A Study on Visualizing Semantically Similar Frequent Patterns in Dynamic Datasets

[1]Y.N.Jyothsna Mallampalli, [2] S.Jayaprada, [3]Dr S.Vasavi
[1]M-Tech II Year, [2]Senior Assistant Professor, [3]Professor
[1,2,3] Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous), Affiliated to JNTU Kakinada, KANURU, Vijayawada, Krishna (DT), Andhra Pradesh, India.

Abstract

Discovering frequent and interesting patterns is an important area of data mining. Transactional databases cannot serve the requirement of analyzing current trends in shopping; it is required to focus on analyzing dynamic data sets. Existing data mining algorithms when applied on dynamic data sets takes lot of time as they generate very huge number of frequent patterns making the analyst with the task to go through all the rules and discover interesting ones. Works that are reported until now in reducing number of rules are either time consuming or does not consider the interestingness of the user and does not focus on analysis of rules. This paper extends SSFPOA algorithm which produces clusters of semantically similar frequent patterns and presents these clusters using data visualization.

Keywords: Clusters, Dynamic Datasets, Ontology, Semantically similar frequent patterns, Data Visualization.

1. Introduction

One important area in data mining is concerned with the discovery of frequent patterns and interesting association rules. While considering the transactional databases, the patterns are extracted over a certain period of time like and at the end of the day. But, as the trends are continuously changing, the patterns extracted by previous day transactions may not suit to the present trends. Hence there is a need in extracting frequent patterns in dynamic datasets. The standard approach to update dynamic dataset is, applying the data mining algorithms continuously for every update in the dataset. The most general algorithm used for this purpose is "Apriori Algorithm". Apriori algorithm extracts all association rules satisfying minimum thresholds of support and confidence. If the threshold is high, some rules may be omitted and if the threshold is low then a large set of all rules will be extracted. Solutions such as frequent closed patterns, using filters, using redundancy rules, ontologies with semantics were proposed to solve this problem. Also compression on appropriate set of frequent patterns, frequent patterns asserting the interestingness of association rules which is evaluated by using relatedness based on relationship between item pair are investigated but there are still many challenges. First, the problem of mining such patterns is difficult as algorithms are very time and memory-consuming. Second, as usual in data mining problems, the number of patterns extracted by current solutions is too large to be easily handled by end-users. Fig.1 presents the traditional approach to association rule mining where support and confidence are used to produce best rules.

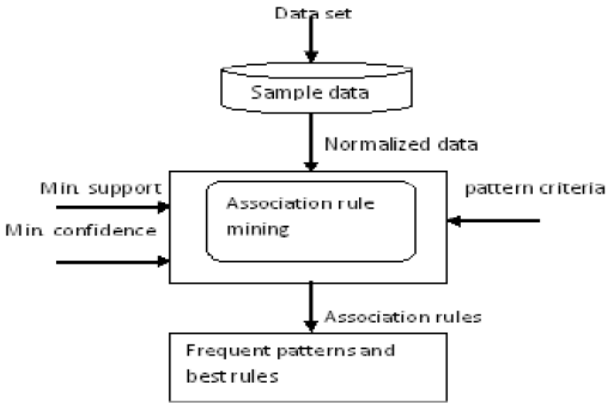

Figure 1: Traditional approach for association rule mining.

Hence, we are considering the methods for finding similarity within frequent patterns and from those patterns we can extract association rules. Then, we will apply clustering techniques on them. i.e. each cluster contains semantically similar patterns. The paper is organised as follows: section 2 discusses about related work, section 3 about Proposed Work, section 4 about Conclusions and Future Work.

2. Related Work

The interesting association rules and frequent patterns can be discovered by using 3 types of measures such as i) Objective Measures (ii) Subjective Measures and (iii) Semantic Measures. Objective measure [1] as given in Table 1 is a data-driven approach for evaluating the quality of association patterns. Since, these measures calculate the frequent patterns based on the statistical parameters, different measures produce different results. So, these are insufficient for determining the interestingness of a discovered rule. Subjective measures [2] generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. But these measures are also not sufficient as the rules generated by these measures are user biased. Semantic measures uses natural language processing techniques such as domain ontologies, web ontologies to identify relationships amongst the patterns. Resnik's approach [3] finds semantic similarity based on the distance from one node to another. i.e.

the path between the 2 nodes is shorter; they are treated as more similar. This approach is mainly based on the domain independent ontology such as Wordnet[4] But problem with Wordnet is, words which are not identified by it is treated as noise. Also, shortest path is not only sufficient to conform on semantic similarity. KEOPS methodology [5] works by comparing the extracted rules with expert's knowledge and it uses IMAK partway interestingness measure that considers relative confidence values and knowledge certainty for determining the rule quality. KEOPS only focused on "Rules based Patterns" and is mainly based on IMAK measure. But, IMAK is computed by quality indices like Support, confidence and lift, which are easily interpretable. Since, the patterns generated by this methodology are heterogeneous, it is very difficult to access as well as analyzing them. The work SSFPOA [6] extracts and clusters semantically similar frequent patterns. It uses both domain dependent and domain independent ontologies, and considers the entire path to conform the semantic similarity between elements along with their structural information such as the number of the children for each node of schema, the number of subclasses for each class within the ontology. We extend SSFPOA by adding visualization techniques for easy analysis of the extracted patterns. This paper proposes visualization techniques for the clusters of rules extracted by SSFPOA.

[7,8] uses Scatter plots to display data as a collection of points. These are used when a variable exists that is under the control of the experimenter. Mosaic plots [8,9] are often used to visualize relationship between two or more categorical values. Parallel co-ordinate plots [8] visualise the distribution of the values of a variable over the different values of another variable (For example, calculation of annual expenditure etc).

In, Matrix-visualization [10], the Antecedent is represented on the X-axis and Consequent on the Y-axis. Intersection of Antecedent and Consequent represents the Selected interest measure. All these techniques can visualize only small amounts of data. In [11] the results of association rule mining algorithms are represented as directed graphs. Items, item-sets and association rules are represented as nodes, where as the links between items and item-sets or association rules are represented as edges. But, this approach can visualize only frequent item-sets and binary association rules derived from transactional data. Line graphs are most useful in displaying data or information that changes continuously over time. Histograms are the special type of bar charts, for visualizing showing a visual impression of the distribution of data. It can display large amounts of data that are difficult to understand in a tabular, or spreadsheet form. So, line graphs and histograms will be more suitable for visualizing semantically similar frequent patterns in dynamic datasets.

3. Proposed Work

Many mining methods use measures such as support and confidence for mining association rules efficiently and also for measuring the quality of the mined rules. While considering the transactional databases, the patterns are extracted over a certain period of time like, at the end of the day. But, as the shopping trends are continuously changing, the patterns extracted by previous day transactions may not be useful to the present trends. So that, we have to consider the problem of extracting frequent patterns in dynamic datasets, in which the dataset is updated in small intervals of time. The measure SSFPOA uses domain dependent as well as domain independent ontologies to construct semantic similarity matrix between all pairs of frequent items and then clusters all the frequent patterns which have high similarity. These clusters are further used to generate association rules with high level abstraction[12]. But, analyzing these association rules is a more difficult task, we are extending this method with visualizing these clusters graphically. Figure 2 shows our extended approach.

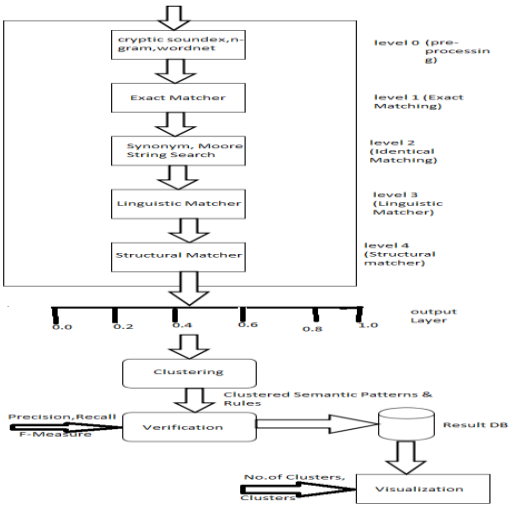

Figure 2: Extended Approach of Visualizing Semantically Similar Frequent Patterns

In this approach Level 1 to level 4 are the hidden layers. Here, each layer evaluates and output of this layer is given as input to the next layer. Layer 1 is an exact string matcher which gives either 0 or 1 depending on equality. If it generates 1 then level 2, 3 are not performed and is directed to level 4 for context matching. Layer 2 uses 2 matchers (neuron units) and layer 3 uses 6 matchers. Structural matcher considers the depth of the node within the tree along with the other structural features such as :

(i) Number of children for each node
(ii) Number of subclasses for each class of ontology

Also semantic similarity is measured in the range of 0-1. The algorithm SSFPOA has focused on interpreting the frequent patterns that are mined, especially extracting semantically similar items and clustering them, instead of generating large number of distinct rules for semantically similar items with separate support values, related rules can be reduced. Our algorithm VSSFPOA extends, SSFPOA by visualizing the clusters of semantically similar frequent item sets using Line graphs and Histograms. Table 2 presents semantically similar frequent patterns produced by SSFPOA for product domain available at [13]. Similar patterns are clustered to generate rules which are in higher level of abstraction. For example, the semantically similar items of Table 2 can be clustered as 3 clusters, (i). Beauty/Health (ii) Foods and (iii) Other items. These can be formed by clustering the similar patterns which are in the lower level of ontology. i.e. Beauty/Heath cluster can be formed by clustering Toothpaste, Creams, Hair products, Health and Soaps. Biscuit/Rusk contains Biscuits, Papad, Dal/pulses, Roti, Snacks and Pickles. The 3rd cluster Other items contains, Vegetables, Miscellaneous and Phone cards. In Line graph Visualization, we represent the cluster numbers on X-axis and the semantic similarity value on Y-axis. By plotting respective clusters with their corresponding similar value, one can easily visualize the clusters for easy analysis as shown in Fig.3.

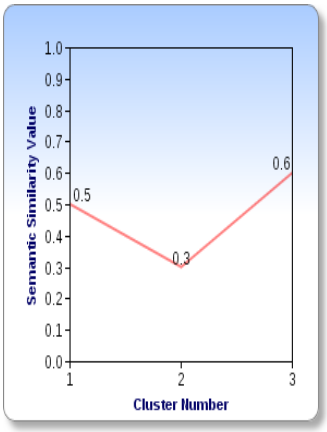

Figure 3: Line graph visualization of semantically similar frequent patterns for Table 2.In Fig.4 visualization, we first calculate the centroid of each cluster. The cluster numbers are represented on X-axis and the centres on the Y- axis. By plotting respective clusters with their centroid, one can easily visualize the clusters for easy analysis as shown in Fig.4.

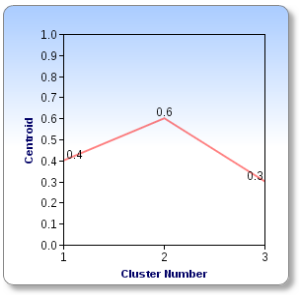

Figure 4: Line graph visualization of semantically similar frequent patterns for Table 2.

In Histogram visualization, we consider the number of rules or items each cluster have. By representing the clusters on X-axis and number of rules on Y-axis, the visualization is more effective for the analysis as shown in Fig.5.

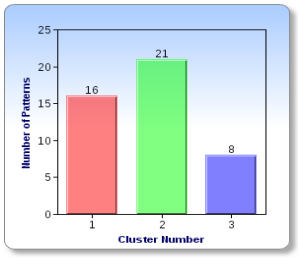

Figure 5: Histogram visualization of semantically similar frequent patterns for Table 2.

4. Conclusion And Future Work

According to the rapidly changing present shopping trends, it is essential to consider dynamic Datasets for extracting frequent patterns. Existing data mining algorithms are very time consuming and they generate very huge number of frequent patterns, hence we extend SSFPOA algorithm which results the semantically similar frequent patterns at higher levels of abstraction and further clusters them. This paper extends SSFPOA as VSSFPOA by adding visualization techniques like line graphs and histograms to clusters for easy analysis of patterns. Our future work concentrates on providing visualizations in 3-D and the comparison of performance of various visualization techniques.

References

[1]. Guillaume, Sylvie, Dhouha Grissa, and Engelbert Mephu Nguifo, Categorization of interestingness measures for knowledge extraction, arXiv preprint arXiv:1206.6741 (2012).
[2]. Bing Liu , Wynne Hsu , Shu Chen , Yiming Ma, Analyzing the Subjective Interestingness of Association Rules, IEEE Intelligent Systems(journal) - Volume 15 Issue 5, September 2000, Pages 47 - 55.
[3]. Resnik P, Using information content to evaluate semantic similarity in a taxonomy(1995).
[4]. http://www.wordnet.princeton.edu
[5]. Laurent Brisson and Martine Collard, How to Semantically Enhance a Data Mining Process?, ICEIS 2008: 103-116.
[6]. S. Vasavi, S. Jayaprada, V. Srinivasa Rao, Extracting Semantically Similar Frequent Patterns Using Ontologies, SEMCCO'11 Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing - Volume Part II, Pages 157-165.
[7]. http://en.wikipedia.org/wiki/Scatter_plot
[8]. Hashler M. and Chelluboina S., Visualizing Association Rules: Introduction to the R-extension Package arulesViz(2011)
[9]. Hofmann H., Siebes A., and Wilhelm A.F.X.,(2000), Visualizing Association Rules with Interactive Mosaic Plots, in KDD, pp. 227-235.
[10]. Michael Hahsler and Sudheer Chelluboina, Visualizing association rules in hierarchical groups, In Computing Science and Statistics, Vol. 42, 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms" (Interface 2011).
[11]. Ertek G. and Demiriz A, (2006), A Framework for Visualizing Association Mining Results, in ISCIS, pp. 593-602
[12]. S.Vasavi, S.Jayaprada, Clustering Semantically Similar Frequent Patterns using SSFPOA, International Journal of Data Ware housing & Mining, Vol 2 Issue 1 pg 45-51,2012 ISSN:2249-7161.
[13]. http://www.indianmart.com.au/shopping/index.php

Table 1: Definition of Objective Measures

N°	Measure	Formula
1	Correlation coefficient	$\frac{p(XY)-p(X)p(Y)}{\sqrt{p(X)p(Y)p(\bar{X})p(\bar{Y})}}$
2	Cohen or Kappa	$2\ \frac{p(XY)-p(X)p(Y)}{p(X)+p(Y)-2p(X)p(Y)}$
3	Confidence or precision	$\frac{p(XY)}{p(X)}$
4	Causal Confidence	$1-\frac{1}{2}\left(\frac{1}{p(X)}+\frac{1}{p(\bar{Y})}\right)p(X\bar{Y})$
5	Centered Confidence or Pavillon	$\frac{p(XY)}{p(\bar{Y})}-p(Y)$
6	Descriptive Confirm Confidence or Ganascia	$1-2\ \frac{p(X\bar{Y})}{p(X)}$
7	Causal Confirm Confidence	$1-\frac{1}{2}\left(\frac{3}{p(X)}+\frac{1}{p(\bar{Y})}\right)p(X\bar{Y})$
8	Causal Confirm	$p(X)+p(\bar{Y})-4p(X\bar{Y})$
9	Descriptive Confirm	$p(XY)-p(X\bar{Y})$
10	Conviction	$\frac{p(X)p(\bar{Y})}{p(X\bar{Y})}$
11	Ccsinus or Ochiai	$\frac{p(XY)}{\sqrt{p(X)p(Y)}}$
12	Coverage	$p(X)$
13	Czekanowski-Dice or F-measure	$2\ \frac{p(XY)}{p(XY)+1-p(\bar{X}\bar{Y})}$
14	Dependency	$\|\|p(\bar{Y})-\frac{p(X\bar{Y})}{p(X)}\|\|$
15	Putative Causal Dependency	$\frac{3}{2}+2p(X)-\frac{3}{2}p(Y)-\left(\frac{3}{2p(X)}+\frac{3}{2p(\bar{Y})}\right)p(X\bar{Y})$
16	Gray and Orlowska's Interestingness Weighting Dependency	$\left(\left(\frac{p(XY)}{p(X)p(Y)}\right)^{k}-1\right)\times p(XY)^{m}$
17	Bayes factor or Odd multiplier	$\frac{p(XY)p(\bar{Y})}{p(X\bar{Y})p(Y)}$
18	Certainty factor or Loevinger or Satisfaction	$\frac{p(XY)-p(X)p(\bar{Y})}{p(X)p(\bar{Y})}$
19	Negative reliability	$\frac{p(X\bar{Y})}{p(\bar{Y})}$
20	Collective Strength	$\frac{p(XY)+\frac{p(X\bar{Y})}{p(X)}}{p(X)p(Y)+p(\bar{X})p(\bar{Y})}\times\frac{1-p(X)p(Y)-p(\bar{X})p(\bar{Y})}{1-p(XY)-\frac{p(X\bar{Y})}{p(X)}}$
21	Fukuda	$n\left(p(XY)-\sigma_{\circ}\ p(X)\right)$
22	Informational gain	$log_2\left(\frac{p(XY)}{p(X)p(Y)}\right)$
23	Gini	$p(X)\left(\frac{p^2(XY)}{p^2(X)}+\frac{p^2(X\bar{Y})}{p^2(X)}\right)+p(\bar{X})\left(\frac{p^2(\bar{X}Y)}{p^2(\bar{X})}+\frac{p^2(\bar{X}\bar{Y})}{p^2(\bar{X})}\right)-p^2(Y)-p^2(\bar{Y})$

Table 1: Definition of Objective Measures(Continued)

24	Goodman-Kruskal	$\frac{\sum_j \max_k P(X_j,Y_k)+\sum_k \max_j P(X_j,Y_k)-\max_j P(X_j)-\max_k P(Y_k)}{2-\max_j P(X_j)-\max_k P(Y_k)}\times\frac{n_{XY}n_{\bar{X}\bar{Y}}-\frac{n_{X\bar{Y}}n_{\bar{X}Y}}{n^2}}{n_{XY}n_{\bar{X}\bar{Y}}+\frac{n_{X\bar{Y}}n_{\bar{X}Y}}{n^2}}$
25	Implication index	$\sqrt{n}\ \frac{p(XY)-p(X)p(Y)}{\sqrt{p(X)p(Y)}}$
26	Probabilistic intensity of deviation from equilibrium (IPEE)	$P\left[N(0,1)\geq\frac{n_{X\bar{Y}}-n_{XY}}{\sqrt{n_X}}\right]$
27	Entropic probabilistic intensity of deviation from equilibrium (IPSE)	$\sqrt{\left[\frac{1}{2}\left((1-h_1(P(X\bar{Y}))^2)\times(1-h_2(P(X\bar{Y}))^2)\right)^{\frac{1}{4}}+1\right]}\times\sqrt{IPEE}\quad with\ h_1(t)=-\left(1-\frac{t}{p(X)}\right)log_2\left(1-\frac{t}{p(X)}\right)-\frac{t}{p(X)}log_2\left(\frac{t}{p(X)}\right)\ for\ t\in\left[0,\frac{p(X)}{2}\right[,\ else\ h_1(t)=1\ h_2(t)=-\left(1-\frac{t}{p(Y)}\right)log_2\left(1-\frac{t}{p(Y)}\right)-\frac{t}{p(Y)}log_2\left(\frac{t}{p(Y)}\right)\ for\ t\in\left[0,\frac{p(Y)}{2}\right[,\ else\ h_2(t)=1$
28	Probabilistic discriminant index (PDI)	$P\left[N(0,1)\geq II^{CR/B}\right]$ where $II^{CR/B}$ indicate that II is reduced-centred according to the values taken by II on the extracted rules set.
29	Mutual Information	$\frac{VS(XY)}{-P(X)log_2P(X)-P(\bar{X})log_2P(\bar{X})}$
30	Intensity of Implication (II)	$P\left[Poisson(nP(X)P(\bar{Y}))\geq P(X\bar{Y})\right]$
31	Entropic intensity of implication (IIE)	$\sqrt{\left[\left(1-h_1(P(X\bar{Y}))^2\right)\times\left(1-h_2(P(X\bar{Y}))^2\right)\right]^{\frac{1}{4}}}\times II$
32	Entropic intensity of revised implication (IIER)	$\sqrt{\left[\left(1-h_1(P(X\bar{Y}))^2\right)\times\left(1-h_2(P(X\bar{Y}))^2\right)\right]^{\frac{1}{4}}}\times\sqrt{max(2\times II-1;0)}$
33	Likelihood discriminant index	$P\left[Poisson(nP(X)P(Y))<P(XY)\right]$
34	Interest or Lift	$\frac{p(XY)}{p(X)p(Y)}$
35	Jaccard	$\frac{p(XY)}{p(X\bar{Y})+p(\bar{X}Y)}$
36	J-Measure	$p(XY)\ log\left(\frac{p(XY)}{p(X)p(Y)}\right)+p(X\bar{Y})\ log\left(\frac{p(X\bar{Y})}{p(X)p(\bar{Y})}\right)$
37	Klosgen	$\sqrt{p(XY)}\left(\frac{p(XY)}{p(X)}-p(Y)\right)$
38	Kulczynski or Agreement and disagreement index	$\frac{p(XY)}{p(X\bar{Y})+p(\bar{X}Y)}$
39	Laplace	$\frac{np(XY)+1}{np(X)+2}$
40	Leverage	$\frac{p(XY)}{p(X)}-p(X)p(Y)$
41	M_{GK}	$If\ P(Y/X)\geq P(Y)\ then\ M_{GK}(X\rightarrow Y)=\frac{p(Y/X)-p(Y)}{1-p(Y)}\ Else\ M_{GK}(X\rightarrow Y)=\frac{p(Y/X)-p(Y)}{p(Y)}$
42	Least contradiction or Surprise	$\frac{p(XY)-p(X\bar{Y})}{p(Y)}$

Table 1: Definition of Objective Measures(Continued)

43	Novelty	$p(XY) - p(X)p(Y)$
44	Pearl	$p(X)\lvert\frac{p(XY)}{p(X)} - p(Y)\rvert$
45	Piatetsky-Shapiro	$n \times \left(p(XY) - p(X)p(Y)\right)$
46	Accuracy	$p(XY) + p(\bar{X}\bar{Y})$
47	Prevalence	$p(Y)$
48	Yule's Q	$\frac{p(XY)p(\bar{X}\bar{Y}) - p(X\bar{Y})p(\bar{X}Y)}{p(XY)p(\bar{X}\bar{Y}) + p(X\bar{Y})p(\bar{X}Y)}$
49	Recall	$\frac{p(XY)}{p(Y)}$
50	Odds Ratio	$\frac{p(XY)p(\bar{X}\bar{Y})}{p(X\bar{Y})p(\bar{X}Y)}$
51	Relative Risk	$\frac{p(Y\lvert X)}{p(Y\lvert\bar{X})}$
52	Sebag-Schoenauer	$\frac{p(XY)}{p(X\bar{Y})}$
53	Specificity	$\frac{p(\bar{X}\bar{Y})}{p(\bar{X})}$
54	Support or Russel and Rao index	$p(XY)$
55	Yao and Liu's One Way Support	$\frac{p(XY)}{p(X)} log_2 \frac{p(XY)}{p(X)p(Y)}$
56	Yao and Liu's Two Way Support	$p(XY) log_2 \frac{p(XY)}{p(X)p(Y)}$
57	Examples and counter-examples rate	$\frac{p(XY) - p(X\bar{Y})}{p(XY)}$
58	Test value VT100	$\phi^{-1}(P[Hypergeometric(100P(X)P(Y)) \leq P(XY)])$
59	Yao and Liu's Two Way Support Variation	$p(XY) log_2 \frac{p(XY)}{p(X)p(Y)} + p(X\bar{Y}) log_2 \frac{p(X\bar{Y})}{p(X)p(\bar{Y})} + p(\bar{X}Y) log_2 \frac{p(\bar{X}Y)}{p(\bar{X})p(Y)} + p(\bar{X}\bar{Y}) log_2 \frac{p(\bar{X}\bar{Y})}{p(\bar{X})p(\bar{Y})}$
60	Yule's Y	$\frac{\sqrt{p(XY)p(\bar{X}\bar{Y})} - \sqrt{p(X\bar{Y})p(\bar{X}Y)}}{\sqrt{p(XY)p(\bar{X}\bar{Y})} + \sqrt{p(X\bar{Y})p(\bar{X}Y)}}$
61	Zhang	$\frac{p(XY) - p(X)p(Y)}{max\left\{p(XY)p(\bar{Y}),\ p(Y)p(X\bar{Y})\right\}}$

Table 2: Semantically similar frequent patterns for product domain

SNO	Semantically similar items	Parent node
1	Vicco vajardanti 100gm 3.75$ Vicco vajardanti 100gm 6.50$	Tooth paste
2	Gori Gori blue fairness bleach 50gm $4.95 Gori Gori fairness bleach 50gm $4.95	Creams
3	Godrej renew hair color cream black 120ml $6.95 Godrej renew hair color cream brown 120ml $6.95	Hair products
4	Supreme Dark brown henna 150gm $2.95 Supreme Maroon henna 150gm $2.95	Hair products
5	Eno lemon 100gm $2.95 Eno regular 100gm $2.95	Health
6	Hajmola imli 120 tbs $3.95 Hajmola regular 120 tbs	Health
7	Godrej No.1 natural $2.75 Godrej No.1 rose $2.75	Soaps
8	Neem active toothepaste 125gm $3.25 Neem toothepaste 125gm $3.25	Tooth paste
9	Pattu Jeera khari biscuit - 200gm $2.95 Pattu Masala khari biscuit - 200gm $2.95 Pattu plain khari biscuit - 200gm $2.95 Pattu Tomato khari biscuit - 200gm $2.95	Biscuit
10	Indian Mart Black Urid 1kg $3.75 Indian Mart Black Urid split 1kg $3.75	Dal/pulses
11	Indian Mart Kabuli Channa (10mm) 1kg $3.75 Indian Mart Kabuli Channa (9mm) 1kg $3.75	Dal/pulses
12	Indian Mart Toor Daal 1kg $3.75 Indian Mart Toor Daal premium 1kg $3.75	Dal/pulses
13	Katoomba Roti Parantha (20 roti) $7.25 Katoomba Roti Parantha lite(20 roti) $7.25	Roti
14	Mezban Gajar Halwa 280gms $6.25 Mezban Loki Halwa 280gms $6.25	Snacks
15	Taj Valor Lilva (indian beans) 400gm $2.00 Taj Valor Papdi (indian beans) 400gm $2.00	Vegetables
16	Food Color - Orange $2.00 Food Color - Red $2.00 Food Color - Yellow $2.00	Miscellaneous
17	Aithra sago papad 200gm $2.25 Aithra sago papad color 200gm $2.25	Papad
18	Lijjat papad - garlic 200gm $2.00 Lijjat papad –red chilli 200gm $2.00 Lijjat papad - Urad 200gm $2.00	Papad
19	South Asia Phone Card $8.00 South Asia Phone Card $16.00 South Asia Phone Card $40.00	Phone cards
20	Priya garlic pickle $2.50 Priya garlic pickle $6.50	Pickles