ISSN 2278 – 0211 (Online)
ISSN 2278 – 7631 (Print)

Post Mining Of Frequent Item Sets Using Mutual Information

Jayanti Mukhopadhyay
Department Of Bachelor in Computer Application, The Calcutta Anglo Gujarati College Kolkata, India
Professor Rohit Kamal Chatterjee
Department Of Computer Science & Engineering, BIT Mesra, Kolkata, India

Abstract:
Buyers' Basket Analysis (BBA or Market Basket Analysis) is a typical example of frequent item set mining that leads to the discovery of association and correlations among items in large transactional or relational data sets to help retailers to develop marketing strategies by gaining insight into which items are frequently purchased together by customers. Association rules are considered interesting if they satisfy both a minimum support threshold and minimum confidence threshold, set by domain experts. Many efficient algorithms like Apriori, Partitioning, Sampling, Eclat etc. are available to generate large number of associated frequent item sets. Additional analysis can be performed to discover interesting statistical correlations between associated items. But the above mentioned correlation measures works in linear relationship between two variables with random distribution; this value alone may not be sufficient to evaluate a system where these assumptions are not valid. Finally we propose a new measure i.e. mutual information that will show dependency between frequent item sets of linear and parabolic datasets and generate stronger associated frequent item sets.

Key words: *Frequent Item sets, Association rules, mutual information*

1. Introduction

Buyers' Basket Analysis (BBA or Market Basket Analysis) is a well known technique of data mining to analyze customer behavior. It is based on association rule mining. The goal of BBA is to identify relationship (i.e. association rules) between groups of products, items or category taken from a large dataset [6]. It helps in increasing sales and maintain inventory by focusing on the point of sale transaction data. Association rule mining is a two step approach i.e. first it finds all frequent item sets using minimum support threshold and then generate strong association rules that satisfy minimum support and minimum confidence. Frequent item set mining and association rule induction [[2],[3],[4]] are powerful methods for Buyers' Basket Analysis. The main problem of association rule induction is that there are many possible rules[3]. It is obvious that such a vast amount of rules cannot be processed by inspecting each one in turn. Therefore efficient algorithms are needed which 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 R. Agrawal & R. Srikant [[3],[4]] for mining frequent item sets for boolean association rules. But the problem of Apriori algorithm is that it requires many scans to generate a large number of frequent item sets. Many other algorithms (partitioning, sampling, Eclat etc.) are present which focus on improving the efficiency of the original Apriori algorithm. But the use of only support and confidence measures to mine associations may generate huge numbers of frequent item sets which can be uninteresting to users [14],[15]. Overviews of measures (subjective and objective) of interestingness are also used to find frequent item sets [13]. There are some additional analysis of post mining data that can be performed to discover interesting statistical correlation (lift, all_confidence, Cosine etc.) between frequently associated items. But all the above mentioned correlation measures works on linear data sets. In this paper we propose a new measure like mutual information to find frequently associated item sets not only from linear data sets but also from parabolic data sets.

2. Background

Buyers' Basket Analysis (BBA) is based upon association rule mining. Let $I = \{I_1, I_2, \dots ,I_m\}$ be a set of Items. Let D, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B\subset I$, and $A\cap B = \emptyset$ [1]. The two measures of association rule mining is Support and confidence.

$$support(A \Rightarrow B) = P(A \cup B) \quad ...(1)$$

$$confidence(A \Rightarrow B) = P(B|A). \quad ...(2)$$

The concept of binary association rules represent presence of item denoted by 1 and absence of item denoted by 0 [[5],[6]]. But association rule mining often generates a huge number of rules, and a majority of them either are redundant or do not reflect the true relationship among data objects [6]. To overcome this difficulty, correlation has been adopted as interesting measures. This leads to correlation rules of the form [12]

A⇨B [support, confidence, correlation].
There are many different correlation measures for mining large data sets:-

1. **Lift** :-The lift between the occurrence of item sets A and B can be measured by computing

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}. \quad ...(3)$$

Positively correlated : lift > 1.
Negatively correlated: lift < 1.
No Correlation : lift = 1.

2. **Chi-Square(X ^{2}):-**

$$\chi^2 = \Sigma \frac{(observed - expected)^2}{expected} \quad ...(4)$$

1. **All confidence:-**

$$all\_conf(X) = \frac{sup(X)}{max\_item\_sup(X)} \quad ...(5)$$

2. **Cosine:-**

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}}$$

$$= \frac{sup(A \cup B)}{\sqrt{sup(A) \times sup(B)}} \quad ..(6)$$

3. **Kulczynski:-**

$$Kulc(A,B) = \tfrac{1}{2} (P (A | B) + P(B | A)) \quad ...(7)$$

4. **Max-confidence:-**

$$Maximum\_confi (A,B) = max \{ P (A | B), P(B | A) \} \quad ...(8)$$

(In the above formulae, A, B, X are the data item sets, P represent probability)

Each of the above six co-relation measure value range from 0 to 1 and higher the value, the closer the relationship between A and B [1].

3. Problems with Objective & Subjective Measure

Objective measures rely on user's ability to choose the right measure for a given scenario out of a huge set of available measures. Some measures produce similar rankings while others almost reverse the order. This poses the problem of choosing the right measure for a given scenario [[8], [9]]. Moreover, due to their rather mathematical foundation most measures lack interpretability and meaningfulness because the rule properties they measure rarely reflect the practical considerations of a user. For a user it is often unclear which measure to choose and how to link its results to his application scenario. Objective measures do not memorize the past and they are unable to identify patterns which have already been discovered multiple times in the past, which are diminishing or emerging [[10], [11]].

Subjective measures, on the other hand, require user’s domain knowledge. A lot of effort is necessary to collect, organize and finally incorporate domain knowledge into a knowledge base against which association rules will be compared. Moreover, domain experts often forget certain key aspects or may not remember others which come into play under rarer circumstances. This problem can be termed ‘expert dilemma’ 1980s [11]. Building a knowledge base can also become a task that can never be finished. Consequently, there is a risk that patterns are regarded as interesting based on outdated knowledge while a user is being left uninformed about the out datedness itself [13].

Here our propose measure i.e. using mutual information, we can get most frequent item set more accurately than using correlation measures.

The correlation co-efficient indicates the strength of a linear relationship between two variables with random distribution; this value alone may not be sufficient to evaluate a system where these assumptions are not valid. The mutual information measures the general dependency while the correlation function measures the linear dependency. Another major difference between mutual information and correlation function is that the former can be applied to symbolic sequences as well as numerical sequences, but the latter can only be used on numerical sequences. For example, we can have zero value of correlation function at some distance d, while the mutual information function at that distance can be any value.

4. Information Theory

The concept of mutual information is quite complex and is the basis of information theory. Information theory is a branch of computer science involving quantification of information. Information theory was developed by Claude E Shannon [[23],[24],[25]]. Since its inception it has broadened to find applications in many other areas, including statistical inference [[17],[22]].

4.1. Mutual Information

Mutual information (also referred to as transinformation) is a quantitative measurement of how much one random variable (Y) tells us about another random variable (X) [16]. In this case, information is thought of as a reduction in the uncertainty of a variable. Thus, the more mutual information between X and Y, the less uncertainty there is in X knowing Y or Y knowing X.

Mutual information is most commonly measured in logarithms of base 2 (bits) but is also found in base e (nats) and base 10 (bans).

The mathematical representation for mutual information of the random variables X and Y are as follows:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right), \quad (9)$$

where, p(x,y) = joint probability distribution function of X and Y.
p(x) = marginal probability distribution function of X.
p(y) = marginal probability distribution function of Y[[18],[19],[20]].

4.2.Kullback-Leibler Divergence Theory

Mutual information can also be expressed as a Kullback-Leibler distance. The mutual information between a random variable X and Y is the Kullback-Leibler distance between the joint distribution P(X,Y) and the product of the marginal P(X) P(Y):

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y)). \quad \ldots(10)$$

If X and Y are independent, so that P(X,Y) = P(X) P(Y), then

$$\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) = \log 1 = 0. \quad \ldots(11)$$

I(X;Y) = 0. That is Y tells no information at all about X.

Using the definitions of mutual information, it is straightforward to show that the mutual information can also be written as
I(X;Y) = H(X) – H(X | Y).

From this definition we can say mutual information is the difference between the average uncertainty in X and the uncertainty in X there still is after measuring Y. Thus it quantifies how much information Y tells about X [22].

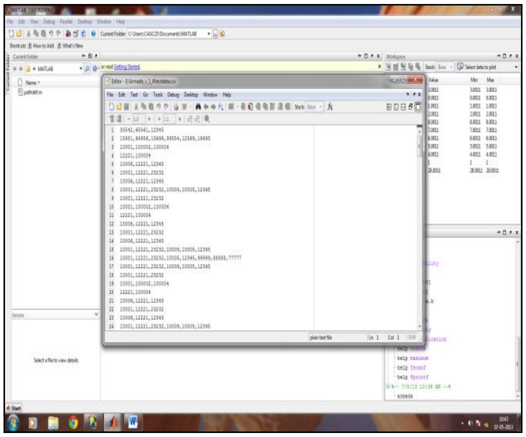
4.3.Properties Of Mutual Information

- A basic property of the mutual information is that knowing Y, we can save an average of bits in encoding X compared to not knowing Y.
- Mutual information is symmetric (i.e. I(X;Y) = I(Y;X)).
- Mutual information I (X;Y) can never be negative(i.e. I(X;Y) ≥ 0;). [D(P || Q) >=0]

From the above discussion it is clear that though Mutual information is a measure of the dependency between two variables, if the two variables are independent, the mutual information between them is zero. If the two are strongly dependent, e.g., one is a function of another; the mutual information between them is large.

5. Experiment and Result

We collect a transactional data (File Test data) and using Matlab version 9.0, we generate frequent item sets using association rule mining algorithm. Then we collect frequent item sets and calculate correlation measures over them. Following are the details:-



```
35542,65561,12345
15931,64956,15699,99554,12569,18695
10001,100002,100004
12221,100004
10006,12221,12345
10001,12221,23232
10006,12221,12345
10001,12221,23232,10009,10005,12345
10001,12221,23232
10001,100002,100004
12221,100004
10006,12221,12345
10001,12221,23232
10006,12221,12345
10001,12221,23232,10009,10005,12345
10001,12221,23232,10005,12345,99999,88888,77777
10001,12221,23232,10009,10005,12345
10001,12221,23232
10001,100002,100004
12221,100004
10006,12221,12345
10001,12221,23232
10006,12221,12345
10001,12221,23232,10009,10005,12345
```

Figure 1: Test Data Element In Matlab 9.0

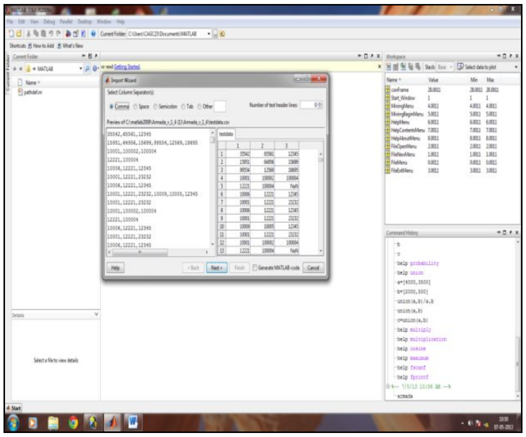
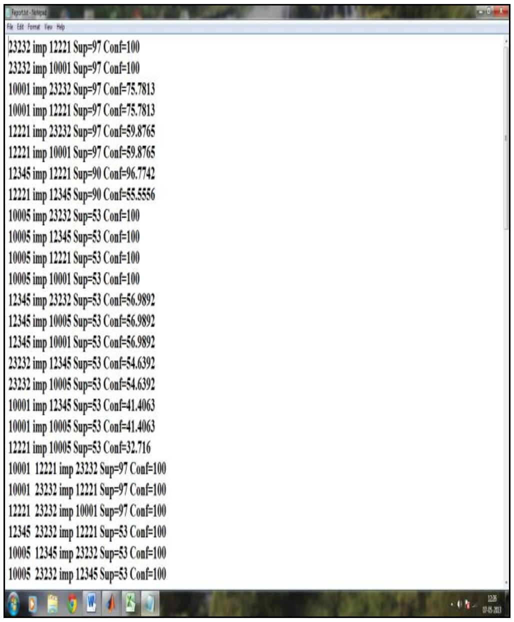
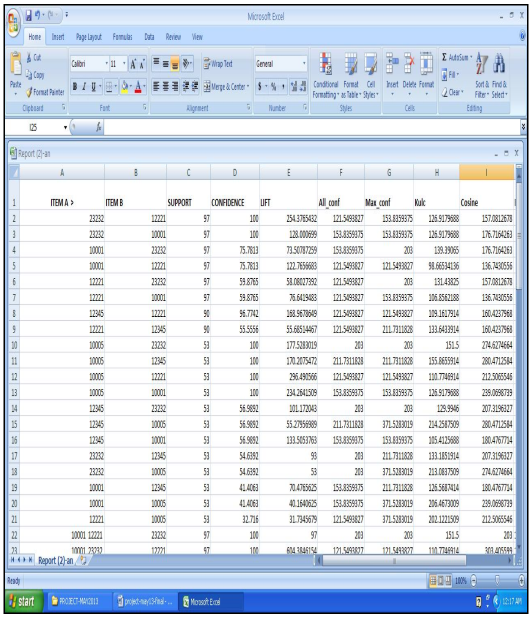


Figure 2: Generate Frequent Item Set Using Association Rule Mining Algorithm IN Matlab 9.0



```
23232 imp 12221 Sup=97 Conf=100
23232 imp 10001 Sup=97 Conf=100
10001 imp 23232 Sup=97 Conf=75.7813
10001 imp 12221 Sup=97 Conf=75.7813
12221 imp 23232 Sup=97 Conf=59.8765
12221 imp 10001 Sup=97 Conf=59.8765
12345 imp 12221 Sup=90 Conf=96.7742
12221 imp 12345 Sup=90 Conf=55.5556
10005 imp 23232 Sup=53 Conf=100
10005 imp 12345 Sup=53 Conf=100
10005 imp 12221 Sup=53 Conf=100
10005 imp 10001 Sup=53 Conf=100
12345 imp 23232 Sup=53 Conf=56.9892
12345 imp 10005 Sup=53 Conf=56.9892
12345 imp 10001 Sup=53 Conf=56.9892
23232 imp 12345 Sup=53 Conf=54.6392
23232 imp 10005 Sup=53 Conf=54.6392
10001 imp 12345 Sup=53 Conf=41.4063
10001 imp 10005 Sup=53 Conf=41.4063
12221 imp 10005 Sup=53 Conf=32.716
10001  12221 imp 23232 Sup=97 Conf=100
10001  23232 imp 12221 Sup=97 Conf=100
12221  23232 imp 10001 Sup=97 Conf=100
12345  23232 imp 12221 Sup=53 Conf=100
10005  12345 imp 23232 Sup=53 Conf=100
10005  23232 imp 12345 Sup=53 Conf=100
```

Figure 3: Frequent Item Sets Generation From "Test Data" File In Matlab 9.0



	A	B	C	D	E	F	G	H	I
1	ITEM A >	ITEM B	SUPPORT	CONFIDENCE	LIFT	All_conf	Max_conf	Kulc	Cosine
2	23232	12221	97	100	254.3765432	121.5493827	153.8359375	126.9179688	157.0812678
3	23232	10001	97	100	128.000699	153.8359375	153.8359375	126.9179688	176.7164263
4	10001	23232	97	75.7813	73.50787259	153.8359375	203	139.39065	176.7164263
5	10001	12221	97	75.7813	122.7656683	121.5493827	121.5493827	98.66534136	136.7430556
6	12221	23232	97	59.8765	58.08027392	121.5493827	203	131.43825	157.0812678
7	12221	10001	97	59.8765	76.6419483	121.5493827	153.8359375	106.8562188	136.7430556
8	12345	12221	90	96.7742	168.9678649	121.5493827	121.5493827	109.1617914	160.4237968
9	12221	12345	90	55.5556	55.68514467	121.5493827	211.7311828	133.6433914	160.4237968
10	10005	23232	53	100	177.5283019	203	203	151.5	274.6274664
11	10005	12345	53	100	170.2075472	211.7311828	211.7311828	155.8655914	280.4712584
12	10005	12221	53	100	296.490566	121.5493827	121.5493827	110.7746914	212.5065546
13	10005	10001	53	100	234.2641509	153.8359375	153.8359375	126.9179688	239.0698739
14	12345	23232	53	56.9892	101.172043	203	203	129.9946	207.3196327
15	12345	10005	53	56.9892	55.27956989	211.7311828	371.5283019	214.2587509	280.4712584
16	12345	10001	53	56.9892	133.5053763	153.8359375	153.8359375	105.4125688	180.4767714
17	23232	12345	53	54.6392	93	203	211.7311828	133.1851914	207.3196327
18	23232	10005	53	54.6392	53	203	371.5283019	213.0837509	274.6274664
19	10001	12345	53	41.4063	70.4765625	153.8359375	211.7311828	126.5687414	180.4767714
20	10001	10005	53	41.4063	40.1640625	153.8359375	371.5283019	206.4673009	239.0698739
21	12221	10005	53	32.716	31.7345679	121.5493827	371.5283019	202.1221509	212.5065546
22	10001 12221	23232	97	100	97	203	203	151.5	203
23	10001 23232	12221	97	100	604.3846154	121.5493827	121.5493827	110.7746914	303.405599

Figure 4: Calculation Of Correlation Over Frequent Item Sets Generated From Test Data Using (1),(2),..(6) Equations

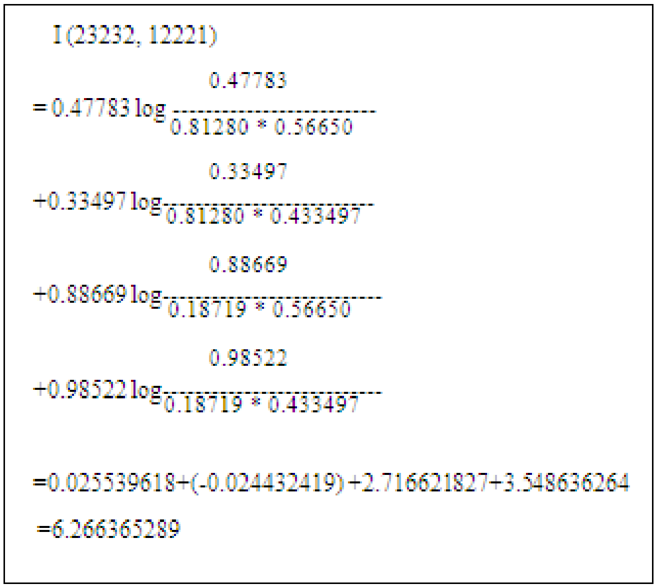
In this paper we already have done correlation analysis on test data and we calculate P(A) and P(B) [Probability of item A and Probability of item B]. From the above example, we take all the possible combination of two items such as ITEM A 23232 and ITEM B 12221 and we make a contingency table below

	12221	$\overline{12221}$	Σ row
23232	97	68	165
$\overline{23232}$	18	20	38
Σ col	115	88	203

Table 1: Contingency table of Item A 23232 and Item B 12221

	12221	$\overline{12221}$	Σrow
23232	97/203 =0.47783	68/203 =0.33497	165/203 =0.81280
$\overline{23232}$	18/203 =0.88669	20/203 =0.98522	38/203 =0.18719
Σ col	115/203 =0.56650	88/203 =0.43349	203/203 =1

TABLE 2: CALCULATION OF MUTUAL INFORMATION



$$
\begin{aligned}
& I(23232, 12221) \\
&= 0.47783 \log \frac{0.47783}{0.81280 * 0.56650} \\
&+ 0.33497 \log \frac{0.33497}{0.81280 * 0.433497} \\
&+ 0.88669 \log \frac{0.88669}{0.18719 * 0.56650} \\
&+ 0.98522 \log \frac{0.98522}{0.18719 * 0.433497} \\
&= 0.025539618 + (-0.024432419) + 2.716621827 + 3.548636264 \\
&= 6.266365289
\end{aligned}
$$

From the above calculation we see that the value of mutual information is large i.e. the above two items are strongly related.
We can calculate the mutual information values taking other items of Test data such as (10001 12221; 23232) or (10001 12221 12345; 23232) or (10001 10005 12221 12345; 23232) in the same way to find the dependency between two variables.

6. Conclusion

So from this above discussion we can say, Correlation measures which is used in the linear relationship (Pearson's correlation) or monotonic relationship between two variables, X and Y, whereas Mutual information is more general and measures the reduction of uncertainty in Y after observing X. Mutual information is based on information theory and it works not only in linear data but also in parabolic datasets. Mutual information helps reduce the range of the probability density function (reduction in the uncertainty) for a random variable X if the variable Y is known. The value of I(X; Y) is relative, and the larger its value, the more information that is known of X. It is generally beneficial to try to maximize the value of I(X; Y), thus minimizing uncertainty. So, if we imply mutual information as a measure to get stronger associated frequent item sets in linear and parabolic item sets, it will work more accurately than other measures which we discussed earlier.

7. Acknowledgement

Many people have cooperated with us during the process of translating a collection of ideas into this research. So, we take this opportunity to thank all for giving us the opportunity to complete this paper.

8. References

1. J. Han, M. Kamber and J. Pei, “Data Mining Concepts and Techniques,” Morgan Kaufmann, 3rd edition.
2. R. Agrawal, T. Imielinski, & A. Swami, 1993, “Mining association rules between sets of items in large databases,” ACM SIGMOD Record 22(2), 207-216.
3. R. Agrawal, & R. Srikant 1994, “Fast Algorithms for Mining Association Rules in Large Databases,” In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile.
4. D. McNicholas Paul, University of Guelph, Ontario, Canada, Z. Yanchang, University of Technology, Sydney, Australia, “Association Rules :An Overview”.
5. A. Savasere, E. Omiecinski, & S. B. Navathe, 1998, “Mining for strong negative associations in a large database of customer transactions,” Proceedings of the 14th International Conference on Data Engineering, Washington DC, USA, 1998, (pp. 494–502).
6. C. Silverstien, S. Brin, & R. Motwani, 1998, “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discover” 2(1), 39-68.
7. D. B. Fogel, 1997, “The advantages of evolutionary computation”. In D. Lundh, B. Olsson, & A. Narayanan, (Eds.), “Bio-Computing and Emergent Computation,” World Scientific Press, Singapore.
8. S. O. Rezende, University of São Paulo, Brazil, E. A. Melanda, Federal University of São Carlos, Brazil, M. Fujimoto, University of São Paulo, Brazil, R. A. Sinoara, University of São Paulo, Brazil, V. O. de Carvalho, University of Oeste Paulista, Brazil, “Combining Data-Driven and User-Driven Evaluation Measures to Identify Interesting Rules”.
9. R. Agrawal, & G. Psaila, 1995, “Active data mining”. In U. M. Fayyad, & R. Uthurusamy, (Eds.), Proceedings of the 1st ACM SIGKDD International Conference on “Knowledge Discovery and Data Mining”, (pp. 3–8), Montreal, Quebec, Canada. AAAI Press, Menlo Park, CA, USA. Au, W.-H., & Chan, K. (2005). “Mining changes in association rules: A fuzzy approach. Fuzzy Sets and Systems,” 149(1), 87–104.
10. C. C. Agarwal, & P. S. Yu (1998), “Mining Large Item sets for Association Rules,” In Bulletin of the Technical Committee, IEEE Computer Society, 21(1) 23-31.
11. S. G. Linoff and M. J. Berry, “Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management,” 3rd edition 1 April 2011.
12. R. Agrawal, and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc VLDB, 1994.
13. M. Boettcher, University of Magdeburg, Germany, G. Ruß, University of Magdeburg, Germany, D. Nauck, BT Group plc, UK, R. Kruse, University of Magdeburg, Germany, “From Change Mining to Relevance Feedback: A Unified View on Assessing Rule Interestingness”.
14. C. R. Feelders, & A. Siebes, 2001, “MAMBO: Discovering association rules based on conditional independencies,” In Proceedings of 4th International Symposium on Intelligent Data Analysis, Cascais, Portugal.
15. C. Silverstien, S. Brin, & R. Motwani, 1998, “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discover,” 2(1), 39-68.
16. P. E. Latham and Y. Roudi, “Mutual information,” 2009 Scholarpedia.
17. C. E. Shannon, “The mathematical theory of communication,” Bell Syst. Techn. Journal 27, (1948).
18. A. Fraser and H. Swinney, “Independent coordinates for strange attractors from mutual information,” Phys. Rev. A 33, (1986).
19. S. Brin, R. Motwani, and C. Silverstein, “Beyond market basket: Generalizing association rules to correlations,” In Proc. SIGMOD, May 1997.
20. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” In Proc. SIGMOD, May 2000.
21. T. K. Moon, 2005, “A Context for Error Correction Coding, in Error Correction Coding: Mathematical Methods and Algorithms,” John Wiley & Sons, Inc., Hoboken, NJ, USA.
22. T. M. Cover and J. A. Thomas “Elements of Information Theory,” Wiley India, 2009.
23. C. E. Shannon, 1948, “A Mathematical Theory of Communication,” Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948.
24. R.V.L. Hartley, “Transmission of Information,” Bell System Technical Journal, July 1928.
25. A. Kolmogorov, 1968, “Three approaches to the quantitative definition of information,” in International Journal of Computer Mathematics.