PARALLEL FP GROWTH

Big Data Processing IT 494

Prof. PM Jat

TEAM

Dev Jadav
202001016

Vikas Patni
202001192

Hemang Joshi
202001212

PROBLEM

Parallelizing the FP-Growth Algorithm

Problems with Traditional

FP-GRWOTH ALGORITHM

STORAGE

COMPUTATION DISTRIBUTION

EXPENSIVE COMMUNICATION

SUPPORT VALUE

SOLUTION

ADVANTAGES OF

PARALLEL FP-GRWOTH

BETTER SCALABILITY

OPTIMIZED SUPPORT VALUE

REDUCED SYNCHRONIZATION CHALLENGES

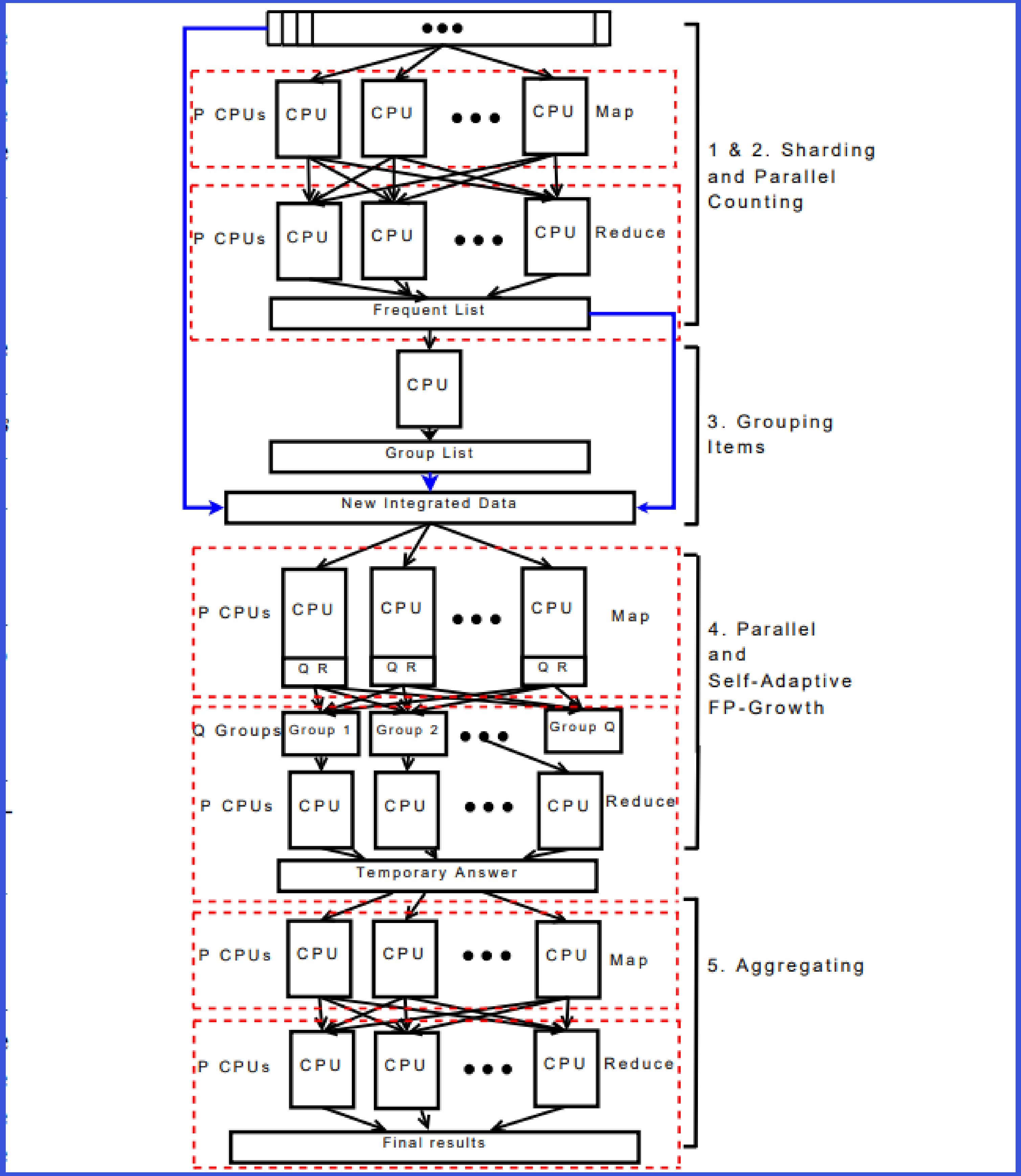
EFFICIENT UTILIZATION OF RESOURCES

ALGORITHM

ALGORITHM

- Sharding
- Parallel Counting
- Grouping Items
- Parallel FP-Growth
- Aggregating

- Sharding
- Parallel Counting
- Grouping Items
- Parallel FP-Growth
- Aggregating



SOURCE: Li, Haoyuan, et al. "Pfp: parallel fp-growth for query recommendation."

EXAMPLE OF ALGORITHM

SHARDING

- Shard 1 (2022-01-01): Transaction_ID 1001
- Shard 2 (2022-01-02): Transaction_IDs 1002, 1003
- Shard 3 (2022-01-03): Transaction_ID 1004

PARALLEL COUNTING

Each machine generates its F-list:

F-list:

- 'Shampoo': 1
- 'Conditioner': 1
- 'Hair Dryer': 1
- 'Milk': 1
- 'Eggs': 1
- 'Cheese': 1
- 'Bread': 1

... (other items)

GROUPING ITEMS

Divide the items into Q groups to form the G-list.

G-list:

- Group 1: ['Shampoo', 'Milk', 'Bread']
- Group 2: ['Conditioner', 'Eggs', 'Butter']
- Group 3: ['Hair Dryer', 'Cheese', 'Juice']
- Group 4: ['Chicken', 'Vegetables']

... (other groups)

PARALLEL FP-GROWTH

- Constructing the FP-Tree
- Mining frequent patterns
- Generating conditional pattern bases

IMPROVEMENT ON ALGORITHM

SPARK

- Frequent itemset
- Information matrix
- FP-tree of the information matrix M
- Mine all frequent itemsets
- Transform M's frequent itemsets into dataset's frequent itemsets

TESTING

DATASET

Transaction_ID	Date	Customer_Name	Product	Total_Items	Total_Cost	Payment_Method	City	Store_Type	Discount_Applied	Customer_Category	Season	Promotion
1000000000	2020-12-21 19:42:52	Cheyenne Newman	['Hair Gel']	6	12.77	Debit Card	New York	Convenience Store	TRUE	Student	Winter	None
1000000001	2020-07-06 7:45:16	Emily Fitzgerald	['Tuna', 'Bread', 'Tiss	5	13.88	Debit Card	Houston	Supermarket	FALSE	Professional	Fall	BOGO (Buy One Get
1000000002	2021-10-02 6:28:44	Michael Webb	['Jam', 'Soap', 'Ketch	7	47.02	Debit Card	Miami	Convenience Store	FALSE	Young Adult	Winter	None
1000000003	2022-01-10 5:39:02	Kimberly Lin	['BBQ Sauce']	9	83.86	Mobile Payment	Seattle	Warehouse Club	TRUE	Senior Citizen	Summer	Discount on Selecte
1000000004	2021-10-13 7:28:47	Cathy Hernandez	['Hand Sanitizer', 'Br	4	30.55	Debit Card	Houston	Warehouse Club	FALSE	Senior Citizen	Spring	None
1000000005	2021-04-26 20:45:13	Elizabeth Cook	['Shower Gel', 'Baby	10	30.19	Debit Card	Atlanta	Supermarket	TRUE	Teenager	Summer	None
1000000006	2023-10-07 23:36:53	Kara Bradley	['Cereal', 'Tuna']	3	5.57	Mobile Payment	Boston	Warehouse Club	TRUE	Student	Winter	Discount on Selecte
1000000007	2022-03-30 0:46:49	Carla Hernandez	['Iron', 'Extension Cor	1	7.15	Debit Card	Dallas	Warehouse Club	FALSE	Student	Fall	BOGO (Buy One Get
1000000008	2020-03-05 23:47:29	Christopher Wang	['Banana', 'Pickles']	2	20.04	Cash	Chicago	Warehouse Club	FALSE	Teenager	Winter	Discount on Selecte
1000000009	2023-03-26 13:28:32	Alisha Hudson	['Ketchup', 'Razors', '	3	88.79	Cash	Dallas	Pharmacy	FALSE	Teenager	Summer	None
1000000010	2023-05-10 16:20:28	Samantha Mcclure	['Shrimp', 'Soda']	3	28.43	Cash	Seattle	Specialty Store	TRUE	Young Adult	Spring	None
1000000011	2020-01-18 8:59:50	Shari Thomas	['Soap', 'Vacuum Cle	9	69.48	Debit Card	San Francisco	Pharmacy	TRUE	Senior Citizen	Fall	None
1000000012	2022-05-30 7:17:29	David Randolph	['BBQ Sauce', 'Soda'	9	21.29	Credit Card	Atlanta	Department Store	TRUE	Middle-Aged	Spring	Discount on Selecte
1000000013	2021-05-21 18:47:34	Maria Munoz	['Ironing Board', 'Lau	2	47.49	Mobile Payment	Seattle	Convenience Store	FALSE	Retiree	Summer	Discount on Selecte
1000000014	2023-02-26 11:38:21	Christopher Barnett	['Lawn Mower', 'Tea']	8	15.67	Credit Card	New York	Warehouse Club	FALSE	Senior Citizen	Fall	None
1000000015	2021-04-28 4:39:49	Jonathan Roach	['Syrup']	9	43.35	Cash	Los Angeles	Pharmacy	TRUE	Retiree	Summer	None
1000000016	2023-04-01 20:30:57	Alexander Hall	['Tea', 'Spinach', 'Mu	9	5.3	Cash	San Francisco	Pharmacy	FALSE	Middle-Aged	Winter	BOGO (Buy One Get
1000000017	2022-12-20 12:56:05	Bryan Smith	['Tuna', 'Bath Towels	8	64.76	Cash	Chicago	Specialty Store	FALSE	Teenager	Winter	Discount on Selecte
1000000018	2022-12-10 9:44:52	Kayla Sanchez	['Syrup', 'Yogurt', 'Eg	5	78.64	Credit Card	Boston	Convenience Store	TRUE	Senior Citizen	Summer	BOGO (Buy One Get
1000000019	2021-01-01 19:48:07	Adam Foster	['Eggs']	1	60.29	Debit Card	San Francisco	Warehouse Club	FALSE	Student	Summer	None
1000000020	2021-10-22 21:05:46	Nancy Mcdonald	['Eggs', 'Razors', 'Pea	7	98.81	Debit Card	Atlanta	Pharmacy	FALSE	Professional	Spring	BOGO (Buy One Get

Source: https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset

MAPREDUCE APPROACH

GLIST

	A	B	C	D	E	F
1	Beef	Bread	Broom	Butter	Canned Soup	Carrots
2	Cereal Bars	Cereal	Cheese	Chicken	Chips	Cleaning Rags
3	Cleaning Spray	Coffee	Deodorant	Diapers	Dish Soap	Dishware
4	Dustpan	Eggs	Extension Cords	Feminine Hygiene P	Garden Hose	Hair Gel
5	Hand Sanitizer	Honey	Ice Cream	Insect Repellent	Iron	Ironing Board
6	Jam	Ketchup	Laundry Detergent	Lawn Mower	Light Bulbs	Mayonnaise
7	Milk	Mop	Mustard	Olive Oil	Onions	Orange
8	Pancake Mix	Paper Towels	Pasta	Peanut Butter	Pickles	Plant Fertilizer
9	Potatoes	Power Strips		Air Freshener	Apple	BBQ Sauce
10	Baby Wipes	Banana	Bath Towels	Razors	Rice	Salmon
11	Shampoo	Shaving Cream	Shower Gel	Shrimp	Soap	Soda
12	Spinach	Sponges	Syrup	Tea	Tissues	Toilet Paper
13	Tomatoes	Toothbrush	Toothpaste	Trash Bags	Trash Cans	Tuna
14	Vacuum Cleaner	Vinegar	Water	Yogurt		

TOP K MOST RECOMMEDED ITEMS FOR EACH PRODUCT

```
"Bread" [0.2130790190735695, ["Bread"]]
"Bread" [0.009445958219800182, ["Bread", "Carrots"]]
"Bread" [0.007992733878292461, ["Bread", "Butter"]]
"Broom" [0.2108991825613079, ["Broom"]]
"Butter"        [0.22343324250681199, ["Butter"]]
"Butter"        [0.007992733878292461, ["Bread", "Butter"]]
"Canned Soup"   [0.1963669391462307, ["Canned Soup"]]
"Carrots"       [0.21326067211625796, ["Carrots"]]
"Carrots"       [0.009445958219800182, ["Bread", "Carrots"]]
"Cereal"        [0.21218836565096952, ["Cereal"]]
"Cheese"        [0.21348107109879963, ["Cheese"]]
"Cheese"        [0.008494921514312095, ["Cheese", "Cleaning Rags"]]
"Chicken"       [0.2090489381348107, ["Chicken"]]
"Chicken"       [0.008125577100646353, ["Chicken", "Chips"]]
"Chips" [0.21329639889196675, ["Chips"]]
"Chips" [0.008125577100646353, ["Chicken", "Chips"]]
"Cleaning Rags" [0.21292705447830101, ["Cleaning Rags"]]
"Cleaning Rags" [0.008494921514312095, ["Cheese", "Cleaning Rags"]]
"Coffee"        [0.2008240773916159, ["Coffee"]]
"Deodorant"     [0.2198136868505912, ["Deodorant"]]
"Deodorant"     [0.007703332139018273, ["Deodorant", "Dishware"]]
"Soap"  [0.2112448356385845, ["Soap"]]
"Soda"  [0.21950781390335908, ["Soda"]]
"Sponges"       [0.22834360275462123, ["Sponges"]]
"Sponges"       [0.008880028996013048, ["Syrup", "Sponges"]]
```

SPARK BASED APPROACH

FREQUENCY TABLE

```
+--------------------+----+
|               items|freq|
+--------------------+----+
|             [Apple]|1052|
|[Apple, Light Bulbs]|  35|
|[Apple, Light Bul...|   3|
|    [Apple, Vinegar]|  32|
|[Apple, Vinegar, ...|   3|
|[Apple, Vinegar, ...|   3|
|[Apple, Vinegar, ...|   4|
|[Apple, Vinegar, ...|   3|
|   [Apple, Tomatoes]|  26|
|[Apple, Tomatoes,...|   3|
|      [Apple, Honey]|  31|
|[Apple, Honey, Ex...|   3|
|[Apple, Honey, Yo...|   3|
|[Apple, Honey, Ca...|   3|
| [Apple, Toothbrush]|  31|
|[Apple, Toothbrus...|   3|
|[Apple, Toothbrus...|   4|
|[Apple, Toothbrus...|   3|
|[Apple, Toothbrus...|   4|
|[Apple, Toothbrus...|   3|
+--------------------+----+
only showing top 20 rows
```

ASSOCIATION RULES

antecedent	consequent	confidence	lift	support
[Tea, Air Freshener]	[Dish Soap]	0.0967741935483871	2.594482400761048	1.0E-4
[Tea, Air Freshener]	[Trash Bags]	0.0967741935483871	2.6273536709969343	1.0E-4
[Tea, Air Freshener]	[Olive Oil]	0.12903225806451613	3.4256351698544103	1.333333333333333...
[Tea, Air Freshener]	[Laundry Detergent]	0.0967741935483871	2.6586316908897554	1.0E-4
[Tea, Air Freshener]	[Cheese]	0.0967741935483871	2.6981652476316107	1.0E-4
[Trash Cans, Toot...	[Apple]	0.04477611940298507	1.2768855343056578	1.0E-4
[Trash Cans, Toot...	[Vacuum Cleaner]	0.05970149253731343	1.7821341055914457	1.333333333333333...
[Trash Cans, Toot...	[Cleaning Rags]	0.05970149253731343	1.6754394538067379	1.333333333333333...
[Trash Cans, Toot...	[Orange]	0.07462686567164178	2.1863339552238803	1.666666666666666...
[Trash Cans, Toot...	[Canned Soup]	0.05970149253731343	1.766316347257794	1.333333333333333...
[Trash Cans, Toot...	[Spinach]	0.05970149253731343	1.5878056525881232	1.333333333333333...
[Trash Cans, Toot...	[Air Freshener]	0.04477611940298507	1.234635645302897	1.0E-4
[Trash Cans, Toot...	[Ironing Board]	0.04477611940298507	1.195092154883943	1.0E-4
[Trash Cans, Toot...	[Milk]	0.04477611940298507	1.2357714646638014	1.0E-4
[Trash Cans, Toot...	[Salmon]	0.04477611940298507	1.2211668928086838	1.0E-4
[Trash Cans, Toot...	[Butter]	0.04477611940298507	1.1824679419802395	1.0E-4
[Trash Cans, Toot...	[Hair Gel]	0.04477611940298507	1.2233912405187177	1.0E-4
[Trash Cans, Toot...	[Toilet Paper]	0.04477611940298507	1.2301131704116777	1.0E-4
[Trash Cans, Toot...	[Rice]	0.04477611940298507	1.2323702587977543	1.0E-4
[Trash Cans, Toot...	[Plant Fertilizer]	0.05970149253731343	1.647695286218402	1.333333333333333...

only showing top 20 rows

PREDICTION TABLE

```
+---+--------------------+--------------------+
| id|               items|          prediction|
+---+--------------------+--------------------+
|  0|          [Hair Gel]|[Apple, Honey, To...|
|  1|[Trash Bags, Brea...|[Tea, Iron, Baby ...|
|  2|[Ketchup, Jam, Soap]|[Tomatoes, Insect...|
|  3|         [BBQ Sauce]|[Apple, Pasta, Va...|
|  4|[Ice Cream, Exten...|[Baby Wipes, Spin...|
|  5|[Baby Wipes, Pape...|[Apple, Pasta, Va...|
|  6|      [Tuna, Cereal]|[Apple, Honey, To...|
|  7|[Extension Cords,...|[Dish Soap, Soda,...|
|  8|   [Pickles, Banana]|[Eggs, Air Freshe...|
|  9|[Lawn Mower, Razo...|[Apple, Tomatoes,...|
| 10|      [Shrimp, Soda]|[Apple, Honey, To...|
| 11|[Mayonnaise, Soap...|[Apple, Tomatoes,...|
| 12|[Lawn Mower, Soda...|[Apple, Tomatoes,...|
| 13|[Laundry Detergen...|[Apple, Honey, To...|
| 14|   [Lawn Mower, Tea]|[Apple, Tomatoes,...|
| 15|             [Syrup]|[Apple, Pasta, Va...|
| 16|[Cleaning Rags, T...|[Dish Soap, Dishw...|
| 17|[Potatoes, Tuna, ...|[Water, Light Bul...|
| 18|[Yogurt, Syrup, E...|[Tomatoes, Milk, ...|
| 19|              [Eggs]|[Apple, Tomatoes,...|
+---+--------------------+--------------------+
only showing top 20 rows
```

REFERENCE

- Li, Haoyuan, et al. "Pfp: parallel fp-growth for query recommendation." Proceedings of the 2008 ACM conference on Recommender systems. 2008.
- Y. Miao, J. Lin and N. Xu, "An improved parallel FP-growth algorithm based on Spark and its application," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019
- https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html

IMPLEMENTATION CODE

[Google Colab Notebook]

THANK YOU!