# A Novel Pattern Merger Algorithm for generating Actionable Rules for Multi-Source Combined Mining

Arti Deshpande
Department of Computer Science and Engineering
G.H. Raisoni College of Engineering
Nagpur, India

Anjali Mahajan
Department of Information Technology
Government Polytechnic
Nagpur, India

## ABSTRACT
The idea of combined mining is very useful and flexible for identifying in-depth patterns through combining sets of items from multiple datasets or using multiple data mining techniques. The identified combined patterns disclose in-depth business intelligence, which are more informative and actionable for business decision-making. The business data is scattered at various locations and to arrive at decisions it needs analysis of data by integrating entire data.

This paper emphasizes on applying Association Rule Mining technique on various data sources located at different locations and patterns so obtained are transported to main site. Finally by applying the proposed novel Pattern Merger Algorithm on the aforesaid patterns, all the generated patterns are merged to obtain actionable rules. These actionable patterns assist in strategic business planning and also pin points various issues arising post application of the data integration technique. Domain Knowledge concept is also included in the Rule generation technique to obtain the final results which are in the form of actionable rules.

## General Terms
Pattern Merger Algorithm , Multisource Combined Mining.

## Keywords
Pattern merger, actionable rules, association rule mining, business intelligence, combined mining.

## 1. INTRODUCTION
Data is very essential part for any business to take strategic decisions for business. In retail industry, whenever company wants to launch a new product or to give any offer on the product, it considers the past transactional data. This data is scattered at different locations. Ultimately all the data is to be clubbed together to get the actual patterns or rules for products sold at various locations by applying appropriate data mining technique. Knowledge can be sort by analyzing the produced patterns and the result assists in launching product offers.

This clubbing of data from various sources is tedious and expensive process as the costs associated with transportation of data from different locations is a costly affair. To optimize the cost predominantly the patterns from various locations are transferred instead of transactional data. Domain-driven data mining is also considered while generating patterns which gives actionable knowledge in a constrained environment for satisfying user preferences [1].

Domain knowledge is the concept in the proposed work where the user can select the features from the dataset for both static and transactional data. This helps to reduce the total number of rules generated using classical rule mining.

Often, extracted mined patterns are non-actionable to cater to the real needs due to lack of interest of such patterns to the business people. Therefore, it is essential to understand the needs so as to recognize interesting links that permit users to react them to create better business objectives [3].

Longbing Cao[2] introduced combined mining as a common way to mining for informative patterns by clubbing together many data sets or many features or many methods as required. The focus of multi-source combined mining is on combining multiple data sets to get more informative knowledge.

Jia and Ning [4] had proposed integrated intelligent e-business portals for multiple data sources. Using the information coming from the huge, distributed, multiple sources, they presented a conceptual model with dynamic multi-level workflows corresponding to a mining-grid centric multi-layer grid architecture, for multi-aspect analysis in building an e-business portal on the Wisdom Web. They had used multi-database mining to find the relationship between the customer's surfing and purchasing behaviors to get the target customers.

Product recommendation for E-commerce was given by Chuen-He Liou [5] for Multi-channel CRM. The sparse problem of customer-product matrix was resolved by considering multiple channel retailers and Web channel. This helped to find similar users for E-commerce. Mobile channels are not considered to get the customer behavior patterns.

In [6] classification on multiple heterogeneous databases is proposed. Inter-database links which are useful for information transfer are used for prediction. A new approach for cross-database classification is given by Xiaoxin Yin and Jiawei Han for classification with data stored in multiple heterogeneous databases.

Charls X. [12] had given the method for classification for multiple data sources. For intelligent business decision making, the author has proposed a novel method to predict the classification for data at multiple sources without class labels in each source. This method bridges the gap between supervised and unsupervised learning.

Sayyadian et al. [13] proposed the technique for Classification to built an accurate prediction model on single database. They combined all the tuples from multiple databases to create the classification model. Those databases were heterogeneous, so they used combine schema matching technique.Our work is focused on applying Multi-Source Combined Mining (MSCM) using association rule mining technique. Framework is being designed for MSCM and a novel Pattern Merger Algorithm is proposed to combine patterns is given in section 2. Association rule mining is applied on each data set and finally the generated rules are combined based on interestingness measure. Experimental results for sample data are shown in section 3.

# 2. MULTI-SOURCE COMBINED MINING

## 2.1 Framework for Multi-Source combined Mining

Multi-Source Combined mining framework is aimed at handling large amount of transactional data from multiple sources as shown in fig. 1. In retail industry, data is scattered at various outlets based at different locations. Association rule mining technique is applied on different data sources D1, D2…..Dn. Feature selection method [7] is applied before association rule mining to select the appropriate and selective features of the available data and in turn reducing the dataset for rule mining . Different pattern sets P1, P2,…..Pn are generated from different data sources D1, D2……Dn respectively post application of the association rule mining technique. By applying subjective and objective interestingness measures [8], patterns are pruned. Finally the pruned patterns so generated are filtered and summarized to produce actionable rules having domain and meta knowledge of the business. These final patterns are called as deliverables or actionable patterns which helps the business people in arriving at appropriate decisions for their business growth.

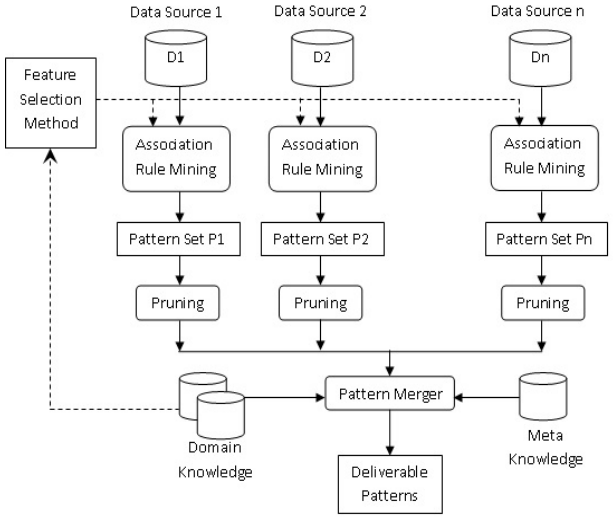


**Fig. 1. Framework for Multi-Source combined Mining**

## 2.2 Pseudo-code for Muli-Source combined mining

The Pseudo-code for MSCM is given below :

Input : Target dataset D, Business problem, threshold for interestingness measures

Step 1 : Generate pattern set P

For Each Dataset D1 to Dn

Apply feature selection method

Employ Association Rule Mining method

Extract the pattern Set Pi

End For

Step 2 : Reduction of general Patterns set using Pruning

For each Pattern Set P1 to Pn

Apply technical and business interestingness measures

Extract the pruned rules

End For

Step 3 : Pattern Merger

By involving domain knowledge and meta data, merge all pruned rules.

Step 4: Get deliverable patterns

Convert the patterns into actionable business rules.

OUTPUT : Actionable pattern set P

## 2.3 Algorithm for Pattern Merger

Pattern Merger combines the patterns generated from various data sources. Select top n patterns from each data source based on support, confidence and lift of that rule.

Assume that the rule is $x \rightarrow y$ where x and y are transaction items [9].

Support: The support of the rule, that is, the number of transactions having both x and y.

$$\text{support } (x \rightarrow y) = \text{ support } (x + y)$$

Confidence: The confidence of the rule.

$$\text{confidence } (x \rightarrow y) = \text{support } ( x + y) / \text{support } (x)$$

Lift: The lift value of the rule is the additional interestingness measures on the rules. Lift is used to give the ranking to the rules as per their importance or to filter out the rules.

$$\text{Lift } (x \rightarrow y) = \text{confidence } (x \rightarrow y) / \text{support } (y)$$

Once the patterns are generated from various sources at the local site only, deploy those patterns to central site. In this case the transportation cost of transfer of data is reduced as only patterns generated are transferred from local site to central site where actual analysis is to be done. Only top n patterns from each source site based on interestingness measures are considered for further processing of merging patterns. The patterns which are common to all sources are filtered out as set of rules R.

Apply the pattern Merger algorithm on those top n rules to get the final deliverable patterns. The algorithm for pattern merger is given below :

Algorithm for Pattern Merger:

Input : Top n rules from k sources,

n is number of rules with lift >=1,

k is number of sources S1,S2,…….,Sk.

R is set of rules which are common for all sources S1,S2,……..Sk

Combine_Rule = NULL

Rule_Found = 0

For source S1

{

For each rule from 1 to n // Rule from Source 1

{

A=Antecedent of Source_Rule //get the left hand side

// of the rule

```
{
        For each source from S2 to Sk  // get top n
        //rules from other remaining  source
  {
        // Check Antecedent of Source1 Rule with
        //antecedent of  Rules from other source
      If A = Antecedent of Other_Source_Rule  then
          Combine_Rule= Combine_Rule  U
          Other_Source_Rule
          Rule_Found = 1
      End If
  }
}
  If Rule_Found = 1 then //only if antecedent found in
                         //any of the source
    Combine_Rule = Combine_Rule U Rule_Source1
  End If
 }
}
```

Output : Combined Rules From Various Sources.

Only the patterns or rules are combined through pattern merger but in all those combined rules, user may not be interested or some of rules may be redundant, so to get the final actionable rules , another algorithm is proposed which gives the final deliverable patterns.

## 2.4 Algorithm to get Actionable Rules

Finally the support of each combined rule is to be checked for all sources. Consider the minimum threshold support value and based on that if all the sites are having the rule $x \rightarrow y$ or $y \rightarrow x$ with above or equal to minimum threshold support value, then that rule is considered as final actionable rule. If the rule $x \rightarrow y$ or $y \rightarrow x$ does not exist on other sources then remove such rule from the combined rule. If the rule $x \rightarrow y$ or $y \rightarrow x$ exists on more than n (where n is some threshold value for number of sources) sources, then consider that rule also the deliverable rule. Finally merge all the deliverable rules together which are actionable rules. The algorithm to get Actionable Rules is given below:

Input : Set of combined rule R with their support on each sources.

AR = NULL // Initally actionable rule AR is NULL

For each rule Ri = $x \rightarrow$ yor $y \rightarrow x$

Get the support for Ri from each source

If Ri exists in all sources then

AR = AR U Ri

End If

If Ri doesn't exists in any other sources then

Remove that rule from R

End If

If Ri exists more than n number of sources then

AR = AR U Ri

End If

End For

Output : Actionable pattern or Deliverable rules

Finally the actionable patterns are generated which are useful for business to take decisions. If the rule mining is applied on integrated data of all sites , then number of rules generated are more than the proposed technique. Understanding of the rules is difficult task and business people may not be interested in those generated rules.

# 3. IMPLEMENTATION AND EXPERIMENTAL RESULTS FOR MSCM

Transaction data is generated for two sources S1 and S2 using Red Gate SQL Data Generator [10]. Data set contains 11 items {Bread ,Butter, Milk, Cheese, Bucket, Washingpowder, Soap, Toothpaste, Shampoo, Juice, jam}. Both the sites S1 and S2 have 1000 transactions each. Third site S3 is considered as central site for analysis.

For experimental purpose Sql Server 2012 and DotNet technology is used. Association Rule Mining i.e Aprori Algorithm [11] is applied on data sets for both the sources S1 and S2 separately with minimum support = 200 transactions and Confidence = 40%. For the source S1 and S2 total 50 and 55 rules are generated respectively and transported to the central site S3 for analysis. Using the interestingness measure Lift, only top 10 rules are considered from both the sites as shown in table 1 and 2. Combined patterns are generated after applying pattern Merger algorithm given in section 2.3 as shown in Table 3.

**Table 1. Patterns Generated from Source 1**

| Sr.No. | Antecedent (x) | Consequent (y) | Support for x & y | % support x and y | Confidence (X->y) % | Lift Ratio |
|---|---|---|---|---|---|---|
| 1 | Bread | Soap | 363 | 18.15 | 43.89 | 1.097 |
| 2 | Soap | Bread | 363 | 18.15 | 45.38 | 1.097 |
| 3 | Butter | Jam | 329 | 16.45 | 42.34 | 1.087 |
| 4 | Jam | Butter | 329 | 16.45 | 42.23 | 1.087 |
| 5 | Butter | Bread | 344 | 17.2 | 44.27 | 1.071 |
| 6 | Bread | Butter | 344 | 17.2 | 41.60 | 1.071 |
| 7 | Bucket | Soap | 356 | 17.8 | 42.79 | 1.070 |
| 8 | Soap | Bucket | 356 | 17.8 | 44.50 | 1.070 |
| 9 | Bread | Juice | 345 | 17.25 | 41.72 | 1.066 |
| 10 | Juice | Bread | 345 | 17.25 | 44.06 | 1.066 |

**Table 2. Patterns Generated from Source 2**

| Sr.No. | Antecedent (A) | Consequent (C) | Support for A & C | % support A and C | Confidence (X->y) % | Lift Ratio |
|---|---|---|---|---|---|---|
| 1 | Juice | Cheese | 349 | 17.45 | 44.74 | 1.113 |
| 2 | Cheese | Juice | 349 | 17.45 | 43.41 | 1.113 |
| 3 | Bucket | Jam | 350 | 17.5 | 43.80 | 1.108 |
| 4 | Jam | Bucket | 350 | 17.5 | 44.25 | 1.108 |
| 5 | Shampoo | Milk | 346 | 17.3 | 43.80 | 1.105 |
| 6 | Milk | Shampoo | 346 | 17.3 | 43.63 | 1.105 |
| 7 | Bread | Butter | 349 | 17.45 | 44.29 | 1.080 |
| 8 | Butter | Bread | 349 | 17.45 | 42.56 | 1.080 |
| 9 | Bread | Soap | 345 | 17.25 | 43.78 | 1.076 |
| 10 | Soap | Bread | 345 | 17.25 | 42.38 | 1.076 |

**Table 3. Combined Patterns Generated from Source 1 and Source 2**

| Sr. No. | Antecedent | SOURCE1 | | SOURCE2 | |
|---|---|---|---|---|---|
| | | ItemSet | Support (%) | ItemSet | Support (%) |
| 1 | Jam | Butter | 16.45 | Bucket | 17.5 |
| 2 | Bucket | Soap | 17.8 | Jam | 17.5 |
| 3 | Juice | Bread | 17.25 | Cheese | 17.45 |

From combining the patterns , final actionable rules are obtained as shown in table 4 by applying the algorithm to get actionable patterns given in section 2.4.

**Table 4. Actionable rules**

| RULES | SOURCE1 | SOURCE2 | Actionable Rules |
|---|---|---|---|
| Jam -> Butter | 16.45 | 16.5 | ✓ |
| Jam -> Bucket | 16.5 | 17.5 | ✓ |
| Bucket -> Soap | 17.5 | 16.95 | ✓ |
| **Bucket -> Jam** | **——-** | **17.5** | |
| **Juice - > Bread** | **17.25** | **——-** | |
| **Juice -> Cheese** | **——-** | **17.45** | |

From table 3, it is observed that the rule Jam → Butter and Jam → Buckets are two different rules obtained on 2 different sources as source 1 and 2 respectively. Here the Antecedent Jam is common in both the rules with different support values. Upon finding such rules from both the sources, it is observed that rule 1, 2 and 3 from table 4 are common on both sites while remaining rules are present either one of the source as indicated in table 4. Summarizing the results it's observed that only first 3 rules from table 4 are the part of actionable rules in addition to the common rules obtained from source 1 and 2 respectively.

Consider highlighted rules from Table 1 and Table 2, which are common on both sources. Upon combining those rules with the marked rules from table 4, the final Actionable Rules are as given below :

1. Bread → Soap
2. Soap → Bread
3. Butter → Bread
4. Bread → Butter
5. Jam → Butter
6. Jam → Bucket
7. Bucket → Soap

To compare and prove the appropriateness of the proposed technique we applied the association rule mining technique on integrated datasets by transferring the transactional data from source 1 and 2 to the central site. Considering same values for support = 200 and confidence = 40% , 62 rules got generated which are above threshold value of which top 10 rules are mentioned in table 5. It is observed that rule Jam → Bucket is missing in table 5 which was present in both the sources but the same can be obtained by applying the proposed method. Similarly rule number 4,7,9,10 from table 5 are unwanted rules and reflected in the output if we apply the data integration method.

**Table 5. Rules obtained from integrated data**

| Sr. No. | Antecedent (A) | Consequent (C) | Support for A & C |
|---|---|---|---|
| **1** | Bread | Soap | **18.15** |
| **2** | Soap | Bread | **18.15** |
| **3** | Jam | Butter | **16.45** |
| **4** | **Butter** | **Jam** | **16.45** |
| **5** | Butter | Bread | **17.2** |
| **6** | Bread | Butter | **17.2** |
| **7** | **Soap** | **Bucket** | **17.8** |
| **8** | Bucket | Soap | **17.8** |
| **9** | **Juice** | **Bread** | **17.25** |
| **10** | **Bread** | **Juice** | **17.25** |

So our proposed algorithm generates more actionable rules as compare to traditional method of integration of data from multiple sources.

## 4. CONCLUSION

The aforesaid method produces the patterns which assist in concluding the appropriate and beneficial business decisions. It minimizes the cost of transportation as only the patterns generated at different sources are sent to the central site. On the other hand while applying the association rule mining for the clubbed data many rules may get neglected which are useful. Time and space complexity in such cases is also higher than the aforesaid algorithm.

Based on the patterns obtained, the experts may arrive at the conclusion for promotion or recommendation by applying appropriate domain knowledge of business. For retail Industry, the products of which the usages are about to expire can be put forth for offers with other items in the pattern.

For telecommunication industry combinations of various voice and data plans can be offered. For insurance company combination of policies may be offered. This work can be further extended for behavioral patterns based on demographic data.

## 5. REFERENCES

[1] Longbing Cao, Philip S. Yu, Chengqi Zhang, Yanchang Zhao, " Domain Driven Data mining ," In Springer ISBN 978-1-4419-5736-8 , science+business media, LLC 2010

[2] Longbing Cao, Huaifeng Zhang Yanchang Zhao, Dan Luo, and Chengqi Zhang, "Combined Mining : Discovering Informative Knowledge in Complex Data, " In IEEE Trans Systems, Man, And Cybernetics, Vol. 41, No. 3, June 2011pp-699-712.

[3] Paul O' Dea, Josephine Griffith, Colm O' Riordan, "Combining Feature Selection and Neural Networks for Solving Classification , "In Article 7/2001 CiteSeer

[4] Jia Hu, NingZhong, "Organizing Multiple Data Sources for Developing Intelligent e-Business Portals," In Springer Science + Business Media, Inc. Manufactured in the United States ,Data Mining and Knowledge Discovery, 12, 127-150, 2006

[5] Chuen-He Liou, "Improve the Quality of Product Recommendation based on Multi-channel CRM for E-commerce," In International Conference on Data Mining, DMIN'13

[6] Xiaoxin Yin and Jiawei Han, "Efficient Classification from Multiple Heterogeneous Databases," In Knowledge Discovery in Databases PKDD 2005, 2005 - Springer

[7] Arti Deshpande and Anjali Mahajan , "Domain Driven Multi-Feature Combined Mining for retail dataset, "InInternational Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 - 8958, Volume-2, Issue-3 February 2013

[8] Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang, and Hans Bohlscheid, "Combined Pattern Mining: From Learned Rules to Actionable Knowledge , " In Lecture Notes of Computer Science, 2008, Volume 5360/2008, pp. 393-403

[9] Arti Deshpande and Anjali Mahajan, "Serial Multimethod Combined Mining," In ICACCI, 2014 International Conference on Advances in Computing, Communications and Informatics 978-1-4799-3080-7/14, IEEE

[10] http://www.red-gate.com/products/sql-development/sql-data-generator/

[11] Mohammed Al-Maolegiand BassamArkok, "An Improved Apriori Algorithm for Association Rules," In International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014

[12] Charls X. Ling and Qiang Yang, "Discovering Classification from Data of Multiple Sources," In Data Mining and Knowledge Discovery, 12, 181–201, 2006.

[13] Sayyadian, Mayssam. "HeteroClass: a framework for effective classification from heterogeneous databases." CS512 Project Report, University of Wisconsin, Madison (2006).