**Στέργιος Τσάνταλης (iis21125)**

1. ***Apriori με το χέρι***

Ξεκινάμε και παίρνουμε τα Ν-στοιχειοσύνολα ξεκινώντας από τα 1-στοιχειοσύνολα,2-στοιχειοσύνολα κλπ.

Υπολογίζουμε για το καθένα τη συχνότητα εμφάνισης δηλαδή το **SUPPORT**. Ο υπολογισμός της συχνότητας αυτής γίνεται ως εξής:

Π.χ. το 1-στοιχειοσύνολο {beer} εμφανίζεται 4 φορές στα 5 αρχεία άρα S=4/5

Το 2-στοιχειοσύνολο {beer, eggs} εμφανίζεται 3 φορές στα 5 αρχεία άρα S=3/5 κλπ.

Πρέπει να δούμε ποια είναι συχνά στοιχειοσύνολα για κάθε Ν-στοιχειοσύνολο που παίρνουμε δηλαδή πάμε πρώτα στα 1-στοιχειοσύνολα, βλέπουμε ποια είναι συχνά και με βάση αυτά φτιάχνουμε τα 2-στοιχειοσύνολα και ούτω καθ’ εξής.

Συχνό λέγεται ένα στοιχειοσύνολο όταν η συχνότητα εμφάνισης S είναι μεγαλύτερη από το min sup που μα δίνεται στην εκφώνηση (σε αυτή την περίπτωση s=40%).

**Υποψήφια 1-στοιχειοσύνολα**

**S**

{beer} 4/5

{eggs} 4/5

{diapers} 3/5

{bread} 1/5

{cheese} 2/5

{milk}2/5

Συχνά είναι όλα τα στοιχειοσύνολα εκτός του {bread}.

Για να δημιουργήσω τα 2-στοιχειοσύνολα βλέπω τα συχνά 1-στοιχειοσύνολα και κάνω το εξής:

{beer} + {eggs} -> {beer, eggs}

Οπότε:

**Υποψήφια 2-στοιχειοσύνολα**

**S**

{beer, eggs} **3/5**

{beer, diapers} **2/5**

{beer, cheese} **2/5**

{beer, milk} **2/5**

{eggs, diapers} **2/5**

{eggs, cheese} **2/5**

{eggs, milk} 1/5

{diapers, cheese} 1/5

{diapers, milk} 0/5

{cheese, milk} 0/5

Αυτά που είναι έντονα είναι συχνά.

Έπειτα δημιουργώ τα υποψήφια 3-στοιχειοσύνολα, στα οποία δεν θα συμπεριλάβω κάποιο το οποίο συνδυάζει οποιοδήποτε ΜΗ-συχνό 2-στοιχειοσύνολο.

Επομένως:

{beer, eggs} + {beer, diapers} -> {beer,eggs,diapers} **ΥΠ.**

{beer, eggs} + {beer, cheese} -> {beer,eggs,cheese} **ΥΠ.**

{beer, eggs} + {beer, milk} -> {beer,eggs,milk} **ΟΧΙ**

{beer, eggs} + {eggs, diapers} -> {beer,eggs,diapers} **ΥΠ.**

{beer, eggs} + {eggs, cheese} -> {beer,eggs,cheese} **ΥΠ.**

{beer, diapers} + {beer, cheese} -> {beer,diapers,cheese} **ΟΧΙ**

{beer, diapers} + {beer, milk} -> {beer,diapers,milk} **ΟΧΙ**

{beer, diapers} + {eggs, diapers} -> {beer,diapers,eggs} **ΥΠ.**

{beer, diapers} + {eggs, cheese} -> {beer,diapers,eggs,cheese} **ΟΧΙ**

{beer, cheese} + {beer, milk} -> {beer,cheese,milk} **ΟΧΙ**

{beer, cheese} + {eggs, diapers} -> {beer,cheese,eggs,diapers} **ΟΧΙ**

{beer, cheese} + {eggs, cheese} -> {beer,cheese,eggs} **ΥΠ.**

{beer, milk} + {eggs, diapers} -> {beer,milk,eggs,diapers} **ΟΧΙ**

{beer, milk} + {eggs, cheese} -> {beer,milk,eggs,cheese} **ΟΧΙ**

{eggs, diapers} + {eggs, cheese} -> {eggs,diapers,cheese} **ΟΧΙ**

**Υποψήφια 3-στοιχειοσύνολα**

**S**

{beer, eggs, diapers} 1/5

{beer, eggs, cheese} **2/5** **ΣΥΧΝΟ**

Άρα δεν υπάρχουν συχνά 4-στοιχειοσύνολα(επειδή υπάρχει μόνο ένα συχνό 3-στοιχειοσύνολο).

**Λίστα συχνών στοιχειοσυνόλων**

{beer} {beer, eggs} {eggs, cheese}

{eggs} {beer, diapers} {beer, eggs, cheese}

{diapers} {beer, cheese}

{cheese} {beer, milk}

{milk} {eggs, diapers}

Ξεκινάμε από τα συχνά 2-στοιχειοσύνολα και βρίσκουμε το C=confidence που υπολογίζεται ως εξής:

C = (Το S του 2-στοιχειοσυνόλου που τα περιλαμβάνει) / (Το S του αριστερού μέλους)

Συχνός λέγεται ένας κανόνας όταν το κλάσμα είναι μεγαλύτερο από το C που δίνεται στην εκφώνηση(σε αυτή την περίπτωση C=70%).

Επομένως υπολογίζουμε(με έντονα οι συχνοί κανόνες):

beer -> eggs: C = (3/5) / (4/5) = 3/4 = **0.75**

eggs -> beer: C = (3/5) / (4/5) = 3/4 = **0.75**

beer -> diapers: C = (2/5) / (4/5) = 2/4 = 0.5

diapers -> beer: C = (2/5) / (3/5) = 2/3 = 0.66

beer -> cheese: C = (2/5) / (4/5) = 2/4 = 0.5

cheese -> beer: C = (2/5) / (2/5) = 2/2 = **1**

beer -> milk: C = (2/5) / (4/5) = 2/4 = 0.5

milk -> beer: C = (2/5) / (2/5) = 2/2 = **1**

eggs -> cheese: C = (2/5) / (4/5) = 2/4 = 0.5

cheese -> eggs: C = (2/5) / (2/5) = 2/2 = **1**

{beer, eggs} -> cheese: C = (2/5) / (3/5) = 2/3 = 0.66

cheese -> {beer, eggs}: C = (2/5) / (2/5) = 2/2 = **1**

{beer, cheese} -> eggs: C = (2/5) / (2/5) = 2/2 = **1**

eggs -> {beer, cheese}: C = (2/5) / (4/5) = 2/4 = 0.5

{eggs, cheese} -> beer: C = (2/5) / (2/5) = 2/2 = **1**

beer -> {eggs, cheese}: C = (2/5) / (4/5) = 2/4 = 0.5

**Λίστα κανόνων συσχέτισης ταξινομημένοι ως προς την εμπιστοσύνη**

cheese -> beer: C = 2/2 = **1**

milk -> beer: C = 2/2 = **1**

cheese -> eggs: C = 2/2 = **1**

cheese -> {beer, eggs}: C = 2/2 = **1**

{beer, cheese} -> eggs: C = 2/2 = **1**

{eggs, cheese} -> beer: C = 2/2 = **1**

beer -> eggs: C = 3/4 = **0.75**

eggs -> beer: C = 3/4 = **0.75**

diapers -> beer: C = 2/3 = 0.66

{beer, eggs} -> cheese: C = 2/3 = 0.66

beer -> diapers: C = 2/4 = 0.5

beer -> cheese: C = 2/4 = 0.5

beer -> milk: C = 2/4 = 0.5

eggs -> cheese: C = 2/4 = 0.5

eggs -> {beer, cheese}: C = 2/4 = 0.5

beer -> {eggs, cheese}: C = 2/4 = 0.5

Τελικοί κανόνες είναι οι συχνοί και το lift υπολογίζεται ως εξής: Lift = C / (Το S του δεξιού μέλους)

cheese -> beer: Lift = 1 / (4/5) = 1 / 0.8 = 1.25

milk -> beer: Lift = 1 / (4/5) = 1 / 0.8 = 1.25

cheese -> eggs: Lift = 1 / (4/5) = 1 / 0.8 = 1.25

cheese -> {beer, eggs}: Lift = 1 / (3/5) = 1 / 0.6 = 1.66

{beer, cheese} -> eggs: Lift = 1 / (4/5) = 1 / 0.8 = 1.25

{eggs, cheese} -> beer: Lift = 1 / (4/5) = 1 / 0.8 = 1.25

beer -> eggs: Lift = 0.75 / (4/5) = 0.75 / 0.8 = 0.93

eggs -> beer: Lift = 0.75 / (4/5) = 0.75 / 0.8 = 0.93

Παρατηρούμε ότι ο ισχυρότερος κανόνας συσχέτισης είναι ο **cheese -> {beer, eggs}**, διότι όχι μόνο έχει από τα μεγαλύτερα Confidence(C=1) αλλά και το Lift του είναι το μεγαλύτερο (Lift=1.66).

1. ***Apriori με το Weka***

Οι 4 από τους 5 κανόνες του ερωτήματος 3 περιέχονται στους κανόνες του ερωτήματος 4. Οι κανόνες του ερωτήματος 4 είναι περισσότεροι (10 έναντι 5). Ακόμη ο ισχυρότερος κανόνας του ερωτήματος 3 είναι ο τρίτος ισχυρότερος κανόνας του ερωτήματος 4. Επομένως, η εξόρυξη κανόνων συσχέτισης με τον Apriori χρησιμοποιώντας την μετρική Lift φαίνεται πως αποδίδει ισχυρότερους, λεπτομερέστερους και ποσοτικά περισσότερους κανόνες σε σχέση με τον Apriori χρησιμοποιώντας την μετρική Confidence.

Τα αποτελέσματα του ερωτήματος 6 μας δείχνουν ότι οι κανόνες μπορεί να φαίνονται οπτικά απλούστεροι, σε σχέση με τα ερωτήματα 3 και 4, αλλά ο βαθμός εμπιστοσύνης μεταξύ των κανόνων είναι αρκετά κοντά (0.86-1). Επομένως αυτοί οι κανόνες θα μπορούσαν να χρησιμοποιηθούν για την κατηγοριοποίηση των ψηφοφόρων ώστε να επιτευχθεί αρκετά μεγάλη ακρίβεια και ορθότητα.