ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ
ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# Ανάλυση Καλαθιού Αγορών με ευφυείς τεχνικές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ
ΚΑΡΟΛΟΥ ΝΙΚΗΤΑΡΑ

**Επιβλέποντες**: Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π
Γεώργιος Σιόλας ΕΔΙΠ Ε.Μ.Π

Αθήνα, Ιούλιος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ
ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# Ανάλυση Καλαθιού Αγορών με ευφυείς τεχνικές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΑΡΟΛΟΥ ΝΙΚΗΤΑΡΑ

| | | |
|---|---|---|
| **Επιβλέποντες**: | Ανδρέας-Γεώργιος Σταφυλοπάτης | Καθηγητής Ε.Μ.Π |
| | Γεώργιος Σιόλας | ΕΔΙΠ Ε.Μ.Π |

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Ιουλίου 2019.

| *(Υπογραφή)* | *(Υπογραφή)* | *(Υπογραφή)* |
|---|---|---|
| ..... | ..... | ..... |
| Ανδρέας-Γεώργιος Σταφυλοπάτης | Γιώργος Στάμου | Παναγιώτης Τσανάκας |
| Καθηγητής Ε.Μ.Π. | Αναπληρωτής Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. |

Αθήνα, Ιούλιος 2019

*(Υπογραφή)*

.....

**ΚΑΡΟΛΟΣ ΝΙΚΗΤΑΡΑΣ**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Περίληψη

Ο σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη των σύγχρονων μεθόδων που χρησιμοποιούνται στον τομέα της Ανάλυσης Καλαθιού Αγορών.Σε αυτό το πλαίσιο ασχοληθήκαμε με συνολοθεωρητικές και γραφοθεωρητικές τεχνικές.

Στην πρώτη κατηγορία, υλοποιήσαμε αλγορίθμους στην περιοχή της Εξόρυξης Συχνών Στοιχειοσυνόλων και Κανόνων Συσχέτισης και μελετήσαμε τη διαδικασία αξιολόγησης των τελευταίων.

Στη δεύτερη κατηγορία τεχνικών, υλοποιήσαμε το Δίκτυο Κανόνων Συσχέτισης που χρησιμοποιεί τους κανόνες της προηγούμενης κατηγορίας και εφαρμόσαμε τη μέθοδο Ανίχνευσης Κοινοτήτων μετά την κατασκευή ενός γράφου προϊόντων.

Τέλος, προχωρήσαμε στην Τμηματοποίηση Πελατών βάσει των αγοραστικών τους συνηθειών, αξιοποιώντας τις Κοινότητες Προϊόντων της προηγούμενης μεθόδου.

**λέξεις κλειδιά: ανάλυση καλαθιού αγορών, συχνά στοιχειοσύνολα, κανόνες συσχέτισης, δίκτυο κανόνων συσχέτισης, ανίχνευση κοινοτήτων, τμηματοποίηση πελατών**

# Abstract

The main aim of this thesis was to study modern techniques that are being used in the field of Market Basket Analysis. In this context, we applied both set-theoretic and graph-theoretic methods.

In the first case, we implemented algorithms in the area of Frequent Itemsets and Association Rules Mining and studied the process of rules ranking based on objective measures.

In the second, we applied Association Rules Network that is built from the rules mined by the previous method and applied Community Detection on a product network.

Finally, we performed Customer Segmentation based on their purchase behavior, reling on the Product Communities that were previously extracted.

**keywords: market basket analysis, frequent itemsets, association rules, association rules network community detection, customer segmentation**

# Ευχαριστίες

Ευχαριστώ τους επιβλέποντες μου, τον κύριο Γεώργιο Σιόλα και τον κύριο Ανδρέα-Γεώργιο Σταφυλοπάτη για την ουσιαστική κατεύθυνση, αλλά και την ελευθερία επιλογών που μου έδωσαν στο πλαίσιο της διπλωματικής μου. Ωστόσο, δεν θα μπορούσα να μην αναφερθώ στα σύγχρονα και γεμάτα γνώση εργαστήρια των μαθημάτων τους, όπου μας δίδαξαν πολλά όντας πάντα ευδιάθετοι. Επίσης, ευχαριστώ τους γονείς μου καθώς ο τρόπος που με μεγάλωσαν αποτελεί τη μεγαλύτερη κληρονομιά που μπορεί να παραδοθεί σε ένα παιδί. Αναμφισβήτητα τους χρωστάω αυτό που είμαι σήμερα. Τέλος, είμαι τυχερός που γνώρισα την Αναστασία και την ευχαριστώ για όλα.

# Πίνακας Περιεχομένων

# Ευρετήριο Εικόνων

# 1 Εισαγωγή

## 1.1 Ανάλυση Καλαθιού Αγορών

Η ανάγκη της εποπτείας της πώλησης των προϊόντων αναγνωρίστηκε νωρίς απ' τις επιχειρήσεις και ήδη το 1974 στο supermarket Marsh, στο Τρόυ του Οχάιο, πραγματοποιήθηκε, χάρη στην ανάπτυξη της τεχνολογίας ραβδωτού κώδικα (barcode), η πρώτη μαζική καταγραφή αγορών μέσω της σάρωσης ετικετών. Αξίζει να σημειωθεί ότι οι σχετικές πληροφορίες καταγράφονταν σε χαρτί καθώς και ότι το επίτευγμα αυτό σήμανε την αρχή της διαχείρισης των προϊόντων σαν ομάδες παρόμοιων ή σχετικών προϊόντων (**category management**).

Ωστόσο, ακόμα δεν υπήρχε η πληροφορία του ποιος αγόρασε το προϊόν. Έτσι, το 1995 στην αλυσίδα supermarket Tesco, στο Λονδίνο, έκανε την εμφάνισή της η πρώτη προσωποποιημένη μαγνητική κάρτα. Μέσω της κάρτας Clubcard, καταγράφονταν δεδομένα που αφορούσαν τόσο τον καταναλωτή όσο και τις αγορές του και ήταν διαθέσιμα σε ηλεκτρονική μορφή. Πλέον, γνωρίζοντας τις ανάγκες και τα ενδιαφέροντα των καταναλωτών, το μέχρι τότε μαζικό marketing (**mass marketing**) πήρε τη μορφή του προσωποποιημένου marketing (**individual marketing**). Ενδεικτικά, πριν την κυκλοφορία της κάρτας αυτής, το Tesco ήταν το τρίτο supermarket σε market value, με τη μισή αξία έναντι του πρώτου Sainsburys. Μέσα σε έναν χρόνο το Tesco έφτασε την πρώτη θέση, με εξαπλάσια αξία σε σχέση με το δεύτερο Sainsburys. Δεν είναι τυχαίο πως παρόμοια εξέλιξη είχε και η αλυσίδα supermarket Kroger στις Ηνωμένες Πολιτείες.

Τα παραπάνω επιτεύχθηκαν χάρις στην εξέλιξη της τεχνολογίας, όπου πλέον είχε καταστεί δυνατή η ψηφιακή αποθήκευση τεράστιων όγκων δεδομένων, κάτι που με τη σειρά του ανέδειξε περαιτέρω την ανάγκη εύρεσης αλγορίθμων που θα αξιοποιήσουν τα δεδομένα αυτά. Τυπικά, το πρόβλημα της Ανάλυσης Καλαθιού Αγορών μπορεί να γίνει κατανοητό ως εξής: απ' τη μία έχουμε τα καλάθια (**baskets**) ή συναλλαγές, απ' την άλλη τα στοιχεία (**items**) και κάθε καλάθι αποτελεί ένα σύνολο στοιχείων (**itemset**). Η Ανάλυση Καλαθιού Αγορών (**Market Basket Analysis**) αποτελεί ένα σύνολο τεχνικών που εφαρμόζονται πάνω σε δεδομένα συναλλαγών με σκοπό την εξαγωγή γνώσης, ικανής να μετατραπεί σε ενέργειες με γνώμονα την καλύτερη λειτουργία μιας επιχείρησης.

Οι εν λόγω τεχνικές είναι προσανατολισμένες είτε στο προϊόν είτε στον καταναλωτή. Παραδείγματα της πρώτης περιπτώσης είναι η Εξόρυξη Συχνών Στοιχειοσυνόλων (**Frequent Itemset Mining**), η Εξόρυξη Κανόνων Συσχέτισης (**Association Rules Mining ή Affinity analysis**), το Δίκτυο Κανόνων Συσχέτισης (**Association Rules Network**), όπως επίσης και η Ανίχνευση Κοινοτήτων (**Community Detection**). Στη δεύτερη περίπτωση, έχουμε την Τμηματοποίηση Καταναλωτών (**Customer Segmentation**) με βάση τις αγοραστικές τους συνήθειες.

Η τεχνική **FIM** αναζητά σύνολα στοιχείων που εμφανίζονται “συχνά” μέσα στα καλάθια, η **ARM** κατασκευάζει σχέσεις μεταξύ συνόλων, τους κανόνες, ενώ η **ARN** δημιουργεί ένα δίκτυο τέτοιων κανόνων που φανερώνει τις άμεσες και έμμεσες συσχετίσεις ενός δεδομένου στοιχείου. Επίσης, η **Ανίχνευση Κοινοτήτων** αφορά στον χωρισμό των προϊόντων σε ομάδες, ώστε τα προϊόντα κάθε ομάδας να είναι περισσότερο όμοια - κατά κάποια έννοια - μεταξύ τους, παρά σε σχέση με αυτά των άλλων ομάδων. Η **Τμηματοποίηση Καταναλωτών** ομαδοποιεί τους καταναλωτές με βάση κοινές αγοραστικές συνήθειες.

Μέσω της εφαρμογής των παραπάνω τεχνικών, οι επιχειρήσεις προβαίνουν σε στρατηγικές marketing όπως ο σχεδιασμός καταλόγου, η διαμόρφωση του καταστήματος, ο σχεδιασμός εκπτώσεων και προσφορών, η αύξηση πωλήσεων (add-on, cross/up selling), καθώς επίσης και η προβλεπτική ανάλυση, όπως για παράδειγμα η πρόβλεψη των προϊόντων στην επόμενης αγορά ενός καταναλωτή με βάση το ιστορικό των αγορών του.

## 1.2 Αντικείμενο Διπλωματικής

Στη διπλωματική αυτή ασχοληθήκαμε με την Ανάλυση Καλαθιού Αγορών (Market Basket Analysis), μία από τις παλαιότερες ερευνητικές περιοχές της Εξόρυξης Δεδομένων (Data Mining). Συγκεκριμένα, μελετήσαμε και υλοποιήσαμε τόσο κάποιες συνολοθεωρητικές όσο και κάποιες γραφοθεωρητικές μεθόδους πάνω σε δεδομένα αγορών από supermarket.

Στην πρώτη κατηγορία, έχουμε τις τεχνικές **FIM** και **ARM** όπως προαναφέρθηκαν. Γύρω από την FIM υλοποιήσαμε τους αλγορίθμους της οικογένειας Apriori, δηλαδή τους Apriori, PCY, Multistage, Multihash καθώς και τον ευρετικό αλγόριθμο Toivonen. Γύρω από την ARM κατασκευάσαμε τους Κανόνες Συσχέτισης, μελετήσαμε τις ιδιότητες των ποσοτικών μεγεθών, ώστε να επιλέξουμε αυτά με τα οποία βαθμολογήσαμε τους κανόνες και τέλος, χρησιμοποιήσαμε την πολυκριτηριακή μέθοδο Pareto για την ανάδειξη των “καλύτερων” κανόνων.

Στη δεύτερη κατηγορία, έχουμε τις τεχνικές **ARN** και **Ανίχνευση Κοινοτήτων.** Για την πρώτη, αξιοποιήσαμε τους κανόνες που προέκυψαν από την **ARM** και κατασκευάσαμε το δίκτυο κανόνων. Για την **Ανίχνευση Κοινοτήτων**, δημιουργήσαμε έναν γράφο προϊόντων, εξαγάγαμε και αξιολογήσαμε τις κοινότητες προϊόντων. Τέλος, βασιζόμενοι στις κοινότητες αυτές προχωρήσαμε στην **Τμηματοποίηση Καταναλωτών** σε δύο κατευθύνσεις. Στην πρώτη ομαδοποιήσαμε τους καταναλωτές με βάση τα χρήματα που ξοδεύουν στις κατηγορίες προϊόντων, ενώ στην δεύτερη με βάση την προτίμησή τους στις κατηγορίες προϊόντων.

## 1.3 Οργάνωση Κειμένου

Η υπόλοιπη διπλωματική οργανώνεται ως εξής. Στο 2ο Κεφάλαιο παρουσιάζουμε παρόμοιες εργασίες. Στο 3ο Κεφάλαιο κάνουμε μία θεωρητική ανάλυση σε διάφορες τεχνικές που έχουν χρησιμοποιηθεί στην Ανάλυση Καλαθιού Αγορών. Στην πρώτη ενότητα έχουμε την Εξόρυξη Συχνών Στοιχειοσυνόλων και Κανόνων Συσχέτισης και το Δίκτυο Κανόνων Συσχέτισης. Στην τρίτη ενότητα αναλύουμε την Ανίχνευση Κοινοτήτων και στην τέταρτη την Αναζήτηση Κοινότητας. Στο 4ο Κεφάλαιο παρουσιάζουμε τα αποτελέσματα των τεχνικών που εκτελέσαμε, την υλοποίησή τους, τα αποτελέσματα που πήραμε καθώς και τα συμπεράσματα που εξάγαμε. Στο 5ο Κεφάλαιο ολοκληρώνουμε συνοψίζοντας την εργασία και προτείνοντας κατευθύνσεις για το μέλλον.

# 2 Θεωρητικό υπόβαθρο

## 2.1 Εξόρυξη Κανόνων Συσχέτισης

Η Εξόρυξη Κανόνων Συσχέτισης (Association Rules Mining) είναι μια μέθοδος μηχανικής μάθησης βασισμένη σε κανόνες (Rule-Based Machine Learning) με σκοπό την εύρεση ενδιαφέρουσων σχέσεων μεταξύ μεταβλητών σε μία μεγάλη βάση δεδομένων. Προτάθηκε (Agrawal et al. 1994) ως μία τεχνική Ανάλυσης Καλαθιού Αγορών (Market Basket Analysis) και για αυτό τον λόγο, συχνά, οι δύο αυτοί όροι συγχέονται.

### 2.1.1 Περιγραφή

Το μοντέλο Καλαθιού-Αγορών ορίζεται με δυο οντότητες, τα καλάθια (baskets) και τα στοιχεία (items). Κάθε καλάθι αποτελεί ένα σύνολο στοιχείων και οποιοδήποτε σύνολο στοιχείων αποτελεί ένα στοιχειοσύνολο (itemset). Τυπικά, ένα σύνολο κανόνων $Rules(T, s, c)$ της μορφής $A \rightarrow B,$ όπου $A,\ B$ δύο αυθαίρετα στοιχειοσύνολα με $A \cap B = \emptyset$, ορίζεται από μία βάση συναλλαγών T, ένα κατώφλι υποστήριξης (support threshold) s και ένα κατώφλι εμπιστοσύνης (confidence threshold) c. Έστω A (όμοια B) το σύνολο των συναλλαγών που περιέχουν το στοιχειοσύνολο $A$ ($B$), τότε για τους παραπάνω κανόνες ισχύουν:

1. $\frac{|A \cap B|}{|T|} \geq s$
2. $\frac{|A \cap B|}{|A|} \geq c$

όπου η πρώτη συνθήκη απαιτεί το στοιχειοσύνολο $A \cup B$ να βρίσκεται σε ικανοποιητικό ποσοστό επί του συνόλου των συναλλαγών, ενώ η δεύτερη σε ικανοποιητικό ποσοστό επί αυτών που περιέχουν το *A.*

Διάφοροι αλγόριθμοι έχουν προταθεί για την εύρεση Κανόνων Συσχέτισης οι οποίοι όμως φέρουν περιορισμούς. Μεγάλες βάσεις συναλλαγών οδηγούν σε τεράστια σύνολα κανόνων συσχέτισης - για λογικές τιμές support, confidence - πολλοί εκ των οποίων χαρακτηρίζονται προφανείς ή περιττοί. Έχουν αναπτυχθεί δύο κύριες κατευθύνσεις μεθόδων στην προσπάθεια ανάδειξης των πιο ενδιαφέρουσων κανόνων. Η πρώτη αφορά στον περιορισμό προφανών κανόνων, ενώ η δεύτερη στη μέτρηση της ποιότητάς τους και είναι αυτή με την οποία εμείς ασχοληθήκαμε.

Στη συνέχεια, μελετάμε την παραγωγή Κανόνων Συσχέτισης, μια διαδικασία δύο βημάτων. Το πρώτο βήμα είναι η Εξόρυξη Συχνών Στοιχειοσυνόλων και το δεύτερο είναι η εξαγωγή, από αυτά, των Κανόνων Συσχέτισης. Ως επόμενο στάδιο, ασχολούμαστε με τα ποσοτικά μεγέθη (interestingness measures) που αξιολογούν τους ευρεθέντες κανόνες,

μελετώντας τις ιδιότητες αυτών και επιλέγοντας ένα μέρος τους. Τέλος, προχωράμε στην πολυκριτηριακή μέθοδο Pareto με σκοπό την ανάδειξη των κανόνων-νικητών.

## 2.1.2 Ορισμοί

### 2.1.2.1 Συχνά Στοιχειοσύνολα

Το πρόβλημα των Συχνών Στοιχειοσυνόλων προτάθηκε ως εξής.

Έστω

$i$ ένα στοιχείο (δυαδική μεταβλητή)

$I = \{i_1, i_2, ..., i_m\}$ ένα σύνολο m στοιχείων

$D = \{t_1, t_2, ..., t_p\}$ ένα σύνολο καλαθιών - η βάση δεδομένων ή βάση συναλλαγών -

Τότε

- κάθε καλάθι - στοιχείο του $D$ - αποτελεί ένα υποσύνολο στοιχείων του $I$.
- κάθε υποσύνολο του $I$ αποτελεί ένα **στοιχειοσύνολο**. (3.1)

| ID | Στοιχεία |
|---|---|
| *1* | *ψωμί , γάλα* |
| *2* | *ψωμί , καφές , μπύρα* |
| *3* | *αυγά , ψωμί , καφές* |
| *4* | *μπύρα , πάνες , γάλα* |

| *ID* | *ψωμί* | *γάλα* | *καφές* | *μπύρα* | *αυγά* | *πάνες* |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 0 | 1 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 | 1 | 0 | 1 |

Εικόνα 3.1: Η βάση δεδομένων περιέχει 4 καλάθια $(p = 4)$ και συνολικά 6 διαφορετικά στοιχεία $(m = 6)$. Οι δύο πίνακες αποτελούν ισοδύναμες αναπαραστάσεις της βάσης συναλλαγών. Στον πίνακα στα δεξιά φαίνεται η χρήση των στοιχείων σαν δυαδικές μεταβλητές.

Έστω

**υποστήριξη** (support) ενός στοιχειοσυνόλου $\sigma(\cdot)$ είναι ο συνολικός αριθμός καλαθιών στα οποία αυτό περιέχεται (3.2)

***κατώφλι υποστήριξης*** $s \in \mathrm{N}^{+}$ (3.3)

Τότε

- ένα στοιχειοσύνολο είναι **συχνό** ανν έχει $\sigma(\cdot) \geq s$. (3.4)

Στην εικόνα 3.1 αν υποθέσουμε $s = 2,$ τότε όλα τα στοιχεία (μονοσύνολα) εκτός από τα αυγά και τις πάνες είναι συχνά. Επίσης, συχνό είναι και το στοιχειοσύνολο $\{\psi\omega\mu\acute{\iota}\,,\ \kappa\alpha\phi\acute{\varepsilon}\varsigma\}$ και δεν υπάρχει κανένα άλλο. Στην περίπτωση όπου $s = 3$, το μοναδικό συχνό στοιχειοσύνολο είναι το $\{\psi\omega\mu\acute{\iota}\}.$

### 2.1.2.2 Κανόνες Συσχέτισης

Έστω

$A, B \subseteq I$

$A \cap B = \emptyset$

Τότε

- οι **κανόνες συσχέτισης** είναι προτάσεις της μορφής $A \rightarrow B$. (3.5)

Έστω

**υποστήριξη** (support) ενός κανόνα $\sigma(A \cup B)$ είναι ο συνολικός αριθμός καλαθιών στα οποία περιέχεται το στοιχειοσύνολο $A \cup B.$ (3.6)

**εμπιστοσύνη** (confidence) ενός κανόνα $c(A \cup B) = \sigma(A \cup B) / \sigma(A)$ είναι ο λόγος του αριθμού των συνολικών καλαθιών στα οποία περιέχεται μαζί το $A \cup B,$ προς αυτού των συνολικών καλαθιών στα οποία περιέχεται το $A$ μόνο του. (3.7)

**κατώφλι εμπιστοσύνης** $c \in [0, 1]$ (3.8)

Τότε

- ψάχνουμε για κανόνες της μορφής $A \rightarrow B$, όπου :
  - $\sigma(A \cup B) \geq s$
  - $c(A \cup B) \geq c$

Στην εικόνα 3.1, ο κανόνας $\{\psi\omega\mu\acute{\iota}\,,\ \kappa\alpha\phi\acute{\varepsilon}\varsigma\} \Rightarrow \{\mu\pi\acute{\upsilon}\rho\alpha\}$ έχει υποστήριξη 1, καθώς το στοιχειοσύνολο $\{\psi\omega\mu\acute{\iota}\,,\ \kappa\alpha\phi\acute{\varepsilon}\varsigma\,,\ \mu\pi\acute{\upsilon}\rho\alpha\}$ συναντάται μόνο στο 2ο καλάθι και εμπιστοσύνη $1/2$ αφού το $\{\psi\omega\mu\acute{\iota}\,,\ \kappa\alpha\phi\acute{\varepsilon}\varsigma\}$ συναντάται τόσο στο 2ο όσο και στο 3ο.

Στη συνέχεια παρουσιάζονται τα μεγέθη που χρησιμοποιούνται παρακάτω, μαζί με την ερμηνεία και τον συμβολισμό τους.

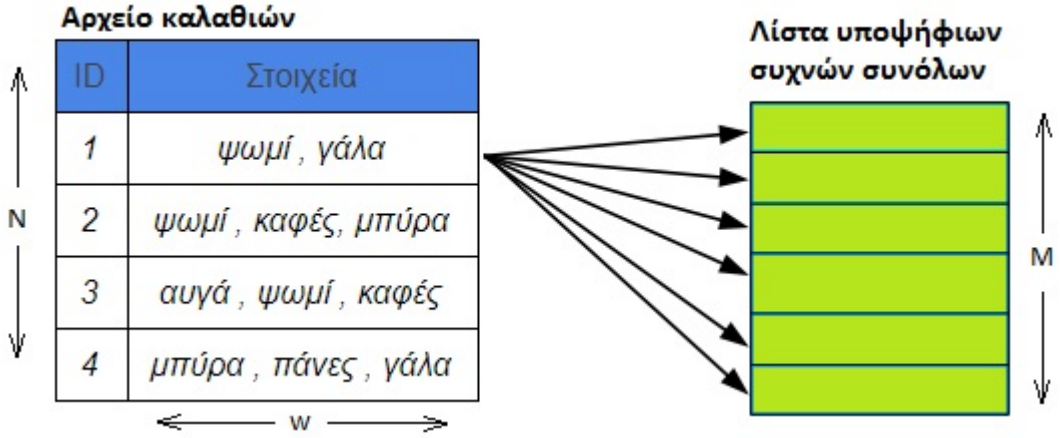
| ΜΕΓΕΘΟΣ \| ΣΥΜΒΟΛΟ | ΕΡΜΗΝΕΙΑ |
|---|---|
| στοιχειοσύνολο | σύνολο στοιχείων |
| k-στοιχειοσύνολο | στοιχειοσύνολο με k στοιχεία |
| υποστήριξη στοιχειοσυνόλου $I$, $\sigma(I)$ | συνολικός αριθμός καλαθιών<br>τα οποία περιέχουν το στοιχειοσύνολο $I$ |
| *κατώφλι υποστήριξης* $s$ | ελάχιστη τιμή υποστήριξης ενός στοιχειοσυνόλου<br>ώστε να θεωρηθεί “συχνό” |
| *συχνό στοιχειοσύνολο* $I$ | στοιχειοσύνολο $I$ *με υποστήριξη* $\sigma(I)$ >= *s*<br>*(θα λέμε ότι το* $I$ *ικανοποιεί το κατώφλι s)* |
| *υποψήφιο “συχνό” στοιχειοσύνολο* | στοιχειοσύνολο που ενδέχεται να ικανοποιεί το κατώφλι s<br>(χρειάζεται να μετρήσουμε τον αριθμό εμφάνισής του) |
| $C_k$ | σύνολο υποψηφίων “συχνών” k-στοιχειοσυνόλων |
| $L_k$ | σύνολο “συχνών” k-στοιχειοσυνόλων |
| *υποστήριξη κανόνα* $\sigma(X \Rightarrow Y)$ | υποστήριξη του στοιχειοσυνόλου $X \cup Y$ |
| *εμπιστοσύνη κανόνα* $c(X \Rightarrow Y)$ | $\frac{\text{υποστήριξη στοιχειοσυνόλου } X \cup Y}{\text{υποστήριξη στοιχειοσυνόλου } X}$ |
| *κατώφλι εμπιστοσύνης c* | ελάχιστη τιμή εμπιστοσύνης ενός κανόνα<br>ώστε να θεωρηθεί μέρος της λύσης |

Εικόνα 3.2: Ερμηνεία και συμβολισμός μεγεθών.

## 2.1.3 Αλγόριθμοι Εξόρυξης Συχνών Στοιχειοσυνόλων

### 2.1.3.1 Απλή προσέγγιση

Ας υποθέσουμε ότι έχουμε την βάση συναλλαγών που φαίνεται στην παρακάτω εικόνα και θέλουμε να βρούμε τα συχνά στοιχειοσύνολα. Μία ιδέα είναι, εξετάζοντας ένα προς ένα τα καλάθια, να παραγάγουμε και να καταμετρήσουμε όλα τα σύνολα στοιχείων που υπάρχουν σε κάθε καλάθι. Στο τέλος, θα έχουμε βρει τον αριθμό εμφάνισης όλων των στοιχειοσυνόλων που υπάρχουν στην βάση και όσα ικανοποιούν το κατώφλι υποστήριξης θα είναι τα “συχνά”.



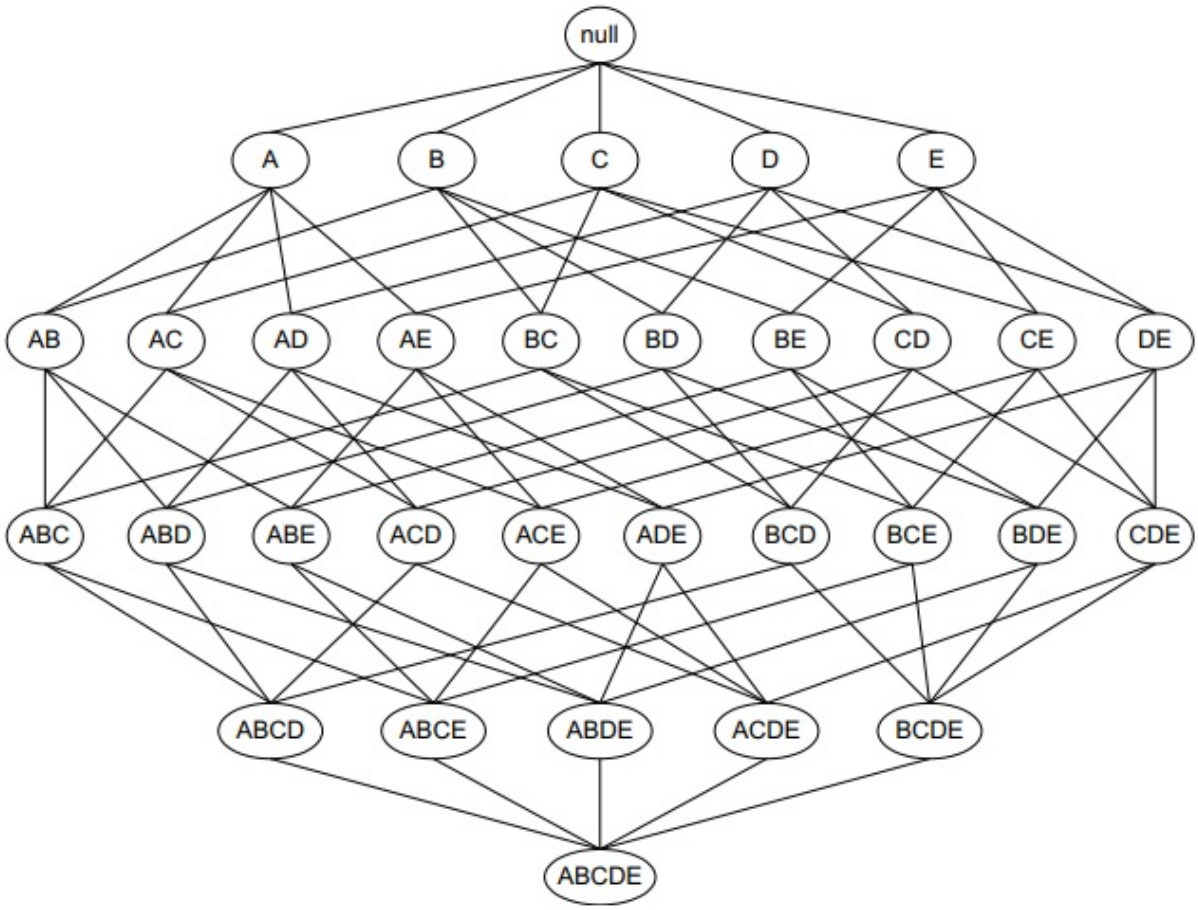

Εικόνα 3.3: *N* ο αριθμός των καλαθιών, w ο αριθμός των στοιχείων του μεγαλύτερου καλαθιού, m ο αριθμός των διαφορετικών στοιχείων στην βάση των συναλλαγών και M ο αριθμός όλων των υποσυνόλων που υπάρχουν τουλάχιστον σε ένα καλάθι.

Ας εξετάσουμε τις απαιτήσεις σε χώρο αποθήκευσης και χρόνο εκτέλεσης της παραπάνω μεθόδου. Γενικά, όλα τα σύνολα ενός καλαθιού μεγέθους w είναι $2^w$ κι εφόσον υπάρχουν N καλαθια, χρειαζόμαστε συνολικό χρόνο ίσο με $O(N * 2^w)$ για την παραγωγή τους και χώρο $\Theta(M)$ για την καταμέτρηση του αριθμού εμφάνισης τους. Στην χειρότερη περίπτωση, όπου ισχύουν $w = m$ και $M = 2^m$, οι προηγούμενες ποσότητες γίνονται $O(N * 2^m)$ και $O(2^m)$, αντίστοιχα. Ενδεικτικά, στην εικόνα 3.4 φαίνονται όλα τα δυνατά υποσύνολα πέντε στοιχείων.

Επιπλέον, ας διερευνήσουμε τον αριθμό των πιθανών κανόνων συσχέτισης. Ουσιαστικά, έχοντας ένα σύνολο $m$ στοιχείων, έστω $X$, θέλουμε να δούμε με πόσους τρόπους μπορούν να προκύψουν δυο μη κενά και ανεξάρτητα υποσύνολα $A$ και $B$, δηλαδή $A, B \subseteq X$, $A, B \neq \emptyset$ και $A \cap B = \emptyset$. Υπάρχουν $\binom{m}{i}$, με $1 \leq i \leq m-1$, σύνολα με $i$ στοιχεία (επιλογές για το $A$), καθένα εκ των οποίων “αφήνει” για το $B$, όλους τους τρόπους επιλογής ενός στοιχείου μέσα από $m - i$,

δηλαδή $\sum_{l=1}^{m-i} \binom{m-i}{l}$. Συνολικά, έχουμε $\sum_{i=1}^{m-1} [\binom{m}{i} * \sum_{l=1}^{m-i} \binom{m-i}{l}] = 3^k - 2^{k+1} + 1$ κανόνες της μορφής $A \rightarrow B$.

Είναι φανερό πως οι παραπάνω ποσότητες είναι μη διαχειρίσιμες. Όπως θα δούμε στη συνέχεια, όταν ένα στοιχειοσύνολο θεωρείται υποψήφιο συχνό, τότε κάθε φορά που το συναντάμε μέσα σε ένα καλάθι, πρέπει να “μετράμε” κάπου την πληροφορία αυτή. Ο αριθμός και κατ’ επέκταση η απαίτηση σε αποθηκευτικό χώρο, περιορίζεται αν θεωρούνται ως υποψήφια συχνά στοιχειοσύνολα, όσα πραγματικά μπορεί να είναι. Για παράδειγμα, στην παραπάνω περίπτωση το στοιχειοσύνολο {μπύρα, αυγά} δεν χρειάζεται να θεωρηθεί υποψήφιο αφού τα δυο αυτά προϊόντα δεν βρίσκονται μαζί σε κανένα καλάθι. Πέρα από την προφανή αυτή περίπτωση, υπάρχουν και άλλες κατά τις οποίες μπορούμε να αποφύγουμε τη δέσμευση χώρου για την καταμέτρηση ενός στοιχειοσυνόλου.Στη συνέχεια, θα χρησιμοποιήσουμε διάφορες ιδιότητες συνόλων στην κατεύθυνση αυτή.




Εικόνα 3.4: Όλα τα $2^5$ υποσύνολα του συνόλου 5 στοιχείων $I = \{A, B, C, D, E\}$.

Έχουν αναπτυχθεί διάφοροι αλγόριθμοι εύρεσης Κανόνων Συσχέτισης που διακρίνονται σε δύο κατηγορίες. Στην πρώτη ανήκουν οι αλγόριθμοι “οριζόντιας διαμοίρασης” και στη δεύτερη οι αλγόριθμοι “κατακόρυφης διαμοίρασης”. Η κύρια διαφορά, όπως φαίνεται στην εικόνα 3.5, έγκειται στην αναπαράσταση της βάσης συναλλαγών, αφού η πρώτη την “βλέπει” ως “καλάθια που περιέχουν στοιχεία”, ενώ η δεύτερη ως “στοιχεία που περιέχονται σε καλάθια”.

| ID | Στοιχεία |
|---|---|
| *1* | *ψωμί , γάλα* |
| *2* | *ψωμί , καφές, μπύρα* |
| *3* | *αυγά , ψωμί , καφές* |
| *4* | *μπύρα , πάνες , γάλα* |

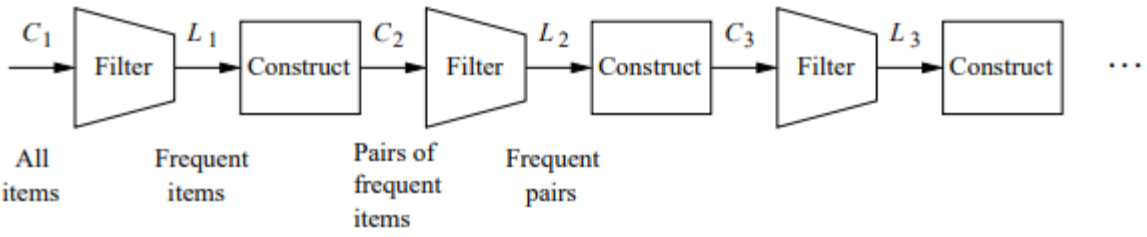
| Στοιχεία | ID |
|---|---|
| *ψωμί* | 1 , 2 , 3 |
| *γάλα* | 1 , 4 |
| *καφές* | 2 , 3 |
| *μπύρα* | 2 , 4 |
| *αυγά* | 3 |
| *πάνες* | 4 |

Εικόνα 3.5: Αναπαράσταση δεδομένων στην “οριζόντια” και “κατακόρυφη” διαμοίραση, αντίστοιχα.

## 2.1.3.2 Apriori οικογένεια αλγορίθμων

### 2.1.3.2.1 Κοινή Βάση

Όλοι οι αλγόριθμοι που ακολουθούν έχουν κάποια κοινά στοιχεία, τα οποία αναφέρονται εδώ. Η κατασκευή “συχνών” στοιχειοσυνόλων πραγματοποιείται, διαδοχικά, με σταθερό αριθμό (συνήθως 1) περασμάτων στην βάση συναλλαγών για κάθε μέγεθος στοιχειοσυνόλων. Συγκεκριμένα, πρώτα βρίσκονται τα “συχνά” στοιχεία (μονοσύνολα ή 1-στοιχειοσύνολα), μετά τα 2-στοιχειοσύνολα κ.ο.κ. όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 3.6: Τα επαναληπτικά βήματα του Apriori αλλά και των βελτιωμένων του εκδοχών. Σε κάθε βήμα ***k*** κατασκευάζεται το σύνολο των υποψηφίων “συχνών” k-στοιχειοσυνόλων $C_k$, καταμετράται ο αριθμός εμφάνισής τους και τελικά προκύπτει το πραγματικό σύνολο “συχνών” k-στοιχειοσυνόλων $L_k$.

Όπως θα δούμε, για την παραγωγή του $L_k$, σε κάθε πέρασμα $k$, απαιτείται αποθηκευτικός χώρος για τη διατήρηση των συχνών στοιχειοσυνόλων που βρέθηκαν στο πρώτο και στο αμέσως προηγούμενο βήμα, δηλαδή τα $L_1 \text{και} L_{k-1}$, αντίστοιχα. Επίσης, χρειαζόμαστε αποθηκευτικό χώρο για την καταμέτρηση του αριθμού εμφάνισης των υποψηφίων συχνών στοιχειοσυνόλων $C_k$. Είναι αναγκαίο, αυτά τα δεδομένα να χωράνε στην κύρια μνήμη και σ` αυτήν την κατεύθυνση, κάθε τεχνική προσπαθεί να μειώσει τον αριθμό των υποψηφίων συχνών στοιχειοσυνόλων $C_k$ με κάποιο κόστος. Το κόστος αυτό είναι είτε χρονικό, δηλαδή κάποιο επιπλέον πέρασμα στο αρχείο καλαθιών, είτε χωρικό, δηλαδή σπατάλη χώρου για χρήση κάποιας δομής δεδομένων.

## 2.1.3.2.1 Apriori

### Κεντρική Ιδέα

Ο δημοφιλής αυτός αλγόριθμος στηρίζεται στην παρακάτω ιδέα.

- $\forall X, Y : X \subseteq Y \Rightarrow \sigma(Y) \leq \sigma(X)$

δηλαδή κάθε υποσύνολο ενός στοιχειοσυνόλου, συναντάται σε ίσο ή μεγαλύτερο αριθμό καλαθιών. Αυτό καθιστά την υποστήριξη (support) $\sigma(\cdot)$**,** μια αντι-μονοτονική συνάρτηση.

Έτσι προκύπτουν οι παρακάτω ιδιότητες.

Ιδιότητα 3.1:

- $\forall X, Y : \sigma(X) < s$, $X \subseteq Y \Rightarrow \sigma(Y) < s$, δηλαδή οποιοδήποτε υπερσύνολο ενός μη “συχνού” στοιχειοσυνόλου, είναι επίσης μη “συχνό”.

Επεξήγηση

Έστω ότι γνωρίζουμε ότι ο αριθμός εμφάνισης του στοιχειοσυνόλου $\{bread, butter, milk\}$ δεν υπερβαίνει το κατώφλι $s$. Αν προσθέσουμε οποιοδήποτε στοιχείο, τότε μέρος των καλαθιών που περιέχουν το $\{bread, butter, milk\}$, μπορεί να μην περιέχουν το στοιχείο “coffee” και επίσης, δεν μπορεί να προκύψει κάποιο άλλο καλάθι που να περιέχει το δεύτερο και και όχι το πρώτο σύνολο. Άρα το $\{bread, butter, milk, coffee\}$ δεν μπορεί παρά να είναι επίσης μη “συχνό”.

Αποτέλεσμα της ιδιότητας 3.1 είναι η αποφυγή σπατάλης χώρου. Συγκεκριμένα, έστω $X$ υποψήφιο “συχνό” k-στοιχειοσύνολο. Όπως για κάθε μέλος του $C_k$, έτσι και για το $X$ θα καταμετρήσουμε τον αριθμό εμφάνισής του. Έστω, επίσης, ότι το $X$ δεν ικανοποιεί το κατώφλι υποστήριξης s και άρα προκύπτει ότι είναι ένα μη ”συχνό” στοιχειοσύνολο. Αν και για το $X$ σπαταλήθηκε χώρος, αυτό δεν θα συμβεί για κανένα υπερσύνολό του.

Ιδιότητα 3.2:

- $\forall X, Y : \sigma(Y) \geq s$, $X \subseteq Y \Rightarrow \sigma(X) \geq s$, δηλαδή οποιοδήποτε υποσύνολο ενός συχνού στοιχειοσυνόλου, είναι επίσης συχνό.

Επεξήγηση

Όμοια, έστω ότι ο αριθμός εμφάνισης του στοιχειοσυνόλου $\{bread, butter, milk\}$ υπερβαίνει το κατώφλι $s$, τότε για καθένα από τα 7 (εξαιρώντας το κενό) υποσύνολά του, όλα τα προηγούμενα καλάθια συνεχίζουν να περιέχουν το εκάστοτε υποσύνολο και επίσης, μπορεί να προκύψουν επιπλέον καλάθια που να το περιέχουν.

Έστω

ένα στοιχειοσύνολο $X$

Τότε

- Κάθε σύνολο $\{S \mid S \subset X, \ |S| = |X| - 1\}$ ονομάζεται άμεσο υποσύνολο (immediate subset) του $X$ (3.9)
- Αντίστοιχα, κάθε σύνολο $\{S \mid S \supset X, \ |S| = |X| + 1\}$ ονομάζεται άμεσο υπερσύνολο (immediate subset) του $X$ (3.10)

Ιδιότητα 3.3:

- Κάθε σύνολο με πληθικό αριθμό k, έχει ακριβώς k άμεσα υποσύνολα.

Με χρήση της ιδιότητας 3.2, για να θεωρηθεί ένα k-στοιχειοσύνολο ως υποψήφιο “συχνό” (candidate, μέλος του $C_k$), πρέπει και τα k k-1-στοιχειοσύνολά του να είναι επίσης “συχνά” (μέλη του $L_{k-1}$). Η συνθήκη αυτή είναι αναγκαία αλλά όχι ικανή, κάτι που φαίνεται στο παρακάτω παράδειγμα.

| ID | Στοιχεία |
|---|---|
| *1* | *ψωμί , γάλα* |
| *2* | *ψωμί , τυρί* |
| *3* | *καφές , κέικ* |
| *4* | *καφές , νερό* |

Εικόνα 3.7: Έχοντας κατώφλι υποστήριξης $s = 2$ , βλέπουμε πως ξεχωριστά, τα στοιχεία ψωμί και καφές είναι “συχνά” αφού περιέχονται σε 2 καλάθια το καθένα, στα καλάθια 1,2 και 3,4 αντίστοιχα, αλλά αν τα εξετάσουμε μαζί, το 2-στοιχειοσύνολο {ψωμί, καφές} δεν είναι συχνό.

Συνδυάζοντας τις ιδιότητες 3.2 και 3.3 προκύπτει η παρακάτω ιδιότητα.

Ιδιότητα 3.4:

- Κάθε στοιχείο ενός συχνού k-στοιχειοσυνόλου συμμετέχει σε τουλάχιστον k συχνά k-1-στοιχειοσύνολα.

Χάρη στην τελευταία ιδιότητα, όπως θα δούμε, μπορεί να ελαχιστοποιήσει ο αριθμός των στοιχείων που δύναται να συμμετέχουν σε συχνά στοιχειοσύνολα. Συγκεκριμένα, αν υποθέσουμε ότι ένα καλάθι περιέχει $n$ συχνά στοιχεία και βρισκόμαστε στο βήμα k, τότε αντί να ελέγξουμε και τα $\binom{n}{k}$ δυνατά υποψήφια “συχνά” στοιχειοσύνολα, χρειάζεται να ελέγξουμε μόνο τα $\binom{n-l}{k}$, όπου l ο αριθμός των στοιχείων που συμμετέχουν συνολικά σε λιγότερα από k συχνά k-1-στοιχειοσύνολα. Ωστόσο, είναι φανερό πως, σε κάθε πέρασμα, χρειάζεται επιπλέον αποθηκευτικός χώρος $O(M)$, όπου $M$ τα συνολικά στοιχεία που περιλαμβάνει η βάση συναλλαγών, όπως έχουμε ήδη ορίσει.

## Υλοποίηση

Η διαδικασία εύρεσης “συχνών” 1-στοιχειοσυνόλων είναι τετριμμένη. Όλα τα στοιχεία που υπάρχουν στα καλάθια μας θεωρούνται υποψήφια “συχνά” και αποτελούν το $C_1$ . Στη συνέχεια, κοιτώντας ένα-ένα κάθε καλάθι μετράμε τον συνολικό αριθμό εμφάνισης κάθε στοιχείου και τελικά, ορίζουμε σαν $L_1$ τα στοιχεία που ικανοποιούν το κριτήριο ελάχιστης υποστήριξης $s$ .

Για την παραγωγή του $L_k$ από το $C_k$ για $k \geq 2$ επαναλαμβάνονται τα παρακάτω βήματα:

1.Συνένωση - υπολογισμός του $C_k$

Έστω $l1, l2$ δύο “συχνά” k-1-στοιχειοσύνολα, μέλη του $L_{k-1}$.
Ας υποθέσουμε ότι υπάρχουν ακριβώς $k-2$ κοινά στοιχεία μεταξύ τους (δηλαδή υπάρχει ακριβώς ένα στοιχείο του $l1$ που δεν ανήκει στο $l2$ και αντίστροφα).
Τότε η συνένωση $l1 \bowtie l2$ δίνει σαν αποτέλεσμα ένα k-στοιχειοσύνολο αποτελούμενο από τα $k-2$ κοινά τους στοιχεία μαζί με τα δύο μη κοινά.
Αν εφαρμόσουμε την πράξη αυτή για κάθε δυνατό συνδυασμό $l1, l2$ τότε παίρνουμε σαν αποτέλεσμα το $C_k$, δηλαδή το σύνολο των υποψηφίων “συχνών” στοιχειοσυνόλων.

2.Κλάδεμα - υπολογισμός του $L_k$

Διατρέχοντας ένα-ένα τα καλάθια, μετράμε τον αριθμό εμφάνισης των στοιχείων του $C_k$. Όσα τελικά, “συναντήσουμε” σε τουλάχιστον $s$ καλάθια θα συντελέσουν το $L_k$.

### *Παραλλαγή*

Εδώ παρουσιάζεται μία παραλλαγή του αλγορίθμου Apriori στην οποία στηρίζονται όλοι οι ακολουθούμενοι αλγόριθμοι που αποτελούν επεκτάσεις-βελτιώσεις αυτού.

Αν παρατηρήσουμε την παραπάνω υλοποίηση, θα δούμε πως υπολογίζοντας το $C_k$ μέσω του $L_{k-1} \bowtie L_{k-1}$ (όπως αναφέρεται στο βήμα της Συνένωσης παραπάνω), ενδέχεται να προκύψουν k-στοιχειοσύνολα που δεν υπάρχουν σε κανένα καλάθι και παρόλα αυτά θεωρούνται υποψήφια “συχνά”. Το πρόβλημα αφορά στη σπατάλη χώρου για την καταμέτρηση του αριθμού εμφάνισής τους και για αυτόν τον λόγο, προχωράμε σε ένα παρόμοιο τρόπο υπολογισμού του $C_k$.

Παρακάτω παρουσιάζεται ο ψευδοκώδικας της παραλλαγής του Apriori. Αξίζει να σημειωθεί πως το $C_k$ δεν κατασκευάζεται ρητά, αντίθετα, χρησιμοποιώντας την αντι-μονοτονική ιδιότητα 3.2, κάνουμε τους απαραίτητους ελέγχους συμμετοχής στο $L_{k-1}$ όπως φαίνεται παρακάτω.

```
Βήμα 1
   ∀ καλάθι
      ∀ στοιχείο
         Μέτρα την εμφάνιση του στοιχείου αυτού (μέλος C_1)

   ∀ στοιχείο του C_1
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_1)


Βήμα k (αρχικά k=2)
   ∀ καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού
      ∀ k-στοιχειοσύνολο των παραπάνω στοιχείων
         αν όλα τα δυνατά k-1-στοιχειοσύνολα είναι στοιχεία του L_{k-1}
         τότε μέτρα την εμφάνισή του k-στοιχειοσυνόλου αυτού (μέλος C_k)

   ∀ στοιχείο του C_k
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι "συχνό" (μέλος L_k)

   Αν το L_k είναι το κενό σύνολο
   τότε Τελος
   αλλιώς   α) k = k + 1
            β) πήγαινε στο Βήμα k
```

Εικόνα 3.8: Ο αλγόριθμος Apriori σε ψευδοκώδικα.

## Παρατηρήσεις

### *Χρονικη πολυπλοκότητα*

Το κύριο κόστος έγκειται στην προσπέλαση των καλαθιών από τον δίσκο. Αν *π.χ.* για δεδομένη τιμή του κατωφλίου υποστήριξης $s$, προκύπτουν μέχρι και “συχνά” k-στοιχειοσύνολα (δηλαδή κανένα “συχνό” σύνολο με τουλάχιστον $k+1$ στοιχεία), τότε θα έχουμε διαβάσει ολόκληρο το αρχείο από τον δίσκο $k$ φορές. Έτσι, ο συνολικός χρόνος εκτέλεσης θα είναι $\frac{k * \text{μέγεθος αρχείου καλαθιών}}{\text{ταχύτητα προσπέλασης του δίσκου}}$.

Ένα άλλο κρίσιμο σημείο, είναι η παραγωγή των υποψήφιων “συχνών” στοιχειοσυνόλων καθώς εξετάζουμε κάθε καλάθι. Αν υποθέσουμε ότι βρισκόμαστε στο βήμα $k$, όπου $k \geq 2$, και το δεδομένο καλάθι περιέχει n “συχνά” στοιχεία, τότε υπάρχουν $\binom{n}{k}$ πιθανά υποψήφια “συχνά” k-στοιχειοσύνολα. Πιθανά διότι, όπως είπαμε, για να θεωρηθεί ένα k-στοιχειοσύνολο ως υποψήφιο “συχνό” και άρα να καταμετρηθεί ο αριθμός εμφάνισης του, πρέπει και τα k k-1-στοιχειοσύνολα του να είναι “συχνά”. Η παραγωγή των $\binom{n}{k}$ αυτών συνόλων προσθέτει την ανάλογη χρονική επιβάρυνση.

Είναι εύκολο να παρατηρήσει κανείς ότι για μεγάλες τιμές του k, τα $\binom{n}{k}$ πιθανά υποψήφια ζευγάρια είναι $\Theta\left(\frac{n^k}{k!}\right)$ και έτσι, ο χρόνος παραγωγής αυτών των συνδυασμών υπερβαίνει τον χρόνο προσπελάσεων του δίσκου. Όμως, έχει νόημα να θέλουμε το k να κυμαίνεται σε μικρές τιμές, κάτι που παρουσιάζονται παρακάτω.

### *Χωρική πολυπλοκότητα*

Η “απάντηση” του αλγορίθμου Apriori είναι όλα τα σύνολα $L_i$, $1 \leq i \leq k$. Ωστόσο, όπως προαναφέραμε, σε κάθε βήμα $k$ μεμονωμένα, ο Apriori χρησιμοποιεί τα $L_1 \text{και} L_{k-1}$, σπαταλά χώρο για το $C_k$. Αυτές είναι οι απαιτήσεις σε χώρο του αλγορίθμου σε κάθε βήμα και, όπως θα δούμε, όλες οι βελτιώσεις του Apriori κινούνται στην κατεύθυνση του περιορισμού του χώρου που καταλαμβάνει το $C_k$.

### *Επιλογή κατωφλίου $s$*

Είναι φανερό πως όσο μεγαλύτερες είναι τιμές παίρνει το $s$, τόσο περισσότερα καλάθια πρέπει να περιέχουν ένα σύνολο ώστε αυτό να θεωρηθεί “συχνό”. Στην άλλη κατεύθυνση, για μικρότερες τιμές του $s$, προκύπτουν περισσότερα στο πλήθος “συχνά” σύνολα, κάτι που τα καθιστά μη διαχειρίσιμα για έναν άνθρωπο. Δεδομένου του παραπάνω συμβιβασμού, είναι ιδιαίτερα κρίσιμη η επιλογή αυτή. Ωστόσο, δεν υπάρχει κοινά αποδεκτή διαδικασία εύρεσης της “κατάλληλης” τιμής του κατωφλίου υποστήριξης.

## 2.1.3.2.2 PCY

### Ιδέα

Ο αλγόριθμος PCY (Park et al. 1995), στηριζόμενος στον Apriori, προσπαθεί να μειώσει τον αριθμό τον υποψηφίων “συχνών” συνόλων $C_k$ και άρα να χρησιμοποιεί λιγότερο χώρο για την καταμέτρηση του αριθμού εμφάνισής τους.

Σε κάθε βήμα $k$, που αναζητά τα “συχνά” k-στοιχειοσύνολα, χρησιμοποιεί μια συνάρτηση κατακερματισμού που αντιστοιχεί στοιχειοσύνολα (κλειδιά) με θέσεις (τιμές κατατεμαχισμού) ενός πίνακα ακεραίων. Κάθε στοιχείο του πίνακα αντιστοιχεί στον συνολικό αριθμό εμφανίσεων όλων των k-στοιχειοσυνόλων που απεικονίζονται, μέσω της συνάρτησης κατακερματισμού, στη θέση αυτή.

Συγκεκριμένα, η εύρεση των “συχνών” k-στοιχειοσυνόλων του βήματος k χωρίζεται σε 2 στάδια. Στο πρώτο, διαβάζοντας κάθε καλάθι από το αρχείο καλαθιών, παράγουμε όλα τα k-στοιχειοσύνολα και αυξάνουμε κατά ένα το στοιχείο στη θέση του πίνακα στην οποία το καθένα αντιστοιχεί. Στο δεύτερο στάδιο, παράγοντας ξανά όλα τα δυνατά k-στοιχειοσύνολα του κάθε καλαθιού, θεωρούμε σαν υποψήφια “συχνα” k-στοιχειοσύνολα αυτά που ικανοποιούν την Apriori συνθήκη, αλλά αντιστοιχούν σε θέση του πίνακα με τιμή που υπερβαίνει το κατώφλι υποστήριξης.

## Υλοποίηση

```
Βήμα 1
   Δημιούργησε τον πίνακα κατακερματισμού για τα 2-στοιχειοσύνολα

   ∀ καλάθι
      ∀ στοιχείο (1-στοιχειοσύνολο)
         Μέτρα την εμφάνιση του στοιχείου αυτού (μέλος C_1)

      ∀ 2-στοιχειοσύνολο
         Βρές τη θέση του πίνακα κατακερματισμού στην οποία αντιστοιχεί
         και αύξησε τον αντίστοιχο μετρητή κατά ένα


   Για κάθε στοιχείο του C_1
      Αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_1)


Βήμα k (αρχικά k=2)
   Δημιούργησε τον νέο πίνακα κατακερματισμού για τα k+1-στοιχειοσύνολα

   ∀ καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού
      ∀ k-στοιχειοσύνολο των παραπάνω στοιχείων
         αν   όλα τα δυνατά k-1-στοιχειοσύνολα του είναι στοιχεία του L_{k-1}
```

```
        και
             η θέση του παλιού πίνακα κατακερματισμού, στην οποία το
             στοιχειοσύνολο αντιστοιχεί, έχει τιμή μεγαλύτερη του s
        τότε μέτρα την εμφάνισή του k-στοιχειοσυνόλου αυτού (μέλος C_k)

    ∀ k+1-στοιχειοσύνολο των παραπάνω στοιχείων
        Βρές τη θέση του νέου πίνακα κατακερματισμού στην οποία
        αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα

∀ στοιχείο του C_k
    αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
    τότε είναι συχνό (μέλος L_k)

Αν το L_k είναι κενό σύνολο
τότε  Τελος
αλλιώς  α) k = k + 1
        β) πήγαινε στο Βήμα k
```

Εικόνα 3.9: Ο αλγόριθμος PCY σε ψευδοκώδικα.

## Παρατηρήσεις

Αν ένα στοιχειοσύνολο είναι συχνό, τότε από μόνο του θα έχει οδηγήσει σε τουλάχιστον $s$ αυξήσεις του μετρητή της θέσης του πίνακα κατακερματισμού, στην οποία αυτό αντιστοιχεί μέσω της συνάρτησης κατακερματισμού. Έτσι, το στοιχειοσύνολο αυτό θα θεωρηθεί υποψήφιο "συχνό", θα καταμετρηθεί ο αριθμός εμφάνισής του στο αρχείο των καλαθιών και τελικά, θα θεωρηθεί "συχνό". Συνεπώς, δεν υπάρχουν ψευδώς αρνητικά (false negatives) αποτελέσματα.

Ωστόσο, διαφορετικά στοιχειοσύνολα μπορούν αντιστοιχούν στην ίδια θέση του πίνακα και να αυξάνουν τον ίδιο μετρητή. Έτσι, αν τελικά ο μετρητής αυτός υπερβεί το κατώφλι $s$, θα θεωρηθούν όλα σαν υποψήφια συχνά και θα καταμετρηθεί ο αριθμός εμφάνισής τους. Ο χώρος που χρειάζεται για την καταμέτρηση του αριθμού εμφάνισής τους αποτελεί σπατάλη αλλά δεν μπορεί να αποφευχθεί. Άρα, υπάρχουν ψευδώς θετικά (false positives) αποτελέσματα.

Συγκεκριμένα, εδώ αναδεικνύεται ο συμβιβασμός μεταξύ του μεγέθους του πίνακα κατακερματισμού και του μεγέθους του χώρου που σπαταλάται για καταμέτρηση μη "συχνών" στοιχειοσυνόλων. Όσο πιο μεγάλο είναι το μέγεθος του πίνακα κατακερματισμού, τόσο λιγότερα στοιχειοσύνολα θα αντιστοιχούν σε κοινές θέσεις. Έτσι, σπαταλώντας μεγαλύτερο χώρο στο μέγεθος του πίνακα αρχικά, μπορούμε να γλιτώσουμε χώρο για την καταμέτρηση της εμφάνισης των υποψηφίων "συχνών" στοιχειοσυνόλων.

## 2.1.3.2.3 Multistage

### Ιδέα

Ο Multistage (Fang et al. 1999) επεκτείνει την ιδέα του PCY, θυσιάζοντας χρόνο επιπλέον περασμάτων στο αρχείο καλαθιών για κάθε μέγεθος στοιχειοσυνόλων, με σκοπό τον περαιτέρω περιορισμό του αριθμού υποψηφίων “συχνών” συνόλων. Για παράδειγμα, με ένα επιπλέον πέρασμα στο αρχείο καλαθιών, χρησιμοποιούμε έναν δεύτερο πίνακα με διαφορετική συνάρτηση κατακερματισμού. Πλέον, ένα στοιχειοσύνολο για να θεωρηθεί υποψήφιο “συχνό” πρέπει, επίσης, να αντιστοιχεί σε θέση του δεύτερου πίνακα με τιμή που να υπερβαίνει το κατώφλι $s$.

Οι συνθήκες ενός k-στοιχειοσυνόλου, λοιπόν, για να θεωρηθεί υποψήφιο “συχνό” είναι:

- όλα τα δυνατά k-1-στοιχειοσύνολα του είναι στοιχεία του $L_{k-1}$ (Apriori συνθήκη)
- η θέση του 1ου πίνακα κατακερματισμού στην οποία το στοιχειοσύνολο αντιστοιχεί, έχει τιμή μεγαλύτερη του $s$ (PCY συνθήκη)
- όμοια για τη θέση του 2ου πίνακα κατακερματισμού (νέα συνθήκη)

### Υλοποίηση

```
Βήμα 1
   Δημιούργησε τον 1ο πίνακα κατακερματισμού για τα 2-στοιχειοσύνολα

   ∀ καλάθι
      ∀ στοιχείο (1-στοιχειοσύνολο)
         Μέτρα την εμφάνιση του στοιχείου αυτού (μέλος C1)

      ∀ 2-στοιχειοσύνολο
         Βρές τη θέση του 1ου πίνακα κατακερματισμού στην οποία
         αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα


   Δημιούργησε τον 2ο πίνακα κατακερματισμού για τα 2-στοιχειοσύνολα

   ∀ καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού
      ∀ 2-στοιχειοσύνολο των παραπάνω στοιχείων
            αν στον 1ο πίνακα κατακερματισμού αντιστοιχεί σε "συχνή" θέση
            τότε βρές τη θέση του 2ου πίνακα κατακερματισμού στην οποία
                 αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα
```

```
   ∀στοιχείο του C_1
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_1)


Βήμα k (αρχικά k=2)
   ∀καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού
      ∀k-στοιχειοσύνολο των παραπάνω στοιχείων
         αν   όλα τα δυνατά k-1-στοιχειοσύνολα του είναι στοιχεία του L_{k-1}
         και  οι θέσεις του 1ου και 2ου πίνακα κατακερματισμού στις οποίες
              το στοιχειοσύνολο αντιστοιχεί, έχουν τιμές μεγαλύτερες του s
         τότε μέτρα την εμφάνισή του k-στοιχειοσυνόλου αυτού (μέλος C_k)


      ∀k+1-στοιχειοσύνολο των παραπάνω στοιχείων
         Βρές τη θέση του νέου 1ου πίνακα κατακερματισμού στην οποία
         αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα


   ∀καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού
      ∀k+1-στοιχειοσύνολο των παραπάνω στοιχείων
            αν στον 1ο πίνακα κατακερματισμού αντιστοιχεί σε συχνή θέση
            τότε βρές τη θέση του νέου 2ου πίνακα κατακερματισμού στην
                 οποία αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά 1

   ∀στοιχείο του C_k
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_k)

   Αν το L_k είναι κενό σύνολο
   τότε Τελος
   αλλιώς α) k = k + 1
          β) πήγαινε στο Βήμα 2
```

Εικόνα 3.10: Ο αλγόριθμος Multistage σε ψευδοκώδικα.

### Παρατηρήσεις

Είναι προφανής ο συμβιβασμός της σπατάλης χρόνου για επιπρόσθετα περάσματα στο αρχείο καλαθιών, προς όφελος της οικονομίας σε χώρο για την καταμέτρηση των πιθανά λιγότερων υποψηφίων “συχνών” συνόλων.

Στην παραπάνω υλοποίηση χρησιμοποιήθηκαν 3 στάδια (2 περάσματα του αρχείου καλαθιών) για κάθε μέγεθος στοιχειοσυνόλων. Είναι εύκολη η γενίκευση σε n στάδια υπό τον περιορισμό της διατήρησης των n-1 πινάκων κατακερματισμού.

Ένα τελευταίο σημείο-κλειδί στον αλγόριθμο αυτόν, είναι η χρήση ανεξάρτητων συναρτήσεων κατακερματισμού ώστε να ελαχιστοποιούνται οι επικαλύψεις των αντιστοιχίσεων των στοιχειοσυνόλων σε θέσεις των πινάκων κατακερματισμού να αφορούν διαφορετικά στοιχειοσύνολα.

## 2.1.3.2.4 Multihash

### Ιδέα

Αντί ενός επιπλέον περάσματος στο αρχείο καλαθιών για κάθε πίνακα κατακερματισμού, ο Multihash (Fang et al. 1999) αξιοποιεί την ίδια λογική σε ίδιο πέρασμα. Συγκεκριμένα, μπορούμε να χρησιμοποιήσουμε δύο (ή όμοια n) πίνακες κατακερματισμού μισού ( $1/n$ ) μεγέθους. Έτσι, υποψήφιο “συχνό” στοιχειοσύνολο θα είναι αυτό που μέσω των διαφορετικών συναρτήσεων κατακερματισμού, αντιστοιχεί σε θέσεις που όλες τους έχουν τιμές που υπερβαίνουν το κατώφλι $s$.

### Υλοποίηση

```
Βήμα 1
   Δημιούργησε τους 2 πίνακες κατακερματισμού για τα 2-στοιχειοσύνολα

   ∀ καλάθι
      ∀ στοιχείο (1-στοιχειοσύνολο)
         Μέτρα την εμφάνιση του στοιχείου αυτού (μέλος C_1)

      ∀ 2-στοιχειοσύνολο
         Βρές τη θέση του 1ου πίνακα κατακερματισμού στην οποία
         αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα
```

```
        Βρές τη θέση του 2ου πίνακα κατακερματισμού στην οποία
        αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα


   ∀ στοιχείο του C_1
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_1 )


Βήμα k (αρχικά k=2)
   Δημιούργησε τους 2 πίνακες κατακερματισμού για τα k+1-στοιχειοσύνολα

   ∀ καλάθι
      Πάρε τα συχνά στοιχεία του καλαθιού

      ∀ k-στοιχειοσύνολο των παραπάνω στοιχείων
         αν   όλα τα δυνατά k-1-στοιχειοσύνολα του είναι στοιχεία του L_{k-1}
         και  οι θέσεις του 1ου και 2ου πίνακα κατακερματισμού στις οποίες
              το στοιχειοσύνολο αντιστοιχεί, έχουν τιμές μεγαλύτερες του s
         τότε μέτρα την εμφάνισή του k-στοιχειοσυνόλου αυτού (μέλος C_k )


      ∀ k+1-στοιχειοσύνολο των παραπάνω στοιχείων
         Βρές τη θέση του νέου 1ου πίνακα κατακερματισμού στην οποία
         αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα

         Βρές τη θέση του νέου 2ου πίνακα κατακερματισμού στην οποία
         αντιστοιχεί και αύξησε τον αντίστοιχο μετρητή κατά ένα


   ∀ στοιχείο του C_k
      αν το στοιχείο μετρήθηκε τουλάχιστον s φορές
      τότε είναι συχνό (μέλος L_k )

   Αν το L_k είναι κενό σύνολο
   τότε Τελος
   αλλιώς α) k = k + 1
          β) πήγαινε στο Βήμα 2
```

Εικόνα 3.11: Ο αλγόριθμος Multihash σε ψευδοκώδικα.

## Παρατηρήσεις

Ο Multihash αναμένουμε να υπερτερεί του Multistage όσο η μέση των πινάκων κατακερματισμού είναι μικρότερη του κατωφλίου $s$ . Αυτό, διότι όσο αυξάνουμε τον αριθμό των πινάκων κατακερματισμού ταυτόχρονα πρέπει να μειώνουμε ανάλογα το μέγεθός τους, προκειμένου να χρησιμοποιούμε την ίδια ποσότητα κύριας μνήμης.

Ωστόσο, το μικρότερο μέγεθος αυξάνει τις επικαλύψεις και άρα τις τιμές των μετρητών όπου στην οριακή περίπτωση όλοι θα υπερβαίνουν το κατώφλι $s$ και άρα δεν θα έχουμε καταφέρει να μειώσουμε καθόλου το πλήθος των υποψηφίων “συχνών” στοιχειοσυνόλων το οποίο για την ακρίβεια, θα ταυτίζεται με αυτό που είχαμε στον Apriori.

## 2.1.3.3 Αλγόριθμοι περιορισμένου αριθμού περασμάτων

Σε εφαρμογές που δεν χρειάζεται απαραίτητα να εντοπίσουμε κάθε “συχνό” στοιχειοσύνολο, ή σε άλλες που, αν και χρειάζεται, το μέγεθος της κύριας μνήμης δεν επαρκεί για την εφαρμογή των παραπάνω αλγορίθμων, μπορούμε να χρησιμοποιήσουμε τις ακόλουθες μεθόδους.

### 2.1.3.3.1 Simple-Randomized

#### Ιδέα

Ο Simple-Randomized επιλέγει ένα μέρος της βάσης συναλλαγών προσαρμόζοντας ανάλογα το κατώφλι $s$. Για παράδειγμα μπορούμε να επιλέξουμε τυχαία το 60% των καλαθιών, θέτοντας το κατώφλι $s$ στό 60% της αρχικής του τιμής, δηλαδή στο $0.6 * s$.

#### Υλοποίηση

Υπάρχει δυνατότητα επιλογής οποιουδήποτε εκ των προηγούμενων αλγορίθμων.

#### Παρατηρήσεις

Αξίζει να σημειωθεί ο τρόπος επιλογής του μέρους του αρχείο καλαθιών. Εάν η σειρά των καλαθιών στο αρχείο δεν είναι τυχαία, θα πρέπει είτε να “ανακατέψουμε” εμείς τα καλάθια και στη συνέχεια να επιλέξουμε τα πρώτα που αντιστοιχούν στο 60% επί του συνόλου, είτε να ορίσουμε την πιθανότητα $p$ (0.6 στο παράδειγμά μας) αποδοχής κάθε καλαθιού καθώς τα διαβάζουμε ένα-ένα. Έτσι, σε κάθε περίπτωση κρατάμε $p * \text{αρχικό αριθμό καλαθιών}$ καλάθια.

Είναι ιδιαίτερα σημαντικό επίσης, να σημειωθεί ότι ένα “συχνό” στοιχειοσύνολο του αρχικού αρχείου να μην είναι ”συχνό” στο κομμάτι που επιλέχθηκε (false negatives) και ένα μη-”συχνό” του αρχικού αρχείου να βρέθηκε “συχνό” στο κομμάτι (false positive), δηλαδή προκύπτουν τόσο false negatives όσο και false positives. Μπορούμε να εξαλείψουμε τα false positives με ένα πέρασμα στο αρχικό αρχείο, δηλαδή ελέγχοντας ποιά απ’ τα “συχνά” στο κομμάτι είναι επίσης “συχνά” σε ολόκληρο το αρχείο. Για τον περιορισμό των false negatives, μπορούμε να θέσουμε το κατώφλι υποστήριξης ίσο με $0.8 * p * s$, δηλαδή κάτι μικρότερο του $p * s$, κάτι που δεν θα τα εξάλειφε απλώς θα τα περιόριζε.

## 2.1.3.3.2 SON

### Ιδέα

Ο αλγόριθμος SON (Savasere et al. 1995) υλοποιεί το MapReduce μοντέλο για να εκμεταλλευτεί τους πολλούς επεξεργαστές, παραλληλοποιώντας τη διαδικασία εύρεσης των “συχνών” στοιχειοσυνόλων.

Χωρίζουμε το αρχείο καλαθιών σε $n$ ισόποσα κομμάτια, όσα και οι επεξεργαστές. Κάθε κομμάτι, συνοδεύεται με το ανάλογα προσαρμοσμένο κατώφλι $s$. Αν εντοπιστούν τα “συχνά” στοιχειοσύνολα κάθε κομματιού, τότε η ένωση αυτών θα αποτελεί τα υποψήφια “συχνά” στοιχειοσύνολα ολόκληρου του αρχείου καλαθιών. Απομένει η καταμέτρηση του αριθμού εμφάνισης των υποψηφίων αυτών στοιχειοσυνόλων σε καθένα από τα αρχικά κομμάτια. Έτσι, για τα συχνά στοιχειοσύνολα θα πρέπει το άθροισμα αριθμού εμφάνισης σε όλα τα κομμάτια Να υπερβαίνει το κατώφλι $s$.

### Υλοποίηση

1.Map συνάρτηση (για κάθε επεξεργαστή)
    Πάρε το κομμάτι καλαθιών
    Προσάρμοσε το κατώφλι $s$ στο ποσοστό του μεγέθους του κομματιού αυτού επί του μεγέθους του αρχικού αρχείου
    Βρες τα “συχνά” στοιχειοσύνολα του κομματιού μέσω οποιουδήποτε αλγόριθμου της Apriori οικογένειας

2.Reduce συνάρτηση
    Υπολόγισε την ένωση όλων των “συχνών” στοιχειοσυνόλων
    (Το αποτέλεσμα αυτό αποτελεί το σύνολο των υποψήφιων “συχνών” συνόλων ολόκληρου του αρχείου καλαθιών)

3.Map συνάρτηση (για κάθε επεξεργαστή)
    Πάρε το ίδιο κομμάτι καλαθιών και το συνολο των υποψήφιων “συχνών” συνόλων
    Καταμέτρησε το αριθμό εμφάνισης των παραπάνω στοιχειοσυνόλων στο κομμάτι αυτό

4.Reduce συνάρτηση
    Άθροισε τους αριθμούς εμφάνισης κάθε υποψήφιου “συχνού” στοιχειοσυνόλου σε όλα τα κομμάτια
    “Συχνά” είναι όσων το άθροισμα υπερβαίνει το κατώφλι $s$

Εικόνα 3.12: Ο αλγόριθμος SON σε ψευδοκώδικα.

### Παρατηρήσεις

Ιδιότητα 3.5:

- Ένα συχνό στοιχειοσύνολο στο αρχικό αρχείο καλαθιών, θα είναι συχνό σε τουλάχιστον ένα κομμάτι καλαθιών.

Απόδειξη

Έστω

το στοιχειοσύνολο $X$ είναι συχνό

χωρίζουμε το αρχείο σε $n$ αυθαίρετα κομμάτια

$o_i$ το ποσοστό του μεγέθους του κομματιού $i$ ως προς το αρχικό, δηλαδή

$\sum_{i=1}^{n} o_i = 1$

προσαρμόζουμε το κατώφλι του $i$-οστού κομματιού σε $s * o_i$

το $X$ προκύπτει μη συχνό σε όλα τα κομμάτια, δηλαδή $k_i < s * o_i \ \forall i$, όπου $k_i$ ο αριθμός εμφάνισης του $X$ στο $i$-οστό κομμάτι

Τότε $\sum_{i=1}^{n} k_i < \sum_{i=1}^{n} s * o_i = s * \sum_{i=1}^{n} o_i = s$, δηλαδή το $X$ δεν είναι “συχνό”, άτοπο!

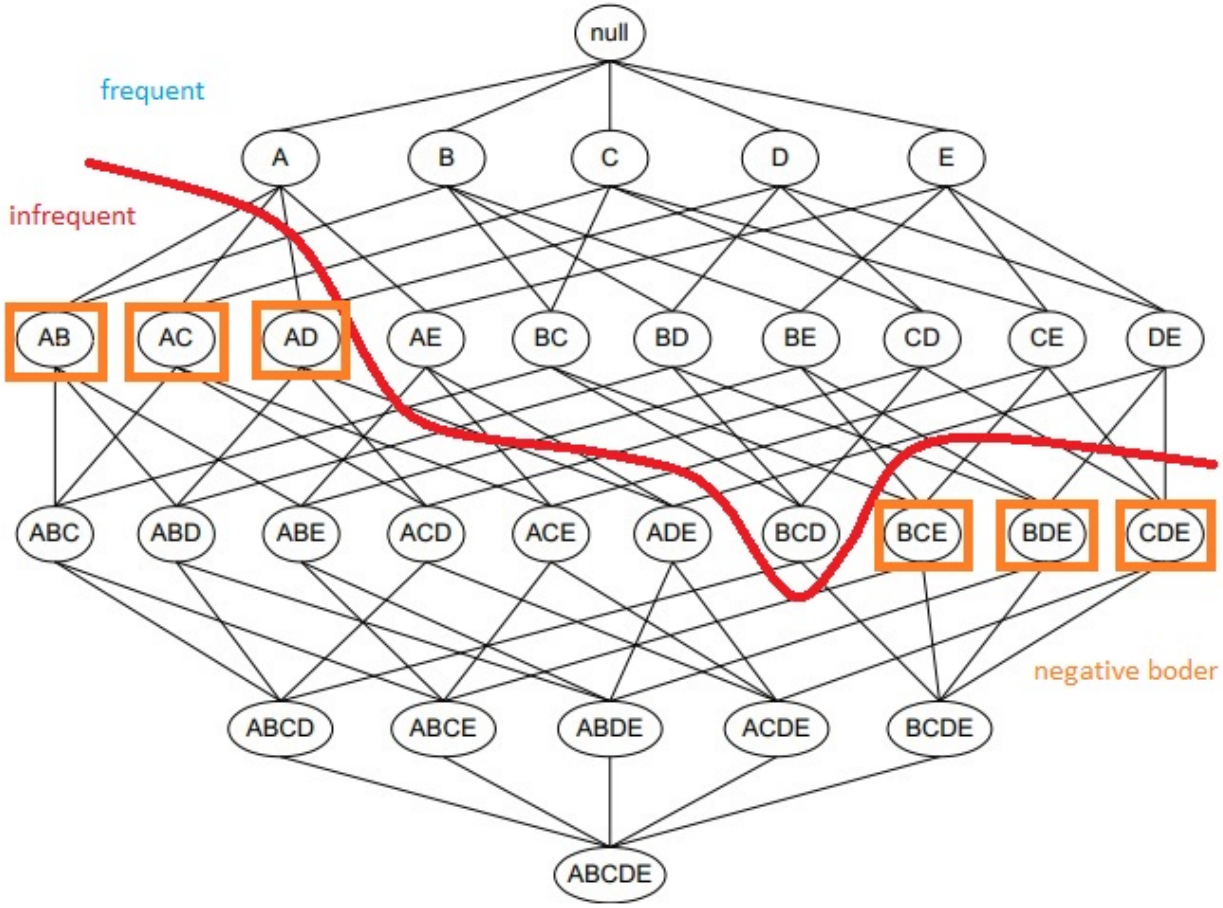
## 2.1.3.3.3 Toivonen

### Ιδέα

Ο αλγόριθμος Toivonen (Toivonen et al. 1996) , εκμεταλλευόμενος κάποιες ιδιότητες της θεωρίας συνόλων, αρκείται σε συνολικά δύο περάσματα, ένα σε κάποιο μικρό κομμάτι του αρχείου καλαθιών και ένα δεύτερο, σε ολόκληρο το αρχείο. Δεδομένου ότι υπάρχει μία μικρή πιθανότητα ο αλγόριθμος να αποτύχει να δώσει απάντηση, θα πρέπει να επαναλαμβάνεται μέχρις ότου να έχουμε αποτέλεσμα. Ωστόσο, αποδεικνύεται ότι κατά μέσο όρο, ο συνολικός αριθμός περασμάτων είναι μικρός.

Απαιτείται η παρουσίαση κάποιων ιδιοτήτων για την κατανόηση του τρόπου λειτουργίας του αλγορίθμου.

- **Αρνητικό όριο** είναι το σύνολο των στοιχειοσυνόλων που αν και δεν είναι “συχνά”, αν τους αφαιρεθεί οποιοδήποτε στοιχείο, τότε προκύπτει “συχνό” στοιχειοσύνολο. (3.11)




Εικόνα 3.13: Αρνητικό όριο (negative border).

Συγκεκριμένα, το πρώτο πέρασμα αφορά στην εύρεση των “συχνών” στοιχειοσυνόλων σε ένα μικρό κομμάτι καλαθιών. Στη συνέχεια υπολογίζουμε το “αρνητικό όριο” αυτών. Τέλος, περνάμε ολόκληρο το αρχείο καλαθιών καταμετρώντας τον αριθμό εμφάνισης των στοιχειοσυνόλων που είτε είναι συχνά στο κομμάτι, είτε ανήκουν στο “αρνητικό όριο”. Αν δεν βρεθεί στοιχειοσύνολο του “αρνητικού ορίου” που να είναι “συχνό” στο αρχείο, τότε “συχνά” στοιχειοσύνολα είναι το υποσύνολο των “συχνών” στο κομμάτι που βρέθηκαν “συχνά” και στο αρχείο. Διαφορετικά, επαναλαμβάνουμε τη διαδικασία.

## Υλοποίηση

1.Πρώτο Πέρασμα
- Πάρε ένα κομμάτι καλαθιών
- Προσάρμοσε το κατώφλι υποστήριξης όπως πριν
- Βρές τα “συχνά” στοιχειοσύνολα στο κομμάτι αυτό
- Υπολόγισε το “αρνητικό όριο” των “συχνών” στοιχειοσυνόλων

2.Δεύτερο Πέρασμα

Καταμέτρησε, σε ολόκληρο το αρχείο, τον αριθμό εμφάνισης των στοιχειοσυνόλων που ανήκουν είτε στα “συχνά” στοιχειοσύνολα του κομματιού είτε στο “αρνητικό όριό” τους.

**αν** δεν βρεθεί κάποιο στοιχειοσύνολο να είναι “συχνό” σε ολόκληρο το αρχείο και
ταυτόχρονα να ανήκει στο “αρνητικό όριο”
**τότε** η απάντηση είναι όσα απ’ τα “συχνά” στοιχειοσύνολα του $1_{ου}$ περάσματος βρέθηκαν επίσης “συχνά” σε ολόκληρο το αρχείο
**αλλιώς** επανάλαβε τη διαδικασία

Εικόνα 3.14: Ο αλγόριθμος Toivonen σε ψευδοκώδικα.

## Παρατηρήσεις

Είναι αξιοσημείωτος ο τρόπος προσαρμογής του κατωφλίου $s$ στο κομμάτι καλαθιών του πρώτου περάσματος. Όπως και πριν, ορίζουμε κάτι μικρότερο, όπως το 90%, του λόγου του μεγέθους του κομματιού προς αυτό του συνολικού αρχείου. Έτσι, ελαχιστοποιούμε την πιθανότητα ο αλγόριθμος να μην μπορέσει να δώσει απάντηση, επιβαρύνοντας ωστόσο, την κύρια μνήμη.

## Απόδειξη αλγορίθμου

Ο αλγόριθμος Toivonen, όπως και ο Simple-Randomized, δεν παράγει false positives χάρη στον έλεγχο που πραγματοποιεί κατά το πέρασμα ολόκληρου του αρχείου. Για να δείξουμε ότι δεν παράγει ούτε false negatives, αρκεί να δείξουμε ότι δεν γίνεται να υπάρχει “συχνό” στοιχειοσύνολο σε ολόκληρο το αρχείο, που ούτε να είναι “συχνό” στο κομμάτι και ούτε να είναι μέλος του αρνητικού ορίου.

‘Εστω

ο αλγόριθμος Toivonen έχει παράξει αποτέλεσμα

υπάρχει σύνολο $S$ το οποίο
είναι “συχνό” σε ολόκληρο το αρχείο
δεν είναι “συχνό” το κομμάτι
δεν είναι μέλος του αρνητικού ορίου

το $S$ δεν είναι μέρος της λύσης

Τότε

επιλέγουμε το σύνολο $T$ το οποίο

$T \subseteq S$

δεν είναι “συχνό” στο κομμάτι

έχει το μικρότερο δυνατό μέγεθος

Είναι φανερό πως

- το $T$ είναι συχνό σε ολόκληρο το αρχείο (ως υποσύνολο συχνού συνόλου)
- το $T$ ανήκει στο “αρνητικό όριο”

Άρα, ο αλγόριθμος δεν μπορεί να παράγει αποτέλεσμα σε αυτόν τον γύρο, ΑΤΟΠΟ!

## 2.1.4 Κανόνες Συσχέτισης

### 2.1.4.1 Εισαγωγή

Έχοντας βρει λοιπόν τα “συχνά” στοχειοσύνολα, για δεδομένο κατώφλι υποστήριξης, προχωράμε στην εξαγωγή Κανόνων Συσχέτισης (Association Rules) με μορφή $A \rightarrow B,$ όπου $A, B$ δύο στοιχειοσύνολα με $A \cap B = \emptyset$. Συγκεκριμένα, από κάθε “συχνό” στοιχειοσύνολο $X$ με πληθικό αριθμό $N$, $|X| = N$, παράγουμε όλους του κανόνες (Ν στο σύνολο) της μορφής $X \setminus \{x\} \rightarrow x$ , $\forall x \in X$. Δηλαδή, ως $A$ ( $B$ ) τίθενται τα άμεσα υποσύνολα - ορισμός 3.9 - του $X$ (το εναπομείναν στοιχείο). Στη συνέχεια, κρατάμε μόνο τους κανόνες που ικανοποιούν το κατώφλι εμπιστοσύνης. Ωστόσο, επειδή το πλήθος των κανόνων παραμένει μεγάλο, προχωράμε στην αξιολόγησή τους.

### 2.1.4.2 Αξιολόγηση Κανόνων

Κάθε κανόνας είναι μία σχέση μεταξύ των στοιχειοσυνόλων *A,B* τα οποία αποτελούν δυαδικές (binary) μεταβλητές, με την έννοια ότι είτε υπάρχουν είτε όχι σε κάθε καλάθι προϊόντων. Η μελέτη της σχέσης μεταξύ δυαδικών μεταβλητών βασίζεται στον $2 \, x \, 2$ πίνακα συνάφειας (contingency table) του εν λόγω κανόνα.

| | B | $\overline{\text{B}}$ | **Total** |
|---|---|---|---|
| A | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\overline{\text{A}}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| **Total** | $f_{+1}$ | $f_{+0}$ | |

Εικόνα 3.15: Πίνακας συνάφειας ενός κανόνα $A \rightarrow B$. Συνοπτικά, $f_{11}$ είναι ο αριθμός των καλαθιών που περιέχουν και τα δύο στοιχειοσύνολα, $f_{10}$ ο αριθμός αυτών που περιέχουν το $A$ και όχι το $B$, ενώ $f_{1+}$ ο αριθμός των καλαθιών που περιέχουν το $A$.

Στην βιβλιογραφία υπάρχει πληθώρα επιλογών ποσοτικών μεγεθών (Tan et al. 2004; Geng et al. 2006; Hahsler 2015), τα αντικειμενικά μέτρα (objective measures), που χρησιμοποιούν τον παραπάνω πίνακα και, βασισμένα στη Θεωρία Πιθανοτήτων, αξιολογούν τους κανόνες συσχέτισης. Ωστόσο, δεν υφίσταται κοινώς αποδεκτή διαδικασία επιλογής τέτοιων μέτρων, κάτι που καθιστά την επιλογή αυτών ένα σημαντικό ζήτημα. Έχουν, λοιπόν, κυριαρχήσει δύο κύριες μέθοδοι για τη σύγκριση και την ανάλυση των αντικειμενικών μέτρων, η βαθμολόγηση (ranking) (Tan et al. 2004; Lenca et al. 2004) και η συσταδοποίησή (clustering) (Vaillant et al. 2004; Lenca et al. 2007) τους. Οι μέθοδοι στηρίζονται είτε στις ιδιότητες των μέτρων είτε σε εμπειρικές παρατηρήσεις, για αυτά, πάνω σε δεδομένα. Εμείς ασχοληθήκαμε με την βαθμολόγηση κανόνων.

Στην κατεύθυνση αυτή, έχουν προταθεί διάφορες ιδιότητες (Piatetsky-Shapiro et al. 1991; Tan et al. 2002; Lenca et al. 2004; Geng et al. 2006) οι οποίες βασίζονται στον πίνακα συνάφειας. Μελετώντας τις ιδιότητες αυτές, μπορεί κανείς να ορίσει τις πιο επιθυμητές, ως προς την εφαρμογή του και έτσι να καταλήξει σε ένα σύνολο αντικειμενικών μέτρων. Παρακάτω, παρουσιάζουμε τις ιδιότητες αλλά και τα ποσοτικά μεγέθη που κρίναμε επιθυμητά (-ές) ή μη για την περιοχή της Ανάλυσης Καλαθιού Αγορών. Για διευκόλυνση, η αναπαράσταση των $2x2$ πινάκων συνάφειας πραγματοποιείται μέσω της ακόλουθης μήτρας.

$$M = \begin{bmatrix} f_{11} & f_{10} \\ f_{01} & f_{00} \end{bmatrix}$$

και έτσι, κάθε αντικειμενικό μέτρο είναι ένας τελεστής, $O$, που εφαρμόζεται σε αυτόν τον πίνακα και τον μετασχηματίζει σε ένα βαθμωτό μέγεθος.

Ιδιότητα 3.6:

- Ένα αντικειμενικό μέτρο $O$ είναι συμμετρικό ως προς τη μετάθεση μεταβλητών ($A \leftrightarrow B$) όταν $O(M) = O(M^{\mathrm{T}})$ για κάθε πίνακα συνάφειας $M$.

Στην εφαρμογή μας υπάρχει ανάγκη διάκρισης μεταξύ των κανόνων $A \rightarrow B$ και $B \rightarrow A$ και για αυτόν τον λόγο η παραπάνω ιδιότητα κρίνεται ανεπιθύμητη.

Έστω

ένα αντικειμενικό μέτρο $O$

Τότε

θα λέμε ότι το $O$ είναι κανονικοποιημένο αν παίρνει τιμές στο διάστημα [-1,1]. Κάθε μη κανονικοποιημένο μέτρο $U$, που εκτείνεται μεταξύ 0 και $+\infty$, μπορεί να κανονικοποιηθεί μέσω των μετασχηματισμών $\frac{U-1}{U+1}$ ή $\frac{tan^{-1}log(U)}{\pi/2}$ . (3.12)

Ιδιότητα 3.7:

- Ένα κανονικοποιημένο αντικειμενικό μέτρο $O$ είναι αντισυμμετρικό ως προς τη μετάθεση γραμμών αν $O(SM) = -O(M)$, ενώ ως προς τη μετάθεση στηλών αν $O(MS) = -O(M)$ για κάθε πίνακα συνάφειας $M$.

Για τη δική μας εφαρμογή, κρίναμε επιθυμητό τα αντικειμενικά μέτρα που θα χρησιμοποιήσουμε να είναι μη συμμετρικά ως προς τη μετάθεση γραμμών και στηλών. Διαφορετικά, δεν θα είχαμε διάκριση μεταξύ θετική και αρνητικής συσχέτισης (correlation) των στοιχειοσυνόλων που συμμετέχουν στους κανόνες μας.

Ιδιότητα 3.8:

- Ένα αντικειμενικό μέτρο $O$ είναι null invariant αν $O(M+C) = O(M)$ όπου C=[0 0; 0 k] και k μια σταθερά.

Αυτή η ιδιότητα εξασφαλίζει ότι η βαθμολόγηση του κανόνα, από μέτρα που φέρουν την ιδιότητα αυτή, θα παραμένει ανεπηρέαστη από την προσθήκη καλαθιών τα οποία δεν περιέχουν ούτε το $A$ ούτε το $B$. Αυτό είναι χρήσιμο σε εφαρμογές όπου η συνύπαρξη δύο μεταβλητών είναι ισχυρότερη της παράλληλης απουσίας τους.

Ιδιότητα 3.9:

- Η αξιολόγηση ενός αντικειμενικού μέτρου $O$ να αυξάνεται μονότονα με το $P[A,B]$ όταν τα $P[A]$ και $P[\mathrm{B}]$ παραμένουν σταθερά. Ισοδύναμα, $O(M_2) > O(M_1)$ αν

$$M_2 = M_1 + \begin{bmatrix} +k & -k \\ -k & +k \end{bmatrix}$$

όπου k μία σταθερά.

Ιδιότητα 3.10:

- Η αξιολόγηση ενός αντικειμενικού μέτρου $O$ να μειώνεται μονότονα με το $P[A]$ (ή $P[B]$) όταν οι υπόλοιπες ποσότητες, $P[A,B]$ και $P[\mathrm{B}]$ (ή $P[A]$) παραμένουν σταθερές. Ισοδύναμα, $O(M_2) < O(M_1)$ αν

$$M_2 = M_1 + \begin{matrix} [0 & +k \\ 0 & -k\ ] \end{matrix}$$

ή

$$M_2 = M_1 + \begin{matrix} [0 & 0 \\ +k & -k\ ] \end{matrix}$$

Στη συνέχεια παρουσιάζονται τα αντικειμενικά μέτρα που επιλέχθηκαν:

- **Confidence** ($A \rightarrow B$) (Agrawal et al. 1993)

$$= \frac{support\,(A \rightarrow B)}{support\,(A)} = \frac{P[A \cap B]}{P[A]} = P[B|A] \quad (3.13)$$

- **Added Value** ($A \rightarrow B$) (Sahar et al. 1999)

$$= \text{confidence}\,(A \rightarrow B) - \text{support (B)} = P[B|A] - P[B] \quad (3.14)$$

- **Laplace** ($A \rightarrow B$) (Clark et al. 1991)

$$= \frac{f_{11} + 1}{f_{+1} + 2} \quad (3.15)$$

- **J-Measure** ($A \rightarrow B$) (Smyth et al. 1992)

$$= P[A \cap B] * \log\left(\frac{P[B|A]}{P[B]}\right) + P[A \cap \overline{B}] * \log\left(\frac{P[\overline{B}|A]}{P[\overline{B}]}\right) \quad (3.16)$$

- **Conviction** ($A \rightarrow B$) (Brin et al. 1997)

$$= \frac{1 - support\,(B)}{1 - support\,(A \rightarrow B)} = \frac{P[A] * P[\overline{B}]}{P[A \cap \overline{B}]} \quad (3.17)$$

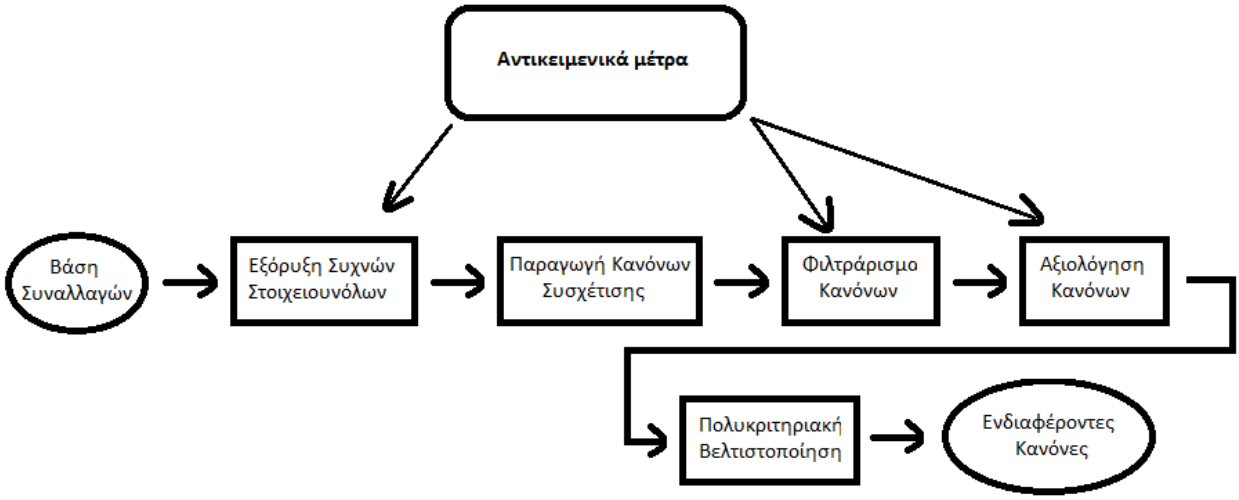
- **Klosgen** ($A \rightarrow B$) (Klösgen 1992)

$$= \sqrt{support(A \cup B)} * \text{AV}\,(A \rightarrow B) = \sqrt{P[A \cap B]} * (P[B|A] - P[B]) \quad (3.18)$$

Ωστόσο, το πρόβλημα δεν έχει λυθεί ακόμα. Μέχρι στιγμής, συνεχίζουμε να έχουμε τον ίδιο (μεγάλο) αριθμό κανόνων και έξι διαφορετικές βαθμολογήσεις για κάθε κανόνα. Ο πιο απλός τρόπος ανάδειξης των καλύτερων κανόνων είναι η χρήση ενός μόνο ποσοτικού μεγέθους εκ των παραπάνω. Ένας άλλος με τον οποίο μπορούν να αναδειχθούν οι κανόνες-νικητές είναι η πολυκριτηριακή μέθοδος Pareto η οποία αναπαριστά κάθε κανόνα ως ένα σημείο του χώρου $R^6$ , όπου φυσικά οι άξονες αντιστοιχούν στα αντικειμενικά μέτρα.Ακολουθούν συνοπτικοί ορισμοί για την εν λόγω μέθοδο.

- **Βελτίωση κατά Pareto** ενός σημείου του χώρου, συντελείται από την ύπαρξη άλλων σημείων, τα οποία παρουσιάζουν άνοδο σε τουλάχιστον έναν άξονα (βαθμολόγηση από κάποιο αντικειμενικό μέτρο), ενώ ταυτόχρονα καμία μείωση στους υπόλοιπους άξονες (3.19)
- **Σημεία-Νικητές κατά Pareto** αποτελούν τα σημεία που δεν επιδέχονται βελτίωσης κατά Pareto (3.20)

Έτσι, οι κανόνες-νικητές ισοδυναμούν με τα σημεία εκείνα, τα οποία αν προσπαθήσουμε να αντικαταστήσουμε με οποιοδήποτε άλλο, θα προκύψει τουλάχιστον ένα αντικειμενικό μέτρο με χαμηλότερη βαθμολόγηση. Συνοψίζουμε τη διαδικασία εξόρυξης ενδιαφερόντων Κανόνων Συσχέτισης, τονίζοντας την χρήση των αντικειμενικών μέτρων μέσα στα επιμέρους βήματα.




Εικόνα 3.16: Βήματα εύρεσης ενδιαφερόντων Κανόνων Συσχέτισης και χρήση αντικειμενικών μέτρων μέσα σε αυτά.

Στην Εξόρυξη Συχνών Στοιχειοσυνόλων χρησιμοποιήθηκε η **υποστήριξη** (support), ενώ στο Φιλτράρισμα των Κανόνων Συσχέτισης η **εμπιστοσύνη** (confidence). Τα δύο αυτά μεγέθη συγκροτούν το γνωστό μοντέλο Support-Confidence. Τέλος, για την αξιολόγηση των ευρεθέντων κανόνων χρησιμοποιήσαμε τα παραπάνω αντικειμενικά μεγέθη.

## 2.1.5 Δίκτυο Κανόνων Συσχέτισης

Η ιδέα αυτής της μεθόδου (Chawla et al. 2003; Pandey et al. 2009; Chawla et al. 2003) είναι η φανέρωση των άμεσων και έμμεσων συσχετίσεων που αφορούν ένα δεδομένο στοιχείο-στόχο. Συγκεκριμένα, κατασκευάζεται ένας γράφος όπου κάθε στοιχείο (κόμβος) συνδέεται με το στοιχείο-στόχο είτε κατευθείαν μέσω ενός κανόνα συσχέτισης, είτε μέσω ενός μονοπατιού κανόνων.

### 2.1.5.1 Ορισμός

Έστω

Ένα συνόλου κανόνων συσχετίσεων $R$ και ένα στοιχείο-στόχος $z$

Τότε

Δίκτυο Κανόνων Συσχέτισης ARN $(R, z)$ είναι ένα κατευθυνόμενο υπεργράφημα $G$ όπου:

- κάθε υπερακμή αντιστοιχεί σε έναν κανόνα $A \rightarrow B,$ όπου $|B| = 1$
- $\exists$ υπερακμή $(u, z)$
- $\nexists$ υπερακμή $(z, u)$
- κάθε κορυφή του $G$ έχει μονοπάτι προς το $z$

### 2.1.5.2 Παρατηρήσεις

#### Ακμές

Το αρχικό σύνολο ακμών $R$, αντιστοιχεί στους κανόνες συσχέτισης που παίρνουμε απ΄΄ το προαναφερθέν μοντέλο Support-Confidence. Για την κατασκευή του $G$, ορίζουμε ένα ποσοτικό μέγεθος που θα χρησιμοποιηθεί σαν βάρος των ακμών αντιστοιχεί σε έναν κανόνα. Για τον καθορισμό του βάρους των ακμών, απαιτείται η επιλογή ενός μη συμμετρικού, ως προς τη μετάθεση μεταβλητών (Ιδιότητα 3.6), ποσοτικού μεγέθους. Εμείς, επιλέξαμε την υποστήριξη (confidence) που, για κάθε κανόνα της μορφής $A \rightarrow B$, εκφράζει την πιθανότητα $P[B|A]$.

#### Επιλογές *R, z*

Η επιλογή του συνόλου Κανόνων Συσχέτισης R ουσιαστικά οδηγεί στην κατάλληλη επιλογή τιμών υποστήριξης (support) και εμπιστοσύνης (confidence) που χρησιμοποιήθηκαν στην Εξόρυξη Κανόνων Συσχέτισης, κάτι που δεν διευθετείται εδώ. Από την άλλη, η επιλογή του στοιχείου-στόχου είναι κομβική, αφού όπως θα δείξουμε αργότερα οδηγεί σε τελείως διαφορετικά δίκτυα. Θα μπορούσε κανείς να επιλέξει το στοιχείο που βρίσκεται στον μεγαλύτερο αριθμό καλαθιών, αυτό που έχει αγοραστεί μαζί με τα περισσότερα διαφορετικά στοιχεία ή αυτό που συμμετέχει στους περισσότερους κανόνες συσχέτισης. Ακολουθήσαμε την τελευταία επιλογή μετρώντας, για κάθε στοιχείο, τον συνολικό αριθμό κανόνων στους οποίους βρίσκεται είτε στο δεξί είτε στο αριστερό μέλος.

#### Μειονεκτήματα

Η τεχνική αυτή κληρονομεί τα μειονεκτήματα του προηγούμενου βήματος, δηλαδή την χρήση των κατωφλίων support και confidence. Επίσης, χρησιμοποιεί και η ίδια ένα ακόμα

ποσοτικό μέγεθος για τον καθορισμό των βαρών, που “κλαδεύει” περαιτέρω τους κανόνες συσχέτισης του συνόλου γραφήματος $G$.

## 2.1.6 Ιστορική Εξέλιξη

Πέραν της Apriori οικογένειας αλγορίθμων, έχουν αναπτυχθεί διάφοροι αλγόριθμοι Εξόρυξης Συχνών Στοιχειοσυνόλων. Η κύρια διαφοροποίηση τους έγκειται στον τρόπο αναπαράστασης της βάσης συναλλαγών. Από τη μια, οι αλγόριθμοι **οριζόντιας** διαμοίρασης, όπως ο Apriori, “βλέπουν” καλάθια που περιέχουν στοιχεία, ενώ οι αλγόριθμοι **κατακόρυφης** διαμοίρασης στοιχεία που περιέχονται σε καλάθια. Στη συνέχεια, παρουσιάζουμε συνοπτικά αλγορίθμους των κατηγοριών αυτών.

Επιπλέον, όπως προαναφέραμε, στην προσπάθεια περιορισμού των προφανών ή περιττών κανόνων και κατ’ επέκταση του συνολικού τους αριθμού, έχουν οριστεί τα κλειστά (closed) και τα μέγιστα συχνά (maximal frequent) στοιχειοσύνολα όπως παρουσιάζονται παρακάτω.

- Κλειστά Στοιχειοσύνολα είναι αυτά που δεν έχουν κανένα άμεσο υπερύνολο (3.10) με ακριβώς την ίδια τιμή υποστήριξης (support) (3.21)
- Μέγιστα Συχνά Στοιχειοσύνολα είναι τα συχνά στοιχειοσύνολα που δεν έχουν κανένα άμεσο υπερσύνολο (3.10) το οποίο να είναι επίσης συχνό. (3.22)

Οι ορισμοί αυτοί μπορούν να χρησιμοποιηθούν στην Εξόρυξη Συχνών Στοιχειοσυνόλων σαν ένα βήμα μετα-επεξεργασίας. Συγκεκριμένα, έχοντας βρει τα Συχνά Στοιχειοσύνολα (αποτέλεσμα οποιουδήποτε αλγορίθμου της οικογένειας Apriori), μπορούμε να εστιάσουμε μόνο στα Κλειστά ή στα Μέγιστα Συχνά και να ακολουθήσουμε τα υπόλοιπα βήματα ως έχουν. Ωστόσο, όπως θα δούμε, υπάρχουν αλγόριθμοι που εξαρχής αναζητούν τα στοιχειοσύνολα με αυτές τις ιδιότητες.

### 2.1.6.1 Αλγόριθμοι οριζόντιας διαμοίρασης

Ο αλγόριθμος Mannila (Mannila et al. 1994), δανειζόμενος την βασική ιδέα του Apriori, προσπαθεί να ανιχνεύει κάποια υποψήφια συχνά στοιχειοσύνολα των επόμενων βημάτων. Ο CAP (Ng et al. 1998) επιτρέπει τον έλεγχο του χρήστη στη διαδικασία εύρεσης Συχνών Στοιχειοσυνόλων. Ο αλγόριθμος UWEP (Ayan et al. 1999), συνεχίζοντας το έργο των FUP $_2$ (Cheung et al. 1997) και Partition-Update (Omiecinski et al. 1998), ασχολείται με την ανανέωση των συχνών στοιχειοσυνόλων κατά την προσθήκη νέων συναλλαγών. Οι TreeProjection (Agarwal et al. 2001), FP-Growth (Han et al. 2000) και COFI (El-Hajj et al. 2003) χρησιμοποιούν τις δικές τους δενδρικές δομές που οργανώνουν τα στοιχεία σε λεξικογραφική διάταξη, την οποία αξιοποιούν με μία ποικιλία τρόπων, όπως η αναζήτηση κατά βάθος (DFS), η αναζήτηση κατά

πλάτος (BFS) ή συνδυασμός αυτών. Οι αλγόριθμοι CLOSE (Pasquier et al. 1999), CLOSET (Pei et al. 1994) και CLOSET+ (Wang et al. 2003) αναζητούν κλειστά στοιχειοσύνολα (3.21), ενώ ο SmartMiner (Zou et al. 2002)μέγιστα συχνά στοιχειοσύνολα (3.22). Τέλος, οι FPmax* και FPclose (Grahne et al. 2005) επιτρέπουν την αποδοτικότερη χρήση παραλλαγών των FP-Trees, μέσω της τεχνικής FP-array.

#### 2.1.6.2 Αλγόριθμοι κατακόρυφης διαμοίρασης

Οι αλγόριθμοι MAFIA (Burdick et al. 2001) και GenMax (Gouda et al. 2001) ασχολούνται με την εξόρυξη μέγιστων συχνών στοιχειοσυνόλων και ο δεύτερος εισήγαγε μία τεχνική ελέγχου της μεγιστότητας. Ο TM (Song et al. 2006) βασίζεται σε δενδρικές δομές στοιχείων. Το Σύστημα Ανάκτησης Πληροφοριών και το Memory-based online pattern (Qiao et al. 2012) ασχολούνται με το πρόβλημα του πραγματικού χρόνου αναζήτησης προτύπων. Οι Charm (Zaki et al. 2005), DCI-CLOSED (Lucchese et al. 2006) και DBV-Miner (Vo et al. 2012) βρίσκουν τα κλειστά στοιχειοσύνολα.

#### 2.1.6.3 Υβριδικοί Αλγόριθμοι

Μία από τις πρώτες προσπάθειες συνδυασμού των δύο παραπάνω τρόπων διαμοίρασης έγινε με τους αλγορίθμους Eclat , MaxEclat, Clique, MaxClique (Zaki et al. 1997) . Επίσης, ο kDCI (Lucchese et al. 2004), επέκταση του DCI (Orlando et al. 2002), χρησιμοποιεί λεξικογραφική ταξινόμηση για να συμπιέσει τα στοιχειοσύνολα.

#### 2.1.6.4 Σύγχρονοι Αλγόριθμοι

Οι αλγόριθμοι FIN (Deng et al. 2014), PrePost (Deng et al. 2012) και PPV (Deng et al. 2010) πρόκειται για αλγορίθμους που χρησιμοποιούν μία δενδρική αναπαράσταση, πραγματοποιούν αναζήτηση κατά βάθος χρησιμοποιώντας τομές συνόλων κόμβων. Ο AprioriDP χρησιμοποιεί δυναμικό προγραμματισμό για τον περιορισμό των υποψήφιων συχνών στοιχειοσυνόλων. Τέλος, οι P-Mine (Baralis et al. 2013), LP-Growth (Pyun et al. 2014), Can-Mining (Jamsheela et al. 2015) και EXTRACT (Feddaoui et al. 2016) είναι εμπνευσμένοι από τον FP-Growth, με την έννοια ότι αυτός αναπαριστά ολόκληρη την βάση συναλλαγών σε μία συμπιεσμένη δενδρική δομή και δεν ασχολείται με την παραγωγή υποψήφιων συχνών στοιχειοσυνόλων.

### 2.1.7 Άλλες Εφαρμογές

Έχει ενδιαφέρον να δούμε άλλες εφαρμογές Εξόρυξης Συχνών Στοιχειοσυνόλων, πέραν της Ανάλυσης Καλαθιού Αγορών. Αυτό μπορεί να γίνει γενικεύοντας τις έννοιες “καλάθι” και “στοιχεία” ανάλογα την εφαρμογή.

## Κοινό περιεχόμενο

Μετά από κατάλληλη προεπεξεργασία κειμένων (απαλοιφή stop words ή/και πολύ συχνών λέξεων, stemming ή lemmatization, κ.α.), μπορούμε να ορίσουμε τα αρχεία σαν καλάθια και τις επεξεργασμένες λέξεις σαν στοιχεία. Έτσι, συχνά σύνολα λέξεων περιμένουμε να αντιπροσωπεύουν κοινό περιεχόμενο.

| | |
|---|---|
| **καλάθια** | αρχεία κειμένου (web pages, blogs, tweets) |
| **στοιχεία** | λέξεις |
| Αποτέλεσμα | συχνά σύνολα λέξεων → κοινό περιεχόμενο κειμένων |

Εικόνα 3.17: Εφαρμογή FIM για εύρεση κοινού περιεχομένου σε κείμενα.

Σημειώνεται ότι σε παρόμοιο συμπέρασμα μπορούμε να καταλήξουμε στην περίπτωση ιστοσελίδων, αν θεωρήσουμε αυτές σαν καλάθια και σαν στοιχεία τις ιστοσελίδες που περιέχουν κάποιο σύνδεσμο (link) προς αυτές.

## Λογοκλοπή

Παραμένωντας στην περίπτωση αρχείων κειμένου, μπορούμε να ορίσουμε τις προτάσεις σαν καλάθια και τα κείμενα σαν στοιχεία. Εδώ, ένα συχνό ζευγάρι κειμένων δηλώνει ότι αυτά μοιράζονται πολλές κοινές προτάσεις, κάτι που αποτελεί ένδειξη λογοκλοπής. Αξίζει να σημειωθεί ότι προτιμάται αυτή έναντι της αντίστροφης επιλογής καλαθιών και στοιχείων, ώστε ο αριθμός των καλαθιών να είναι πολύ μεγαλύτερος συγκριτικά με το μέγεθός τους.

| | |
|---|---|
| **καλάθια** | προτάσεις κειμένων |
| **στοιχεία** | κείμενα |
| Αποτέλεσμα | συχνά σύνολα κειμένων → κείμενα με πολλές κοινές προτάσεις |

Εικόνα 3.18: Εφαρμογή FIM για ανίχνευση λογοκλοπής σε κείμενα.

## Βιοδείκτες

Στην περίπτωση ασθενών για τους οποίους έχουμε στη διάθεσή μας διάφορους βιοδείκτες, μπορούμε να ορίσουμε σαν καλάθια τους ίδιους τους ασθενείς και σαν στοιχεία τις ασθένειες και τους βιοδείκτες τους. Ένα συχνό στοιχειοσύνολο αποτελούμενο από μία ασθένεια και κάποιους βιοδείκτες, θα μπορούσε να αποτελέσει έλεγχο για την ασθένεια αυτή.

| | |
|---|---|
| **καλάθια** | ασθενείς |
| **στοιχεία** | ασθένεια (-ες) και βιοδείκτες |
| Αποτέλεσμα | συχνά σύνολα μιας ασθένειας και κάποιων βιοδεικτών<br>↓<br>βιοδείκτες που πιθανώς συσχετίζονται με τη δεδομένη ασθένεια |

Εικόνα 3.19: Εφαρμογή FIM για εύρεση συσχέτισης βιοδεικτών με ασθένειες.

Σημειώνεται ότι αν αντί για ασθένειες και βιοδείκτες, είχαμε παρενέργειες και φάρμακα ασθενών, τότε ένα συχνό στοιχειοσύνολο με μια (περισσότερες) παρενέργεια (-ες) και κάποια φάρμακα, θα μας έδινε φάρμακα που σχετίζονται με κοινή παρενέργεια (-ες).

Τέλος, το FIM, με κάποιες παραλλαγές, έχει εφαρμοστεί σε έξυπνα συστήματα (Zhao et al. 2018), στην ταξινόμηση εικόνων (Patel et al. 2015) καθώς επίσης και σε ασφάλεια εικόνων (Swetha et al. 2018).

### 2.1.8 Μειονεκτήματα

Αξίζει να συνοψισθούν τα μειονεκτήματα της Εξόρυξης Κανόνων Συσχέτισης:

- η χρήση των κατωφλίων οδηγεί στην αγνόηση στοιχειοσυνόλων και κανόνων που ενδέχεται να έφεραν ουσία
- η υποκειμενικότητα στην αξιολόγηση των κανόνων μέσω των ποσοτικών μεγεθών καθιστά μη δυνατή την ύπαρξη μιας, κοινώς αποδεκτής, διαδικασίας εύρεσης των πιο ενδιαφερόντων κανόνων
- η μη αξιοποίηση της πληροφορίας των ποσοτήτων αλλά και του καταναλωτή που σχετίζονται με κάθε καλάθι προϊόντων

## 2.2 Ανίχνευση Κοινοτήτων

Η Αναζήτηση Κοινοτήτων (Community Detection) είναι η διαδικασία οργάνωσης των κορυφών $V$ ενός γράφου $G(V,E)$ σε ισχυρές κοινότητες. Ορίζουμε σαν $C = \{C_1, C_2, \dots, C_k\}$ μία διαμέριση του $V$ και ονομάζουμε το $C$ συσταδοποίηση του $G$ και $C_i$, που πρέπει να είναι μη κενά, τις κοινότητες ή συστάδες. Συχνά, προσπαθούμε να κατασκευάσουμε το $C$ μεγιστοποιώντας μια αντικειμενική συνάρτηση $f(G)$.

Η ακριβής λύση του προβλήματος μεγιστοποίησης του modularily ενός γράφου είναι NP-hard[1] πρόβλημα, κάτι που ενισχύει την ανάδειξη προσεγγιστικών και ευρετικών μεθόδων, όπως είναι η άπληστη Louvain μέθοδος. Συνοπτικά, έχουν αναπτυχθεί και άλλες ευρετικές μέθοδοι βελτιστοποίησης του Modularity όπως συσσωρευτικές που αναδρομικά συγχωνεύουν παρόμοιες κορυφές-κοινότητες, διαιρετικές (divisive) που εντοπίζουν και διαγράφουν ακμές μεταξύ των κοινοτήτων όπως επίσης και η προσομοιωμένη ανόπτηση (simulated annealing) που μπορεί να ερμηνευθεί σαν αργή μείωση της πιθανότητας αποδοχής “χειρότερων” λύσεων καθώς εξερευνάται ο συνολικός χώρος λύσεων.

## 2.2.1 Ορισμοί

Έστω

$G(V, E)$ ένας μη-κατευθυνόμενος γράφος

Τότε

- Διαμέριση $C = \{C_1, C_2, \dots, C_k\}$ του $V$ είναι η διάσπασή του σε k αμοιβαία αποκλειόμενες ομάδες, τις κοινότητες (4.1)
- Ενδοκοινοτικές ακμές μιας κοινότητας $i$, είναι οι ακμές που έχουν άκρα στην ίδια κοινότητα $C_i$ (4.2)
- Διακοινοτικές ακμές των κοινοτήτων $i$ και $j$ είναι οι ακμές που έχουν το ένα τους άκρο στο $i$ και το άλλο άκρο στο $j$ (4.3)
- Modularity $Q$ είναι ένας δείκτης ποιότητας της διαμέρισης $C$ και ορίζεται σαν

$$Q(C) = \sum_{i=1}^{k} (e_{ii} - a_i^2) \quad (4.4)$$

όπου

$e_{ii}$ : ποσοστό των ακμών που έχουν και τα δύο τους άκρα στην κοινότητα $i$

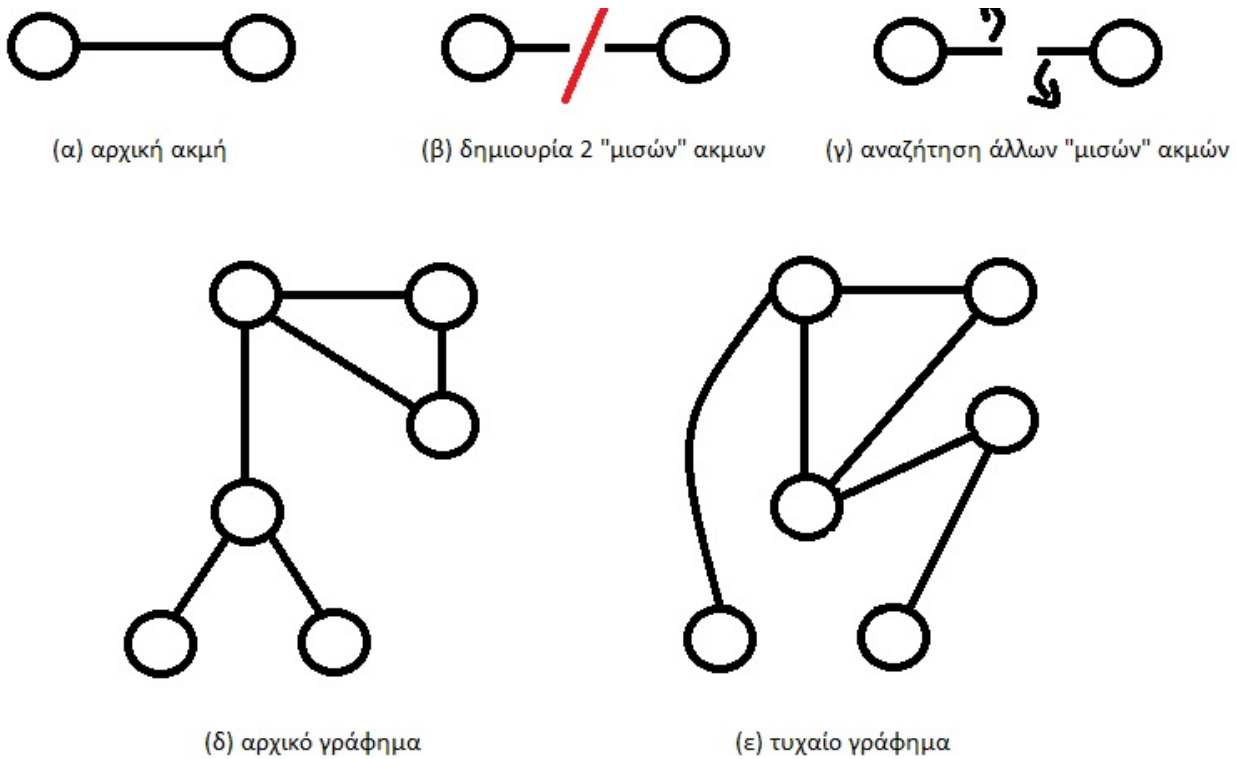
$a_i$ : ποσοστό των άκρων των ακμών που αντιστοιχούν σε κόμβους της κοινότητας $i$

Το μέγεθος Modularity έχει εύρος τιμών το διάστημα [-1,1] και υπολογίζει την πυκνότητα των ενδοκοινοτικών ακμών σε σχέση με αυτή των διακοινοτικών. Θετικές τιμές αν ο αριθμός των ακμών της πρώτης κατηγορίας υπερβαίνει τον αναμενόμενο αριθμό με βάση την τύχη, δηλαδή ενός τυχαίου γραφήματος με την ίδια κατανομή βαθμών των ακμών. Με απλά λόγια, η κατασκευή ενός τυχαίου γραφήματος με την παραπάνω ιδιότητα, θα μπορούσε να γίνει ως εξής.

Έστω μια ακμή $(u, v)$ ϵ $E$. Κόβουμε την ακμή στη μέση και έτσι οι κόμβοι $u$ και $v$ έχουν μείνει με μία “μισή” ακμή ο καθένας. Κάνοντας ταυτόχρονα το ίδιο πράγμα σε όλες τις ακμές, έστω $m$ αρχικά, καταλήγουμε σε $2 * m$ “μισές” ακμές. Τώρα, γυρνώντας ταυτόχρονα και με

[1] https://perso.uclouvain.be/vincent.blondel/research/louvain.html

τυχαίο τρόπο τις “μισές” αυτές ακμές, ενώνουμε ένα ζευγάρι “μισών” ακμών όταν αυτές ευθυγραμμίζονται. Έτσι, υπό τον περιορισμό την ενοποίησης όλων των “μισών” ακμών, προκύπτει ένα τυχαίο γράφημα όπου κάθε κόμβος διατηρεί τον αρχικό βαθμό ακμών του.




Εικόνα 4.1: Παράδειγμα παραγωγής τυχαίου γραφήματος με την ίδια κατανομή βαθμών.

Η σχέση 4.4 φανερώνει έναν έμφυτο συμβιβασμό: η μεγιστοποίηση του πρώτου όρου απαιτεί πολλές ακμές μέσα στις κοινότητες, ενώ αυτή του δεύτερου απαιτεί τη διαμέριση του γραφήματος σε πολλές κοινότητες με μικρό συνολικό άθροισμα βαθμών η καθεμία. Διαισθητικά, μια διαμέριση $C$ του $V$ κατά την οποία υπάρχουν πολλές ενδοκοινοτικές ακμές και άρα σχετικά λίγες διακοινοτικές, αναπαριστά μία ισχυρή δομή κοινοτήτων.

Χρειάζεται, ωστόσο, να μελετήσουμε τη περίπτωση ύπαρξης βαρών στις ακμές.

- Έστω $G(V, E, w)$ ένας μη-κατευθυνόμενος γράφος με βάρη στις ακμές του
- Ορίζουμε $C = \{C_1, C_2, \dots, C_k\}$ μία διαμέριση του $V$ σε k κοινότητες

Τώρα, το Modularity ορίζεται σαν

$$Q(C) = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i * k_j}{2m}) \, \delta(c_i, c_j) \qquad (4.5)$$

όπου

$A_{ij}$ : το βάρος της ακμής μεταξύ των κόμβων $i, j$

$m$: το άθροισμα των βαρών όλων των ακμών, ίσο με $\frac{1}{2}\sum_{i,j} A_{ij}$

$c_i$: η κοινότητα που ανήκει ο κόμβος $i$

$k_i$: το άθροισμα των βαρών των ακμών με ένα άκρο τον κόμβο $i$, ίσο με $\sum_j A_{ij}$

$\delta(c_i, c_j)$: 1 όταν $c_i = c_j$ , αλλιώς 0

## 2.2.2 Αλγόριθμος Louvain

Ο αλγόριθμος Louvain (Blondel et al. 2008) αναπτύσσεται σε δύο φάσεις τις οποίες υλοποιεί επαναληπτικα.

### 2.2.2.1 Ψευδοκώδικας

#### Πρώτη φάση

```
while true
   ∀ κόμβο i
      ∀ γειτονικό κόμβο j
         Βρες το κέρδος στο Total Modularity που θα έδινε
         η μετακίνηση του i στην κοινότητα c_j

       αν το μέγιστο κέρδος είναι θετικό
       τότε εφάρμοσε την αντίστοιχη μετακίνηση


    Αν δεν πραγματοποιήθηκε καμία μετακίνηση στο πρώτο πέρασμα
    Τότε Τέλος

    Αν δεν πραγματοποιήθηκε καμία μετακίνηση στο τρέχων πέρασμα
    Τότε πήγαινε στη δεύτερη φάση
```

#### Δεύτερη φάση

```
Κατασκεύασε ένα νέο δίκτυο όπου:
     οι νέοι κόμβοι είναι οι διαμορφωμένες κοινότητες της πρώτης φάσης

     τα βάρη των ακμών μεταξύ των νέων κόμβων είναι ίσα με το άθροισμα των
     παλιών ακμών μεταξύ των αντίστοιχων κοινοτήτων
```

Εικόνα 4.2: Ο αλγόριθμος Louvain σε ψευδοκώδικα.

### 2.2.2.2 Παρατηρήσεις

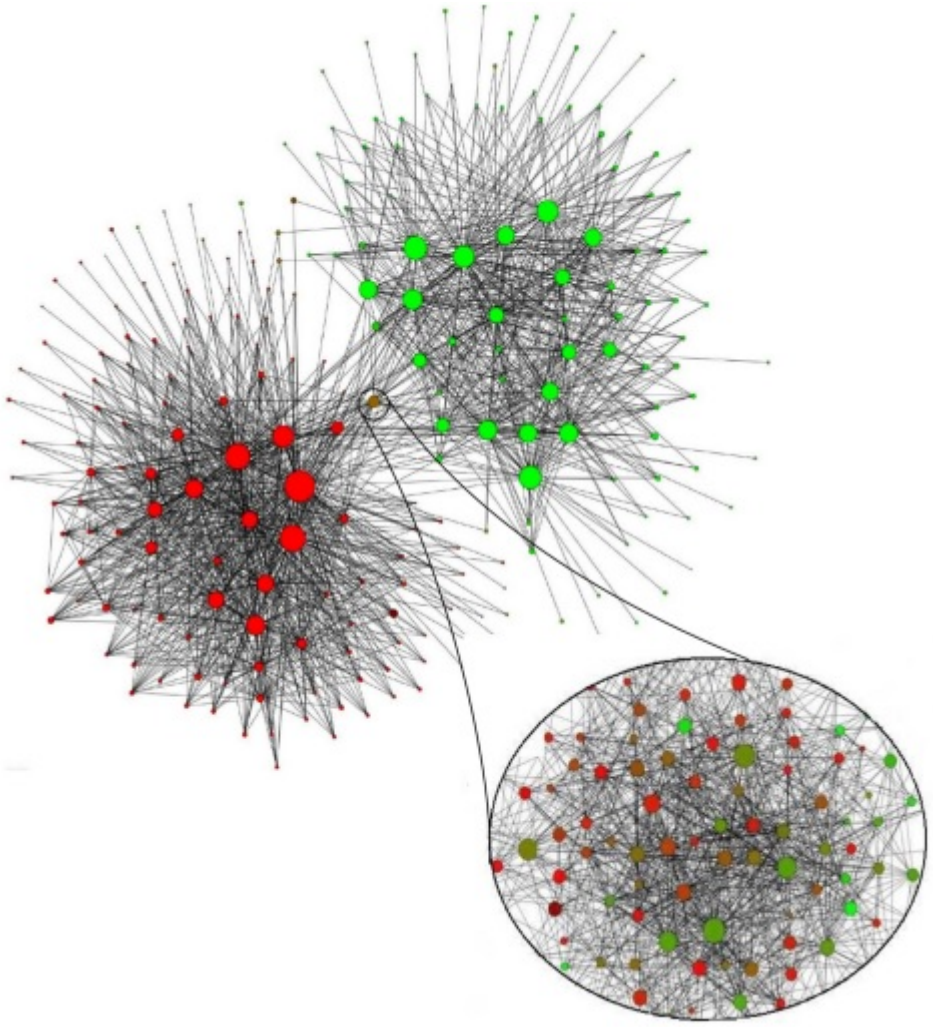
Αρχικά, κάθε κόμβος αποτελεί μια ξεχωριστή κοινότητα. Κατά την πρώτη φάση περνάμε έναν-έναν τους υπάρχοντες κόμβους (κοινότητες) και ψάχνουμε για μετακινήσεις αυτών που συνεισφέρουν στη συνολική τιμή του $Q$. Ένα σημείο κλειδί είναι ο γρήγορος υπολογισμός της διαφοράς $\Delta Q$ που προκύπτει από' τη μετακίνηση ενός κόμβου μιας κοινότητας σε μία άλλη.

Επίσης, όταν η μέγιστη συνεισφορά από τη μετακίνηση ενός κόμβου μπορεί να προκύψει από την εισαγωγή του σε διαφορετικές κοινότητες, τότε μπορεί να εφαρμοστεί κάποιος κανόνας για να σπάει τις ισοπαλίες αυτές, όπως για παράδειγμα μετακίνηση στην κοινότητα με τον μικρότερο αριθμό κόμβων. Τέλος, η μετάβαση από την πρώτη στη δεύτερη φάση σηματοδοτεί ένα τοπικό μέγιστο στο συνολικό Modularity, εφόσον δεν μπορούμε να επωφεληθούμε από καμία μετακίνηση κόμβου σε κάποια άλλη κοινότητα.

Στη δεύτερη φάση, συγχωνεύουμε τις ομάδες κόμβων που προέκυψαν στην πρώτη, διαμορφώνοντας έτσι τις νέες κοινότητες. Αξίζει να σημειωθεί ότι ακμές μεταξύ κόμβων που ανήκουν στις ίδιες ομάδες της πρώτης φάση, οδηγούν σε αυτο-ακμές (self-loops) των συγχωνευμένων κόμβων και ότι οι δύο αυτές φάσεις εκτελούνται επαναληπτικά. Τέλος, ο αλγόριθμος αυτός έχει σαν αποτέλεσμα μια πλήρως ιεραρχική δομή των κοινοτήτων και όχι απλά τις τελικές κοινότητες. Αυτό συμβαίνει χάρη στην επαναληπτική συγχώνευση κόμβων και έτσι έχουμε πρόσβαση σε διαφορετικά επίπεδα των κοινοτήτων όπως φαίνεται παρακάτω.

Επιπλέον, ο αλγόριθμος φαίνεται να τρέχει σε $O(n * logn)$[2], περνώντας τον περισσότερο χρόνο στις πρώτες επαναλήψεις. Αυτό, αφενός είναι αποτέλεσμα του γρήγορου υπολογισμού της μεταβολής του συνολικού Modularity, αφετέρου του ότι ο αριθμός των κοινοτήτων μειώνεται δραστικά μετά από μόλις μερικές συγχωνεύσεις της δεύτερης φάσης. Το τελευταίο έχει σαν αποτέλεσμα ο περισσότερος χρόνος εκτέλεσης να συγκεντρώνεται στις πρώτες επαναλήψεις.

[2] https://perso.uclouvain.be/vincent.blondel/research/louvain.html



Εικόνα 4.3: Μεγέθυνση σε υψηλότερη ανάλυση για ανάδειξη υπο-κοινοτήτων.

### 2.2.2.3 Μειονεκτήματα Αλγορίθμου

Οι αλγόριθμοι μεγιστοποίησης του Modularity παρουσιάζουν το πρόβλημα της ανάλυσης (resolution limit problem, (Fortunato et al. 2007), δηλαδή αδυνατούν να ανιχνεύσουν κοινότητες μικρής κλίμακας. Αυτό ισχύει μερικώς στον αλγόριθμο Louvain καθώς η πιθανότητα μετακίνησης, στην ίδια επανάληψη της πρώτης φάσης, όλων των κόμβων μιας κοινότητας σε μία άλλη, είναι πολύ μικρή. Έτσι, αυτές οι κοινότητες πιθανότατα θα συγχωνευτούν σε επόμενα βήματα, όπου πλέον κομμάτια της κάθε μίας θα έχουν ήδη ενωθεί και αυτό είναι ορατό σε μάς, χάρη στην ιεραρχία που προσφέρει ο αλγόριθμος.

### 2.2.2.4 Άλλες Εφαρμογές

Είναι ενδιαφέρον το πόσο δημοφιλής είναι ο αλγόριθμος Louvain. Συγκεκριμένα, χρησιμοποιήθηκε από εταιρείες όπως Twitter, LinkedIn, Flickr καθώς επίσης σε εφαρμογές[3] Δικτύων Κινητής Τηλεφωνίας, Δικτύων Παραπομπών, Δικτύων Ανθρώπινου Εγκεφάλου. Συνεπώς, ο αλγόριθμος Louvain, καθώς και γενικότερα η Ανίχνευση Κοινοτήτων, έχει εφαρμοστεί σε κοινωνικά, βιολογικά και τηλεφωνικά δίκτυα, όπως επίσης σε δίκτυα ιστοσελίδων και άρθρων.

[3] https://perso.uclouvain.be/vincent.blondel/research/louvain.html

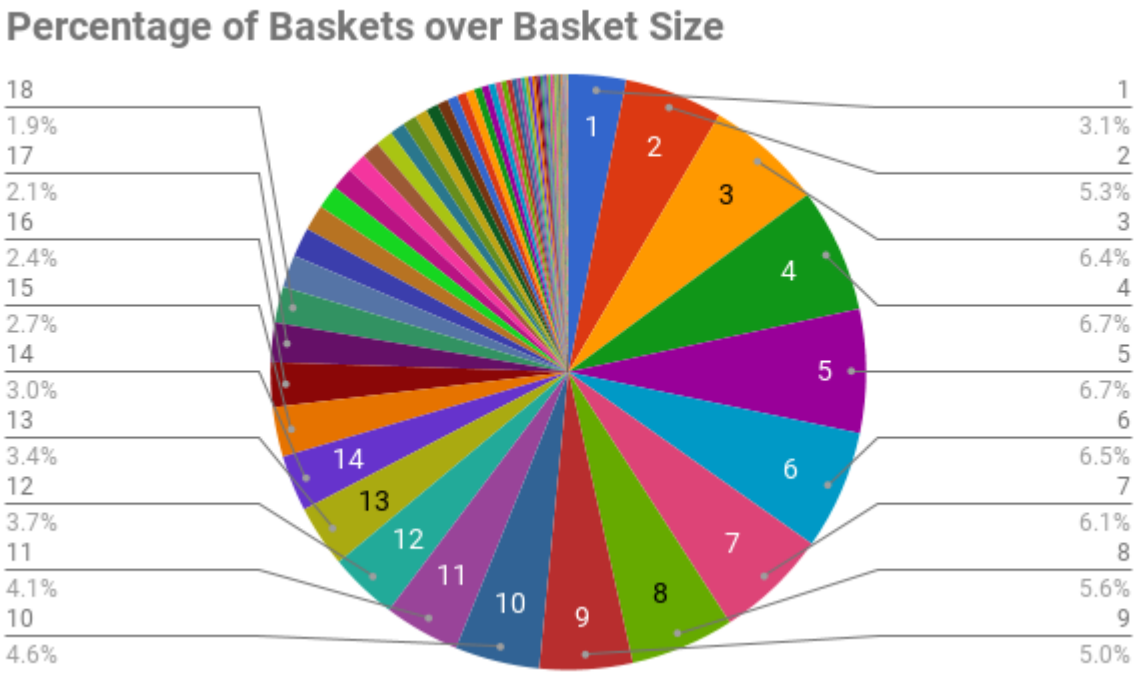
# 3 Πειράματα

## 3.1 Δεδομένα

Οι προαναφερθείσες τεχνικές εφαρμόστηκαν σε δεδομένα αγορών από supermarket. Οι διαθέσιμες πληροφορίες περιλαμβάνουν το αναγνωριστικό και την ποσότητα των προϊόντων που περιλαμβάνονται σε κάθε αγορά, καθώς και την χρονική στιγμή κατα την οποία αυτές πραγματοποιήθηκαν. Επίσης, για το υποσύνολο των συναλλαγών στις οποίες χρησιμοποιήθηκε κάρτα μέλους, έχουμε διαθέσιμο το αναγνωριστικό των καταναλωτών. Τέλος, τα προϊόντα είναι κατηγοριοποιημένα σε ομάδες διαφόρων επιπέδων όπως παρουσιάζεται στην παρακάτω εικόνα.

### 3.1.1 Χαρακτηριστικά

Συγκεκριμένα, τα δεδομένα συναλλαγών αφορούν αγορές που πραγματοποιήθηκαν στη διάρκεια ενός χρόνου, περιέχουν συνολικά σχεδόν 475.000 αγορές και 22.000 διαφορετικά προϊόντα και 23.000 αναγνωριστικά καταναλωτών.




Εικόνα 5.1: Κατανομή μεγέθους των καλαθιών

Όπως μπορούμε να παρατηρήσουμε, το πολύ 5 διαφορετικά προϊόντα περιλαμβάνονται στο 25% των καλαθιών, 9 στο 50% ενώ το πολύ 16 στο 75%. Επίσης, το μεγαλύτερο καλάθι περιλαμβάνει 158 προϊόντα και το μέσο καλάθι 13.

### 3.1.2 Αφαιρετικότητα

Η κυκλοφορία ενός προϊόντος από πολλές εταιρείες, οδηγεί σε αλλοίωση των αποτελεσμάτων κάποιων τεχνικών. Ενδεικτικά, στην Εξόρυξη Κανόνων Συσχέτισης, έστω $A, B$ δύο προϊόντα, από τα οποία το πρώτο κυκλοφορεί σε μία μάρκα, ενώ το δεύτερο σε $n$ ( $B_1, B_2, \dots, B_n$) . Η υποστήριξη (confidence) για τον κανόνα $A \rightarrow B$ είναι ίση με

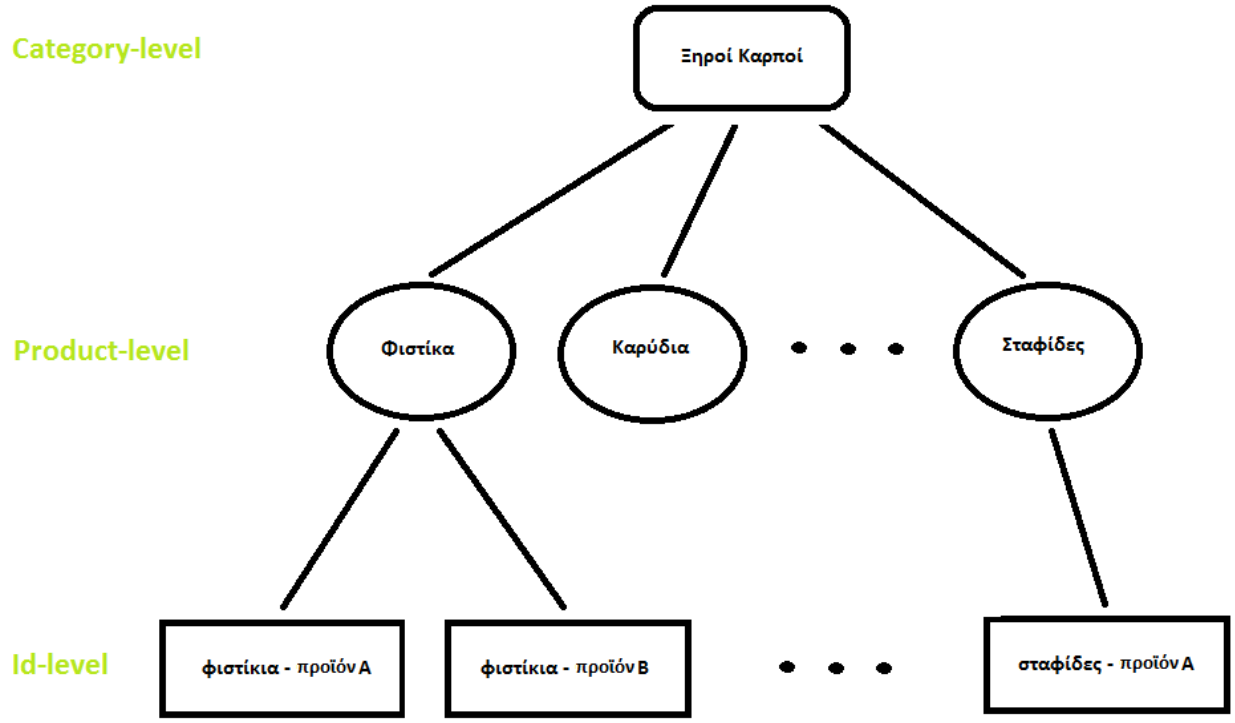
$$c(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)} = \kappa$$

Επιπλέον, χωρίς βλάβη της γενικότητας, έστω τα προϊόντα $B_i$ έχουν αγοραστεί ισόποσα μαζί με το $A$, τότε για τους κανόνες $A \rightarrow B_i$ θα ισχύει

$$c(A \rightarrow B_i) = \frac{\sigma(A \rightarrow B_i)}{\sigma(A)} = \frac{\frac{\sigma(A \cup B)}{n}}{\sigma(A)} = \frac{\kappa}{n}$$

Είναι φανερό πως, στη δεύτερη περίπτωση, οι κανόνες είναι αισθητά αποδυναμωμένοι. Για τον λόγο αυτόν, καθώς και για την βελτίωση της ποιότητας των αποτελεσμάτων, όπως αναλύουμε στη συνέχεια, προχωρήσαμε σε ομαδοποίηση των προϊόντων.

Η επιλογή μας, σχετικά με την αφαιρετικότητα των προϊόντων, συνίσταται σε τρία επίπεδα. Το κατώτερο αφορά στο αναγνωριστικό των προϊόντων (**Id-level**), το οποίο ακολουθείται από το επίπεδο προϊόντος (**Product-level**) και το επίπεδο κατηγορίας προϊόντων (**Category-level**). Σε αυτή την κατεύθυνση, στο **Product-level** προσπαθήσαμε να εξαλείψουμε τις διαφορετικές εκδοχές των ίδιων προϊόντων. Συγκεκριμένα, αντιμετωπίζουμε ως ίδιο ένα προϊόν που κυκλοφορεί από διαφορετικές εταιρείες, σε διαφορετικά μεγέθη ή βρίσκεται σε προσφορά.




Εικόνα 5.2: Τα επίπεδα αφαιρετικότητας των προϊόντων.

Είναι γεγονός ότι κάποιες ομάδες καταναλωτών, όπως οι εργένηδες, ενδέχεται να προτιμούν προϊόντα μικρότερων συσκευασιών ή συγκεκριμένων εταιρειών. Με την παραπάνω επιλογή, η εξαγωγή συμπερασμάτων βάσει των παραπάνω χαρακτηριστικών μπορεί να γίνει μόνο σε δεύτερο στάδιο. Για παράδειγμα, πρώτα θα βρούμε τη σχέση μεταξύ γάλακτος και δημητριακών και στη συνέχεια, χρησιμοποιώντας το αφαιρετικό επίπεδο **Size-level,** θα μπορούσαμε να ερευνήσουμε την το αν όντως υπάρχει σχέση μεταξύ μικρών συσκευασιών γάλακτος και δημητριακών.

Αξίζει να σημειωθεί ότι οι παραπάνω πληροφορίες δεν αξιοποιούνται από κάθε τεχνική. Για παράδειγμα, η FIM δεν χρησιμοποιεί ούτε τις ποσότητες των προϊόντων ούτε το αναγνωριστικό του καταναλωτή, ωστόσο, όλες οι τεχνικές μπορούν να εφαρμοστούν σε οποιαδήποτε ομαδοποίηση προϊόντων. Κατά την παρουσίαση των αποτελεσμάτων κάθε τεχνικής, αναφέρουμε το αφαιρετικό επίπεδο των δεδομένων που χρησιμοποιήθηκε είτε λόγω της ίδιας της μεθόδου, είτε από δική μας επιλογή με σκοπό την καλύτερη ποιότητα αποτελεσμάτων.

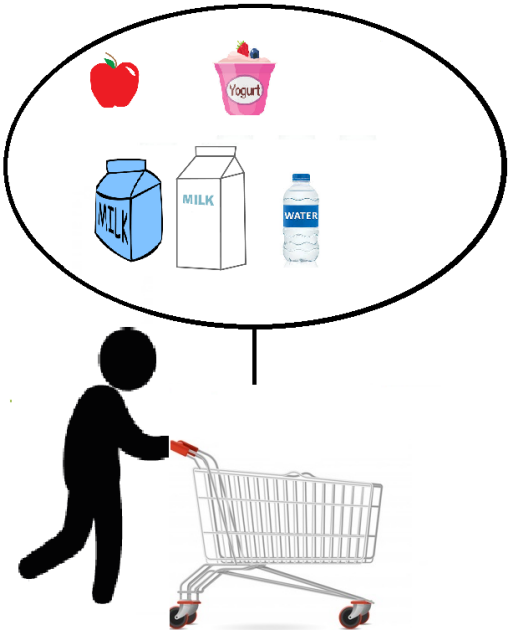
Τέλος, υπάρχουν περιπτώσεις όπως τα βιολογικά προϊόντα και τα προϊόντα διαίτης, τα οποία είναι δημοφιλή σε ένα μεγάλο εύρος κατηγοριών, αλλά και άλλες που έχουν εφαρμογή σε μεμονωμένες κατηγορίες, όπως τα γάλατα μακράς διαρκείας, ομαδοποιήθηκαν ξεχωριστά. Αυτό αποσκοπεί στην ανάδειξη πιο σύνθετων σχέσεων. Για παράδειγμα, όσοι αγοράζουν βιολογικά φρούτα και λαχανικά, είναι πιθανόν να προτιμούν βιολογικά προϊόντα και σε άλλες οικογένειες

προϊόντων, κάτι που περιμένουμε να αποτυπωθεί στα αποτελέσματα των τεχνικών. Ωστόσο, κάτι τέτοιο εισάγει υποκειμενικότητα στην ανάλυσή μας, αφού για παράδειγμα δεν υπάρχει σαφές όριο επιπέδου λιπαρών που πρέπει να έχει ένα τυρί για να θεωρηθεί light. Επίσης, υποκειμενικότητα υπάρχει και σε φαινομενικά απλούστερες περιπτώσεις, όπως τα εμφιαλωμένα νερά που κυκλοφορούν τόσο σαν “επιτραπέζια νερά” όσο και σαν “νερά πηγής”.

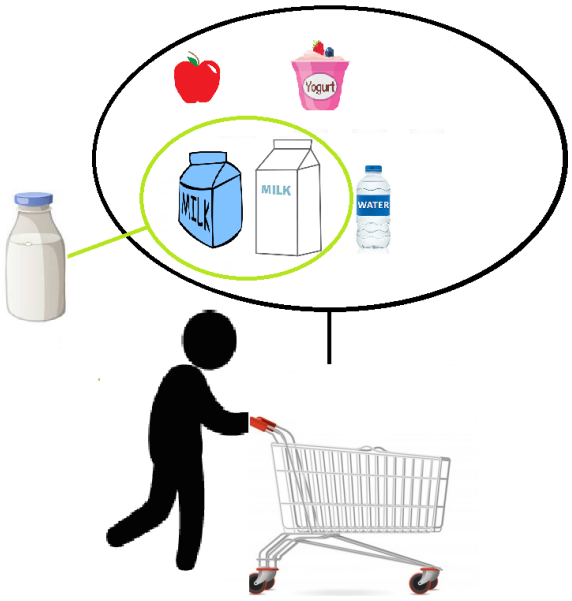
Το συμπέρασμά μας, μέσω της εποπτείας των αποτελεσμάτων, είναι ότι η ομαδοποίηση προϊόντων ήταν ωφέλιμη σε όλες τι τεχνικές. Είναι κρίσιμο να επισημανθεί πως η εκάστοτε επιλογή επιπέδου ομαδοποίησης, πρέπει να καθορίζεται από το επίπεδο στο οποίο επιθυμούμε να αναλύσουμε τα προϊόντα. Στα πλαίσια της διπλωματικής, επιλέξαμε την εξάλειψη των εταιρειών καθώς και ανάδειξη ειδικών περιπτώσεων προϊόντων όπως τα βιολογικά, τα light, τα ολικής άλεσης καθώς και κάποιες άλλες

Αρχική Πληροφορία


Μη αξιοποίηση ποσοτήτων



Ομαδοποίηση Προϊόντων (Product-level)




Εικόνα 5.3: Παράδειγμα χρήσης των διαφορετικών αφαιρετικών επιπέδων των προϊόντων. Στην πρώτη περίπτωση, έχουμε διαθέσιμο το αναγνωριστικό του πελάτη, τα προϊόντα που περιλαμβάνονται στο καλάθι, καθώς και τις ποσότητές τους. Στη δεύτερη, δεν έχουμε το αναγνωριστικό του πελάτη ούτε τις ποσότητες των προϊόντων. Στην τρίτη, ομαδοποιούμε διαφορετικά προϊόντα γάλακτος σε ένα.

# 3.2 Αποτελέσματα

## 3.2.1 Κανόνες Συσχέτισης

Όπως αναφέρθηκε παραπάνω, η Εξόρυξη Κανόνων Συσχέτισης $Rules(T, s, c)$ είναι διαδικασία που τυπικά ορίζεται από μία βάση συναλλαγών T, ένα κατώφλι υποστήριξης (support threshold) s και ένα κατώφλι εμπιστοσύνης (confidence threshold) c. Η παραγωγή κανόνων της μορφής $A \rightarrow B$ έγινε σε 2 βήματα. Στο πρώτο βρήκαμε τα συχνά στοιχειοσύνολα, δηλαδή αυτά που ικανοποιούν το κατώφλι support, και στο δεύτερο παράξαμε, από τα συχνά, κανόνες της παραπάνω μορφής με $|B| = 1$ .

### 3.2.1.1 Εξόρυξη Συχνών Στοιχειοσυνόλων και Κανόνων Συσχέτισης

Στα πλαίσια της Εξόρυξης Συχνών Στοιχειοσυνόλων υλοποιήθηκαν, απ' την αρχή, οι αλγόριθμοι Apriori, PCY, Multistage, Multihash και Toivonen. Καθένας από αυτούς τους

αλγορίθμους έχει σαν έξοδο τα συχνά στοιχειοσύνολα, καθώς και τον αριθμό εμφάνισής τους μέσα στην βάση συναλλαγών. Παρακάτω παρουσιάζονται τα 15 πιο συχνά προϊόντα ( σε **Id-level** και **Product-level)** με τον αριθμό εμφάνισής τους στο σύνολο των καλαθιών**.**

| | |
|---|---|
| ΜΠΑΝΑΝΕΣ DOLE ΕΙΣΑΓΩΓΗΣ CAVENDISH | 8.87% |
| ΖΑΧΑΡΗ ΛΕΥΚΗ ΚΡΥΣ/ΚΗ | 6.82% |
| ΤΟΜΑΤΕΣ ΧΥΜΑ ΕΛΛΗΝΙΚΕΣ | 6.65% |
| ΛΟΥΜΙΔΗΣ ΚΑΦΕΣ ΠΑΠΑΓ.ΠΑΡΑΔΟΣ.194GR | 4.5% |
| GOUDA ΓΕΡΜΑΝΙΑΣ ΦΡΑΤΖΟΛΑ | 4.4% |
| ΑΓΓΟΥΡΙΑ ΤΕΜΑΧΙΟ ΕΛΛΗΝΙΚΑ | 4.35% |
| ΜΠΑΝΑΝΕΣ DULCE BUONA ΕΙΣΑΓΩΓΗΣ | 4.22% |
| ΧΑΡΤ/ΤΕΣ ΛΕΥΚΕΣ 28Χ30 80Φ | 3.86% |
| ΧΑΡΤΙ ΚΟΥΖΙΝΑΣ 500GR | 3.84% |
| ΨΩΜΙ ΧΩΡΙΑΤΙΚΟ ΚΛΑΣΙΚΟ | 3.51% |
| ΜΩΡΟΜΑΝΤΗΛΑ REFILL 80ΤΕΜ | 3.24% |
| ΝΕΡΟ ΖΑΓΟΡΙ 1.5LT ΦΥΣΙΚΟ 5+1ΔΩΡΟ | 3.22% |
| ΚΑΡΑΜ.ΤΟΣΤ ΣΤΑΡΕΝΙΟ 680G 0.60Ε ΦΘΗΝ | 3.1% |
| ΤΥΡΙ MILNER ΟΛΛΑΝΔΙΑΣ ΦΡΑΝΤΖΟΛΑ 17% | 3.03% |
| ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΑΓΓΕΛΑΚΗΣ | 2.96% |

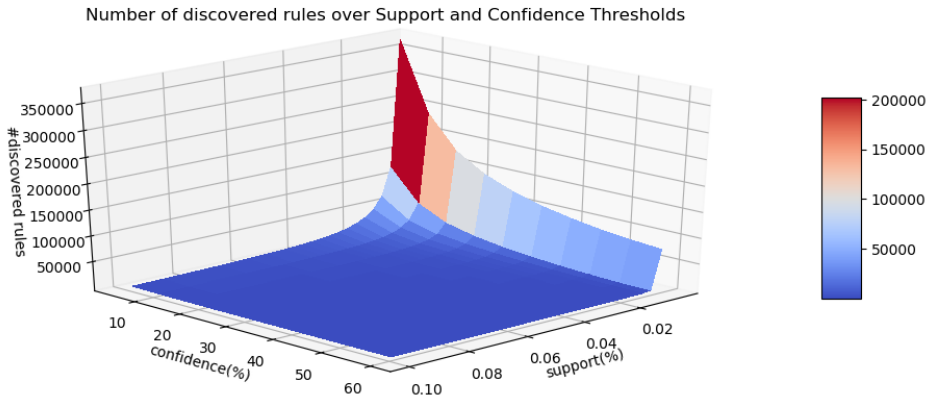
α) **Id-level**

| | |
|---|---|
| ΨΩΜΙΑ ΓΙΑ ΤΟΣΤ | 19.48% |
| ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | 16.3% |
| ΝΕΡΑ | 15.69% |
| ΤΥΠΟΥ COLA | 13.39% |
| TABLETS ΣΟΚΟΛΑΤΕΣ ΑΠΛΕΣ | 13.17% |
| ΜΠΑΝΑΝΕΣ | 13.1% |
| ΕΛΛΗΝΙΚΟΣ ΚΑΦΕΣ | 12.14% |
| GOUDA ΣΕ ΦΕΤΕΣ | 11.73% |
| ΤΟΜΑΤΟΠΟΛΤΟΙ | 11.42% |
| ΣΠΑΓΓΕΤΙ | 11.14% |
| ΦΡΥΓΑΝΙΕΣ | 11.01% |
| ΥΨΗΛΗΣ ΠΑΣΤΕΡΙΩΣΗΣ ΓΑΛΑ ΛΕΥΚΟ | 10.7% |
| ΧΑΡΤΟΠΕΤΣΕΤΕΣ | 10.31% |
| ΑΥΓΑ ΚΟΤΑΣ | 9.89% |
| ΦΡΕΣΚΟ ΓΑΛΑ ΛΕΥΚΟ | 9.79% |

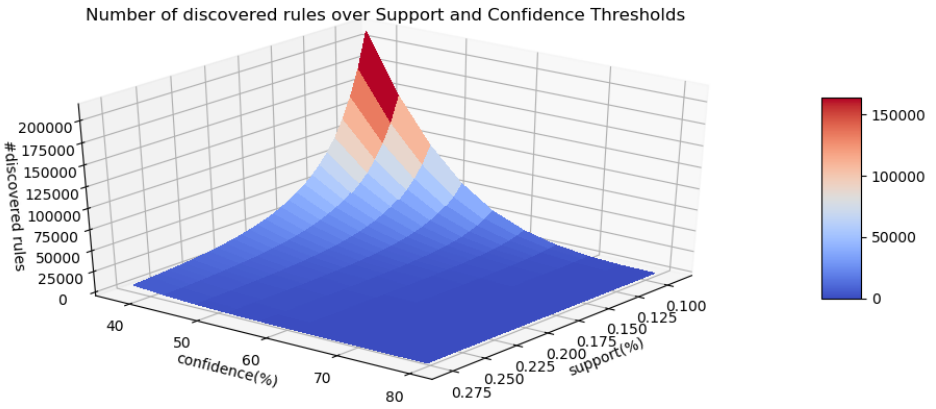
β) **Product-level**

Εικόνα 5.4: Τα 15 δημοφιλέστερα προϊόντα σε **Id-level** και **Product-level**

Στη συνέχεια, από κάθε “συχνό” στοιχειοσύνολο $X$ με $|X| = N > 1$, παράγουμε και τους $N$ κανόνες της μορφή $X \setminus \{x\} \rightarrow x$ , $\forall\, x \in X$. Παρακάτω παρουσιάζουμε τον αριθμό των συνολικών κανόνων σαν συνάρτηση των κατωφλίων υποστήριξης και εμπιστοσύνης.




α) Id-level




β) Product-level

Εικόνα 5.5: Συνολικός αριθμός Κανόνων Συσχέτισης συναρτήσει των κατωφλίων υποστήριξης (support) και εμπιστοσύνης (confidence).

Είναι εύκολο να παρατηρήσει κανείς ότι καθώς μεγαλώνουν οι τιμές των δύο κατωφλίων είτε ξεχωριστά είτε μαζί, ο αριθμός των ευρεθέντων κανόνων συσχέτισης ελαττώνεται με γρήγορους ρυθμούς. Επίσης, για λογικές τιμές κατωφλίων ο αριθμός των κανόνων είναι μη διαχειρίσιμος για κάποιον που θέλει να αναλύσει τα αποτελέσματα. Έτσι, χρησιμοποιήσαμε τα

ποσοτικά μεγέθη, που αναλύθηκαν προηγουμένως, με σκοπό να βαθμολογήσουμε τους κανόνες και προκύψουν οι καλύτεροι.

Στο σημείο αυτό, χρησιμοποιήσαμε το επίπεδο **Product-level** και ορίσαμε support threshold ίσο με 300 (0,07%) και confidence threshold ίσο με 0.1. Ως αποτέλεσμα είχαμε περίπου 1.400.000 κανόνες και τον πίνακα συνάφειας καθενός εξ αυτών.Τέλος, για κάθε κανόνα υπολογίσαμε αφενός τις αξιολογήσεις των 6 αντικειμενικών μέτρων, αφετέρου αποθηκεύσαμε τα διαφορετικά **Product** και **Category levels** που συμμετέχουν είτε στο μέλος $A$ είτε στο μέλος $B$.

### 3.2.1.2 Αξιολόγηση Κανόνων Συσχέτισης

Η εξαγωγή κανόνων συσχέτισης $A \rightarrow B$ είναι μία διαδικασία που μπορεί να πραγματοποιηθεί κάτω από διάφορες συνθήκες. Για παράδειγμα, αν ένα supermarket πρόκειται να διακόψει τη συνεργασία του με κάποιον προμηθευτή, είναι χρήσιμο να βρει τα προϊόντα των οποίων οι πωλήσεις ενδέχεται να επηρεαστούν. Κάτι τέτοιο είναι δυνατό να προκύψει από κανόνες που έχουν σαν μέλη του $A$, τα προϊόντα του προμηθευτή. Επίσης, η αύξηση των πωλήσεων του προϊόντος $y$, θα μπορούσε να σχεδιαστεί μέσω των κανόνων $A \rightarrow y$.

Μια ακόμα ενδιαφέρουσα περίπτωση αποτελεί αυτή των κανόνων της μορφής $A \rightarrow y$, όπου $x \in A$. Κάτι τέτοιο, θα υποδείκνυε στο supermarket τα προϊόντα που πρέπει να πωληθούν μαζί με το $x$ ώστε να ευνοηθεί η πώληση του $y$. Τέλος, θα ήταν ουσιώδης η εξαγωγή κανόνων μεταξύ προϊόντων που ανήκουν σε συγκεκριμένες κατηγορίες, όπως για παράδειγμα γαλακτοκομικών και κρεατικών.

Οι παραπάνω περιπτώσεις χρήσης μας οδήγησαν στην παραμετρική υλοποίηση εξαγωγής κανόνων συσχέτισης. Οι κύριες παράμετροι αφορούν στη δυνατότητα επιλογής προϊόντων-στόχων σε κάθε αφαιρετικό επίπεδο (**Product-level, Category-level)** και ενός ή περισσοτέρων αντικειμενικών μέτρων για την αξιολόγηση των κανόνων. Στην περίπτωση επιλογής ενός μέτρου πραγματοποιείται ταξινόμηση ως προς αυτό, ενώ η επιλογή περισσότερων οδηγεί στην χρησιμοποίηση της πολυκριτηριακής μεθόδου Pareto.

### 3.2.1.3 Περιπτώσεις Χρήσης

#### 3.2.1.3.1 Κατηγορία 1

Κανόνες $A \rightarrow B$ με δεδομένο προϊόν-στόχο ως $B$.

Στην περίπτωση αυτή, έχουμε σαν αποτέλεσμα προϊόντα ή συνδυασμούς αυτών (σύνολο $A$) των οποίων η αγορά, ευνοεί αυτή του $B$. Ένα supermarket θα μπορούσε να προβεί στο εξής πλάνο. Αφού επιλεγεί ο καλύτερος κανόνας, μπορεί το προϊόν $A$ να διαφημιστεί σε προσφορά, ενώ ταυτόχρονα το $B$ να υποστεί μικρή αύξηση στην τιμή του. Όταν οι καταναλωτές

έρθουν στο κατάστημα για το φθηνό προϊόν $A$, τότε συχνά, όπως φαίνεται από την ισχύ του κανόνα, θα αγοράσουν και το $B$. Αν τυχόν εντοπίσουν τη μικρή αύξηση του $B$, τότε πιθανότατα θα θεωρήσουν ότι δεν αξίζει να μεταβούν σε άλλο κατάστημα απλά για φθηνότερο $B$. Είναι φανερό πως όλοι κερδίζουν, οι καταναλωτές δεν ξοδεύουν παραπάνω χρήματα και ταυτόχρονα το κατάστημα προσελκύει περισσότερους καταναλωτές.

### Περίπτωση α

Το σύνολο $A$ να αποτελείται από 1 μόνο στοιχείο.

*Παράδειγμα 1*

| **Αντικειμενικά μέτρα** | **Κατηγορία-Στόχος Β** | **Μέγεθος Α** |
|---|---|---|
| Confidence | Ρύζια | 1 |

Στην κατηγορία “Ρύζια” ανήκουν τα προϊόντα ΡΥΖΙ ARBORIO, BASMATI, JASMINE, PARBOILED, ΑΓΡΙΟ, ΓΛΑΣΣΕ, ΚΑΡΟΛΙΝΑ, ΜΠΛΟΥ ΡΟΖ, ΝΥΧΑΚΙ, ΣΟΥΠΕ.

| | | | |
|---|---|---|---|
| 1 | ΚΥΒΟΙ ΚΟΤΑΣ | → | ΡΥΖΙ PARBOILED |
| 2 | ΡΕΒΥΘΙΑ | → | ΡΥΖΙ PARBOILED |
| 3 | ΦΑΚΕΣ | → | ΡΥΖΙ PARBOILED |
| 4 | ΚΥΒΟΙ ΛΑΧΑΝΙΚΩΝ | → | ΡΥΖΙ PARBOILED |
| 5 | ΜΑΝΙΤΑΡΙΑ ΣΕ ΚΟΝΣΕΡΒΑ | → | ΡΥΖΙ ARBORIO |
| 6 | ΣΚΟΝΗ ΠΟΥΡΕ | → | ΡΥΖΙ PARBOILED |
| 7 | ΦΑΣΟΛΙΑ | → | ΡΥΖΙ PARBOILED |
| 8 | ΜΙΧ ΛΑΧΑΝΙΚΩΝ ΚΑΤΕΨΥΓΜΕΝ | → | ΡΥΖΙ PARBOILED |
| 9 | ΡΕΒΥΘΙΑ | → | ΡΥΖΙ ΚΑΡΟΛΙΝΑ |
| 10 | BALSAMIKO ΞΙΔΙ | → | ΡΥΖΙ PARBOILED |

Εικόνα 5.6: Οι καλύτεροι κανόνες $A \rightarrow B$, με $B$ ϵ “Ρύζια” και $|A| = 1$.

Σημειώνεται ότι θέσαμε έναν επιπλέον περιορισμό, τα προϊόντα του $A$ να μην ανήκουν στην κατηγορία “Ρύζια”.

*Παράδειγμα 2*

| Αντικειμενικά μέτρα | Προϊόν-Στόχος Β | Μέγεθος Α |
|---|---|---|
| Confidence | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ | 1 |

Αποτελέσματα

| | | | |
|---|---|---|---|
| 1 | ΧΟΙΡΙΝΟΣ ΛΑΙΜΟΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 2 | ΜΟΣΧΑΡΙ ΝΩΠΟ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 3 | ΧΟΙΡΙΝΕΣ ΜΠΡΙΖΟΛΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 4 | ΧΟΙΡΙΝΟ ΜΠΟΥΤΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 5 | ΧΟΙΡΙΝΗ ΣΠΑΛΑ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 6 | ΜΟΣΧΑΡΙΣΙΑ ΣΠΑΛΑ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 7 | ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 8 | ΜΠΑΜΙΕΣ ΚΑΤΕΨΥΓΜΕΝ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 9 | ΚΥΒΟΙ ΚΟΤΑΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 10 | ΧΥΛΟΠΙΤΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |

Εικόνα 5.7: Οι καλύτεροι κανόνες $A \rightarrow B$, με $B$ ="ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ" και $|A| = 1$.

## Περίπτωση β

Το σύνολο $A$ να αποτελείται από τουλάχιστον δύο στοιχεία.

*Παράδειγμα 1*

| Αντικειμενικά μέτρα | Προϊόν-Στόχος Β | Μέγεθος Α |
|---|---|---|
| Confidence | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ | $\geq 2$ |

Αποτελέσματα

| | | | | | |
|---|---|---|---|---|---|
| 1 | | ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 2 | | ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ | ΣΠΑΓΓΕΤΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 3 | | ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 4 | | ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ | ΨΩΜΙΑ ΓΙΑ ΤΟΣΤ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 5 | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | ΣΠΑΓΓΕΤΙ | ΠΑΤΑΤΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 6 | ΣΠΑΓΓΕΤΙ | ΠΑΤΑΤΕΣ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 7 | ΠΑΤΑΤΕΣ | ΨΩΜΙΑ ΓΙΑ ΤΟΣΤ | ΦΡΥΓΑΝΙΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 8 | ΑΥΓΑ ΚΟΤΑΣ | ΣΠΑΓΓΕΤΙ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 9 | | ΦΑΚΕΣ | ΦΡΥΓΑΝΙΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| 10 | ΚΡΕΜΜΥΔΙΑ | ΠΑΤΑΤΕΣ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |

Εικόνα 5.8: Οι καλύτεροι κανόνες $A \rightarrow B$, με $B$ =“ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ” και $|A| \geq 2$.

Αξίζει να παρατηρήσουμε πως το προϊόν ΨΩΜΙΑ ΓΙΑ ΤΟΣΤ αποτελεί ένα από τα πιό δημοφιλή προϊόντα του καταστήματος και αυτός είναι ο λόγος που συμμετέχει στους παραπάνω κανόνες, αποτελώντας ουσιαστικά θόρυβο. Έτσι, δοκιμάσαμε την επιλογή παραπάνω του ενός, αντικειμενικών μέτρων, κάνοντας χρήση της πολυκριτηριακής βελτιστοποίησης κατά Pareto.

### Περίπτωση γ

Χρήση, περισσότερων του ενός, αντικειμενικών μέτρων.

*Παράδειγμα 1*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Β** | **Μέγεθος Α** |
|---|---|---|

| | | |
|---|---|---|
| Όλα | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ | - |

Αποτελέσματα

| | | | | |
|---|---|---|---|---|
| | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΠΕΝΝΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| ΤΟΜΑΤΕΣ | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| ΤΟΜΑΤΟΠΟΛΤΟΙ | ΑΛΕΥΡΙ | ΤΥΠΟΥ COLA | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
| ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | GOUDA ΣΕ ΦΕΤΕΣ | ΕΛΛΗΝΙΚΟΣ ΚΑΦΕΣ | → | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |

Εικόνα 5.9: Οι καλύτεροι, κατά Pareto, κανόνες $A \rightarrow B$, με $B$ =“ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ”

Υπενθυμίζουμε ότι οι κανόνες-νικητές της μεθόδου Pareto δεν βρίσκονται σε απόλυτη κατάταξη.

### 3.2.1.3.2 Κατηγορία 2

Κανόνες $A \rightarrow B$ με δεδομένο προϊόν-στόχο ως $A$.

Στην περίπτωση αυτή, έχουμε προϊόντα (σύνολο $B$) η αγορά των οποίων ευνοείται από την αυτή είτε αποκλειστικά του προϊόντος $A$ είτε του $A$ σε συνδυασμό με άλλα. Τέτοιοι κανόνες είναι χρήσιμοι, για ένα supermarket, όταν αναζητά ισχυρούς κανόνες που πηγάζουν από συγκεκριμένα προϊόντα. Για παράδειγμα, σε ενδεχόμενο διακοπής της συνεργασίας με κάποιον προμηθευτή, τα προϊόντα στο $B$ αντιστοιχούν σε αυτά που μπορεί να επηρεαστούν από τη διακοπή της κυκλοφορίας των προϊόντων στο $A$.

Αξίζει να αναφέρουμε ότι σε αυτήν την κατηγορία, όπου δεν υπάρχει περιορισμός στο σύνολο $B$, το αντικειμενικό μέτρο **confidence** ($P[B|A]$) δεν αποτελεί καλή επιλογή. Ο λόγος είναι ότι προϊόντα που περιλαμβάνονται σχεδόν σε κάθε καλάθι, είναι αναμενόμενο να περιλαμβάνονται και σε καλάθια που περιέχουν το σύνολο $A$. Έτσι, επιλέξαμε το μέτρο **added value** ($P[B|A]$ - $P[B]$) που αφαιρεί από την confidence τιμή, την apriori πιθανότητα του $B$.

### Περίπτωση α

Το σύνολο $A$ να αποτελείται από 1 μόνο στοιχείο.

*Παράδειγμα 1*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Α** | **Μέγεθος Α** |
|---|---|---|
| Added Value | ΚΟΥΝΟΥΠΙΔΙ | 1 |

Αποτελέσματα

| | | | |
|---|---|---|---|
| 1 | ΚΟΥΝΟΥΠΙΔΙ | → | ΜΠΡΟΚΟΛΟ |
| 2 | ΚΟΥΝΟΥΠΙΔΙ | → | ΚΑΡΟΤΑ |
| 3 | ΚΟΥΝΟΥΠΙΔΙ | → | ΠΑΤΑΤΕΣ |
| 4 | ΚΟΥΝΟΥΠΙΔΙ | → | ΜΠΑΝΑΝΕΣ |
| 5 | ΚΟΥΝΟΥΠΙΔΙ | → | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |
| 6 | ΚΟΥΝΟΥΠΙΔΙ | → | ΜΗΛΑ |
| 7 | ΚΟΥΝΟΥΠΙΔΙ | → | ΤΟΜΑΤΕΣ |
| 8 | ΚΟΥΝΟΥΠΙΔΙ | → | ΚΡΕΜΜΥΔΙΑ |
| 9 | ΚΟΥΝΟΥΠΙΔΙ | → | ΑΓΓΟΥΡΙΑ |
| 10 | ΚΟΥΝΟΥΠΙΔΙ | → | ΑΥΓΑ ΚΟΤΑΣ |

Εικόνα 5.10: Οι καλύτερο κανόνες $A \rightarrow B$ , με “ΚΟΥΝΟΥΠΙΔΙ” $\epsilon$ $A$ και $|A| = 1$

*Παράδειγμα 2*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Α** | **Μέγεθος Α** |
|---|---|---|
| Added Value | ΣΟΛΩΜΟΣ | 1 |

Αποτελέσματα

| | | | |
|---|---|---|---|
| 1 | ΣΟΛΩΜΟΣ | → | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |

| | | | |
|---|---|---|---|
| 2 | ΣΟΛΩΜΟΣ | → | PHILADELPHIA |
| 3 | ΣΟΛΩΜΟΣ | → | ΑΛΛΑΝΤΙΚΑ ΚΑΠΝΙΣΤΗΣ ΓΑΛΟΠΟΥΛΑΣ |
| 4 | ΣΟΛΩΜΟΣ | → | TABLETS ΣΟΚΟΛΑΤΕΣ ΑΠΛΕΣ |
| 5 | ΣΟΛΩΜΟΣ | → | ΑΥΓΑ ΚΟΤΑΣ |
| 6 | ΣΟΛΩΜΟΣ | → | ΓΡΑΒΙΕΡΑ |
| 7 | ΣΟΛΩΜΟΣ | → | ΣΤΡΑΓΓΙΣΤΑ ΓΙΑΟΥΡΤΙΑ LIGHT |
| 8 | ΣΟΛΩΜΟΣ | → | ΦΡΥΓΑΝΙΕΣ |
| 9 | ΣΟΛΩΜΟΣ | → | ΜΠΑΝΑΝΕΣ |
| 10 | ΣΟΛΩΜΟΣ | → | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ |

Εικόνα 5.11: Οι καλύτερο κανόνες $A \rightarrow B$ , με “ΣΟΛΩΜΟΣ” ϵ $A$ και $|A| = 1$

*Παράδειγμα 3*

| **Αντικειμενικά μέτρα** | **Κατηγορία-Στόχος Α** | **Μέγεθος Α** |
|---|---|---|
| Added Value | Χορταρικά | 1 |

Στην κατηγορία “Χορταρικά” περιλαμβάνονται προϊόντα όπως ΠΡΑΣΣΑ, ΧΟΡΤΑ ΑΝΤΙΔΙΑ, ΣΠΑΝΑΚΙ, ΣΕΛΕΡΥ, ΣΠΑΡΑΓΓΙΑ, ΣΕΛΙΝΟ, ΡΑΔΙΚΙΑ, ΠΑΤΖΑΡΙΑ, ΣΕΣΚΟΥΛΑ, ΒΛΗΤΑ.

Αποτελέσματα

| | | | |
|---|---|---|---|
| 1 | ΣΠΑΝΑΚΙ | → | ΚΡΕΜΜΥΔΙΑ |
| 2 | ΣΕΛΙΝΟ | → | ΚΑΡΟΤΑ |
| 3 | ΣΕΛΕΡΥ | → | ΚΑΡΟΤΑ |
| 4 | ΣΠΑΝΑΚΙ | → | ΑΝΗΘΟΣ |
| 5 | ΠΡΑΣΣΑ | → | ΚΡΕΜΜΥΔΙΑ |

| | | | |
|---|---|---|---|
| 6 | ΠΡΑΣΣΑ | → | ΑΝΗΘΟΣ |
| 7 | ΠΡΑΣΣΑ | → | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |
| 8 | ΣΕΛΕΡΥ | → | ΚΡΕΜΜΥΔΙΑ |
| 9 | ΣΠΑΝΑΚΙ | → | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ |
| 10 | ΠΡΑΣΣΑ | → | ΚΑΡΟΤΑ |
| 11 | ΣΕΛΙΝΟ | → | ΚΡΕΜΜΥΔΙΑ |
| 12 | ΣΕΛΙΝΟ | → | ΠΑΤΑΤΕΣ |
| 13 | ΣΠΑΝΑΚΙ | → | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |
| 14 | ΠΡΑΣΣΑ | → | ΠΙΠΕΡΙΕΣ |
| 15 | ΣΕΛΕΡΥ | → | ΠΑΤΑΤΕΣ |
| 16 | ΣΠΑΝΑΚΙ | → | ΚΑΡΟΤΑ |
| 17 | ΠΡΑΣΣΑ | → | ΠΑΤΑΤΕΣ |
| 18 | ΣΕΛΕΡΥ | → | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |
| 19 | ΣΠΑΝΑΚΙ | → | ΤΟΜΑΤΕΣ |
| 20 | ΣΠΑΝΑΚΙ | → | ΑΛΕΥΡΙ |

Εικόνα 5.12: Οι καλύτερο κανόνες $A \rightarrow B$ , με “Χορταρικά” ϵ $A$ και $|A| = 1$

### Περίπτωση β

Το σύνολο $A$ να αποτελείται από τουλάχιστον δύο στοιχεία.

*Παράδειγμα 1*

| **Αντικειμενικά μέτρα** | **Κατηγορία-Στόχος Α** | **Μέγεθος Α** |
|---|---|---|
| Added Value | Χορταρικά | $\geq 2$ |

Αποτελέσματα

| | | | | |
|---|---|---|---|---|
| 1 | ΑΝΗΘΟΣ | ΣΠΑΝΑΚΙ | → | ΚΡΕΜΜΥΔΙΑ |

| | | | | |
|---|---|---|---|---|
| 2 | ΣΕΛΙΝΟ | ΦΑΣΟΛΙΑ | → | ΚΑΡΟΤΑ |
| 3 | ΠΡΑΣΣΑ | ΑΝΗΘΟΣ | → | ΚΡΕΜΜΥΔΙΑ |
| 4 | ΣΠΑΝΑΚΙ | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | → | ΚΡΕΜΜΥΔΙΑ |
| 5 | ΣΕΛΙΝΟ | ΤΟΜΑΤΟΠΟΛΤΟΙ | → | ΚΑΡΟΤΑ |
| 6 | ΚΡΕΜΜΥΔΙΑ | ΣΠΑΝΑΚΙ | → | ΑΝΗΘΟΣ |
| 7 | ΣΕΛΙΝΟ | ΠΑΤΑΤΕΣ | → | ΚΑΡΟΤΑ |
| 8 | ΣΕΛΙΝΟ | ΑΓΓΟΥΡΙΑ | → | ΤΟΜΑΤΕΣ |
| 9 | ΣΕΛΙΝΟ | ΣΠΑΓΓΕΤΙ | → | ΚΑΡΟΤΑ |
| 10 | ΣΕΛΙΝΟ | ΚΡΕΜΜΥΔΙΑ | → | ΚΑΡΟΤΑ |
| 11 | ΣΕΛΙΝΟ | ΥΨΗΛΗΣ ΠΑΣΤΕΡΙΩΣΗΣ ΓΑΛΑ ΛΕΥΚΟ | → | ΚΑΡΟΤΑ |
| 12 | ΤΟΜΑΤΕΣ | ΣΕΛΙΝΟ | → | ΚΑΡΟΤΑ |
| 13 | ΣΕΛΙΝΟ | ΑΥΓΑ ΚΟΤΑΣ | → | ΚΑΡΟΤΑ |
| 14 | ΜΑΙΝΤΑΝΟΣ | ΣΕΛΙΝΟ | → | ΚΑΡΟΤΑ |
| 15 | ΠΡΑΣΣΑ | ΚΡΕΜΜΥΔΙΑ | → | ΑΝΗΘΟΣ |
| 16 | ΣΕΛΙΝΟ | ΦΡΥΓΑΝΙΕΣ | → | ΚΑΡΟΤΑ |
| 17 | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ | ΠΡΑΣΣΑ | → | ΚΡΕΜΜΥΔΙΑ |
| 18 | ΣΕΛΙΝΟ | ΣΠΑΓΓΕΤΙ | → | ΤΟΜΑΤΟΠΟΛΤΟΙ |
| 19 | ΤΟΜΑΤΕΣ | ΣΕΛΙΝΟ | → | ΑΓΓΟΥΡΙΑ |
| 20 | ΠΡΑΣΣΑ | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | → | ΚΡΕΜΜΥΔΙΑ |

Εικόνα 5.13: Οι καλύτερο κανόνες $A \rightarrow B$ , με “Χορταρικά” ϵ $A$ και $|A| \geq 2$

#### 3.2.1.3.3 Κατηγορία 3

Κανόνες $A \rightarrow B$ με δεδομένο προϊόν-στόχο ως $A$ και ως $B$ .

Στην περίπτωση αυτή, έχουμε σαν αποτέλεσμα επιπλέον προϊόντα (σύνολο $A$ ), που σε συνδυασμό με τα προϊόντα-στόχους, δίνουν ισχυρούς κανόνες. Για παράδειγμα, αν είναι γνωστή

η σχέση μεταξύ των προϊόντων $A$ και $B$, τότε ένα supermarket μπορεί να αναζητήσει προϊόντα που όταν επιλέγονται μαζί με το $A$, ευνοείται η αγορά του $B$. Τέτοιοι συνδυασμοί μπορούν να διαμορφώσουν προσφορές. Επιπλέον, πάλι για ένα supermarket θα μπορούσε να διερευνηθεί το κατά πόσο συσχετίζονται τα προϊόντα μεταξύ των ραφιών ή των κατηγοριών $A$ και $B$. Στη συνέχεια, περιορίζουμε το μέγεθος του $A$ σε 2 προϊόντα.

*Παράδειγμα 1*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Α** | **Προϊόν-Στόχος Β** | **Μέγεθος Α** |
|---|---|---|---|
| Confidence | ΑΥΓΑ ΚΟΤΑΣ | ΑΛΕΥΡΙ | 2 |

Αποτελέσματα

| | | | | |
|---|---|---|---|---|
| 1 | ΑΥΓΑ ΚΟΤΑΣ | ΜΑΓΙΑ | → | ΑΛΕΥΡΙ |
| 2 | ΑΥΓΑ ΚΟΤΑΣ | ΒΑΝΙΛΙΕΣ | → | ΑΛΕΥΡΙ |
| 3 | ΑΥΓΑ ΚΟΤΑΣ | ΜΑΓΕΙΡΙΚΗ ΣΟΔΑ | → | ΑΛΕΥΡΙ |
| 4 | ΑΥΓΑ ΚΟΤΑΣ | ΜΑΓΕΙΡΙΚΑ ΛΙΠΗ | → | ΑΛΕΥΡΙ |
| 5 | ΑΥΓΑ ΚΟΤΑΣ | ΣΤΑΓΟΝΕΣ ΣΟΚΟΛΑΤΑΣ | → | ΑΛΕΥΡΙ |
| 6 | ΑΥΓΑ ΚΟΤΑΣ | ΒΟΥΤΥΡΑ ΠΑΚΕΤΟ | → | ΑΛΕΥΡΙ |
| 7 | ΑΥΓΑ ΚΟΤΑΣ | ΜΑΡΓΑΡΙΝΕΣ ΠΑΚΕΤΟ | → | ΑΛΕΥΡΙ |
| 8 | ΑΥΓΑ ΚΟΤΑΣ | ΛΕΥΚΗ ΖΑΧΑΡΗ | → | ΑΛΕΥΡΙ |
| 9 | ΑΥΓΑ ΚΟΤΑΣ | ΡΟΦΗΜΑ ΚΑΚΑΟ ΣΚΟΝΗ | → | ΑΛΕΥΡΙ |
| 10 | ΑΥΓΑ ΚΟΤΑΣ | ΚΑΝΕΛΛΑ | → | ΑΛΕΥΡΙ |
| 11 | ΑΥΓΑ ΚΟΤΑΣ | ΣΙΜΙΓΔΑΛΙ | → | ΑΛΕΥΡΙ |
| 12 | ΑΥΓΑ ΚΟΤΑΣ | ΚΟΥΒΕΡΤΟΥΡΑ | → | ΑΛΕΥΡΙ |
| 13 | ΑΥΓΑ ΚΟΤΑΣ | ΑΡΑΒΟΣΙΤΕΛΑΙΟ | → | ΑΛΕΥΡΙ |
| 14 | ΑΥΓΑ ΚΟΤΑΣ | ΧΑΡΤΙ ΨΗΣΙΜΑΤΟΣ | → | ΑΛΕΥΡΙ |
| 15 | ΑΥΓΑ ΚΟΤΑΣ | ΗΛΙΕΛΑΙΟ | → | ΑΛΕΥΡΙ |
| 16 | ΑΥΓΑ ΚΟΤΑΣ | ΑΝΘΟΣ ΑΡΑΒΟΣΙΤΟΥ | → | ΑΛΕΥΡΙ |

Εικόνα 5.14: Οι καλύτερο κανόνες $A \rightarrow B$, με “Χορταρικά” ϵ $A$, $|A| = 2$ και $B =$ ”ΑΛΕΥΡΙ”

*Παράδειγμα 2*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Α** | **Προϊόν-Στόχος Β** | **Μέγεθος Α** |
|---|---|---|---|
| Confidence | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΜΠΕΙΚΟΝ | 2 |

Αποτελέσματα

| | | | | |
|---|---|---|---|---|
| 1 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΛΟΥΚΑΝΙΚΑ ΒΡΑΣΤΑ | → | ΜΠΕΙΚΟΝ |
| 2 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΩΜΟΠΛΑΤΕΣ | → | ΜΠΕΙΚΟΝ |
| 3 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | GOUDA | → | ΜΠΕΙΚΟΝ |
| 4 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΜΑΝΙΤΑΡΙΑ ΣΕ ΚΟΝΣΕΡΒΑ | → | ΜΠΕΙΚΟΝ |
| 5 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΣΑΛΑΜΙΑ ΑΕΡΟΣ | → | ΜΠΕΙΚΟΝ |
| 6 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΠΕΝΝΕΣ | → | ΜΠΕΙΚΟΝ |
| 7 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΖΑΜΠΟΝ | → | ΜΠΕΙΚΟΝ |
| 8 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΠΑΡΙΖΕΣ | → | ΜΠΕΙΚΟΝ |
| 9 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΚΕΤΣΑΠ | → | ΜΠΕΙΚΟΝ |
| 10 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | REGATO | → | ΜΠΕΙΚΟΝ |
| 11 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΣΚΛΗΡΟ ΤΥΡΙ ARLA | → | ΜΠΕΙΚΟΝ |
| 12 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΖΥΜΑΡΙΚΑ ΣΕΛΙΝΟ | → | ΜΠΕΙΚΟΝ |
| 13 | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΣΠΑΓΓΕΤΙ | → | ΜΠΕΙΚΟΝ |

Εικόνα 5.15: Οι καλύτεροι κανόνες $A \rightarrow B$, με “ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ” ϵ $A$, $|A| = 2$ και $B =$ ”ΑΛΕΥΡΙ”

*Παράδειγμα 3*

| **Αντικειμενικά μέτρα** | **Προϊόν-Στόχος Α** | **Προϊόν-Στόχος Β** |
|---|---|---|
| Όλα | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΜΠΕΙΚΟΝ |

Αποτελέσματα

| | | | | |
|---|---|---|---|---|
| | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | PHILADELPHIA | → | ΜΠΕΙΚΟΝ |
| | ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΦΡΕΣΚΟ ΓΑΛΑ ΛΕΥΚΟ LIGHT | → | ΜΠΕΙΚΟΝ |
| ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΑΥΓΑ ΚΟΤΑΣ | → | ΜΠΕΙΚΟΝ |
| ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ | ΜΑΡΓΑΡΙΝΕΣ SOFT | ΑΛΕΥΡΙ | → | ΜΠΕΙΚΟΝ |

Εικόνα 5.16: Οι καλύτεροι, κατά Pareto, κανόνες $A \rightarrow B$, με "ΚΡΕΜΕΣ ΓΑΛΑΚΤΟΣ" $\epsilon$ $A$, και $B =$"ΑΛΕΥΡΙ"

### 3.2.1.3.4 Κατηγορία 4

Κανόνες $A \rightarrow B$ χωρίς δεδομένα προϊόντα-στόχους.

Το πρόβλημα των δημοφιλών προϊόντων, το οποίο παρουσιάστηκε στην Κατηγορία 2, ήταν αρκετά εντονότερο σε αυτήν την κατηγορία. Υπενθυμίζουμε πως ούτε στην περίπτωση της Κατηγορίας 2 είχαμε προϊόντα-στόχους ως $B$, ωστόσο είχαμε ως $A$.

Ενδεικτικά, οι 200 καλύτεροι κανόνες, κατά **Confidence**, έχουν ως μέλη $B$ ένα απ' τα στοιχεία ΑΛΕΥΡΙ, ΛΕΥΚΗ ΖΑΧΑΡΗ, ΣΠΑΓΓΕΤΙ, ΤΟΜΑΤΟΠΟΛΤΟΙ, τα οποία ανήκουν στα 15 δημοφιλέστερα. Επιπλέον, το μέγεθος του $A$ είναι μεγάλο, γεγονός που ενισχύει την αρχική παρατήρηση σχετικά με τις συγκυριακές συσχετίσεις προϊόντων.

Οι 200 καλύτεροι κανόνες, κατά **Added Value,** πάλι έχουν ως $B$ ένα μεταξύ οκτώ προϊόντων, τα οποία είναι λιγότερο δημοφιλή. Οι κανόνες φέρουν φυσικό νόημα. Τέλος, το **Lift**, όπως και το **Added Value,** συγκρίνει την "a posteriori" $P[B|A]$, με την "a priori" $P[B]$ και αποτελεί ένα συμμετρικό αντικειμενικό μέτρο, πράγμα που όπως αναλύσαμε δεν το καθιστά κατάλληλο για την αξιολόγηση κανόνων. Ωστόσο, οι 200 καλύτεροι κανόνες, κατά **Lift,** έχουν ως $B$ ένα μεταξύ εξήντα εννέα προϊόντων, χωρίς να ανήκουν σε αυτά τα δημοφιλή. Παρακάτω παραθέτουμε τους 62 καλύτερους κανόνες κατά **Lift.**

| | | | | |
|---|---|---|---|---|
| 1 | ΑΛΕΥΡΙ | ΜΑΣΤΙΧΑ | → | ΜΑΧΛΕΠΙ |
| 2 | ΜΑΧΛΕΠΙ | ΑΛΕΥΡΙ | → | ΜΑΣΤΙΧΑ |

| | | | | | |
|---|---|---|---|---|---|
| 3 | | | ΜΑΣΤΙΧΑ | → | ΜΑΧΛΕΠΙ |
| 4 | | | ΜΑΧΛΕΠΙ | → | ΜΑΣΤΙΧΑ |
| 5 | | | ΠΛΑΣΤΙΚΑ ΚΟΥΤΑΛΙΑ | → | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ |
| 6 | | | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ | → | ΠΛΑΣΤΙΚΑ ΚΟΥΤΑΛΙΑ |
| 7 | | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ | ΧΑΡΤΟΠΕΤΣΕΤΕΣ | → | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ |
| 8 | | ΚΡΟΥΑΣΑΝ ΒΕΡΙΚΟΚΟ | ΨΩΜΙΑ ΓΙΑ ΤΟΣΤ | → | ΚΡΟΥΑΣΑΝ ΚΕΡΑΣΙ |
| 9 | | ΚΑΝΕΛΛΑ | ΜΑΓΕΙΡΙΚΗ ΣΟΔΑ | → | ΓΑΡΥΦΑΛΛΟ |
| 10 | | ΚΡΟΥΑΣΑΝ ΒΕΡΙΚΟΚΟ | ΚΡΟΥΑΣΑΝ ΣΟΚΟΛΑΤΑ | → | |
| 11 | | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ | ΤΥΠΟΥ COLA | → | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ |
| 12 | | ΚΡΟΥΑΣΑΝ ΚΕΡΑΣΙ | ΚΡΟΥΑΣΑΝ ΣΟΚΟΛΑΤΑ | → | ΚΡΟΥΑΣΑΝ ΒΕΡΙΚΟΚΟ |
| 13 | | ΒΑΝΙΛΙΕΣ | ΜΑΓΕΙΡΙΚΑ ΛΙΠΗ | → | ΑΜΜΩΝΙΑ |
| 14 | ΚΑΝΕΛΛΑ | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΑΛΕΥΡΙ | → | ΓΑΡΥΦΑΛΛΟ |
| 15 | | | ΕΤΟΙΜΑ ΦΑΓΗΤΑ ΜΕ ΠΟΥΛΕΡΙΚΑ | → | ΠΑΤΑΤΕΣ ΦΡΕΣΚΑ ΕΤΟΙΜΑ ΦΑΓΗΤΑ |
| 16 | | | ΠΑΤΑΤΕΣ ΦΡΕΣΚΑ ΕΤΟΙΜΑ ΦΑΓΗΤΑ | → | ΕΤΟΙΜΑ ΦΑΓΗΤΑ ΜΕ ΠΟΥΛΕΡΙΚΑ |
| 17 | | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ | ΤΥΠΟΥ COLA | → | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ |
| 18 | | | ΚΡΟΥΑΣΑΝ ΚΕΡΑΣΙ | → | ΚΡΟΥΑΣΑΝ ΒΕΡΙΚΟΚΟ |
| 19 | | | ΚΡΟΥΑΣΑΝ ΒΕΡΙΚΟΚΟ | → | ΚΡΟΥΑΣΑΝ ΚΕΡΑΣΙ |
| 20 | | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ | ΧΑΡΤΟΠΕΤΣΕΤΕΣ | → | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ |
| 21 | | | ΠΙΚΛΕΣ | → | ΤΑΡΑΜΑΣ |
| 22 | | | ΤΑΡΑΜΑΣ | → | ΠΙΚΛΕΣ |

| | | | | | |
|---|---|---|---|---|---|
| 23 | | | ΜΗΛΑ ΒΙΟΛΟΓΙΚ | → | ΑΧΛΑΔΙΑ ΒΙΟΛΟΓΙΚ |
| 24 | | | ΑΧΛΑΔΙΑ ΒΙΟΛΟΓΙΚ | → | ΜΗΛΑ ΒΙΟΛΟΓΙΚ |
| 25 | | ΚΑΝΕΛΛΑ | ΑΛΕΥΡΙ | → | ΓΑΡΥΦΑΛΛΟ |
| 26 | | ΚΑΝΕΛΛΑ | ΛΕΥΚΗ ΖΑΧΑΡΗ | → | ΓΑΡΥΦΑΛΛΟ |
| 27 | | | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ | → | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ |
| 28 | | | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ | → | ΠΛΑΣΤΙΚΑ ΠΙΑΤΑ |
| 29 | | | ΠΛΑΣΤΙΚΑ ΚΟΥΤΑΛΙΑ | → | ΠΛΑΣΤΙΚΑ ΠΟΤΗΡΙΑ |
| 30 | | ΜΑΓΕΙΡΙΚΑ ΛΙΠΗ | ΑΛΕΥΡΙ | → | ΑΜΜΩΝΙΑ |
| 31 | | | ΕΤΟΙΜΑ ΦΑΓΗΤΑ ΜΕ ΚΡΕΑΣ | → | ΠΑΤΑΤΕΣ ΦΡΕΣΚΑ ΕΤΟΙΜΑ ΦΑΓΗΤΑ |
| 32 | | | ΠΑΤΑΤΕΣ ΦΡΕΣΚΑ ΕΤΟΙΜΑ ΦΑΓΗΤΑ | → | ΕΤΟΙΜΑ ΦΑΓΗΤΑ ΜΕ ΚΡΕΑΣ |
| 33 | | | ΧΑΛΒΑΣ ΜΕ ΒΑΝΙΛΙΑ | → | ΧΑΛΒΑΣ ΜΕ ΚΑΚΑΟ |
| 34 | | | ΧΑΛΒΑΣ ΜΕ ΚΑΚΑΟ | → | ΧΑΛΒΑΣ ΜΕ ΒΑΝΙΛΙΑ |
| 35 | | | ΖΩΜΟΣ ΛΑΧΑΝΙΚΩΝ | → | ΖΩΜΟΣ ΚΟΤΑΣ |
| 36 | | | ΖΩΜΟΣ ΚΟΤΑΣ | → | ΖΩΜΟΣ ΛΑΧΑΝΙΚΩΝ |
| 37 | ΓΑΡΥΦΑΛΛΟ | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΑΛΕΥΡΙ | → | ΚΑΝΕΛΛΑ |
| 38 | | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΜΑΓΕΙΡΙΚΑ ΛΙΠΗ | → | ΑΜΜΩΝΙΑ |
| 39 | | ΓΑΡΥΦΑΛΛΟ | ΜΑΓΕΙΡΙΚΗ ΣΟΔΑ | → | ΚΑΝΕΛΛΑ |
| 40 | | ΓΑΡΥΦΑΛΛΟ | ΛΕΥΚΗ ΖΑΧΑΡΗ | → | ΚΑΝΕΛΛΑ |
| 41 | | | ΛΟΥΚΑΝΟΠΙΤΑ ΚΑΤΕΨΥΓΜΕΝ | → | ΤΥΡΟΠΙΤΑ ΚΑΤΕΨΥΓΜΕΝ |
| 42 | | | ΤΥΡΟΠΙΤΑ ΚΑΤΕΨΥΓΜΕΝ | → | ΛΟΥΚΑΝΟΠΙΤΑ ΚΑΤΕΨΥΓΜΕΝ |

| | | | | | |
|---|---|---|---|---|---|
| 43 | | ΜΠΑΝΑΝΕΣ ΒΙΟΛΟΓΙΚ | ΚΑΡΟΤΑ ΒΙΟΛΟΓΙΚ | → | ΜΗΛΑ ΒΙΟΛΟΓΙΚ |
| 44 | | | ΣΥΚΩΤΑΡΙΑ ΑΡΝΙΩΝΚΑΤΕΨΥΓΜΕΝ | → | ΑΡΝΑΚΙ |
| 45 | | ΓΑΡΥΦΑΛΛΟ | ΑΛΕΥΡΙ | → | ΚΑΝΕΛΛΑ |
| 46 | | | ΚΥΒΟΙ ΒΟΔΙΝΟ | → | ΚΥΒΟΙ ΚΟΤΑΣ |
| 47 | | | ΚΥΒΟΙ ΚΟΤΑΣ | → | ΚΥΒΟΙ ΒΟΔΙΝΟ |
| 48 | | | ΚΑΝΕΛΛΑ | → | ΓΑΡΥΦΑΛΛΟ |
| 49 | | | ΓΑΡΥΦΑΛΛΟ | → | ΚΑΝΕΛΛΑ |
| 50 | ΤΟΜΑΤΕΣ | ΚΟΛΟΚΥΘΙΑ | ΠΙΠΕΡΙΕΣ | → | ΜΕΛΙΤΖΑΝΕΣ |
| 51 | | | ΤΟΜΑΤΕΣ ΒΙΟΛΟΓΙΚ | → | ΑΓΓΟΥΡΙΑ ΒΙΟΛΟΓΙΚ |
| 52 | | | ΑΓΓΟΥΡΙΑ ΒΙΟΛΟΓΙΚ | → | ΤΟΜΑΤΕΣ ΒΙΟΛΟΓΙΚ |
| 53 | | ΜΗΛΑ ΒΙΟΛΟΓΙΚ | ΚΑΡΟΤΑ ΒΙΟΛΟΓΙΚ | → | ΜΠΑΝΑΝΕΣ ΒΙΟΛΟΓΙΚ |
| 54 | ΚΟΛΟΚΥΘΙΑ | ΠΙΠΕΡΙΕΣ | ΠΑΤΑΤΕΣ | → | ΜΕΛΙΤΖΑΝΕΣ |
| 55 | ΚΟΛΟΚΥΘΙΑ | ΠΙΠΕΡΙΕΣ | ΚΡΕΜΜΥΔΙΑ | → | ΜΕΛΙΤΖΑΝΕΣ |
| 56 | ΚΟΛΟΚΥΘΙΑ | ΠΙΠΕΡΙΕΣ | ΦΕΤΑ ΑΙΓΟΠΡΟΒΕΙΑ | → | ΜΕΛΙΤΖΑΝΕΣ |
| 57 | | ΚΟΥΒΕΡΤΟΥΡΑ | ΓΑΛΑ ΣΥΜΠΥΚΝΩΜΕΝΟ | → | ΣΑΝΤΙΓΥ |
| 58 | ΒΑΝΙΛΙΕΣ | ΛΕΥΚΗ ΖΑΧΑΡΗ | ΑΛΕΥΡΙ | → | ΑΜΜΩΝΙΑ |
| 59 | | | ΤΑΡΑΜΑΣ | → | ΧΑΛΒΑΣ ΜΕ ΚΑΚΑΟ |
| 60 | | | ΧΑΛΒΑΣ ΜΕ ΚΑΚΑΟ | → | ΤΑΡΑΜΑΣ |
| 61 | | | ΔΡΑΚΟΥΛΙΝΙΑ | → | ΠΑΤΑΤΑΚΙΑ ΜΕ ΠΙΤΣΑ |
| 62 | | | ΠΑΤΑΤΑΚΙΑ ΜΕ ΠΙΤΣΑ | → | ΔΡΑΚΟΥΛΙΝΙΑ |

Εικόνα 5.17: Οι καλύτεροι κανόνες $A \rightarrow B$ χωρίς προϊόντα στόχους

Στα παραπάνω αποτελέσματα μπορούμε να παρατηρήσουμε πολλούς κανόνες $A \rightarrow B$ με $|A| = 1$, να παρατίθενται μαζί με τους συμμετρικούς τους. Αυτό συμβαίνει αφενός διότι αμφότεροι ικανοποιούν τα κατώφλια confidence και support που αρχικά χρησιμοποιήθηκαν, αφετέρου λόγω της ίδιας αξιολόγησης, από το **Lift,** λόγω της συμμετρικής του ιδιότητας.

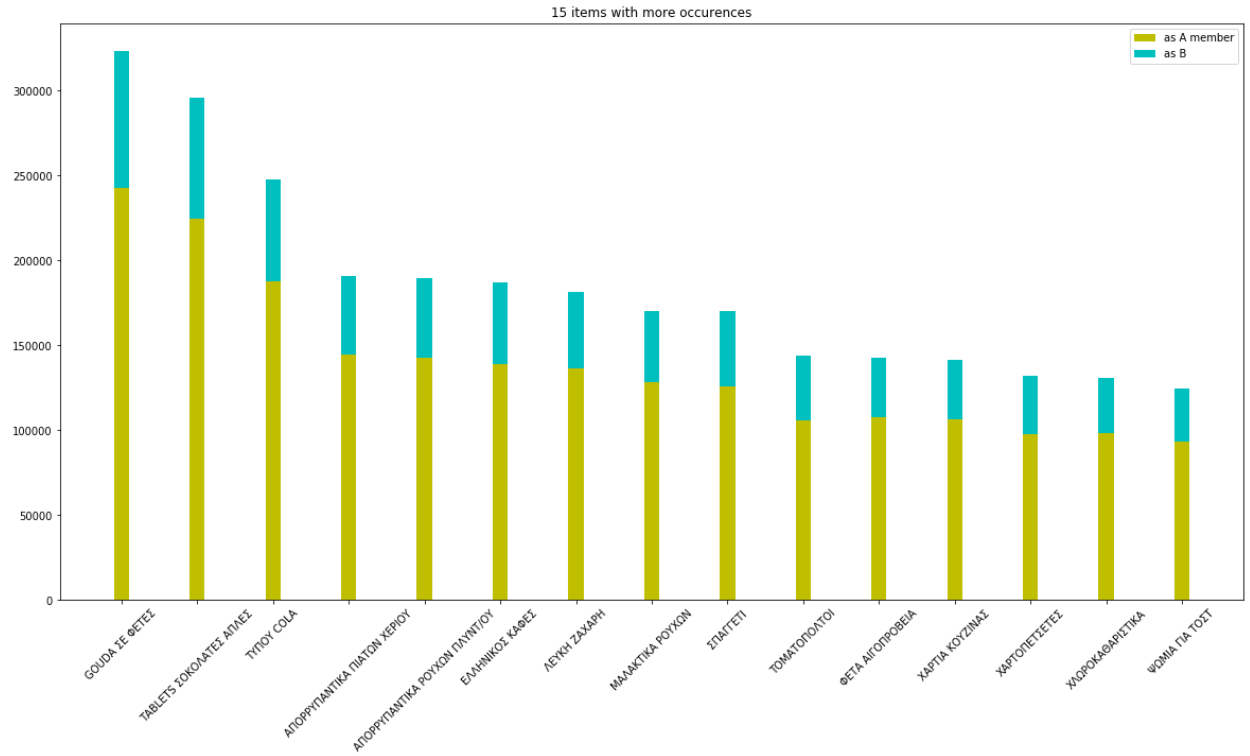
Αξίζει να σημειωθεί πως δοκιμάσαμε συνδυασμούς του **Lift** και άλλων κατάλληνων, για την Κατηγορία 4, αντικειμενικών μέτρων. Ωστόσο, κατά την εφαρμογή της πολυκριτηριακής βελτιστοποίησης κατά Pareto, οι κανόνες-νικητές παρουσίαζαν ξανά το προηγούμενο πρόβλημα.

### 3.2.1.2 Δίκτυο Κανόνων Συσχέτισης

Η μέθοδος $ARN(Rules, z)$ αναδεικνύει την ροή των κανόνων προς κάποιο προϊόν-στόχο $z$, μέσω των άμεσων και των έμμεσων συσχετίσεων προς αυτό. Για τον σκοπό αυτό, επιλέξαμε το αφαιρετικό επίπεδο των δεδομένων **Product-level** και κατασκευάσαμε έναν γράφο από τους Κανόνες Συσχέτισης που έχουν προηγουμένως βρεθεί. Κάθε κανόνας αποτελεί ακμή μεταξύ δύο στοιχειοσυνόλων (κόμβων) και φέρει βάρος ίσο με την **confidence** τιμή του. Αποτέλεσμα της μεθόδου αυτής είναι ένας κατευθυνόμενος γράφος, όπου όλοι οι κόμβοι έχουν μονοπάτι προς το προϊόν $z$.

Η επιλογή του $z$ είναι καθοριστική για τον $ARN$ γράφο. Το ενδιαφέρον για κάποιο συγκεκριμένο προϊόν είναι ένας τρόπος επιλογής. Σε διαφορετική περίπτωση, θα μπορούσε κανείς να επιλέξει το πιο δημοφιλές προϊόν, το προϊόν που έχει πωληθεί με τα περισσότερα άλλα ή το προϊόν που συμμετέχει στους περισσότερους Κανόνες Συσχέτισης. Παρατηρήσαμε πως η τελευταία περίπτωση είναι αυτή που οδηγεί σε γράφους με περισσότερη “πληροφορία”. Αξίζει να σημειωθεί, ότι σε πολλές περιπτώσεις, θέσαμε confidence κατώφλι με σκοπό την ανάδειξη λιγότερο ή περισσότερου ισχυρού γράφου. Φυσικά ο γράφος πρέπει να είναι διαχειρίσιμος, από άποψη όγκου, με σκοπό την απεικόνιση και την περαιτέρω μελέτη του.

Στη συνέχεια παρουσιάζουμε τα προϊόντα που συμμετέχουν στους περισσότερους Κανόνες Συσχέτισης $A \rightarrow B$ είτε στο μέλος $A$ είτε στο μέλος $B$.




Εικόνα 5.18: Τα 15 προϊόντα που συμμετέχουν στους περισσότερους Κανόνες Συσχέτισης $A \rightarrow B$

### 3.2.1.2.1 Περιπτώσεις Χρήσης

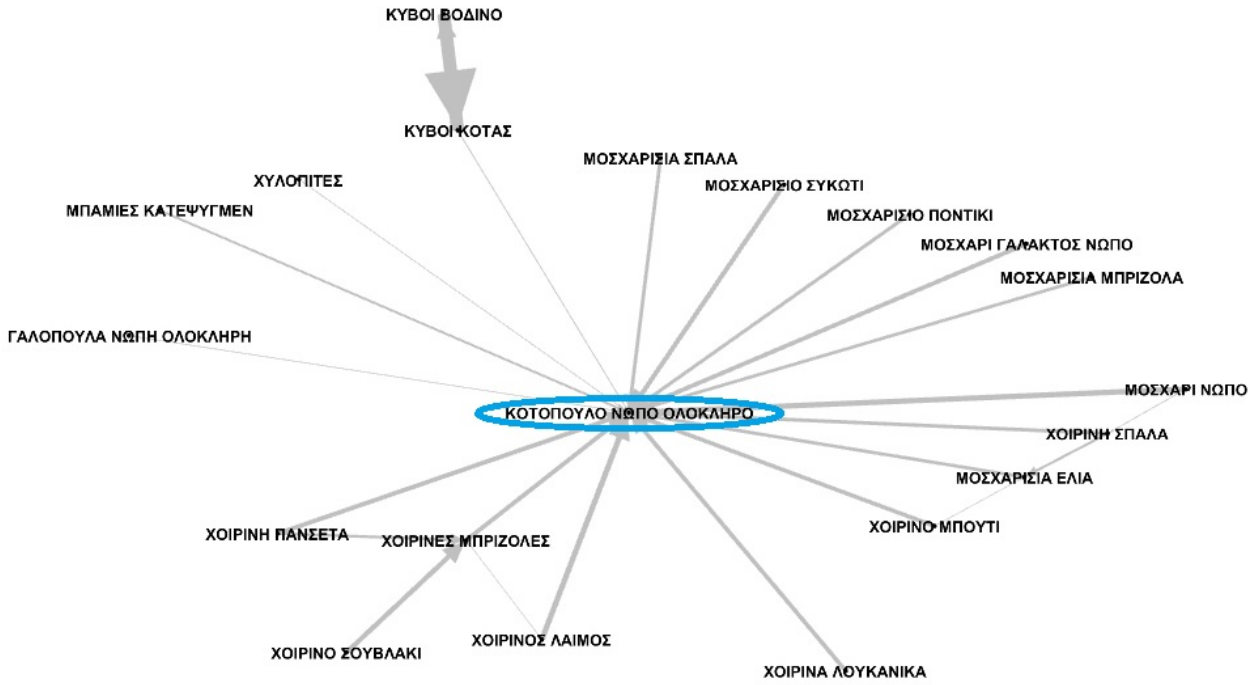
#### 3.2.1.2.1.1 Κατηγορία 1

Μικροί ARN Γράφοι.

Σε αυτήν την κατηγορία χρησιμοποιήσαμε υψηλό confidence κατώφλι για τους κανόνες-ακμές, με σκοπό την εξαγωγή γράφου αποτελούμενου από λίγους και ισχυρά συνδεδεμένους κόμβους.

Παράδειγμα 1

| **Προϊόν-Στόχος** | ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ |
|---|---|

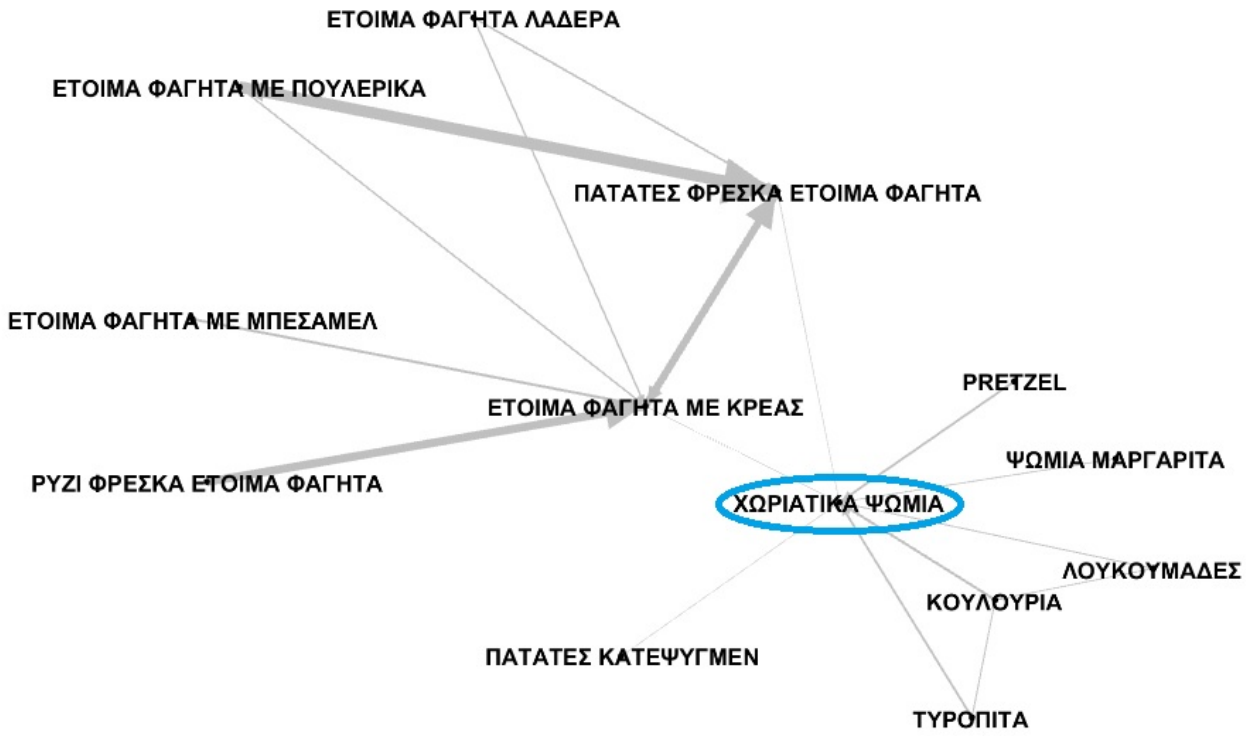



Εικόνα 5.19: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΚΟΤΟΠΟΥΛΟ ΝΩΠΟ ΟΛΟΚΛΗΡΟ"

Στην παραπάνω εικόνα φαίνεται η σχέση του κοτόπουλου με διάφορα χοιρινά και μοσχαρίσια κομμάτια κρέατος. Έχουν ενδιαφέρον οι δύο μεγαλύτερες ομάδες που συνδέονται με το κοτόπουλο. Στα αριστερά, βλέπουμε μία ομάδα λιπαρών κομματιών όπως ο ΧΟΙΡΙΝΟΣ ΛΑΙΜΟΣ, ΧΟΙΡΙΝΗ ΠΑΝΣΕΤΑ, ΧΟΙΡΙΝΟ ΣΟΥΒΛΑΚΙ και ΧΟΙΡΙΝΕΣ ΜΠΡΙΖΟΛΕΣ. Στα δεξιά, έχουμε άλιπα κρέατα όπως ΜΟΣΧΑΡΙ ΝΩΠΟ, ΧΟΙΡΙΝΗ ΣΠΑΛΑ, ΜΟΣΧΑΡΙΣΙΑ ΕΛΙΑ, ΧΟΙΡΙΝΟ ΜΠΟΥΤΙ.

Παράδειγμα 2

| **Προϊόν-Στόχος** | ΧΩΡΙΑΤΙΚΑ ΨΩΜΙΑ |
|---|---|

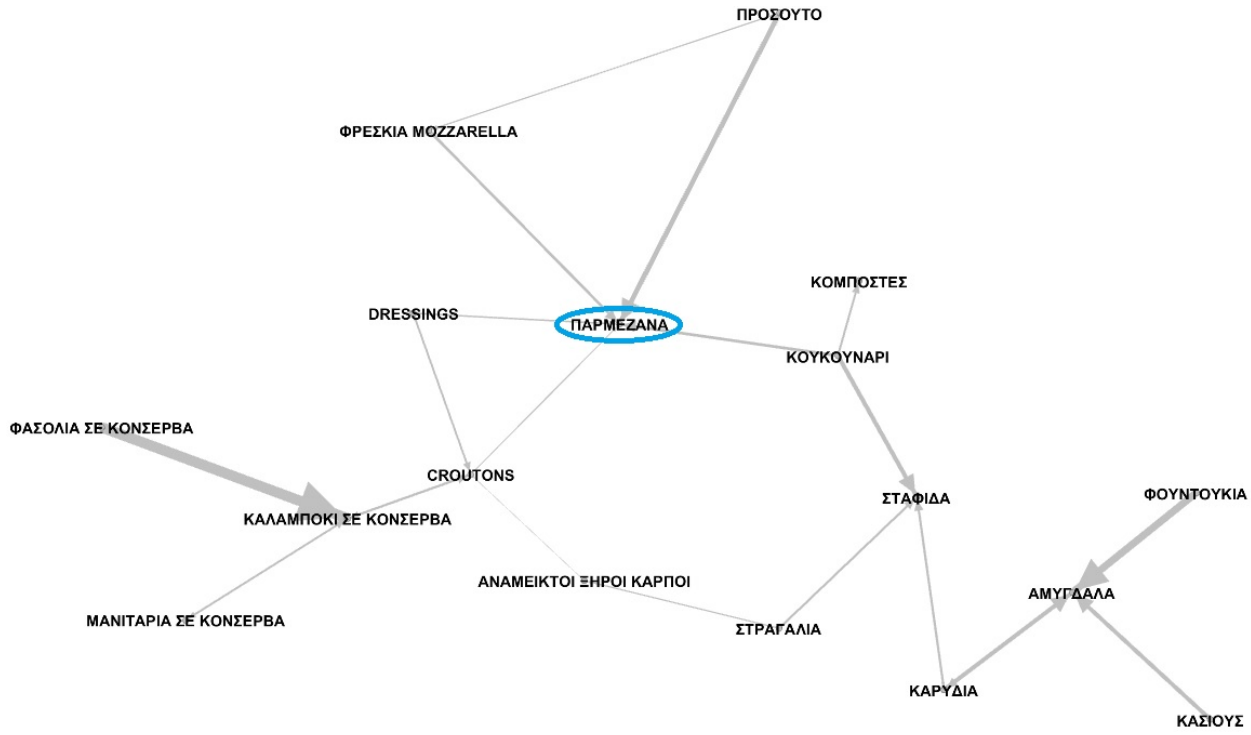



Εικόνα 5.20: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΧΩΡΙΑΤΙΚΑ ΨΩΜΙΑ"

Στην παραπάνω εικόνα φαίνονται διάφορα έτοιμα φαγητά, καθώς επίσης και διάφορα αρτοειδή προϊόντα. Έχει ενδιαφέρον η μεταξύ τους σχέση, που πραγματοποιείται μέσω της συσχέτισης των φαγητών με κρέας και τις σκέτες πατάτες, με το ψωμί.

Παράδειγμα 3

| **Προϊόν-Στόχος** | ΠΑΡΜΕΖΑΝΑ |
|---|---|

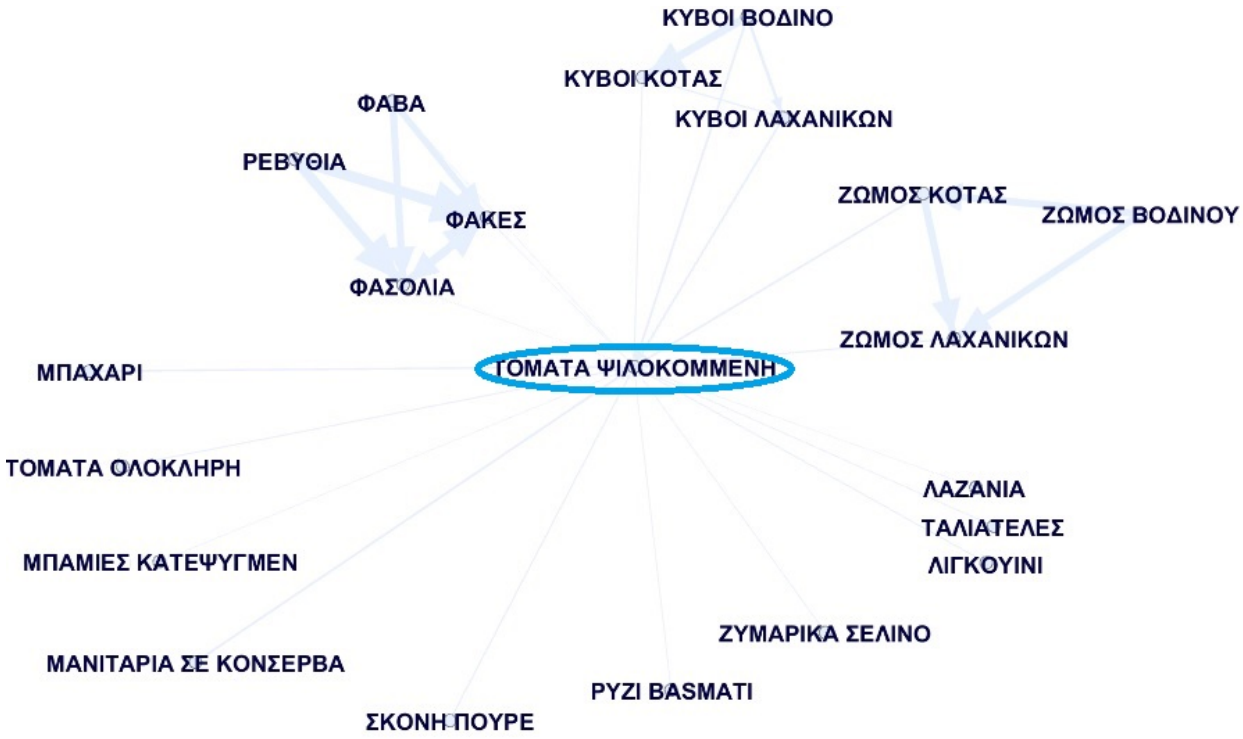



Εικόνα 5.21: Δίκτυο Κανόνων Συσχέτισης $ARN(R,z)$ για $z =$"ΠΑΡΜΕΖΑΝΑ"

Σε αυτόν τον γράφο κυριαρχούν προϊόντα που χρησιμοποιούνται σε σαλάτες. Η ΠΑΡΜΕΖΑΝΑ είναι ένα τυρί που χρησιμοποιείται ευρέως σε πράσινες σαλάτες. Έχει ενδιαφέρον ότι η σχέση της με τα CROUTONS και το ΚΟΥΚΟΥΝΑΡΙ, έχει σαν αποτέλεσμα τη συμμετοχή διαφόρων προϊόντων κονσέρβας και ξηρών καρπών, αντίστοιχα.

Παράδειγμα 4

| **Προϊόν-Στόχος** | ΤΟΜΑΤΑ ΨΙΛΟΚΟΜΜΕΝΗ |
|---|---|




Εικόνα 5.22: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΤΟΜΑΤΑ ΨΙΛΟΚΟΜΜΕΝΗ"

Στην εικόνα αυτή φαίνεται η σχέση της τομάτας με ομάδες προϊόντων όπως όσπρια, ζωμοί και κύβοι, καθώς επίσης και διάφορα είδη ζυμαρικών.

Παράδειγμα 5

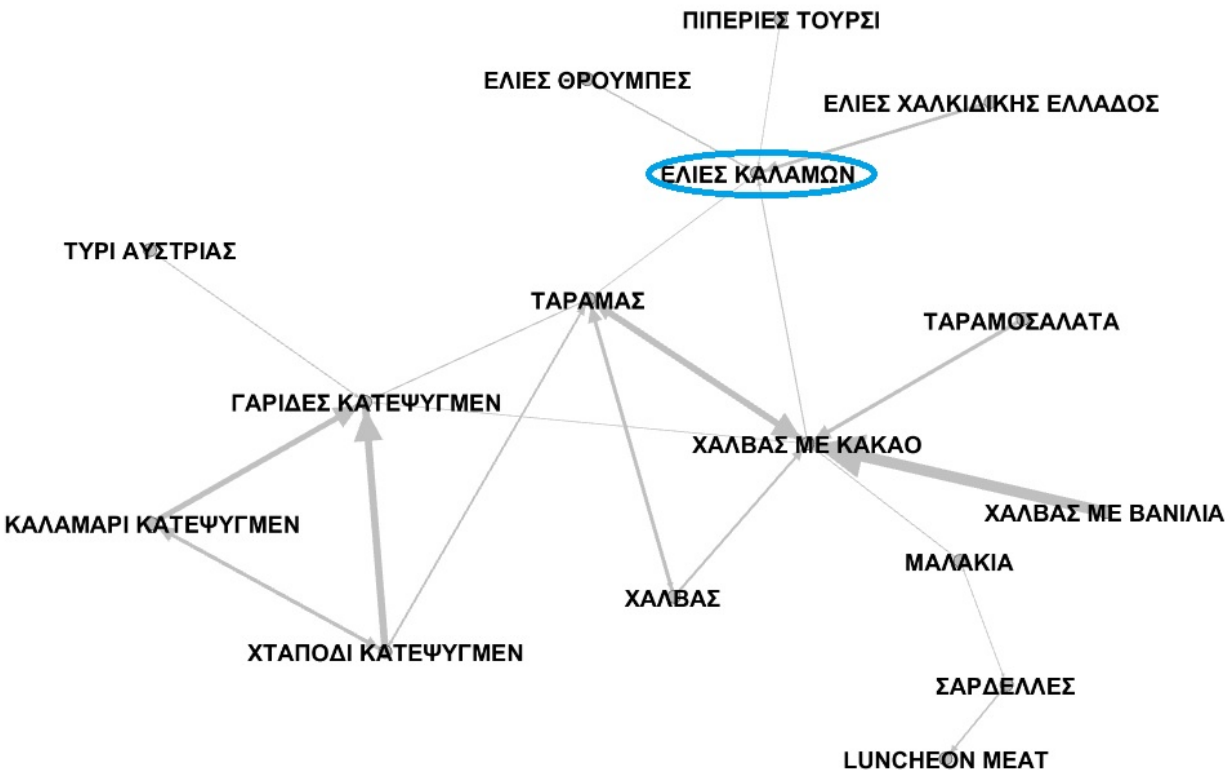
| **Προϊόν-Στόχος** | ΠΡΟΒΕΙΟ ΓΙΑΟΥΡΤΙ |
|---|---|


Εικόνα 5.23: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΠΡΟΒΕΙΟ ΓΙΑΟΥΡΤΙ"

Στην παραπάνω εικόνα φαίνονται οι σχέσεις προϊόντων με το ΠΡΟΒΕΙΟ ΓΙΑΟΥΡΤΙ. Όπως είναι αναμενόμενο, συμμετέχον άλλα είδη γιαουρτιών και γενικότερα, γαλακτοκομικά προϊόντα και προϊόντα ψυγείου. Το ενδιαφέρον είναι πως όλα τα προϊόντα αυτά αποτελούν επιλογές ανθρώπων που ακολουθούν υγιεινή διατροφή.

Παράδειγμα 6

| **Προϊόν-Στόχος** | ΕΛΙΕΣ ΚΑΛΑΜΩΝ |
|---|---|

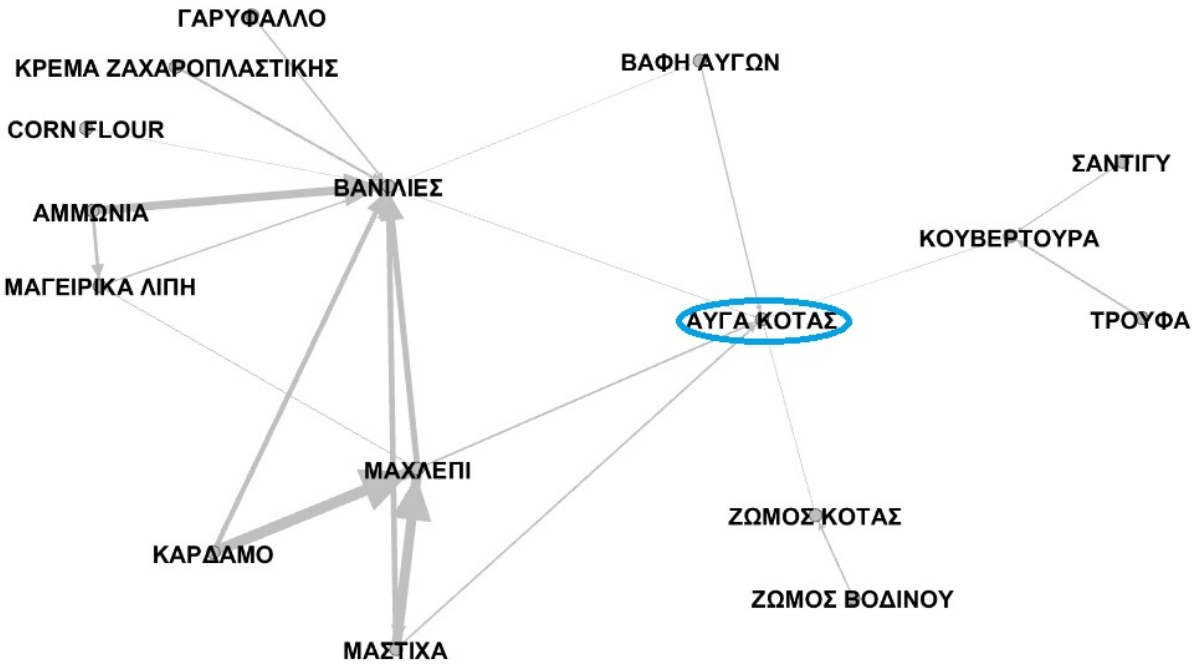



Εικόνα 5.24: Δίκτυο Κανόνων Συσχέτισης $ARN(R,z)$ για $z =$"ΠΡΟΒΕΙΟ ΓΙΑΟΥΡΤΙ"

Η σχέση των ελιών με τον ταραμά και τον χαλβά, οδήγησε στην ανάδειξη νηστίσιμων φαγητών.

Παράδειγμα 7

| **Προϊόν-Στόχος** | ΑΥΓΑ ΚΟΤΑΣ |
|---|---|




Εικόνα 5.25: Δίκτυο Κανόνων Συσχέτισης $ARN(R,z)$ για $z =$"ΑΥΓΑ ΚΟΤΑΣ"

Στην παραπάνω εικόνα φαίνεται η σχέση των αυγών με προϊόντα αρτοζαχαροπλαστικής, μπαχαρικά, ζωμούς και προϊόντα παρασκευής γλυκών.

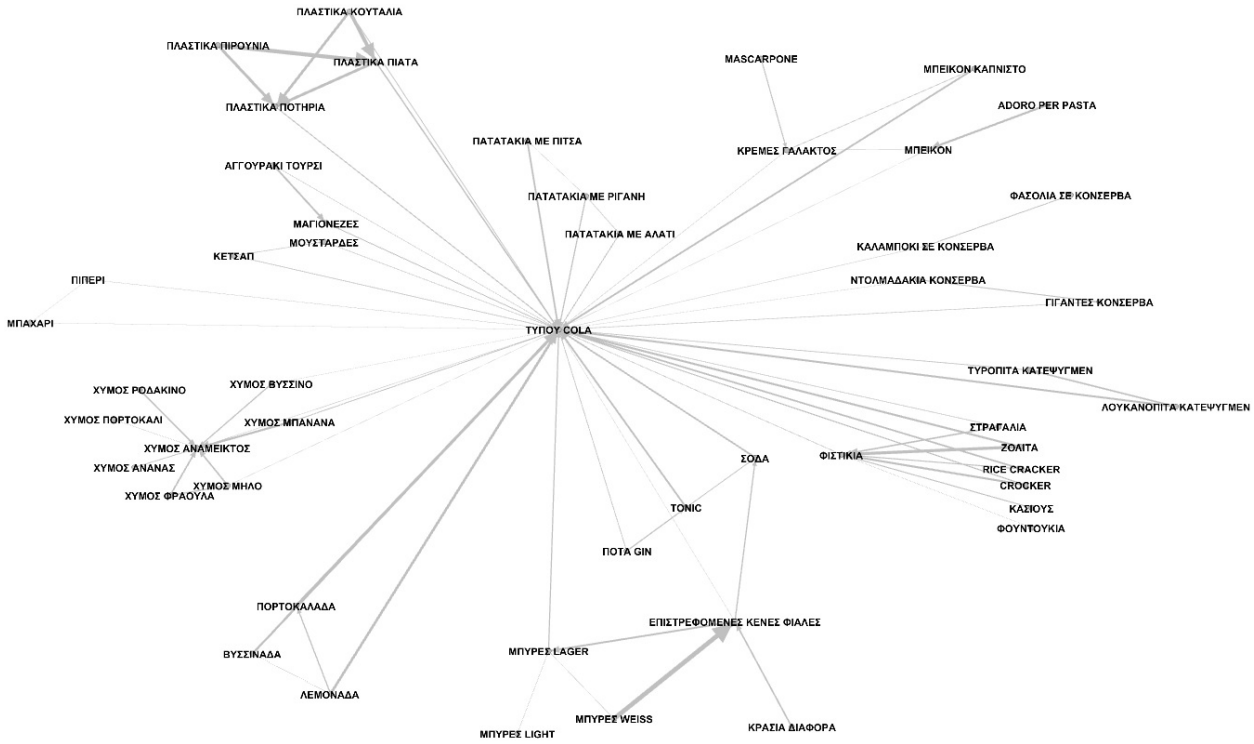
### 3.2.1.2.1.2 Κατηγορία 2

Μεγάλοι ARN Γράφοι.

Σε αυτήν την κατηγορία θέσαμε μικρότερες τιμές κατωφλίου confidence, με σκοπό την ανάδειξη σχέσεων μεταξύ διαφόρων ομάδων προϊόντων. Ωστόσο, όπως αναφέρουμε παρακάτω, γράφοι με πολλούς κόμβους-προϊόντα είναι πλέον δυσκολότερο να απεικονιστούν. Έτσι, η χρήση **Category-level** κρίνεται καταλληλότερη**.**

Παράδειγμα 1

| **Προϊόν-Στόχος** | ΤΥΠΟΥ ΚΟΛΑ |
|---|---|




Εικόνα 5.26: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΤΥΠΟΥ COLA"

Τα προϊόντα ΤΥΠΟΥ ΚΟΛΑ αποτελούν ιδιαίτερα δημοφιλή προϊόντα και ταιριάζουν με διάφορα φαγητά. Όπως φαίνεται, τα προϊόντα ΤΥΠΟΥ ΚΟΛΑ συνδέονται με προϊόντα μιας χρήσης (πλαστικά πηρούνια, πιάτα κ.α.), χυμούς, μπύρες, ποτά και ξηρούς καρπούς. Επίσης, με την κρέμα γάλακτος και διάφορα αλλαντικά που ταιριάζουν με αυτήν, διάφορες σάλτσες, καθώς επίσης και προϊόντα κονσέρβας και κατεψυγμένα προϊόντα.

Παράδειγμα 2

| **Προϊόν-Στόχος** | ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ |
|---|---|

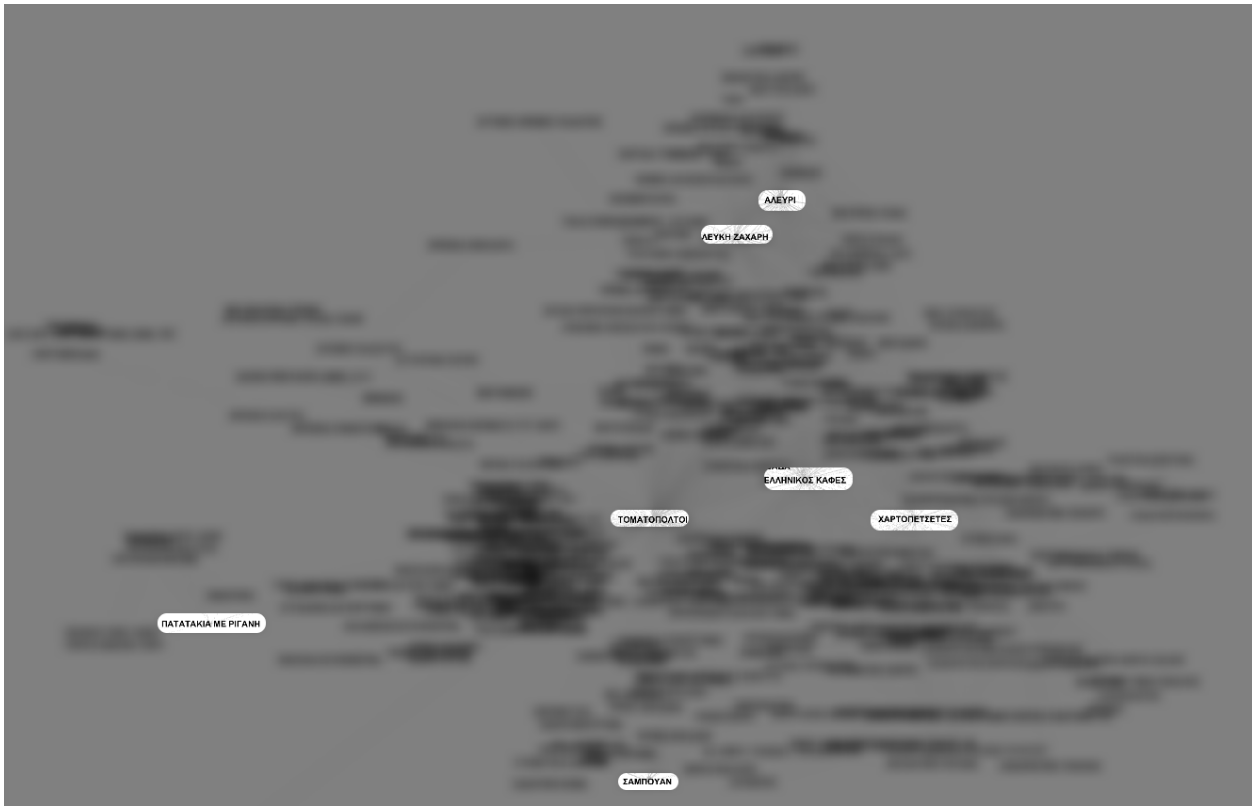

Εικόνα 5.27: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ"

Απομονώσαμε τα κύρια προϊόντα-κόμβους που οδηγούν στη σύνθεση του γράφου αυτού. Τα ΠΑΙΔΙΚΑ ΓΙΑΟΥΡΤΙΑ συνδέονται κυρίως με άλλα ποϊόντα για παιδιά, όπως γάλατα, μπισκότα, παιχνίδια, βιβλία, σαμπουάν κ.α. Η σχέση τους με τις ΜΠΑΝΑΝΕΣ οδηγεί στην ανάδειξη φρούτων, λαχανικών και χορταρικών, όπου οι κύριοι κόμβοι είναι οι ΜΠΑΝΑΝΕΣ, οι ΤΟΜΑΤΕΣ και τα ΚΡΕΜΜΥΔΙΑ. Τέλος, οι ΜΠΑΝΑΝΕΣ συνδέονται, μέσω φρέσκων και αποξηραμένων φρούτων με τις ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ, όπως επίσης και τα ΚΡΕΜΜΥΔΙΑ συνδέονται, μέσω μυρωδικών και σαλσών, με τις ΕΤΟΙΜΕΣ ΣΑΛΑΤΕΣ.

Παράδειγμα 3

| **Προϊόν-Στόχος** | ΤΟΜΑΤΟΠΟΛΤΟΙ |
|---|---|

\




Εικόνα 5.28: Δίκτυο Κανόνων Συσχέτισης $ARN(R, z)$ για $z =$"ΤΟΜΑΤΟΠΟΛΤΟΙ"

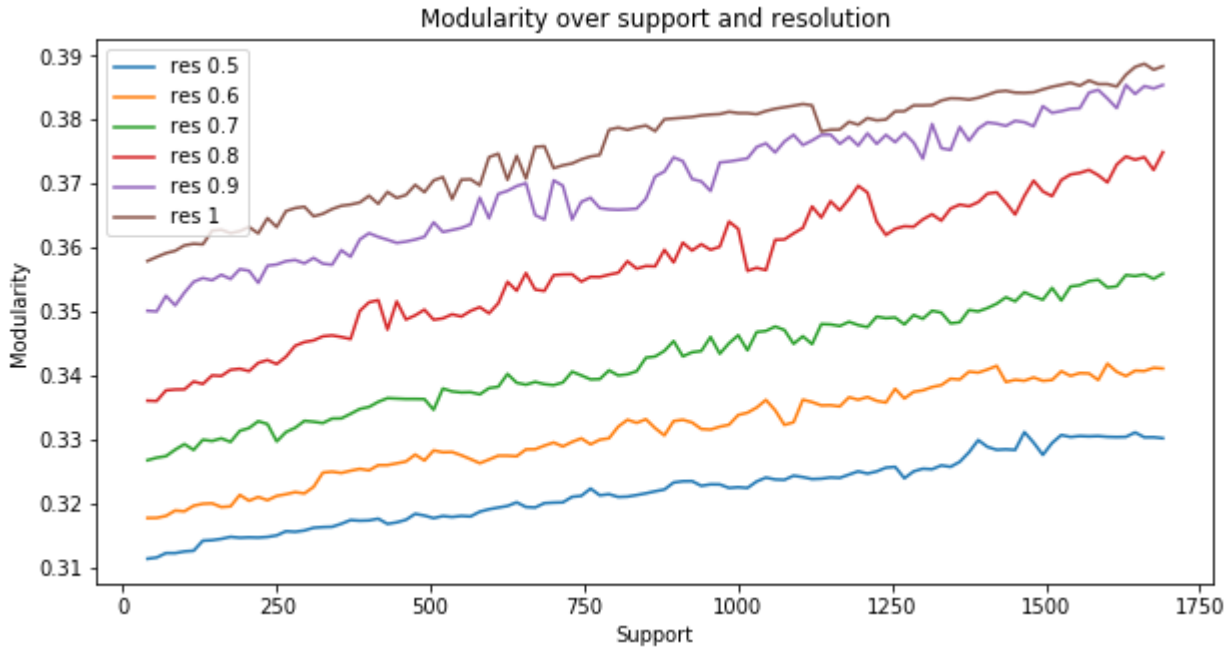
Στην παραπάνω εικόνα φαίνεται ότι το **Product-level**, δεν είναι κατάλληλο για την ανάδειξη σχέσεων μεταξύ κατηγοριών προϊόντων και πρέπει να εργαστούμε σε **Category-level**. Ο μεγάλος αριθμός των προϊόντων-κόμβων δυσχεραίνει την απεικόνιση, ωστόσο θα περιγράψουμε τη δομή του γράφου.

Οι κόμβοι που κυριαρχούν στον γράφο αυτό είναι οι ΤΟΜΑΤΟΠΟΛΤΟΙ, τα ΚΡΕΜΜΥΔΙΑ, η ΖΑΧΑΡΗ, το ΑΛΕΥΡΙ, ο ΕΛΛΗΝΙΚΟΣ ΚΑΦΕΣ, οι ΧΑΡΤΟΠΕΤΣΕΤΕΣ, το ΣΑΜΠΟΥΑΝ και ΠΑΤΑΤΑΚΙΑ. Καθένας εκ των προαναφερθέντων κόμβων, συνδέεται με σχετικά προϊόντα. Έτσι, συμμετέχουν λαχανικά, μυρωδικά, προϊόντα ψυγείου, κατεψυγμένα ψυγείου, προϊόντα προσωπικής υγιεινής και καθαρισμού και προϊόντα αρτοζαχαροπλαστικής. Η ομαδοποίηση των παραπάνω ομάδων, μέσω της χρήσης **Category-level**, θα είχε αποτέλεσμα την εύρωστη απεικόνιση των σχέσεων μεταξύ τους.

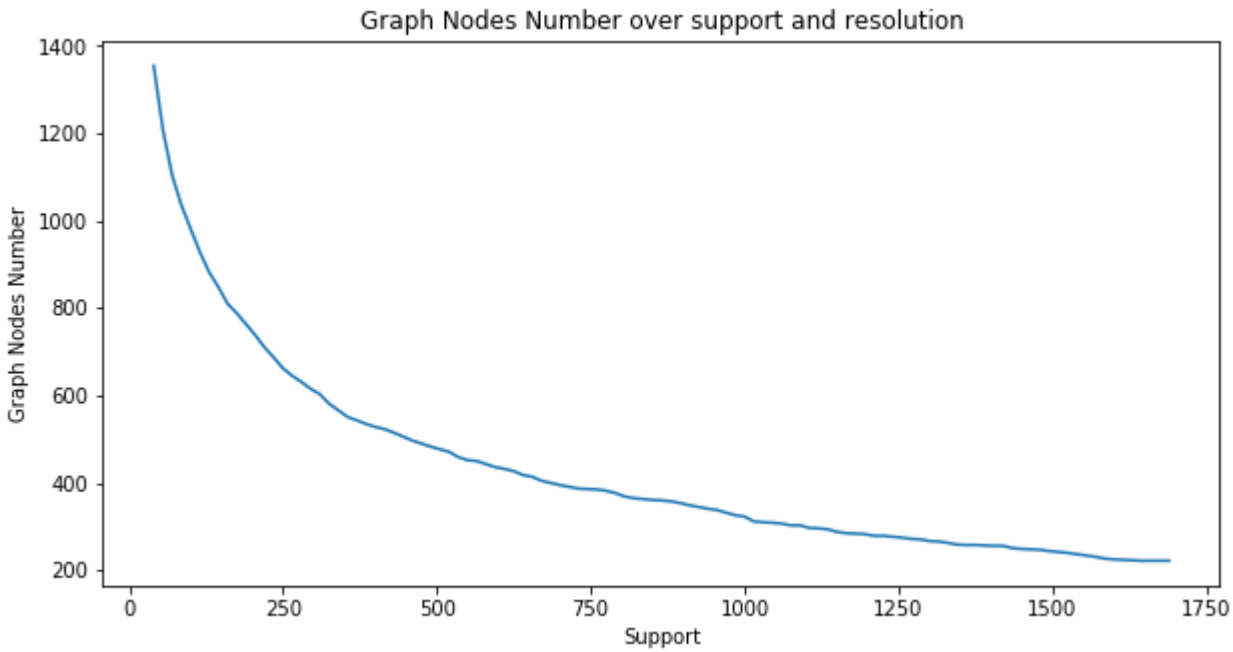
### 3.2.2 Ανίχνευση Κοινοτήτων

Για την τεχνική αυτή εφαρμόσαμε τη μέθοδο Louvain, χρησιμοποιώντας το Python package NetworkX. Συγκεκριμένα, κατασκευάσαμε έναν μη κατευθυνόμενο γράφο προϊόντων, σε επίπεδο **Product-level,** όπου κάθε ζευγάρι προϊόντων ενώνεται με μία ακμή που φέρει βάρος ίσο με τον αριθμό των συναλλαγών στις οποίες αυτά συνυπάρχουν.

Υπενθυμίζουμε πως και αυτή η τεχνική επιβάλλει την χρήση του κατωφλίου support, η επιλογή του οποίου είναι ένα κρίσιμο σημείο και αναλύεται παρακάτω. Επίσης, όπως αναφέραμε, η μέθοδος Louvain ανιχνεύει κοινότητες σε διάφορες αναλύσεις, δηλαδή παρέχει ιεραρχικά τις ομαδοποιήσεις προϊόντων και όχι μόνο τις τελικές κοινότητες. Ωστόσο, η υλοποίηση αυτή δεν προσφέρει κάτι τέτοιο. Τέλος, η υπερ-παράμετρος **resolution** του αλγορίθμου Louvain, αν και δεν συμπεριλαμβανόταν στο αρχικό paper, καθορίζει το μέγεθος των κοινοτήτων.




Εικόνα 5.29: Συνολική τιμή Modularity ως προς το κατώφλι support και της υπερ-παραμέτρου resolution




Εικόνα 5.30: Αριθμός κόμβων στον γράφο προϊόντων ως προς το κατώφλι support

Παραπάνω, φαίνεται η συνολική τιμή Modularity του γράφου μετά από ανίχνευση κοινοτήτων για διαφορετικές τιμές του κατωφλίου support και της υπερπαραμέτρου resolution. Σκοπός είναι να χρησιμοποιήσουμε τις τιμές εκείνες που δίνουν όσο πιο κοντά στη μέγιστη τιμή του Modularity και ταυτόχρονα, που έχουν σαν αποτέλεσμα κοινότητες αρκετά μικρές, ώστε να είναι διαχειρίσιμες και αρκετά μεγάλες, ώστε να φέρουν πληροφορία.

Βλέπουμε ότι η τιμή του resolution (default 1), είναι αντιστρόφως ανάλογη της συνολικής τιμής Modularity. Δηλαδή, όσο μικρότερες κοινότητες προσπαθούμε να πάρουμε σαν αποτέλεσμα, τόσο μικραίνει το Modularity. Σε μία άλλη κατεύθυνση, βλέπουμε ότι αύξηση του κατωφλίου support επιδρά θετικά στο Modularity, ωστόσο, οι εναπομείναντες κόμβοι-προϊόντα μειώνονται εκθετικά.

Λαμβάνοντας υπόψην τα παραπάνω, επιλέξαμε support 300 και resolution 0.85. Το κατώφλι support είχε ως αποτέλεσμα ο γράφος να αποτελείται από 612 κόμβους και 29369 ακμές. Με την εφαρμογή της μεθόδου Louvain, βρήκαμε 25 κοινότητες τις οποίες στη συνέχεια αξιολογήσαμε ώστε να προκύψει μία ιεραρχία αυτών. Για την αξιολόγηση χρησιμοποιήσαμε το μέτρο Utility που αποτελεί τον αρμονικό μέσο των μέτρων Information και Information density.

Έστω

$C_i$ η i-οστή κοινότητα

$|V_i|$ το σύνολο των κόμβων της

$|E_i|$ το σύνολο των κόμβων της

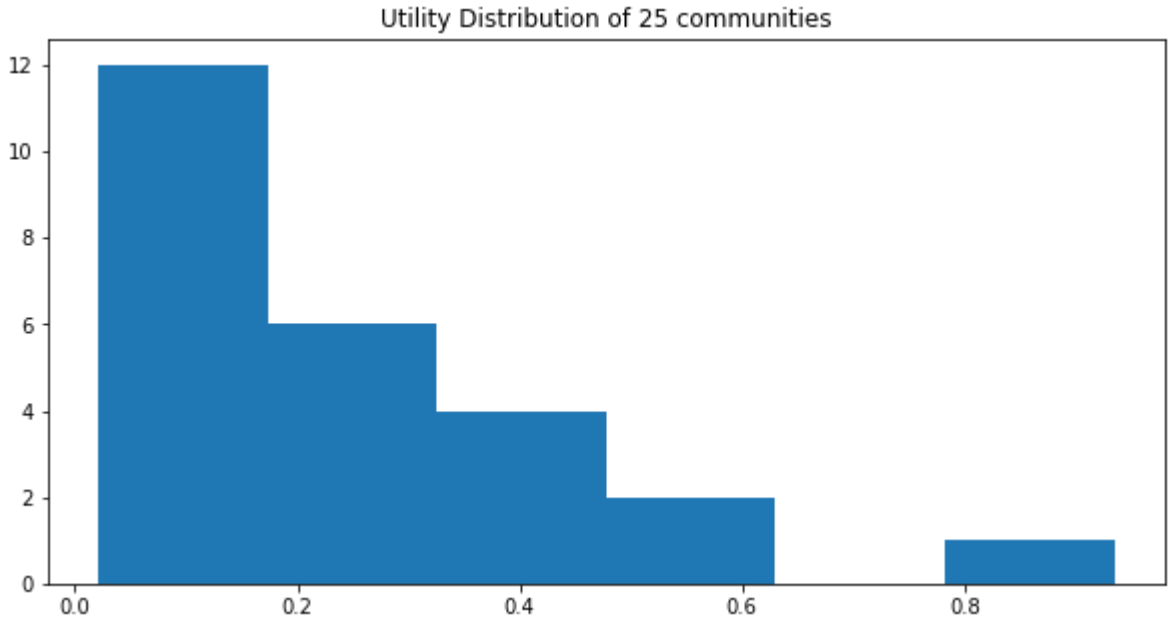
Τότε

**Information:** $$I(C_i) = \sum_{u,v \in E_i} min(P[v|u], P[u|v]) \quad (5.1)$$

**Information Density:** $$D(C_i) = \frac{I(C_i)}{|V_i|} \quad (5.2)$$

**Utility:** $$U(C_i) = \frac{2*D(C_i)*I(C_i)}{D(C_i)+I(C_i)} \quad (5.3)$$

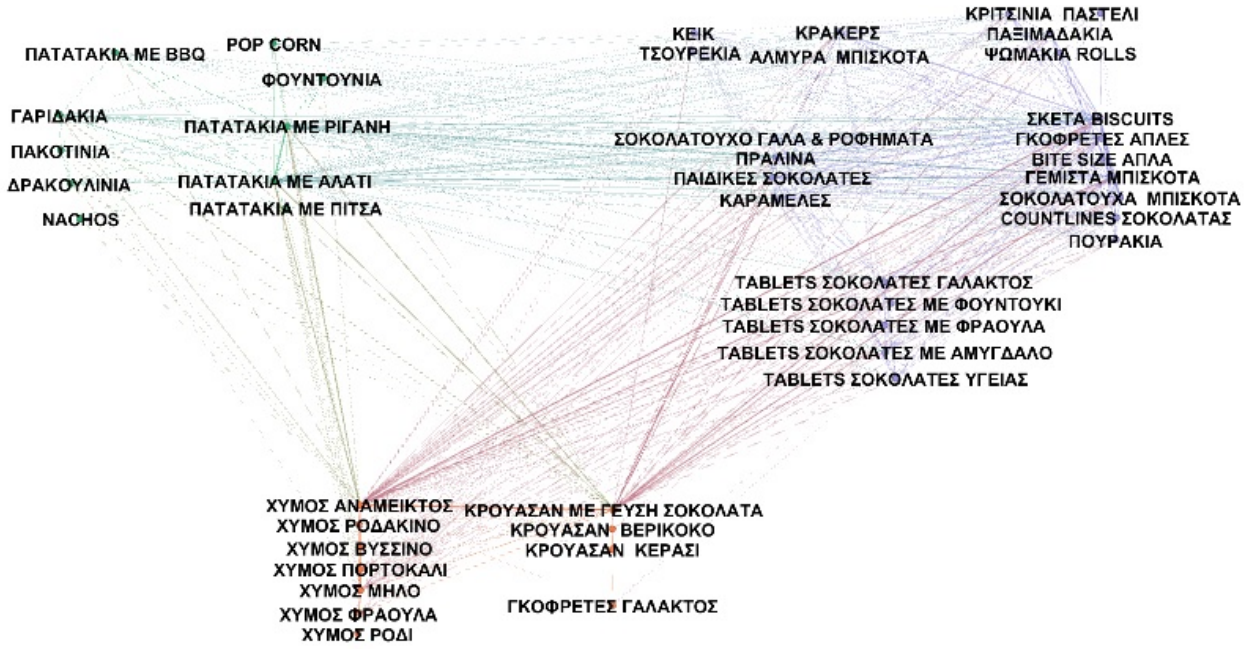
Αξίζει να σημειωθεί ότι χρησιμοποιήθηκε το min για τον υπολογισμό του Information, με σκοπό τον περιορισμό του προβλήματος των δημοφιλών προϊόντων, όπως αναπτύχθηκε προηγουμένως. Επίσης, με τη χρήση του utility, μεταξύ δύο κοινοτήτων με ίδιο Information Density, προτιμάται η μεγαλύτερη. Τέλος, παρουσιάζουμε την κατανομή του Utility για τις 25 κοινότητες που ανιχνεύθηκαν και στη συνέχεια, τις πιο ενδιαφέρουσες κοινότητες.




Εικόνα 5.31: Κατανομή του Utility των 25 ανιχνευμένων κοινοτήτων

Τονίζουμε πως σε περιπτώσεις που εκτιμήσαμε ότι μια κοινότητα περιέχει παραπάνω από μία κατηγορίες προϊόντων, αποφασίσαμε για λόγους ελέγχου και επιβεβαίωσης, να ομαδοποιήσουμε τις επιμέρους κατηγορίες. Αν και ο αλγόριθμός Louvain παρέχει την ιεραρχική ενοποίηση των κόμβων κάθε τελικής κοινότητας, η υλοποίηση που χρησιμοποιήσαμε δεν περιείχε κάτι τέτοιο. Έτσι, τρέξαμε εκ νέου τον αλγόριθμο Louvain για τις παραπάνω κοινότητες, των οποίων οι επιμέρους υπο-κοινότητες απεικονίζονται με ξεχωριστό χρώμα. Τέλος, σε κάθε κοινότητα ή υποκοινότητα, επιλέξαμε να απεικονίσουμε κοντά τα προϊόντα που είναι πιο ¨όμοια¨, για μια εύρωστη οπτικοποίηση.

## Κοινότητα 1




Εικόνα 5.32: Κοινότητα “Πατατάκια - Χυμοί - Σοκολατοειδή”, 1η θέση

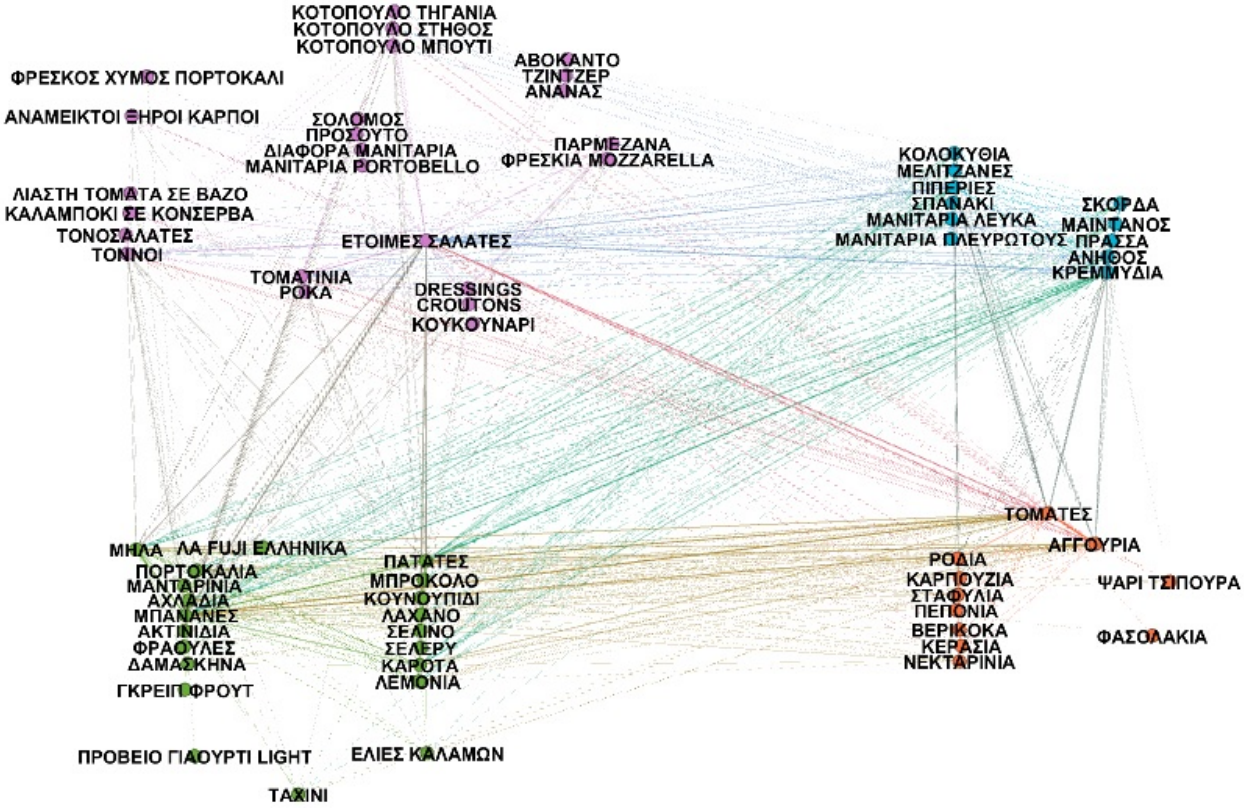
Η παραπάνω κοινότητα αποτελείται από 3 υπο-κοινότητες. Η πρώτη είναι τα πατατάκια, η δεύτερη οι χυμοί και τα κρουασάν και η τρίτη διάφορα σοκολατοειδή προϊόντα.

## Κοινότητα 2


Εικόνα 5.33: Κοινότητα “Έτοιμα Φαγητά - Ψωμοειδή”, 2η θέση

Η κοινότητα αυτή αναδεικνύει τη σχέση των έτοιμων φαγητών και των ψωμιών, κάτι που είχαμε δει στην $ARN$ μέθοδο με προϊόν-στόχο τα χωριάτικα ψωμιά. Είναι φανερό ότι τα έτοιμα φαγητά συνδέονται έντονα με τις έτοιμες πατάτες, ενώ μεταξύ τους αμυδρά, ενώ το ΚΟΤΟΠΟΥΛΟ ΨΗΤΟ αγοράζεται με διάφορα ψωμοειδή προϊόντα. Σαν αποτέλεσμα, ο αλγόριθμος Louvain που χρησιμοποιήθηκε για λόγους απεικόνισης, κατέταξε το ΚΟΤΟΠΟΥΛΟ ΨΗΤΟ με τα ψωμοειδή, αν και θα μπορούσαμε να περιμέναμε το αντίθετο.

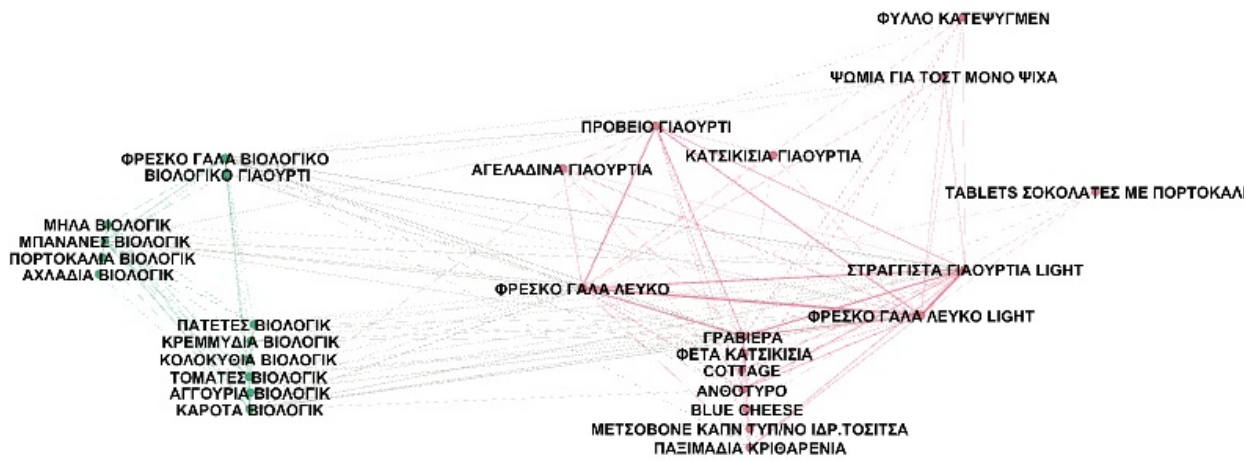
Κοινότητα 3




Εικόνα 5.34: Κοινότητα “Φρούτα -Λαχανικά - Έτοιμες Σαλάτες”, 3η θέση

Η κοινότητα αυτή αποτελείται από διάφορα φρούτα, λαχανικά και έτοιμες σαλάτες μαζί με κοντινά τους προϊόντα. Επίσης, αξίζει να αναφέρουμε το ότι ο αλγόριθμος Louvain, που τρέξαμε για την εύρεση υπο-κοινοτήτων, διαχώρισε τα καλοκαιρινά φρούτα από τα υπόλοιπα.

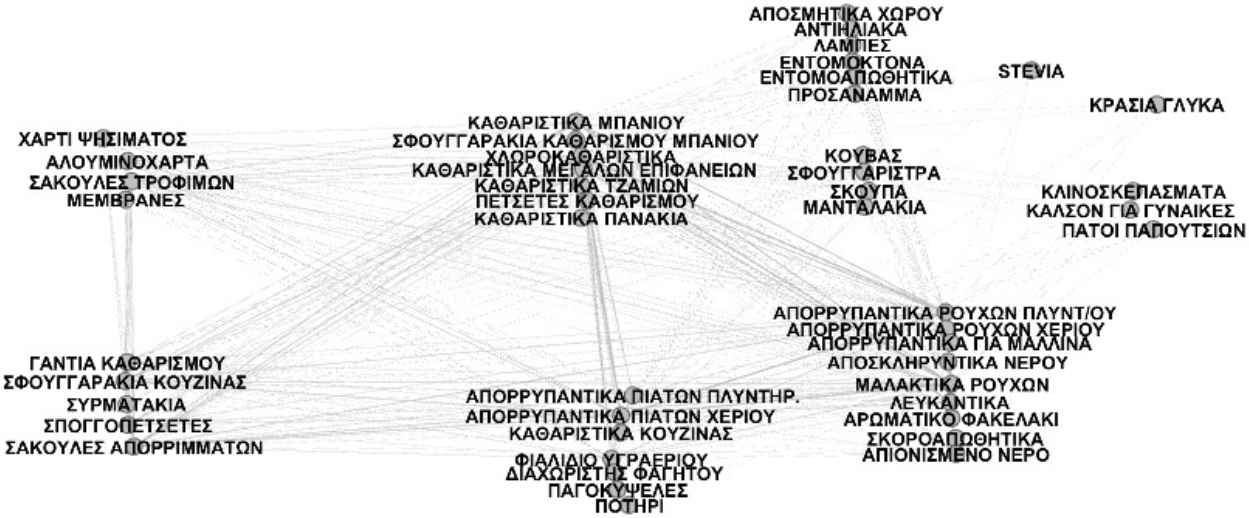
## Κοινότητα 4




Εικόνα 5.35: Κοινότητα “Βιολογικά - Γαλακτοκομικά”, 4η θέση

Η κοινότητα αυτή αποτελείται από τις υπο-κοινότητες των βιολογικών και των γαλακτοκομικών προϊόντων. Αξίζει να σημειωθεί πως τα βιολογικά φρούτα και λαχανικά, τοποθετήθηκαν μαζί και ταυτόχρονα σε διαφορετική κοινότητα από τα μη βιολογικά. Αυτό, είναι αναμενόμενο διότι οι καταναλωτές που επιλέγουν βιολογικά προϊόντα, τα προτιμούν σε όλες τις κατηγορίες.

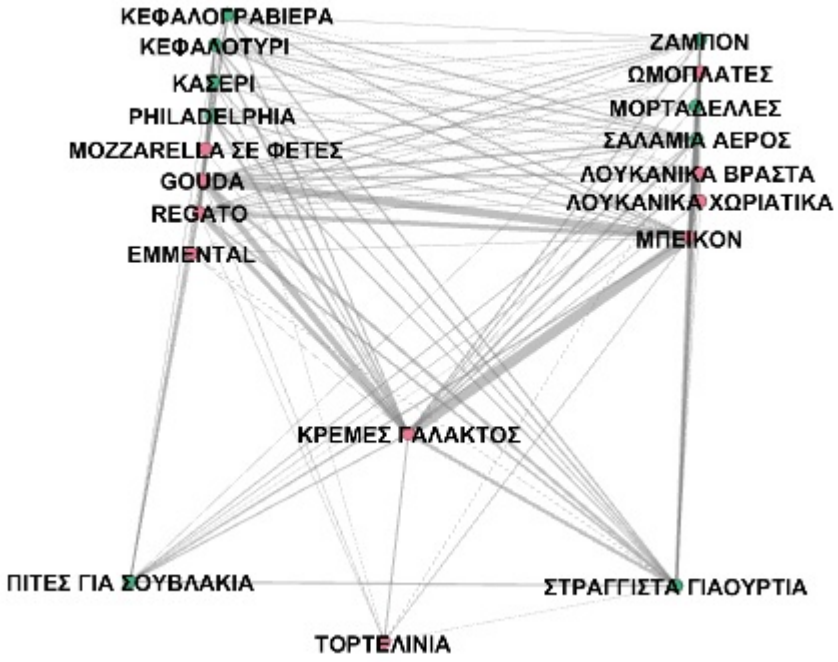
## Κοινότητα 5




Εικόνα 5.36: Κοινότητα “Προϊόντα Οργάνωσης Σπιτιού - Καθαριστικά”, 5η θέση

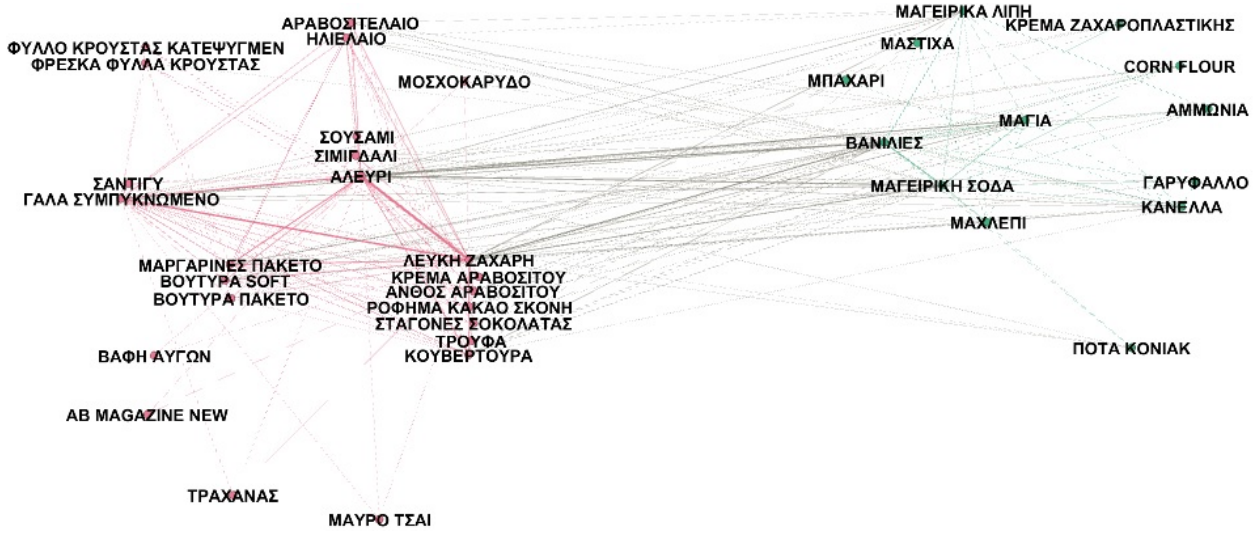
Η κοινότητα αυτή αποτελείται από προϊόντα οργάνωσης κουζίνας (αριστερά) και καθαριστικά (δεξιά).

## Κοινότητα 6




Εικόνα 5.37: Κοινότητα “Τυριά - Αλλαντικά”, 6η θέση

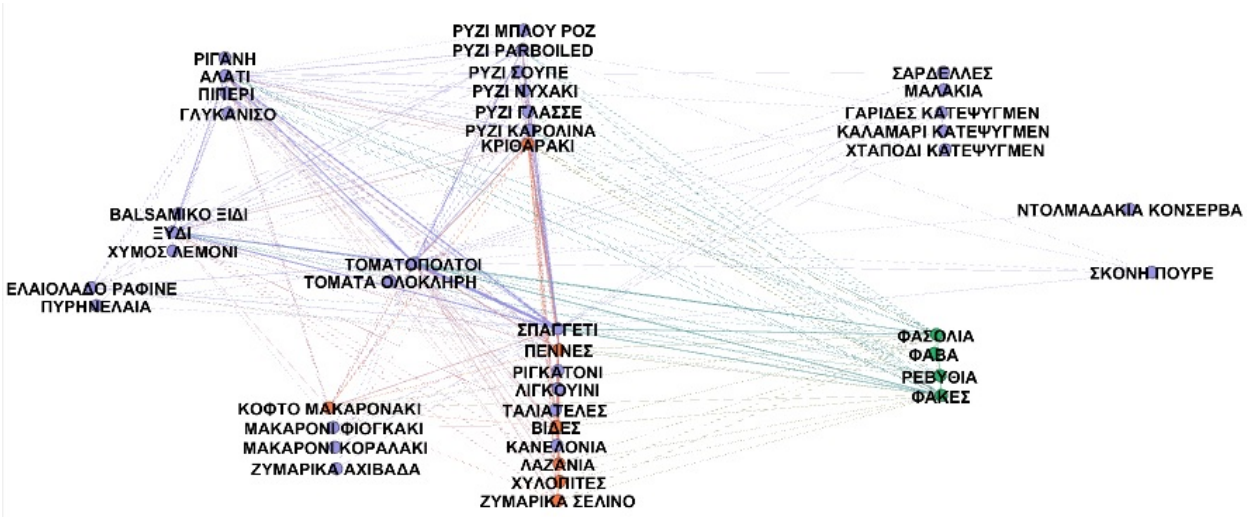
## Κοινότητα 7




Εικόνα 5.38: Κοινότητα “Προϊόντα Αρτοζαχαροπλαστικής”, 7η θέση

Η κοινότητα αυτή αποτελείται από προϊόντα ψησίματος και υλικά ζαχαροπλαστικής. Τα κύρια προϊόντα κάθε μιας υπο-κοινότητας φαίνεται να είναι το αλεύρι και η ζάχαρη (αριστερά), η μαγιά και οι βανίλιες (δεξιά).

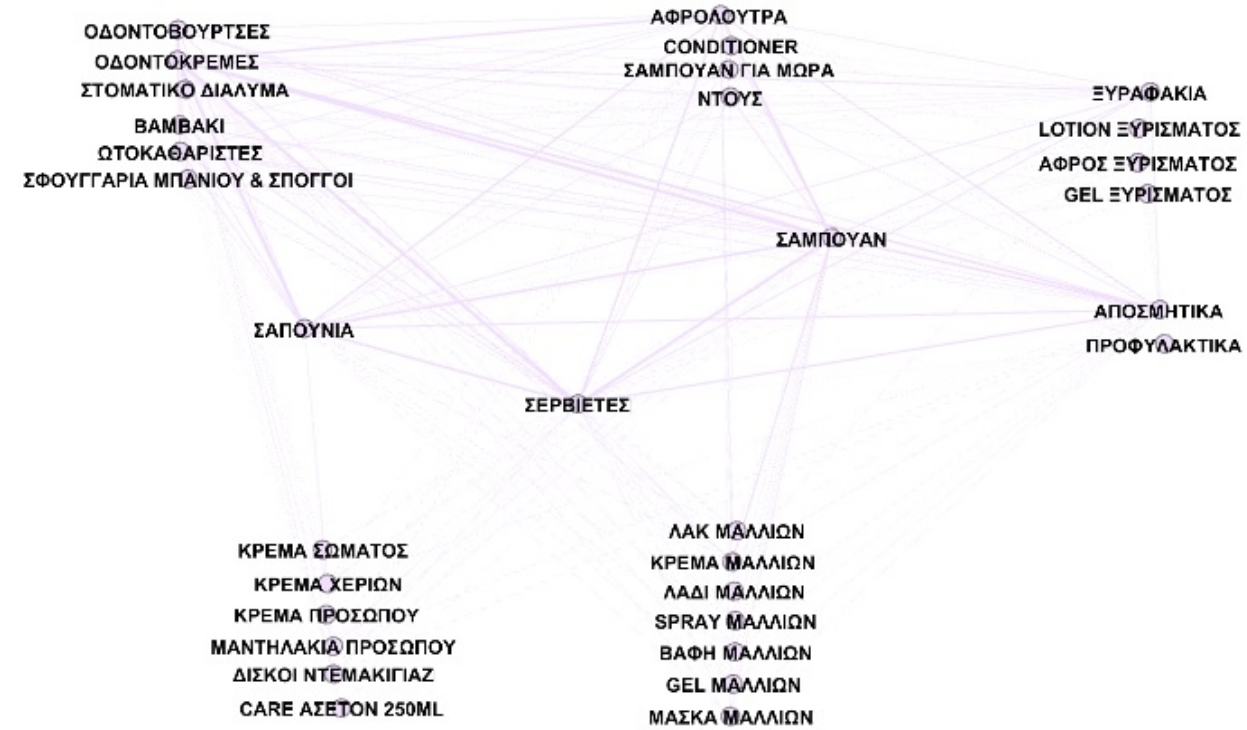
## Κοινότητα 8




Εικόνα 5.39: Κοινότητα “Ρύζια - Ζυμαρικά”, 8η θέση

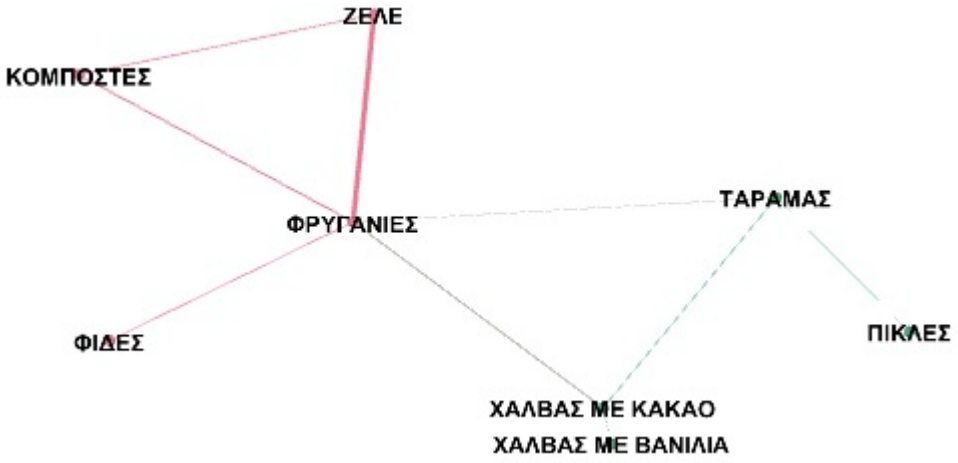
Στην κοινότητα αυτή κεντρικό ρόλο παίζουν οι τοματοπολτοί. Συνδέονται άμεσα με μπαχαρικά, είδη λαδιού και ξυδιού, όπως επίσης ζυμαρικά και ρύζια. Τα δύο τελευταία συνδέονται με όσπρια και κατεψυγμένα θαλασσινά.

## Κοινότητα 9




Εικόνα 5.40: Κοινότητα “Προσωπική Υγιεινή”, 10η θέση

## Κοινότητα 10

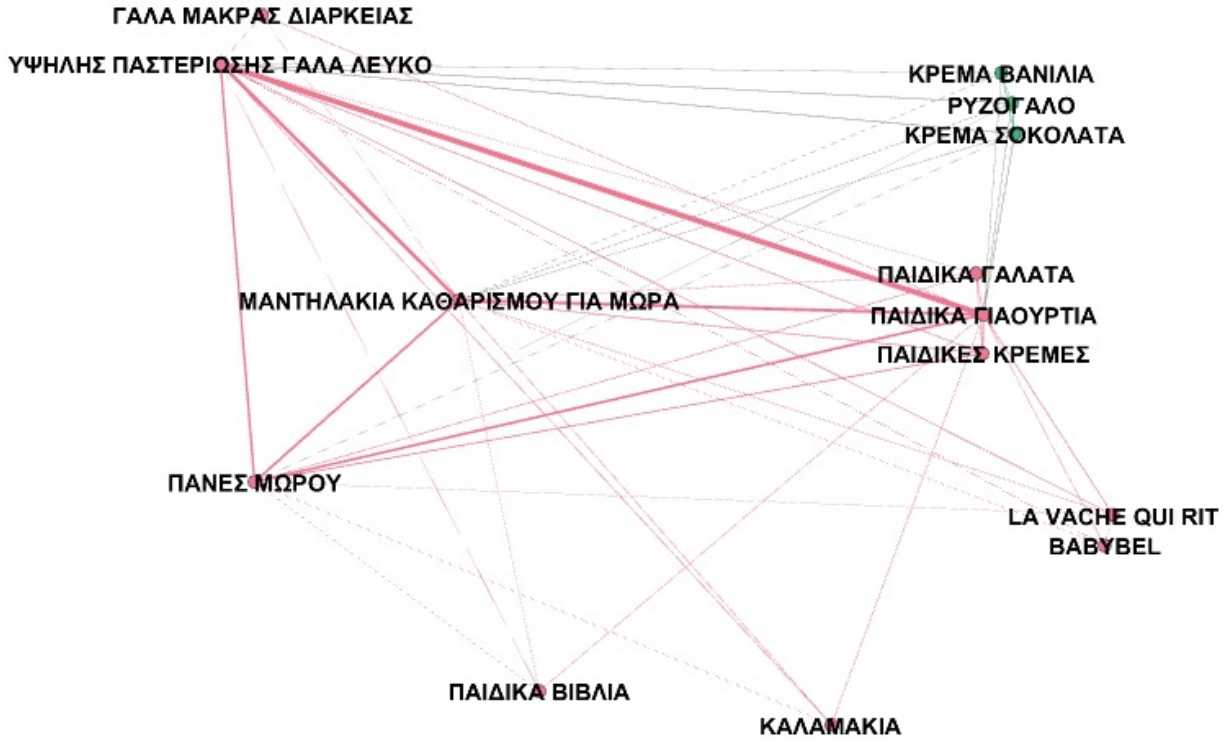



Εικόνα 5.41: Κοινότητα “Hub Φρυγανιών”, 11η θέση

Η σχέση των προϊόντων της δεξιά υπο-κοινότητας έχει εντοπιστεί ήδη από την $ARN$ τεχνική και αφορά σε νηστίσιμα προϊόντα. Η σύνδεσή τους με τις φρυγανιές δεν είναι ιδιαίτερα ισχυρή. Ωστόσο, οι φρυγανιές συνδέονται έντονα με τις κομπόστες και το ζελέ, προϊόντα δημοφιλή σε ανθρώπους της τρίτης ηλικίας.

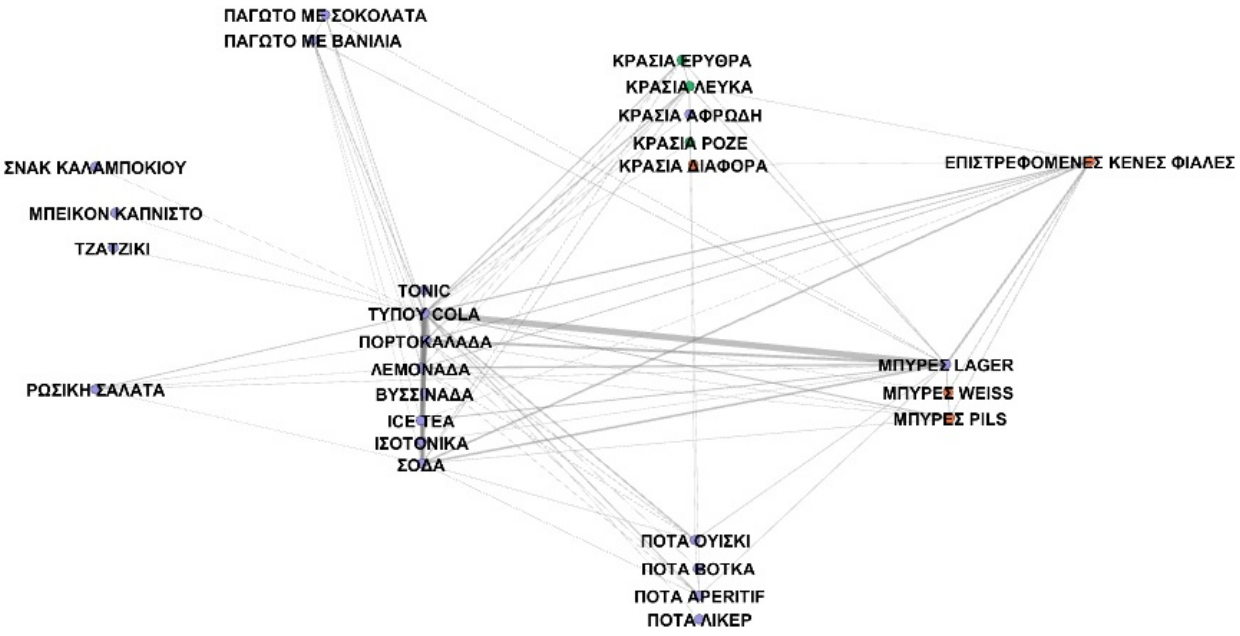
Σε αυτό το σημείο αξίζει να τονιστεί το εξής. Οι φρυγανιές, προφανώς αποτελούν το μοναδικό σημείο κοπής του γράφου και απ’ την αφαίρεσή τους προκύπτουν τρεις συνεκτικές συνιστώσες. Ωστόσο, πρέπει να είναι ξεκάθαρο ότι η παραπάνω κοινότητα αποτελεί μέρος της ανίχνευσης κοινοτήτων ολόκληρου του γράφου προϊόντων και όχι την “καλύτερη” κοινότητα συγκεκριμένα για τις φρυγανιές. Το τελευταίο πρόβλημα είναι γνωστό ως Αναζήτηση Κοινότητας (Community Search, ) και δεν το μελετήσαμε εδώ.

## Κοινότητα 11



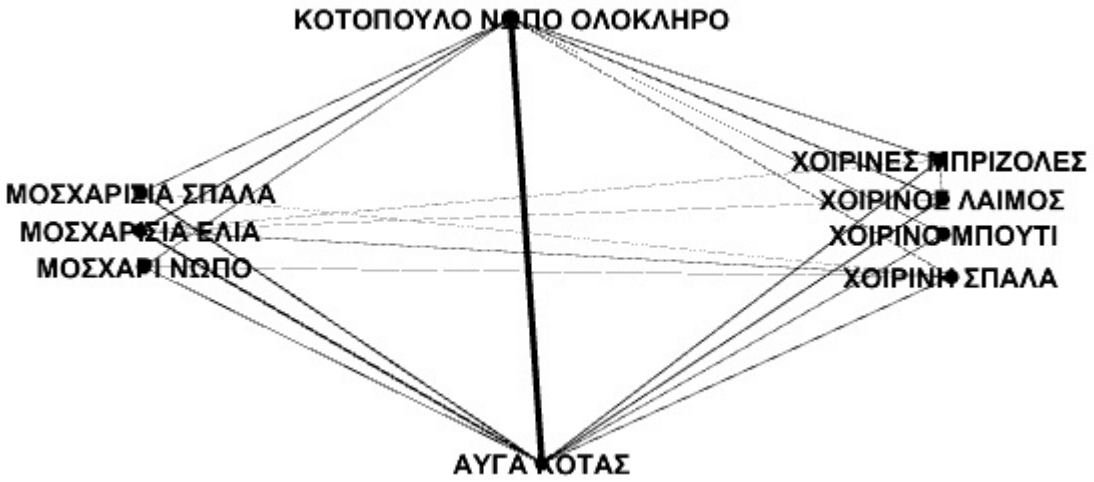

Εικόνα 5.42: Κοινότητα “Προϊόντα για παιδιά”, 12η θέση

## Κοινότητα 12




Εικόνα 5.43: Κοινότητα “Ποτά - Αναψυκτικά”, 13η θέση

## Κοινότητα 13




Εικόνα 5.44: Κοινότητα “Κρέατα”, 14η θέση

Η κοινότητα αυτή αποτελείται από διάφορα είδη κρέατος τα οποία συνδέονται στενά με τα αυγά. Όπως αναφέραμε και πριν, αυτή δεν είναι η καλύτερη κοινότητα των αυγών, αλλά μέρος της ανίχνευσης κοινοτήτων κατά τη διαχείριση ολόκληρου του γράφου προϊόντων.

## Κοινότητα 14

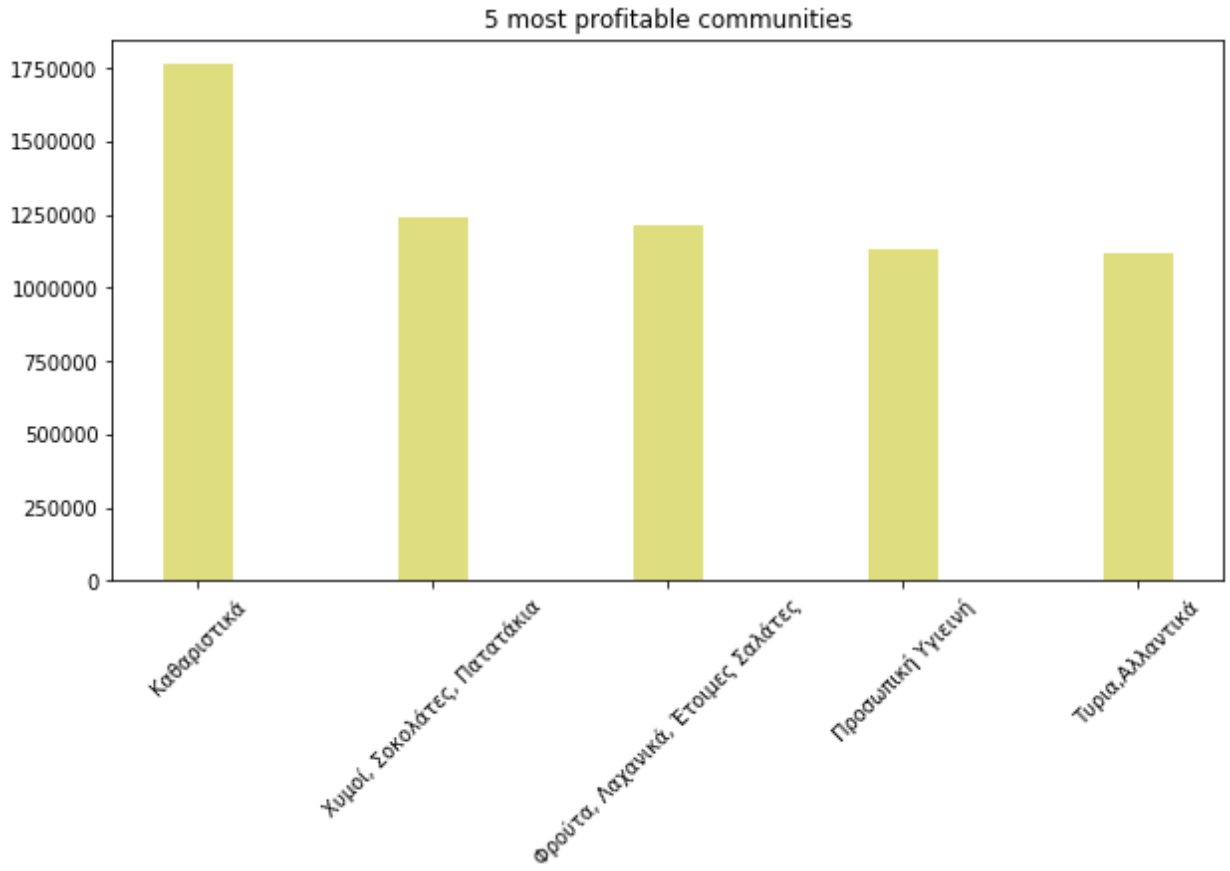

Εικόνα 5.45: Κοινότητα “Καφές”, 15η θέση

## 3.2.3 Τμηματοποίηση Καταναλωτών

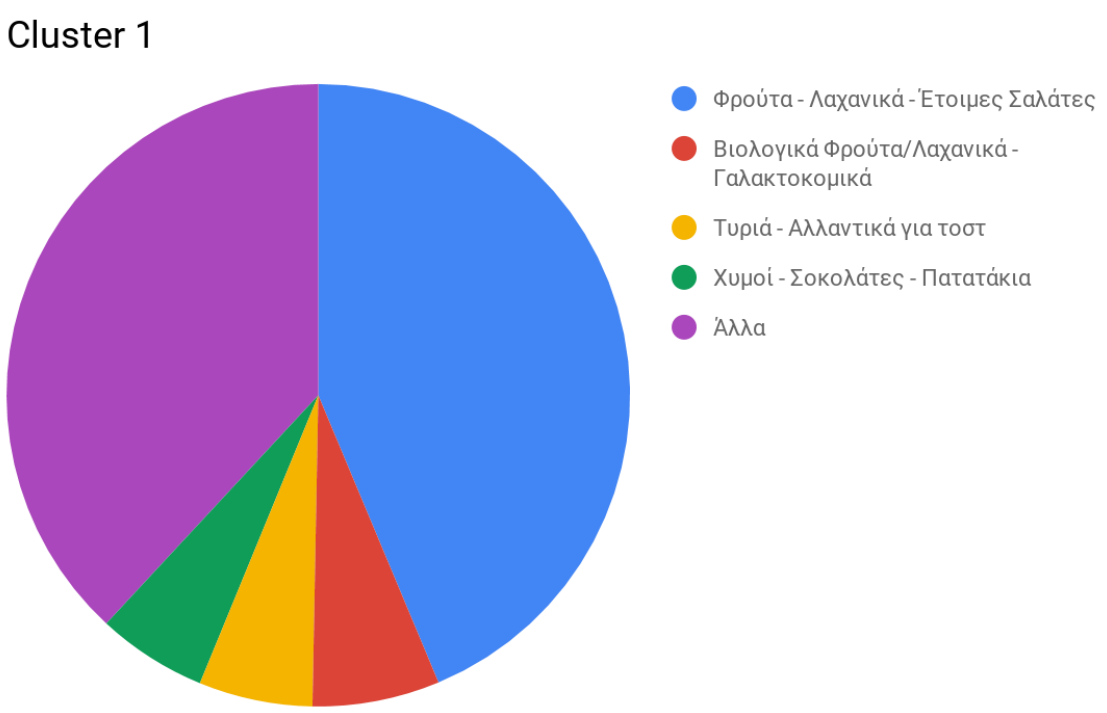
Η Τμηματοποίηση Καταναλωτών αφορά στην ομαδοποίηση των καταναλωτών με βάση τις αγοραστικές τους συνήθειες και πραγματοποιήθηκε σε δύο κατευθύνσεις. Αξιοποιήσαμε τις Κοινότητες Προϊόντων που ανιχνεύσαμε στην προηγούμενη ενότητα ώστε να βρούμε προφιλ καταναλωτών με κοινές κατηγορίες προϊόντων στις οποίες ξοδεύουν τα χρήματά τους ή τις προτιμούν.

### 3.2.3.1 Ομαδοποίηση με βάση την κερδοφορία

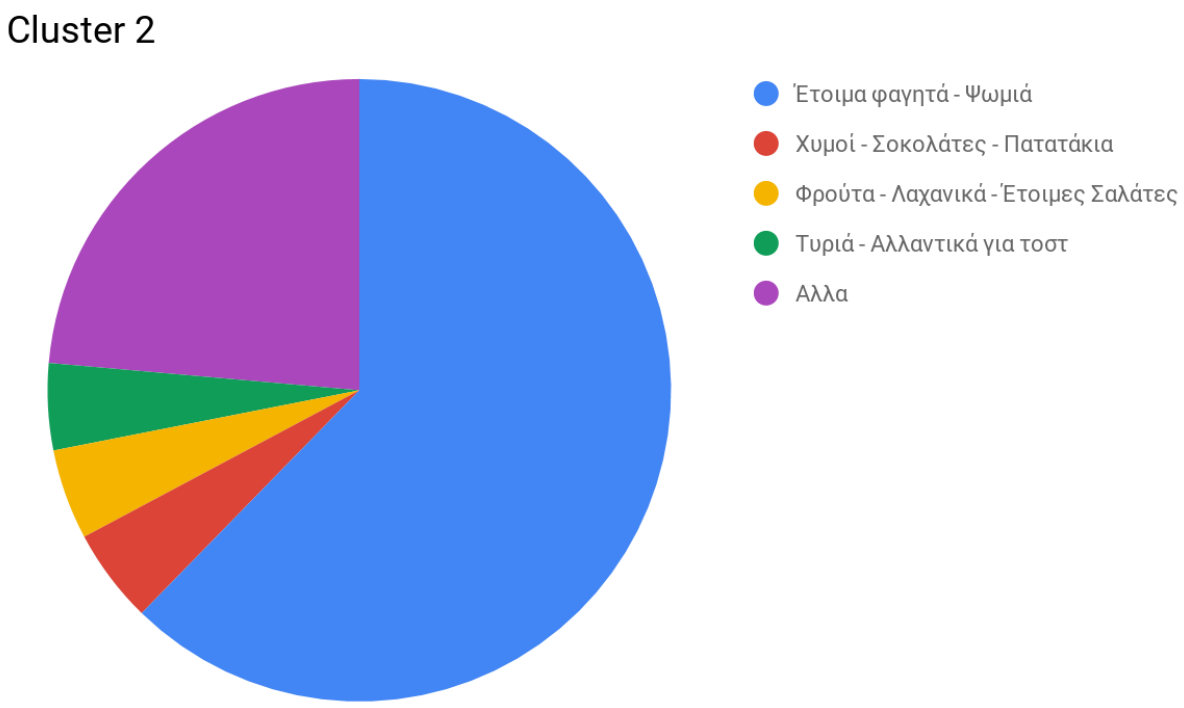
Στην περίπτωση αυτή αναζητήσαμε προφίλ καταναλωτών με βάσει τα χρήματα που ξοδεύουν ανά κατηγορία προϊόντων. Ισχύει ότι ένα άτομο, μια μικρή ή μια μεγάλη οικογένεια δύναται να ξοδεύουν τα χρήματά τους στις ίδιες αναλογίες μέσα στις Κοινότητες Προϊόντων. Ωστόσο, τα επιμέρους ποσά θα διαφέρουν σημαντικά. Έτσι υπολογίσαμε για κάθε καταναλωτή τι μέρος των συνολικών του χρημάτων ξοδεύει σε κάθε μια από τις 25 κοινότητες προϊόντων, καταλήξαμε σε διανύσματα του χώρου $R^{25}$ και προχωρήσαμε σε Clustering με τον αλγόριθμο k-Means.



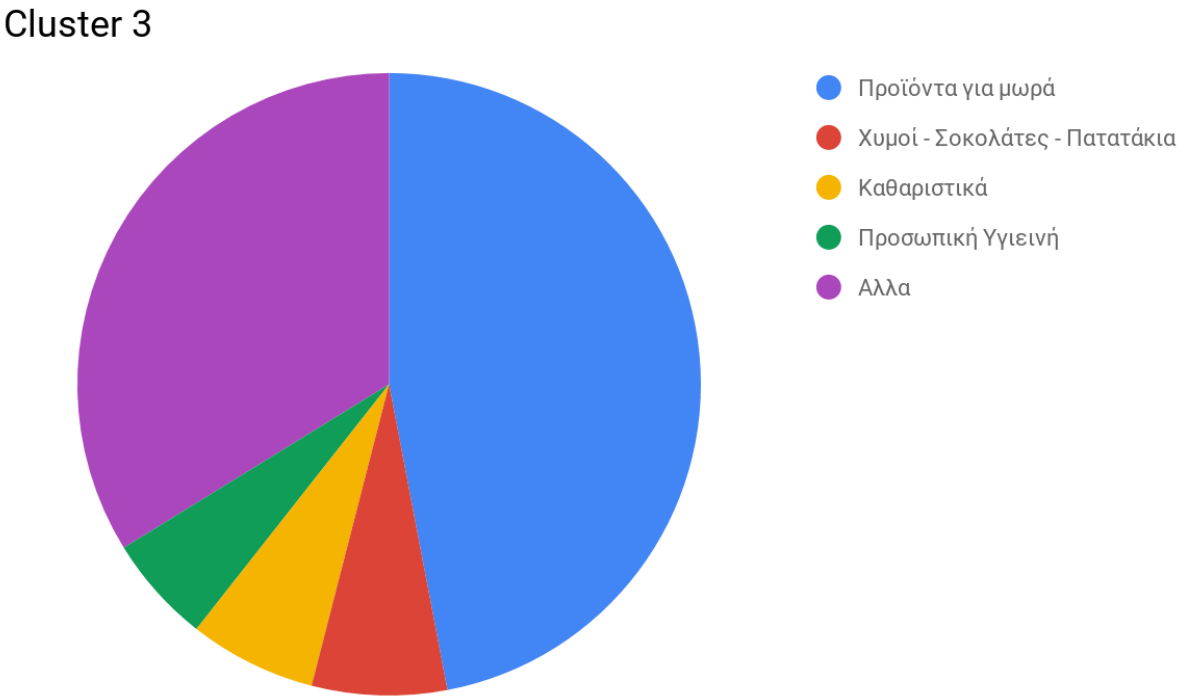

Εικόνα 5.46: Οι 5 πιο κερδοφόρες Κοινότητες Προϊόντων




Εικόνα 5.47: Cluster “Φρούτα - Λαχανικά - Έτοιμες Σαλάτες”

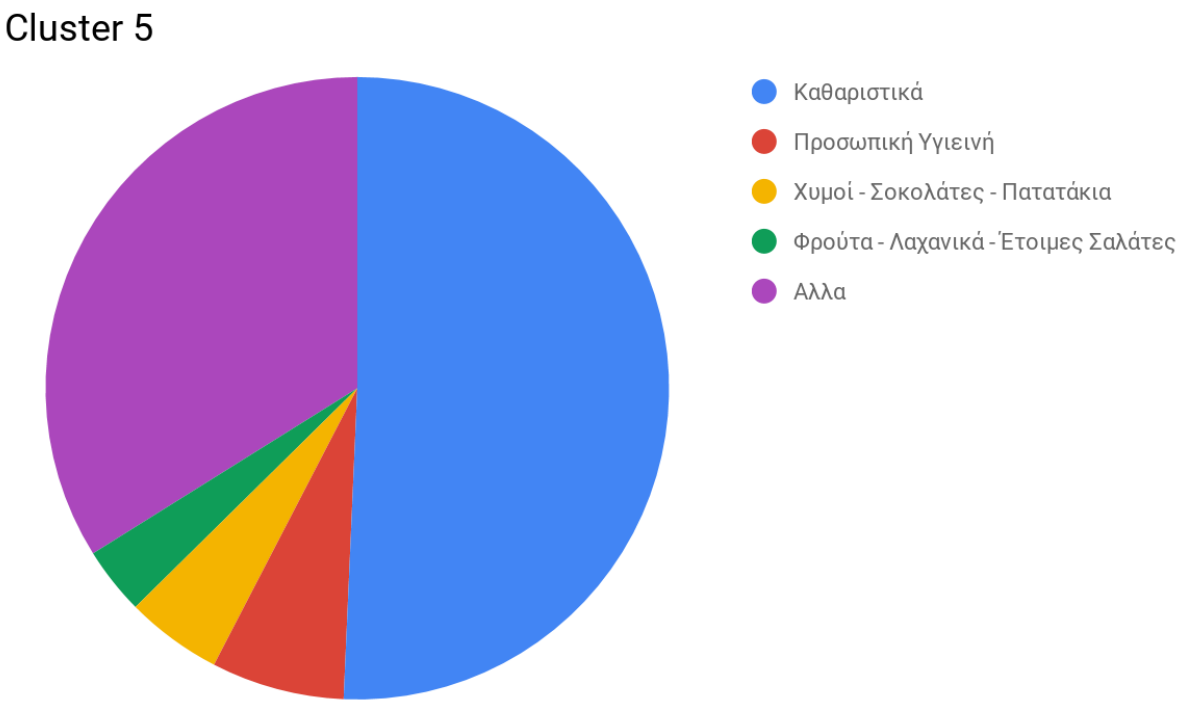



Εικόνα 5.48: Cluster “Έτοιμα Φαγητά - Ψωμιά”

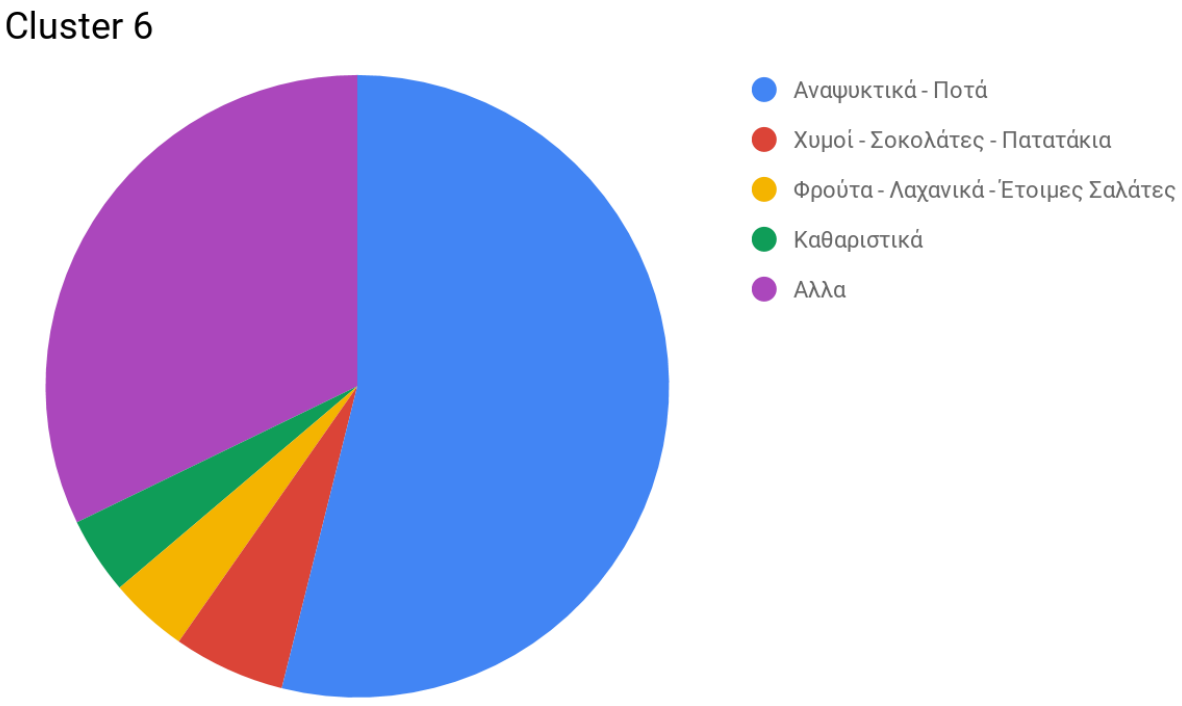



Εικόνα 5.49: Cluster “Προϊόντα για μωρά”

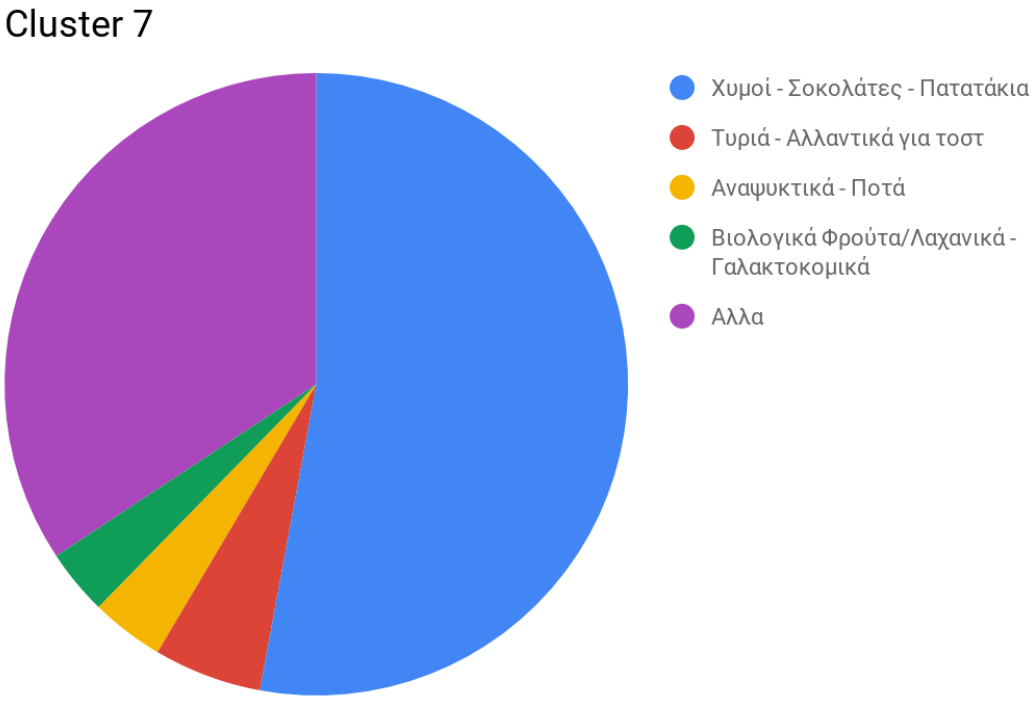

Εικόνα 5.50: Οι 5 πιο κερδοφόρες Κοινότητες Προϊόντων



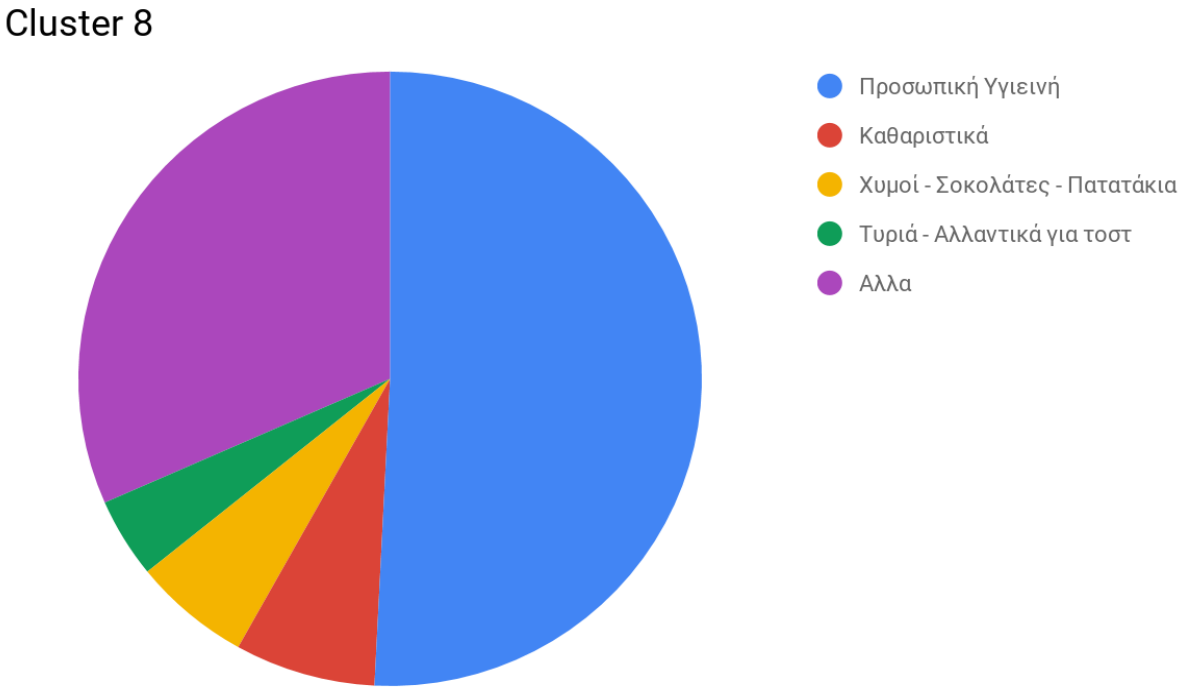

Εικόνα 5.50: Cluster “Καθαριστικά”



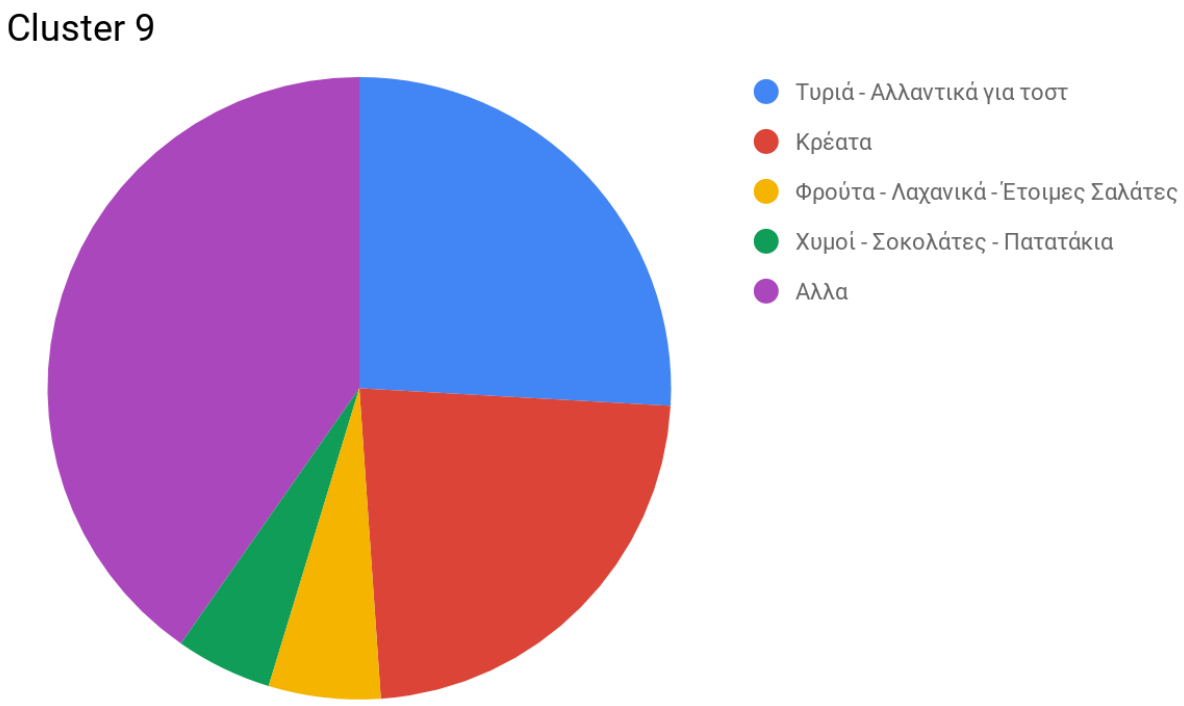

Εικόνα 5.51: Cluster “Αναψυκτικά - Ποτά”




Εικόνα 5.52: Cluster “Χυμοί - Σοκολάτες - Πατατάκια”



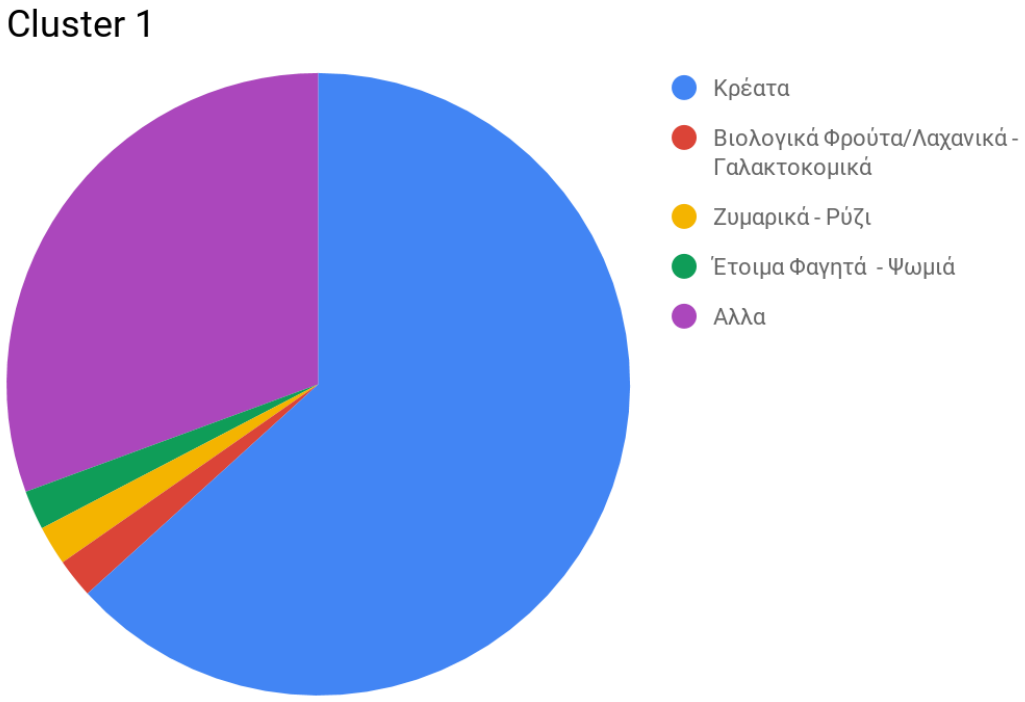

Εικόνα 5.53: Cluster “Προσωπική Υγιεινή”



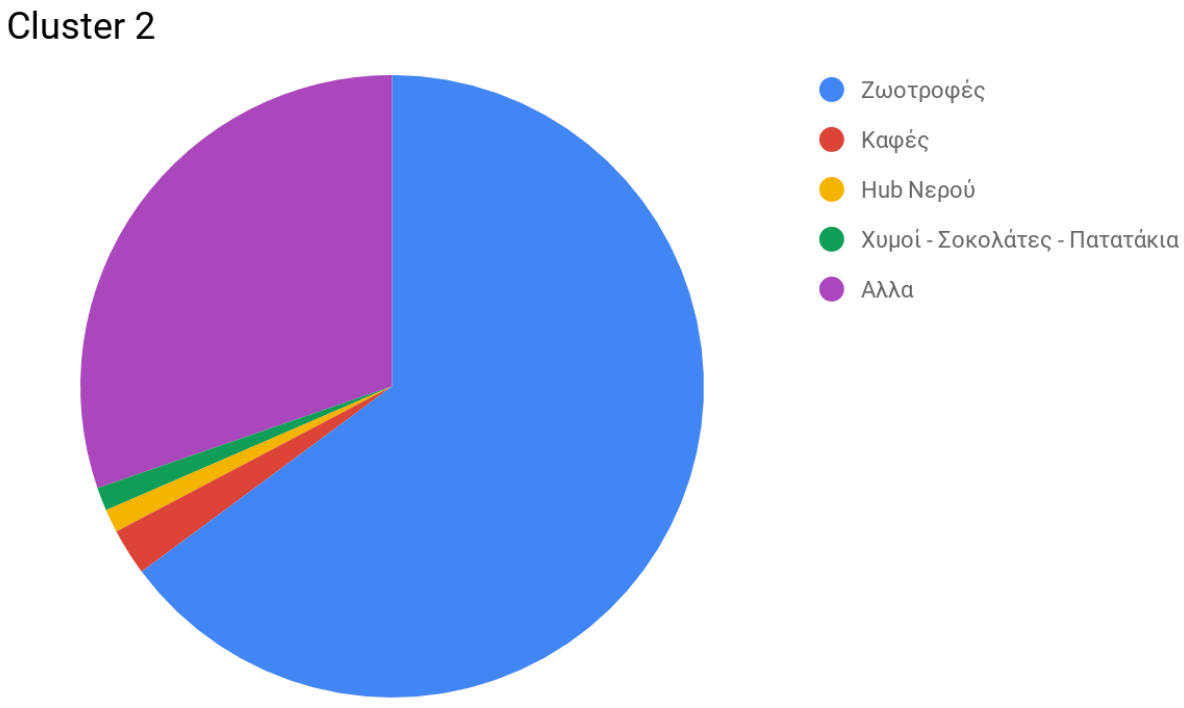

Εικόνα 5.54: Cluster “Τυριά - Αλλαντικά για τοστ”

### 3.2.3.2 Ομαδοποίηση με βάση την προτίμηση

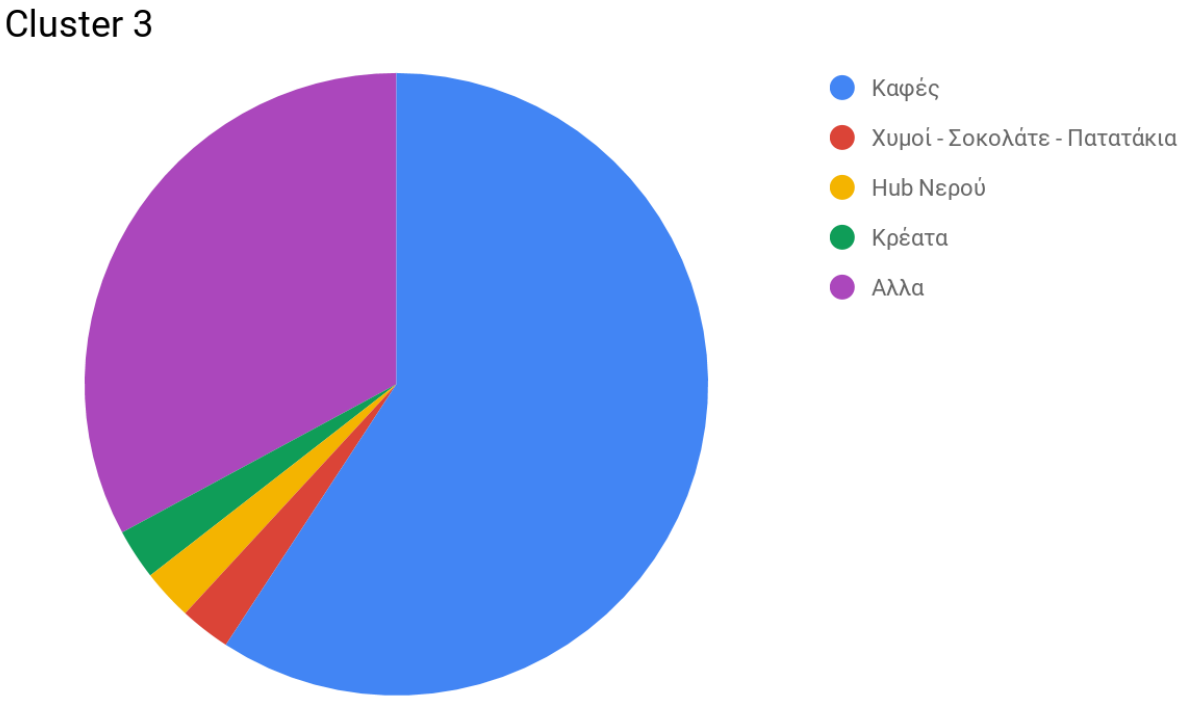
Στην περίπτωση αυτή αναζητήσαμε προφίλ καταναλωτών με βάσει την προτίμησή τους σε προϊόντα των προηγούμενων κοινοτήτων. Υπολογίσαμε, για κάθε καταναλωτή, τι μέρος των συνολικών προϊόντων που αγοράζει, αντιστοιχεί σε κάθε κοινότητα. Επίσης, για να εξαλείξουμε τον παράγοντα του μεγέθους κάθε κοινότητας, διαιρέσαμε τους αντίστοιχους όρους με το μέγεθος αυτό. Όπως και πριν, χρησιμοποιήσαμε τον k-Means.



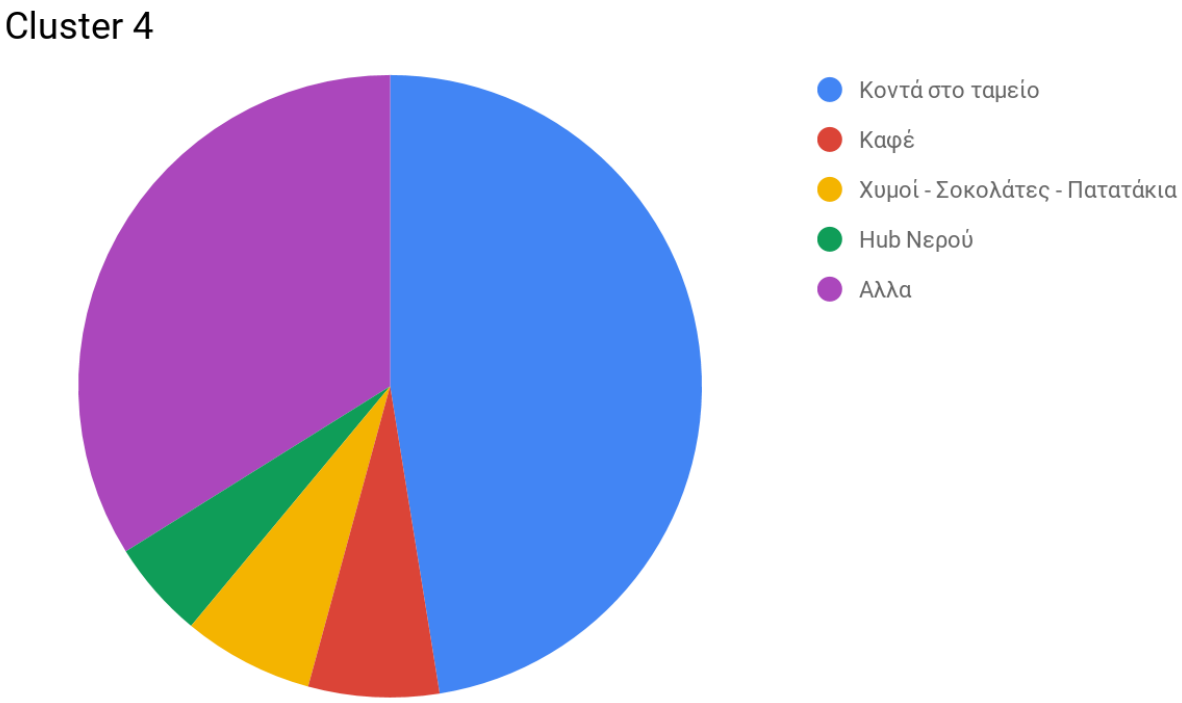

Εικόνα 5.55: Cluster “Κρέατα”



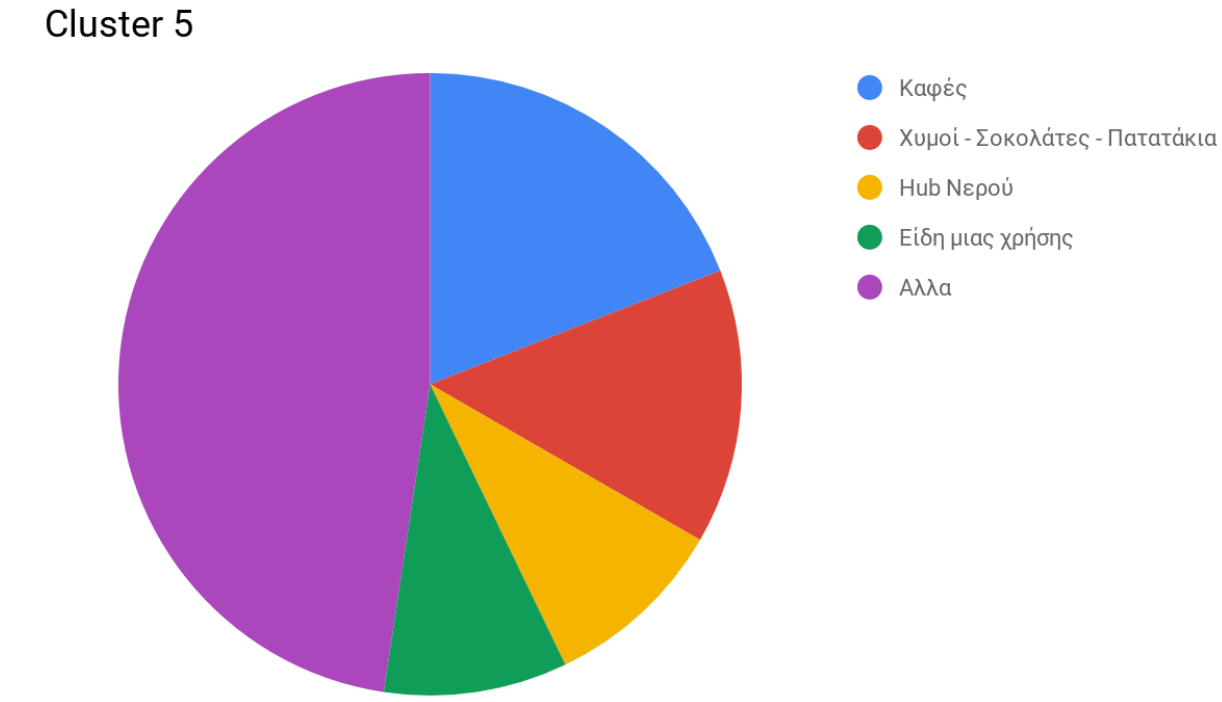

Εικόνα 5.56: Cluster “Ζωοτροφές”



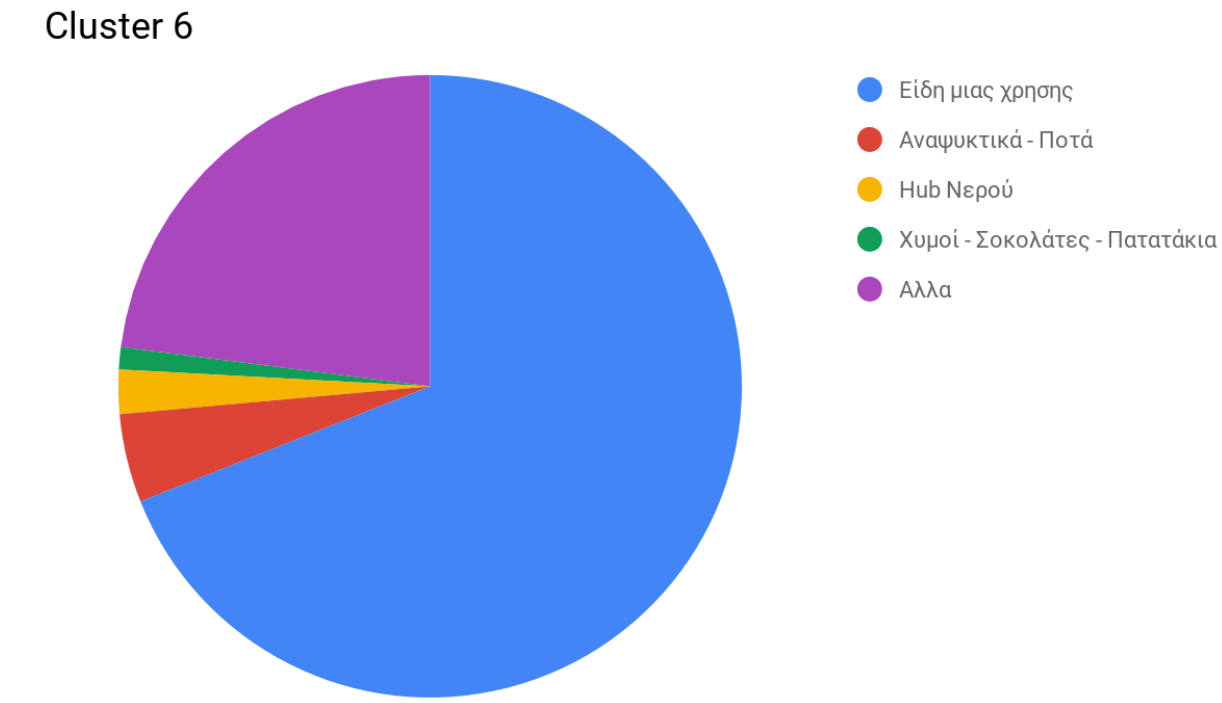

Εικόνα 5.57: Cluster “Καφές”



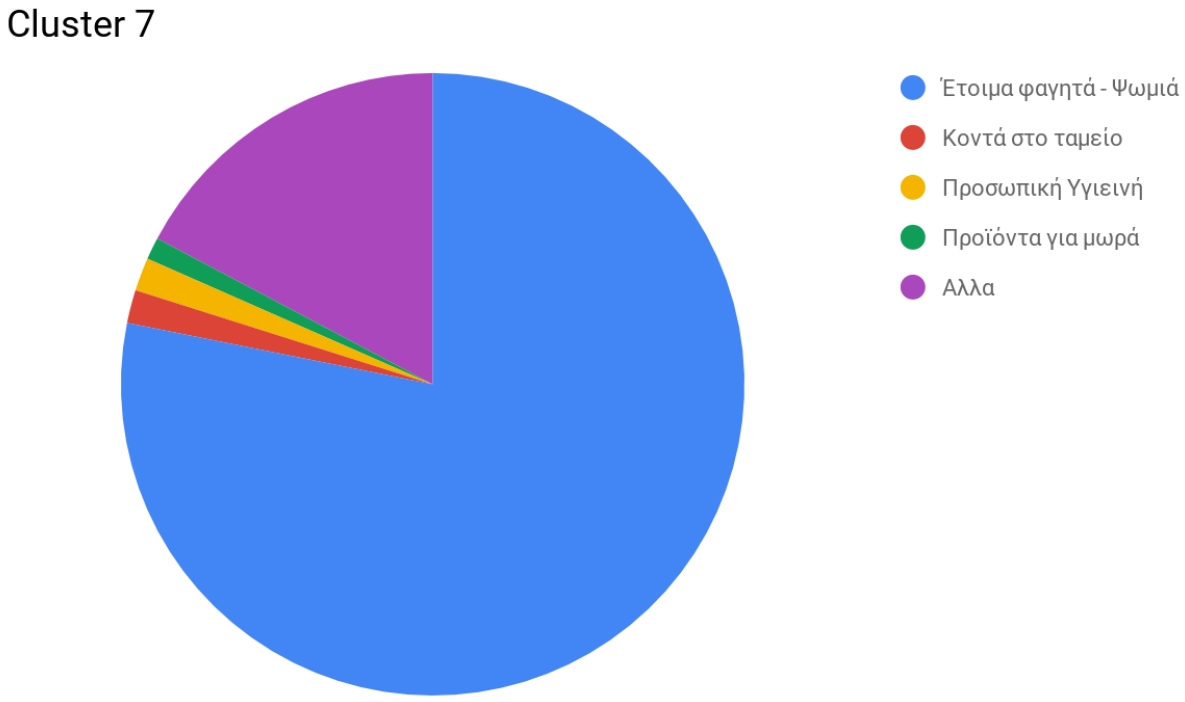

Εικόνα 5.58: Cluster “Κοντά στο ταμείο”




Εικόνα 5.59: Cluster Ουδέτερο




Εικόνα 5.60: Cluster “Είδη μιας χρήσης”




Εικόνα 5.61: Cluster “Έτοιμα Φαγητά - Ψωμιά”

# 4 Επίλογος

## 4.1 Σύνοψη και Συμπεράσματα

Σε αυτή την εργασία μελετήθηκαν διάφορες τεχνικές στην περιοχή της Ανάλυσης Καλαθιού Αγορών πάνω σε δεδομένα συναλλαγών από supermarket. Η ομαδοποίηση των προϊόντων, σε διάφορα επίπεδα, κρίθηκε αναγκαία για την βελτίωση των αποτελεσμάτων όλων των τεχνικών.

Στην τεχνική FIM υλοποιήσαμε αλγορίθμους της οικογένειας Apriori καθώς και τον αλγόριθμο Toivonen. Από τα συχνά στοιχειοσύνολα παράξαμε Κανόνες Συσχέτισης της μορφής $A \rightarrow B,\ \mu\varepsilon\ |\mathrm{B}| = 1.$ Στην συνέχεια, για κάθε κανόνα κρατήσαμε τις κατηγορίες των προϊόντων που συμμετέχουν σε αυτόν και τον βαθμολογήσαμε βάσει έξι αντικειμενικών μέτρων. Τέλος, δημιουργήσαμε διάφορες περιπτώσεις χρήσης κατά τις οποίες η Εξόρυξη Κανόνων έχει νόημα, παρέχοντας παραδείγματα. Παράμετροι την παραπάνω υλοποίησης είναι το προϊόν ή η κατηγορία προϊόντων ως $A$ και ως $B$, τα αντικειμενικά μέτρα με τα οποία θα βαθμολογηθούν οι κανόνες και το μέγεθος του $A$. Σημειώνεται ότι στην περίπτωση επιλογής δύο ή παραπάνω μέτρων, χρησιμοποιείται η πολυκριτηριακή μέθοδος Pareto.

Παρατηρήσαμε το πρόβλημα των δημοφιλών προϊόντων, δηλαδή την τάση κάποιοι κανόνες που έχουν ως $B$ ένα δημοφιλές προϊόν, να δέχονται υψηλή αξιολόγηση. Το φαινόμενο αυτό εξαλείφεται αν τεθεί κάποιο συγκεκριμένο προϊόν ως $B$, αλλά επίσης περιορίζεται αισθητά στην περίπτωση προϊόντος-στόχου ως $A$. Ένας ακόμη τρόπος αντιμετώπισης του προβλήματος είναι η χρήση μέτρων, όπως το “Added Value” ( $P[B|A]$ - $P[B]$ ) , που αφαιρεί από την “a posteriori” την “a priori”, αποδυναμώνοντας έτσι τους κανόνες που αποτελούν πρόβλημα.

Στην τεχνική ARN βρήκαμε τις έμμεσες συσχετίσεις για δεδομένα προϊόντα στόχους. Αξίζει να σημειωθεί ότι σε περιπτώσεις που ο αριθμός των κόμβων των ARN γράφων ήταν μεγάλος αλλά και για την ανάδειξη συσχετίσεων υψηλότερου επιπέδου, κρίθηκε απαραίτητη η χρήση ομαδοποιήσεων υψηλότερων επιπέδων.

Στην τεχνική Community Detection βρήκαμε κοινότητες προϊόντων τα οποία αγοράζονται συχνά μαζί, σε σχέση με προϊόντα άλλων κοινοτήτων. Προέκυψαν σχέσεις προϊόντων που προηγουμένως δεν είχαν φανεί και συγχρόνως επιβεβαιώθηκαν άλλες.

Αξιοποιώντας τις παραπάνω κοινότητες προχωρήσαμε σε Τμηματοποίηση Καταναλωτών, ομαδοποιώντας τους ως προς τις προτιμήσεις τους στις παραπάνω κοινότητες και τα χρήματα που ξοδεύουν σε αυτές.

## 4.2 Μελλοντικές επεκτάσεις

Παραθέτουμε παρακάτω μερικά προβλήματα που συναντήσαμε και θα θέλαμε να αντιμετωπίσουμε, καθώς και θέματα για μελλοντική έρευνα:

- Η Κατηγοριοποίηση Προϊόντων, ως βασικό στάδιο προ-επεξεργασίας, αποτελεί καθοριστική παράμετρο για την ποιότητα των αποτελεσμάτων των τεχνικών μας. Μια όχι τόσο καλή κατηγοριοποίηση μπορεί είτε να μην ομαδοποιήσει καν κάποια προϊόντα είτε να συμπεριλάβει επιπλέον προϊόντα σε κάποιες ομαδοποιήσεις. Έτσι, κάποιες ομάδες προϊόντων δύναται να ενισχυθούν ή να αποδυναμωθούν και για αυτό, η βοήθεια από κάποιον που γνωρίζει τον χώρο των πωλήσεων στα supermarket ,που έχουν ευρύ φάσμα κατηγοριών προϊόντων, θα ήταν ενισχυτική.

- Όπως έχουμε αναφέρει, το σύνολο των τεχνικών με τις οποίες ασχοληθήκαμε κάνει χρήση κατωφλίων. Αποτέλεσμα είναι ότι υπάρχουν προϊόντα, οι σχέσεις των οποίων δεν μπορούν να αναδειχθούν. Από την άλλη, η Αναζήτηση Κοινότητας (Community Search, Raeder et a. 2011), η οποία λόγω της χαμηλής σχετικά πολυπλοκότητας δεν κάνει κάποιο pruning, αναζητά τους κοντινότερους γείτονες, ως προς κάποιο ποσοτικό μέγεθος, ενός συνόλου κόμβων.

- Δύο επιπλέον μέθοδοι με ιδιαίτερο ενδιαφέρον σε πραγματικές εφαρμογές, είναι η Πρόβλεψη και η Σύσταση Προϊόντων. Η πρώτη αναζητά τα προϊόντα που θα συμμετέχουν στο επόμενο καλάθι ενός καταναλωτή, δηλαδή προϊόντα που έχει αγοράσει στο παρελθόν. Η δεύτερη ενδέχεται να έχει σαν αποτέλεσμα και νέα προϊόντα που με υψηλή πιθανότητα ενδέχεται να ενδιαφέρουν τον καταναλωτή με βάση τις καταναλωτικές του συνήθειες.

# 5 Βιβλιογραφία


Agarwal, R.C., Aggarwal, C.C. & Prasad, V.V.V., 2001. A Tree Projection Algorithm for Generation of Frequent Item Sets. *Journal of parallel and distributed computing*, 61(3), pp.350–371.

Agrawal, R., Imielinski, T. & Swami, A., 1993. Mining associations between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. pp. 207–216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*. pp. 487–499.

Ayan, N.F., Tansel, A.U. & Arkun, E., 1999. An Efficient Algorithm to Update Large Itemsets with Early Pruning. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '99. New York, NY, USA: ACM, pp. 287–291.

Baralis, E., Cerquitelli T., Chiusano S., Grand A., 2013. P-Mine: Parallel itemset mining on large datasets. *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. Available at: http://dx.doi.org/10.1109/icdew.2013.6547461.

Blondel, V.D., Guillaume J., Lambiotte R., Lefebvre E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), p.P10008. Available at: http://dx.doi.org/10.1088/1742-5468/2008/10/p10008.

Brin, S., Motwani R., Ullman J., Tsur S., 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *SIGMOD Rec.*, 26(2), pp.255–264.

Burdick, D., Calimlim, M. & Gehrke, J., MAFIA: a maximal frequent itemset algorithm for transactional databases. *Proceedings 17th International Conference on Data Engineering*. Available at: http://dx.doi.org/10.1109/icde.2001.914857.

Chawla, S., Arunasalam, B. & Davis, J., 2003. Mining Open Source Software (OSS) Data Using Association Rules Network. In *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, pp. 461–466.

Chawla, S., Davis, J. & Pandey, G., On local pruning of association rules using directed hypergraphs. *Proceedings. 20th International Conference on Data Engineering*. Available at: http://dx.doi.org/10.1109/icde.2004.1320063.

Cheung, D.W., Lee, S.D. & Kao, B., 1997. A General Incremental Technique for Maintaining Discovered Association Rules. In *Database Systems for Advanced Applications '97*. Advanced Database Research and Development Series. WORLD SCIENTIFIC, pp. 185–194.

Clark, P. & Boswell, R., 1991. Rule induction with CN2: Some recent improvements. In *Machine Learning — EWSL-91*. Springer Berlin Heidelberg, pp. 151–163.

Deng, Z.-H. & Lv, S.-L., 2014. Fast mining frequent itemsets using Nodesets. *Expert Systems with Applications*, 41(10), pp.4505–4512. Available at: http://dx.doi.org/10.1016/j.eswa.2014.01.025.

Deng, Z. & Wang, Z., 2010. A New Fast Vertical Method for Mining Frequent Patterns. *International Journal of Computational Intelligence Systems*, 3(6), p.733. Available at: http://dx.doi.org/10.2991/ijcis.2010.3.6.4.

Deng, Z., Wang, Z. & Jiang, J., 2012. A new algorithm for fast mining frequent itemsets using N-lists. *Science China Information Sciences*, 55(9), pp.2008–2030. Available at: http://dx.doi.org/10.1007/s11432-012-4638-z.

El-Hajj, M. & Zaiane, O.R., 2003. COFI-tree mining: a new approach to pattern growth with reduced candidacy generation. In *Workshop on Frequent Itemset Mining Implementations (FIMI'03) in conjunction with IEEE-ICDM*. Available at: http://www.academia.edu/download/30503821/fimi03.pdf.

Fang, M., Shivakumar N., Garcia M., Motwani R., Ullman D. 1999. Computing Iceberg Queries Efficiently. *Internaational Conference on Very Large Databases (VLDB'98), New York, August 1998*. Available at: http://ilpubs.stanford.edu:8090/423/ [Accessed May 13, 2019].

Feddaoui, I., Felhi, F. & Akaichi, J., 2016. EXTRACT: New extraction algorithm of association rules from frequent itemsets. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Available at: http://dx.doi.org/10.1109/asonam.2016.7752322.

Fortunato, S. & Barthelemy, M., 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), pp.36–41. Available at: http://dx.doi.org/10.1073/pnas.0605965104.

Geng, L. & Hamilton, H.J., 2006. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.*, 38(3). Available at: http://doi.acm.org/10.1145/1132960.1132963.

Gouda, K. & Zaki, M.J., Efficiently mining maximal frequent itemsets. *Proceedings 2001 IEEE International Conference on Data Mining*. Available at: http://dx.doi.org/10.1109/icdm.2001.989514.

Grahne, G. & Zhu, J., 2005. Fast algorithms for frequent itemset mining using FP-trees. *IEEE Transactions on Knowledge and Data Engineering*, 17(10), pp.1347–1362. Available at: http://dx.doi.org/10.1109/tkde.2005.166.

Hahsler, M., 2015. A probabilistic comparison of commonly used interest measures for association rules. *United States. Southern Methodist University*. Available at: https://michael.hahsler.net/research/association_rules/measures.html.

Han, J., Pei, J. & Yin, Y., 2000. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2), pp.1–12. Available at: http://dx.doi.org/10.1145/335191.335372.

Jamsheela, O. & G., R., 2015. Frequent itemset mining algorithms: A literature survey. *2015 IEEE International Advance Computing Conference (IACC)*. Available at:

http://dx.doi.org/10.1109/iadcc.2015.7154874.

Klösgen, W., 1992. Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora. *International Journal of Intelligent Systems*, 7(7), pp.649–673.

Lenca, P., Meyer P., Vaillant B., Lallich S. 2004. A multicriteria decision aid for interestingness measure selection. Available at: https://hal.archives-ouvertes.fr/hal-01853661/.

Lenca, P., Vaillant B., Meyer P., Lallich S., 2007. Association Rule Interestingness Measures: Experimental and Theoretical Studies. In F. J. Guillet & H. J. Hamilton, eds. *Quality Measures in Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 51–76.

Lucchese, C., Orlando, S. & Perego, R., 2006. Fast and memory efficient mining of frequent closed itemsets. *IEEE Transactions on Knowledge and Data Engineering*, 18(1), pp.21–36. Available at: http://dx.doi.org/10.1109/tkde.2006.10.

Lucchese, C., Orlando, S., Palmerini, P., Perego, R., Silvestri, F., 2004. kDCI: a Multi-Strategy Algorithm for Mining Frequent Sets.

Mannila, H., Toivonen, H. & Verkamo, A.I., 1994. Efficient algorithms association for discovering rules. In *Knowledge Discovery in Databases: AAAI Workshop*. pp. 181–192.

Ng, R. T., Lakshmanan, L. V. S., Pang, A., & Han, J., 1998. Exploratory Mining and Pruning Optimizations of Constrained Associations Rules. *SIGMOD Rec.*, 27(2), pp.13–24.

Omiecinski, E. & Savasere, A., 1998. Efficient mining of association rules in large dynamic databases. In *Advances in Databases*. Springer Berlin Heidelberg, pp. 49–63.

Orlando, S., Palmerini P., Perego R., Silvestri F. Adaptive and resource-aware mining of frequent sets. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* Available at: http://dx.doi.org/10.1109/icdm.2002.1183921.

Pandey, G., Chawla S., 2009. Association Rules Network: Definition and Applications. *Statistical analysis and data mining*, 1(4), pp.260–279.

Park, J.S., Chen, M.S. & Yu, P.S., 1995. An effective hash-based algorithm for mining association rules. Available at: https://dl.acm.org/citation.cfm?id=223813.

Pasquier, N., Bastide Y., Taouil R., Lakhal L., 1999. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1), pp.25–46. Available at: http://dx.doi.org/10.1016/s0306-4379(99)00003-4.

Patel, V. & Sahani, G.J., 2015. Image Classification using Frequent Itemset Mining. *International Journal of Computer Applications*, 121(15), pp.7–11. Available at: http://dx.doi.org/10.5120/21614-4880.

Pei, J., Han, J., Mao, R., 2000. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets

Piatetsky-Shapiro G., 1991. Discovery, Analysis, and Presentation of Strong Rules. *Knowledge*

*Discovery in Databases*, pp.229–238.

Pyun, G., Yun, U. & Ryu, K.H., 2014. Efficient frequent pattern mining based on Linear Prefix tree. *Knowledge-Based Systems*, 55, pp.125–139. Available at: http://dx.doi.org/10.1016/j.knosys.2013.10.013.

Qiao, M. & Zhang, D., 2012. EFFICIENTLY MATCHING FREQUENT PATTERNS BASED ON BITMAP INVERTED FILES BUILT FROM CLOSED ITEMSETS. *International Journal on Artificial Intelligence Tools*, 21(03), p.1250011. Available at: http://dx.doi.org/10.1142/s021821301250011x.

Raeder T., Chawla N.V., 2011 Market basket analysis with networks. Interdisciplinary Center for Network Science and Applications

Sahar, S. & Mansour, Y., 1999. Empirical evaluation of interest-level criteria. In *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*. Data Mining and Knowledge Discovery: Theory, Tools, and Technology. International Society for Optics and Photonics, pp. 63–74.

Savasere, A., Omiecinski, E.R. & Navathe, S.B., 1995. *An efficient algorithm for mining association rules in large databases*, Georgia Institute of Technology. Available at: https://smartech.gatech.edu/handle/1853/6678.

Smyth, P. & Goodman, R.M., 1992. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4), pp.301–316. Available at: http://dx.doi.org/10.1109/69.149926.

Song, M. & Rajasekaran, S., 2006. A transaction mapping algorithm for frequent itemsets mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(4), pp.472–481. Available at: http://dx.doi.org/10.1109/tkde.2006.1599386.

Swetha, M.H., Sivaselvan, B. & Oswald, C., 2018. Closed Frequent Itemset Mining Approach to Image Security Enhancement. *2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP)*. Available at: http://dx.doi.org/10.1109/icccsp.2018.8452846.

Tan, P.-N., Kumar, V. & Srivastava, J., 2002. Selecting the Right Interestingness Measure for Association Patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’02. New York, NY, USA: ACM, pp. 32–41.

Tan, P.-N., Kumar, V. & Srivastava, J., 2004. Selecting the right objective measure for association analysis. *Information systems*, 29(4), pp.293–313.

Toivonen, H., 1996. Sampling large databases for association rules. In *VLDB*. pp. 134–145.

Vaillant, B., Lenca, P. & Lallich, S., 2004. A Clustering of Interestingness Measures. In *Discovery Science*. Springer Berlin Heidelberg, pp. 290–297.

Vo, B., Hong, T.-P. & Le, B., 2012. DBV-Miner: A Dynamic Bit-Vector approach for fast mining frequent closed itemsets. *Expert Systems with Applications*, 39(8), pp.7196–7206. Available

at: http://dx.doi.org/10.1016/j.eswa.2012.01.062.

Wang, J., Han, J., Pei, J., 2003. CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets

Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W., 1997. New Algorithms for Fast Discovery of Association Rules

Zaki, M.J. & -J. Hsiao, C., 2005. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), pp.462–478. Available at: http://dx.doi.org/10.1109/tkde.2005.60.

Zhao, X., Zhang X., Wang P., Chen S., Sun Z., 2018. A weighted frequent itemset mining algorithm for intelligent decision in smart systems. *IEEE Access*, 6, pp.29271–29282. Available at: http://dx.doi.org/10.1109/access.2018.2839751.

Zou, Q., Chu, W.W. & Lu, B., SmartMiner: a depth first algorithm guided by tail information for mining maximal frequent itemsets. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* Available at: http://dx.doi.org/10.1109/icdm.2002.1184003.