Mining change in customers' buying behavior on different time snapshot datasets

Niti Desai, Amit Ganatra

Abstract— Customers' buying behavior is changing because of change in Life style, liking and disliking, priorities of purchase. Such changes have occurred due to demographical changes, technological advancement etc. Earlier, people typically bought TV→ VCR. The buying pattern is replaced by new technological invention TV→ DVD. Apart from purchase behavior of customer in terms of item(s), purchase time span has also changed. Earlier purchase period of TV→DVD was 6 months to 1 year. This gap is reduced to 3 months - 6 months. This reduction might be because of revised income structure, change in life style, increase purchase power etc. This paper attempted to capture such behavioral changes of customers over time. Real-time Retail dataset capture at different time slot is used to understand changing behavior of customer. Proposed algorithm is working in two phase's namely sequential rule of purchase and frequent purchase. Frequent item purchase is mine based on Support (frequency of items) and confidence (highly co-related items) of items in huge sequential dataset. Proposed sequential pattern mining algorithm is working on FP- Growth based PrefixSpan for time stamp based sequential dataset. Later phase of proposed framework compare two datasets which are captured at different time. It's able to evaluate three interesting things of purchase: 1) New patterns captured, 2) Obsolete patterns and 3) Emerging patterns (EPs). Empirical experiment has conducted on real-time sequential dataset *'retail'*. Data has been captured on two different time shot. Various experiment evaluates, 12.33% more purchase sequences are generated by latest dataset. Out of them 13% purchase rules are newly discovered.33% buying patterns are obsolete patterns, which are no longer in existence.

Index Terms— Customers' buying behavior, Emerging Patterns (EPs), New Patterns, Obsolete Patterns, PrefixSpan, Sequential Pattern Mining (SPM), Time snapshot sequential dataset

1 LITRATURE SURVEY

SEQUENTIAL patterns have been applied in many domains like retail marketing, fraud detection, understanding of customer, web click analysis and medical research. Plenty of work has been done on data mining in changing environment [1],[2].Association rule mining is used to find out interesting relationship of items from large transaction dataset [3],[4].Incremental updating techniques are proposed for efficient maintenance of discovered association rules when new transaction data are added in original transaction dataset. But it does not reflect on changes of association rules[5].Certain patterns are identified whose support are significantly change from one dataset to another dataset, known as 'Emerging Patterns' (EPs) [6]. There are certain application in which occurrence of event is equally important known as sequential pattern mining (SPM) which can be identified by Apriori and FP-Growth techniques.[1],[7].Apriori based — GSP[8], SPAM[9], SPADE[10], SPRIT[11] and FP-Growth based Freespan[12] and PrefixSpan[13] are well known SPM techniques. Among those methods PrefixSpan defeated other SPM methods in context of execution time, memory consumption and pattern generation [13],[14].Current SPM techniques emphasis on such buying patterns which are mostly preferable by customers presently. Customers' buying behavior is changing over time, it is essential to track such buying priorities of customer. Customers' changing behavior is captured by association rule mining. Changes of patterns are defied as Emerging Patterns (EPs), Unexpected change in patterns and added/perished rules [15]. Paper focused on how customers' buying behavior is changing at different time snapshot using sequential pattern mining. *Three* interesting things of purchase are captured using FP growth based PrefixSpan algorithm. Paper focused on following interesting analysis of purchase: 1) New purchase patterns 2) Obsolete buying patterns and 3) Change in occurrence of purchase at various time snapshot. Above experiment has been conducted on *retail* real-time sequential dataset.

2 THEORETICAL FRAMEWORK

2.1 Discovery of Sequential patterns

FP-Growth based Prefixspan is use to find out frequent sequences or frequent purchase. Paper tried to focus on change of buying habit of customer as different point in time. So, two datasets are taken which are captured at different time slots for experiment. Sequential patterns based on support threshold are used to capture buying behavior of customer. Here, support and confidence threshold values are kept very less. So, the patterns which are suffering of low support values can't be rejected.

Support: The Support of an item set expresses how often the

Niti Desai *is pursuing Ph.D in Computer Engineering from Uka Tarsada University, Bardoli, Surat, Gujarat from 2012. She has 8+ years of teaching experience.Currentlty working as Assistant Professor in computer science and technology depertment at Usha Mital Institute of Technology – Mumbai.She has publishes and presented 10+ papers in national/international conferences andreputed journals. E-mail:* nitiadesai@gmail.com

Dr. Amit Ganatra is Professor, Head in computer Engineering Department at CSPIT, CHARUSAT and Dean in Faculty of Technology-CHARUSAT, Gujarat (since Jan 2011 to till date). He is a member of Board of Studies (BOS), Faculty Board, Academic Council and Governing Body for CHARUSAT and member of BOS for Gujarat Technological University (GTU), KSV and Indus University. He has 15+ years of teaching experience. He is having good research record. He has published and contributed over 90+ papers (as an Author and a Co-author) referred journals and presented in various international conferences. E-mail: amitganatra.ce@charusat.ac.in

item set appears in a single transaction in the database i.e. the support of an item is the percentage of transaction in which that items occurs.

$$\mathbf{I=P\ (X\cap Y) = (X\cap Y)\ /\ N} \quad (1)$$

Range: [0, 1] If I=1 then Most Interesting If I=0 then Least Interesting

Confidence: Confidence or strength for an association rule is the ratio of the number of transaction that contain both antecedent and consequent to the number of transaction that contain only antecedent.

$$\mathbf{I=P\ (Y/X) =P\ (X\cap Y)/\ P(X)} \quad (2)$$

Range: [0, 1] If I=1 then Most Interesting If I=0 then Least Interesting.

2.2 Comparison of sequential patterns at different time snapshot

Buying habit of customer is changing over a time. It is necessary to capture such buying habit. It is important to understand such change. Paper tries to capture *emerging trend* of buying by changes in frequency. Paper is also able to capture *completely new purchase trend* and *out dated purchase trend.*

Following notation are used to represent dataset at different time snapshot:

D^t, D^{t+k} :Dataset at time *t* and *t+k*

R^t,R^{t+k} :Discovered ruleset at time *t* and *t+k*

SEQ^{t+k} , SEQ^t *:Discovered set of frequent sequences t+k and t.*

r_i^t, r_j^{t+k} :Each rule from corresponding ruleset where i=1,2,3.... | R^t | j=1,2,3... | R^{t+k} |

seq_i^t, seq_j^{t+k}:Each sequences from corresponding sequences where i=1,2,3.... | SEQ^t | and j=1,2,3... | SEQ^{t+k} |

A) Emerging Patterns (EPs) [6],[15]

For rule r_j^{t+k} following two conditions are met, than we call it the rule of emerging Pattern with respect to r_i^t

1. Conditional and consequent parts are same between r_i^t and r_j^{t+k}
2. Supports of two rules are significantly different.

$$\delta = \frac{\text{Support}^{t+k}(\text{ri}) - Support^t(\text{ri})}{\alpha_{DB}} \quad (3)$$

where , α_{DB} *is growth in size of database in* (%)

$$\alpha_{DB} = \frac{n_{t+k} - n_t}{n_t} \times 100 \quad (4)$$

Where, n_{t+k} is sequence generated by growth in size of database D^{t+k} and n_t is sequence generated by growth in size of database D^t. Support of any sequence is growing as per growth of dataset is normal. $\alpha_{DB} \geq 0$. If there is no change in database captured at time t and at t+k, then $\alpha_{DB} = 0$. Expected growth rate δ, is changed because of change in size of dataset. If dataset is not changing over time period than δ=0. δ ≤0.2 is considered as normal growth rate for sequence or rule. Rules contain δ ≥ 0.2 are considered as Emerging Patterns.

B) New Patterns

There are several patterns which are detected in dataset captured at later time span. These patterns are called as newly generated patterns. New_SEQ is set of patterns contain new patterns which are not detected before.

$$\text{New\_SEQ}^{t+k} = SEQ^{t+k} - SEQ^t \quad (5)$$

$\forall seq_i^{t+k} \in \text{New\_SEQ}^{t+k}$ and $\forall seq_i^{t+k} \notin SEQ^t$ where i={1,2..n}, $\text{New\_SEQ}^{t+k} \subset SEQ^{t+k}$

C) Obsolete Patterns

There are several patterns which are detected only in historical dataset but should not present in dataset captured at later time span. Obs_SEQ is set of patterns contain such patterns which are not detected in later dataset.

$$\text{Obs\_SEQ}^t = SEQ^t - SEQ^{t+k} \quad (6)$$

$\forall seq_i^t \in \text{Obs\_SEQ}^t$ and $\forall seq_i^t \notin SEQ^{t+k}$ where i={1,2..n}, $\text{Obs\_SEQ}^t \subset SEQ^t$

[Note: (A) is partially novel but (B) and (c) are absolutely novel idea of authors]

D) Representation of Sequence Dataset

Data-sequence *A* is represented as <(a_1, t_1), (a_2, t_2), …,(a_n, t_n)>, where (a_j, t_j) means that item a_j is purchased at time t_j *where,* 1 ≤ *j* ≤ *n*, and t_j*-1* ≤ t_j for 2 ≤ *j* ≤ *n*. In the data-sequence, if items occur at the same time, they are ordered alphabetically.

Dataset Format (SequenceDatabase.txt)
<1> 1 -1 <2> 1 2 3 -1 <3> 1 2 -1 -2
<1> 1 -1 <2> 1 2 -1 <3> 1 2 -1 <4> 1 3 -1 -2
<1> 1 2 -1 <2> 1 2 -1 -2
<1> 2 -1 <2> 1 3 -1 -2
<1>, <2> ….. - TimeStamp (day at which transaction occurred)
-1 - End of Transaction
-2 - End of Sequence
1,2,3 - Actual Items

2.3 System Framework

Figure 1 describes system architecture to detect change in customers' purchase behavior at different time snapshot. System architecture is divided in two phases namely;

Phase-1: *Sequential patterns of purchase:* sequential rule generation at different time snapshot.

Phase-2: *Compare patterns* generated at different time snapshot which extracts:

1) Emerging Patterns (EPs)
2) New Patterns
3) Obsolete Patterns

Following things can be analyzed from above mentioned phases:

1) New patterns captured, which are not detected earlier.
2) Obsolete patterns are no longer in existence.
3) Change in occurrence of patterns based on its Frequency.
4) Drastic change in frequency of patterns from one dataset to another dataset is used to capture Emerging patterns (EPs).

3 PROPOSED ALGORITHM

Input: D^t, D^{t+k} : Dataset at time *t* and *t+k*
Output: (i)New Rules New_R^{t+k}
(ii)Obsolete Rules Obs_R^t
(iii)Emerging patterns (EPs)

Phase-I

Input: D^t, D^{t+k} : Dataset at time *t* and *t+k*

Procedure: call Seq(D^t , sup)
call Seq(D^{t+k} , sup)
Output:R^t,R^{t+k} :Discovered ruleset at time *t* and *t+k*
SEQ^t,SEQ^{t+k} :Discovered Complete set of sequences at time *t* and *t+k*

Phase-II

Input: R^t , R^{t+k} :Discovered ruleset at time *t* and *t+k*
SEQ^t,SEQ^{t+k} :Discovered Complete set of sequences at time *t* and *t+k* ***(Output of Phase-I)***
Procedure: Call CompareSeq(SEQ^t, SEQ^{t+k}, sup)
Call EPs(SEQ^t, SEQ^{t+k})
Output: New Rules New_R^{t+k} :Discovered rules at time *t+k* (5)
Obsolete Rules Obs_R^t :Discovered rules at time *t* (6)
Emerging patterns (EPs) (7)

Procedure: Seq(Dataset D, Support sup)

Step-1: Find length-1 patterns and remove irrelevant sequences.

i. Scan the sequence database SDB once to count f-support for each itemset.
ii. Identify patterns as length-1 patterns.
iii. Infrequent items are removed and generate pseudo-sequence database.

Step-2: Divide the set of sequential patterns into subsets

i. Without considering constraint C, the complete set of sequential patterns should be divided into subsets without overlap according to the set of length-1 sequential patterns (prefix).

Step-3: Construct projected database and mine subsets recursively

i. Construct projected database for each prefix.
ii. Recursively generate projected database for each new prefix and mine it to find local frequent patterns.

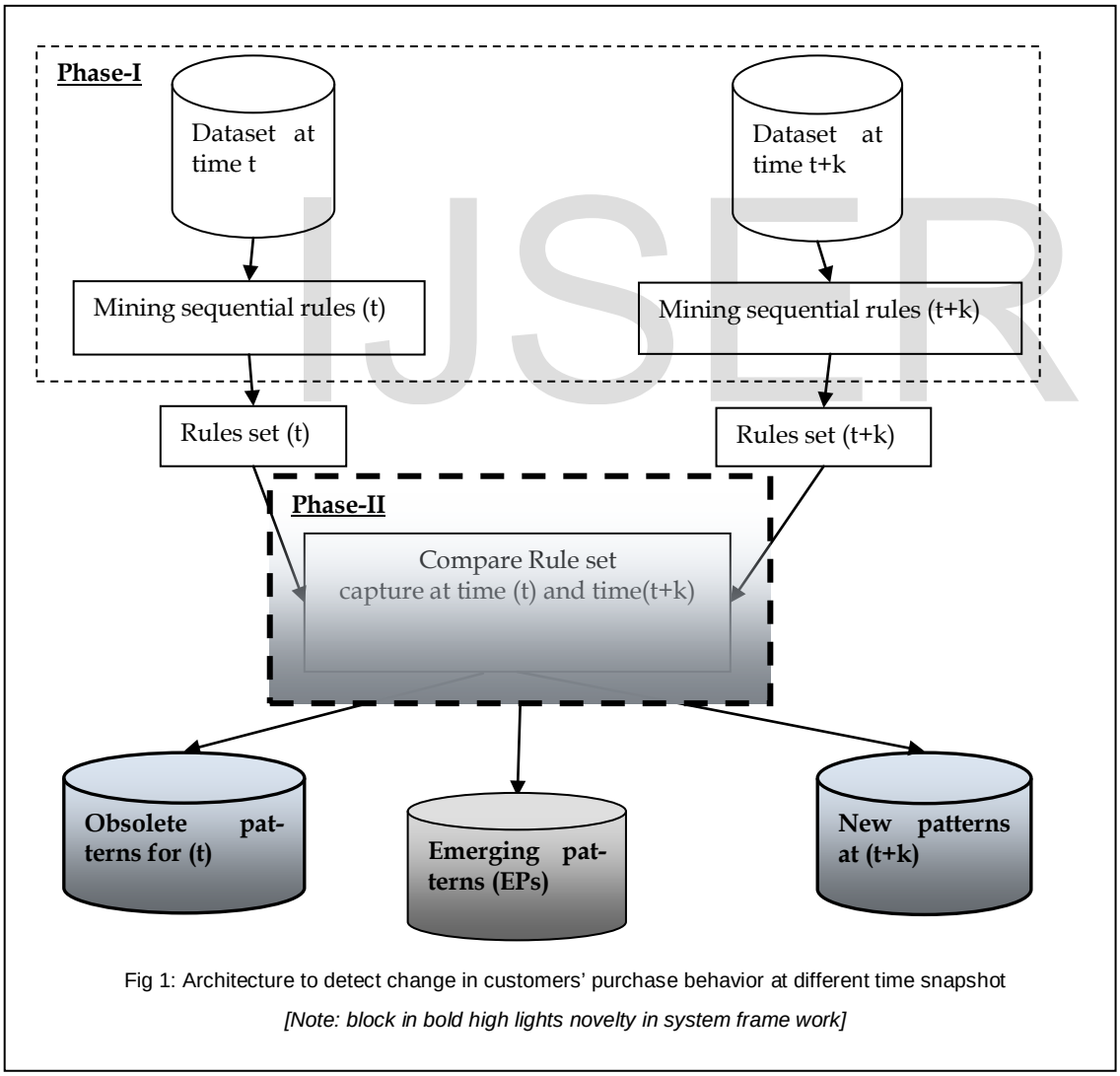

Fig 1: Architecture to detect change in customers' purchase behavior at different time snapshot

[Note: block in bold high lights novelty in system frame work]

Step-4: Mine sequential Pattern from Projected database and mine subset recursively.

For each frequent item i append it to prefix to generate new prefix in such a way that

a) i can be assembled to the last element of prefix to form a sequential pattern

or

b) <i> can be appended to prefix to form a sequential pattern.

Return* Complete set of sequential patterns *SEQ* and Rules Set *R

Procedure: CompareSeq(SEQ^t, SEQ^{t+k}, sup)

$Obs\_SEQ^t = SEQ^t - SEQ^{t+k}$ *(refer eq.6)*

$New\_SEQ^{t+k} = SEQ^{t+k} - SEQ^t$ *(refer eq. 5)*

seq_i^t, seq_j^{t+k} :Each sequences from corresponding sequences where i=1,2,3....|SEQ^t| and j=1,2,3...|SEQ^{t+k}|

$\forall seq_i^t \in Obs\_SEQ^t$ *and* $\forall seq_i^t \notin SEQ^{t+k}$ *where i={1,2..n},* $Obs\_SEQ^t \subset SEQ^t$

$\forall seq_i^{t+k} \in Obs\_SEQ^{t+k}$ *and* $\forall seq_i^{t+k} \notin SEQ^t$ *where i={1,2..n},* $New\_SEQ^{t+k} \subset SEQ^{t+k}$

Return (i)New Rules New_R^{t+k}
(ii)Obsolete Rules Obs_R^t

(iii)Emerging patterns (EPs)

Procedure: EPs(SEQ^t, SEQ^{t+k}) *(refer eq. 7)*

$SEQ_{common} \leftarrow$ Null
$SEQ_{EPs} \leftarrow$ Null
$SEQ_{common} = SEQ^t \cap SEQ^{t+k}$

$$\alpha_{DB} = \frac{n_{t+k} - n_t}{n_t} \times 100$$

(Where n_{t+k} and n_t is total sequences generated at time t+k and t respectively.)

$\forall seq_i \in SEQ_{common}$ *where i=1,2,3…..|* SEQ_{common} *|*
*For seq*i (i : 1→| *SEQcommon* |)

$$\delta = \frac{\text{Support}^{t+k}(seqi) - Support^t(seqi)}{\alpha_{DB}}$$

IF δ > 0.2 then
$SEQ_{EPs} \leftarrow SEQ_{EPs} \cup \{ seq_i \}$
Where,SEQ_{EPs} contain all emerging patterns.
END IF
END

Return SEQ_{EPs}

Table 1: Statistical Description

Parameters to evaluate	New Dataset (captured at later time slot)	Old Dataset (captured at earlier time slot)
Number of sequences	88162	12000
Number of distinct items	16470	9004
Average number of itemsets per sequence	10.30575	10.00
Average number of distinct item per sequence	10.30575	10.00
Average number of occurrences in a sequence for each item appearing in a sequence	1.0	1.0
Average number of items per itemset	1.0	1.0

4. EMPIRICAL FINDINGS AND DATA ANALYSIS

Algorithm is implemented in Java and tested on an Intel Core Duo Processor with 2GB main memory under Windows XP operating system.

4.1. Description of Real-time *Retail* Dataset [16]

Duration of data:
The data was obtained from a Belgian retail supermarket store.
The data are collected over three non-consecutive periods.
First period: starts from half December 1999 to half January 2000.
Second period: 2000 to the beginning of June 2000.
Third period: End of August 2000 to the end of November 2000.

Contain of Data:
The total amount of receipts being collected is 88,163. Data set contains information about the date of purchase:date, receipt number: receipt_nr, article number: article_nr, the number of items purchased: amount, the article price in Belgian Francs: price and the customer number: customer_nr.5,133 customers have purchased at least one product in the supermarket during the data collection period.

Two experiments have been performed on retail real-time datasets which are captured at different time period. First experiment evaluates: 1) execution time, 2) sequence generation and 3) memory utilization of datasets captured on different time slot. Old and new datasets are 914 KB and 6990 KB respectively in size. Table 1 shows statistical description of both datasets. Experiment has been done for 0.1%, 0.08% and 0.05% of support values. Average 12.13% more patterns are generated by dataset captured at later time slot is shown in figure 2. Execution time and memory consumption are increased with increment of size, is shown in figure 3 and figure 4 respectively.
Second experiment evaluates and analyzes, change in customers' buying behavior with respect to different time period. Customers' buying behavior captured at different time slots

for various supports like 0.1%,0.08% and 0.05%. Figure 5 discussed buying behavior of customers' for 0.1% of support value. 33% patterns are obsolete patterns, which are no longer in existence for later captured datset.12% (single) items are not preferred by customers.13% totally new preferences (sequences) are chosen by customers. Out of these new patterns, 31% new items are chosen alone.28.75% patterns are generated by *item 48* or sequences generated by *item 48* w.r.t historical patterns, which are not in existence for dataset captured at later stage.

Business manager can follow the changing trends using proposed framework. It can be utilized at macro and micro level business analysis.

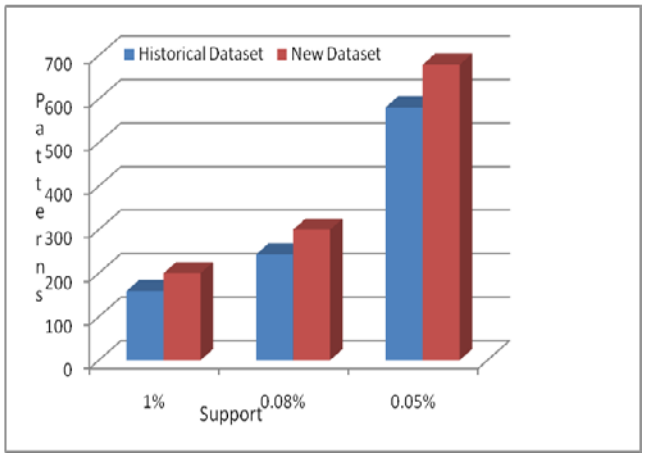

Fig 2: Sequence generation w.r.t support

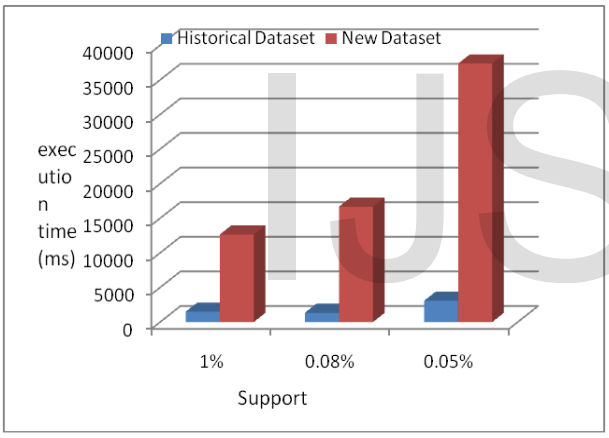

Fig 3: Execution time Vs. support

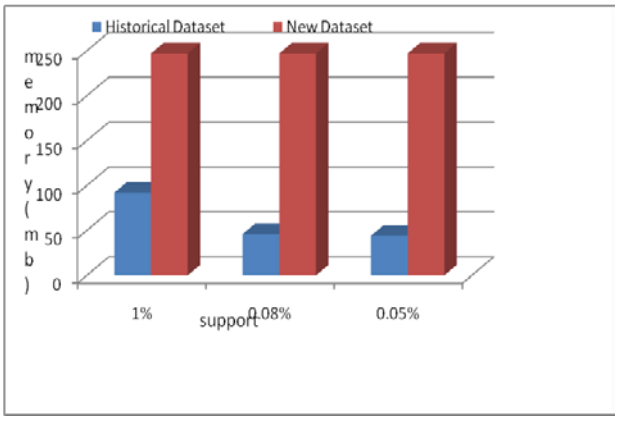

Figure 4: Memory consumption w.r.t support

5. BUSINESS IMPLICATION

5.1. Macro aspect

In macro aspect, business manager can track changing trend. It is important for them to detect their customers' changing behavior in order to provide products and services, which help them to with stand in competitive market. Detection of obsolete and new buying behavior helps them to focus on accumulation of goods. Manager can understand the decreasing rate of co-purchase of two certain products and then decision maker can examine the reason and develop a prevention plan to preserve the trend.

5.2. Micro aspect

In micro aspect, to track the changes are not only the sufficient factor for business manager but to understand *how?* and *why?*, parts in depth of such changes are equally important. Data captured at various time snapshot can help in detection of obsolete buying patterns, as well helps to understand the reasons of its drop off. Like, due to technological advancement, poor service or due to another high competitive product leads the item or co-purchase toward declining. Same way incremental changes in preferences help to predict forming stage of any purchase, which lead business man one step ahead than competitor. Proposed framework is useful in marketing of product or service and helpful in target customer group with the accumulation of demographics data.

6. DISCUSSIONS AND CONCLUSIONS

Evaluation and survey analysis of data captured at various time snapshots are not only helps companies, to understand customers' ever-changing needs, but also provides important information of future trend and incremental development of purchase tendency. For such purpose, growth of change is measured and on bases of that emerging patterns (EPs) are tracked. Proposed framework has also provides solution to detect new and obsolete trend. Paper has also outline practical use of proposed change detection framework for expansion and enhancement of business in macro and micro aspects.

Later stage of paper illustrates, experiments performed on real time retail data of Belgian supermarket store. Datasets are captured at two different time snapshot. Average 12.13% more patterns are generated by dataset captured at later time slot. 33% patterns are obsolete patterns and 13% absolute new patterns are detected. Paper is not more focused on *how* and *why* part of change detection in purchase because of lack of demographics description of captured data.

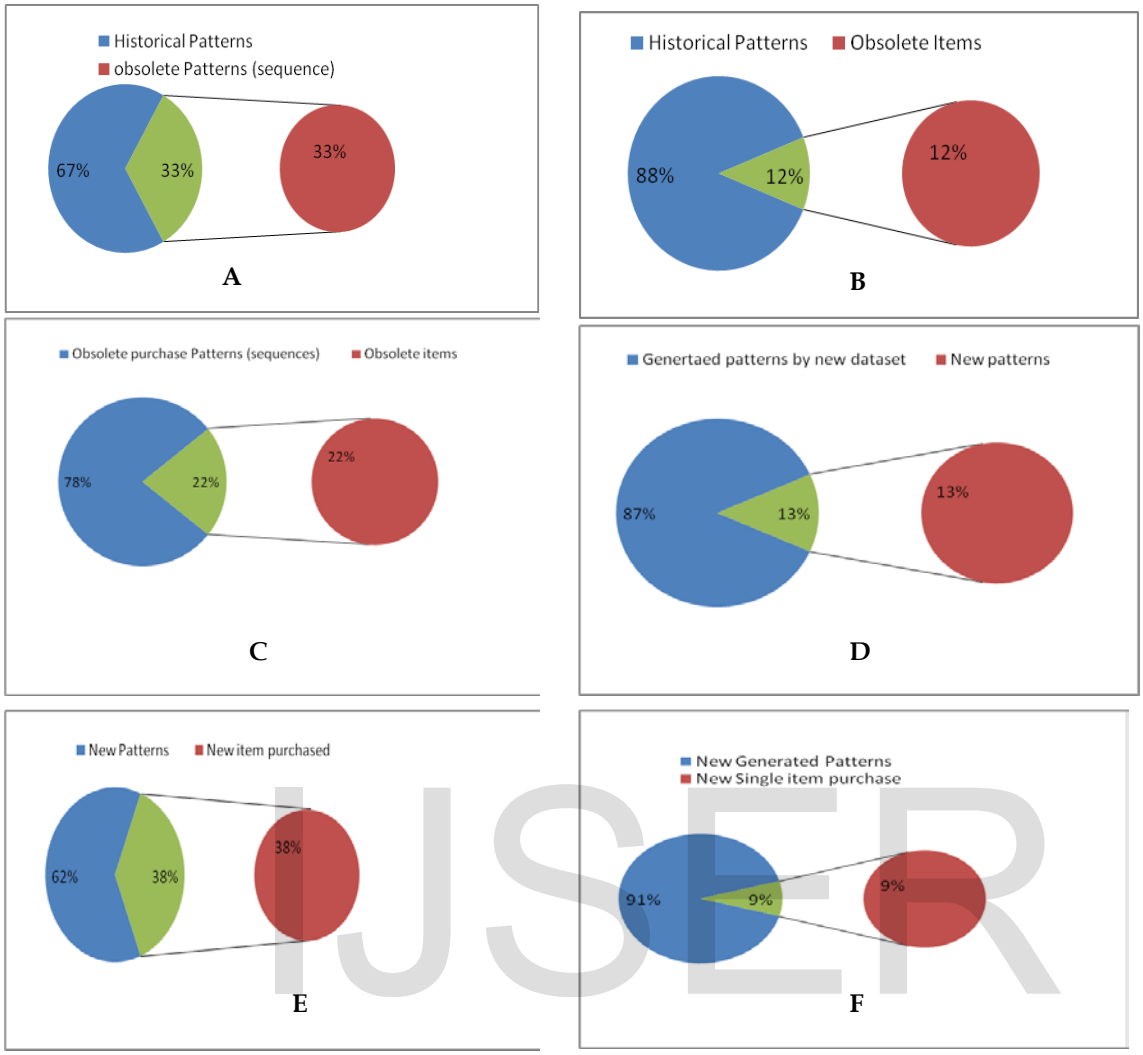

Fig 5: Study of Historical Patterns (A,B,C) and New Patterns (D,E,F)

REFERENCES

[1]Lui B.Hsu W., Han H.S. Mining changes for real-life applications. Second International conference on Data warehousing and Knowledge Descovery.2000 pp.337-346

[2] Ganti V ,Gehrke J ,Ramakrishnan R. Framework for measuring changes in Data characteristics. Proceding of the eighteenth ACM SIGACT- SIGMOD- SIGART symposium of principles of Database System (1999) pp.126-137

[3] R. Agrawal and R. Srikant.Fast Algorithms for Mining Association Rules. Preceding of 1994 International Conference of Very Large Data Bases (VLDB '94). pp. 487-499

[4] Agrawal R. And Srikant R. Mining Sequential Patterns. Proceding of the 11th International Conference on Data Engineering.Taipei-Taiwan.March 1995.

[5]Cheung D.W,,Han J.,Ng V T & Wong. Maintenance of large association rules in large database: In incremental updating techniques Proceeding of 12th International conference on data engineering (1996). pp.106-114.

[6] Dong Ji & Li . Efficient mining of emerging patterns: Discovering Trend and differences. Proceeding of fifth international conference on Knowledge Discovery and Data Mining 1999 pp.43-52.

[7] J. Han, J. Pei, and Y. Yin.Mining Frequent Patterns without Candidate Generation. Preceding 2000 ACM-SIGMOD. International Conference on Management of Data (SIGMOD "00).2000. pp. 1 -12.

[8] Srikant R. and Agrawal R.Mining sequential patterns: Generalizations and performance improvements. Proceedings of the 5th International Conference Extending Database Technology.1996.pp- 3-17.

[9] AYRES, J., FLANNICK, J., GEHRKE, J., AND YIU, T.Sequential pattern mining using a bitmap representation. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2002.

[10] M. Zaki. SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning. vol. 40.2001. pp. 31-60.

[11] M. Garofalakis, R. Rastogi, and K. Shim.SPIRIT: Sequential pattern mining with regular expression constraints. VLDB'99.

[12] Han J., Dong G., Mortazavi B., Chen Q., Dayal U., Hsu M.-C. Freespan: Frequent pattern-projected sequential pattern mining. Proceedings 2000 International Conference on Knowledge Discovery and Data Mining (KDD"00). 2000. pp. 355 -359.

[13] J. Pei, J. Han, B. Mortazavi, H. Pino. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix- Projected Pattern

Growth.ICDE'01.2001.

[14] Niti Desai, Amit Ganatra Sequential Pattern Mining Methods: A Snap Shot. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727.Volume 10. Issue 4,2013. PP 12-20

[15] Hee soek song,Jae Kyeong Kim, Soung Hie kim. Mining the change of customer behaviour in an internet shopping mall Expert system with Applications. Elesvair Science Ltd .2001.pp. 157-168

[16] Tom Brijs and Gilbert Swinnen and Koen Vanhoof and Geert Wets. Using Association Rules for Product Assortment Decisions: A Case Study .Knowledge Discovery and Data Mining,.1999.pp. 254-260.