WORKING PAPER NO: 546

An Affordance-based Substitution Mechanism to Mine Positive and Negative Association Rules

Rupal Sethi
Doctoral Student
Decision Sciences and Information Systems Area,
Indian Institute of Management Bangalore,
Bangalore, 560 076, India
rupal.sethi14@iimb.ernet.in

B Shekar
Professor
Decision Sciences and Information Systems Area
Indian Institute of Management Bangalore
Bannerghatta Road, Bangalore – 5600 76
Ph: 080-2699 3093
shek@iimb.ernet.in

Year of Publication – March 2017

An Affordance-based Substitution Mechanism to Mine Positive and Negative Association Rules

Abstract

Negative association among products is an upcoming topic of interest in the field of association rule mining. Although researchers have given considerable attention to this idea, we have analyzed some key unaddressed issues in negative association rule discovery and generation. Substitution of items based on their position in the taxonomy is an interesting method to link positive and negative rules. However, we argue that the definition of substitution is not limited to just sibling positions in the taxonomy. Two items can act as substitutes based on the goals and past experiences of the customer. We capture this idea of substitutability using the lens of Affordances to link *Product, Purpose and Person* notion of Marketing to the generation of negative rules in data mining. This is done by matching the affordances (action possibilities) of products specified by the managers with the customer's goals to form substitutable sets of items corresponding to each goal. These sets of substitutable items are then used to form substitution rules. The deviation from the expected confidence of generating these substitute rules, result in formation of interesting negative rules.

Keywords: association rule mining; negative rules; substitution; affordances; marketing; categories; taxonomy; subjective interestingness.

1. Introduction

Association Rule Mining (ARM) is one of the most popular techniques of data mining that discovers relationships between group of items (itemsets) of a particular transaction database. ARM finds its applications in many domains such as Market Basket Analysis, Demographic and Profile Analysis, Recommendation Systems and Collaborative Filtering, Web Log Analysis and Bioinformatics (Charu Aggarwal, 2015). These popular models for finding association rules use *frequency count* as quantification of level of association between items. For example, if we consider Market Basket Analysis, the frequency count of various items in the transaction set is a surrogate for measuring customer behaviour related to that supermarket. These 'interesting' patterns of data enable managers of the store to formulate important strategic decisions such as shelf-positioning or bundling of items that are purchased together.

There have been many algorithms like Apriori (Agrawal and Srikant, 1994), FP growths (Han, 2000) or Eclat (Zaki, 1997) that mine association rules based on frequency counts in transaction data. However, all these algorithms are restricted to positive association rules only. By positive association rule, we mean relationship between items or group of items that exist in the transaction set. For instance, a person who is buying bread is likely to buy butter as well. Contrary to this, we have another category of association rules which are negative association rules. These depict relationship between items that conflict each other, like people who buy Pepsi *do not* buy Coke. Thus, in market-basket analysis, negative association rules identify those items that a customer is not likely to buy given that he buys a certain set of items (Savasere, Omiesinski and Navathe, 1998). These negative rules provide valuable information to managers for making strategic decisions concerning products that are getting obsolete and are being replaced by newer ones (substitutes) and products which are bought together across categories (complements).

In this paper, we attempt to understand the concept of substitution in order to discover and generate negative association rules. We use the theoretical lens of functional affordances to link user goals and manager's knowledge to define sets of substitutable items corresponding to a particular goal. Once we have the sets of substitutable items, we use them to define pair of association rules which we call substitution rules. This pair of rules helps us link the knowledge combined from a positive and negative AR. We also define a measure of objective interestingness based on expected confidence levels for negative rules generated.

2. Literature Review

The work in the area of discovery and generation of negative association rules is limited. Since a typical Apriori algorithm cannot be used for generation of negative AR, there has been some work in developing new algorithms. Brin, Motwani and Silverstein (1997) argued the need for extending the support-confidence framework and using correlation coefficient for tapping negative relationships between itemsets. Antonie and Zaiane (2004), extending the work of Brin et al (1997), used Pearson's correlation coefficient as a measure of negative association and developed an algorithm to generate negative AR with a sliding correlation coefficient threshold.

The idea of using the concept of substitutability in negative AR came from the work of Savasere, Omiecinski and Navathe (1998). They used the taxonomy approach to define substitutability based on the position of the item in the taxonomy. They restricted their definition only to sibling substitution wherein items that belong to the sibling positions in the taxonomy exhibit similar behaviour and hence can be substitutable. They use "unexpectedness" as an objective measure of interestingness for negative AR. A negative rule is interesting if it deviates from the expectation based on previous belief relating to the itemset.

A similar approach to Savasere et al (1998) has been adopted by Buckles, Yuan and Zhang (2002). They also use the concept of locality of similarity in defining siblings rules from the taxonomy. Sibling rules are a pair of positive association rules where both the siblings are expected to be related to the same consequent. For example, if *Pepsi* → *Chips* is a rule that is generated through Apriori, then *Coke* → *Chips* should also be generated. If the confidence measures of *Coke* → *Chips* is less than the expected confidence (equal to *Pepsi* → *Chips*), then a negative rule *Coke* → ¬ *Chips* is generated.

Given our limited knowledge of the literature on negative association rules, we find a limited and ambiguous definition of substitutability being incorporated (Savasere et al, 1998; Buckles et al, 2002). Hence, in this paper we attempt to define substitutable using the lens of affordances to link the subjective notions of users' goals and managers' domain knowledge.

3. Substitution in Market Basket Analysis

There are various definitions of substitutes available in the literature of marketing, information systems and economics. A Substitute is defined as a product that is same or similar to another product in customers' perceptions. Another definition is that two goods are substitutes, if one good may, as a result of changed conditions, replace the other in use (Nicholson and Snyder, 2011). In both these definitions, there is one thing common that defines the conditions for substitution- either customer perceptions or changed environment. This shows that one needs to focus on what function defines substitution of two products.

Hence, the question we are addressing here is:

When do two products become substitutes to each other?

Clearly, we shall look at the definition of substitution of products in the domain of market basket analysis. A lot of literature on association rule mining look at the taxonomy structure of the data to define interesting patterns (Srikant and Agrawal, 1995; Liu, Hsu, Chen and Ma, 2000; Savasere et al, 1998).

Taxonomy in market basket analysis is a scheme of classification of products that belong to a particular supermarket. It is represented as a hierarchical tree structure with parents as classes and children as specialized products in that class. The taxonomy captures the domain knowledge of the manager of that supermarket in terms of classification of all the products. However, the basic question on taxonomy that has received limited attention in the marketing literature is two fold: 1) How do you define categories in a taxonomy? 2) How do you distinguish between various categories? (Ratneshwar and Shocker, 1991). The categorization has different connotations for the buyer as well as the producer. According to the buyer, categories should reflect similarities within the same categories and differences across different categories. Categorization is a means for simplifying information, better decision making and efficient interpersonal communication for the buyers (Shocker, Bayus and Kim, 2004). Some of the parameters on which the buyers categorize products are physical resemblance, perceived similarity of producers or fit with category label (Day, Shocker and Srivastava, 1979). From the seller's point of view, categorization should promote new products among the buyers. Also, it should be easy to integrate communications with other stakeholders like distributors and logistics. Two other factors that shape the dynamic categorization of products are word of mouth from other customers and seasonality. Thus, product categorization is dependent on behaviors of both buyers and sellers. Table 1 below shows the factors of product categorization from both buyer and seller's perspectives.

Buyer	Seller
➢ Physical resemblance	➢ New Product Development
➢ Brand resemblance	➢ Word of Mouth
➢ Perceived similarity of producers	➢ Integrate with distributors
➢ Better decision making and communication	➢ Seasonality of products

Table 1: Factors of product categorization from buyer and seller's perspectives

Likewise, one may develop an argument for substitution of products across categories also. While category definition involves the above mentioned factors, substitutions from one or different categories may also be contingent on the following factors:

- Between different brands like Pepsi and Coke
- Between physically resembling products like Butter and Margarine
- Based on buyer's goals and motives like Chocolates and Teddy bear
- Based on seasonality or events like Chips and Pepsi

On the prima facie level, buyers tend to substitute products in the same category because of the similarity of function served by the products. However, the role of substitution can be more abstract, based on inter-category replacement of products. Depending on the usage situation or the customer's goals, two products in distant categories can act as substitutes. For example, on the occasion of valentine's day, a guy may purchase a teddy bear or chocolates, whichever is available or less expensive. Hence the substitution here is based on the situation or the context rather than brand or physical resemblance of products.

4. Affordances

Thus far, we have argued that customers' purpose for buying a product cannot be neglected when we are defining substitutability. As a result, we use Affordances as lens to develop our theoretical model. The notion of Affordances was originally introduced by psychologist James Gibson in his article "The Theory of Affordances" (Gibson, 1977). According to Gibson, Affordances are potential benefits and disadvantages of an object, relational to the agent using that object. For example, a sweet *affords* chewing if the user has teeth. Since this definition is applied to cognitive psychology, it highlights the complementarity between the organism (agent) and its environment necessitating that the studies on the organism (humans, in this case) should be conducted in its natural environment rather than in isolation. This idea of agent and environment was picked up by Donald Norman in redefining the concept of Affordances in the context of Human-Machine Interaction. According to Gibson, the "action possibilities" depended on the agent's physical capabilities. However, Norman extended it, adding that actor's goals, plans, values, beliefs and past experiences also account for while evaluating the benefits or disadvantages from the use of that object (Norman, 1999).

Norman's definition has been extensively adopted by system-theoretic approaches and design problems using a human-factors approach. In marketing and consumer behavior, Ratneshwar, Shocker, Cotte, and Srivastava (1999) have come up with a theoretical model linking **product, person and purpose** using the lens of affordances. Their proposition states that: *The perceived benefits of a product are a function of its affordances to an individual consumer in relation to that person's purpose. Thus, perceived benefits represent interactions between a product's affordances and a consumer's purpose.*

We use this notion of product, person and purpose to define substitutability of products in market basket. Product substitutability cannot be considered without controlling for the effects of purpose. Since products serve multiple purposes, user can 'afford' them based on their goals. In addition to the goals, the buyer's past experience and knowledge is also an important factor in defining product categories and hence substitutability between items (Shocker et al, 2004). For instance, a person who is already using a windows operating system in his laptop and is satisfied with it, will buy a Nokia mobile phone with the same OS. Thus, the past experience and knowledge of the user regarding the OS helped him in his new purchase.

The literature in association rule mining focusses on subjective interestingness from the point-of-view of the manager. The subjective measures take into account the domain knowledge of the manager and associate it with the data semantics from the transaction set. However, in this paper we attempt to view subjective interestingness in case of product substitutability from both manager's and user's point-of-view.

5. Framework

We try to capture user's previous knowledge and current goals of purchase. This shall result in knowing the context of product purchase by that user and in turn help us in finding his rationale for substitution between two sets of products. The substitution decision shall be made as a result of the mapping between the products' affordances and users' goals. The affordances or the potential action possibilities of a product will be subject to the manager's domain knowledge. Hence, subjectively of the manager is instrumental in describing various affordances related to all products in his supermarket. These affordances of products shall be matched with customer goals forming substitutable sets of products.

Let products be denoted by P_1 , P_2 …. P_n each with a set of affordances identified by the manager (a_{11} , a_{12} … a_{1m}). The set of goals captured by the users is denoted by g_l to g_k .

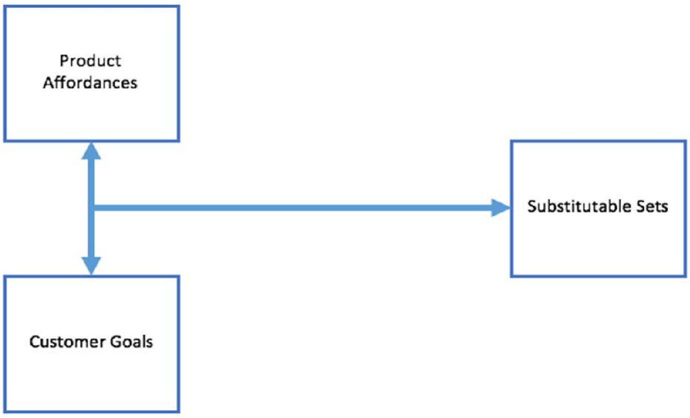

Figure 1: Conceptual framework for linking manager and user's knowledge in substitution

A matching algorithm is used to map product affordances with customer goals. The matched products are combined in sets that are substitutable for those specific goals.

5.1 An Illustration

Consider a sample of Product-Affordances set of 3 products in Table 2.

Product	**Affordances**
P_1 : Chocolate	a_{11} : Sweet
	a_{12} : Gift
	a_{13} : Waxing
	a_{14} : Medicinal Purpose
P_2 : Flowers	a_{21} : Decoration
	a_{22} : Gift
	a_{23} : Ingredient in food
	a_{24} : Medicinal Purpose
P_3 : Ice-cream	a_{31} : Sweet
	a_{32} : Ingredient in food

Table 2: Manager's knowledge of products as affordances

Let the goals captured by three different users be denoted by Table 3.

g_1	Gift for Valentine's day
g_2	Sweet for Diwali
g_3	Ingredient in making some food item

Table 3: Sample goals from selected users

We now match the affordances of each product with the three goals to get the set of similar items based on potential actions and user goals.

Goals	Set of Substitutable items
g_1 : Gift for Valentine's day	Chocolate and Flowers
g_2 : Sweet for Diwali	Chocolate and Ice-cream
g_3 : Ingredient in making some food item	Flowers and Ice-cream

Table 4: Generation of substitutable sets for each goal

Thus, a customer who comes for purchasing a gift for Valentine's day may choose between either chocolate or flowers, based on other factors such as user preferences and quality of product.

The goals specified by the user will have seasonality element as well. Those nuances need to be specified as refinement in the matching algorithm.

6. Substitution Rules

We now use the substitution sets generated based on each goal of the user to mine substitution rules from the transaction dataset. Buckles et al (2002) define sibling rules based on the position of the items in the taxonomy. One of the gaps in their approach is that siblings in the taxonomy may not be substitutable items for a particular customer. Hence, our approach matches the manager's knowledge in terms of the affordances provided by the product and the customer goals for purchase.

The items that belong to one substitution set based on goal g_i can now be a part of the substitution rules similar to the method defined by Buckles et al (2002).

Let $Y \rightarrow X$ be a positive rule that is generated by Apriori algorithm such that

$$supp(Y \rightarrow X) \geq minsupp$$
$$conf(Y \rightarrow X) \geq minconf$$

where Y is any itemset and X itemset contains items $\{i_1, i_2, \ldots i_k, \ldots i_m\}$.

Now, let X' be another itemset denoted as $\{i_1, i_2, \ldots i_h, \ldots i_m\}$ such that

$X' = X - i_k + i_h$ where i_h and i_k are items in the substitution set for goal g_i.

Consider for a possibility that the inequality

$$conf(Y \rightarrow X') \geq minconf \qquad (1)$$

As a result, $Y \rightarrow X$ and $Y \rightarrow X'$ are substitution association rules.

Please note that in our case, we do not consider the converse rule as mentioned by Buckles et al (2002) due to the uncertainty in the individual probabilities of X and X'. We assume that keeping the antecedent same in the set of substitution rules, the likelihood of confidence being greater than the minimum threshold is higher as compared to rules $\{X \rightarrow Y, X' \rightarrow Y\}$.

Using the hypothetical transactions we consider in Table 2, consider the following example of substitution rules.

R1: *Greeting card* → *Chocolate*
R2: *Greeting card* → *Flowers*

According to our substitution set of {*Chocolates, Flowers*}, a person who wishes to buy a gift for any occasion will equally-likely choose between chocolates and flowers along with greeting card.

6.1 Linking Positive and Negative Rules

The set of substitution rules are formed only if (1) holds true.

However, if the association is not supported, it means that *X'* is not related with *Y* itemset. In that case, we generate negative association rule *Y*→ ¬ *X'* based on our expectation of the base rule *Y* →*X*.

There would be many cases where condition (1) would not hold true simply because we are dealing with users' purchase behavior and not an objective metric. As a result, one needs to incorporate some measure of interestingness to the negative rules generated so that they are actionable by the manager to study the consumer behavior.

$$Rule\ Interestingness, RI = conf(Y \rightarrow X) - conf(Y \rightarrow X')$$

Rule Interestingness for the negative rule *Y*→¬ *X'* is the distance between the confidence levels of the base rule and its substitute rule. It refers to the deviation in expectation of the confidence in both substitution rules.

For instance, consider minimum confidence to be 50%.

Given below are the confidence values for the frequent itemsets:

R1: {*Greeting card , Chocolate*} conf = 0.75
R2: {*Greeting card , Flowers*} conf = 0.25

R1 rule would be generated since $conf(R1) \geq minconf$.

But, $conf(R2) < minconf$

Thus, negative rule *Greeting card*→¬ *Flowers* would be generated with rule interestingness

$$RI = 0.75 - 0.25 = 0.5$$

The deviation is quite high; hence the negative rule seems interesting.

7. Conclusions

We analyze the previous research done in mining negative association rules. We come up with a few important unaddressed issues that could be viewed for a better understanding of negative association between items. Particularly the case of market basket analysis has been taken up by researchers to map associations of one product with absence of another. Objective as well as subjective measures of interestingness have been used to evaluate and generate negative relationships among items. One key concept that has received least attention while mining negative rules is substitution. Savasere et al (1998) and Buckles et al (2002) use the taxonomy approach to generate negative relationships based on the position of the product in the taxonomy. They assume that siblings in a taxonomy are perfect substitutes and their relationship with the other itemsets should be identical. Any deviation in this expected relationship shall result in negative rules. However, we extend the definition of substitution of products beyond the taxonomy approach. We argue that substitution is also contingent on user's goals and context of purchase. For example, a customer might substitute chocolates for flowers if he has to gift it to someone on Valentine's day. Since chocolates and flowers may not lie close to each other in any generic taxonomy of a supermarket, this rule might not be generated using the approach of Buckles et al (2002).

In order to develop this definition of substitution in generation of positive and negative rules, we use the lens of Affordances. Affordances is an upcoming concept in the area of Information Systems, Marketing and Consumer Behavior. Affordances are *action possibilities* of an object that depend on its agent's physical capabilities along with his goals, plans, values, beliefs and past experiences. We use affordances of products specified by the manager using his domain knowledge and match it with the customer's goals and past experiences involved in the purchase. This matching gives us sets of substitutable items contingent on goals specified by the customer. Hence, the items in these sets act as substitutes for those segment of customers who purchase for that particular goal.

These substitution sets are then used to generate substitution rules similar to sibling rules introduced by Buckles et al (2002). If the substitution set does not fulfil the minimum confidence threshold, it is categorized as a negative rule with converse consequent. We also define a rule interestingness measure of the negative rules that are generated based on our expectation from the base substitution rule. Illustrations are provided at each step of the paper for simplicity and we plan to develop an algorithm for the same subsequently.

References

1. Aggarwal, Charu C. (2015). Data Mining: The Textbook. Springer

2. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

3. Antonie, M. L., & Zaïane, O. R. (2004). Mining positive and negative association rules: an approach for confined rules. In Knowledge Discovery in Databases: PKDD 2004 (pp. 27-38). Springer Berlin Heidelberg.

4. Brin, S., Motwani, R., & Silverstein, C. (1997a, June). Beyond market baskets: Generalizing association rules to correlations. In ACM SIGMOD Record (Vol. 26, No. 2, pp. 265-276). ACM.

5. Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997b, June). Dynamic itemset counting and implication rules for market basket data. In ACM SIGMOD Record (Vol. 26, No. 2, pp. 255-264). ACM.

6. Day, G. S., Shocker, A. D., & Srivastava, R. K. (1979). Customer-oriented approaches to identifying product-markets. The Journal of Marketing, 8-19.

7. Gibson, J. J. (1977). The theory of affordances. Hilldale, USA.
8. Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In ACM SIGMOD Record (Vol. 29, No. 2, pp. 1-12). ACM.

9. Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing the subjective interestingness of association rules. Intelligent Systems and their Applications, IEEE, 15(5), 47-55.

10. Nicholson, W., & Snyder, C. (2011). *Microeconomic theory: basic principles and extensions*. Cengage Learning.

11. Norman, D. A. (1999). Affordance, conventions, and design. *interactions*, *6*(3), 38-43.

12. Pearson, K. (1895). Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 240-242.

13. Ratneshwar, S., & Shocker, A. D. (1991). Substitution in use and the role of usage context in product category structures. Journal of Marketing Research, 281-295.

14. Ratneshwar, S., Shocker, A. D., Cotte, J., & Srivastava, R. K. (1999). Product, person, and purpose: putting the consumer back into theories of dynamic market behaviour. Journal of Strategic Marketing, 7(3), 191-208.

15. Savasere, A., Omiecinski, E., & Navathe, S. (1998, February). Mining for strong negative associations in a large database of customer transactions. In Data Engineering, 1998. Proceedings., 14th International Conference on (pp. 494-502). IEEE.

16. Shocker, A. D., Bayus, B. L., & Kim, N. (2004). Product complements and substitutes in the real world: The relevance of "other products". Journal of Marketing, 68(1), 28-40.

17. Srikant, R., & Agrawal, R. (1995). Mining generalized association rules (pp. 407-419). IBM Research Division.

18. Yao, Y. Y., & Zhong, N. (1999). An analysis of quantitative measures associated with rules. In Methodologies for knowledge discovery and data mining (pp. 479-488). Springer Berlin Heidelberg.

19. Yuan, X., Buckles, B. P., Yuan, Z., & Zhang, J. (2002). Mining negative association rules. In Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on (pp. 623-628). IEEE.

20. Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997, August). New Algorithms for Fast Discovery of Association Rules. In KDD (Vol. 97, pp. 283-286).