General Advance Review of Borgelt's Apriori Rule Mining Algorithm Implementation : Business Intelligence and Data Science Perspective

Niket Bhargava
Department of Computer Science and Engineering
Mewar University
Chittaurgarh, India
niket.bhargava.datascience@gmail.com

Dr. Manoj Shukla
Professor, Department of Electronics and Communication
BCE, Mandideep
Bhopal, India
dr.manojsh@gmail.com

Abstract—**In this paper we studied Finding Frequent Item Sets and Association Rules with the Apriori Algorithm. This comparative study refers to Borgelt's Apriori implementation version 6.12 in particular and original Apriori by R. Agrawal and R. Srikant. This paper covers introduction to Apriori, an Association Rule Mining algorithm and discussed related issues like Basic Notions of Items and Transactions, Support of an Item Set, Confidence of an Association Rule, Support of an Association Rule, and also discussed Target Types like Frequent Item Sets, Closed Item Sets, Maximal Item Sets, Generators/Free Item Sets, Association Rules, and Extended Rule Selection and their variations, for confidence, lift, conviction, χ^2-Measure, p-value, and importance of Fisher's Exact test for table probability, χ^2-Measure, information gain, support, and in the end item set selection methods are discussed. All these measures are very important to invent new data mining algorithms and they all have very strong utilization and application in data science and business intelligence. This paper also discuss The possible application of association rule mining from the perspective of Business Intelligence and Data Science. This review is important as this implementation is useful for Business Analysts, Data Scientists and data miners who want to exploit association rules to take decisions using R programming language to find solution fo business problems.**

Keywords- Apriori , Implementation, Rule Type, Target Type, Interestingness, Measures

I. Introduction

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).[1] Data Mining sometimes also known as Big Data and Data Science. Business Intelligence is use of computation techniques to improve business reporting. Business intelligence (BI) is a technology-driven process for analyzing data and presenting actionable information to help corporate executives, business managers and other end users make more informed business decisions.[2]

Frequent item set mining and association rule induction are powerful methods for so-called market basket analysis, which aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, online shops etc[9].

Induction is the process of finding general conclusion from a set of specific examples[3]. Deduction is the reverse process of induction[3]. With the induction of frequent item sets and induction of association rules one tries to find sets of products that are frequently bought together, so that from the presence of certain products in a shopping cart one can infer (with a high probability) that certain other products are present.

Such information, especially if expressed in the form of rules, can often be used to increase the number of items sold, for instance, by appropriately arranging the products on the shelves of a supermarket or on the pages of a mail-order catalog (they may, for example, be placed adjacent to each other in order to invite even more customers to buy them together) or by directly suggesting items to a customer, which may be of interest for him/her.

An association rule is a rule like "If a customer buys wine and bread, he/she often buys cheese, too." It expresses an association between (sets of) items, which may be products of a supermarket or a mail-order company, special equipment options of a car, optional services offered by telecommunication companies etc. An association rule states that if we pick a customer at random and find out that he/she selected certain items (bought certain products, chose certain options etc.), we can be confident, quantified by a percentage, that he/she will also select certain other items (will buy certain other products, will chose certain other options etc.).

Of course, we do not want just any association rules, we want "good" useful and interesting rules, rules that are "useful", "expressive" and "reliable". The standard measures to assess association rules are the support and the confidence of a rule known as support-confidence framework for rule evaluation, support and confidence both of which are computed from the support of certain item sets. These notions are discussed in the next section 2 in more detail. However, these standard criteria are often not sufficient to restrict the set of rules to the interesting ones. Therefore some additional rule evaluation measures are considered.

This paper is further divided into following sections, section 2 discuss basic notions. Section 3 discuss Target Types, section 4 discuss extended rule selection, section 5 extended rule selection discuss, section 6 discuss Transaction Prefix Tree, section 7 discuss future work.

II. Basic Notations

This section introduces the basic notions needed to talk about frequent item sets and association rules. These notions are item, transaction, support and confidence.

A. Items and Transactions

On an abstract level, the input to frequent item set mining and association rule induction consists of a bag or multiset of transactions that are defined over a set of items, sometimes called the item base. These items may be products, special equipment items, service options etc. For an abstract treatment it suffices that items have an identity, that is, it is possible to distinguish one item from another. The item base is the set of all considered items, for example, the set of all products that are sold by a given supermarket, mail-order company or online shop.

Any subset of the item base is called an item set.

A transaction is simply an item "set" and it represents, for example, the set of products bought by a customer. Since two or more customers may, in principle, buy the exact same set of products, we cannot model the whole of all "shopping baskets" or "shopping carts" (bought, say, in a given week) as a set of transactions, since in a set each element is unique.

There are several solutions to this problem: one may, for example, model the whole of all transactions as a bag or multiset (a generalization of a set, which allows for multiple occurrences of the same element) or as a vector (where elements at different positions may be the same, but are still distinguished by their position), or by extending each transaction with a unique transaction identifier(or tid for short; note that the position of a transaction in a vector representation is an implicit transaction identifier). Still another possibility consists in using a standard set of (unique) transactions and assigning to each of them an occurrence counter. This page exploits the bag/multiset terminology, even though an occasional "set of transactions" may have slipped through. (This should always be read as "bag/multiset of transactions".)

Note that the item base (the set of all considered items) is often not given explicitly, but only implicitly as the union of all given transactions. This is also the case for Borgelt's Apriori program, which by default only takes a transaction file as input. However, it is also possible to specify the item base explicitly with an optional item appearances file. This can be useful, for example, if one wants to restrict the analysis to a subset of all items. It can also be used to specify that certain items should only appear in the antecedents or only in the consequents of reported association rules.

B. Support of an Item Set

Let S be an item set and T the bag/multiset of all transactions under consideration. Recall cardinality and multiplicity terms from set theory. Multiplicity is used with bag/multiset. Then the "absolute" support(or simply the support) of the item set S is the "number" of transactions in T that contain S. Likewise, the "relative" support of S is the fraction (or "percentage") of the transactions in T which contain S.

More formally, let S be an item set and U = {X□T|S□t} the bag/multiset of all transactions in T that have S as a subset (i.e. contain all of the items in S and possibly some others). Then

$\text{supp}_{abs}(S)$ = |U| =|{X□T|S□t}|,

is the absolute support of S, and

$\text{supp}_{rel}(S)$=(|U|/|T|)*100%, is the relative support of S.

Here |U| and |T| are the number of elements in U and T, respectively.

In a supermarket setting, the item set S may be a set like S={bread, wine, cheese} and T may be the bag/multiset of all "baskets" or "carts" of products bought by the customers of a supermarket – in a given week if you like. U is the bag/multiset of all transactions in T that contain all items in S (and maybe also some other items). For example, if a customer buys the set X = {milk, bread, apples, wine, sausages, cheese, onions, potatoes}, then S, S={bread, wine, cheese}, is obviously a subset of X, hence X is in U. If there are 318 customers, each giving rise to one transaction, and 242 of these customers bought such a set X or a similar one that contains S, while the other customers bought sets of products that lacked at least one of the items in S, then $\text{supp}_{abs}(S)$ = 242 and $\text{supp}_{rel}(S)$ =242/318 * 100 =76.10%.

The goal of frequent item set mining is to find all item sets (that is, all subsets of the item base) that occur in the given bag/multiset of transactions with at least a user-specified minimum support supp_{min}. Such item sets are called frequent item sets.

The default value for the minimum support in Borgelt's Apriori program is 10% (the percentage indicates implicitly that it refers to "relative" support). This value can be changed "with" the option -s. Note that the argument to this option is interpreted as a percentage if it is “positive” example -s20, here 20 is positive, but if it is “negative” i.e. -s-20, here -20 is negative, it is interpreted as an absolute number (number of transactions) rather than a percentage. That is, -s20 means a minimum relative support of 20%, while -s-20 means a minimum absolute support of 20 transactions.

Note that the default operation mode i.e “target type” of Borgelt's Apriori program is to find such frequent item sets. In order to find association rules, the “target type” has to be changed.

C. Confidence of an association rule

If we search for association rules, we do not want just any association rules, but "good" association rules. To measure the “quality of association rules”, [Agrawal and Srikant 1994], the inventors of the Apriori algorithm, introduced the confidence of a rule. The confidence of an association rule R = "X→Y" (with item sets X and Y) is the support of the set of all items that appear in the rule (here: the support of S = X □ Y) divided by the support of the antecedent (also called "if-part" or "body") of the rule (here X). That is,

$$\text{conf}(R)= \frac{\text{supp}(X \,\square\, Y)}{\text{supp}(X)}$$

(Note that it does not matter whether the confidence is computed from the absolute or the relative support of an item set, as long as the same “support type” is used in both the numerator and the denominator of the fraction.)

More intuitively, the confidence of a rule is the number of cases in which the rule is “correct relative” to the number of cases in which it is applicable. For example, let R = "wine and bread → cheese". If a customer buys wine and bread (and maybe some other items), then the rule is applicable and it says that he/she can be expected to buy cheese. “If he/she does not buy wine or does not buy bread or buys neither, then the rule is not applicable” and thus (obviously) does not say anything about this customer.

“If the rule is applicable, it says that the customer can be expected to buy cheese”. But he/she may or may not buy cheese, that is, the rule may or may not be correct (for this customer). Naturally, we are interested in how good the “prediction” of the rule is, that is, how often its prediction that the customer buys cheese is correct. The “rule confidence” measures this: it states the percentage of cases in which the rule is correct. It states this percentage relative to the number of cases in which the antecedent holds, since these are the cases in which the rule makes a prediction that can be true or false. If the antecedent does not hold, then the rule does not make any prediction, so these cases are excluded.

Rules are reported as association rules if their confidence reaches or exceeds a given lower limit (“minimum confidence”; to be specified by a user). That is, we look for rules that have a high probability of being true: we look for "good" rules, which make correct (or very often correct) predictions. Apriori program always uses a minimum confidence to select association rules. The “default value for the minimum confidence is 80%”. This value can be changed with the option-c. (Note that for the minimum confidence, the argument is “always interpreted as a percentage”. Negative values cause an error message, because there is no "absolute confidence".)

In addition to the rule confidence, Borgelt's Apriori program lets you select from several other (additional) rule evaluation measures, which are explained in section 2.4, but it will also use rule confidence. If you want to rely entirely on some other measure, you can do so by setting the minimal rule confidence to zero. (Attention: If you have a large number of items, setting the minimal rule confidence to zero can result in very high memory consumption. Therefore: use this possibility with a lot of care, if at all.)

D. Support of an association rule

The support of association rules may cause some confusion, because Borgelt used this term in a different way than [Agrawal and Srikant 1994] do. For them, the support of an association rule "A and B→C" is the support of the set S={A,B,C}. This may be fine if rule confidence is the only evaluation measure, but it causes problems if some other measure is used. For these other measures it is often much more appropriate to call the support of the antecedent of the association rule, that is, the support of X ={A,B} in the example above, the support of the association rule "A and B→C".

The difference can also be stated in the following way: for [Agrawal and Srikant 1994], the support of the rule is the (absolute or relative) number of cases in which the rule is correct (that is, in which the presence of the item C follows from the presence of the items A and B), whereas for Borgelt's (and thus Borgelt's Apriori program) the support of a rule is the (absolute or relative) number of cases in which it is applicable (that is, in which the antecedent of the rule holds), although in some of these cases it may be false (because only the items A and B are present, but the item C is missing).

One reason for this choice, as already mentioned, is that the definition of [Agrawal and Srikant 1994] does not work well for evaluation measures other than rule confidence. This is explained in more detail below. Another reason is that Borgelt preferred the support of a rule to say something about the "statistical" support of a rule and its confidence, that is, from how many cases the confidence is computed in order to express how well founded the statement about the confidence is.

Maybe an example will make this clearer. Suppose you have a die which you suspect to be biased. To test this hypothesis, you throw the die, say, a thousand times. 307 times the 6 turns up. Hence you assume that the die is actually biased, since the relative frequency is about 30% although for an unbiased die it should be around 17% (1/6 in 1000 so in 100 it is 16.66) . Now, what is the "statistical" support of this statement, that is, on how many experiments does it rest? Obviously it rests on all 1000 experiments and not only on the 307 experiments in which the 6 turned up. This is so, simply because you had to do 1000 experiments to find out that the relative frequency is around 30%, and not only the 307 in which a 6 turned up (doing only these experiments is obviously impossible).

Or suppose you are doing an opinion poll to find out about the acceptance of a certain political party, maybe with the usual question "If an election were held next Sunday ...?" You ask 2000 persons, of which 857 say that they would vote for the party you are interested in. What is the support of the assertion that this party would get around 43% of all votes? It is the size of your sample, that is, all 2000 persons, and not only the 857 that answered in the positive. Again you had to ask all 2000 people to find out about the percentage of 43%. Of course, you could have asked fewer people, say, 100, of which, say, 43 said that they would vote for the party, but then your statement would be less reliable, because it is less "supported". The number of votes for the party could also be 40% or 50%, because of some random influences. Such deviations are much less likely, if you asked 2000 persons, since then the random influences can be expected to cancel out.

E. The rule support can be used to filter association rules by stating a lower bound for the support of a rule (minimum support). This is equivalent to saying that “you are interested only in such rules that have a large enough statistical basis” (since Borgelt's Apriori program uses the term "support" in Borgelt's interpretation and not in the one used by [Agrawal and Srikant 1994]. The default value for this support limit is 10%. It can be changed with the option -s. Note that the argument, if positive, is interpreted as a percentage. If, however, the given argument is negative, it is interpreted as an absolute number (number of transactions) rather than a percentage.

The minimum support is combined with the minimum confidence to filter association rules. That is, Borgelt's Apriori program generates only association rules, the confidence of which is greater than or equal to the minimum confidence and the support of which is greater than or equal to the minimum support.

Despite the above arguments in favor of Borgelt's definition of the support of an association rule, a rule support compatibility mode is available (due to the overwhelming pervasiveness of the original definition). With the option -o the original rule support definition can be selected. In this case the support of an association rule is the support of the set with all items in the antecedent (also called "if-part" or "body") and the consequent (also called "then-part" or "head") of the association rule, that is, the support of an association rule as defined in [Agrawal and Srikant 1994].

III. TARGET TYPE

An annoying problem in frequent item set mining is that the number of frequent item sets is often huge and thus the output can easily exceed the size of the transaction database to mine. In order to mitigate this problem, several restrictions of the set of frequent item sets have been suggested. These restrictions are covered by the target type.

The target type, which can be selected via the option -t, is either frequent item sets (default, option -ts), closed item sets (option -tc), maximal item sets (option -tm), generators (also called free item sets, option -tg) or association rules (option -tr).

More detailed information about the different target types of frequent item set mining can be found in the survey [Borgelt 2012].

Note that association hyperedges, which were a separate target type in earlier versions of Borgelt's Apriori program, are now covered by the target type of frequent item sets. In order to find association hyperedges, choose rule confidence as the additional evaluation measure (option -ec) and averaging as the aggregation mode (option -aa). Bastien Duclaux asked for originally requesting the possibility to generate association hyperedges.

A. Frequent Item Sets (default, option -ts)

Often one only wants to find frequent item sets. That is, one wants to find all item sets with a support exceeding a certain threshold, the so-called minimum support $supp_{min}$. For Borgelt's Apriori program this is the default operation mode (since version 4.36, earlier versions had association rules as the default mode). However, this mode can also be selected explicitly with the option -ts.

B. Closed Item Sets (option -tc)

A frequent item set is called closed if no superset has the same support (or, in other words, if all supersets have a lower support). Formally, an item set I is called closed iff □ J □ I: supp(J) < supp(I). If the option -tc is given, the found frequent item sets are subsequently filtered and only the closed item sets are reported.

C. Maximal Item Sets (option -tm)

A frequent item set is called maximal if no superset is frequent, that is, has a support reaching or exceeding the minimum support. Formally, an item set I is called maximal iff □ J □ I: supp(J) < $supp_{min}$.If the option -tm is given, the found frequent item sets are subsequently filtered and only the maximal item sets are reported.

D. Generators/Free Item Sets (option -tg)

A frequent item set is called a generator or free if no subset has the same support (or, in other words, if all subsets have a larger support). Formally, an item set I is called a generator iff □ J □ I: supp(J) > supp(I). If the option -tg is given, the found frequent item sets are subsequently filtered and only the generators / free item sets are reported.

E. Association Rules (option -tr)

Borgelt's Apriori program generates association rules if the the option -tr is given. Note, however, that it produces only association rules with a single item in the consequent (also

niket.bhargava.datascience@gmail.com

called "then-part" or "head" of the rule). This restriction is due to the following considerations:

In the first place, association rule mining usually produces too many rules even if one confines oneself to rules with only one item in the consequent. So why should one make the situation worse by allowing more than one item in the consequent? (It merely blows up the size of the output.)

Secondly, much applications are not observed in which rules with more than one item in the consequent are of any real use. The reason is that such more complex rules add almost nothing to the insights about the data set. To understand this, consider the simpler rules that correspond to a rule with multiple items in the consequent, that is, rules having the same antecedent, but consequents with only single items from the consequent of the complex rule. All of these rules must necessarily be in the output, because neither their support (in Borgelt's interpretation) nor their confidence can be less than that of the more complex rule. That is, if you have a rule AB→CD, you will necessarily also have the rules AB→C and AB→D in the output. Of course, these latter two rules together do not say the same as the more complex rule; they do contain additional information. However, what do you gain from the additional information the more complex rule gives you? How can you use it? And is this little extra information worth having to cope with a much bigger rule set? the Borgelt's answer is a clear no.

IV. EXTENDED RULE SELECTION

If association rules are selected using a minimum confidence, the following problem arises: "Good" rules (rules that are often true) are not always "interesting" rules (rules that reveal something about the interdependence of the items). You certainly know the examples that are usually given to illustrate this fact. For instance, it is easy to discover in a medical database that the rule "pregnant→female" is true with a confidence of 100%. Hence it is a perfect rule, it never fails, but, of course, this is not very surprising. Although the measures explained below cannot deal with this problem (which is semantical), they may be able to improve the results in a related case.

Let us look at the supermarket example again and let us assume that 60% of all customers buy some kind of bread. Consider the rule "cheese→bread", which holds with a confidence of, say, 62%. Is this an important rule? Obviously not, since the fact that the customer buys cheese does not have a significant influence on him/her buying bread: The percentages are almost the same. But if you had chosen a confidence limit of 60%, you would get both rules "ø→bread" (confidence 60%) and "cheese→bread" (confidence 62%), although the first would suffice (the first, because it is the simpler of the two). The idea of all measures, which can be used in addition or instead of rule confidence, is to handle such situations and to suppress the second rule.

In addition, consider the following case: Assume that the confidence of the rule "cheese → bread" is not 62% but 35%. With a confidence limit of 60% it would not be selected, but it may be very important to know about this rule! Together with cheese, bread is bought much less frequently than it is bought at all. Is cheese some kind of substitute for bread, so that one does not need any bread if one has cheese? OK, maybe this is not a very good example. However, what can be seen is that a “rule with low confidence can be very interesting”, since it may capture an important influence. Furthermore, this is a way to express negation (although only in the consequent of a rule), since "cheese →bread" with confidence 35% is obviously equivalent to "cheese→no bread" with confidence 65%. This also makes clear why the support of the item set that contains all items in the antecedent ("if-part", "body") and the consequent ("then-part", "head") of the rule is not appropriate for this measure. An important rule may have confidence 0 and thus a support (in the interpretation of [Agrawal and Srikant 1994]) of 0. Hence it is not reasonable to set a lower bound for this kind of support.

Potentially interesting rules differ significantly in their confidence from the confidence of rules with the same consequent, but a simpler antecedent. “Adding an item to the antecedent is informative only if it significantly changes the confidence of the rule. Otherwise the simpler rule suffices.”

Unfortunately the measures other than rule confidence do not solve the rule selection problem in the very general form in which it was stated above. It is not that easy to deal with all rules that have a simpler antecedent, to keep track of which of these rules were selected (this obviously influences the selection of more complicated rules), to deal with the special type of Poincare paradox that can occur etc. Hence the measures always compare the confidence of a rule with the confidence of the rule with an empty antecedent, that is, with the relative frequency of the consequent.

Borgelt called the confidence of an association rule with an empty antecedent the prior confidence (of any rule with the same consequent), since it is the confidence that the item in the consequent of the rule will be present in a transaction prior to any information about other items that may be present. (Note that the prior confidence of a rule is obviously equal to the (relative) support of its consequent item.) The confidence of an association rule with non-empty antecedent (and the same consequent) Borgelt call this posterior confidence (or simple the confidence) of the rule, since it is the confidence that the item in the consequent of the rule will be present after it becomes known that the items in the antecedent of the rule are present.

Most measures, which can be computed with Borgelt's Apriori program and can be used for filtering in addition to the rule confidence, are basically computed from these two values: the prior confidence and the posterior confidence. Only those association rules are reported for which the value of the chosen additional evaluation measure is in a certain relation to certain limit (that is, either (1) meets or exceeds the limit or (2) does not exceed the limit – which relation applies depends on the chosen measure). The measures are chosen with the option -e, the limit is passed to the program via the option -d. The default value for the limit is 10%. Note that the option -d always interprets its argument as a percentage. As a consequence, the desired limit value may sometimes have to be multiplied by 100 in order to obtained the needed argument value.

Note that all additional rule evaluation measures are combined with the limits for rule confidence and rule support. That is, Borgelt's Apriori program reports only those rules, the confidence of which is greater than or equal to the minimum confidence, the support of which is greater than or equal to the minimum support,and for which the additional evaluation value (if selected) is in the measure-specific relation to the limit for this measure. The default is to use no additional evaluation measure (but this may also be specified explicitly with the option -ex), that is, to rely only on rule confidence and rule support. Of course you can remove the restriction that the rule support and the rule confidence must meet or exceed certain limits by simply setting either (or both) of these limits to zero. In this case rules are selected using only the limit for the additional evaluation measure. (Attention: If you have a large number of items, setting the minimum rule support or the minimum rule confidence to zero can result in very high memory consumption. Therefore: use this possibility with a lot of care, if at all.)

A. Absolute Confidence Difference to Prior (option -ed)

The simplest way to compare the posterior and the prior confidence of an association rule (since the former should differ considerably from the latter to make the first one “posterior” a rule which is interesting one) is to compute the absolute value of their difference. That is, if "□→bread" has a confidence of 60% and "cheese→ bread" has a confidence of 62%, then the value of this measure is 2% (for the second rule). The parameter given with the option -d to the program states a lower bound for this difference (in percentage points). It follows that this measure selects rules, the (posterior) confidence of which differs more than a given threshold from the corresponding prior confidence.

For example, with the option -d20 (and, of course, the option -ed to select this measure) only rules with a “confidence less than 40% (60(prior confidence) – 20) or greater than 80% (60(prior confidence) + 20) would be selected” for the item "bread" in the consequent. As a consequence, the selected rules are those, for which the antecedent considerably changes the confidence. (Note that, of course, for other items, with a different prior confidence, the upper and lower bounds are different. For example, they are 10% (30(priot confidence) – 20 (-d20 parameter)) and 50% (30(prior confidence) + 20(-d20 parameter)) for a rule with a prior confidence of 30%.).

B. Lift Value (Confidence Quotient, option -el)

The lift value is the quotient of the posterior and the prior confidence of an association rule. That is, if "□→bread" has a confidence of 60% and "cheese→bread" has a confidence of 72%, then the lift value (of the second rule) is 72/60 = 1.2. Obviously, if the posterior confidence equals the prior confidence, the value of this measure is 1. If the posterior confidence is greater than the prior confidence, the lift value exceeds 1 (the presence of the antecedent items raises the confidence), and if the posterior confidence is less than the prior confidence, the lift value is less than 1 (the presence of the antecedent items lowers the confidence).

More formally, the lift of the rule R = X→Y is

$$\text{lift}(R) = \frac{\text{conf}(X \to Y)}{\text{conf}(\text{Ø} \to Y)} = \frac{\text{supp}(X\square Y)/\text{supp}(X)}{\text{supp}(Y)/\text{supp}(\text{ø})}$$

where supp(ø)=|T|, the size of the transaction database (number of transactions).

The value that can be passed to the program with the option -d is a lower limit for this measure: only rules for which this measure meets or exceeds the given value are reported. As a consequence, the selected rules are those that raise the confidence by at least a given minimum factor. Note, however, that the option -d always interprets its arguments as a percentage. Therefore, in order to filter rules with a lift value of at least 1.2, the option -d120 must be provided (that is, the argument must be the limit for the lift value times 100%).

C. Absolute Difference of Lift Value to 1 (option -ea)

With the lift value (see preceding section) it is not possible to filter association rules for which the lift value differs more than a given amount from 1, because the limit passed to the program with the option -d is a lower limit. Hence only rules with a lift value that exceeds 1 by a given minimum amount (and thus, for which the presence of the antecedent items raises the confidence by at least a given factor) can be properly filtered.

In order to be able to filter rules for which the lift value differs by more than a given amount from 1, the absolute difference of the lift value to 1 can be used (option -ea). For example, if the prior confidence of an association rule (the support of its consequent) is 60% and the option -d20 is given, rules with a confidence less than or equal to (1–20%)*60% = 0.8 *60% = 48% or a confidence greater than or equal to (1+20%)*60%=1.2*60%=72% are selected. Similarly, if the prior confidence is 30%, the numbers are 0.8*30%=24% and 1.2*30%=36%.

As these examples show, the main difference between this measure and the absolute difference of the posterior and prior confidences is that the “deviation” that is considered to be significant depends on the prior confidence (12% deviation for 60% prior confidence, but only 6% deviation for 30% prior confidence). The idea is that for a high prior confidence the deviation of the posterior confidence must also be high, and if it is low, the deviation only needs to be low. (Roland Jonscher, S-Rating GmbH, Berlin, Germany, who pointed out these measure.)

D. Difference of Lift Quotient to 1 (option -eq)

Instead of simply forming the absolute difference of the lift value of an association rule, one may also compute the difference of "either the lift value or its reciprocal, whichever is smaller, to one". Since either the lift value or its reciprocal must be less than or at most equal to one (and necessarily non-negative), this yields a value between 0 and 1. This measure can be selected with the option -eq, a lower limit can, as usual, be specified with the option -d.

For example, if the prior confidence of an association rule (the support of its consequent) is 60% and the option -d20 is given, association rules with a confidence less than or equal to (1 –20%)*60%=0.8*60%=48% or a confidence greater than or equal to 60%/(1–20%) = 60%/0.8 =75% are selected. On the other hand, if the prior confidence is only 30%, rules with a (posterior) confidence of no more than 0.8*30%=24% or at least 30%/0.8=37.5% are selected (with -d20).

Note that the main difference to the absolute difference of the lift value to 1 (see the preceding section) is that positive and negative deviations are treated differently. With the absolute difference of the lift value to 1, positive and negative deviations (posterior confidence exceeding and falling short of the prior confidence, respectively) are treated the same. The deviation must be by the same amount (in percentage points) in order to select the rule. "With this measure (difference of lift quotient to 1), however, a negative deviation by a certain number of percentage points can lead to the selection of the rule, while the same positive deviation (in percentage points) need not lead to a selection of the rule." For the example above, with a prior confidence of 60%, the positive deviation limit is 15%, the negative deviation limit only 12%. Likewise, for a prior confidence of 30%, the positive deviation limit is 7.5%, while the negative deviation limit is only 6%.

E. Conviction (Inverse Lift for Negated Head, option -ev)

The conviction of an association rule R = X→Y is the inverse lift of the rule R' = X→not Y. That is, while the lift of the rule R is

$$\text{lift}(R) = \frac{\text{conf}(X\rightarrow Y)}{\text{conf}(\varnothing\rightarrow Y)}$$

and hence, the conviction of the rule R is

$$\text{cvct}(R) = \frac{1 - \text{conf}(\varnothing \rightarrow Y)}{1 - \text{conf}(X\rightarrow Y)} = \frac{\text{conf}(\varnothing\rightarrow \text{not } Y)}{\text{conf}(X\rightarrow \text{not } Y)}$$

where supp(ø) = |T|, the size of the transaction database.

Intuitively, the conviction states by what factor the correctness of the rule (as expressed by its confidence) would reduce if the antecedent and the consequent of the rule were independent. A high value therefore means that the consequent depends strongly on the antecedent. Consequently, the value passed with the option -d is interpreted as a lower limit. Note, however, that the option -d always interprets its arguments as a percentage (cf. the discussion of the lift value). Therefore, in order to filter rules with a conviction value of at least 1.2, the option -d120 must be provided (that is, the argument must be the limit for the conviction value times 100%).

F. Absolute Difference of Conviction to 1 (option -ee)

Analogous to Absolute Difference of Lift Value to 1, only with conviction instead of lift.

G. Difference of Conviction Quotient to 1 (option -er)

Analogous to Difference of Lift Quotient to 1, only with conviction instead of lift.

H. Certainty Factor (option -ez)

The certainty factor of a rule states by how much the prior confidence changes to the posterior confidence relative to the maximumally possible change in the same direction. That is, the certainty factor distinguishes whether the posterior confidence is larger or smaller than the prior confidence. If it is larger, it relates the change to a hypothetical change to the maximally possible posterior confidence 100%. That is, in this case the certainty factor is

$$\text{cf}(R) = \frac{\text{conf}(X\rightarrow Y) - \text{conf}(\varnothing\rightarrow Y)}{100\% - \text{conf}(\varnothing\rightarrow Y)}$$

On the other hand, that is, if the posterior confidence is smaller than the prior confidence, it relates the change to a hypothetical change to the minimally possible posterior confidence 0%. That is, in this case the certainty factor is

$$\text{cf}(R) = \frac{\text{conf}(\varnothing\rightarrow Y) - \text{conf}(X\rightarrow Y)}{\text{conf}(\varnothing\rightarrow Y) - 0\%}$$

In this way the certainty factor assesses a small increase of an already high prior confidence as more significant than the same increase of a smaller prior confidence. For example, an increase from 90% to 92% gives rise to a certainty factor of 0.2 = 20% (2% increase divided by 10% maximally possible change from 90% to 100%). The same increase from 60% to 62%, however, only gives rise to certainty factor of 0.05 = 5% (2% increase divided by 40% maximally possible change from 60% to 100%).

I. Normalized χ^2-Measure (option -en)

The χ^2-measure is well known from statistics. It is often used to measure the difference between a conjectured independent distribution of two discrete variables and the actual joint distribution in order to determine how strongly two variables depend on each other (or whether they are independent). The two (binary) variables considered here are the consequent and the antecedent of the rule, namely whether (all) the items contained in them are present or not (variable X: all antecedent items are present (X=1) versus at least one absent (X=0), and variable Y: consequent item present (Y=1) versus consequent item absent (Y=0)).Formally, the χ^2-measure can be defined as

$$\chi^2(R) = n \frac{(n01n10n11n)^2}{n01(n - n01)\ n10(n - n10)}$$

where the first index of the n's refers to X and the second to Y and n=|T| is the total number of transactions in the database both x, y not present than n_{00}. Clearly, the χ^2-measure contains the number n of cases it is computed from as a factor. Even though this is, of course, highly relevant in statistics, this is not very convenient if one wants to filter association rules that can have different support. Hence in Borgelt's Apriori program this factor is removed by simply dividing the measure by the total number n of transactions. With this normalization, the χ^2-measure can have values between 0 (no dependence) and 1 (very strong – or actually perfect – dependence). The value that can be passed with the -d option is a lower bound for the strength of the dependence of the consequent on the antecedent in percent (0 – no dependence, 100 – perfect dependence). Only those association rules are selected, in which the consequent depends on the antecedent with this or a higher degree of dependence.

J. p-Value Computed from χ^2 Measure (option -ep)

Provided that the two considered variables are independent, the χ^2-measure has a χ^2-distribution, which, for two binary variables, has one degree of freedom. Hence it is possible to compute the probability that the χ^2-value of an association rule under consideration can be observed by chance under the assumption that the antecedent and the consequent of the rule are independent. Since with this measure "small p-values indicate interesting rules," the value given as a parameter to the option -d is an upper limit. Note also that the option -d interprets its parameter as a percentage, so that -d1 means that only association rules with a p-value smaller than 1% = 0.01 are reported.

K. Normalized χ^2-Measure with Yates' Correction (option -ey)

The χ^2-distribution, that is used to compute a p-value from the χ^2-measure is a continuous distribution, but the χ^2-measure is computed from the discrete entries of a 2x2 contingency table. As a consequence, the assumption that the distribution of the (discrete) χ^2-measure can be approximated by the (continuous) χ^2-distribution is not quite correct and can lead to some approximation error. To reduce this error, Yates suggested a correction of the χ^2-measure, which is also known as Yates' correction for continuity. Its intention is to prevent an overestimation of significance (that is, p-values smaller than justified) if the counters in the contingency table have small values (small underlying data set). Formally, the Yates-corrected χ^2-measure can be defined as

$$\chi^2(R) = n \frac{(n01n10\ n11n - 0.5n)^2}{n01\,(n - n01)\,n10\ \ (n - n10)}$$

where the first index of the n's refers to X and the second to Y and n = |T| is the total number of transactions in the database. Like the standard χ^2-measure, The Yates-corrected χ^2-measure contains the number n of cases it is computed from as a factor. Even though this is, of course, highly relevant in statistics, this is not very convenient if one wants to filter association rules that can have different support. Hence in Borgelt's Apriori program this factor is removed by simply dividing the measure by the total number n of transactions. With this normalization, the Yates-corrected χ^2-measure can have values between 0 (no dependence) and 1 (very strong – or actually perfect – dependence). The value that can be passed with the -d option is a lower bound for the strength of the dependence of the consequent on the antecedent in percent (0 – no dependence, 100 – perfect dependence). Only those association rules are selected, in which the consequent depends on the antecedent with this or a higher degree of dependence.

L. p-Value Computed from Yates-corrected χ^2-Measure (option -et)

Analogous to the p-value computed from the standard χ^2-measure, but with the Yates-corrected χ^2-measure.

M. Information Difference to Prior (option -ei)

The information difference to the prior is simply the information gain criterion (also called mutual information that is also used, for example, in decision tree learners like C4.5 to select the split attributes. Its basic idea is as follows: Without any further information about other items in the set, we have a certain probability (or, to be exact, a relative frequency) distribution for, say "bread" and "no bread". Let us assume it is 60% : 40% (prior confidence of the item "bread", just as above). This distribution has a certain entropy

H = - P(bread) $\log_2$ P(bread) - P(no bread) $\log_2$ P(no bread),

where P(bread) is equivalent to the (relative) support of "bread", which in turn is equivalent to the prior confidence of "bread". The entropy of a probability distribution is, intuitively, a lower bound on the number of yes-no-questions you have to ask in order to determine the actual value. This cannot be understood very well with only two possible values, but it can be made to work for this case too. After we get the information that the items in the antecedent of the association rule are present (say, cheese), we have a different probability distribution, say 35% : 65%. I.e., P(bread | cheese) = 0.35 and P(no bread|cheese) = 0.65. If we also know the support of the item "cheese" (let it be P(cheese) = 0.4 and P(no cheese) = 0.6), then we can also compute the probabilities P(bread|no cheese) = 0.77 and P(no bread|no cheese) = 0.23. Hence we have two posterior probability distributions: one in case cheese is also present, and one in case cheese is not present. The question now is: How much information do we receive by observing whether the antecedent of the rule holds or not? Information is measured as a reduction of entropy. Hence the entropies of the two conditional probability distributions (one for "cheese" and one for "no cheese") are computed and summed weighted with the probability of their occurrence (that is, the relative frequency of "cheese" and "no cheese", respectively). This gives the (expected value of the) posterior or conditional entropy. The difference of this value to the prior entropy (see above) is the gain in information from the antecedent of the rule or, as I called it here, the information difference to the prior.

The value that can be given via the -d option is a lower bound for the information gain, measured in hundreds of a bit (percent of one bit). (Note that, since all items can only be present or absent, the information gain can be at most one bit.)

N. p-Value Computed from G-Statistic (option -eg)

The so-called G-statistic is a less well-known alternative to the χ^2-measure that is based on information gain (or mutual information). Formally, it is proportional to the product of the information gain and the number of cases the information gain is computed from (here: the (absolute) support of the antecedent of the association rule). Under independence, the G-statistic also has a χ^2-distribution and thus it can be used in a similar fashion as the χ^2-measure to compute a p-value. This measure (p-value computed from G-statistic) can be selected with the option -eg. Since with this measure small p-values indicate interesting rules, the value given as a parameter to the option -d is an upper limit. Note also that the option -d interprets its parameter as a percentage, so that -d1 means that only association rules with a p-value smaller than 1% = 0.01 are reported.

O. Fisher's Exact Test; Table Probability (option -ef)

Like the p-value computed from the χ^2-measure, Fisher's exact test co3mputes a p-value from a contingency table. It is often used to measure the difference between a conjectured independent distribution of two discrete variables and the actual joint distribution in order to determine how strongly two variables depend on each other (or whether they are independent). The two (binary) variables considered here are the consequent and the antecedent of the rule, namely whether (all) the items contained in them are present or not (variable X: all antecedent items are present (X=1) versus at least one absent (X=0), and variable Y: consequent item present (Y=1) versus consequent item absent (Y=0)).

In Fisher's exact test, the marginals of the contingency table are fixed, that is, the support of the antecedent of the rule (that is, supp(X)) and the support of the consequent of the rule (that is, supp(Y)) as well as the total number of transactions, are seen as fixed. Under this assumption, the distribution of the numbers in the cells of the contingency table that respect the marginals follows a hyper geometric distribution. Fisher's exact test is based on the probability of finding, under the assumption of independence, a contingency table that is at least as extreme as the one that was actually observed. Of course, the notion "at least as extreme" needs clarification. Unfortunately, there are several ways of making this notion mathematically precise, which give rise to different forms of Fisher's exact test. The most common is to order the possible contingency tables based on the probability that is assigned to them by the hyper geometric distribution. In this case the p-value is the sum of the probabilities (as computed from the hyper geometric distribution) of all possible contingency tables that are no more probable than the one actually observed. Since with this measure small p-values indicate interesting rules, the value given as a parameter to the option -d is an upper limit. Note also that the option -d interprets its parameter as a percentage, so that -d1 means that only association rules with a p-value smaller than 1% = 0.01 are reported.

P. Fisher's Exact Test; χ^2-Measure (option -eh)

Analogous to Fisher's exact text (table probability), but contingency tables ordered by the χ^2-measure. In this case the p-value is the sum of the probabilities of all possible contingency tables that have a value of the χ^2-measure that is no smaller than the value of the χ^2-measure of the table that was actually observed.

Q. Fisher's Exact Test; Information Gain (option -em)

Analogous to Fisher's exact text (table probability), but contingency tables ordered by the Information Gain. In this case the p-value is the sum of the probabilities of all possible contingency tables that have an information gain that is no larger than the information gain of the table that was actually observed.

R. Fisher's Exact Test; Support (option -es)

Analogous to Fisher's exact text (table probability), but contingency tables ordered by the rule support in its original definition, that is, by the support of all items in the antecedent (body) and the consequent (head) of the rule. In this case the p-value is the sum of the probabilities of all possible contingency tables that have a support for all the items in the antecedent and consequent that is no larger than the support in the table that was actually observed.

S. Original Rule Support (body & head; option -eo)

The original rule support can be chosen as an additional evaluation measure to make it possible to use both the support of the antecedent (Borgelt's interpretation of rule support) and of all items occurring in the rule (original interpretation of rule support) to select rules. Note that this measure deviates from the general scheme of comparing the prior and the posterior confidence and thus serves different purposes.

T. Rule Confidence (option -ec)

The standard rule confidence can also be chosen as an additional rule evaluation measure. Obviously, this is not really useful for filtering association rules (because the rule confidence is always used by default already). Indeed, this option is not relevant for association rule selection, but can be useful for evaluation item sets.

V. EXTENDED RULE SELECTION

Since versions 4.20 (binary logarithm of support quotient) and 4.36 (other measures) there are also extended selection possibilities for frequent item sets. (These were added due to a cooperation with Sonja Gruen, RIKEN Brain Science Institute, Tokyo, Japan.)

A. Binary Logarithm of Support Quotient

A simple evaluation measure for item sets is the “quotient” of the “actual support” (as computed from the given transactions) and the “expected support” of an item set. The expected support of an item set can be computed from the support values of the individual items by assuming that the occurrences of "all items are independent". Under independence we expect an item set to occur with a relative frequency that is the product of the relative occurrence frequencies of the individual items contained in it. Since this product quickly becomes very small (it decreases basically exponentially with the size of the item set), and thus the quotient of the actual and the expected frequency can become very large, it is advisable not to use the described support quotient directly, but its binary logarithm, so that its values stay in a manageable range. Computing the logarithm also has the advantage that the measure has the value zero for an item set that occurs exactly as often as expected under an assumption of item independence. The binary logarithm of the quotient of the actual and the expected support can be selected with the option -eb.

As for the additional rule "evaluation measures", a "minimum value" for this measure can be set with the option -d. In this case only frequent item sets for which this measure exceeds the given threshold are reported. Note, however, that the option -d generally interprets its argument as a percentage. That is, if you only want item sets reported that occur at least twice as often as expected under independence (and thus desire the binary logarithm of the quotient of the actual and the expected support to be at least 1), you have to specify -d100.

Additional Rule Evaluation Measures

Apart from the measure described in the preceding section (which is specific to item sets), all evaluation measures for association rules, as they were described in this section (including rule confidence, option -ec), can be used to filter item sets. The idea is to form all possible rules with "one item in the consequent" and all other items of the given item set in the antecedent. Each of these rules is then evaluated with the chosen measure and the results are aggregated. The aggregated value is the evaluation of the item set, and item sets can now be filtered by requiring a minimum value for this aggregation with the option -d. There are four different aggregation modes for the evaluations of the rules that can be formed from an item set, which can be selected via the option -a:

-ax	no aggregation (use first value)
-am	minimum of individual measure values
-an	maximum of individual measure values
-aa	average of individual measure values
-as	split into equal size subsets

Here "no aggregation (use first value)" means that only one association rule is formed. This rule has that item in the consequent that comes last in the order of the items. The evaluation of this rule is the evaluation of the item set. For the next three other options all possible rules are formed. For the last option a single rule with half of the items in the antecedent and the other half in the consequent is formed. If the total number of items is not even, the antecedent contains one item more.

The order of the items is determined with the option -q. By default the items are sorted ascendantly w.r.t. the sum of the sizes of the transactions they are contained in. (Note that this choice is based on efficiency issues: it usually leads to the fastest search. However, for use with this aggregation mode, you may want to choose a different order.) Other sorting options are ascendantly w.r.t. the item frequency or descending w.r.t. item frequency or the sum of the sizes of the transactions the items are contained in. With the option -q0 the order in which the items are first encountered when the transactions are read is used. A specific, user-defined order may also be chosen, namely by using the option -q0 and using the optional item selection/appearances file, since the item appearances file is read before the transactions file. Therefore, in this case, the order of items in the selection/appearances file determines the order of the items.

B. Pruning with Additional Measures

By default only the minimum support is used to prune the search for frequent item sets. That is, if it is discovered that an item set does not meet the user-specified minimum support, no supersets of this item set are considered. (Note that this is a safe pruning rule, because no superset of an infrequent item set can be frequent.)

With the option " -pthe" additional item set evaluation measure can also be used for pruning the search additionally. Pruning with an additional item set evaluation measure comes in two flavors: forward and backward.

Backward pruning (which is switched on with the option -p-1), means that all frequent item sets (item sets that reach or exceed the user-specified minimum support) are evaluated with the additional measure, but those item sets, the evaluation of which is less than the user-specified limit (option -d), are discarded. Note that providing or not providing the option -p-1 makes no difference for finding frequent item sets, since for this target type simply selecting an additional evaluation measure and a limit for its value already have the effect of discarding item sets for which the additional evaluation is too low. However, backward pruning can make a huge difference for finding closed or maximal item sets. The reason is that the option -p changes the order in which item sets are filtered by the additional evaluation measure and are filtered for closed (or maximal) item sets: by default (or if explicitly requested with the option -p0), the item sets are first filtered for closed (or maximal) item sets. Then only the closed (or maximal) item sets are evaluated with the additional measure and thus further reduced to those closed (or maximal) item sets that reach or exceed the user-specified evaluation limit. With the option -p-1 (or any other negative number as the argument - all negative numbers have the same effect) all frequent item sets are first evaluated with the additional measure and those that do not reach or exceed the user-specified limit are discarded. Only afterwards the closed (or maximal) item sets of this reduced set of item sets are determined.

Forward pruning, on the other hand, means that in addition to a check whether the additional evaluation of an item set reaches or exceeds the user-specified limit, it is tested whether all of its subsets (with at least a certain size) reach or exceed the limit. Only if this is the case, the item set is reported. (Technically, this is achieved by not considering any superset of item sets that failed to reach or exceed the user-specified limit for the additional measure - hence the name "forward pruning".)

Since some additional measures (like, for example, the χ^2-measure) cannot reasonably be used to evaluate the empty set or single element sets, but also because this restriction can be useful for certain applications, the option -p allows for a parameter that states the minimum size of the subsets that are to be considered. For example, -p2 means that only subsets with at least two elements are required to reach or exceed the user-specified limit; single element sets and the empty set need not do so. Similarly, -p4 means that all subsets up to a size of 3 are ignored in the check; only subsets with at least 4 elements must reach or exceed the limit. Note that the option -p1 usually produces no output, because all single element sets and the empty set are formally assigned an additional evaluation of 0 for most measures. Hence 2 is usually the smallest useful parameter for this option.

Note also that, for example, -p4 does not mean that all reported item sets must have at least four elements. It rather means that there is no additional requirement for item sets up to four elements. If you want to require a minimum size for reported item sets, use the option -m.

Difference of Support Quotient to 1

As with the preceding measure the quotient of actual support and expected support of an item set is computed and compared to 1 (a value of 1 signifies independence of the items). A minimum value for this measure can be set with the option -d. In this case only frequent item sets for which this measure exceeds the given threshold are kept.

VI. TRANSACTION PREFIX TREE

The main problem of association rule induction is that there are so many possible rules. For example, for the product range of a supermarket, which may consist of several thousand different products, there are billions of possible association rules. It is obvious that such a vast amount of rules cannot be processed by inspecting each one in turn. Therefore efficient algorithms are needed that restrict the search space and check only a subset of all rules, but, if possible, without missing important rules. One such algorithm is the Apriori algorithm, which was developed by [Agrawal and Srikant 1994]and which is implemented in a specific way in c language by Borgelts, development started from 1996 and continued till date. For an overview of frequent item set mining in general and several specific algorithms (including Apriori), see the survey [Borgelt 2012]. This page describes the Apriori implementation. It uses a prefix tree to organize the support counters and a doubly recursive procedure to process the transaction to count the support of candidate item sets. Some implementation details can be found in [Borgelt and Kruse 2002], [Borgelt 2003], and [Borgelt 2004].The counting process can be speed up by organizing the transactions into a prefix tree. That is, the items in each transaction are sorted and then transactions with the same prefix are grouped together and are counted, as one may say, in parallel. This way of organizing the transactions was added in version 4.03 c program developed by Borgelt and is the default behavior now. If you prefer that the transactions are treated individually (i.e., the transactions are stored in a simple list and only one transaction is counted at a time), use the option -h.

Earlier versions of Borgelt's Apriori program are incorporated in the well-known data mining tool Clementine (Apriori version 1.8 in Clementine version 5.0, Apriori version 2.7 in Clementine version 7.0), available from SPSS. Newer versions of Clementine still use Borgelt's program. His program is also accessible through .Call function the arules package of the statistical software package R. Furthermore it can be used through the Python interface provided by the PyFIM library. In R arules package provide a function apriori, which further use .Call and call rapriori. A graphical user interface for this program (ARuleGUI), written in Java is also available.

VII. FUTURE WORK

Concepts discussed in this paper have strong application in Business and also in many other domains. Good quality reporting of the current status of business, past summary, and almost exact prediction of future events are critical to business decision making. Now a days, many organization are struggling with huge amount of data. Processing this data and finding useful, interesting, good, applicable to real world, rules from the the data is real challenge. In future we will try to develop new measure of interestingness and fast and efficient methods for frequent item mining and rule generation based on calendar schema. Data science can help in BI reporting using proper use of statistics, artificial intelligence, simulation and modeling, operations research, and machine learning tools and techniques for small data set and for big data sets of all sorts like transaction database, relational database, sequence database, hierarchical database and text corpus, videos, audios and image databases etc.

References

[1] http://documents.software.dell.com/statistics/textbook/data-mining-techniques, dated 16, Jan, 2016

[2] http://searchdatamanagement.techtarget.com/definition/business-intelligence, dated 16, Jan, 2016

[3] http://www.personal.kent.edu/~rmuhamma/Algorithms/MyAlgorithms/DeductInduct.htm,dated 16 Jan 2016

[4] Christian Borgelt, “Frequent Item Set Mining,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6):437-456. J. 2012, doi:10.1002/widm.1074 wiley.com

[5] Christian Borgelt, “Recursion Pruning for the Apriori Algorithm,” 2nd Workshop of Frequent Item Set Mining Implementations (FIMI 2004, Brighton, UK). fimi_04.pdf

[6] Christian Borgelt, “Efficient Implementations of Apriori and Eclat,” Workshop of Frequent Item Set Mining

Implementations (FIMI 2003, Melbourne, FL, USA). fimi_03.pdf

[7] Christian Borgelt and Rudolf Kruse, “Induction of Association Rules: Apriori Implementation,” 15th Conference on Computational Statistics (Compstat 2002, Berlin, Germany), 395-400 cstat_02.pdf

[8] D Picado-Muiño, C Borgelt, D Berger, George Gerstein, and Sonja Grün, “Finding Neural Assemblies with Frequent Item Set Mining,” Frontiers in Neuroinformatics 7:article 9, accfim.pdf

[9] Emiliano Torre, David Picado-Muiño, Michael Denker, Christian Borgelt, and Sonja Grün, “Statistical Evaluation of Synchronous Spike Patterns extracted by Frequent Item Set Mining” Frontiers in Computational Neuroscience, 7:article 132 Frontiers Media, 2013

[10] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, Proc. 20th Int. Conf. on Very Large Databases (VLDB 1994, Santiago de Chile), 487-499 Morgan Kaufmann, CA, USA 1994

[11] *R.* Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, “Fast Discovery of Association Rules”, In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, 307-328,AAAI Press / MIT Press, Cambridge, CA, USA 1996

AUTHORS PROFILE

Niket Bhargava, born in state is MP in the India, on July 6, 1978. He graduated from the Oriental Institute of Science And Technology, Bhopal, and after qualifying in GATE, completed his Master in Technology from the RGTU- the technical university of MP, India. His employment experience included the 10 years as Teacher in Top Most Ranking Institutions of state of MP,and Industry experience of BI product development. Presently, he is pursuing his PhD in CSE with focus on business applications of data mining , data science, big data techniques for Business Intelligence.

Dr. MANOJ **SHUKLA**, is working as professor in BCE, Mandiddeep. He is having more than 15 years of experience in teaching and hold positions like principal, and others in various institutes in Madhya Pradesh, India.