A soft set theory approach for mining category wise sequential patterns

P. Nirmala Kumari[a] & Dr.D.V.S.R. Anil Kumar[b]
[a]Lecturer in Mathematics, KRK Government Degree College, Addanki.
[b]Professor of Mathematics, TKR College of Engineering & Technology, Hyderabad.

Received: July 01, 2018 Accepted: August 20, 2018

ABSTRACT *This paper presentsthe method for miningfrequent categories of the items in the largedatabase using soft sets. In this approach, the different categories of the items are treated as paramanothers and the item set as the universal set of the soft set.Theembedded patterns are discovered by usof equivalence classes of the preference value vectors of the soft item sets.*

Keywords: *Soft element, soft element set, soft sequence, soft sequence matrix.*

Introduction

Majority of the research work on sequential data mining was finding all common patterns hidden in the database of a sequence of items. They [1,4] indeed worked on finding all the maximal sequences whose frequency is greater than a specified threshold. But none of the work was done on the category wise mining of the patterns in the sequence of the items.Therefore we introduce the problem of mining sequential patterns based on the category of the items in the item set. We intend to mine frequent categories of the items in the sequences using soft sets. Since a soft set is a parameterized family of a universal set [2,3], by considering all the items in the database as a universal set it is possible to parameterize the items based on their category. Therefore the data is transformed into parameter wise subsets and we define a new method for extracting sequential patterns. We define the soft item, the frequency of soft item, soft item set, soft sequence, and soft sequence matrix. This paper is organized as follows. Section II presents preliminaries on soft sets and an introduction to sequential patterns. Section III presents the basic definitions required for the mining of sequential patterns using soft sets. Then calculating the frequency of items based on their categories is detailed in section IV and the paper ends with the conclusion.

Section II

Definition 2.1: [2]: Let U be an initial universal set and E be a set of parameters. Let P(U) denotes the power set of U and A ⊂ E. A pair (F,A) is called a soft set over U if and only if F is a mapping of A into P(U). Hence a soft set (F,A) = {(e, F(e)) / e ∈ A}.

Definition 2.2:[3]: The soft set (F,A) over U is said to be a null soft set if F(e) = ∅∀ $e \in A$ and is denoted by $\tilde{\emptyset}_A$. We usually ignore the suffix A and write $\tilde{\emptyset}$for $\emptyset_A$.

Definition 2.3:[3]: The soft set (F,A) over U is said to be an absolute soft set if F(e) = U ∀ $e \in A$ and is denoted by $\tilde{A}$.

Definition 2.4:[3] Let (F,A) be a soft set over U. Then the relative complement of (F,A) is defined by $(F,A)^c$ = (F^c, A) where F^c: A→P(U) is given by $F^c(e)$ = U- F(e), ∀ $e \in A$. Hence
$(F,A)^c$ = (F^c,A) = {(e, U- F(e)) / $e \in A$}.

Definition 2.5:[3] Let (F,A) and (G,B) be soft sets defined over the common universe U. Then (F,A) is said to be a soft subset of (G,B) if

(i) A ⊂ B, and
(ii) for all e ∈ A, F(e) and G(e) are identical approximations.

We write (F,A) $\tilde{\subseteq}$ (G,B)

Definition 2.6: [3] Let (F,A) and (G,B) be soft sets over the common universe U. We say that (F,A) and (G,B) are soft equal if (F,A) is a soft subset of (G,B) and (G,B) is a soft subset of (F,A).

Definition 2.7:[3] The union of soft sets (F,A) and (G,B) over the common universe U is the soft set (H,C) where C = A ∪ B and ∀ $e \in C$

$$H(e) = \begin{cases} F(e) & if\ e \in A - B \\ G(e) & if\ e \in B - A \\ F(e) \cup G(e) & if\ e \in A \cap B \end{cases}$$

We write (H,C) = (F,A) $\tilde{\cup}$ (G,B)

Definition 2.8:[3] The intersection of soft sets of (F,A) and (G,B) over a common universe U is the soft set (H,C), where C = A∩B and ∀ $e \in C$, H(e) = F(e) ∩ G(e).

We write (F,A) $\widetilde{\cap}$ (G,B) = (H,C).

Section III

This section lays the foundation for the mining of sequential patterns basing on their categories by introducing the concepts of the soft data set, a soft item, soft item set, soft sequence. These definitions are further detailed with suitable examples.

Definition 3.1:

Soft dataset: Let U be the set of all products (Items) available in the database and E be the set of all the broad categories of the products in the data set. A soft data set is a mapping from E into power set of U.

Example: Let U = { $i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{11}, i_{12}, i_{13}, i_{14}, i_{15}, i_{16}, i_{17}, i_{18}, i_{19}$} be the set of all products available in the supermarket and E be the set of its categories.

E= { e_1=cosmetics, e_2 = Pulses, e_3 = vegetables, e_4 = Frozen food, e_5 = edible oils, e_6= milk products}

The mapping F: E→P(U) defined as follows.

(F,E) = { (e_1, {i_1, i_2, i_3}), (e_2, {i_4, i_5, i_6, i_7}), (e_3, {i_8, i_9}), (e_4, {i_{10}, i_{11}, i_{12}}), (e_5, {$i_{13}, i_{14}, i_{15}, i_{16}$}), ($e_6$, {$i_{17}, i_{18}, i_{19}$}) }. Then (F,E) is a soft data set.

Definition 3.2: Soft item:A soft item is an ordered pair of items and its corresponding category. In other words, the ordered pair (e, I) called a soft item provided I is a subset of F(e).

In the above example (e_1, {i_1}), (e_4, {i_{10}, i_{11}}) are soft items.

Definition 3.3: Soft item set:The set of soft items is a soft item set. It is denoted by (A,X) where A ⊆ E and X⊆ U.

For example { (e_1, {i_1}), (e_4, {i_{10}, i_{11}}) } is a soft item set.

Suppose a customer purchases a sunscreen lotion, body soaps, toordal, moong dal, sunflower oil in one transaction in a day. The soft items corresponding to this customer are (Cosmetics, {sun screen lotion, body soaps}), (pulses, { toor dal, moong dal}), (edible oils, { sunflower oil}) and the soft item set is {(Cosmetics, {sun screen lotion, body soaps}), (pulses, { black gram, green gram}), (edible oils, { sunflower oil})}.

Frequency computation:

Assume that the set of parameters, E consists of n number of elements and number of soft sets with respect to each object is m.

Definition 3.4: The cardinality of the soft item $(e_i, F_k(e_i))_{O_r}$ is the cardinality of$(F_k(e_i))_{O_r}$ with respect to the soft item set$(F_k, E)_{O_r}$ for k = 1,2,3,…….m and with respect to the object O_r, for r= 1,2,3,……..t where t is the number of objects in the data set.

Definition 3.5: The preference value of a soft item $(e_i, F_k(e_i))_{O_r}$ is computed as the ratio to the cardinality of that soft item to the total cardinality of soft items of a soft item set and is denoted by $p\left((F_k(e_i))_{O_r}\right)$.

$$p\left((F_k(e_i))_{O_r}\right) = \frac{|(F_k(e_i))_{O_r}|}{\sum_{j=1}^{n}|(F_k(e_j))_{O_r}|}, \text{ i≠j}$$

Definition 3.6: The preference value vector of a soft item set is an ordered n-tuple $(p\left((F_k(e_1))_{O_r}\right), p\left((F_k(e_2))_{O_r}\right), p\left((F_k(e_3))_{O_r}\right), \ldots\ldots\ldots\ldots\ldots, p\left((F_k(e_n))_{O_r}\right))$ and is simply denoted by $(\alpha_k)_{O_r}$

If $(F_1,E)_{O_r}$, $(F_2,E)_{O_r}$, $(F_3,E)_{O_r}$, ……. $(F_m,E)_{O_r}$ are soft item sets with respect to the object O_r over a period of time then their preference value vectors are $(\alpha_1)_{O_r}$, $(\alpha_2)_{O_r}$, $(\alpha_3)_{O_r}$, ……. $(\alpha_m)_{O_r}$ respectively.

Definition 3.7: We say that a soft item set $(F_{m_i}, E)_{O_r}$ with preference value vector $(\alpha_{m_i})_{O_r}$ supports another soft item set $(F_{m_j}, E)_{O_r}$ withpreference value vector $(\alpha_{m_j})_{O_r}$ with respect to an object O_r if whenever $p(F_{m_i}(e_{i_1})) \leq p(F_{m_i}(e_{i_2})) \leq p(F_{m_i}(e_{i_3})) \leq \ldots\ldots\ldots\ldots\ldots\ldots\ldots p(F_{m_i}(e_{i_q}))$ implies that $p(F_{m_j}(e_{i_1})) \leq p(F_{m_j}(e_{i_2})) \leq p(F_{m_j}(e_{i_3})) \leq \ldots\ldots\ldots\ldots\ldots\ldots\ldots p(F_{m_j}(e_{i_q}))$.

Meaning to say that the preference values of each soft item in the two soft item sets follow the same order, Here $e_{i_1}, e_{i_2}, e_{i_3}, \ldots\ldots\ldots\ldots e_{i_q}$ are nothing but parameters (categories)$e_1, e_2, e_3, \ldots\ldots\ldots\ldots e_n$ written according to the ascending order of their preference values.

The relation "supports" defined above is an equivalence relation and this equivalence relation partitions the set of preference value vectors into equivalence classes denoted by $[(\alpha_m)_{O_r}]$ with respect to the object O_r. Here one has to observe that the preference value vector in the equivalence class, $[(\alpha_m)_{O_r}]$ follows the same order as in the $(\alpha_m)_{O_r}$.Let the pattern of the categories, $e_1, e_2, e_3, \ldots\ldots\ldots\ldots e_n$ of the preference value vector $(\alpha_m)_{O_r}$ in the descending order of their preferences be $(e_{i_1}, e_{i_2}, e_{i_3}, \ldots\ldots\ldots\ldots e_{i_q})$.

Definition 3.8: The frequency of the class $[(\alpha_m)_{O_r}]$ is the ratio to the number of elements in the class to the total number of preference value vectors of the objectO_r.

Thecategory wise frequent sequential pattern with respect to the object O_r is the pattern of the preference value vector $(\alpha_m)_{O_r}$ of the class $[(\alpha_m)_{O_r}]$ having the highest frequency.In this way category wise frequent sequential pattern for each object can be calculated.

The category wise frequent sequential pattern in the data set is the pattern followed by thehighest number of objects.

Section IV

In the present section, we applied the concepts of category wise sequential pattern mining to the data of a supermarket for finding the most frequent category of items sold in their store.

Suppose a supermarket has different categories of items like cosmetics, edible oils, vegetables, cleaning products etc. The shopkeeper intends to introduce new products in his store and his desire is to find the sequential order of categories of items purchased by the customers over a period of time.

Let U be the set of all items available in the store and E be the set of all categories of items available in the store. Let U = { c_1,c_2 ,c_3 , c_4 ,b_1,b_2,b_3,b_4, v_1,v_2, v_3} be the items available in the store where

c_1= face wash, c_2 = body lotion, c_3 = shampoo, c_4 = face powder, b_1= coco cola, b_2= Thums up, b_3= minute maid, v_1= potato, v_2= tomato, v_3= carrot.

Let E= { e_1,e_2,e_3} be the set of all categories of elements of U, E is the set of parameters. Here $e_1$represents cosmetics category, e_2 represents beverages, $e_3$represents vegetables. That means all the cosmetics in the store comes under the categorye_1, beverages under e_2 and vegetables undere_3.

The database of customer transactions in the store during a particular period is as follows.

Date	Customer Id	Items bought
March 23, 2016	o_1	c_1,c_2,c_3 , c_4 ,b_2,b_1 v_1
March 23, 2016	o_2	c_1,c_2,c_3 , c_4 ,b_1,b_2, b_3 v_1, v_2
March 23, 2016	o_3	c_1,c_2,c_3 , c_4 , , b_1,b_2, b_3 v_3
March 24, 2016	o_3	c_2,v_1,v_2,v_3
March 24, 2016	o_1	c_2,c_3 , c_4 ,b_1, b_3 , v_2
March 25, 2016	o_2	c_1,c_2,c_3 , c_4 ,b_1,b_4,v_3
March 25, 2016	o_3	c_2,c_3 ,v_1,v_2,v_3
March 28, 2016	o_1	c_1,c_2,c_3 , c_4 ,b_2,b_1 ,b_3,v_1
March 28, 2016	o_2	c_2,c_3 , b_2, v_1,v_2,v_3
March 28, 2016	o_3	c_3 ,v_2,v_3
March 30, 2016	o_1	c_1,v_2
March 30, 2016	o_3	c_3,v_1,v_2,v_3

Soft items and soft itemsets of the customers

Customer Id	Items bought	Soft items	Soft item set
o_1	c_1,c_2,c_3 , c_4 ,b_2,b_1 v_1	$(F_1(e_1))_{o_1}$= $(e_1,\{c_1,c_2,c_3,c_4,\})$ $(F_1(e_2))_{o_1}$=$(e_2,\{b_2,b_1\})$ $(F_1(e_3))_{o_1}$ = $(e_3,\{v_1\})$	$(F_1,E)_{O_1}$= $\{(e_1,\{c_1,c_2,c_3,c_4,\}),(e_2,\{b_2,b_1\}),$ $(e_3,\{v_1\})\}$
	c_2,c_3 , c_4 ,b_1, b_3 , v_2	$(F_2(e_1))_{o_1}$ = $(e_1,\{c_2,c_3$, c_4 $\})$ $(F_2(e_2))_{o_1}$ = $(e_2,\{$ $b_1,b_3\})$ $(F_2(e_3))_{o_1}$ =$(e_3,\{v_2\})$	$(F_2,E)_{O_1}$ = $\{(e_1,\{c_2,c_3$, c_4 $\})$, $(e_2,\{$ $b_1,b_3\})$, $e_3,\{v_2\})$ }
	c_1,c_2,c_3 , c_4 ,b_2,b_1 ,b_3,v_1	$(F_3(e_1))_{o_1}$=$(e_1,\{c_1,c_2,c_3,c_4,\})$ $(F_3(e_2))_{o_1}$=$(e_2,$ $\{$ b_2,b_1 , $b_3\})$ $(F_3(e_3))_{o_1}$ = $(e_3,\{v_1\})$	$(F_3,E)_{O_1}$=$\{(e_1,\{c_1,c_2,c_3,$ $c_4,\})$ $(e_2,$ $\{b_2\})$, $(e_3,\{v_2\})\}$
	c_1, v_1,v_2	$(F_4(e_1))_{o_1}$= $(e_1,\{c_1\})$, $(F_4(e_3))_{o_1}$ = $(e_3,\{v_2\})$	$(F_4,E)_{O_1}$ = $\{(e_1,\{c_1\})$, $(e_3,$ $\{v_1,$ $v_2\})\}$
o_2	c_1,c_2,c_3 ,c_4 ,b_1,b_2, *b3 v1, v2*	$(F_1(e_1))_{o_2}$=$(e_1,\{c_1,c_2,c_3$,c_4 $\})$ $(F_1(e_2))_{o_2}$=$(e_2,\{b_1,b_2,$ $b_3\})$ $(F_1(e_3))_{o_2}$=$(e_3,\{v_1,$ $v_2\})$	$(F_1,E)_{O_2}$=$\{(e_1,\{c_2\})$,$(e_2,\{b_1,b_2,b_3\})$, $(e_3,\{v_1,$ $v_2\})$
	c_1,c_2,c_3 , c_4 ,b_1,b_4,v_3	$(F_2(e_1))_{o_2}$= $(e_1,\{c_1,c_2,c_3,c_4$ $\})$ $(F_2(e_2))_{o_2}$ = $(e_2,\{b_1,b_4\})$ $(F_2(e_3))_{o_2}$ = $(e_3,\{v_3\})$	$(F_2,E)_{O_2}$ = $\{(e_1,\{c_1,c_2,c_3,c_4$ $\})$, $(e_2,\{b_1,b_4\})$, $(e_3,\{v_3\})\}$
	c_1,c_3 ,b_2,v_1,v_2,v_3	$(F_3(e_1))_{o_2}$ = $(e_1,\{c_1,c_3,\})$	$(F_3,E)_{O_2}$ =$\{$ $(e_1,\{c_1,c_3,\})$, (e_2, $\{$

		$(F_3(e_2))_{o_2} = (e_2, \{b_2\})$ $(F_3(e_3))_{o_2} = (e_3, \{v_1, v_2, v_3\})$	$b_2\}), (e_3, \{v_1, v_2, v_3\})\}$
o_3	$c_1, c_2, c_3, c_4, b_1, b_2, b_3, v_3$	$(F_1(e_1))_{o_3} =$ $(e_1, \{c_1, c_2, c_3, c_4\})$ $(F_1(e_2))_{o_3} = (e_2, \{b_1, b_2, b_3\})$ $(F_1(e_3))_{o_3} = (e_3, \{v_3\})$	$(F_1, E)_{O_3} = \{(e_1, \{c_1, c_2, c_3, c_4\}),$ $(e_2, \{b_1, b_2, b_3\}), (e_3, \{v_3\})\}$
	c_2, v_1, v_2, v_3	$(F_2(e_1))_{o_3} = (e_1, \{c_2\})$ $(F_2(e_2))_{o_3} = (e_2, \emptyset)$ $(F_2(e_3))_{o_3} = (e_3, \{v_1, v_2, v_3\})$	$(F_1, E)_{O_3} = \{e_1, \{c_2\}), (e_2, \emptyset),$ $(e_3, \{v_1, v_2, v_3\})\}$
	c_2, c_3, v_1, v_2, v_3	$(F_3(e_1))_{o_3} = (e_1, \{c_2, c_3\})$ $(F_3(e_2))_{o_3} = (e_2, \emptyset)$ $(F_3(e_3))_{o_3} = (e_3, \{v_1, v_2, v_3\})$	$(F_2, E)_{O_3} = \{e_1, \{c_2, c_3\}), (e_2, \emptyset),$ $(e_3, \{v_1, v_2, v_3\})\}$
	c_3, v_2, v_3	$(F_4(e_1))_{o_3} = (e_1, \{c_3\})$ $(F_4(e_2))_{o_3} = (e_2, \emptyset)$ $(F_4(e_3))_{o_3} = (e_3, \{v_2, v_3\})$	$(F_3, E)_{O_3} = \{(e_1, \{c_3\}), (e_2, \emptyset),$ $(e_3, \{v_2, v_3\})\}$
	c_3, v_1, v_2, v_3	$(F_2(e_1))_{o_3} = (e_1, \{c_3\})$ $(F_2(e_2))_{o_3} = (e_2, \emptyset)$ $(F_2(e_3))_{o_3} = (e_3, \{v_1, v_2, v_3\})$	$(F_1, E)_{O_3} = \{e_1, \{c_3\}), (e_2, \emptyset),$ $(e_3, \{v_1, v_2, v_3\})\}$

The preference values of soft items and their corresponding preference value vectors are as follows.

Customer	Soft item set	Preference value vector
o_1	$(F_1, E)_{O_1}$	$(\alpha_1)_{O_1} = (4/7, 2/7, 1/7)$
	$(F_2, E)_{O_1}$	$(\alpha_2)_{O_1} = (3/6, 2/6, 1/6)$
	$(F_3, E)_{O_1}$	$(\alpha_3)_{O_1} = (4/8, 3/8, 1/8)$
	$(F_4, E)_{O_1}$	$(\alpha_4)_{O_1} = (1/2, 0, 1/2)$
o_2	$(F_1, E)_{O_2}$	$(\alpha_1)_{O_2} = (4/9, 3/9, 2/9)$
	$(F_2, E)_{O_2}$	$(\alpha_2)_{O_2} = (4/7, 2/7, 1/7)$
	$(F_3, E)_{O_2}$	$(\alpha_3)_{O_2} = (2/6, 1/6, 3/6)$
o_3	$(F_1, E)_{O_3}$	$(\alpha_1)_{O_3} = (4/8, 3/8, 1/8)$
	$(F_2, E)_{O_3}$	$(\alpha_2)_{O_3} = (1/4, 0, 3/4)$
	$(F_3, E)_{O_3}$	$(\alpha_3)_{O_3} = (2/5, 0, 3/5)$
	$(F_4, E)_{O_3}$	$(\alpha_4)_{O_3} = (1/3, 0, 2/3)$
	$(F_5, E)_{O_3}$	$(\alpha_5)_{O_3} = (1/4, 0, 3/4)$

Now the equivalence classes, their frequencies,and their corresponding category wise patterns according to their importance are as follows.

$[(\alpha_1)_{O_1}] = \{(\alpha_1)_{O_1}, (\alpha_2)_{O_1}, (\alpha_3)_{O_1}\}$, the frequency of $(\alpha_1)_{O_1} = 3/4$ and its pattern is (e_1, e_2, e_3). $[(\alpha_1)_{O_2}] = \{(\alpha_1)_{O_2}, (\alpha_2)_{O_2}\}$the frequency of $(\alpha_1)_{O_2} = 2/3$ and its pattern is (e_1, e_2, e_3)

$[(\alpha_2)_{O_3}] = \{(\alpha_2)_{O_3}, (\alpha_3)_{O_3}, (\alpha_4)_{O_3}\}$, the frequency of $(\alpha_2)_{O_3} = 4/5$ and its pattern is (e_2, e_1, e_3)

Here the category wise sequential pattern with respect to the customer $o_1$is (e_1, e_2, e_3). The customer $o_1$preferred to purchase cosmetics, then beverages and then vegetables in the store.

The choice value of the pattern (e_1, e_2, e_3) is 2/3 and this pattern of the categories is preferred by the majority of the customers of the store.

Conclusion:

This paper presents a new method of mining category wise sequential patterns in the data set using soft sets. This method is achieved by defining preference values of each category of a soft item set and their corresponding preference value vectors. The method derived is supported by an example in the supermarket analysis.

References:

1. Celine Fiot, Student Member, IEEE, Anne Laurent,andMaguelonneTeisseire, "From Crispness to Fuzziness":Three algorithms for the soft sequential pattern mining, IEEE Transactions on Fuzzy Systems, Vol 15, No 6, December 2007.
2. D. Molodtsov (1999), Soft set theory - First Results, Computers and Mathematics with Applications,Vol.37,Issues4-5,Pp.19-31. https://doi.org/10.1016/S0898-1221(99)00056-5
3. P.K.Maji, R.Biswas& A.R. Roy (2003), Soft Set Theory, Computers and Mathematics with Applications, Vol. 45, Pp. 555-562. https://doi.org/10.1016/S0898-1221(03)00016-6
4. R. Agrawal, R. Srikanth, " Mining sequential patterns " in Proc. 11th IEEE Int. Conf. Data Eng., 1995, pp. 3-14.