4	Source excluded. Reason: Low overlapping percentage.
[43]	0%	0,38%	http://myweb.sabanciuniv.edu/rdehkh... http://myweb.sabanciuniv.edu	27 Mar 2022	Search module INTERNET PLUS	0	4	Source excluded. Reason: Low overlapping percentage.
[44]	0%	0,37%	Next Market Basket Prediction Based o...	29 Apr 2022	Модуль поиска "ВШЭ"	0	2	Source excluded. Reason: Low overlapping percentage.
[45]	0%	0,36%	An opcode-based technique for polym... https://doi.org	25 Mar 2020	Wiley Open Library	0	3	Source excluded. Reason: Low overlapping percentage.
[46]	0%	0,36%	https://etu.ru/assets/files/nauka/disser... https://etu.ru	06 Oct 2020	Search module INTERNET PLUS	0	6	Source excluded. Reason: Low overlapping percentage.
[47]	0%	0,35%	Using object relational extensions for m... http://freepatentsonline.com	06 Nov 2016	Patents collection	0	3	Source excluded. Reason: Low overlapping percentage.
[48]	0%	0,35%	Method and apparatus for deriving an a... http://freepatentsonline.com	06 Nov 2016	Patents collection	0	2	Source excluded. Reason: Low overlapping percentage.
[49]	0%	0,35%	Method and apparatus for computing a... http://freepatentsonline.com	05 Nov 2016	Patents collection	0	2	Source excluded. Reason: Low overlapping percentage.
[50]	0%	0,34%	260911 http://e.lanbook.com	10 Mar 2016	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[51]	0%	0,34%	Информационное обеспечение актуа... https://book.ru	03 Jul 2017	ELS joint collection	0	2	Source excluded. Reason: Low overlapping percentage.
[52]	0%	0,34%	https://www.cs.cornell.edu/home/klein... https://cs.cornell.edu	17 Apr 2021	Search module INTERNET PLUS	0	5	Source excluded. Reason: Low overlapping percentage.
[53]	0%	0,33%	Mining association rules on significant r... http://elibrary.ru	26 Aug 2003	eLIBRARY.RU collection	0	2	Source excluded. Reason: Low overlapping percentage.
[54]	0,33%	0,33%	.[PDF] http://suraj.lums.edu.pk	06 Jan 2018	Search module of Internet paraphrases (EN)	1	1	
[55]	0%	0,33%	http://repository.unima.ac.id/jspui/bitst... http://repository.unima.ac.id	06 Oct 2022	Search module INTERNET PLUS	0	3	Source excluded. Reason: Low overlapping percentage.

[56]	0%	0,32%	Ковалев, Дмитрий Александрович дис... http://dlib.rsl.ru	earlier than 2011	RSL collection	0	2	Source excluded. Reason: Low overlapping percentage.
[57]	0%	0,32%	Зудов, Антон Борисович Модельные ... http://dlib.rsl.ru	22 Aug 2019	RSL collection	0	2	Source excluded. Reason: Low overlapping percentage.
[58]	0%	0,31%	Depth first method for generating items... http://freepatentsonline.com	06 Nov 2016	Patents collection	0	2	Source excluded. Reason: Low overlapping percentage.
[59]	0%	0,31%	Бузмаков, Алексей Владимирович Мо... http://dlib.rsl.ru	14 Sep 2020	RSL collection	0	2	Source excluded. Reason: Low overlapping percentage.
[60]	0%	0,29%	A Survey of Outlier Detection Methodol... https://dl.acm.org	29 Jan 2023	Search module INTERNET PLUS	0	5	Source excluded. Reason: Low overlapping percentage.
[61]	0%	0,28%	Review of Statistical Methodologies for ... https://frontiersin.org	13 Jan 2021	Media collection	0	3	Source excluded. Reason: Low overlapping percentage.
[62]	0%	0,28%	Incremental association rule mining: a s... https://doi.org	31 May 2013	Search module of Wiley Open Library paraphrases	0	2	Source excluded. Reason: Low overlapping percentage.
[63]	0%	0,28%	Ontology Integration for Linked Data | ... https://link.springer.com	29 Jan 2023	Search module INTERNET PLUS	0	3	Source excluded. Reason: Low overlapping percentage.
[64]	0%	0,27%	Contents http://fit.vutbr.cz	09 Jan 2018	Search module of Internet paraphrases (EN)	0	1	Source excluded. Reason: Low overlapping percentage.
[65]	0%	0,27%	An Efficient Algorithm to Automated Dis... http://thesai.org	06 Jan 2018	Search module of Internet paraphrases (EN)	0	1	Source excluded. Reason: Low overlapping percentage.
[66]	0%	0,24%	Методы и алгоритмы обработки и ан... http://dep.nlb.by	11 Nov 2016	Dissertations and abstracts of National Library of Belarus	0	2	Source excluded. Reason: Low overlapping percentage.
[67]	0%	0,23%	Пан, Константин Сергеевич диссерта... http://dlib.rsl.ru	earlier than 2011	RSL collection	0	2	Source excluded. Reason: Low overlapping percentage.
[68]	0%	0,22%	https://arxiv.org/pdf/1703.02819.pdf https://arxiv.org	15 Mar 2022	Search module INTERNET PLUS	0	2	Source excluded. Reason: Low overlapping percentage.
[69]	0%	0,22%	Машинное обучение. Наука и искусс... https://e.lanbook.com	22 Jan 2020	ELS joint collection	0	1	Source excluded. Reason: Low overlapping percentage.
[70]	0%	0,22%	Машинное обучение. Наука и искусс... http://studentlibrary.ru	20 Dec 2016	Medical collection	0	1	Source excluded. Reason: Low overlapping percentage.
[71]	0%	0,22%	Application of Association Rules to Dete... https://frontiersin.org	18 Aug 2020	Media collection	0	2	Source excluded. Reason: Low overlapping percentage.
[72]	0%	0,22%	Программные продукты и системы: н... http://biblioclub.ru	21 Jan 2020	ELS joint collection	0	1	Source excluded. Reason: Low overlapping percentage.
[73]	0%	0,22%	THE MFFP-TREE FUZZY MINING ALGOR... https://doi.org	28 Feb 2014	Search module of Wiley Open Library paraphrases	0	1	Source excluded. Reason: Low overlapping percentage.
[74]	0%	0,22%	A survey on association rules mining usi... https://doi.org	31 Jul 2019	Search module of Wiley Open Library paraphrases	0	1	Source excluded. Reason: Low overlapping percentage.
[75]	0%	0,21%	Применение методов генетического ... http://elibrary.ru	01 Feb 2021	eLIBRARY.RU collection	0	1	Source excluded. Reason: Low overlapping percentage.
[76]	0%	0,2%	Parallel Semi-supervised enhanced fuzz... https://doi.org	25 Jul 2019	Wiley Open Library	0	1	Source excluded. Reason: Low overlapping percentage.
[77]	0%	0,2%	Exploring the Mechanism of Flavonoids ... https://frontiersin.org	09 Feb 2021	Media collection	0	3	Source excluded. Reason: Low overlapping percentage.
[78]	0%	0,2%	Parallel Semi-supervised enhanced fuzz... https://doi.org	25 Jul 2019	Search module of Wiley Open Library paraphrases	0	1	Source excluded. Reason: Low overlapping percentage.
[79]	0%	0,2%	https://decisionsciences.org/wp-conten... https://decisionsciences.org	17 Feb 2022	Search module INTERNET PLUS	0	1	Source excluded. Reason: Low overlapping percentage.
[80]	0%	0,18%	Abramyan-WinForms-2021 Учебник	08 Jun 2021	Institutes Unity collection	0	2	Source excluded. Reason: Low overlapping percentage.
[81]	0%	0,17%	gerasimov_v_a_preimushchestva-i-ned...	20 May 2020	Модуль поиска "ВШЭ"	0	1	Source excluded. Reason: Low overlapping percentage.
[82]	0%	0,16%	Interactive mining of most interesting r... http://freepatentsonline.com	05 Nov 2016	Patents collection	0	1	Source excluded. Reason: Low overlapping percentage.
[83]	0%	0,15%	Indices for concept assessment	01 May 2022	Модуль поиска "ВШЭ"	0	1	Source excluded. Reason: Low overlapping percentage.
[84]	0%	0,12%	260857 http://e.lanbook.com	10 Mar 2016	ELS joint collection	0	1	Source excluded. Reason: Low overlapping percentage.
[85]	0%	0,12%	260916 http://e.lanbook.com	10 Mar 2016	ELS joint collection	0	1	Source excluded. Reason: Low overlapping percentage.
[86]	0%	0,12%	https://petsymposium.org/popets/202... https://petsymposium.org	07 Sep 2022	Search module INTERNET PLUS	0	1	Source excluded. Reason: Low overlapping percentage.
[87]	0%	0,11%	Method and system for harvesting feed... http://freepatentsonline.com	06 Nov 2016	Patents collection	0	1	Source excluded. Reason: Low overlapping percentage.
[88]	0%	0,11%	Байбулатов, Артур Арсенович Исслед... http://dlib.rsl.ru	01 Jan 2018	RSL collection	0	1	Source excluded. Reason: Low overlapping percentage.
[89]	0%	0,11%	Алгоритмы декомпозиции булевых ф... http://dep.nlb.by	04 Jul 2017	Dissertations and abstracts of National Library of Belarus	0	1	Source excluded. Reason: Low overlapping percentage.
[90]	0%	0,11%	Approach to Detecting Anomalies in We...	30 Apr 2022	Модуль поиска "ВШЭ"	0	1	Source excluded. Reason: Low overlapping percentage.
[91]	0%	0,11%	Attribute-Aware Recommender System ... https://frontiersin.org	20 Jan 2021	Media collection	0	1	Source excluded. Reason: Low overlapping percentage.
[92]	0%	0,11%	Каршиев, Зайнидин Абдувалиевич ди... http://dlib.rsl.ru	earlier than 2011	RSL collection	0	1	Source excluded. Reason: Low overlapping percentage.
[93]	0%	0,11%	https://gbata.org/wp-content/uploads/... https://gbata.org	10 Jan 2022	Search module INTERNET PLUS	0	1	Source excluded. Reason: Low overlapping percentage.
[94]	0%	0,1%	Generating conditional functional depe... http://freepatentsonline.com	09 Nov 2016	Patents collection	0	1	Source excluded. Reason: Low overlapping percentage.

[95]	0%	0,09%	Методика и алгоритмы поиска медиц... http://dep.nlb.by	11 Nov 2016	Dissertations and abstracts of National Library of Belarus	0	1	Source excluded. Reason: Low overlapping percentage.
[96]	0%	0,09%	Неклюдов Кирилл Олегович; [Место з... http://dlib.rsl.ru	12 Jan 2021	RSL collection	0	1	Source excluded. Reason: Low overlapping percentage.
[97]	0%	0,08%	Vertigo, nausea, tinnitus and hypoacusi... http://emll.ru	21 Dec 2016	Medical collection	0	1	Source excluded. Reason: Low overlapping percentage.
[98]	0%	0,08%	Animal Rights Movement in Russia and ...	30 Apr 2022	Модуль поиска "ВШЭ"	0	1	Source excluded. Reason: Low overlapping percentage.
[99]	0%	0,06%	not specified	13 Jan 2022	Citations	0	1	Source excluded. Reason: Low overlapping percentage.
[100]	0%	0,06%	https://icestconf.org/wp-content/uploa... https://icestconf.org	15 Sep 2022	Search module INTERNET PLUS	0	1	Source excluded. Reason: Low overlapping percentage.

National Research University Higher School of Economics

Faculty of Computer Science

Programme' Master of Data Science'

MASTER'S THESIS

Detecting patterns in purchase history using association rule learning methods

---

Student:

David Schmid

Supervisor:

HSE Lecturer Anastasia Maximovskaya

Moscow, 2023

ABSTRACT

Association rules were introduced in 1993. It is a powerful data mining method, associating items bought together in itemsets and rules. The most popular methods are A-priori, Eclat and FP-Growth algorithms. The author develops a modified FP-growth algorithm, enriching it with a date-decay and profit-function. Moreover, the association rules will get restricted to the most frequent associative path. The author suggests how to consider the same item per transaction multiple times by averaging its profit. The

experiments show that the modified FP-growth algorithm can generate more relevant and stronger itemsets and rules.

Table of Contents

ABSTRACT *2*

1. Introduction *6*

2 Discussion of related works *9*

3 Baseline algorithm *12*

3.1 Simple dataset **12**

3.2 Formal model of association rules **13**

3.2.1 Support 13

3.2.2 Paths: 14

3.2.3 Confidence: 14

3.2.4 Application of association rules to the simple dataset 15

3.3 A-priori **16**

3.4 FP-growth **19**

4 Developing a modified Algorithm based on the FP-growth algorithm *24*

4.1 FP-growth Model adjustments **24**

4.1.1 Introduction 24

4.1.1 Date-decay-function 25

4.1.2 Profit function 26

4.2 Modified FP-growth reconciliation with TAR **28**

4.3 Modified FP-growth with date decay **30**

4.4 Modified FP-growth with profit **33**

4.5 Modified FP-growth with combined date decay and profit **37**

4.6 Performance, Strengths, and weaknesses of the author's modified FP-growth algorithm **39**

5 Experiments/ Analysis of big datasets *41*

5.1 E-Commerce purchase history from an electronics store **41**

5.1.1 Dataset 41

5.1.2 Analysis 42

5.2 E-Commerce purchase history from a jewellery store **45**

5.2.1 Dataset 45

5.2.2 Analysis 47

5.3 E-Commerce purchase history from a cosmetics store **49**

5.3.1 Dataset 49

5.2.2 Analysis 51

6 Conclusion *53*

7 References *55*

8 Applications *57*

1. Introduction

The concept of association rules (AR) was popularized due to a 1993 article by Agrawal et al. [1]. This article shows the application of AR, namely the "A-priori-algorithm" for shelf management to maximize profit. In the evolution of basket data mining, the application of AR belongs to the early methods, which are simple but powerful. AR has not lost its significance in the research. More efficient and modified versions of the base algorithm are developed to mine the relevant and significant basket relations more efficiently. Compared to more complex ML or DL models, the most significant advantage of this relatively simple algorithm is its ease of use and generalization of pre-defined rules, making it scalable to millions of items and Big Data. Moreover, it is simple enough to explain to businesses conceptionally.

The main driver for the business is ultimately not the frequency of items but the profitability. One can assume that more frequent items are lower in price and higher in margin but can be equal with not frequent but highly-priced articles. In an Online-Article from Chou, T [15], a conclusion to AR is written the following:

> "I would like to share an interesting story. While I was writing this article after work, a full-time engineer walked by, and he saw that I was writing something about A-priori and Fp Growth. He said, "Interesting, but not realistic." He further explained that this algorithm doesn't take weighing under consideration. For example, what about a transaction with multiple same items? There are also more subtle conditions that are not included in this algorithm. And that's why companies won't implement this in their business."

That conclusion is not in academic language, but it concludes the real-world experience with AR. The author's thesis wants to address that issue, that it is not frequently implemented in real business, and especially proposes a solution for a consideration of transactions transaction with multiple same items.

Frequency is only one part, but not the ultimate driver for business. The interference of profit and frequency could be solved by post-application of associative rules when connecting all frequencies with profit from articles later. However, in traditional association rules (TAR), we have a **first problem**: we already sorted out the least frequent items and lost crucial relations.

Moreover, so this approach would not be that straightforward. Instead, the author suggests considering profit already within the algorithm as decision criteria, whether to keep the item for getting relational rules.

The second problem not considered in TAR is that items can occur multiple times within one transaction. The support counting does not work in such a scenario. Even such a scenario is dependent on the nature of items quite frequently.

The author suggests adding the profit of these items and dividing by the unique count of these items, leading to a new average profit per item. In such a way, we consider these weights in the profit but not in the count, keeping the AR's primal structural integrity based on simple counts of each item per transaction.

A third problem poses the question: Why should we consider all transactions with the exact counting if more recent items have higher relevance than older items?

The author suggests introducing a date-decay function to consider the relevance of more recent items.

The fourth and last problem the author addresses is algorithm efficiency. We are primarily interested in the itemsets themselves; instead of repetitions leading to the same items. Even the improved FP-Tree algorithm must loop through the whole dataset repeatedly to find the exact relation between the identified itemsets branches.

The author suggests that by simplifying counting directly from the original tree structure (one loop through the whole dataset), we can derive the complete itemsets already with the most count to the minor count of items (1 path). Mostly the other paths

are repetitions and do not have many gains. That approach will make the algorithm super fast, with some loss of information (loss of other less relevant paths).

2 Discussion of related works

AR was popularized and, in effect, introduced to the public due to a 1993 article by Agrawal et al. [1]. The articles describe the advances in barcode technology and the recording of transactions. Before that, a computer-recorded transaction was not even possible. According to Agrawal et al. [1], the analysis aims to "how to place merchandise on shelves to maximize the profit." They introduced a formal association rule model.

Interestingly this model has not named an A-priori algorithm yet but has been labelled as such in a paper one year later by Agrawal, R. and Srikant, R. [18]. Fundamental terms like antecedent, consequent, resulting itemsets and support are introduced. Agrawal et al. [1] already optimized how to find association rules. Instead of finding all possible candidates for an association rule, it can be defined if at least one of the children (paths) of a new itemsets candidate is not frequent, the resulting itemsets will for sure not be frequent anymore, limiting the number of possible itemsets candidates and improving the performance. Agrawal et al. [1] make the following limitation: "The algorithm proposed in this paper is targeted at discovering qualitative rules. However, the rules we discover are not classification rules. We have no pre-specified classes. Rather, we find all the rules that describe the association between sets of items." In the author's opinion, that is a crucial statement because in finding relevant items, the dataset should be pre-processed in categories to find the relevant connection in each classification; otherwise, the algorithm would not be able to handle the sheer mass of associations, if the support-parameter is chosen too low. Details about the fundamental A-priori algorithm will be presented in chapter 3.

Some years later, in a paper called "Association Rules Mining: A Recent Overview "Kotsiantis, S. and Kanellopoulos, D (2006) [5] state: "In many cases, the algorithms

generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such a large number of complex association rules, thereby limiting the usefulness of the data mining results." They state that there are four main directions of research to address this issue:

1. by reducing the number of passes over the database
2. by sampling the database
3. by adding extra constraints on the structure of patterns
4. by parallelization.

The first point is addressed by the much more efficient FP-growth algorithm developed by Han, J. and Pei, J. (2000) [6], breaking the bottleneck of the A-priori algorithm. Toivonen, H, has developed on the second point. A sample is used to represent the whole dataset. Therefore, the choice of support level has to be much lower to find the same association rules as in the whole dataset. However, the main problem here is what sampling method should be chosen. Sampling Error Estimation (SEE) plays a key role and has to match a certain confidence level. The third point can be implemented as pre-processing the source file, filtering during the algorithm, or post-processing. The key in all methods is to find the "relevant patterns." Rapid Association Rule Mining (RARM) is an association rule mining method developed by Das, A. (2001) [8] is an association rule mining method that uses the tree structure to represent the original database and avoids the candidate generation process. According to Das, A., the bottleneck is the 2-candidate itemsets because the number of candidates in A-priori are all combinations of items that are greater than the minimum support, meaning

Num of 2item candidates=n*(n-1)

Das, A. (2001) [8] states that the algorithm is up to 100 times faster than the A-priori algorithm. It can be seen as a predecessor of the FP-growth algorithm, entirely relying on a tree structure. The fourth point is the most technical and is used to improve an

already existing algorithm; there is much research on parallelizing A-priori and FP-growth algorithms. That is especially important when using the algorithm in production and strict time and efficiency constraints.

In a paper by Slimani, T. and Lazzez, A. (2014) [9], different association rule mining approaches and variants are discussed. It shows that research in association rules was and is quite active. When we look at the development of association rules, all early algorithm variants try to increase the efficiency of the original A-priori algorithm by reducing the number of passes over the database like A-prioriTID (1994), DHP (1995), FDM (1996), GSP (1996), DIC (1997) and PincerSearch (1998). Further, CHARM (1999), Depth-project (2000), Eclat (2000) [20] and FP-growth (2004) [6] are all based on a tree-based search algorithm improving the efficiency significantly. With FP-growth, the peak of algorithm efficiency is nearly reached; therefore, the research for efficiency in a general algorithm is reduced. The author uses the FP-growth algorithm as a baseline algorithm for his modifications. Only three algorithms are widely known and implemented by common programming language libraries like mlxtend [16]: A-priori, Eclat and FP-Growth.

Very innovative is the "Sporadic Rules algorithm" proposed by Koh, S. Y. and Rountree, N. (2005) [19]. It proposes finding the not frequent rulesets by defining the maximum support instead of defining the minimum support but still having high confidence. In their paper, they suggest that this method can be used, for example, to find a rare association of two symptoms, which leads to the recognition of a rare disease.

We see that there are a vast number of different approaches. However, all can be categorized in one or several of the four main directions to address the performance issue and find the relevant itemsets and rules.

Recent papers try to generate association rules based on DL. 10. Patel H. K. and Yadav K. P. (2022) [10] show the increased efficiency of an autoencoder algorithm called DAENMF-ARM, a" denoising autoencoder and non-negative matrix factorization based

on association rule mining" to increase the efficiency compared to the classical baseline algorithms A-priori, Eclat, and FP-growth.

Al Shehabi S. et al. (2021) [11] introduce an algorithm called MARC to extract association rules, which is based on a Multi Self-Organizing Map (MultiSOM). The goal is to increase efficiency and only find relevant association rules.

The author spent some time rebuilding some new metrics suggested by Bao, F. et al. [12]. They suggest introducing new association rules and measures for validation, called Bi-support, Bi-lift, Bi-improvement, and Bi-confidence. They consider the correlation of non-support. If item A has a support of 1000 and item B has a support of 100, and they have joint support of 100 as itemsets, then item A is just a very popular overall item but has no direct correlation to B. That can be found with these new measures.

3 Baseline algorithm

3.1 Simple dataset

For demonstrating the method of the basic associative algorithm, a simple dataset, the author uses the following dataset. The dataset is has chosen to be simply enough to explain the concept of the algorithms. The dataset consists of 10 items, denoted as I=Item and 10 Transactions as T=Transaction .

Index	I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I
T_0	1	1	1	1	1	1	0	1	0	0	7
T_1	1	0	1	1	1	1	1	1	0	0	7
T_2	1	0	1	1	0	0	0	0	1	0	4
T_3	1	1	0	0	1	0	1	1	1	1	7
T_4	1	1	0	0	1	0	0	0	1	1	5
T_5	1	1	0	0	0	0	0	1	1	1	5
T_6	1	0	0	0	0	1	0	0	1	0	3

T_7	1	1	0	1	0	0	0	0	0	1	4
T_8	0	1	1	0	0	0	0	0	1	0	3
T_9	0	1	0	0	1	0	0	0	0	1	3
T	**8**	**7**	**4**	**4**	**5**	**3**	**2**	**4**	**6**	**5**	**48**

Figure 1: Simple dataset

3.2 Formal model of association rules

The author takes the terms and general description of the association rules model by Agrawal et al. [1].

Let I be the itemsets

Let I0, I1, . . . , I9 $\subset$ I

Each item is binary. If an item is in the itemsets, it is denoted by 1; if missing by 0.
Let T be the database/dataset of transactions.

Let T0, T1, . . . , T9 $\subset$ T

Let X be a set of items in each transaction.
A transaction T_n satisfies X if all X items commonly occur in T_n.

3.2.1 Support

The number of elements of X is 10.

1. If {I0 }$\subset$ X .

T satisfies X in 8 transactions.
Therefore the support of X is 0.8.

2. If {I1 }$\subset$ X .

T satisfies X in 7 transactions.
Therefore the support of X is 0.7.

3. If $\{I0, I1\} \subset X$.

T satisfies X in 5 transactions.

If X consists of at least two items, we have two paths to get to itemsets X. Agrawal et al. [1] name the existing itemsets "antecedent" and the newly added item "consequent."

3.2.2 Paths:

Suppose we have an itemsets $\{I0, I1\} \subset X$ with support 2. The number of paths leading to an itemset is factorial n!. Therefore for two items, there are two paths.

First path:

$I0 \rightarrow I1$, where I_0 is the antecedent, and I_1 is the consequent, building the itemsets X.

Second path:

$I1 \rightarrow I0$, where I_1 is the antecedent, and I_0 is the consequent, building the itemsets X.

3.2.3 Confidence:

The metric confidence is calculated by dividing the itemsets' support by the antecedent's support.

First path's confidence:

Support of $I0 = 0.8$; Support of $X = 0.5$

$\text{Confidene Path 1} = \frac{XI0}{I0} = \frac{0.5}{0.8} = 0.625$

Second path's confidence:

Support of $I1 = 0.7$; Support of $X = 0.5$

$\text{Confidence Path 1} = \frac{XI1}{I1} = \frac{0.5}{0.7} \approx 0.714$

We see there is not the same confidence for both paths, even if they lead to the identical itemsets X.

3.2.4 Application of association rules to the simple dataset

We can construct and filter out all association rules depending on the itemsets' minimum support and confidence.

Suppose we want to find all itemsets with:

Support ≥0.5; Confidence ≥0.5

Index	Itemsets	Support
0	(I_0)	0.8
1	(I_1)	0.7
2	(I_8)	0.6
3	(I_4)	0.5
4	(I_9)	0.5
5	(I_0, I_1)	0.5
6	(I_0, I_8)	0.5
7	(I_1, I_9)	0.5

Figure 2: Simple dataset - support

For illustration, the author chose enough high support min_support of 0.5 to filter only six frequent itemsets. If there is no min_support, there will be 308 itemsets. No restrictions on confidence are applied.

Generated association rules:

Index	Itemsets	Support	Ant.	Ant. Supp.	Cons	Cons. Supp.	Confidence
0	(I_0, I_1)	0.5	(I_1)	0.7	(I_0)	0.8	0.714
1	(I_0, I_1)	0.5	(I_0)	0.8	(I_1)	0.7	0.625
2	(I_0, I_8)	0.5	(I_8)	0.6	(I_0)	0.8	0.833
3	(I_0, I_8)	0.5	(I_0)	0.8	(I_8)	0.6	0.625
4	(I_1, I_9)	0.5	(I_9)	0.7	(I_1)	0.5	0.714

5	(I_1, I_9)	0.5	($_1$)	0.5	(I_9)	0.7	1.000

Figure 3: Simple dataset – association rules

When choosing confidence as min threshold, only itemsets of at least two items get filtered out because single items have no confidence. The confidence threshold has been chosen to be small enough to show all possible combinations of the three frequent itemsets with two items:

{I_0,I_1},
{I_0,I_8},
{I_1,I_9}

Two paths exist for each of the three itemsets of precisely two items; in total, 3 * 2 = 6 possible combinations.

If no minimum support is chosen, 308 itemsets build the basis for the rules. If no confidence threshold is chosen, these 308 itemsets generate 5,183 rules. Considering the straightforward dataset T, introduced in chapter 2, is overwhelming. That reflects problem 1 (relevance) and problem 4 (efficiency and order) described in the introduction.

3.3 A-priori

A-priori, Eclat, and FP-growth, all the TAR algorithms, result in the same frequent itemsets and rules. There is no difference in the outcome, but there is a difference in the efficiency. A short example shall illustrate the difference between the least efficient A-priori Algorithm by Agrawal, R. and Srikant, R. [18] and the most efficient FP-growth algorithm by Han, J. and Pei, J [6].

The author recalculates the outcome from the Python library (A-priori) step by step.

Step 1:

Let us consider all frequent itemsets >= 0.5 minimum support.

{I_0} Support: 0.8

{I_1} Support: 0.7

{I_4} Support: 0.5

{I_8} Support: 0.6

{I_9} Support: 0.5

Let us call that **List 1**

Notice: For the first round, one loop in the database is enough; count all occurrences of all items and compare with minimal support

Step 2:

Building all possible combinations out of List1 gives back the following candidate itemsets:

{I_0,I_1},{ I_0,I_4},{I_0,I_8},{I_0,I_9}

{I_1,I_4},{I_1,I_8},{I_1,I_9}

{I_4,I_8},{I_4,I_9}

{I_8,I_9}

Step 3:

Repeat Step 1 for candidate itemsets 1 >=0.5 minimum support

{I_0,I_1} Support: 0.5

{I_0,I_8} Support: 0.5

{I_1,I_9} Support: 0.5

Let us call that **List 2.**

Notice: Instead of simply looping once, we need to check all possible candidates to be a subset of the transactions for every transaction. If their list of candidates grows to several 1,000 and the database has over 1 million rows, it can become very inefficient and slow on an ordinary Desktop computer (without distributed calculations)

Step 4:

From List 2, we can create a new candidate set by joining all possible combinations with a length of itemsets three together.

{I_0,I_1,I_8},{I_0,I_1,I_9},{I_0,I_8,I_9},{I_1,I_8,I_9}

Step 5:

By pruning, which means we build all possible subsets with two elements (1 element less than the candidate itemsets), we can figure out if the new candidates already disqualify as not above minimum support. If one of the subsets is not in the already found frequent item list, the resulting itemsets will not be frequent either. This way, not all candidates have to be looped in the database.

Let Frequent = F, Not Frequent = NF
Subsets for {I_0,I_1,I_8}: {I_0,I_1}:F, {I_0,I_8}:F, {I_1,I_8}: NF -> Disqualified
Subsets for {I_0,I_1,I_9}: {I_0,I_1}:F, {I_0,I_9}:NF ,{I_1,I_9}: F -> Disqualified
Subsets for {I_0,I_8,I_9}: {I_0,I_8}:F, {I_0,I_9}:NF ,{I_8,I_9}: NF -> Disqualified
Subsets for {I_1,I_8,I_9}: {I_1,I_8}:NF, {I_1,I_9}:F ,{I_8,I_9}: NF -> Disqualified

All four candidate itemsets contain at least one subset, which is not frequent. Therefore, the A-priori algorithm will stop here and give back the frequent itemsets and rules. If there are no constraints to any other metrics, the possible rules per itemsets are factorial n!

If at least one candidate is left, steps 3-5 are repeated until no frequent itemsets can be found.

Time-complexity: Tahyudin I. et al. [13] show that the time-complexity of the A-priori algorithm is $O(2n)$, while "n" is the total number of all unique items above the minimum support threshold. We will have many problems if the minimum support is too low and all items are interconnected. Indeed in a real dataset, A-priori still can perform okay, but in the worst case, when all items are associated and a too-low minimum support has been chosen, it becomes nearly impossible to proceed with this algorithm.

3.4 FP-growth

The FP-growth algorithm is a big step toward a much more efficient solution of finding frequent itemsets. FP stands for "Frequent Pattern," structured in tree nodes, which grow with each new combination and are dynamically incremented.

The most significant change is that each candidate set has to be looped through the whole dataset. Important to notice is that both algorithms, A-priori and FP-growth, do not change the methodology of frequent itemsets and association rules but have different approaches to getting the result more or less efficiently.

Step 1 is the same as for A-priori:
Let us consider all frequent itemsets >= 0.5 minimum support.
{I_0} Support: 0.8
{I_1} Support: 0.7
{I_8} Support: 0.6
{I_4} Support: 0.5
{I_9} Support: 0.5
Let us call that **List 1**

Notice: In FP-growth, the outcome of the list is kept in strict order by count. If two items have the exact count, it does not matter, but most important, once fixed, the order is kept over the whole algorithm.

In Theory, it would be possible to have a random order, which is fixed, or even the most minor frequent item at the top, and it still would work and lead to the same result. However, we have an upside-down or bizarre-looking tree that would take away some of the efficiency and waste memory and computing power. Therefore there is no reason not to order it from most frequent to least frequent.

Step 2:
The root of the tree is denoted as null.

Building the tree

T0 ={ I0 , I1 , I2 , I3 , I4 , I5 , I7 }

T_0 is filtered by minimum support and ordered by the strictly ordered list of Step 1= T0 modified={ I0 , I1 , I4 }

Figure 4: FP-Tree after the first transaction

T1 ={ I0 , I2 , I3 , I4 , I5 , I6 , I7 }

T_0 is filtered by minimum support and ordered by the strictly ordered list of Step 1= T0 modified={ I0 , I4 }

Figure 5: FP-Tree after the second transaction

The T0 modified share the same node I_0. Therefore, this node is incremented by 1. I_4 is a direct child of I_0. This combination still does not yet exist and therefore builds a new branch.

After adding all transactions into the tree, the tree looks like the following:

Figure 6: Final FP-Tree after all transactions

Step 3:

A recursive function results in the following sets best gather the frequent itemsets:

I1:2 I1,I4:1 I1,I8:1 I0:8 I0,I8:1 I0,I4:1 I0,I1:5 I0,I1,I2:2 I0,I1,I4:1 I0,I1,I4,I9:1 I0,I1,I8:4 I0,I1,I8,I9:1 I0,I1,I8,I4:2 I0,I1,I8,I4,I9:1

Step 4:

From the itemsets in Step 3, we can build all the frequent itemsets.

For example, finding the association rule of I_1 -> I_9:

To find the support total of I_1, we sum up all generated sets of Step 3, which have I_1 as

the last item. These are

I1:2 ; I0,I1:5 → I1 =7

Afterwards, we sum up all generated sets of Step 3, which have I_1 in it and I_9 as the last item.

I0,I1,I4,I9:1 ; I0,I1,I8,I9:1 ; I0,I1,I8,I4,I9:1 → I9 =3

Therefore for AR I_1 -> I_9, the antecedent has the support of 7, and the common itemsets have the support of 3.

For finding the opposite association rule of I_9 -> I_1, in short:

I0,I1,I4,I9:1 ; I0,I1,I8,I9:1 ; I0,I1,I8,I4,I9:1 → I1 = 3

I0,I1,I4,I9:1 ; I0,I1,I8,I9:1 ; I0,I1,I8,I4,I9:1 → I9 =3

Therefore for AR I_9 -> I_1, the antecedent has the support of 3, and the common itemsets have the support of 3. There is no other I_9 besides all the sets in which A_1 is commonly present. Therefore, both the antecedent and the itemsets have the support of 3.

Time-complexity: The maximum time complexity of FP-growth is O(Number of items in header table * maximum depth of tree)= O(n*n) . [14] That is substantially better than O(n2) . In small datasets, A-priori and FP-growth will perform both well. However, as soon as there is an extensive dataset, FP-growth will outperform A-priori by far, and A-priori becomes "unusable."

4 Developing a modified Algorithm based on the FP-growth algorithm

4.1 FP-growth Model adjustments

4.1.1 Introduction

As discussed in chapter 2, there are four main directions of AR development [5]

1. by reducing the number of passes over the database

2. by sampling the database
3. by adding extra constraints on the structure of patterns
4. by through parallelization.

The author's modified FP-growth algorithm's adjustments belong to the second point. We mainly consider the transaction's date and the profit of the transaction during the algorithm. Question: Why does the author not consider it a pre- or post-processing step?

For the date-decay, it is thinkable to do it in a pre-processing step. The date-decay function will change the support of the transaction weight; the more current, the more weight. The function's curve can be adjusted flexibly by a lambda function argument. However, it is much handier to do it inside the algorithm for practical reasons; we only have to think about the parameters, like the date function. Post-processing with a date decay function is impossible because the date information will get lost in the itemsets and association rules. Moreover, date-decay support serves as a filtering criterion for the minimum support.

For the profit function, pre-processing of the data is thinkable, but then we do not have the author's so-called "associated profit" of itemsets, which can be a handy feature to see the overall profit of an item set. It is the most practical way to process the data inside the algorithm. Post-processing is not thinkable because profit-based support serves as a filtering criterion for the minimum support.

Let us discuss the adjustment steps in summary. The details of "the simple dataset" are shown in chapters 4.3 and 4.4. There is no subchapter for the single path modification, as explained in chapter 4.2. The modified algorithm & experiments can be observed in the author's GitHub references in chapter 8, Applications.

4.1.1 Date-decay-function

Step 1 – Parameters:

1. date_col:

The date column position selection, beginning with 0. 2, will be the third column of the

pandas' data frame.

2. max_date:

A dataset with a date range of 2022-01-01 until 2022-01-31 will be 2022-01-31.

3. date_range:

If a dataset has a date range of 2022-01-01 until 2022-01-31, it will be 365.

4. date_sensitivity

A lambda function must be given as an input values parameter [0,1]. For example

```
lambda x: 1 / (1+math.ex p-10*x+5 )
```

The date-decay function is processed inside the algorithm as follows:

Step 2 – Date-decay function:

```
for i in range(date_range):
    date_sensitivity(( date_range – i ) / date_range))
```

Therefore the input values will always result in a range [0,1]. The lambda function can be flexibly chosen for each dataset anew, if necessary.

Step 3 – Multiplying the resulting weights with transactions support:

The resulting weights are multiplied with the support of the transaction, directly manipulating the support. That will modify the standard AR methodology and interpretation. Therefore it should be treated with caution.

Time-complexity: The time complexity is the same as for FP-growth, which is O(Number of items in header table * maximum depth of tree)= O(n*n) [14]. It should run faster by showing only the most frequent single path. However, it does not change the time complexity fundamentally.

4.1.2 Profit function

In the author's opinion, the profit function is the more powerful and practical function because it solves the fundamental problem of having more than one of the same items in the same transaction.

Step 1 – Parameters:

1. profit_col:

Selection of the profit column position, beginning with 0. 3, will be then the fourth column of the pandas' dataframe.

2. max_profit:

That is the max profit of all items. That is used to convert each item's profit in percentage values.

3. profit_sensitivity

A lambda function must be given as an input values parameter [0,1]. For example

```
lambda x: x *1
```

This function will be embedded to:

```
profit_sensitivity(items_profit/max_profit)*max_profit
```

The lambda function is just a straight linear function and does not weigh the item's profit in any way. However, it could be thinkable to have a non-linear curve for special effects or finding rules, which otherwise could not be found.

Step 2 – Minimum profit for filtering out qualified items for association rules

The overall profit per item is used instead of support (frequency) to decide whether an item qualifies as a candidate. That is a primary trigger to find more relevant associated itemsets. Therefore an expensive item with fewer transactions will be qualified for association rules, which dropped out with TAR.

Minimum profit is calculated from the parameter support. Therefore:

```
minimum profit=support*overall profit
```

The overall profit will be diluted if additional profit and date-decay function are active:

```
diluted overall profit=overall profit*  date dependent supporttotal support
```

Step 3 – Calculating average profit:

The profit function is like a separate metric, summing up the profit by the lambda

function of each item in the first part of the algorithm. The difference to support summing up is that we sum up the transaction in the support. In the profit metric, we sum up each item, even the multiple items. However, we need a single count of each item once in a transaction to calculate its average later. For these reasons, we save both profit per item and single count per transaction per item in a dictionary.

In the end, we calculate the average profit like:

averageProfit= profitPerItem / singleCountPerItem

Step 4 - Associated profit:

For the FP-growth, we still use the support for TAR. The profit is only a second metric. We calculate the associated profit at the end as follows:

I=Item; AssociatedProfit=supportItemset* avgProfit I1+avgProfit I2+avgProfit I3

There is tension between supportItemsets and avgProfit. The more items we have an in a set, the higher the sum of avgProfit will be. However, most probably, the support for that new itemsets will shrink. However, it cannot be predicted if the resulting profit will rise or shrink; the change in profit of a new item added is an interesting metric to observe.

4.2 Modified FP-growth reconciliation with TAR

First, the author has to prove that the modified FP-growth, the author's implementation based on FP-growth, will give the same results as the traditional ones.

Input parameters in modified F-Growth algorithm:
1. data = T (Simple Dataset, introduced in chapter 3)
2. minSupRatio = 0.5
3. minConf = 0.5

Resulting in the following table:

Index	Itemsets	Support	Ant.	Ant. Supp.	Cons	Cons. Supp.	Confidence

0	(I_0)	0.8	-	-	(I_0)	0.8	-
1	(I_1)	0.7	-	-	(I_1)	0.7	-
2	(I_8)	0.6	-	-	(I_8)	0.6	-
3	(I_4)	0.5	-	-	(I_4)	0.5	-
4	(I_9)	0.5	-	-	(I_9)	0.5	-
5	(I_1, I_9)	0.5	(I_1)	0.7	(I_9)	0.5	0.714
7	(I_1, I_8)	0.5	(I_1)	0.8	(I_8)	0.6	0.625
8	(I_0, I_1)	0.5	(I_0)	0.8	(I_1)	0.7	0.625

Figure 7: Result of modified FP-growth (TAR parameters)

Generally, the author does not modularize frequent itemsets as in the popular FP-growth library from mlxtend [16] and rules but gives back both attributes at once. One difference, which can be observed, is that we have only three rules for frequent itemsets with two items instead of 6 rules. Per frequent itemsets, the descending order of total count is taken to show the "most relevant path."

I1,I9 : Path1 : I1 → I9 :s h own, Path2 : I9 → I1 :not s h own I0,I8 : Path1 : I0 → I8 :s h own, Path2 : I8 → I0 :not s h own I0,I1 : Path1 : I0 → I1 :s h own, Path2 : I1 → I0 :not s h own

That restriction is a conscious decision. One can argue that it takes away information. However, if we focus on solving business problems and later on focus on profit, this will have the following advantages:

- Generally, a more efficient algorithm
- Clearer rules. If there are frequent datasets with many items, the factorial n! will be shown because all subsets are frequent, too. That will spam the conscious analysis and take away the focus. If we do not have minimum support, 308 frequent itemsets will generate 5,183 rules for this specific dataset. Instead, in the modified algorithm, we show only the most relevant path, 308 rules for 308 item sets.

- With additional elements such as profit metrics, it will become chaotic otherwise and take away the essence and focus.

4.3 Modified FP-growth with date decay

The author has a second suggestion, how to improve the traditional association rules. This suggestion refers to the date decay function. We can assume that more recent transactions may have a higher relevance than older ones. That is especially true for products with a very short lifecycle, like consumer electronics. We have two options if the transaction period is two years and the typical product lifecycle is only one year. Either cut the transactions to adjust to the product lifecycle or a smarter one, to have a date decay function. The oldest items weigh 0.7, and the most recent items weigh 1. That is purely speculation, and we must research and find the ideal date decay function for every new dataset. There cannot be only one perfect date decay function for all datasets; instead, it must be calibrated for every dataset again. This date decay function must be implemented in the frequent items and association rules itself because later, it cannot be calculated anymore.

Let us assume the following date-enriched simple database for the observation:

Index	Items	date
T_0	$\{I_0, I_1, I_2, I_3, I_4, I_5, I_7\}$	2022-11-01
T_1	$\{I_0, I_2, I_3, I_4, I_5, I_6, I_7\}$	2022-11-02
T_2	$\{I_0, I_2, I_3, I_8\}$	2022-11-03
T_3	$\{I_0, I_1, I_4, I_6, I_7, I_8, I_9\}$	2022-11-04
T_4	$\{I_0, I_1, I_4, I_8, I_9\}$	2022-11-05
T_5	$\{I_0, I_1, I_7, I_8, I_9\}$	2022-11-06
T_6	$\{I_0, I_5, I_8\}$	2022-11-07
T_7	$\{I_0, I_1, I_3, I_9\}$	2022-11-08
T_8	$\{I_1, I_2, I_8\}$	2022-11-09

T_9	{I_1, I_4, I_9}	2022-11-10

Figure 8: Simple dataset, enriched with exemplary date

Let us see how it works with the simple dataset. T_1 has the date of 2022-11-01. T_{10} has the date of 2022-11-10. The range of dates is ten days. Now we take a lambda function. We take as an example a modified sigmoid and an extreme function showing date decay relevance.

lambda x: 1 / (1+math.ex p-10*x+5)

x will be between 0 and 1. 0 for the oldest date and 1 for the most recent date. Between this range, it applies the lambda. In this case, lambda looks like that.

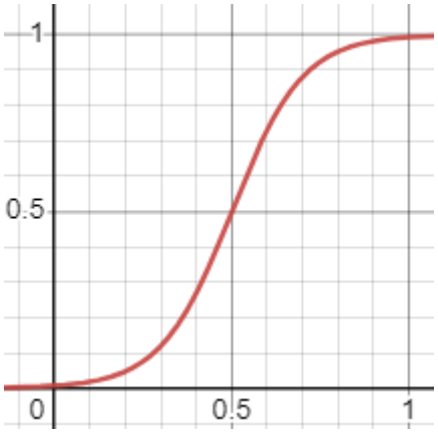


Figure 9: Function: f(x)= 11+e(-10x+5) , plotted in desmos [17]

We get the following weights for the transactions:

T9 Weight ≈0.9933 T8 Weight ≈0.9820 T7 Weight ≈0.9526 T6 Weight ≈0.8808 T5 Weight ≈0.7311 T4 Weight =0.5000 T3 Weight ≈0.2689 T2 Weight ≈0.1192 T1 Weight ≈0.0474 T0 Weight ≈0.0180

The function with the following parameters in the modified FP-growth:

1. data = T (Simple dataset enriched with date)
2. min_sup = 0.4
3. min_conf = 0.5
4. max_date = 2022-11-10

5. date_range = 10

6. date_sensitivity = lambda x: 1 / (1+math.ex p-10*x+5)

Index	Itemsets	Support	Ant.	Ant. Supp.	Cons	Cons. Supp.	Confidence
0	(I_1)	0.820	-	-	(I_1)	0.820	-
1	(I_8)	0.642	-	-	(I_8)	0.642	-
2	(I_0)	0.636	-	-	(I_0)	0.636	-
3	(I_9)	0.636	-	-	(I_9)	0.636	-
4	(I_1, I_9)	0.636	(I_1)	0.820	(I_9)	0.636	0.775
5	(I_1, I_8)	0.458	(I_1)	0.820	(I_8)	0.642	0.558
6	(I_0, I_1)	0.456	(I_1)	0.820	(I_0)	0.636	0.556
7	(I_0, I_9)	0.452	(I_0)	0.636	(I_9)	0.636	0.712
8	(I_0, I_1, I_9)	0.452	(I_0, I_1)	0.456	(I_9)	0.636	0.993
9	(I_0, I_8)	0.448	(I_8)	0.642	(I_0)	0.636	0.697

Figure 10: Result of modified FP-growth (Date decay parameters)

We see here that the equivalence on the transaction level is not destroyed. All items of the same transaction have the same weight. However, the frequent items and rules are now not an absolute frequency but an interpreted date decay frequency.

It seems very powerful; however, when not calibrated, especially on one database at a time, it will lead to misleading results. Therefore, it should be used carefully.

Later transactions are still frequent, while items like I_0, which was the most frequent in normal mode, are now ranked only as the third most frequent item and cannot build a frequent itemset with another item.

It is essential to mention that the date decay function dilutes the row count. We do not count every transaction anymore with 1, but the modified count after the date decay function.

4.4 Modified FP-growth with profit

Why does the author include profit as a metric inside the frequent items/ rules? Firstly, we can replace the decision with minimum profit instead of minimum support. There is somehow an equilibrium between more frequent, the same time cheaper, mostly higher margined items and expensive, less frequent, and lower margined items. We do not get the most relevant business rules if we consider frequency a primary driver.

The first use-case explains why profit as a metric will improve association rules:

1'000 items A with a profit of 10 equals 100 items B with a profit of 100. However, in traditional rules, one would never have a frequent association rule between both items because the minimum support is probably chosen high enough that item B drops out. However, in sight of profit, they are equal, and in the author's opinion, they should both be selected. Instead, let us think about a very frequent item C, with 100,000 occurrences and 0.1 profit per item. Should it be selected? Most probably, it is irrelevant. In traditional association rules, it will spam everywhere and have "not relevant" associations.

Secondly, we solve a fundamental problem. Usually, TAR cannot handle transactions like:

T1 :{A,A,A,A,A,A,A,A,A,A,B} T2 :{A,A,A,A,A,A,B} What would be the TAR outcome?

T1 :{A,B} T2 :{A,B}

In TAR, the outcome would be very falsified. We could build subgroups for frequencies of items and create new items for these combinations, but that would dilute the item count.

If we consider profit as a second metric, we count it once as in TAR, but we add the profit to the item's overall profit; at the end, we calculate a new average profitability and create an optimal representation of the item's weight, not in the frequency of count,

but in profit; it is like a shift from the strict count to the dynamic profit. It does not destroy association rules but represents their weight in a higher average profit.

For example, if one item always is bought five times, then at the end, the item's profit is five times higher for the single item, representing that it is, on average, bought five times in association rules

Surely in edge cases applying an average could be problematic if it is not representative; however, it is still better than simply not considering and counting only once.

Let us see how it works with the simple dataset.
The author enriched the items with profit data, I_0 with 10, I_1 with 20, ..., I_9 with 100, so that the simple dataset after implementing the profit looks like that:

Index	Items	profits
T_0	{I_0, I_1, I_2, I_3, I_4, I_5, I_7}	{10, 20, 30, 40, 50, 60, 80}
T_1	{I_0, I_2, I_3, I_4, I_5, I_6, I_7}	{10, 30, 40, 50, 60, 70, 80}
T_2	{I_0, I_2, I_3, I_8}	{10, 30, 40, 90}
T_3	{I_0, I_1, I_4, I_6, I_7, I_8, I_9}	{10, 20, 50, 70, 80, 90, 100}
T_4	{I_0, I_1, I_4, I_8, I_9}	{10, 20, 50, 90, 100}
T_5	{I_0, I_1, I_7, I_8, I_9}	{10, 20, 80, 90, 100}
T_6	{I_0, I_5, I_8}	{10, 60, 90}
T_7	{I_0, I_1, I_3, I_9}	{10, 20, 40, 100}
T_8	{I_1, I_2, **I_8, I_8, I_8**}	{20, 30, **90, 90, 90**}
T_9	{I_1, I_4, I_9}	{20, 50, 100}

Figure 11: Simple dataset, enriched with exemplary profit

The author modified the simple dataset with three times I_8 in Transaction T_9. Let us see how we handle this.

We use the following parameters:

1. data = T (Simple dataset enriched with profit)
2. min_sup = 0.1
3. max_profit = 100
4. profit_sensitivity = lambda x: 1*x

That will give us back the following result:

Index	Itemsets	Support	Profit associated	Perc of total profit	Cons.	Add. Profit Cons.	Decr. Profit Cons.	Net Change Cons.
0	(I_8)	0.6	720	0.276	(I_8)	720	0	720
1	(I_8, I_9)	0.3	660	0.253	(I_9)	300	-360	-60
2	(I_7, I_8, I_9)	0.2	600	0.230	(I_7)	160	-220	-60
3	(I_9)	0.5	500	0.192	(I_9)	500	0	500
4	(I_7, I_8)	0.2	400	0.153	(I_7)	160	-480	-320
5	(I_7, I_9)	0.2	360	0.138	(I_7)	160	-300	-140
6	(I_7)	0.4	320	0.123	(I_7)	320	0	320

Figure 12: Result of modified FP-growth (Profit parameters)

The minSupRatio is not anymore the minimum frequency but is calculated in this way min_ sup: min _supRatio*total_profit It can be interpreted as the most negligible percentage of the profit of a single item or association we still see as relevant. The descending order is defined according to the profit associated. The count and FP-tree construction are strictly according to traditional rules. We do not break the algorithm but enrich it.

Let us recalculate the profit associated with I_8.
The count of I_8 is 6. That is the same as in TAR.
In T_9, we have the item 3 times.
In total, the item count is, therefore, 8.

The item's profit is 90.

Let us recalculate the average profit of the I_8:

I8 avg:itemOriginalProfit* itemTotalCountitemSingleCount =90* 86 =120

For I_8, as a single item without any other items combined, we have counted six and an associated profit of 720 (=6 *120). That represents the "real weight" of the item by profit; in frequency alone, it is not reflected.

Secondly, let us have a look at frequent itemset

{ I7 , I8 , I9 } .

The support for the itemset is 2.

That would fall out in the TAR because of low frequency. However, when we consider the combined associated profit, this combination will still be relevant for business. Associated profit=supportItemset* avgProfit I7+avgProfit I8+avgProfit I9 =2* 80+120++100 =600

Profit net_change represents the newly associated profit minus the total before the newly added item. If the new association has a higher profit than the previous one, it shows synergies. If it is lower, it is potentially nothing to focus on to " maximize" the overall profit. In our simple artificial dataset, all net changes are negative; however, this is not a must; there could be positive net profit changes by extending the itemset.

4.5 Modified FP-growth with combined date decay and profit

When we combine the date decay and the profit function, we have the full range of parameters:

1. data = T (Simple dataset enriched with profit and date)
2. min_sup = 0.1
3. min_conf = 0.5
4. max_profit = 100
5. profit_sensitivity = lambda x: 1*x

6. max_date = 2022-11-10

7. date_range = 10

8. date_sensitivity = lambda x: 1 / (1+math.ex p-10*x+5)

Index	Itemsets	Support	Profit associated	Perc of total profit	Cons.	Add. Profit Cons.	Decr. Profit Cons.	Net Change Cons.
0	(I_8)	0.642	417.8	0.361	(I_8)	417.8	0.0	417.8
1	(I_9, I_1)	0.636	413.5	0.357	(I_9)	344.6	-20.0	324.6
2	(I_9, I_1, I_8)	0.636	360.0	0.311	(I_9)	150.0	-137.5	12.5
3	(I_1, I_8)	0.642	347.5	0.300	(I_8)	297.8	-39.2	258.6
4	(I_9)	0.636	344.6	0.298	(I_9)	344.6	0.0	344.6
5	(I_7, I_9, I_1, I_8)	0.197	320.0	0.277	(I_7)	80	-120	-40.0
6	(I_7, I_9, I_8)	0.197	300.0	0.259	(I_7)	80	-110	-30.0
7	(I_4, I_9, I_1)	0.337	299.6	0.259	(I_4)	88.1	202.0	-113.9
8	(I_4, I_9)	0.337	264.3	0.228	(I_4)	88.1	-168.4	-80.3

9	(I_4, I_9, I_1, I_8)	0.337	223.0	0.193	(I_4)	38.4	-175.4	-137.0
10	(I_4, I_9, I_8)	0.33	207.6	0.179	(I_4)	38.4	-160.8	-122.4
11	(I_4, I_9, I_1, I_7, I_6, I_8)	0.058	118.3	0.102	(I_6)	18.8	0.0	18.8

Figure 13: Result of modified FP-growth (Profit & Date-decay parameters)

According to classical TAR, the support (calculated) will be the same as in chapter 4.2. However, the fundamental selection of the "min support" items is chosen according to 4.3. Therefore, we do not find the same elements anymore. At the same time, the profit view gives us the great advantage of equalizing multiple items over the profit measure. The essential part and maybe a disadvantage of the custom date decay and profit lambda functions is the correct representative function for each dataset. Profit at easiest and most logically will be linear; however, there could be other means of measurement, which do not strictly measure profitability as such, but like a modified profit as a means of marginal profitability; for example, as a decreasing curve.

4.6 Performance, Strengths, and weaknesses of the author's modified FP-growth algorithm

The author checked the performance of the cosmetics dataset with a limitation of transactions no longer than five items as follows, and 77'264 (155'617 original) transactions, 40'777 unique items, support of 0.0001 (resulting in minimum support of 16) with 2'567 found itemsets/rules. Please notice that loading the prepared dataset in memory (dataframe) is omitted, which takes a big part of the work with smaller datasets.

Modified FP-growth (normal parameters): 14.1 seconds
Modified FP-growth (profit parameters): around the same, 14.3 seconds
FP-growth library from mlxtend: 5.0 seconds
A-priori library from mlxtend: A-priori failed with minimum support of 0.0001 due to insufficient memory in the pandas dataframe. After decreasing minimum support to

0.001, it succeeded but needed 142.2 seconds, while the same, minor minimum support in FP-growth needed only 3.1 seconds.

We see that in a real-world dataset, A-priori becomes unusable and can fail due to memory limitations. That is especially true with the diverse and rich cosmetics dataset, with many different, low-frequent items.

The author has to admit that the implementation of the FP-growth library from mlxtend is highly efficient, and the own implementation lacks some code optimizations. In academia, the A-priori algorithm is much easier to grasp and modify to explain fundamental concepts. However, without digging deeper into performance, it can be concluded that every real-world algorithm for business should be mandatorily implemented as a variant of the FP-growth algorithm.

The strengths of the modified FP-growth algorithm see the author as follows:

- The single-path approach increases efficiency and gives more clarity about the association rules, meaning obsolete rules are not repeated, but the "most relevant" path is shown.
- The date-decay function is a feature that can be used in specific datasets; however, it is not recommended to use it in general. It is a highly specialized modification, which can make sense. The usefulness is, therefore, somewhat limited.
- Profit-function solves the problem of the correct weight of multiple items in one transaction; which otherwise gets lost
- By the profit function, in the author's opinion, we have better filtering criteria to find relevant datasets and associations. Moreover, the associated profit of itemsets is an interesting metric for profit-driven businesses.

The weaknesses of the modified FP-growth algorithm see the author as follows:

- When the author speaks of profit, he assumes 10% of revenue. That simple approach may be enough; revenue can be chosen instead of profit. However, to

find the actual profit, the author would recommend determining the average profit per item category; in such a way, it will be better and more powerful. However, without insider data to an external dataset, it is nearly impossible to assume correct profit margins; or it is very time-consuming, which would not be worth it.

- In the date-decay function, we manipulate the support, and the meaning of support becomes different. Support of 1 does not mean one transaction but could be three transactions together, weighted by the date-decay function. Therefore that modification should be applied only for exceptional cases.
- We lose association rules by finding only one potentially relevant path. However, there are pros and cons; situational, however, it could have disadvantages
- The data preparation to get profit and date-decay functionality need more data pre-processing in the pandas dataframe; it has more parameters, meaning it is more difficult to use.

5 Experiments/ Analysis of big datasets

Let us experiment with the same datasets and compare the new profit-based AR algorithm with the traditional one.

5.1 E-Commerce purchase history from an electronics store

5.1.1 Dataset

Let us observe the dataset "E-Commerce purchase history from electronic store" [2] from Kaggle. The dataset author [2] states: "This dataset contains purchase data from April 2020 to November 2020 from a large home appliances and electronics online store."

The dataset has 2,633,521 rows. It has 1,435,266 transactions. Thereof are 25,113 unique items. 649 transactions contain the same item multiple times. The TAR is not

able to solve this discrepancy. However, with our improved algorithm, we can balance it out by the profit measure. The distribution of multiple items per transaction is shown below:

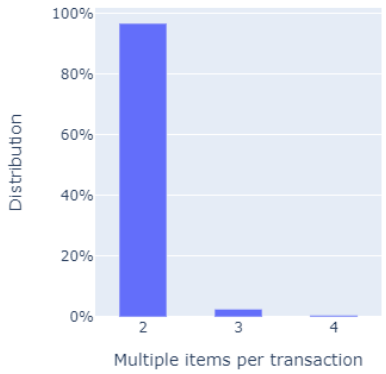


Figure 14: Electronics store – Distribution of multiple items per transaction

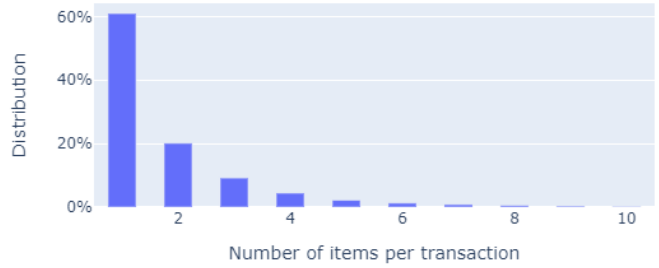


Figure 15: Electronics store – Distribution of the number of items per transaction

In the dataset, more than 60% are single-item transactions, therefore not relevant in finding association rules. However, their support still counts inside the minimum support. Around 20% of the dataset are 2-item transactions, and around 10% are 3-item transactions; the residual 10% is split in small numbers until 60 items of transactions are outliers.

5.1.2 Analysis

An essential part and maybe a disadvantage of the custom date decay and profit lambda functions is the correct representative function for each dataset. Profit at easiest and most logically will be linear; however, there could be other means of measurement,

which do not strictly measure profitability as such, but like a modified profit as means of marginal profitability; for example, as a decreasing curve. The dataset is not the most ideal for finding association rules. One reason could be that it is an online store, and most items are "higher priced." A client does not buy goods like in a grocery store. However, it is an excellent start to analyze. Later, let us compare a grocery store database and see the differences.

itemsets	support	product_id	category	brand	price
0	44,491	1515966223523303302	57.87	-	-
1	41,076	1515966223523303301	16.18	-	-
2	38,786	1515966223523303308	29.63	-	-
3	38,472	1515966223523303310	8.10	-	-
4	36,530	1515966223523303314	9.26	-	-
5	31,652	1515966223523303321	16.18	-	-
6	26,080	1515966223523303310	8.10	-	-
6	26,080	1515966223523303314	9.26	-	-
7	19,844	1515966223523303301	16.18	-	-
7	19,844	1515966223523303321	16.18	-	-
8	15,994	1515966223523303312	6.94	-	-
9	14,284	1515966223509117074	-	samsung	30.07

Figure 16: Electronics store – TAR - Top overall ten itemsets

If we observe Figure 16, its data is either incomplete or does not have a price. We generally can say that these are highly frequent, but generally low prices articles. Are we interested in them? The author would assume they do not drive the business and are like "spam" in recognizing the relevant rules.

itemsets	support	product_id	category	brand	price
0	4,336	1515966223509088567	electronics.smartphone	apple	856.23
1	2,587	1515966223509088671	electronics.smartphone	apple	925.67
2	1,967	1515966223509089284	electronics.smartphone	samsung	1215.25
3	6,466	1515966223509088532	electronics.smartphone	samsung	300.9
4	2,242	1515966223509088628	electronics.smartphone	samsung	856.46
5	4,502	1515966223509130879	appliances.kitchen.washer	samsung	381.92
6	12,058	1515966223509106786	electronics.smartphone	samsung	138.87
7	8,692	2273948218662322995	-	grohe	184.72
8	1,778	1515966223509089438	electronics.smartphone	apple	856.23
9	2,747	1515966223509088610	electronics.smartphone	samsung	532.38

Figure 17: Electronics store – Profit-based modified FP-Growth - Top overall ten itemsets

In Figure 17, the only familiar item in line with traditional association rules is the item with rule 6. It is a Samsung smartphone with a relatively low price but high frequency, which is still relevant in total profit. Generally, when we observe the list, it looks much

more relevant than the previous one. According to our expectation, when we would not know anything about consumer electronic goods, smartphones of different brands will be the top sellers/ top articles for stores. That already has some proof that the method with profit considered is advantageous.

itemsets	support	product_id	category	brand	price
0	26,080	1515966223523303310	8.10	-	-
0	26,080	1515966223523303314	9.26	-	-
1	19,844	1515966223523303301	16.18	-	-
1	19,844	1515966223523303321	16.18	-	-
2	12,109	1515966223523303310	8.10	-	-
2	12,109	1515966223523303312	6.94	-	-
3	10,401	1515966223523303314	9.26	-	-
3	10,401	1515966223523303312	6.94	-	-
4	9,031	1515966223523303310	8.10	-	-
4	9,031	1515966223523303312	6.94	-	-
4	9,031	1515966223523303314	9.26	-	-
5	8,726	1515966223523303310	8.10	-	-
5	8,726	1515966223523303308	29.63	-	-
6	8,304	1515966223523303314	9.26	-	-
6	8,304	1515966223523303301	16.18	-	-

Figure 18: Electronics store – TAR - Top overall seven itemsets with more than 1 item

In Figure 18, unfortunately, there are not many insights about the products because of missing brands and prices. Maybe it could be transportation costs or VAT taxes; the categories are listed as numbers. In the context of some special investigation, these items with pre-knowledge could make sense for analysis. However, with our missing insight, a deeper analysis is pure speculation. In conclusion, we generally do not want to find these items in association rules. They are not interesting to us.

itemsets	support	product_id	category	brand	price
0	3,863	2273948218662322995	-	grohe	184.72
0	3,863	2273948218662322996	-	grohe	143.98
1	1,007	2273948218662322995	-	grohe	184.72
1	1,007	2273948218662322996	-	grohe	143.98
1	1,007	2273948186248741817	-	incase	97.20
2	1,386	2273948218662322995	-	grohe	184.72
2	1,386	2273948186248741817	-	incase	97.20
3	1,001	1515966223509088532	electronics.smartphone	samsung	300.90
3	1,001	1515966223509117074	-	samsung	30.07
4	359	1515966223509088567	electronics.smartphone	apple	856.23
4	359	2273948297037087396	-	goodride	48.82
5	1,605	1515966223509117074	-	samsung	30.07
5	1,605	2273948316473492113	electronics.smartphone	samsung	162.01
6	1,208	2273948218662322996	-	grohe	143.98
6	1,208	2273948186248741817	-	incase	97.20

Figure 19: Electronics store – Profit-based modified FP-Growth - Top overall seven itemsets with more than 1 item

In Figure 19, Itemsets 0 from Grohe are very popular. Grohe belongs to plumbing, a necessary, but rather a B2B good instead of B2C. The installation is rarely done by end customers but by plumbers. The combination of the incase item and 3 of the items is interesting. It seems like a plumber service company charges for the installation costs. That is a strong pattern. Itemsets 4, 5, and 6 are smartphones with a high price (primary product) and a complimentary product like a smartphone cover or car appliance for smartphones. It makes absolute sense to buy these items together. However, it makes sense that Smartphones are mostly bought alone (single items) and that two smartphones (basic) products are not bought together as a single transaction.

5.2 E-Commerce purchase history from a jewellery store

5.2.1 Dataset

The Kaggle provider (3) of this dataset states: "This dataset contains purchase data from December 2018 to December 2021 (3 years) from a medium-sized jewellery online store."

The dataset has 95,911 rows. It has 74,760 transactions. Thereof are 9,613 unique items. 2,298 transactions contain multiple items. The distribution of multiple items per transaction is shown below:

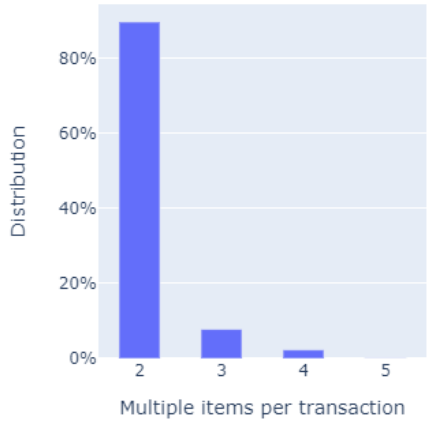


Figure 20: Jewellery store – Distribution of multiple items per transaction

The length of transactions is distributed like that:

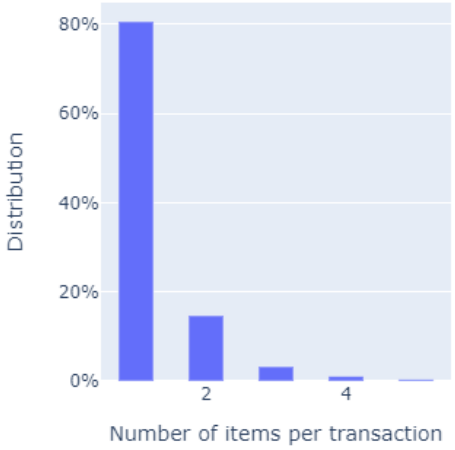


Figure 21: Jewellery store – Distribution of the number of items per transaction

In the dataset, more than 80% are single-item transactions, therefore not relevant in finding association rules. However, their support still counts inside the minimum support. Around 15% of the dataset are 2-item transactions, and around 3% are 3-item transactions; the residual 2% is split in small numbers until 27 items of transactions are outliers. Reasonably, this pattern is expected because jewellery is a luxury good and is not expected to be bought as would be goods in a grocery store.

5.2.2 Analysis

The second dataset is even worse than the first for finding association rules. The reason this time is quite clear. Jewellery is a luxury not that often bought in a single transaction.

itemsets	product_id	support	category	gender	metal	gem	price
0	1515966223577083202	468	jewellery.brooch	-	silver	-	10.27
1	1956663847725040099	362	jewellery.ring	-	gold	-	200.45
2	1956663840242401751	301	jewellery.ring	f	gold	-	259.97
3	1956663840309510725	274	jewellery.ring	-	gold	-	133.55
4	1956663830872326617	266	nan	-	gold	-	140.62
5	1956663847666319760	266	jewellery.ring	-	gold	-	188.70
6	1352907200745439279	265	jewellery.ring	-	gold	-	215.14
7	1956663836207481430	259	jewellery.ring	-	gold	-	119.77
8	1956663845787271453	258	nan	-	gold	-	445.95
9	1956663831283368958	253	jewellery.ring	-	gold	-	242.90

Figure 22: Jewellery store – TAR - Top overall ten itemsets

In Figure 22, all the itemsets are single items. We can observe that the most common items are relatively low priced compared to an average item price of 405.49 and in the lower half compared to the median of 261.60. Moreover, none of the rings has a "gem".

itemsets	product_id	support	category	gender	metal	gem	price
0	1806829191514031042	88	-	nan	gold	mix	2,780.68
1	1956663836668854462	95	-	nan	gold	nan	2,265.52
2	1956663836668854461	159	-	nan	gold	mix	1,258.51
3	1956663847708262860	107	-	nan	gold	mix	1,727.03
4	1313614230015967859	170	jewellery.earring	f	gold	diamond	903.97
5	1956663831425974504	111	-	nan	gold	mix	1,357.79
6	1956663831375642763	129	jewellery.ring	f	gold	diamond	1,020.41
7	1515966223500113689	5	jewellery.ring	nan	gold	mix	26,424.52
8	1944945390285488528	221	jewellery.ring	f	gold	diamond	534.11
9	1956663846449972073	146	-	nan	gold	nan	829.16

Figure 23: Jewellery store – Profit-based modified FP-Growth - Top overall ten itemsets

In Figure 23, the top 10 profitable itemsets are single-item transactions, the same as in TAR. However, the price is either on the upper side of the median or the opposite of TAR. The support on the other side varies, ranging from 5 to 221. The item with only five support is still relevant because of its high price of 26,425. The item with 221 support has the lowest price of 534. We see similarly that rings and earrings are popular, each in support and profitability. That is different from the electronic goods store in the

first analysis, where we found utterly different kinds of items. However, one must admit that the jewellery category range is small. That could bring a reasonable explanation.

itemsets	product_id	support	category	gender	metal	gem	price
0	1313614230015967859	51	jewellery.earring	f	gold	diamond	903.97
0	1944945390285488528	51	jewellery.ring	f	gold	diamond	534.11
1	1944422271028298685	34	jewellery.ring	f	gold	diamond	493.01
1	1937902773722939556	34	jewellery.earring	f	gold	diamond	856.03
2	1956663836207481431	29	jewellery.ring	-	gold	-	145.62
2	1956663836207481430	29	jewellery.ring	-	gold	-	119.77
3	1313614230015967859	26	jewellery.earring	f	gold	Diamond	903.97
3	1343446704099164925	26	jewellery.pendant	f	gold	Diamond	410.82
4	1956663836207481430	25	jewellery.ring	-	gold	-	119.77
4	1956663840309510725	25	jewellery.ring	-	gold	-	133.55
5	1956663848287077290	24	jewellery.earring	-	gold	amethyst	287.53
5	1956663848287077291	24	jewellery.ring	-	gold	amethyst	198.49
6	1956663846349308653	23	jewellery.ring	-	gold	diamond	328.63
6	1956663840242401751	23	jewellery.ring	f	gold	-	259.97

Figure 24: Jewellery store – TAR - Top overall seven itemsets with more than 1 item

In Figure 24, interestingly, the price of the itemsets pair has some positive correlation, and the price range of these item pairs is relatively homogenous. Unfortunately, we do not have more specific data about the product, but it seems like either the jewellery itemsets are in a collection of a set or maybe something very classical like wedding rings, a smaller woman's ring, and a bigger man's ring. We are missing the high-priced combinations in the traditional association rules, which are expected to be much lower in joint support but will be relevant because of the overall price. Moreover, we do not consider items bought in a quantity of more than 1. All found items are rings or earrings. It seems to be the most popular overall item.

itemsets	product_id	support	category	gender	metal	gem	price
0	1313614230015967859	51	jewellery.earring	f	gold	diamond	903.97
0	1944945390285488528	51	jewellery.ring	f	gold	diamond	534.11
1	1944422271028298685	34	jewellery.ring	f	gold	diamond	493.01
1	1937902773722939556	34	jewellery.earring	f	gold	diamond	856.03
2	1956663846340920038	16	jewellery.ring	-	gold	diamond	1,082.05
2	1956663845602721816	16	jewellery.earring	-	gold	diamond	1,198.49
3	1860421116053422140	18	jewellery.ring	-	gold	diamond	671.1
3	1845870160691331313	18	jewellery.earring	-	gold	diamond	1,280.68
4	1806829191514031042	7	-	-	gold	mix	2,780.68
4	1956663836668854462	7	-	-	gold	-	2,265.52
5	1313614230015967859	26	jewellery.earring	f	gold	diamond	903.97
5	1343446704099164925	26	jewellery.pendant	f	gold	diamond	410.82

6	1956663845736939738	17	jewellery.ring	-	gold	diamond	575.21
6	1956663836392031040	17	jewellery.earring	-	gold	diamond	1,294.38

Figure 25: Jewellery store – Profit-based modified FP-Growth - Top overall seven itemsets with more than 1 item

In Figure 25, the result is surprising. The results of TAR correlate quite a bit with profit-based AR. The top 2 itemsets are identical. Out of 7 itemsets, three are identical to AR. What pitches the eye are the itemsets with only seven support but are highly-priced. These items are relevant even if the support is low. Surely there should be a hard lower threshold of irrelevance. Below support of 5, maybe, even if the price is high, the relevance is diminished because of randomness. A rule would not be strict enough and be evident to predict this relation for future transactions.

In this jewellery dataset, the advantages of a profit AR are lower than in the consumer electronic stores. The author concludes this because of the much higher homogeneity of jewellery products.

5.3 E-Commerce purchase history from a cosmetics store

5.3.1 Dataset

The Kaggle provider [4] of this dataset states: "This dataset contains behaviour data for five months (Oct 2019 – Feb 2020) from a medium cosmetics online store. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relations between products and users." Originally, the dataset contained the events "view", "cart", "remove_from_cart", and "purchase". For the author's analysis, only purchases are considered.

The dataset has 1'287'007 rows. It has 155'617 transactions. Thereof are 40'777 unique items. 10'123 transactions contain multiple items. The TAR is not able to solve this discrepancy. However, with our improved algorithm, we can balance it out by the profit measure. The distribution of multiple items per transaction is shown below:

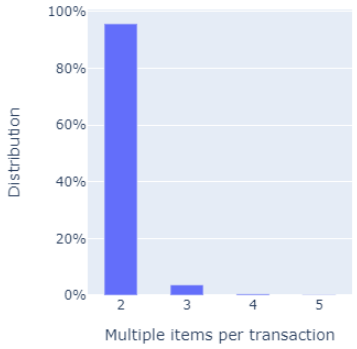


Figure 26: Cosmetics store – Distribution of multiple items per transaction

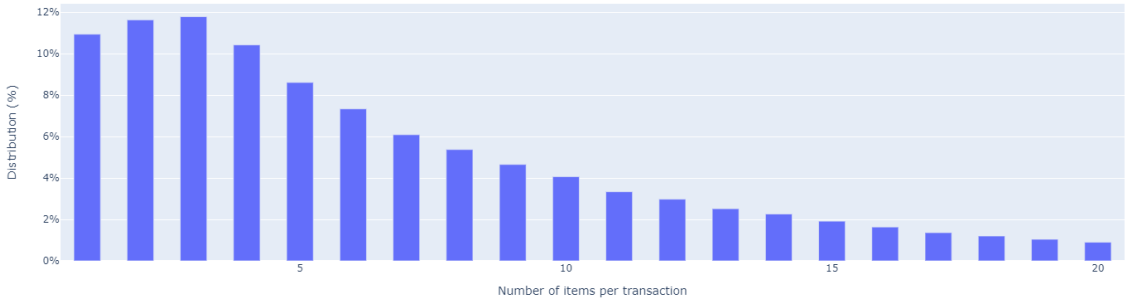


Figure 27: Cosmetics store – Distribution of the number of items per transaction

The cosmetics dataset is quite different from the previous ones. It has a relatively low number of single-item transactions. Cosmetics are most frequently bought together as a transaction set of 3 items with a frequency of around 11%. The curve is relatively flat, meaning there are a lot of combined items per transaction. That is ideal for finding association rules. One can expect that distribution because cosmetics are cheaper than electronics and jewellery. Moreover, in cosmetics, we will have many complementary products bought together.

5.2.2 Analysis

itemsets	product_id	support	category_id	brand	price
0	5809910	7,458	1602943681873052386	grattol	5.24
1	5854897	4,593	1487580009445982239	irisk	0.32
2	5700037	3,623	1487580009286598681	runail	0.40
3	5802432	3,486	1487580009286598681	-	0.32
4	5751422	3,457	1487580005268456287	uno	10.95

5	5809912	3,272	1602943681873052386	grattol	5.24
6	5815662	3,206	1487580006317032337	-	0.92
7	5304	3,102	1487580009471148064	runail	0.32
8	5751383	2,912	1487580005092295511	uno	10.32
9	5849033	2,753	1487580005092295511	uno	10.32

Figure 28: Cosmetics store – TAR - Top overall ten itemsets

In figure 28, none of the associated itemsets (2-itemsets) got into the top 10. The image varies, ranging from prices from 0.32 to 10.32, support from 2'753 (1.77%) to 7'458 (4.79%). For such an extensive dataset, these are relatively high numbers. The most popular item category from brand names seems to be nail polish and nail care products. From a price perspective, these items represent the median, which is 4.44 for the whole dataset, while the mean is 7.05

itemsets	product_id	support	category_id	brand	price
0	5560754	365	1487580006300255120	strong	194.44
1	5809910	7,458	1602943681873052386	grattol	5.24
2	5751422	3,457	1487580005268456287	uno	10.95
3	5751383	2,912	1487580005092295511	uno	10.32
4	5850281	209	1487580006300255120	marathon	137.78
5	5849033	2,753	1487580005092295511	uno	10.32
6	5792800	2,683	1487580005268456287	nan	10.32
7	5877454	603	1487580006300255120	jessnail	44.29
8	5560756	117	1487580006300255120	strong	207.94
9	89343	79	2193074740619379535	nan	299.81

Figure 29: Cosmetics store – Profit-based modified FP-Growth - Top overall ten itemsets

In figure 29, the price ranges from 5.24 to 299.81. Compared to TAR, the price and frequency variance is very big. Nevertheless, these mixed items have an equilibrium on profitability. The cheap product, with a price of 5.24, makes a similar revenue as the expensive product, with a price of 299.81.

itemsets	product_id	support	category_id	brand	price
0	5809910	1180	1602943681873052386	grattol	5.24
0	5809912	1180	1602943681873052386	grattol	5.24
1	5751422	837	1487580005268456287	uno	10.95
1	5751383	837	1487580005092295511	uno	10.32
2	5809910	756	1602943681873052386	grattol	5.24
2	5809911	756	1602943681873052386	grattol	5.24
3	5809910	569	1602943681873052386	grattol	5.24
3	5816170	569	1602943681873052386	grattol	5.24
4	5809912	546	1602943681873052386	grattol	5.24
4	5809911	546	1602943681873052386	grattol	5.24
5	5751422	529	1487580005268456287	uno	10.95
5	5849033	529	1487580005092295511	uno	10.32
6	5809912	411	1602943681873052386	grattol	5.24

6	5816170	411	1602943681873052386	grattol	5.24

Figure 30: Cosmetics store – TAR - Top overall seven itemsets with more than 1 item

In figure 30, the itemsets are very interesting. They are different products but have the same brands and prices. It can be assumed that these items are like a variant of each other, like a change of nail polish or lipstick colour. The prices are standard compared to the median and average; for shelf management, these products should be presented directly to each other or, in e-commerce, should be recommended in an online store.

itemsets	product_id	support	category_id	brand	price
0	5751383	837	1487580005092295511	uno	10.32
0	5751422	837	1487580005268456287	uno	10.95
1	5809912	1,180	1602943681873052386	grattol	5.24
1	5809910	1,180	1602943681873052386	grattol	5.24
2	5849033	529	1487580005092295511	uno	10.32
2	5751422	529	1487580005268456287	uno	10.95
3	5809911	756	1602943681873052386	grattol	5.24
3	5809910	756	1602943681873052386	grattol	5.24
4	3762	178	1487580005411062629	cnd	19.37
4	4185	178	1487580005411062629	cnd	19.37
5	5528034	340	1487580005553668971	nan	9.52
5	5528035	340	1487580005553668971	nan	9.52
6	5809910	381	1602943681873052386	grattol	5.24
6	5809912	381	1602943681873052386	grattol	5.24
6	5809911	381	1602943681873052386	grattol	5.24

Figure 31: Cosmetics store – Profit-based modified FP-Growth - Top overall seven itemsets with more than 1 item

In figure 31, Interestingly, profit-based AR delivers nearly precisely the same itemsets. Only the itemsets with 88 support and a higher price are different. One can assume that high-priced cosmetics are not bought in frequent combinations with other products, and regular products have a quite homogenous price, resulting in nearly the same outcome. In the case of cosmetics, we get, therefore, a similar result for both profit-based and TAR for itemsets with more than 1 item.

6 Conclusion

We saw much research done on association rules. However, the critical problem of successful implementation for businesses remains. The author addressed four fundamental problems to which he wants to suggest solutions.

1. Finding relevant rules based on a secondary profit metric instead of frequency

2. Consider Transactions with multiple same items
3. Date-Decay function as a feature for exceptional cases
4. Increase clarity of AR by reducing itemsets' AR to the most frequent path

The author showed in the experiments that stronger and more relevant rules could be found with the modified FP-growth algorithm. In the cosmetics store dataset, we found very similar top itemsets with TAR and modified FP-growth, while in the electronic good and jewellery, we found some overlapping but included the higher-priced, average frequent items and skipped the highly frequent but very low-priced items.

In the author's opinion, the experiments successfully showed an improvement in finding relevant and strong itemsets and association rules.

Future research, when implementing a profit metric as filter criteria and solving the multi-item transaction, can further focus on materializing and measuring the benefit of such an approach and further modify the function to standardize a proven approach for businesses in the real world. Moreover, implementing a more efficient modified FP-growth algorithm can be further optimized.

7 References

1. Agrawal, R., Imieliński, T., Swami, A. 1993. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207
2. E-Commerce purchase history from electronic store (Retrieved: 02.10.2022 from https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-electronics-store)
3. E-Commerce purchase history from jewellery store (Retrieved: 11.12.2022 from https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewellery-store)
4. E-Commerce purchase history from cosmetics shop (Retrieved: 12.12.2022 from https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop)
5. Kotsiantis, S., Kanellopoulos, D. 2006. Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering. Vol.32 (1), 2006, pp. 71-82
6. Han, J., Pei, J. 2000. Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter 2, 2, 14-20.
7. Toivonen, H. 1996. Sampling large databases for association rule. The VLDB Journal, pp. 134-145.
8. Das, A., Ng, W.-K., Woon, Y.-K. 2001. Rapid association rule mining. Proceedings of the tenth international conference on information and knowledge management. ACM Press, 474-481
9. Slimani, T., Lazzez, A. 2014. Efficient Analysis of Pattern and Association Rule Mining Approaches. International Journal of Information Technology and Computer Science (IJITCS), vol.6, no.3, pp.70-81, 2014

10. Patel, H, K., Yadav, K, P. 2022. An Innovative Approach for Association Rule Mining in Grocery Dataset Based On Non-Negative Matrix Factorization And Autoencoder. Journal of Algebraic Statistics, Volume 13, No. 3, 2022, p. 2898 – 2905
11. Al Shehabi, S., Baba, A. 2021. MARC: Mining Association Rules from datasets by using Clustering models. Published: 2021-07-14, Retrieved 2022-09-22 from: https://arxiv.org/abs/2107.08814
12. Bao, F., Mao, L., Zhu, Y., Xiao, C., Xu, C. 2021. An Improved Evaluation Methodology for Mining Association Rules. Axioms, Published: 2021-12-30, Retrieved from https://www.mdpi.com/2075-1680/11/1/17
13. Tahyudin, I., Haviluddin, I,. Nanbo, H. 2019. Time Complexity Of A Priori And Evolutionary Algorithm For Numerical Association Rule Mining Optimization. International Journal of Scientific & Technology Research Volume 8, Issue 11, November 2019
14. Retrieved 2022-12-22 from: https://stackoverflow.com/questions/9869884/what-is-the-time-and-space-complexity-of-FP-growth-algorithm
15. Chou, T. 2020. FP Growth - Frequent Pattern Generation in Data Mining with Python Implementation, Retrieved 2022-10-23 from: https://towardsdatascience.com/FP-growth-frequent-pattern-generation-in-data-mining-with-python-implementation-244e561ab1c3
16. Raschka, S. 2022,. fpgrowth: Frequent itemsets via the FP-growth algorithm, Retrieved 2022-12-29 from http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/
17. Desmos. Retrieved 2022-12-29 from https://www.desmos.com/
18. Agrawal, R. Srikant, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, 12-15 September 1994, pp. 487-499.

19. Koh, S. Y., Rountree, N. 2005. Finding sporadic rules usingapriori-inverse InProceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'05), Tu Bao Ho, David Cheung, and Huan Liu (Eds.). Springer-Verlag, Berlin, Heidelberg, 2005: 97-106.

20. Zaki, M. J. 2000. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12, 2000:372–390

8 Applications

1. Datasets, Algorithms, Analyses, and Master Thesis are saved in the author's repository: https://github.com/dawei7/Detecting-patterns-in-purchase-history-using-association-rule-learning-methods