



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης  
Σχολή Θετικών Επιστημών  
Τμήμα Πληροφορικής  
Πρόγραμμα Μεταπτυχιακών Σπουδών  
«Πληροφορική & Επικοινωνίες»

## Εξατομικευμένες προτάσεις προϊόντων σε διαδικτυακές αγορές

Διπλωματική εργασία του  
**Σταμάτη Σταματιάδη,**  
Μεταπτυχιακός φοιτητής  
(ΑΕΜ 641)

Σεπτέμβριος 2017



## Περίληψη

Η παρούσα εργασία αποτελεί μία προσπάθεια αντιμετώπισης του προβλήματος της ελλιπούς πυκνότητας το οποίο παρατηρείται κατά την εφαρμογή των συστημάτων συστάσεων σε περιβάλλοντα ηλεκτρονικού εμπορίου. Ενώ οι περισσότερες εργασίες βασίζονται σε ειδικά στάδια προεπεξεργασίας άμεσης ανατροφοδότησης, εναλλακτικές προσεγγίσεις προτείνουν την αξιοποίηση ενός μικρού ποσοστού των δεδομένων έμμεσης ανατροφοδότησης. Σχεδόν σε όλες τις περιπτώσεις, τα δεδομένα περιλαμβάνουν πληροφορίες σχετικά με τις αγορές των χρηστών ενός ηλεκτρονικού καταστήματος. Η εργασία παρουσιάζει μία νέα λύση στο πρόβλημα ελλιπούς πυκνότητας αξιοποιώντας μεγαλύτερο όγκο δεδομένων έμμεσης ανατροφοδότησης που σχετίζονται με επιπλέον ενέργειες των χρηστών.

**Λέξεις κλειδιά:** σύστημα συστάσεων, έλλειψη πυκνότητας, έμμεση ανατροφοδότηση, συνεργατικό φιλτράρισμα, ανάλυση πίνακα σε ιδιάζουσες τιμές

## Abstract

This work is an attempt to alleviate the problem of sparsity, observed during the application of recommender systems in an e-commerce environment. While most of the relevant work is based on special preprocessing of explicit feedback, alternative solutions propose exploiting only a tiny fraction of available implicit feedback. In most cases, the data includes only the purchase events of e-shop users. This work presents a novel solution to the sparsity problem by taking advantage of extra implicit data based on the actions users make during their online shopping experience.

**Keywords:** recommender system, sparsity, implicit feedback, collaborative filtering, singular value decomposition

**Translated title:** Personalized product recommendations for e-shopping

## Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστώ τον κ. Γούναρη Αναστάσιο για την καθοδήγηση που μου παρείχε στην εκπόνηση της διπλωματικής μου εργασίας ως επιβλέπων καθηγητής. Παράλληλα θα ήθελα να εκφράσω τις ευχαριστίες μου προς όλο το διδακτικό και διοικητικό προσωπικό του μεταπτυχιακού προγράμματος. Τα δύο αυτά χρόνια αποτέλεσαν μία πλούσια εμπειρία γεμάτη νέες γνώσεις και συνέβαλαν αναμφισβήτητα στην διεύρυνση των επαγγελματικών μου ικανοτήτων.

Επίσης θα ήθελα να ευχηθώ καλή επαγγελματική σταδιοδρομία σε όλους τους συνάδελφους συμφοιτητές. Η ανταλλαγή απόψεων, εντυπώσεων και εμπειριών σχετικά με το αντικείμενο της πληροφορικής αποτέλεσαν τεράστιο όφελος κατά τη διάρκεια των σπουδών μου.

Τέλος, με την ολοκλήρωση της παρούσας εργασίας θα ήθελα να εκφράσω την τεράστια ευγνωμοσύνη μου απέναντι στην οικογένειά μου καθώς στους πολύ κοντινούς μου ανθρώπους. Η ηθική τους στήριξη με κρατάει όρθιο μέσα σε αυτόν τον περίεργο κόσμο.

Σταμάτης Σταματιάδης  
Θεσσαλονίκη, 27/09/2017



## Περιεχόμενα

Περίληψη.....	3
Abstract .....	4
Ευχαριστίες.....	5
Περιεχόμενα.....	7
Κατάλογος εικόνων .....	9
Κατάλογος πινάκων.....	10
Κατάλογος διαγραμμάτων .....	11
1 Εισαγωγή .....	13
1.1 Περιγραφή προβλήματος.....	13
1.2 Προτεινόμενη λύση.....	13
1.3 Οργάνωση εργασίας.....	14
2 Συστήματα συστάσεων .....	16
2.1 Εισαγωγή .....	16
2.2 Δεδομένα εισόδου .....	18
2.2.1 Άμεση ανατροφοδότηση.....	18
2.2.2 Έμμεση ανατροφοδότηση .....	19
2.2.3 Συσχέτιση άμεσης και έμμεσης ανατροφοδότησης .....	20
2.3 Αλγόριθμοι & τεχνικές.....	20
2.3.1 Προσέγγιση βασισμένη στο περιεχόμενο.....	20
2.3.2 Προσέγγιση συνεργατικού φιλτραρίσματος.....	21
2.3.3 Προσέγγιση ανάλυσης πίνακα σε ιδιάζουσες τιμές .....	22
2.3.4 Προσέγγιση παραγοντοποίησης πινάκων.....	24
2.3.5 Προσεγγίσεις κατασκευής μοντέλου .....	25
2.4 Προβλήματα & προκλήσεις.....	26
2.4.1 Έλλειψη πυκνότητας .....	26
2.4.2 Αδυναμία κλιμάκωσης .....	27
2.4.3 Πολυσημία.....	27
2.4.4 Απώλεια μεταβατικής σχέσης.....	27
2.4.5 Άγνωστα προϊόντα & χρήστες.....	27
2.4.6 Υπερεπάρκεια δεδομένων .....	27
2.5 Αξιολόγηση συστημάτων σύστασης .....	27
2.5.1 Πρόβλημα πρόβλεψης .....	28
2.5.2 Πρόβλημα n-καλύτερων προτάσεων .....	28
2.6 Εφαρμογές.....	29
2.6.1 Συστήματα συστάσεων στο ηλεκτρονικό εμπόριο .....	29

2.6.2 Συστήματα συστάσεων σε άλλα πεδία εφαρμογής.....	31
3 Αξιολόγηση αντίκτυπου έμμεσης ανατροφοδότησης στο ηλεκτρονικό εμπόριο .....	33
3.1 Κίνητρο .....	33
3.2 Σχετική έρευνα .....	33
3.3 Σύνολα δεδομένων.....	34
3.3.1 Πηγές δεδομένων.....	34
3.3.2 Ανάλυση αρχικών δεδομένων.....	35
3.4 Προεπεξεργασία Δεδομένων .....	36
3.4.1 Παραδοχές.....	36
3.4.2 Στάδια προεπεξεργασίας .....	37
3.4.3 Ανάλυση επεξεργασμένων δεδομένων .....	37
3.4.4 Ανάθεση τιμών προτίμησης .....	39
3.5 Αλγόριθμοι παραγωγής συστάσεων .....	41
3.5.1 Συνεργατικό φιλτράρισμα.....	41
3.5.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές .....	43
3.6 Πειράματα .....	44
3.6.1 Σύνολα ελέγχου .....	44
3.6.2 Παραμετροποίηση αλγορίθμων.....	45
3.6.3 Οργάνωση πειραμάτων & αξιολόγηση .....	45
3.7 Αποτελέσματα πειραμάτων .....	46
3.7.1 Συνεργατικό φιλτράρισμα.....	46
3.7.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές .....	49
3.7.3 Σύγκριση των δύο μεθόδων .....	52
4 Δοκιμαστική ιστοσελίδα.....	55
4.1 Τεχνολογίες υλοποίησης.....	55
4.2 Παρουσίαση διεπαφής .....	55
5 Επίλογος .....	60
5.1 Τελικά συμπεράσματα .....	60
5.2 Μελλοντικές επεκτάσεις .....	60
Βιβλιογραφία .....	62



## Κατάλογος εικόνων

Εικόνα 1: Πίνακας εισόδου άμεσης ανατροφοδότησης - πηγή: διαδίκτυο .....	18
Εικόνα 2: Παραγωγή συστάσεων με βάση το περιεχόμενο - πηγή: διαδίκτυο .....	20
Εικόνα 3: Συνεργατικό φιλτράρισμα - πηγή: διαδίκτυο .....	22
Εικόνα 4: Ανάλυση πίνακα σε ιδιάζουσες τιμές - πηγή: διαδίκτυο .....	23
Εικόνα 5: Συστάσεις προϊόντων της Amazon - πηγή: amazon.co.uk .....	29
Εικόνα 6: Παροχή άμεσης ανατροφοδότησης στην Amazon - πηγή: amazon.co.uk.....	30
Εικόνα 7: Συστάσεις προϊόντων του Banggood - πηγή: banggood.com .....	30
Εικόνα 8: Συστάσεις προϊόντων του Banggood - πηγή: banggood.com .....	31
Εικόνα 9: Συστάσεις βίντεο του YouTube - πηγή: YouTube.com .....	31
Εικόνα 10: Συστάσεις μουσικής του Spotify - πηγή: Spotify .....	31
Εικόνα 11: Απαλοιφή οντοτήτων με μόνο μία ενέργεια αγοράς.....	38
Εικόνα 12: Ταξινόμηση ενεργειών ηλεκτρονικού καταστήματος σε ομάδες – πηγή: [19] ....	39
Εικόνα 13: Μορφές πινάκων εισόδου, (α) μόνο αγορές (β) όλες οι ενέργειες - πηγή:[25] ...	44
Εικόνα 14: Οθόνη εισόδου, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος .....	55
Εικόνα 15: Κεντρικό μενού, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος .....	56
Εικόνα 16: Κεντρικό μενού 2, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος .....	56
Εικόνα 17: Οθόνη αγαπημένων, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος.....	57
Εικόνα 18: Καλάθι αγορών, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος .....	57
Εικόνα 19: Επιτυχημένη αγορά, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος .....	57
Εικόνα 20: Οθόνη διαχείρισης, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος.....	58

## Κατάλογος πινάκων

Πίνακας 1: Μεγέθη των συνόλων δεδομένων.....	35
Πίνακας 2: Ενέργειες έμμεσης ανατροφοδότησης ανά σύνολο δεδομένων .....	35
Πίνακας 3: Μεγέθη των συνόλων δεδομένων μετά την προεπεξεργασία.....	37
Πίνακας 4: Πυκνότητα συνόλων δεδομένων .....	38
Πίνακας 5: Συνολικός αριθμός κάθε τύπου ενέργειας ανά σύνολο δεδομένων .....	40
Πίνακας 6: Τελικά σκορ που αντικαθιστούν κάθε ενέργεια με μία τιμή προτίμησης .....	40
Πίνακας 7: Μέγεθος συνόλου ελέγχου ανά σύνολο δεδομένων .....	45
Πίνακας 8: Τιμές παραμέτρων για τις οποίες εκτελέστηκαν τα πειράματα.....	45

## Κατάλογος διαγραμμάτων

Διάγραμμα 1: Συχνότητα εμφάνισης κάθε τύπου ενέργειας ανά σύνολο δεδομένων.....	38
Διάγραμμα 2: Συνεργατικό φιλτράρισμα, Retailrocket, σχέση topS και F1 με topN = 1.....	46
Διάγραμμα 3: Συνεργατικό φιλτράρισμα, Retailrocket, σχέση topN και F1 με topS = 15.....	47
Διάγραμμα 4: Συνεργατικό φιλτράρισμα, Alibaba, σχέση topS και F1 με topN = 1.....	47
Διάγραμμα 5: Συνεργατικό φιλτράρισμα, Alibaba, σχέση topN και F1 με topS = 20.....	48
Διάγραμμα 6: Συνεργατικό φιλτράρισμα, Tmall, σχέση topS και F1 με topN = 1 .....	48
Διάγραμμα 7: Συνεργατικό φιλτράρισμα, Tmall, σχέση topN και F1 με topS = 20 .....	49
Διάγραμμα 8: Ανάλυση σε ιδιάζουσες τιμές, Retailrocket, σχέση k και F1 με topN = 1 .....	49
Διάγραμμα 9: Ανάλυση σε ιδιάζουσες τιμές, Retailrocket, σχέση topN και F1 με k = 50 .....	50
Διάγραμμα 10: Ανάλυση σε ιδιάζουσες τιμές, Alibaba, σχέση k και F1 με topN = 1 .....	50
Διάγραμμα 11: Ανάλυση σε ιδιάζουσες τιμές, Alibaba, σχέση topN και F1 με k = 50 .....	51
Διάγραμμα 12: Ανάλυση σε ιδιάζουσες τιμές, Tmall, σχέση k και F1 με topN = 1.....	51
Διάγραμμα 13: Ανάλυση σε ιδιάζουσες τιμές, Tmall, σχέση topN και F1 με k = 55.....	52
Διάγραμμα 14: Σύγκριση τεχνικών, Retailrocket, σχέση topN και F1 με topS = 15, k = 50....	52
Διάγραμμα 15: Σύγκριση τεχνικών, Alibaba, σχέση topN και F1 με topS = 20, k = 50 .....	53
Διάγραμμα 16: Σύγκριση τεχνικών, Tmall, σχέση topN και F1 με topS = 20, k = 55.....	53



## 1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μία εισαγωγή στο αντικείμενο της παρούσας εργασίας. Ο αναγνώστης εισάγεται στη φύση του προβλήματος που καλείται να επιλύσει η εργασία καθώς και στη λύση που προτείνεται. Επίσης παρουσιάζεται η οργάνωση και δομή της εργασίας.

### 1.1 Περιγραφή προβλήματος

Αδιαμφισβήτητο το διαδίκτυο χαρακτηρίζεται από τεράστιους όγκους δεδομένων, ο ρυθμός αύξησης των οποίων ενισχύεται ιδιαίτερα τα τελευταία χρόνια. Το γεγονός αυτό οδήγησε σε μία νέα ανάγκη του ανθρώπου να δαμάσει τους τεράστιους αυτούς όγκους πληροφοριών προκειμένου να αξιοποιήσει την αποθηκευμένη γνώση του διαδικτύου.

Το πρόβλημα της εύρεσης κατάλληλου περιεχομένου, αποτελεί ένα από τα σημαντικότερα προβλήματα που καλείται να λύσει όχι μόνο η ακαδημαϊκή αλλά ολόκληρη η τεχνολογική και επιχειρηματική κοινότητα. Το πρόβλημα εμφανίζεται σε διάφορα πεδία τεχνολογικών εφαρμογών με διάφορους τύπους περιεχομένου (βίντεο, μουσική, άρθρα κ.α.). Υπάρχει μεγάλη ανάγκη για την δημιουργία συστημάτων που βοηθούν τον άνθρωπο στην αναζήτηση, εξέταση και επιλογή περιεχομένου που τον ενδιαφέρει.

Τα συστήματα συστάσεων έρχονται ως λύση στο παραπάνω πρόβλημα. Τα συστήματα αυτά παράγουν εξατομικευμένες προτάσεις προς τους χρήστες με περιεχόμενο που θα ικανοποιήσει τις ανάγκες τους.

Η παρούσα εργασία εστιάζει στα συστήματα παραγωγής συστάσεων με εφαρμογή τα ηλεκτρονικά καταστήματα και τις δυσκολίες που αντιμετωπίζουν. Η εφαρμογή των συστημάτων συστάσεων στον τομέα του ηλεκτρονικού εμπορίου χαρακτηρίζεται συνήθως από έλλειψη δεδομένων εισόδου. Τα δεδομένα αυτά αφορούν την εικόνα που έχουν οι χρήστες για τα προϊόντα του καταστήματος και συγκεντρώνονται υπό την μορφή αξιολογήσεων.

### 1.2 Προτεινόμενη λύση

Λύση στο παραπάνω πρόβλημα αποτελεί η εξασφάλιση μεγαλύτερου όγκου δεδομένων εισόδου. Δυστυχώς όμως το εγχείρημα αυτό είναι ιδιαίτερα δύσκολο και εξαρτάται από την προθυμία των χρηστών να μοιραστούν τις εμπειρίες αγοράς τους με το ηλεκτρονικό κατάστημα.

Στην παρούσα εργασία προτείνουμε την αξιοποίηση επιπλέον δεδομένων τα οποία περιγράφουν την συμπεριφορά των χρηστών κατά την πλοήγησή τους σε ένα ηλεκτρονικό κατάστημα. Τα δεδομένα αυτά είναι σχεδόν πάντα διαθέσιμα στις βάσεις δεδομένων των ηλεκτρονικών καταστημάτων. Σε αντίθετη περίπτωση, η καταγραφή τους δεν αποτελεί δύσκολο εγχείρημα και μπορεί εύκολα να ενσωματωθεί στη λειτουργία των διαθέσιμων πληροφοριακών συστημάτων.

Με βάση τα παραπάνω, διατυπώνεται η παρακάτω υπόθεση προς επιβεβαίωση:

*Η αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης βελτιώνει την ποιότητα των αποτελεσμάτων που παράγουν τα συστήματα συστάσεων σε σύγκριση με την αξιοποίηση δεδομένων που αφορούν μόνο τις αγορές των χρηστών.*

Η λύση που προτείνεται λειτουργεί ενισχυτικά στην απόδοση των συστημάτων συστάσεων, γεγονός που επιβεβαιώνεται μέσω των πειραμάτων που εκτελούνται στην εργασία.

### 1.3 Οργάνωση εργασίας

Στο πρώτο κεφάλαιο γίνεται μία εισαγωγή στο αντικείμενο της παρούσας εργασίας. Ο αναγνώστης εισάγεται στη φύση του προβλήματος που καλείται να επιλύσει η εργασία καθώς και στη λύση που προτείνεται. Επίσης παρουσιάζεται η οργάνωση και δομή της εργασίας.

Στο δεύτερο κεφάλαιο πραγματοποιείται μία βιβλιογραφική μελέτη επί των συστημάτων παραγωγής συστάσεων. Πιο συγκεκριμένα, περιγράφονται οι προσεγγίσεις που χρησιμοποιούνται για την παραγωγή συστάσεων, οι τύποι δεδομένων εισόδου, οι δυσκολίες που αντιμετωπίζουν τα συστήματα συστάσεων, οι τρόποι αξιολόγησης καθώς και σύγχρονες εφαρμογές των συστημάτων αυτών.

Στο τρίτο κεφάλαιο περιγράφεται η λύση που προτείνεται μέσω της συγκεκριμένης εργασίας. Αναλύονται τα στάδια της προεπεξεργασίας των δεδομένων εισόδου και των αλγορίθμων που χρησιμοποιήθηκαν. Επίσης αναφέρεται ο τρόπος με τον οποίο οργανώθηκαν τα πειράματα εκτέλεσης και παρουσιάζονται τα αποτελέσματα των πειραμάτων.

Στο τέταρτο κεφάλαιο παρουσιάζεται η υλοποίηση μίας δοκιμαστικής ιστοσελίδας ηλεκτρονικού καταστήματος που δημιουργήθηκε με σκοπό την μελλοντική συγκέντρωση δεδομένων.

Στο πέμπτο κεφάλαιο συγκεντρώνονται τα συμπεράσματα που προέκυψαν από την παρούσα εργασία καθώς και οι μελλοντικοί ερευνητικοί στόχοι.



## 2 Συστήματα συστάσεων

Το κεφάλαιο αυτό περιλαμβάνει μία σύνοψη της βιβλιογραφικής έρευνας που πραγματοποιήθηκε με αντικείμενο τα συστήματα συστάσεων. Αναλύονται οι πιο διαδεδομένες τεχνικές παραγωγής συστάσεων, οι τύποι των δεδομένων που αξιοποιούνται καθώς και οι δυσκολίες που αντιμετωπίζουν τα συγκεκριμένα συστήματα. Τέλος παρουσιάζονται μερικές σύγχρονες εφαρμογές συστημάτων συστάσεων.

### 2.1 Εισαγωγή

Η ταχεία ανάπτυξη και εξέλιξη του παγκόσμιου ιστού έδωσε τις βάσεις για την παραγωγή, συλλογή και αποθήκευση τεράστιου όγκου δεδομένων. Σύντομα παρουσιάστηκε η ανάγκη για εύκολη και άμεση αναζήτηση στον τεράστιο αυτό όγκο δεδομένων. Τη λύση στο συγκεκριμένο πρόβλημα έδωσαν οι μηχανές αναζήτησης. Η μηχανή της Google, ξεκινώντας τον Ιανουάριο του 1996 ως ερευνητικό πρόγραμμα από τους Λάρρυ Πέιτζ (Larry Page) και Σεργκέι Μπριν (Sergey Brin), αποτελεί πλέον την πιο διαδεδομένη μηχανή αναζήτησης.

Η ανάπτυξη του παγκόσμιου ιστού και του διαδικτύου όμως δημιούργησε ευκαιρίες και σε άλλους τομείς, με πρώτο και σημαντικότερο αυτόν της πώλησης αγαθών και υπηρεσιών. Το 1994 ιδρύεται το διαδικτυακό κατάστημα της Amazon από τον Τζεφ Μπέζος (Jeff Bezos). Η σταδιακή ενσωμάτωση και αποδοχή των ηλεκτρονικών καταστημάτων από τους χρήστες του διαδικτύου ανέδειξε την Amazon το 2015 ως την πιο επικερδή εταιρία λιανικής πώλησης στις Ηνωμένες Πολιτείες Αμερικής, ξεπερνώντας την παραδοσιακή αλυσίδα λιανικής πώλησης προϊόντων Walmart [1].

Το ηλεκτρονικό εμπόριο (e-commerce) και η διαδικτυακή παρουσία χαρακτηρίζονται πλέον ως βασικοί παράγοντες επιτυχημένης κερδοφορίας ενός καταστήματος που δραστηριοποιείται ή πρόκειται να δραστηριοποιηθεί στην πώληση αγαθών και υπηρεσιών. Οι βασικοί λόγοι που οδηγούν στο παραπάνω συμπέρασμα αφορούν την κατάργηση των φυσικών περιορισμών στους οποίους υπόκειντο μέχρι πριν λίγα χρόνια τα παραδοσιακά καταστήματα. Πέρα από τη δυνατότητα προσέλκυσης πελατών από ολόκληρη την υφήλιο, ένα ηλεκτρονικό κατάστημα μπορεί πλέον να παρουσιάσει αναλυτικά όλα τα προϊόντα του χωρίς να περιορίζεται από το μέγεθος και την έκταση ενός παραδοσιακού καταστήματος. Τα οφέλη είναι πολλά, όχι μόνο για τις επιχειρήσεις που μπορούν πλέον να εξασφαλίσουν μεγαλύτερη προβολή και πρόσβαση στα προϊόντα τους αλλά και για τους ίδιους τους καταναλωτές οι οποίοι καλούνται με τη σειρά τους να επιλέξουν μέσα σε έναν ωκεανό διαφορετικών προϊόντων.

Η νέα ανάγκη που προέκυψε, ανάλογη με αυτήν που ικανοποίησε η μηχανή αναζήτησης της Google μέσω της τεχνολογίας PageRank, είναι η μελέτη και δημιουργία τεχνολογιών που εξυπηρετούν τον σύγχρονο διαδικτυακό χρήστη στην εύκολη αναζήτηση, έρευνα και επιλογή των προϊόντων που θα του εξασφαλίσουν τη μεγαλύτερη χρησιμότητα και ικανοποίηση. Τα συστήματα που προσπαθούν να λύσουν το συγκεκριμένο πρόβλημα χαρακτηρίζονται στη βιβλιογραφία ως συστήματα συστάσεων (recommender systems).

Τα συστήματα συστάσεων αποτελούν υλοποιήσεις ειδικών αλγορίθμων που επιχειρούν να προβλέψουν ποια προϊόντα είναι πιθανό να ικανοποιήσουν περισσότερο τον διαδικτυακό χρήστη. Οι αλγόριθμοι που χρησιμοποιούνται λαμβάνουν ως είσοδο δεδομένα που αφορούν τους χρήστες (π.χ. δημογραφικά στοιχεία, ιστορικό αναζητήσεων ή αγορών, προηγούμενες αξιολογήσεις προϊόντων) ή και τα προϊόντα (π.χ. είδος, περιγραφή, χαρακτηριστικά) με σκοπό να υπολογίσουν το εκτιμώμενο ενδιαφέρον των χρηστών για κάθε προϊόν. Στη



συνέχεια, ο κάθε χρήστης έχει στη διάθεσή του μία λίστα με προϊόντα, ταξινομημένα με βάση το εκτιμώμενο ενδιαφέρον. Με αυτό τον τρόπο, ένα ηλεκτρονικό κατάστημα μπορεί να προσαρμόσει και να εξατομικεύσει το περιεχόμενό του για κάθε χρήστη συστήνοντας του συγκεκριμένα προϊόντα για αγορά.

Τα συστήματα συστάσεων ενισχύουν το ηλεκτρονικό εμπόριο με τρεις τρόπους [2]:

- Μετατρέπουν τους «περιπλανώμενους» χρήστες (αυτούς που συνήθως περιηγούνται στις σελίδες ενός ηλεκτρονικού καταστήματος αλλά δεν πραγματοποιούν αγορές) σε αγοραστές. Με τη βοήθεια των συστημάτων συστάσεων είναι πολύ πιο πιθανό ένας χρήστης να ανακαλύψει άμεσα τα προϊόντα που τον ενδιαφέρουν περισσότερο.
- Αυξάνουν τον αριθμό των αγορών ανά χρήστη. Με την αγορά ή και ακόμα την προβολή ενός συγκεκριμένου προϊόντος το σύστημα αναδιαμορφώνει τις συστάσεις του προς τον χρήστη, παρουσιάζοντάς του επιπλέον χρήσιμα προϊόντα που ίσως του τραβήξουν το ενδιαφέρον και τα αγοράσει. Η διαδικασία αυτή πραγματοποιείται συνήθως κατά το τελικό στάδιο μιας ηλεκτρονικής αγοράς (checkout), όπου το σύστημα γνωρίζει τα προϊόντα που έχει εισάγει ο χρήστης στο καλάθι του.
- Ενισχύουν την αφοσίωση των αγοραστών προς το κατάστημα (loyalty). Οι χρήστες που επιστρέφουν στο κατάστημα για να πραγματοποιήσουν περισσότερες αγορές ενισχύουν παράλληλα τη γνώση που διατηρεί το σύστημα συστάσεων για αυτούς. Με αυτό τον τρόπο, το κατάστημα εξατομικεύεται ακόμα περισσότερο στις ανάγκες του κάθε χρήστη και ελαχιστοποιεί τις πιθανότητες να επισκεφτεί ο χρήστης ανταγωνιστικά καταστήματα και να επενδύσει στην εκπαίδευση των δικών τους συστημάτων συστάσεων.

Ωστόσο, στην περίπτωση που τα συστήματα αυτά εξάγουν ανακριβείς συστάσεις, οι συνέπειες στην αγοραστική συμπεριφορά των χρηστών είναι άμεσες και συνήθως μη αναστρέψιμες. Υπάρχουν δύο είδη σφαλμάτων που μπορεί να προκύψουν από ένα ανακριβές σύστημα συστάσεων [3]:

- **Εσφαλμένα αρνητικά σφάλματα (false negatives):** Είναι προϊόντα που είναι χρήσιμα και ικανοποιούν τον χρήστη αλλά δεν εξάγονται ως προτεινόμενα από το σύστημα.
- **Εσφαλμένα θετικά σφάλματα (false positives):** προϊόντα που δεν ικανοποιούν τον χρήστη αλλά εξάγονται από το σύστημα ως προτεινόμενα προς τον χρήστη.

Σκοπός ενός συστήματος είναι η αποφυγή των εσφαλμένων θετικών σφαλμάτων. Η συγκεκριμένη κατηγορία λαθών έχει ως αποτέλεσμα την απογοήτευση των αγοραστών και την οριστική εγκατάλειψη του ηλεκτρονικού καταστήματος.

Το πρόβλημα που καλούνται να λύσουν τα συστήματα συστάσεων καταγράφεται στη βιβλιογραφία με δύο διαφορετικούς τρόπους [3]:

- **Πρόβλημα πρόβλεψης (prediction problem):** Στο συγκεκριμένο πρόβλημα τα συστήματα συστάσεων χρησιμοποιούνται για να προβλέψουν κατά πόσο ένας χρήστης θα ικανοποιηθεί με την αγορά ενός συγκεκριμένου (άγνωστου για τον χρήστη) προϊόντος.

- **Πρόβλημα n-καλύτερων προτάσεων (top-n recommendation problem):** Στο πρόβλημα αυτό τα συστήματα συστάσεων χρησιμοποιούνται για να εξάγουν ένα φθίνουσα ταξινομημένο σύνολο n προϊόντων (που δεν έχουν αγοραστεί στο παρελθόν από τον χρήστη) με κριτήριο την εκτιμώμενη ικανοποίηση που θα προσφέρουν στον χρήστη με την αγορά τους.

## 2.2 Δεδομένα εισόδου

Οι βασικές οντότητες που κυριαρχούν σε ένα σύστημα συστάσεων είναι ο χρήστης και (στην περίπτωση του ηλεκτρονικού εμπορίου) το προϊόν. Τα δεδομένα που απαιτούνται από τους αλγόριθμους και τις τεχνικές που ενσωματώνονται στα συστήματα αυτά αναφέρονται κυρίως στη σχέση μεταξύ ενός χρήστη και ενός προϊόντος. Ζωτικής σημασίας είναι η εξασφάλιση ενός βασικού, αρχικού όγκου πληροφοριών που περιγράφουν την άποψη ενός χρήστη για ένα σύνολο προϊόντων. Βάσει αυτών των πληροφοριών, ένα σύστημα συστάσεων μπορεί να εξάγει νέες συστάσεις για προϊόντα άγνωστα μέχρι στιγμής στον χρήστη.

Στις περισσότερες περιπτώσεις, τα δεδομένα εισάγονται στους αλγόριθμους οργανωμένα σε δισδιάστατο πίνακα (Εικόνα 1). Η μία διάσταση (συνήθως οι γραμμές του πίνακα) περιγράφει τους χρήστες του συστήματος, ενώ η δεύτερη (συνήθως οι στήλες του πίνακα) περιγράφει τα προϊόντα. Κάθε τιμή του πίνακα αναπαριστά την άποψη, σχέση, προτίμηση ή αξιολόγηση ενός συγκεκριμένου χρήστη για ένα συγκεκριμένο προϊόν.

### 2.2.1 Άμεση ανατροφοδότηση

Επικρατούσα μορφή δεδομένων στη σχετική βιβλιογραφία αλλά και σε υπάρχοντα συστήματα συστάσεων είναι οι αξιολογήσεις (ratings) χρηστών επί προϊόντων που ήδη βρίσκονται στην κατοχή τους.

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	0	3	0	3	0
User 2	4	0	0	2	0
User 3	0	0	3	0	0
User 4	3	0	4	0	3
User 5	4	3	0	4	0

Εικόνα 1: Πίνακας εισόδου άμεσης ανατροφοδότησης - πηγή: διαδίκτυο

Η αξιολόγηση (Εικόνα 1) πραγματοποιείται με την ανάθεση μιας αριθμητικής τιμής σε ένα προϊόν υπό κλίμακα Likert πέντε βαθμίδων, με την ελάχιστη τιμή «1» να χαρακτηρίζει τη δυσαρέσκεια και μέγιστη τιμή «5» την πλήρη ικανοποίηση του χρήστη (π.χ. Amazon, Netflix). Ένας εξίσου διαδεδομένος τρόπος αξιολόγησης είναι η ανάθεση μίας δυαδικής τιμής σε ένα προϊόν που περιγράφει την αρέσκεια ή δυσαρέσκεια του χρήστη (π.χ. Facebook, YouTube). Η παραπάνω μορφή δεδομένων αναφέρεται στη βιβλιογραφία ως άμεση (ή ρητή) ανατροφοδότηση (explicit feedback).

Η χρήση της συγκεκριμένης μορφής δεδομένων ωστόσο παρουσιάζει κάποια μειονεκτήματα. Βασικό πρόβλημα αποτελεί η έλλειψη ικανοποιητικού αριθμού αξιολογήσεων για την ορθή λειτουργία ενός συστήματος συστάσεων. Ο χαμηλός αριθμός δεδομένων οδηγεί σε

δημιουργία σφαλμάτων και συνεπώς δυσaréσκεια των χρηστών. Πολλά ηλεκτρονικά καταστήματα προσπαθούν να εξασφαλίσουν αξιολογήσεις από τους χρήστες ανταμείβοντάς τους ή ασκώντας τεχνικές μάρκετινγκ και προωθητικές ενέργειες. (π.χ. έκπτωση στην επόμενη αγορά). Στην περίπτωση αυτή όμως ελλοχεύει ο κίνδυνος οι αξιολογήσεις να μην είναι αντιπροσωπευτικές και να εισαχθεί θόρυβος στο σύνολο των αρχικών δεδομένων.

### 2.2.2 Έμμεση ανατροφοδότηση

Εναλλακτική μορφή δεδομένων αποτελεί η έμμεση (μη ρητή) ανατροφοδότηση (implicit feedback). Τα δεδομένα που ανήκουν στη συγκεκριμένη κατηγορία δεν απαιτούν τη ρητή αξιολόγηση του χρήστη και συνεπώς δεν τον επιβαρύνουν μετά την αγορά και τη λήψη των προϊόντων.

Ως έμμεση ανατροφοδότηση μπορούν να χαρακτηριστούν δεδομένα που περιγράφουν τη συμπεριφορά και την πλοήγηση ενός χρήστη σε ένα ηλεκτρονικό κατάστημα. Ο βασικός όγκος δεδομένων που αξιοποιείται από τα συστήματα σύστασης, σύμφωνα με την παρούσα βιβλιογραφία, αφορά κυρίως δεδομένα αγορών. Ωστόσο ελάχιστες είναι οι καταγεγραμμένες περιπτώσεις όπου αξιοποιούνται δεδομένα που αφορούν 1) ενέργειες πλοήγησης των χρηστών όπως αναζήτηση, περιήγηση, προβολή προϊόντος, προσθήκη στα αγαπημένα, προσθήκη στο καλάθι ή 2) ενέργειες συμπεριφοράς όπως αναλογία κλικ για ένα συγκεκριμένο τύπο προϊόντων, διάρκεια ανάγνωσης της περιγραφής ενός προϊόντος, αριθμός επισκέψεων σε συγκεκριμένο προϊόν [4].

Ο τεράστιος όγκος των διαθέσιμων δεδομένων αποτελεί το βασικό πλεονέκτημα της έμμεσης ανατροφοδότησης. Ωστόσο είναι αδύνατο τα δεδομένα αυτά να αξιοποιηθούν υπό την αρχική τους μορφή και συνεπώς απαιτείται προεπεξεργασία των δεδομένων και μετατροπή σε μορφή αποδεκτή από ένα σύστημα συστάσεων.

Η έμμεση ανατροφοδότηση παρουσιάζει κάποια ιδιαίτερα χαρακτηριστικά τα οποία δυσχεραίνουν την αξιοποίησή τους [5]:

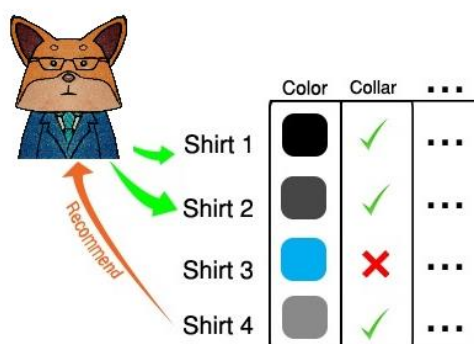
- **Απουσία αρνητικής ανατροφοδότησης:** είναι αδύνατο να προσδιοριστεί ποια προϊόντα δεν αρέσουν σε έναν χρήστη. Η απουσία ανατροφοδότησης μπορεί να χαρακτηριστεί είτε ως δυσaréσκεια είτε ως άγνοια ύπαρξης του συγκεκριμένου προϊόντος. Η συγκεκριμένη ασυμμετρία δεν παρουσιάζεται στην άμεση ανατροφοδότηση.
- **Θόρυβος:** τα δεδομένα συμπεριφοράς και πλοήγησης που καταγράφονται από ένα ηλεκτρονικό κατάστημα εισάγουν θόρυβο στα δεδομένα που θα αξιοποιηθούν από το σύστημα συστάσεων. Περιπτώσεις όπως η αγορά δώρων ή λάθη κατά τη διαδικτυακή πλοήγηση, δε συμβάλλουν στη δημιουργία ενός αντιπροσωπευτικού προφίλ για έναν χρήστη.
- **Εμπιστοσύνη:** σε αντίθεση με την άμεση ανατροφοδότηση μέσω της οποίας οι χρήστες εκφράζουν την προτίμησή τους για τα προϊόντα, η έμμεση ανατροφοδότηση περιγράφει την εμπιστοσύνη των χρηστών προς τα προϊόντα. Για παράδειγμα, η μεγάλη συχνότητα προβολής ενός συγκεκριμένου προϊόντος μπορεί να ερμηνευτεί ως αυξημένο ενδιαφέρον από τον χρήστη. Όμως χαμηλή συχνότητα προβολής δε συνεπάγεται απαραίτητα με χαμηλό ενδιαφέρον ή δυσaréσκεια.

### 2.2.3 Συσχέτιση άμεσης και έμμεσης ανατροφοδότησης

Η αξιοποίηση της έμμεσης ανατροφοδότησης ως εναλλακτική λύση έναντι της άμεσης ανατροφοδότησης προϋποθέτει την ύπαρξη ισχυρής συσχέτισης ανάμεσα στις δύο μορφές δεδομένων. Το συγκεκριμένο ζήτημα έχει απασχολήσει την ακαδημαϊκή κοινότητα [6]. Συγκεκριμένα διαπιστώθηκε πως 1) ο χρόνος προβολής ενός προϊόντος από έναν χρήστη συσχετίζεται σημαντικά με τη θετική αξιολόγηση του και 2) ο αριθμός των επισκέψεων σε ένα προϊόν ή κατηγορία προϊόντος συσχετίζεται επίσης σημαντικά με το ύψος των θετικών αξιολογήσεων ενός χρήστη.

### 2.3 Αλγόριθμοι & τεχνικές

Στην παρούσα εργασία αναλύονται διεξοδικά οι τεχνικές που αναφέρονται στη βιβλιογραφία ως αλγόριθμοι βασισμένοι στη μνήμη (memory based algorithms). Οι προσεγγίσεις αυτές παρουσιάζουν καλύτερα αποτελέσματα κατά την παραγωγή συστάσεων αλλά χαρακτηρίζονται από μεγαλύτερη πολυπλοκότητα έναντι των αλγορίθμων που κατασκευάζουν μοντέλα, τους οποίους αναφέρουμε συνοπτικά στην ενότητα 2.3.5.



Εικόνα 2: Παραγωγή συστάσεων με βάση το περιεχόμενο - πηγή: διαδίκτυο

#### 2.3.1 Προσέγγιση βασισμένη στο περιεχόμενο

Η συγκεκριμένη τεχνική εστιάζει στα βασικά χαρακτηριστικά και τις ιδιότητες των προϊόντων (content based approach). Απαραίτητη προϋπόθεση είναι η δημιουργία ενός προφίλ για κάθε προϊόν ξεχωριστά, το οποίο περιγράφεται από ένα κοινά ορισμένο σύνολο μεταβλητών. Οι μεταβλητές αυτές ορίζονται από τον δημιουργό του συστήματος, και μπορούν να αφορούν γνωρίσματα όπως την κατηγορία, την περιγραφή ή τη μάρκα του προϊόντος κ.α. Με αυτό τον τρόπο, καθίσταται δυνατή η αναπαράσταση ενός προϊόντος ως ένα διάνυσμα σε έναν περατό χώρο διαστάσεων.

Δεύτερο βήμα αποτελεί η δημιουργία προφίλ για κάθε χρήστη λαμβάνοντας υπόψιν τα δεδομένα εισόδου. Κάθε χρήστης περιγράφεται επίσης από ένα διάνυσμα (αντίστοιχου αριθμού διαστάσεων με αυτό των προϊόντων) το οποίο προκύπτει συνδυάζοντας τα διανύσματα αντικειμένων για τα οποία έχει εκφράσει προτίμηση (π.χ. συναθροίζοντας τα διανύσματα ή εκτελώντας κάποια άλλη διαδικασία της επιλογής μας).

Στη συνέχεια με τη βοήθεια μετρικών αποστάσεων (π.χ. ομοιότητα συνημίτονου), μπορούμε να εκτιμήσουμε ποια άγνωστα προς τον χρήστη προϊόντα είναι περισσότερο όμοια με το προφίλ του σε σύγκριση με τα υπόλοιπα (Εικόνα 2).

Τα συστήματα βασισμένα στο περιεχόμενο υποφέρουν από σημαντικούς περιορισμούς που τα καθιστούν δύσχρηστα, συνεπώς τις περισσότερες φορές συνδυάζονται με άλλες τεχνικές και μεθόδους. Βασικό πρόβλημα αποτελεί η δυσκολία ορισμού των μεταβλητών που

περιγράφουν τα προφίλ χρηστών και προϊόντων. Η συλλογή δεδομένων για κάθε προϊόν ξεχωριστά καθώς και η προεπεξεργασία των δεδομένων αυτών είναι χρονοβόρα διαδικασία. Επιπλέον, δύσκολο εγχείρημα αποτελεί και η επιλογή των κατάλληλων μεταβλητών (feature selection), η οποία συνήθως προϋποθέτει ιδιαίτερη γνώση και εμπειρία για τη φύση του προβλήματος (domain knowledge & expertise). Εξίσου σημαντικό πρόβλημα είναι το πρόβλημα της εξειδίκευσης. Τα προϊόντα που προτείνονται από το σύστημα περιορίζονται κυρίως στα χαρακτηριστικά των προϊόντων που παρουσίασε αρχικά ο χρήστης. Με αυτό τον τρόπο, είναι πρακτικά αδύνατο να παρουσιαστούν νέα προϊόντα στον χρήστη που να καλύπτουν διαφορετικές ανάγκες του. Τέλος ιδιαίτερο πρόβλημα αποτελεί η αδυναμία δημιουργίας προφίλ και παροχής συστάσεων στους νέους χρήστες του συστήματος εξαιτίας της έλλειψης σε δεδομένα που περιγράφουν προηγούμενες αγορές ή προτιμήσεις τους.

### 2.3.2 Προσέγγιση συνεργατικού φιλτραρίσματος

Η προσέγγιση αυτή μπορεί να αναλυθεί από τη σκοπιά των χρηστών ή των προϊόντων.

#### 2.3.2.1 Συνεργατικό φιλτράρισμα βασισμένο στους χρήστες

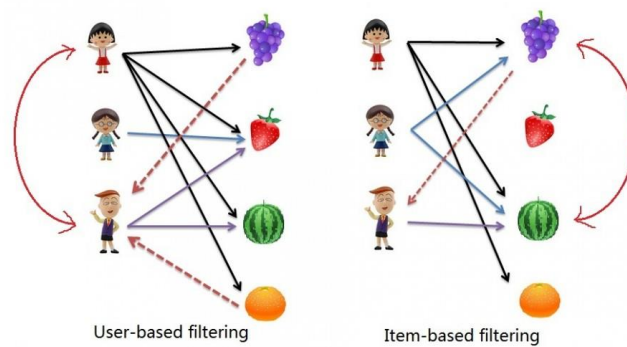
Η προσέγγιση του συνεργατικού φιλτραρίσματος (collaborative filtering) εστιάζει στην εύρεση παρόμοιων χρηστών και τη σύσταση των προϊόντων που έχουν προμηθευτεί. Αποτελεί μία από τις πιο διαδεδομένες τεχνικές συστάσεων και εφαρμόστηκε αρχικά στο πρώτο σύστημα συστάσεων Tapestry [7]. Η τεχνική αυτή δεν απαιτεί τη δημιουργία ξεχωριστού προφίλ για κάθε προϊόν ή χρήστη αλλά προσπαθεί να εκτιμήσει άμεσα τον βαθμό ομοιότητας ανάμεσα στους χρήστες αξιοποιώντας τα δεδομένα εισόδου.

Κάθε χρήστης αναπαρίσταται από ένα διάνυσμα μεγέθους ίσο με τον αριθμό των προϊόντων, και περιλαμβάνει τις τιμές από τον διδιάστατο πίνακα εισόδου. Χρησιμοποιώντας μία μετρική απόστασης, υπολογίζονται αρχικά όλες οι αποστάσεις μεταξύ των χρηστών του συστήματος. Έπειτα προσδιορίζεται ένα σύνολο όμοιων χρηστών (γειτονιά του χρήστη) προς κάθε χρήστη είτε επιλέγοντας έναν αριθμό των πιο όμοιων χρηστών είτε χρήστες που παρουσιάζουν ομοιότητα άνω ενός ορισμένου ποσοστού. Τα βήματα αυτά αποτελούν το πρώτο στάδιο της προσέγγισης (στάδιο προσδιορισμού γειτονιάς) το οποίο εκτελείται περιοδικά και τα αποτελέσματά αφού αποθηκευτούν, επαναχρησιμοποιούνται στο επόμενο στάδιο.

Έχοντας προσδιορίσει τη γειτονιά ενός χρήστη, μπορούμε να υπολογίσουμε μία τιμή σύστασης για κάθε προϊόν που εξετάστηκε από τη γειτονιά του, συναθροίζοντας τις τιμές των όμοιων χρηστών ανά προϊόν. Τα προϊόντα ταξινομούνται σε φθίνουσα σειρά βάσει της τιμής σύστασης και απαλείφονται τα προϊόντα που έχουν ήδη εξεταστεί από τον χρήστη. Έτσι ολοκληρώνεται και το δεύτερο στάδιο (στάδιο παραγωγής συστάσεων) της διαδικασίας, το οποίο σε αντίθεση με το πρώτο εκτελείται όποτε υπάρχει αίτημα από το σύστημα για συστάσεις.

Τα βασικά πλεονεκτήματα της συγκεκριμένης μεθόδου είναι η δυνατότητα υλοποίησής της με οποιοδήποτε τύπο περιεχομένου (ηλεκτρονικό εμπόριο, ταινίες, μουσική, βίντεο) καθώς και ότι παρουσιάζει τα καλύτερα αποτελέσματα σε σύγκριση με άλλες προσεγγίσεις.

Ωστόσο η τεχνική του συνεργατικού φιλτραρίσματος παρουσιάζει αξιοσημείωτα μειονεκτήματα, με πιο σημαντικό αυτό της περιορισμένης κλιμάκωσης (scalability). Η αύξηση του αριθμού των χρηστών αυξάνει δραματικά την πολυπλοκότητα εκτέλεσης κατά τον υπολογισμό των αποστάσεων μεταξύ των χρηστών (στάδιο προσδιορισμού γειτονιάς).



Εικόνα 3: Συνεργατικό φιλτράρισμα - πηγή: διαδίκτυο

### 2.3.2.2 Συνεργατικό φιλτράρισμα βασισμένο στα προϊόντα

Μια εναλλακτική προσέγγιση συνεργατικού φιλτραρίσματος βασισμένη στα προϊόντα αποτελεί τη λύση στα προβλήματα που αναφέρθηκαν στην προηγούμενη προσέγγιση. Η βασική διαφορά έγκειται στον υπολογισμό των αποστάσεων ανάμεσα στα προϊόντα, τα οποία αναπαρίστανται ως διανύσματα που περιλαμβάνουν τιμές αξιολόγησης από το σύνολο των χρηστών. Στη συνέχεια, υπολογίζονται τιμές σύστασης για προϊόντα τα οποία παρουσιάζουν μεγάλη ομοιότητα με αυτά που έχει ήδη εξετάσει ο χρήστης.

Το βασικό πλεονέκτημα της παραπάνω προσέγγισης είναι η μείωση της πολυπλοκότητας των υπολογισμών. Ο αριθμός των προϊόντων αυξάνεται συνήθως πιο αργά από τον αριθμό των χρηστών (περισσότεροι νέοι χρήστες, λιγότερα νέα προϊόντα). Επίσης, τα διανύσματα των χρηστών αλλάζουν πιο συχνά από τα διανύσματα των αντικειμένων. Συνεπώς, στην περίπτωση του συνεργατικού φιλτραρίσματος βασισμένο στους χρήστες, απαιτείται πιο συχνά ο εκ νέου υπολογισμός των αποστάσεων ανάμεσα στους χρήστες (στάδιο προσδιορισμού γειτονιάς). Τέλος όταν η προσέγγιση βασίζεται στα προϊόντα, είναι εφικτό να εξηγήσουμε σε ποια προϊόντα οφείλεται η πρόταση ενός άγνωστου προϊόντος. Η δυνατότητα αυτή συμβάλλει στη δημιουργία εμπιστοσύνης με τον χρήστη αλλά και στον έλεγχο της ορθής λειτουργίας του συστήματος συστάσεων.

Και οι δύο μεθοδολογίες του συνεργατικού φιλτραρίσματος υποφέρουν από τα παρακάτω δύο προβλήματα. Πρώτον, οι μέθοδοι υστερούν σε απόδοση όταν ζητούνται προβλέψεις για νέους χρήστες καθώς και σε περιπτώσεις όπου εισάγονται νέα αντικείμενα στο σύστημα. Δεύτερον, εάν τα δεδομένα εισόδου δεν είναι αρκετά και συνεπώς δεν παρουσιάζονται επικαλύψεις στις τιμές των προϊόντων από διάφορους χρήστες, προκύπτει η αδυναμία εύρεσης παρόμοιων προϊόντων ή χρηστών, η οποία οδηγεί δυσκολίες στην παραγωγή συστάσεων.

### 2.3.3 Προσέγγιση ανάλυσης πίνακα σε ιδιάζουσες τιμές

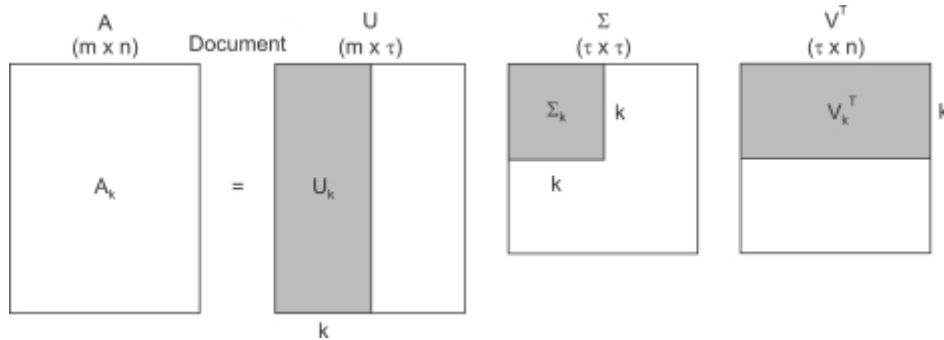
Η ανάλυση πίνακα σε ιδιάζουσες τιμές (singular value decomposition) αποτελεί μία τεχνική παραγοντοποίησης πινάκων σύμφωνα με την οποία ένας πίνακας  $M$  με πραγματικά στοιχεία, διαστάσεις  $m \times n$  και βαθμό  $r$  παραγοντοποιείται στο παρακάτω γινόμενο τριών πινάκων (Εικόνα 4):

$$SVD(M) = U \times \Sigma \times V^T, \text{ όπου}$$

- $U$  είναι ένας ορθογώνιος πίνακας διαστάσεων  $m \times r$
- $V^T$  ο ανάστροφος  $V$  ορθογώνιος πίνακας διαστάσεων  $n \times r$

- $\Sigma$  ένας διαγώνιος πίνακας διαστάσεων  $r \times r$ , με θετικές τιμές στη διαγώνιο
- $r \leq \min(m, n)$

Τα στοιχεία του  $\Sigma$  είναι γνωστά ως ιδιάζουσες τιμές του  $M$ , και είναι ταξινομημένα σε φθίνουσα σειρά. Ο πίνακας  $\Sigma$  καθορίζεται με μοναδικό τρόπο από τον  $M$ , αλλά το ίδιο δε συμβαίνει με τους πίνακες  $U$  και  $V$ . Με τη βοήθεια των πινάκων  $U$  και  $V$ , οι χρήστες και τα προϊόντα μπορούν να αναπαρασταθούν ως διανύσματα σε έναν χώρο  $r$  διαστάσεων.



Εικόνα 4: Ανάλυση πίνακα σε ιδιάζουσες τιμές - πηγή: διαδίκτυο

Η χρησιμότητα της συγκεκριμένης τεχνικής έγκειται στη δυνατότητα διατήρησης των  $k$  μεγαλύτερων ιδιαιζουσών τιμών του διαγώνιου πίνακα  $\Sigma$  (Εικόνα 4) και η αποβολή των υπόλοιπων  $r - k$  τιμών (και αντίστοιχα η διατήρηση των  $k$  πρώτων στηλών των πινάκων  $U$  και  $V$  και η αποβολή των υπόλοιπων  $r - k$  στηλών). Το γινόμενο των νέων αυτών πινάκων αποτελεί την καλύτερη δυνατή προσέγγιση του αρχικού πίνακα  $M$ , με τη μικρότερη δυνατή απόσταση από τον  $M$  με βάση τη νόρμα Frobenius που ορίζεται ως εξής:

$$\|M - \tilde{M}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m |M[i, j] - \tilde{M}[i, j]|^2}$$

Τα πλεονεκτήματα που προκύπτουν από τη χρήση της ανάλυσης πίνακα σε ιδιάζουσες τιμές συμβάλλουν σημαντικά στην απόδοση και ακρίβεια του συστήματος συστάσεων. Πρώτον, σημαντική είναι η ανάλυση των λανθανουσών σχέσεων που υπάρχουν ανάμεσα στους χρήστες και τα προϊόντα. Κάθε ιδιάζουσα τιμή αναφέρεται σε έναν άγνωστο παράγοντα, που χαρακτηρίζει σημαντικά τη σχέση ανάμεσα στους χρήστες και τα προϊόντα. Διατηρώντας τις  $k$  μεγαλύτερες ιδιάζουσες τιμές, εξασφαλίζεται η διατήρηση των πιο σημαντικών παραγόντων, άσχετα εάν η ερμηνεία τους είναι κατά πάσα πιθανότητα αδύνατη. Επιπλέον, η τεχνική αυτή εξασφαλίζει μείωση διαστάσεων (dimensionality reduction) του αρχικού προβλήματος. Εξαλείφοντας τις  $r - k$  μικρότερες ιδιάζουσες τιμές αρκεί να αποθηκεύσουμε μόνο τους τρεις παραγόμενους πίνακες μειώνοντας το σύνολο των αποθηκευμένων στοιχείων από  $(m \times n)$  σε  $(m \times k + n \times k + k)$ . Συνεπώς η επιλογή της τιμής  $k$  έχει άμεσες συνέπειες στην ακρίβεια συστάσεων αλλά και στη απόδοσή του συστήματος ως προς τις απαιτήσεις σε αποθηκευτικό χώρο.

Τα μειονεκτήματα της προσέγγισης αφορούν τον χρόνο εκτέλεσής της, ο οποίος είναι συνήθως μεγαλύτερος από τον υπολογισμό των αποστάσεων ανάμεσα στους χρήστες ή στα προϊόντα (προσέγγιση συνεργατικού φιλτραρίσματος). Ωστόσο τα αποτελέσματα της ανάλυσης του αρχικού πίνακα μπορούν να αποθηκευτούν και να επαναχρησιμοποιηθούν για την παραγωγή προβλέψεων. Μάλιστα, εξαιτίας της μείωσης των διαστάσεων,



παρουσιάζονται πιο σύντομοι χρόνοι εκτέλεσης κατά το στάδιο δημιουργίας συστάσεων σε σύγκριση τα αντίστοιχα στάδια των υπόλοιπων προσεγγίσεων. Δεύτερο μειονέκτημα αποτελεί η ανάγκη για εκ νέου εκτέλεση της διαδικασίας ανάλυσης εξαιτίας αλλαγών στον αρχικό πίνακα δεδομένων (προσθήκη νέων χρηστών ή αντικειμένων). Στη βιβλιογραφία όμως παρατηρούνται προσπάθειες για τη δημιουργία τεχνικών ανανέωσης και επέκτασης των τριών παραγόμενων πινάκων (SVD updating & folding in). Τέλος η προσέγγιση απαιτεί συνήθως απαλοιφή τον μηδενικών τιμών του αρχικού πίνακα μέσω εξειδικευμένης προεπεξεργασίας.

Η συγκεκριμένη τεχνική μπορεί να αξιοποιηθεί στα συστήματα συστάσεων με δύο διαφορετικούς τρόπους:

- Ο πρώτος τρόπος αφορά την αξιοποίηση του νέου πίνακα  $U$  με διαστάσεις  $m \times k$  για τον υπολογισμό των αποστάσεων μεταξύ των χρηστών (ή αντίστοιχα του νέου πίνακα  $V$  διαστάσεων  $n \times k$  για τον υπολογισμό των αποστάσεων μεταξύ των προϊόντων) στην προσέγγιση συνεργατικού φιλτραρίσματος. Οι νέοι πίνακες περιγράφουν τη σχέση του κάθε χρήστη ή προϊόντος με τις λανθάνουσες μεταβλητές που προκύπτουν από την ανάλυση του αρχικού πίνακα.
- Ο δεύτερος τρόπος περιλαμβάνει τη δημιουργία προβλέψεων εκτελώντας το γινόμενο των τριών νέων μειωμένων πινάκων. Οι γραμμές (ή στήλες, ανάλογα με τη μορφή του αρχικού πίνακα  $M$ ) του νέου πίνακα  $M$  αποτελούν τα νέα διανύσματα των χρηστών του συστήματος. Η παραγωγή συστάσεων για έναν χρήστη πραγματοποιείται αφαιρώντας από το διάνυσμά του τα προϊόντα που έχει ήδη εξετάσει και ταξινομώντας τις τιμές κατά φθίνουσα σειρά.

#### 2.3.4 Προσέγγιση παραγοντοποίησης πινάκων

Τα τελευταία χρόνια, υπό την ενθάρρυνση διαφόρων διαγωνισμών όπως αυτός της διαδικτυακής πλατφόρμας ταινιών Netflix [8], αναπτύχθηκαν νέες τεχνικές παραγωγής συστάσεων βασισμένες στην παραγοντοποίηση πινάκων (matrix factorization). Οι προσεγγίσεις αυτές απορρέουν από τη μέθοδο ανάλυσης πίνακα σε ιδιάζουσες τιμές αλλά διαφέρουν ως προς το γινόμενο πινάκων που παράγουν. Πιο συγκεκριμένα, οι τεχνικές αυτές προσπαθούν να παραγοντοποιήσουν τον αρχικό πίνακα  $M$ , διαστάσεων  $m \times n$ , στο παρακάτω γινόμενο δύο πινάκων αξιοποιώντας μόνο τις γνωστές τιμές του πίνακα εισόδου  $M$ :

$$UV - decomposition(M) = U \times V, \text{ όπου}$$

- $U$  είναι ένας ορθογώνιος πίνακας διαστάσεων  $m \times d$
- $V$  είναι ένας ορθογώνιος πίνακας διαστάσεων  $d \times n$

Στόχος των τεχνικών είναι η ελαχιστοποίηση των διαφορών μεταξύ των γνωστών τιμών του πίνακα εισόδου  $M$  και των τιμών που προκύπτουν από το γινόμενο των πινάκων  $U$  και  $V$ . Συνεπώς οι τεχνικές αυτές στοχεύουν στην παραγωγή των πινάκων  $U$  και  $V$  υπό τον παρακάτω περιορισμό:

$$\min \sum_{(u,i) \in \kappa} (r_{ui} - q_i^T p_u)^2 + \lambda (\|q_i\|^2 + \|p_u\|^2)$$

Όπου  $\lambda$  μία παράμετρος γενίκευσης έτσι ώστε να αποφεύγεται η εξειδίκευση (overfitting) της μεθόδου στις γνωστές τιμές του αρχικού πίνακα και  $\kappa$  το σύνολο εκπαίδευσης. Με τον τρόπο αυτό εξασφαλίζονται νέες προβλέψεις για τις άγνωστες τιμές του αρχικού πίνακα.



Μέσω των πινάκων  $U$  και  $V$ , οι χρήστες και τα προϊόντα μπορούν να αναπαρασταθούν ως διανύσματα σε έναν χώρο  $d$  διαστάσεων. Η παραγωγή τιμής σύστασης πραγματοποιείται εκτελώντας το εσωτερικό γινόμενο των διανυσμάτων του χρήστη και του προϊόντος.

Τα πλεονεκτήματα της παραγοντοποίησης πινάκων αποτελούν λύση για τα προβλήματα της ανάλυσης πίνακα σε ιδιάζουσες τιμές. Η μέθοδος εστιάζει στις γνωστές τιμές του πίνακα εισόδου και δεν απαιτεί την προεπεξεργασία των δεδομένων εισόδου για τη συμπλήρωση των τιμών που λείπουν.

Υπάρχουν δύο στρατηγικές ελαχιστοποίησης του περιορισμού που παρουσιάστηκε παραπάνω:

#### 2.3.4.1 Στοχαστική ελάττωση παραγώγου

Σύμφωνα με τη στρατηγική (stochastic gradient descent) αυτή ο αλγόριθμος εκτελεί μία επανάληψη επί όλων των γνωστών τιμών του αρχικού πίνακα και υπολογίζει το σφάλμα πρόβλεψης επί των τιμών αυτών:

$$e_{ui} = r_{ui} - q_i^T p_u$$

Στη συνέχεια τροποποιεί τις τιμές των πινάκων  $U$  και  $V$  σύμφωνα με τις παρακάτω εξισώσεις (η τιμή  $\gamma$  αποτελεί παράμετρο βελτιστοποίησης):

$$q_i \leftarrow q_i + \gamma \cdot (e_{ui} \cdot p_u - \lambda \cdot q_i)$$

$$p_u \leftarrow p_u + \gamma \cdot (e_{ui} \cdot q_i - \lambda \cdot p_u)$$

#### 2.3.4.2 Εναλλασσόμενα ελάχιστα τετράγωνα

Η συγκεκριμένη στρατηγική (alternating least squares) αποτελεί επίσης επαναληπτική διαδικασία. Σε κάθε επανάληψη, ο αλγόριθμος διατηρεί σταθερές τις τιμές ενός εκ των δύο πινάκων  $U$  και  $V$ , και υπολογίζει εκ νέου τις τιμές του δευτέρου λύνοντας ουσιαστικά ένα πρόβλημα ελαχίστων τετραγώνων. Κάθε επανάληψη μειώνει την τιμή του περιορισμού που παρουσιάστηκε παραπάνω, εξασφαλίζοντας σύγκλιση σε μία ελάχιστη τιμή.

Το βασικό πλεονέκτημα της στρατηγικής αυτής είναι η εύκολη παραλληλοποίησή της, συνεπώς μπορεί να χρησιμοποιηθεί σε μεγάλα σύνολα δεδομένων τα οποία προκύπτουν συνήθως με έμμεση ανατροφοδότηση [5], [9].

#### 2.3.5 Προσεγγίσεις κατασκευής μοντέλου

Οι προσεγγίσεις που αναφέρθηκαν στις προηγούμενες ενότητες χαρακτηρίζονται συνήθως ως αλγόριθμοι βασισμένοι στη μνήμη (memory based algorithms). Οι προσεγγίσεις που θα αναφερθούν στην παρακάτω ενότητα χαρακτηρίζονται στη βιβλιογραφία ως προσεγγίσεις βασισμένες στην κατασκευή ενός μοντέλου (model based algorithms).

##### 2.3.5.1 Κανόνες συσχέτισης & σειριακοί κανόνες

Η εξόρυξη κανόνων συσχέτισης εξελίχθηκε ως μία από τις σημαντικότερες τεχνικές εξόρυξης γνώσης εξαιτίας της συνεχούς προσπάθειας για ανάλυση και κατανόηση των καλαθιών αγοράς. Στόχος της τεχνικής είναι η ανακάλυψη και η εξήγηση συνδυασμών προϊόντων που αγοράζονται από τους καταναλωτές. Τα δεδομένα εισάγονται συνήθως σε μορφή συναλλαγών (μία εγγραφή περιλαμβάνει προϊόντα που αγοράστηκαν ταυτόχρονα κατά την επίσκεψη σε ένα κατάστημα) ενώ η γνώση που εξάγεται, έχει τη μορφή κανόνων και συνεπώς μπορεί να αξιοποιηθεί για την παραγωγή συστάσεων. Εναλλακτική τεχνική αποτελεί η εξαγωγή σειριακών κανόνων η οποία βασίζεται σε αγορές που πραγματοποιούνται κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου. Σε αντίθεση με τους κανόνες συσχέτισης, η

προσέγγιση αυτή δεν εστιάζει στα προϊόντα που αποκτήθηκαν μαζί αλλά στη σειρά με την οποία αγοράστηκαν.

Το βασικό μειονέκτημα της συγκεκριμένης μεθόδου είναι η αδυναμία παραγωγής συστάσεων για προϊόντα που δεν εμφανίζονται στους κανόνες που εξάχθηκαν από το σύστημα.

#### 2.3.5.2 Συσταδοποίηση

Μέσω της συσταδοποίησης (clustering) οι χρήστες οργανώνονται σε έναν ορισμένο αριθμό συστάδων. Η διαδικασία είναι επαναληπτική και κατά την εκτέλεσή της κάθε χρήστης ταξινομείται σε μία συγκεκριμένη συστάδα βάσει της απόστασής του από τους χρήστες που την απαρτίζουν (η σύγκριση γίνεται με τη βοήθεια μετρικών αποστάσεων). Η συσταδοποίηση μπορεί να χρησιμοποιηθεί για τον υπολογισμό της γειτονιάς ενός χρήστη και τη σύσταση προϊόντων που εξετάστηκαν από τη γειτονιά αυτή.

Βασικό πλεονέκτημα της προσέγγισης είναι η γρήγορη εκτέλεση και ταξινόμηση ενός νέου χρήστη σε μία υπάρχουσα συστάδα, ωστόσο οι περισσότεροι αλγόριθμοι συσταδοποίησης δεν παρουσιάζουν καλά αποτελέσματα ως προς την παραγωγή συστάσεων.

#### 2.3.5.3 Δίκτυα Bayes

Η συγκεκριμένη προσέγγιση περιλαμβάνει τη δημιουργία ενός πιθανοτικού μοντέλου υπό τη μορφή δέντρου απόφασης (Bayes network). Κάθε κόμβος και ακμή του δέντρου αναπαριστά κάποιο χαρακτηριστικό του χρήστη. Τα μοντέλα που παράγονται είναι μικρά σε μέγεθος, εκτελούνται γρήγορα και παρουσιάζουν ακρίβεια παρόμοια με τις προσεγγίσεις συνεργατικού φιλτραρίσματος.

#### 2.3.5.4 Κατασκευή γράφων

Το πρόβλημα παραγωγής συστάσεων προσεγγίζεται επίσης από σύνολο τεχνικών που βασίζεται στην κατασκευή γράφων. Αντιπροσωπευτική τεχνική είναι η μέθοδος Horting σύμφωνα με την οποία οι κόμβοι ενός γράφου αναπαριστούν τους χρήστες ενώ οι ακμές περιγράφουν την ομοιότητα μεταξύ των χρηστών. Η παραγωγή συστάσεων γίνεται με τη διάσχιση του γράφου και τον συνδυασμό των απόψεων των γειτονικών χρηστών. Η μέθοδος Horting διαφέρει από τις τεχνικές συνεργατικού φιλτραρίσματος ως προς τον συνυπολογισμό γειτόνων που δεν έχουν αξιολογήσει κοινά προϊόντα με αυτά του εξεταζόμενου χρήστη.

### 2.4 Προβλήματα & προκλήσεις

Τα παρακάτω φαινόμενα αποτελούν σημαντικά προβλήματα στην κατασκευή των συστημάτων συστάσεων και έχουν μελετηθεί εκτενώς στη σχετική βιβλιογραφία [3].

#### 2.4.1 Έλλειψη πυκνότητας

Τα περισσότερα ηλεκτρονικά καταστήματα παρέχουν χιλιάδες προϊόντα σε χιλιάδες χρήστες. Ακόμα και οι πιο ενεργοί χρήστες, μπορεί να αξιολογήσουν ή να αγοράσουν ένα πολύ μικρό μέρος των προϊόντων (συνήθως αρκετά κάτω του 1% - στα 100.000 προϊόντα, το 1% ισούται με 1.000 προϊόντα). Συνεπώς πάγιο πρόβλημα των συστημάτων συστάσεων είναι η έλλειψη πυκνότητας (sparsity) στα δεδομένα εισόδου. Το πρόβλημα αυτό επηρεάζει ιδιαίτερα τις τεχνικές συνεργατικού φιλτραρίσματος (αδυναμία προσδιορισμού γειτονιάς) και συνήθως λύνεται μέσω της προεπεξεργασίας των δεδομένων, της αξιοποίησης της έμμεσης ανατροφοδότησης ή της σύνθεσης πολλών διαφορετικών μεθοδολογιών για την υλοποίηση ενός υβριδικού συστήματος συστάσεων.

#### 2.4.2 Αδυναμία κλιμάκωσης

Τα συστήματα συστάσεων καλούνται να παράγουν προβλέψεις για σύνολα χρηστών τα οποία συνήθως αυξάνονται ενώ παράλληλα αυξάνονται και τα σύνολα των προϊόντων που παρέχονται από τα ηλεκτρονικά καταστήματα. Το γεγονός αυτό, δημιουργεί σημαντικές επιπτώσεις στην κλιμάκωση (scalability) των αλγορίθμων που χρησιμοποιούνται. Λύση στο συγκεκριμένο πρόβλημα παρέχουν κυρίως οι νέες προσεγγίσεις που αξιοποιούν τεχνικές παραγοντοποίησης πινάκων ενώ στις υπόλοιπες προσεγγίσεις γίνεται προσπάθεια για τον διαχωρισμό της αλγοριθμικής διαδικασίας σε στάδια εκτέλεσης και στη συνέχεια την αποθήκευση και επαναχρησιμοποίηση των ενδιάμεσων αποτελεσμάτων.

#### 2.4.3 Πολυσημία

Οι παραδοσιακές τεχνικές συστάσεων αδυνατούν να αναγνωρίσουν τη λανθάνουσα σχέση που υπάρχει ανάμεσα σε προϊόντα ίδιας φύσεως αλλά διαφορετικής ονομασίας ή μάρκας (polysemy & synonymy). Η αδυναμία αυτή αποτελεί σημαντικό πρόβλημα στην παραγωγή συστάσεων και συνήθως αντιμετωπίζεται με τεχνικές παραγοντοποίησης πινάκων.

#### 2.4.4 Απώλεια μεταβατικής σχέσης

Υπάρχει η πιθανότητα ένας χρήστης να συσχετιστεί υψηλά (μέσω των μετρικών απόστασης) με έναν δεύτερο χρήστη και αυτός με τη σειρά του με έναν τρίτο. Ωστόσο το γεγονός αυτό δεν εξασφαλίζει ότι ο πρώτος χρήστης θα συσχετιστεί υψηλά με τον τρίτο επειδή δεν έχουν εξετάσει κοινά προϊόντα. Το φαινόμενο αυτό (αγγλιστί «loss of neighbor transitivity») εμφανίζεται συχνά στις τεχνικές συνεργατικού φιλτραρίσματος και συνεπώς οδηγεί σε μικρότερη ακρίβεια προβλέψεων, αλλά επιλύεται όταν γίνεται χρήση προσεγγίσεων που βασίζονται στην ανάλυση γράφων.

#### 2.4.5 Άγνωστα προϊόντα & χρήστες

Ένα σύστημα συστάσεων αδυνατεί να παράγει προβλέψεις για χρήστες που δεν έχουν πραγματοποιήσει αξιολογήσεις (ή αγορές) καθώς και για προϊόντα που δεν έχουν αξιολογηθεί (ή αγοραστεί). Πολλά συστήματα απαιτούν έναν αριθμό αγορών ή αξιολογήσεων από τον χρήστη πριν δημιουργήσουν συστάσεις για τον ίδιο. Επίσης το πρόβλημα αντιμετωπίζεται με αξιοποίηση της έμμεσης ανατροφοδότησης.

#### 2.4.6 Υπερεπάρκεια δεδομένων

Συχνό φαινόμενο είναι η ύπαρξη μεγάλου όγκου αξιολογήσεων που πραγματοποιήθηκαν από έναν συγκεκριμένο χρήστη κατά τη διάρκεια ενός μεγάλου χρονικού διαστήματος. Τα περισσότερα συστήματα συστάσεων λαμβάνουν υπόψιν όλες τις διαθέσιμες πληροφορίες για έναν χρήστη, γεγονός που οδηγεί σε παραγωγή συστάσεων που δεν αντικατοπτρίζουν τα πιο πρόσφατα ενδιαφέροντα του χρήστη.

### 2.5 Αξιολόγηση συστημάτων σύστασης

Σημαντικό ζήτημα αποτελεί η αξιολόγηση των συστημάτων συστάσεων. Η απόδοση των συστημάτων αυτών μπορεί να εκτιμηθεί από διαφορετικές σκοπιές. Στο μεγαλύτερο κομμάτι της βιβλιογραφίας, κατά την διαδικασία της αξιολόγησης, δίνεται βάση κυρίως στην ποιότητα των παραγόμενων συστάσεων. Ωστόσο εξίσου σημαντικά κριτήρια για την ορθή λειτουργία ενός συστήματος συστάσεων αποτελούν ο υπολογιστικός χώρος και χρόνος που καταναλώνουν για την εξυπηρέτηση των χρηστών.

Στην παρούσα εργασία θα εστιάσουμε στα μέτρα που αξιολογούν την ποιότητα των παραγόμενων συστάσεων. Τα μέτρα αυτά χωρίζονται σε δύο διαφορετικές κατηγορίες ανάλογα με τον τρόπο που αντιμετωπίζεται το πρόβλημα της παραγωγής συστάσεων.

### 2.5.1 Πρόβλημα πρόβλεψης

Στην περίπτωση του προβλήματος πρόβλεψης η ποιότητα των παραγόμενων συστάσεων εκτιμάται υπολογίζοντας το μέσο απόλυτο σφάλμα (mean absolute error). Το συγκεκριμένο μέτρο εκτιμάει την απόκλιση ανάμεσα στις πραγματικές αξιολογήσεις των χρηστών και σε αυτές που εκτιμώνται από το σύστημα παραγωγής συστάσεων. Ο υπολογισμός πραγματοποιείται σύμφωνα με τον παρακάτω τύπο (όπου  $i$  ο εξεταζόμενος χρήστης και  $j$  ένα προϊόν που ανήκει στο σύνολο των προϊόντων που έχει εξετάσει ο χρήστης  $i$ ):

$$MAE_i = \frac{\sum_{j \in \text{προϊόντα}} |\text{πραγματική τιμή}_{ij} - \text{εκτιμώμενη τιμή}_{ij}|}{|\text{προϊόντα}|}$$

Έπειτα υπολογίζεται το συνολικό απόλυτο σφάλμα, ως μέσος όρος των μέσων απόλυτων σφαλμάτων όλων των χρηστών:

$$MAE = \frac{\sum_{i \in \text{χρήστες}} MAE_i}{|\text{χρήστες}|}$$

Σημειώνεται πως το παραπάνω μέτρο υπολογίζεται μόνο για γνωστές τιμές αξιολόγησης. Μικρότερες τιμές του συγκεκριμένου μέτρου υποδεικνύουν καλύτερη ποιότητα παραγόμενων συστάσεων.

### 2.5.2 Πρόβλημα n-καλύτερων προτάσεων

Στην περίπτωση που το πρόβλημα παραγωγής συστάσεων αντιμετωπίζεται ως πρόβλημα n-καλύτερων προτάσεων χρησιμοποιούνται μέτρα που αξιολογούν την κάλυψη (recall) και την ακρίβεια του συστήματος (precision).

Βασική προϋπόθεση για την ορθή αξιολόγηση είναι ο χωρισμός των αρχικών δεδομένων σε δύο σύνολα, εκ των οποίων το ένα θα χρησιμοποιηθεί στην εκπαίδευση του αλγορίθμου (σύνολο εκπαίδευσης) ενώ το δεύτερο στην αξιολόγησή του (σύνολο ελέγχου). Με αυτόν τον τρόπο το σύστημα αξιολογείται επί ενός συνόλου επιβεβαιωμένων αγορών, το οποίο αποκρύψαμε αφαιρώντας το από τα διαθέσιμα δεδομένα εισόδου. Μετά την παραγωγή των συστάσεων, τα στοιχεία που εμφανίζονται στις n-καλύτερες συστάσεις και ταυτόχρονα περιλαμβάνονται στο σύνολο ελέγχου αποτελούν το σύνολο επιτυχίας.

Το μέτρο της κάλυψης περιγράφει κατά πόσο το σύστημα παραγωγής συστάσεων προέβλεψε τις αξιολογήσεις που ανήκουν στο σύνολο ελέγχου. Μαθηματικά ορίζεται ως:

$$\text{recall} = \frac{|\text{σύνολο ελέγχου} \cap \text{σύνολο επιτυχίας}|}{|\text{σύνολο ελέγχου}|}$$

Το μέτρο της ακρίβειας περιγράφει κατά πόσο οι n συστάσεις που παράγει το σύστημα για τον κάθε χρήστη ανήκουν στο σύνολο ελέγχου. Μαθηματικά ορίζεται ως:

$$\text{precision} = \frac{|\text{σύνολο ελέγχου} \cap \text{σύνολο επιτυχίας}|}{|\text{σύνολο παραγόμενων συστάσεων}|}$$

Τα δύο αυτά μέτρα είναι αλληλοσυγκρουόμενα. Η αύξηση του αριθμού των παραγόμενων συστάσεων ανά χρήστη συνεπάγεται συνήθως αύξηση του μέτρου της κάλυψης αλλά παράλληλη μείωση του μέτρου της ακρίβειας.

Σύμφωνα με τη βιβλιογραφία [3] τα δύο αυτά μέτρα συνδυάζονται συνήθως σε ένα κοινό, γνωστό και ως μέτρο F1 το οποίο ορίζεται ως εξής:

$$F1 = \frac{2 * recall * precision}{recall + precision}$$

Τα μέτρα αυτά υπολογίζονται για κάθε χρήστη ξεχωριστά και έπειτα υπολογίζεται ο μέσος όρος για να προκύψει η αξιολόγηση του συστήματος.

## 2.6 Εφαρμογές

Στην ενότητα αυτή θα παρουσιάσουμε κάποια από τα συστήματα συστάσεων που ενσωματώνονται σε γνωστές διαδικτυακές πλατφόρμες. Στη βιβλιογραφία υπάρχουν αρκετές μελέτες εφαρμογής των συστημάτων συστάσεων [2], ωστόσο οι περισσότερες αναφέρονται σε ηλεκτρονικά καταστήματα και ιστοσελίδες που πλέον δε λειτουργούν.

### 2.6.1 Συστήματα συστάσεων στο ηλεκτρονικό εμπόριο

Αναμφισβήτητα η εφαρμογή των συστημάτων συστάσεων σε ηλεκτρονικά καταστήματα κρίνεται πλέον άκρως απαραίτητη. Παρακάτω περιγράφονται δύο περιπτώσεις χρήσης συστημάτων συστάσεων από δύο γνωστά ηλεκτρονικά καταστήματα.

#### 2.6.1.1 Amazon.co.uk

Ως κυρίαρχο κατάστημα στον χώρο του ηλεκτρονικού εμπορίου, το Amazon αποτελεί ένα από τα πρώτα καταστήματα που υιοθέτησε τεχνικές παραγωγής συστάσεων για να διευκολύνει τους χρήστες του στην επιλογή προϊόντων προς αγορά.

Όπως φαίνεται και στην Εικόνα 5, η ιστοσελίδα περιλαμβάνει ειδική οθόνη για τον χρήστη, όπου παρουσιάζονται συστάσεις προϊόντων βάσει της προηγούμενης συμπεριφοράς του. Δίνεται επίσης η δυνατότητα επιλογής συγκεκριμένης κατηγορίας προϊόντων, καθώς και νέων ή σύντομα διαθέσιμων προϊόντων.

Your Amazon.co.uk > Recommended for you  
(If you're not Stamatiadis Stamatis, click [here](#).)

**Recommendations**


- Amazon Video
- Appstore for Android
- Baby
- Books
- Books on Kindle
- Car
- Clothing
- Computers & Accessories
- DIY & Tools
- DVD & Blu-ray
- Electronics
- Garden & Outdoors
- Grocery
- Home & Garden
- Jewellery
- Large Appliances
- Lighting
- MP3 Downloads
- Music
- Musical Instruments
- PC & Video Games
- Pet Supplies
- Software
- Sports & Outdoors
- Toys & Games
- Video
- Watches

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon [More results](#)

---

1.




**Hydrograd** [EXPLICIT LYRICS]  
~ Stone Sour (30 Jun. 2017)  
Average Customer Review: ★★★★★ (50)  
In stock  
**Price: £9.99**  
[39 used & new](#) from £7.33

☐ I own it ☐ Not interested ☒ ★★★★★ Rate this item  
Recommended because you purchased **FOR WE ARE MANY** and more ( [Fix this](#) )

[Add to Basket](#) [Add to Wish List](#)

---

2.




**Incarnate**  
~ Killswitch Engage (11 Mar. 2016)  
Average Customer Review: ★★★★★ (34)  
In stock  
**Price: £9.99**  
[33 used & new](#) from £6.45

☐ I own it ☐ Not interested ☒ ★★★★★ Rate this item  
Recommended because you purchased **FOR WE ARE MANY** and more ( [Fix this](#) )

[Add to Basket](#) [Add to Wish List](#)

---

3.



**THE FALL OF IDEALS** [CD]  
~ All That Remains (27 Mar. 2015)  
Average Customer Review: ★★★★★ (9)  
In stock  
**Price: £10.16**  
[34 used & new](#) from £4.81

☐ I own it ☐ Not interested ☒ ★★★★★ Rate this item  
Recommended because you purchased **Overcome** and more ( [Fix this](#) )

[Add to Basket](#) [Add to Wish List](#)

Εικόνα 5: Συστάσεις προϊόντων της Amazon - πηγή: amazon.co.uk

Επιπλέον η ιστοσελίδα παροτρύνει τον χρήστη να αξιολογήσει αντικείμενα που έχει αγοράσει (μέσω άμεσης ανατροφοδότησης) έτσι ώστε να βελτιωθούν οι εξαγόμενες συστάσεις (Εικόνα 6).

Your Amazon.co.uk > Improve Your Recommendations  
(If you're not Stamatiadis Stamatias, click here.)

Help us make better recommendations. You can refine your recommendations by rating items or adjusting the checkboxes.

**EDIT YOUR COLLECTION**

**Items you've purchased**

Videos you've watched



Items you've marked "I own it"

Items you've rated

Items you've marked "Not interested"

**Need Help?**  
Visit our help area to learn more.

**Items you've purchased**

			Your Rating:
1.	 <p><b>MOTOROLA Moto G XT 1032 8 GB Black (Charger sold seperately)</b> by Motorola</p>		Rate this item <input checked="" type="checkbox"/> ★★★★★ <input type="checkbox"/> Don't use for recommendations
2.	 <p><b>Terrapin TPU Gel Skin Case/Cover for Motorola Moto G - Solid Black</b> by TERRAPIN</p>		Rate this item <input checked="" type="checkbox"/> ★★★★★ <input type="checkbox"/> Don't use for recommendations


Εικόνα 6: Παροχή άμεσης ανατροφοδότησης στην Amazon - πηγή: amazon.co.uk

### 2.6.1.2 Banggood

Ένα πολύ γνωστό ηλεκτρονικό κατάστημα με έδρα την Ασία είναι το Banggood. Το Banggood δημιουργήθηκε το 2006 και περιλαμβάνει πάνω από 150.000 προϊόντα. Αποτελεί διαδεδομένο κατάστημα στην Ελλάδα εξαιτίας των χαμηλών τιμών αλλά και της υψηλής αξιοπιστίας προς τους πελάτες.

Το κατάστημα παρουσιάζει συγκεκριμένο αριθμό συστάσεων τα οποία συνήθως είναι προϊόντα σε έκπτωση ή προσφορά (Εικόνα 7). Από την χρήση της ιστοσελίδας διαπιστώθηκε πως οι συστάσεις προκύπτουν όχι μόνο από τις αγορές του χρήστη αλλά και από τα προϊόντα που εξετάζει κατά την περιήγησή του στην ιστοσελίδα (αξιοποίηση έμμεσης ανατροφοδότησης).


Recommendations
Top Selling



Original Xiaomi Rolling Brush...

€10.15

★★★★★ (4)




21% OFF

BlitzWolf BW-F4 xBASS 10...

€46.57

€58-84


★★★★★ (411)



Original Xiaomi 2\*10W Smart...

€101.63

★★★★★ (5)




48% OFF

2Pcs Xiaomi Soocare-X3 Too...

€10.15

€19-47

★★★★★ (8)




42% OFF

IPRee Portable Travel Lazy...

€10.58

€18-20

★★★★★ (664)




25% OFF

ANENG AN8008 True RMS ...

€15.24

€20-23

★★★★★ (95)




28% OFF

Eachine E010 Mini 2.4G 4CH...

€11.00

€15-24

★★★★★ (4301)




38% OFF

Eachine VTX03 Super Mini 5...

€8.94

€14-39


★★★★★ (1132)



Drill Holster Cordless Tool Ho...

€5.07

★★★★★ (1)



17% OFF

DOOGEE MIX 5.5 Inch Andr...

€160.91

€194-79

★★★★★ (918)

Εικόνα 7: Συστάσεις προϊόντων του Banggood - πηγή: banggood.com



Επιπλέον στην Εικόνα 8, διαπιστώθηκε πως κατά την εξέταση ενός συγκεκριμένου προϊόντος παρουσιάζονται επιπλέον προϊόντα υπό τη μορφή συστάσεων, γεγονός που δικαιολογεί την αξιοποίηση τεχνικών συνεργατικού φιλτραρίσματος βασισμένο στα προϊόντα.

Customers Who View This Item Also View:



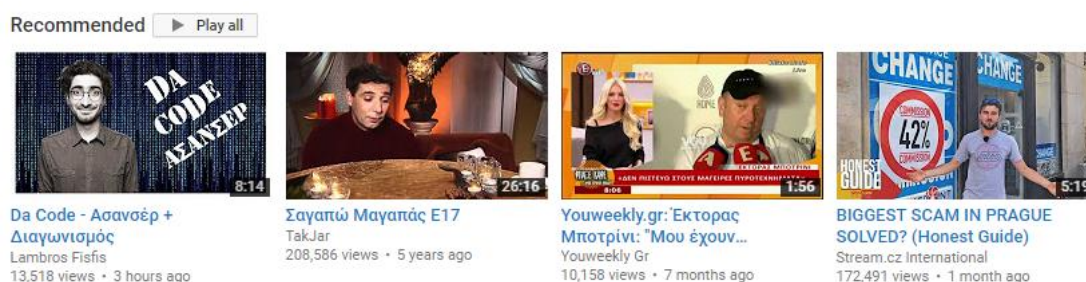
Εικόνα 8: Συστάσεις προϊόντων του Banggood - πηγή: banggood.com

## 2.6.2 Συστήματα συστάσεων σε άλλα πεδία εφαρμογής

Πέρα από το ηλεκτρονικό εμπόριο, τα συστήματα συστάσεων μπορούν να αξιοποιηθούν σε οποιαδήποτε διαδικτυακή πλατφόρμα που φιλοξενεί και οργανώνει με συστηματικό τρόπο οποιοδήποτε είδος περιεχομένου.

### 2.6.2.1 YouTube

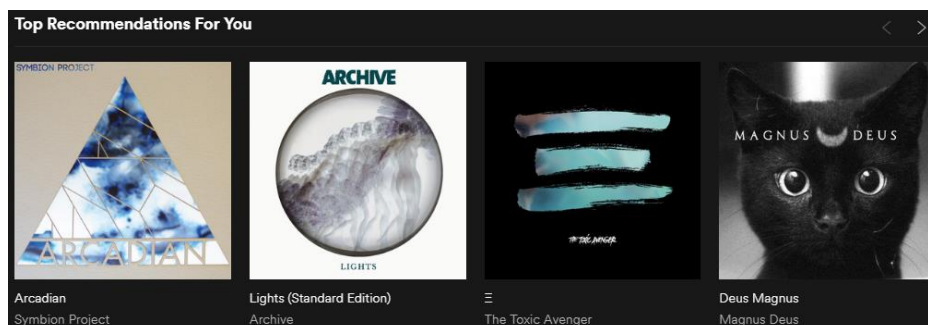
Το YouTube αποτελεί τη μεγαλύτερη πλατφόρμα φιλοξενίας οπτικοακουστικού υλικού. Κατά την επίσκεψη του χρήστη στην ιστοσελίδα παρουσιάζονται άμεσα συστάσεις για την προβολή βίντεο (Εικόνα 9). Οι συστάσεις βασίζονται σε περιεχόμενο που έχει εξετάσει στο παρελθόν ο χρήστης καθώς και στις ρητές προτιμήσεις που έχει ο ίδιος δηλώσει.



Εικόνα 9: Συστάσεις βίντεο του YouTube - πηγή: YouTube.com

### 2.6.2.2 Spotify

Το Spotify χαρακτηρίζεται ως η πιο επιτυχημένη διαδικτυακή πλατφόρμα μουσικής. Όπως και στις προηγούμενες περιπτώσεις, η παρουσίαση των συστάσεων γίνεται στο κεντρικό μενού της εφαρμογής και βασίζεται στις προηγούμενες μουσικές επιλογές του χρήστη.



Εικόνα 10: Συστάσεις μουσικής του Spotify - πηγή: Spotify





### 3 Αξιολόγηση αντίκτυπου έμμεσης ανατροφοδότησης στο ηλεκτρονικό εμπόριο

Στο κεφάλαιο αυτό αναλύεται η ερευνητική συνεισφορά της παρούσας εργασίας. Γίνεται λόγος για τα σύνολα δεδομένων που χρησιμοποιήθηκαν, τα στάδια προεπεξεργασίας και τους αλγορίθμους που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων. Τέλος παρουσιάζονται τα αποτελέσματα των πειραμάτων.

#### 3.1 Κίνητρο

Η παρούσα εργασία αποτελεί μία απόπειρα επινόησης λύσεων για κάποια από τα προβλήματα που εμφανίζονται κατά τη δημιουργία και χρήση των συστημάτων συστάσεων. Πιο συγκεκριμένα, ανάμεσα στα προβλήματα που αναφέρθηκαν στην ενότητα 2.4, μας απασχολεί κυρίως το πρόβλημα της έλλειψης πυκνότητας στον πίνακα δεδομένων εισόδου. Ωστόσο οι λύσεις που παρουσιάζονται στις επόμενες ενότητες συνεισφέρουν παράλληλα και στην επίλυση των προβλημάτων της κλιμάκωσης και της πολυσημίας εξαιτίας της φύσης των τεχνικών που χρησιμοποιούνται. Η εργασία εστιάζει σε δεδομένα που αφορούν ηλεκτρονικά καταστήματα, αλλά τα ευρήματα μπορούν να εφαρμοστούν εύκολα σε οποιοδήποτε πεδίο προβλήματος.

Το πρόβλημα έλλειψης πυκνότητας οφείλεται κυρίως στην έλλειψη ικανοποιητικού όγκου δεδομένων άμεσης ανατροφοδότησης. Παράλληλα η εξασφάλιση περισσότερων αξιολογήσεων από τους χρήστες μιας ιστοσελίδας αποτελεί δύσκολο εγχείρημα. Η λύση που προτείνεται αφορά την αξιοποίηση της έμμεσης ανατροφοδότησης που προκύπτει από την καταγραφή της συμπεριφοράς των χρηστών κατά την πλοήγησή τους σε μια διαδικτυακή πλατφόρμα. Πολλά συστήματα αξιοποιούν ήδη δεδομένα έμμεσης ανατροφοδότησης τα οποία όμως περιορίζονται μόνο στις αγορές των χρηστών και πολλές φορές αντιμετωπίζονται από τη βιβλιογραφία ως δεδομένα άμεσης ανατροφοδότησης. Στην εργασία αυτή, παρουσιάζουμε τρόπους που καθιστούν δυνατή την αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης όπως για παράδειγμα δεδομένα που καταγράφουν τα προϊόντα που εξέτασε ο χρήστης ή προϊόντα που πρόσθεσε στο καλάθι του.

Τα πειράματα που πραγματοποιήθηκαν αφορούν τη μελέτη των επιδόσεων που παρουσιάζουν δύο διαφορετικές τεχνικές παραγωγής συστάσεων, αυτή του συνεργατικού φιλτραρίσματος και αυτή της ανάλυσης πίνακα σε ιδιάζουσες τιμές, σε δύο διαφορετικά σενάρια. Στο πρώτο γίνεται αξιοποίηση μόνο των δεδομένων που αφορούν τις αγορές των χρηστών ενώ στο δεύτερο αξιοποιούνται δεδομένα που αφορούν επιπλέον ενέργειες των χρηστών (συμπεριλαμβανομένης και της ενέργειας αγοράς). Τα αποτελέσματα που προκύπτουν επιβεβαιώνουν την υπόθεση ότι η αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης ενισχύουν την ποιότητα των παραγόμενων συστάσεων.

#### 3.2 Σχετική έρευνα

Στην ενότητα αυτή αναφέρονται σχετικές δημοσιεύσεις και ερευνητικό υλικό που μελετήθηκε κατά τη συγγραφή της εργασίας και την εκτέλεση των πειραμάτων.

Μελετώντας τη σχετική βιβλιογραφία διαπιστώθηκε πως η πιο διαδεδομένη τεχνική παραγωγής συστάσεων είναι η προσέγγιση του συνεργατικού φιλτραρίσματος [10], [3], [11]. Οι πρώτες δημοσιεύσεις αφορούσαν συνεργατικό φιλτράρισμα βασισμένο στους χρήστες αλλά στην πορεία προτάθηκαν προσεγγίσεις βασισμένες στα προϊόντα οι οποίες παρουσιάζουν βελτιώσεις ως προς την ποιότητα των συστάσεων αλλά και την απόδοση των συστημάτων [10], [12], [13].

Σημαντικός όγκος δημοσιεύσεων αφορά και την προσέγγιση ανάλυσης πίνακα σε ιδιάζουσες τιμές, μέθοδος που επιλύει το πρόβλημα της πολυσημίας [14], [15], [16]. Μάλιστα, υπάρχουν ακόμα και δημοσιεύσεις που συνδυάζουν τις δύο προαναφερόμενες προσεγγίσεις [17], [18].

Ωστόσο οι περισσότερες από τις παραπάνω εργασίες αξιοποιούν μόνο άμεση ανατροφοδότηση. Όσο αναφορά την χρήση έμμεσης ανατροφοδότησης, οι μεθοδολογίες που χρησιμοποιούνται εστιάζουν σε διάφορες προσεγγίσεις όπως παραγοντοποίηση πινάκων (εναλλασσόμενα ελάχιστα τετράγωνα) [5] ή κανόνες συσχέτισης [4], [19]. Οι εργασίες που συνδυάζουν τεχνικές συνεργατικού φιλτραρίσματος και έμμεση ανατροφοδότηση [20], [21], [22] συνήθως χαρακτηρίζονται από σύνθετα στάδια προεπεξεργασίας ή πεδίο εφαρμογής διαφορετικό από το ηλεκτρονικό εμπόριο. Επίσης σε αρκετές περιπτώσεις παρουσιάζονται υβριδικά συστήματα που αξιοποιούν δεδομένα έμμεσης ανατροφοδότησης [23], [24] τα οποία παρουσιάζουν επίσης υψηλή πολυπλοκότητα. Η μόνη δημοσίευση που βρέθηκε με κοινούς ερευνητικούς στόχους [25] αξιοποιεί μόνο τις τεχνικές συνεργατικού φιλτραρίσματος και χαρακτηρίζεται από σύνθετα στάδια προεπεξεργασίας.

### 3.3 Σύνολα δεδομένων

Στο διαδίκτυο διατίθεται πληθώρα συνόλων δεδομένων που μπορούν να αξιοποιηθούν για τη δοκιμή και αξιολόγηση ενός αλγορίθμου παραγωγής συστάσεων. Ωστόσο τα περισσότερα από αυτά περιέχουν στοιχεία άμεσης ανατροφοδότησης υπό τη μορφή αξιολογήσεων. Ακόμα και στην περίπτωση της έμμεσης ανατροφοδότησης, τα περισσότερα σύνολα δεδομένων καταγράφουν μόνο τις αγορές των χρηστών. Υπάρχει σημαντική έλλειψη συνόλων που να περιλαμβάνουν πληροφορίες σχετικά με τη συμπεριφορά των χρηστών και να αφορούν παράλληλα περιπτώσεις ηλεκτρονικού εμπορίου. Τα παρακάτω σύνολα είναι τα μόνα που διανέμονται ελεύθερα στο διαδίκτυο (την παρούσα στιγμή) και εξυπηρετούν ταυτόχρονα τις ανάγκες της εργασίας.

#### 3.3.1 Πηγές δεδομένων

Κοινό χαρακτηριστικό όλων των συνόλων δεδομένων που παρουσιάζονται αποτελεί η ανάρτησή τους σε διαδικτυακές πλατφόρμες διαγωνισμών μηχανικής μάθησης.

##### 3.3.1.1 Σύνολο δεδομένων *Retailrocket*

Το πρώτο σύνολο δεδομένων που χρησιμοποιήθηκε για τις ανάγκες της παρούσας εργασίας προσφέρεται ελεύθερα μέσω της διαδικτυακής πλατφόρμας Kaggle [26]. Τα δεδομένα προέρχονται από άγνωστο αλλά πραγματικό ηλεκτρονικό κατάστημα και περιλαμβάνουν πληροφορίες για τις ιδιότητες των προϊόντων, τις κατηγορίες προϊόντων και τη συμπεριφορά των χρηστών που καταγράφηκε σε χρονικό διάστημα 4,5 μηνών. Η συμπεριφορά ενός χρήστη χαρακτηρίζεται από τρεις διαφορετικές ενέργειες: 1) προβολή προϊόντος, 2) προσθήκη προϊόντος στο καλάθι και 3) αγορά προϊόντος. Για τις ανάγκες της εργασίας αξιοποιήθηκε μόνο το αρχείο «events.csv» το οποίο περιλαμβάνει τα παρακάτω πεδία: «timestamp, visitorid, event, itemid, transactionid». Από τα προαναφερόμενα συκρατήθηκαν μόνο τα πεδία: «visitorid», «event» και «itemid».

##### 3.3.1.2 Σύνολο δεδομένων *Alibaba*

Το δεύτερο σύνολο δεδομένων προσφέρεται μέσω της διαδικτυακής πλατφόρμας Tianchi [27]. Πρόκειται για δεδομένα που συλλέχθηκαν από το ηλεκτρονικό κατάστημα Alibaba κατά το έτος 2014 και εστιάζουν κυρίως στο ηλεκτρονικό εμπόριο μέσω κινητών συσκευών (mobile commerce). Τα δεδομένα περιλαμβάνουν πληροφορίες για τις κατηγορίες προϊόντων, τη συμπεριφορά αλλά και τη γεωγραφική τοποθεσία των χρηστών. Όσο αναφορά τη

συμπεριφορά των χρηστών, καταγράφονται οι ενέργειες: 1) προβολή προϊόντος, 2) προσθήκη προϊόντος στα αγαπημένα, 3) προσθήκη προϊόντος στο καλάθι και 4) αγορά προϊόντος. Για τις ανάγκες της εργασίας αξιοποιήθηκε μόνο το αρχείο «tianchi mobile recommend train user.csv» το οποίο περιλαμβάνει τα παρακάτω πεδία: «user\_id, item\_id, behavior\_type, user\_geohash, item\_category, time». Από τα παραπάνω συγκρατήθηκαν μόνο τα πεδία: «user\_id», «item\_id» και «behavior\_type».

### 3.3.1.3 Σύνολο δεδομένων Tmall

Το τελευταίο σύνολο δεδομένων προσφέρεται επίσης από τη διαδικτυακή πλατφόρμα Tianchi [28], [29]. Τα δεδομένα παρέχονται από το ηλεκτρονικό κατάστημα Tmall και περιλαμβάνουν πληροφορίες για τη συμπεριφορά των χρηστών σε χρονικό διάστημα 6 μηνών. Οι ενέργειες είναι όμοιες με αυτές του συνόλου δεδομένων Alibaba. Για τις ανάγκες της εργασίας αξιοποιήθηκε μόνο το αρχείο «user\_log\_format1.csv» το οποίο περιλαμβάνει τα παρακάτω πεδία: «user\_id, item\_id, cat\_id, seller\_id, brand\_id, time\_stamp, action\_type». Από τα προαναφερόμενα συγκρατήθηκαν μόνο τα πεδία: «user\_id», «item\_id» και «action\_type».

### 3.3.2 Ανάλυση αρχικών δεδομένων

Στην παρούσα ενότητα παρουσιάζονται τα βασικά χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν κατά τη δοκιμή των μεθόδων παραγωγής συστάσεων. Και τα τρία σύνολα περιλαμβάνουν δεδομένα υπό μορφή ημερολογίου συμβάντων (event log). Στον Πίνακα 1 αναγράφονται στοιχεία για τα μεγέθη των δεδομένων.

	Retailrocket	Alibaba	Tmall
μέγεθος αρχείου (KB)	133	705	2.048
αριθμός εγγραφών	2.756.101	12.256.906	54.925.330
αριθμός χρηστών	11.719	8.886	333.204
αριθμός προϊόντων	12.025	92.753	234.592

Πίνακας 1: Μεγέθη των συνόλων δεδομένων

Επιπλέον, στον επόμενο πίνακα (Πίνακας 2) καταγράφονται οι διαθέσιμοι τύποι συμβάντων (ενέργειες χρηστών επί των προϊόντων) ανά σύνολο δεδομένων.

	Retailrocket	Alibaba	Tmall
προβολή	✓	✓	✓
προσθήκη στα αγαπημένα	✗	✓	✓
προσθήκη στο καλάθι	✓	✓	✓
αγορά	✓	✓	✓

Πίνακας 2: Ενέργειες έμμεσης ανατροφοδότησης ανά σύνολο δεδομένων

Σύμφωνα με την προκαταρκτική μελέτη των δεδομένων και των πινάκων 1 και 2 καταλήγουμε στις παρακάτω διαπιστώσεις:

- Τα τρία σύνολα δεδομένων διαφέρουν σημαντικά σε μέγεθος, αριθμό εγγραφών καθώς και αριθμό χρηστών/προϊόντων. Το γεγονός αυτό διευρύνει τις δοκιμές των αλγορίθμων παραγωγής συστάσεων σε τρία διαφορετικά σενάρια, όπου κάθε σενάριο χαρακτηρίζεται από διαφορετικό όγκο δεδομένων. Ωστόσο, όπως θα δούμε και παρακάτω, ο χαμηλός όγκος δεδομένων των συνόλων Retailrocket και Alibaba έχει ως συνέπεια τη μειωμένη ακρίβεια των συστάσεων κατά την αξιολόγηση των τεχνικών παραγωγής συστάσεων.

- Η αναλογία χρηστών και προϊόντων του συνόλου δεδομένων Alibaba διαφέρει σημαντικά σε σύγκριση με τα άλλα δύο σύνολα. Τα προϊόντα είναι δέκα φορές περισσότερα σε αριθμό από τον αριθμό των χρηστών.
- Πολλοί χρήστες επαναλαμβάνουν την ίδια ενέργεια προς ένα συγκεκριμένο προϊόν, π.χ. επαναλαμβανόμενες προβολές ή αγορές ενός προϊόντος.
- Πολλοί χρήστες εκτελούν διαφορετικές ενέργειες προς το ίδιο προϊόν, μερικές φορές χωρίς την ύπαρξη ορθής χρονικής αλληλουχίας, π.χ. καταγραφή συμβάντος προβολής ακολουθούμενο από συμβάν αγοράς χωρίς να προηγείται ενέργεια προσθήκης στο καλάθι.
- Τα δεδομένα περιλαμβάνουν διπλότυπα.

### 3.4 Προεπεξεργασία Δεδομένων

Σκοπός της παρούσας φάσης είναι η μετατροπή των συνόλων δεδομένων από την αρχική τους μορφή (μορφή ημερολογίου συμβάντων) σε μορφή αξιοποιήσιμη από τους αλγορίθμους παραγωγής συστάσεων. Πιο συγκεκριμένα, απαιτείται η οργάνωση των δεδομένων σε μορφή δισδιάστατου πίνακα όπου κάθε τιμή αντιπροσωπεύει τη σχέση μεταξύ ενός χρήστη και ενός προϊόντος. Το συγκεκριμένο εγχείρημα είναι δύσκολο αλλά παράλληλα ύψιστης σημασίας εξαιτίας της ιδιαίτερης φύσης των δεδομένων (έμμεση ανατροφοδότηση).

#### 3.4.1 Παραδοχές

Η μετατροπή των δεδομένων από την αρχική τους μορφή σε μορφή χρήσιμη για τα επόμενα βήματα απαιτεί την υιοθέτηση ξεκάθαρα ορισμένων παραδοχών. Έπειτα, με τη βοήθεια των παραδοχών αυτών, δημιουργούμε μία αλγοριθμική διαδικασία για την αξιοποίηση ακόμα και νέων δεδομένων. Λαμβάνοντας υπόψιν τις διαπιστώσεις της ενότητας 3.3.2 και βασιζόμενοι σε προηγούμενες βιβλιογραφικές εργασίες [25] ορίζουμε τις παρακάτω παραδοχές:

- Δεν λαμβάνεται υπόψιν η συχνότητα εκτέλεσης μίας ενέργειας από συγκεκριμένο χρήστη προς συγκεκριμένο προϊόν. Το γεγονός ότι ένας χρήστης έχει αγοράσει επανειλημμένα ένα συγκεκριμένο προϊόν δεν επηρεάζει τον βαθμό αποδοχής του.
- Δεν αξιοποιείται η χρονική στιγμή εκτέλεσης του κάθε συμβάντος καθώς και η κατηγορία στην οποία ανήκει το κάθε προϊόν.
- Θεωρείται ότι οι διαθέσιμες ενέργειες ταξινομούνται με συγκεκριμένη σειρά και ιεραρχία όπως στο [19]. Η εκτέλεση μιας ενέργειας (με εξαίρεση την προβολή ενός προϊόντος) υπονοεί ότι προηγήθηκαν οι/η προηγούμενες/-η ενέργειες/-α.
- Διατηρούμε μόνο την πιο σημαντική ενέργεια για κάθε συνδυασμό χρήστη και προϊόντος. Εάν για παράδειγμα έχουν καταγραφεί οι ενέργειες προβολής, προσθήκης στα αγαπημένα και αγοράς μεταξύ ενός χρήστη και ενός προϊόντος, συγκρατείται μόνο το συμβάν της αγοράς.
- Απαλείφουμε χρήστες και προϊόντα για τα οποία έχει καταγραφεί μόνο μία ενέργεια αγοράς. Αν και είναι εφικτό να παραχθούν συστάσεις για τους χρήστες που έχουν πραγματοποιήσει μόνο μία αγορά, η αξιοποίησή τους κατά την αξιολόγηση είναι αδύνατη. Οι συγκεκριμένες οντότητες εάν ενσωματωθούν στα σύνολα ελέγχου θα είναι αδύνατο να προβλεφθούν ενώ εάν ενσωματωθούν στα σύνολα εκπαίδευσης θα εισάγουν θόρυβο και άχρηστη γνώση. Συνεπώς μας ενδιαφέρουν οντότητες που

χαρακτηρίζονται από τουλάχιστον 2 ενέργειες αγοράς. Στην περίπτωση των δεδομένων του Tmall διατηρούμε οντότητες με τουλάχιστον 6 ενέργειες αγοράς.

- Δεν πραγματοποιείται καμία επιπλέον προεπεξεργασία πέρα από αυτές που αναφέρονται στις επόμενες ενότητες. Σύμφωνα με τη βιβλιογραφία [3] προτείνεται η αντικατάσταση των μηδενικών του πίνακα με κάποια προεπιλεγμένη τιμή (για παράδειγμα ο μέσος όρος των αξιολογήσεων του χρήστη). Επίσης πολλές φορές προτείνεται η κανονικοποίηση των δεδομένων εισόδου. Στην παρούσα εργασία κρίνεται πως κάτι τέτοιο δεν είναι αναγκαίο εξαιτίας της φύσης των δεδομένων.

### 3.4.2 Στάδια προεπεξεργασίας

Σύμφωνα με τις παραδοχές της ενότητας 3.4.1 ορίζεται η παρακάτω ροή προεπεξεργασίας των αρχικών δεδομένων:

1. Απομόνωση των χρήσιμων πεδίων (αναγνωριστικό χρήστη, προϊόντος, ενέργεια).
2. Απαλοιφή διπλοτύπων και διατήρηση μόνο της πιο σημαντικής ενέργειας.
3. Απαλοιφή χρηστών και προϊόντων που χαρακτηρίζονται από μόνο μία μόνο αγορά.
4. Εκ νέου αρίθμηση και ανακατανομή των χρηστών και προϊόντων.
5. Υπολογισμός και ανάθεση τιμών που θα εισαχθούν στον δισδιάστατο πίνακα.

### 3.4.3 Ανάλυση επεξεργασμένων δεδομένων

Αφού ολοκληρωθεί η προεπεξεργασία των δεδομένων και η μετατροπή τους σε μορφή αποδεκτή από τους αλγορίθμους παραγωγής συστάσεων προκύπτουν τα παρακάτω νέα στοιχεία για τα μεγέθη των συνόλων δεδομένων.

	Retailrocket	Alibaba	Tmall
αριθμός ελάχιστων αγορών ανά οντότητα	2	2	6
αριθμός χρηστών	799	3.387	7.155
αριθμός προϊόντων	1.532	5.005	4.069
αριθμός συμβάντων	12.998	61.673	174.700
μείωση χρηστών (%)	-93,18%	-61,88%	-97,85%
μείωση προϊόντων (%)	-87,26%	-94,60%	-98,27%

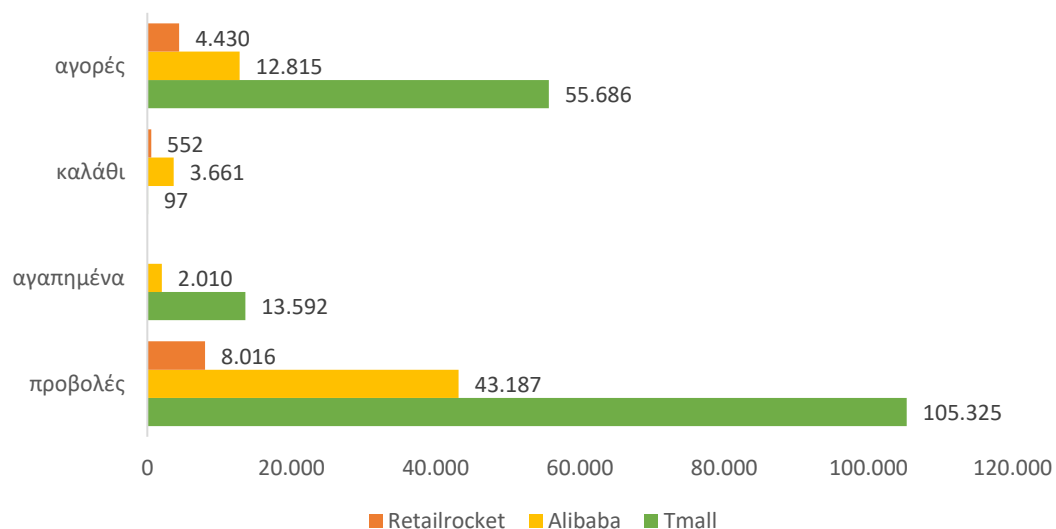
Πίνακας 3: Μεγέθη των συνόλων δεδομένων μετά την προεπεξεργασία

Σύμφωνα με τον Πίνακα 3 παρατηρείται δραματική μείωση στους αριθμούς των χρηστών αλλά και των προϊόντων. Το γεγονός αυτό οφείλεται στην αναδρομική διαδικασία απαλοιφής οντοτήτων που χαρακτηρίζονται από μόνο μία ενέργεια αγοράς. Στην Εικόνα 11 παρουσιάζεται η συνάρτηση που εκτελεί την προαναφερθείσα εργασία.

Δυστυχώς η απαλοιφή των μη επιθυμητών οντοτήτων μπορεί να δημιουργήσει νέες οντότητες που πρέπει να απαλειφθούν επίσης. Για παράδειγμα, έστω ένας χρήστης ο οποίος έχει πραγματοποιήσει δύο αγορές διαφορετικών προϊόντων εκ των οποίων το ένα έχει αγοραστεί μόνο από αυτόν. Η διαγραφή του προϊόντος καθιστά τον συγκεκριμένο χρήστη υποψήφιο για απαλοιφή.

```
@tailrec
def cleanDataset(eventsBuy: Iterable[(Int, Int, Int)],
                 users: Set[Int], items: Set[Int],
                 minTransactions: Int):
  (Set[Int], Set[Int]) = {
    println((users.size, items.size))
    val eventsNew = eventsBuy.filter(X => users.contains(X._1) && items.contains(X._2))
    val usersNew = eventsNew.groupBy(_._1).filter(_._2.size >= minTransactions).keySet
    val itemsNew = eventsNew.groupBy(_._2).filter(_._2.size >= minTransactions).keySet
    if(usersNew.equals(users) && itemsNew.equals(items))
      (users, items)
    else
      cleanDataset(eventsNew, usersNew, itemsNew, minTransactions)
  }
```

Εικόνα 11: Απαλοιφή οντοτήτων με μόνο μία ενέργεια αγοράς



Διάγραμμα 1: Συχνότητα εμφάνισης κάθε τύπου ενέργειας ανά σύνολο δεδομένων

Στο Διάγραμμα 1 αποτυπώνεται η συχνότητα εμφάνισης κάθε ενέργειας ανά σύνολο δεδομένων. Παρατηρείται ότι υπερισχύουν οι ενέργειες της προβολής και αγοράς.

	Retailrocket	Alibaba	Tmall
πυκνότητα πίνακα (αγορές)	0,00362	0,00076	0,00191
πυκνότητα (όλες οι ενέργειες)	0,01062	0,00364	0,00600
αύξηση πυκνότητας (%)	193,41%	381,26%	213,72%

Πίνακας 4: Πυκνότητα συνόλων δεδομένων

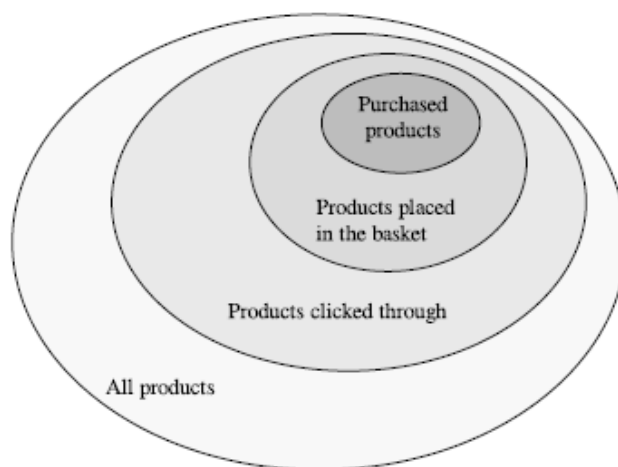
Τέλος στον Πίνακα 4 παρατηρούμε την πυκνότητα των πινάκων εισόδου για όλα τα σύνολα δεδομένων σε δύο διαφορετικές περιπτώσεις. Στην πρώτη περίπτωση, υπολογίζεται η πυκνότητα λαμβάνοντας υπόψιν μόνο τις αγορές προϊόντων ενώ στη δεύτερη περίπτωση συνυπολογίζονται όλες οι ενέργειες. Η αξιοποίηση των επιπλέον ενεργειών ενισχύει δραματικά την αύξηση της πυκνότητας του πίνακα εισόδου. Επιπλέον, στο σύνολο δεδομένων Alibaba κατά την αξιοποίηση μόνο των αγορών σημειώνεται ιδιαίτερα χαμηλή πυκνότητα.



#### 3.4.4 Ανάθεση τιμών προτίμησης

Το τελευταίο στάδιο της προεπεξεργασίας αφορά την ανάθεση τιμών στις θέσεις του πίνακα εισόδου. Τα δεδομένα που έχουμε στη διάθεσή μας αφορούν ενέργειες χρηστών επί των προϊόντων του καταστήματος.

Βασιζόμενοι στην εργασία των Cho, Kim και Kim [19], υιοθετούμε την ταξινόμηση των ενεργειών των χρηστών που παρουσιάστηκε στην εργασία τους. Σύμφωνα με το [19], θεωρούμε ότι όλοι οι χρήστες ενός ηλεκτρονικού καταστήματος πραγματοποιούν τις αγορές τους ακολουθώντας τρία ή τέσσερα διαδοχικά βήματα. Με αυτό τον τρόπο, μπορούμε να οργανώσουμε όλα τα προϊόντα σε συγκεκριμένο αριθμό διαφορετικών ομάδων, βάσει των ενεργειών που τους έχουν ασκηθεί, όπως ακριβώς φαίνεται στην Εικόνα 12. Επιπλέον θεωρούμε ότι ορισμένες ομάδες αποτελούν υποσύνολο κάποιων άλλων. Για παράδειγμα η ομάδα των προϊόντων που αγοράστηκαν αποτελεί υποσύνολο της ομάδα των προϊόντων που τοποθετήθηκαν στο καλάθι επειδή η ενέργεια προσθήκης στο καλάθι είχε προηγηθεί σε προγενέστερο στάδιο κατά την αγορά τους.



Εικόνα 12: Ταξινόμηση ενεργειών ηλεκτρονικού καταστήματος σε ομάδες – πηγή: [19]

Σκοπός μας είναι η ανάθεση μιας τιμής που χαρακτηρίζει τη σχέση αναμεσά σε έναν χρήστη και ένα προϊόν. Κοινά αποδεκτή πρακτική στη βιβλιογραφία είναι η ανάθεση της τιμής 1 στα προϊόντα που έχουν αγοραστεί από τον χρήστη και 0 στα προϊόντα που δεν έχουν εξεταστεί και είναι άγνωστα για τον χρήστη. Μας ενδιαφέρει λοιπόν η εύρεση τιμών εντός του συνεχούς διαστήματος τιμών  $[0,1]$  που θα αντιστοιχηθούν στις ενέργειες προβολή, προσθήκη στα αγαπημένα και προσθήκη στο καλάθι.

Στην εργασία των Kim, Yum, Song και Kim [25] ακολουθείται η ίδια ακριβώς διαδικασία. Για την ενέργεια προσθήκης προϊόντος στο καλάθι, οι συγγραφείς υπολογίζουν την αναμενόμενη πιθανότητα αγοράς μετά από προσθήκη στο καλάθι σύμφωνα με τον παρακάτω τύπο:

$$p = \frac{\text{συνολικός αριθμός περιπτώσεων όπου το προϊόν αγοράζεται}}{\text{συνολικός αριθμός περιπτώσεων όπου το προϊόν τοποθετείται στο καλάθι}}$$

Ωστόσο για τον υπολογισμό της τιμής που θα ανατεθεί στα συμβάντα προβολής προϊόντος, οι συγγραφείς ακολουθούν μία πιο σύνθετη προσέγγιση. Πιο συγκεκριμένα, γίνεται προσπάθεια υπολογισμού της αναμενόμενης πιθανότητας προσθήκης στο καλάθι μετά από προβολή του προϊόντος μέσω της κατασκευής μοντέλων μηχανικής μάθησης (δοκιμάζονται τρεις διαφορετικές προσεγγίσεις: δέντρα αποφάσεων, νευρωνικά δίκτυα, παλινδρόμηση).

Στα συγκεκριμένα μοντέλα, αξιοποιούνται επιπλέον δεδομένα έμμεσης ανατροφοδότησης (αριθμός επισκέψεων, διάρκεια ανάγνωσης κ.α.) τα οποία όμως δεν είναι διαθέσιμα στη δική μας περίπτωση.

Η εναλλακτική προσέγγιση που παρουσιάζεται στην παρούσα εργασία αφορά τον υπολογισμό των αναμενόμενων πιθανοτήτων αγοράς μετά από κάθε τύπο ενέργειας χρήση. Έχοντας γνώση του συνολικού αριθμού των ενεργειών που πραγματοποιήθηκαν σε κάθε σύνολο δεδομένων, μπορούμε εύκολα να εκτελέσουμε τους απαραίτητους υπολογισμούς. Στον Πίνακα 5 παρουσιάζεται ο συνολικός αριθμός κάθε τύπου ενέργειας που καταγράφηκε στα διαθέσιμα σύνολα δεδομένων.

	<b>Retailrocket</b>	<b>Alibaba</b>	<b>Tmall</b>
προβολές	8.016	43.187	105.325
αγαπημένα	-	2.010	13.592
καλάθι	552	3.661	97
αγορές	4.430	12.815	55.686

Πίνακας 5: Συνολικός αριθμός κάθε τύπου ενέργειας ανά σύνολο δεδομένων

Για τον υπολογισμό της αναμενόμενης πιθανότητας μιας ενέργειας  $\varepsilon$  χρησιμοποιούμε τον παρακάτω τύπο:

$$p(\varepsilon) = \frac{\text{αριθμός αγορών}}{\text{αριθμός συμβάντων που περιλαμβάνουν την ενέργεια } \varepsilon}$$

Για παράδειγμα εάν θέλουμε να υπολογίσουμε την αναμενόμενη πιθανότητα αγοράς μετά από προβολή προϊόντος υπολογίζουμε το παρακάτω αποτέλεσμα:

$$p(\text{προβολής}) = \frac{\# \text{αγορών}}{\# \text{προβολών} + \# \text{αγαπημένα} + \# \text{καλάθι} + \# \text{αγορών}}$$

Το οποίο στην περίπτωση του συνόλου δεδομένων Alibaba ισούται με:

$$p(\text{προβολής}) = \frac{12.815}{43.187 + 2.010 + 3.661 + 12.815} = \frac{12.815}{61.673} = 0,2077 \cong 0,21$$

Στον Πίνακα 6 παρουσιάζονται οι τελικές αναμενόμενες πιθανότητες για κάθε είδος ενέργειας και σύνολο δεδομένων.

	<b>Retailrocket</b>	<b>Alibaba</b>	<b>Tmall</b>
προβολές	0,34	0,21	0,32
αγαπημένα	-	0,69	0,80
καλάθι	0,89	0,78	1,00
αγορές	1,00	1,00	1,00

Πίνακας 6: Τελικά σκορ που αντικαθιστούν κάθε ενέργεια με μία τιμή προτίμησης

Η συγκεκριμένη προσέγγιση παρουσιάζει τα παρακάτω οφέλη:

- Δεν είναι υπολογιστικά ασύμφορη. Τα δεδομένα που απαιτούνται για τον υπολογισμό των τιμών προτιμήσεων είναι άμεσα διαθέσιμα εφόσον καταγράφονται από τη βάση δεδομένων του ηλεκτρονικού καταστήματος.



- Δίνεται η δυνατότητα εκ νέου υπολογισμού των παραπάνω τιμών σε περίπτωση που αλλάξει δραστικά η συμπεριφορά των χρηστών ως προς την εκτέλεση των διάφορων ενεργειών.

Εκτελώντας την παραπάνω διαδικασία μπορούμε πλέον να δημιουργήσουμε τον πίνακα εισόδου σε μορφή αποδεκτή από τους αλγορίθμους παραγωγής συστάσεων.

### 3.5 Αλγόριθμοι παραγωγής συστάσεων

Στην ενότητα αυτή θα παρουσιαστούν οι τεχνικές παραγωγής συστάσεων που χρησιμοποιήθηκαν κατά τη διενέργεια των πειραμάτων. Όπως αναφέρθηκε και σε προηγούμενες ενότητες, επιλέχθηκαν οι παρακάτω δύο διαδεδομένες προσεγγίσεις: 1) προσέγγιση του συνεργατικού φιλτραρίσματος και 2) προσέγγιση ανάλυσης πίνακα σε ιδιάζουσες τιμές.

Η υλοποίηση των διαδικασιών παραγωγής συστάσεων πραγματοποιήθηκε με τη βοήθεια των τεχνολογιών Scala 2.11.11 και Apache Spark 2.1.1. Η τεχνολογία Scala αποτελεί μία γλώσσα προγραμματισμού που συνδυάζει στοιχεία αντικειμενοστρέφειας αλλά και συναρτησιακού προγραμματισμού. Εξαιτίας των παραπάνω γνωρισμάτων, η Scala παρέχει ένα από τα πιο πρακτικά μοντέλα σχεδίασης και συγγραφής κώδικα για παράλληλη εκτέλεση. Η τεχνολογία Apache Spark αποτελεί το πιο διαδεδομένο εργαλείο για παράλληλη επεξεργασία μεγάλων όγκων δεδομένων με εφαρμογές στη μηχανική μάθηση, την ανάλυση γράφων και την επεξεργασία δεδομένων ροής.

Τέλος αναφέρεται πως στην παρούσα εργασία το πρόβλημα της παραγωγής συστάσεων αντιμετωπίζεται ως πρόβλημα n-καλύτερων προτάσεων (βλέπε ενότητα 2.1).

#### 3.5.1 Συνεργατικό φιλτράρισμα

Παρακάτω αναλύονται οι δύο εκδοχές του συνεργατικού φιλτραρίσματος (βασισμένο σε χρήστες και βασισμένο σε προϊόντα). Οι διαφορές στις δύο υλοποιήσεις είναι μικρές καθώς και οι δύο εκδοχές βασίζονται σε παρόμοια βήματα.

##### 3.5.1.1 Υπολογισμός πινάκων ομοιότητας

Όπως αναφέρθηκε και στην ενότητα 2.3.2.1, πρώτο βήμα στη διαδικασία του συνεργατικού φιλτραρίσματος είναι η επιλογή μιας μετρικής απόστασης. Η μετρική αυτή χρησιμοποιείται για τον υπολογισμό της απόστασης (ή ομοιότητας) ανάμεσα σε δύο διανύσματα, ανεξαρτήτως εάν τα διανύσματα αυτά αποτελούν διανύσματα χρηστών ή προϊόντων. Στη βιβλιογραφία καταγράφεται η χρήστη διάφορων μετρικών απόστασης όπως για παράδειγμα ο συντελεστής συσχέτισης του Pearson [3]. Στην παρούσα εργασία χρησιμοποιήθηκε η ομοιότητα συνημίτονου ως μετρική ομοιότητας:

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Χρησιμοποιώντας την παραπάνω μετρική μπορούμε να υπολογίσουμε την ομοιότητα ανάμεσα σε όλα τα διανύσματα χρηστών ή προϊόντων. Το Apache Spark μας δίνει τη δυνατότητα να υπολογίσουμε έναν πίνακα ομοιότητας συνημίτονου με πολύ εύκολο τρόπο, εκτελώντας απλά τη συνάρτηση «columnSimilarities» επί ενός αντικειμένου τύπου «IndexerRowMatrix» το οποίο περιλαμβάνει τα δεδομένα εισόδου [30].

### 3.5.1.2 Συναρτήσεις αξιολόγησης

Σύμφωνα με τη βιβλιογραφία [3], η πιο διαδεδομένη συνάρτηση αξιολόγησης για την πρόβλεψη μιας τιμής σύστασης ενός χρήστη  $\alpha$  για ένα άγνωστο προϊόν  $j$  είναι η παρακάτω (όπου  $sim$  η ομοιότητα ανάμεσα στους χρήστες  $\alpha$  και  $k$  ενώ  $rat$  η αξιολόγηση του χρήστη  $k$  για το προϊόν  $j$ ):

$$pr_{aj} = \frac{\sum_{k \in \text{γειτονιά του } \alpha} sim_{ak} * rat_{kj}}{\sum_{k \in \text{γειτονιά του } \alpha} |sim_{ak}|}$$

Στην περίπτωση του συνεργατικού φιλτραρίσματος βασισμένο σε προϊόντα, η συνάρτηση αναδιαμορφώνεται ως εξής (όπου  $sim$  η ομοιότητα ανάμεσα στα προϊόντα  $j$  και  $k$  ενώ  $rat$  η αξιολόγηση του χρήστη  $\alpha$  για το προϊόν  $k$ ):

$$pr_{aj} = \frac{\sum_{k \in \text{γειτονιά του } j} sim_{jk} * rat_{ak}}{\sum_{k \in \text{γειτονιά του } j} |sim_{jk}|}$$

Μετά από δοκιμές διαπιστώθηκε ότι οι παραπάνω συναρτήσεις δεν παρουσιάζουν ικανοποιητικά αποτελέσματα ως προς την ποιότητα των παραγόμενων συστάσεων. Υπενθυμίζεται ότι για τις ανάγκες της εργασίας απαιτείται η εκτέλεση πειραμάτων σε δύο διαφορετικά σενάρια. Στο πρώτο σενάριο αξιοποιούνται μόνο τα δεδομένα αγοράς (συνεπώς τα δεδομένα μας αποτελούνται μόνο από μονάδες) ενώ στη δεύτερη περίπτωση αξιοποιούνται όλα τα δεδομένα (με πεδίο τιμών το  $[0,1]$ ). Ο προσεκτικός αναγνώστης θα παρατηρήσει ότι στην περίπτωση που τα δεδομένα μας χαρακτηρίζονται μόνο από μονάδες οι παραπάνω συναρτήσεις θα παράγουν πάντα την τιμή 1 (αγνοούνται οι τιμές ομοιότητας). Το ίδιο πρόβλημα παρατηρήθηκε και στο [31], όπου προτείνεται η χρήση μιας διαφορετικής συνάρτησης αξιολόγησης την οποία υιοθετούμε για τις ανάγκες του πρώτου σεναρίου δοκιμής:

$$pr_{aj} = \sum_{k \in \text{γειτονιά του } \alpha} sim_{ak}$$

Στην περίπτωση αυτή, η τιμή ενός προϊόντος προκύπτει από το άθροισμα των ομοιοτήτων των χρηστών που έχουν εξετάσει το συγκεκριμένο προϊόν. Στην περίπτωση συνεργατικού φιλτραρίσματος βασισμένο σε προϊόντα, αθροίζονται οι ομοιότητες των παρόμοιων προϊόντων που έχει εξετάσει ο χρήστης.

Για τις ανάγκες του δεύτερου σεναρίου (αξιοποίηση όλων των ενεργειών των χρηστών) χρησιμοποιήθηκε η παρακάτω συνάρτηση (η οποία προσαρμόζεται αναλόγως όταν το συνεργατικό φιλτράρισμα βασίζεται στα προϊόντα):

$$pr_{aj} = \sum_{k \in \text{γειτονιά του } \alpha} sim_{ak} * rat_{kj}$$

Οι παραπάνω συναρτήσεις παρουσιάζουν καλύτερα αποτελέσματα από τις παραδοσιακές συναρτήσεις αξιολόγησης.

### 3.5.1.3 Παραγωγή συστάσεων

Στην περίπτωση εκτέλεσης συνεργατικού φιλτραρίσματος βασισμένο στους χρήστες ακολουθείται η παρακάτω διαδικασία:

1. Προσδιορισμός των πιο όμοιων χρηστών ως προς τον εξεταζόμενο χρήστη (προσδιορισμός γειτονιάς του εξεταζόμενου χρήστη).

2. Προσδιορισμός των προϊόντων που έχουν εξετάσει οι χρήστες που ανήκουν στη γειτονιά του χρήστη.
3. Απαλοιφή των προϊόντων που βρίσκονται στο παραπάνω σύνολο και έχουν εξεταστεί από τον χρήστη για τον οποίο θέλουμε να εξάγουμε συστάσεις.
4. Για κάθε προϊόν υπολογίζεται μία τιμή σύστασης σύμφωνα με την κατάλληλη συνάρτηση αξιολόγησης.
5. Ταξινόμηση προϊόντων σε φθίνουσα σειρά ως προς την τιμή σύστασης.

Ενώ στην περίπτωση συνεργατικού φιλτραρίσματος βασισμένο στα προϊόντα εκτελούνται τα παρακάτω βήματα:

1. Εύρεση των όμοιων προϊόντων με αυτά που έχει ήδη αξιολογήσει ο εξεταζόμενος χρήστης.
2. Απαλοιφή τυχόν προϊόντων που έχει εξετάσει ο χρήστης.
3. Για κάθε προϊόν, λαμβάνοντας υπόψιν μόνο τα πιο όμοια προϊόντα από αυτά που έχει αξιολογήσει ο εξεταζόμενος χρήστης και υπολογίζεται μία τιμή σύστασης σύμφωνα με την κατάλληλη συνάρτηση αξιολόγησης.
4. Ταξινόμηση προϊόντων σε φθίνουσα σειρά ως προς την τιμή σύστασης.

### 3.5.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές

Όπως αναφέρθηκε και στην ενότητα 2.3.3, υπάρχουν δύο τρόποι με τους οποίους μπορεί να αξιοποιηθεί η ανάλυση πίνακα σε ιδιάζουσες τιμές. Στην παρούσα εργασία, αξιοποιούμε τη συγκεκριμένη τεχνική με τον δεύτερο τρόπο.

Το πρώτο βήμα της διαδικασίας περιλαμβάνει την ανάλυση του πίνακα εισόδου στο γινόμενο των τριών πινάκων  $U$ ,  $\Sigma$  και  $V$ . Το Apache Spark περιλαμβάνει υλοποιημένο τον αλγόριθμο SVD, ο οποίος μπορεί να εκτελεστεί μέσω της συνάρτησης «computeSVD» επί ενός αντικειμένου τύπου «IndexerRowMatrix» το οποίο περιλαμβάνει τα δεδομένα εισόδου [30]. Η συγκεκριμένη συνάρτηση δέχεται ως όρισμα το επιθυμητό μέγεθος του πίνακα  $\Sigma$ .

Το επόμενο βήμα είναι η επιλογή των  $k$  μεγαλύτερων ιδιζουσών τιμών του πίνακα  $\Sigma$ . Ο προσδιορισμός της τιμής  $k$  αποτελεί δύσκολο εγχείρημα καθώς επηρεάζει την ποιότητα των παραγόμενων συστάσεων και συνήθως η βέλτιστη τιμή επιλέγεται μετά την εκτέλεση των πειραμάτων. Μας ενδιαφέρει η τιμή  $k$  να είναι αρκετά μεγάλη έτσι ώστε να καταγραφούν όσο το δυνατόν περισσότερες λανθάνουσες σχέσεις αλλά παράλληλα αρκετά μικρή έτσι ώστε να αποφεύγονται λάθη εξαιτίας της εξειδίκευσης και να εξασφαλισθεί χαμηλή πολυπλοκότητα εκτέλεσης [17].

Αφού προσδιορίσουμε το σύνολο των διατηρητέων τιμών του πίνακα  $\Sigma$  μηδενίζουμε τις υπόλοιπες τιμές. Επίσης από τον πίνακα  $U$  διατηρούμε μόνο τα διανύσματα των χρηστών για τους οποίους θέλουμε να εξάγουμε συστάσεις. Στη συνέχεια εκτελούμε το γινόμενο των νέων πινάκων  $U$  και  $\Sigma$  με τον παραγόμενο πίνακα  $V$ .

Τέλος για κάθε χρήστη, απομονώνεται το διάνυσμά του από τον νέο πίνακα δεδομένων και αφαιρούνται τα προϊόντα που έχει εξετάσει. Οι τιμές που απομένουν, ταξινομημένες σε φθίνουσα σειρά αποτελούν τις παραγόμενες συστάσεις του χρήστη.

### 3.6 Πειράματα

Μέσω της παρούσας εργασίας (όπως αναφέρθηκε και στην ενότητα 1.2) τίθεται η παρακάτω υπόθεση προς επιβεβαίωση:

*Η αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης βελτιώνει την ποιότητα των αποτελεσμάτων που παράγουν τα συστήματα συστάσεων σε σύγκριση με την αξιοποίηση δεδομένων που αφορούν μόνο τις αγορές των χρηστών.*

Προκειμένου να επιβεβαιώσουμε την παραπάνω υπόθεση μας ενδιαφέρει η δοκιμή των αλγορίθμων που αναφέρθηκαν στην ενότητα 3.5 σε δύο διαφορετικά σενάρια. Στο πρώτο σενάριο αξιοποιούνται μόνο τα δεδομένα αγορών ενώ στο δεύτερο σενάριο χρησιμοποιούνται όλα τα δεδομένα (Εικόνα 13).

	CD1	CD2	CD3	CD4		CD1	CD2	CD3	CD4
Customer 1	1	0	1		Customer 1	1	0.15	1	
Customer 2		1	0	0	Customer 2		1	0.82	0.44
Customer 3	1	0			Customer 3	1	0.15		
Customer 4	0		0	1	Customer 4	0.82		0.44	1
Customer 5		0		1	Customer 5		0.15		1

(a) Conventional Recommender System      (b) Proposed Recommender System

Εικόνα 13: Μορφές πινάκων εισόδου, (α) μόνο αγορές (β) όλες οι ενέργειες - πηγή:[25]

#### 3.6.1 Σύνολα ελέγχου

Για τις ανάγκες της διαδικασίας αξιολόγησης απαιτείται η απομόνωση συγκεκριμένου όγκου δεδομένων και η δημιουργία συνόλου ελέγχου. Τα στοιχεία που ενσωματώνονται στο σύνολο ελέγχου αφαιρούνται από το σύνολο των αρχικών δεδομένων και δεν χρησιμοποιούνται κατά την εκτέλεση των αλγορίθμων.

Στην βιβλιογραφία παρουσιάζονται διάφοροι τρόποι για την δημιουργία συνόλων ελέγχου, ωστόσο στη παρούσα εργασία ακολουθούμε την στρατηγική που χρησιμοποιήθηκε στο [13]. Πιο συγκεκριμένα, για κάθε χρήση αφαιρείται τυχαία μία από τις αγορές που έχει πραγματοποιήσει. Σημειώνεται πως στην περίπτωση των συνόλων δεδομένων RetailRocket και Alibaba, όπου διατηρούνται οντότητες με τουλάχιστον δύο αγορές, παρουσιάζεται το ενδεχόμενο να «εξαφανιστούν» προϊόντα κατά την δημιουργία του συνόλου ελέγχου. Για την αντιμετώπιση του παραπάνω προβλήματος, εξασφαλίζεται πως σε περίπτωση επιλογής της τελευταίας αγοράς ενός προϊόντος, η αγορά αυτή αντικαθίσταται από μία άλλη τυχαία αγορά του χρήστη.

Επιπλέον, για να εξασφαλιστεί μία πιο αντικειμενική αξιολόγηση των αλγορίθμων, εφαρμόζεται η μέθοδος «10-fold cross validation» επί το σύνολο των χρηστών. Οι χρήστες χωρίζονται τυχαία σε δέκα σύνολα, εκ των οποίων το ένα χρησιμοποιείται ως σύνολο ελέγχου (σύμφωνα με τον τρόπο που αναφέρθηκε παραπάνω), ενώ τα υπόλοιπα χρησιμοποιούνται για την εκτέλεση των αλγορίθμων. Η διαδικασία επαναλαμβάνεται δέκα φορές (κάθε φορά επιλέγεται διαφορετικό σύνολο ελέγχου) και έπειτα υπολογίζεται ο μέσος όρος των αποτελεσμάτων.

	Retailrocket	Alibaba	Tmall
συνολικός αριθμός χρηστών	799	3.387	7.155
αριθμός χρηστών ανά σύνολο ελέγχου	80	339	716

Πίνακας 7: Μέγεθος συνόλου ελέγχου ανά σύνολο δεδομένων

Στον Πίνακα 7 παρουσιάζονται τα μεγέθη των συνόλων ελέγχου ανά σύνολο δεδομένων.

### 3.6.2 Παραμετροποίηση αλγορίθμων

Οι αλγόριθμοι που παρουσιάστηκαν στην ενότητα 3.5 δέχονται αριθμητικές παραμέτρους που επηρεάζουν την εκτέλεση τους και κατ' επέκταση τα αποτελέσματα που παράγουν. Η σωστή επιλογή των τιμών των παραμέτρων απαιτεί την διενέργεια πειραμάτων για την επιλογή των πιο αποδοτικών τιμών.

Οι αλγόριθμοι του συνεργατικού φιλτραρίσματος επιτρέπουν τη ρύθμιση δύο παραμέτρων. Η πρώτη παράμετρος (από εδώ και στο εξής θα αναφέρεται ως «topS») περιγράφει τον αριθμό των όμοιων χρηστών (ή προϊόντων στην περίπτωση του συνεργατικού φιλτραρίσματος βασισμένο σε προϊόντα) που λαμβάνονται υπόψιν κατά τον προσδιορισμό της «γειτονιάς» του χρήστη. Η δεύτερη παράμετρος (από εδώ και στο εξής θα αναφέρεται ως «topN») περιγράφει τον αριθμό των συστάσεων που παράγονται για κάθε χρήστη.

Στην περίπτωση της ανάλυσης πίνακα σε ιδιάζουσες τιμές, ο αλγόριθμος δέχεται επίσης δύο παραμέτρους. Πέρα από την παράμετρο που καθορίζει των αριθμό παραγόμενων συστάσεων ανά χρήστη (topN), ορίζεται και μία δεύτερη παράμετρος (από εδώ και στο εξής θα αναφέρεται ως «k») που περιγράφει τον αριθμό των ιδιάζουσών τιμών του πίνακα  $\Sigma$  που συγκρατούνται για τον υπολογισμό του νέου πίνακα  $U \times \Sigma \times V$ .

### 3.6.3 Οργάνωση πειραμάτων & αξιολόγηση

Για τις ανάγκες τις εργασίας πραγματοποιήθηκαν πειράματα λαμβάνοντας υπόψιν δύο βασικά σενάρια. Στο πρώτο σενάριο, τα δεδομένα που εισάγονται στους αλγορίθμους αφορούν μόνο τις αγορές των χρηστών ενώ στο δεύτερο εισάγονται επιπλέον δεδομένα έμμεσης ανατροφοδότησης τα οποία περιλαμβάνουν επιπλέον ενέργειες χρηστών.

Όσο αναφορά τις τεχνικές παραγωγής συστάσεων, αξιοποιήθηκαν δύο διαφορετικές προσεγγίσεις. Η μέθοδος συνεργατικού φιλτραρίσματος εκτελέστηκε και με τους δύο διαθέσιμους τρόπους (βασισμένο στους χρήστες και στα προϊόντα) και επιπλέον χρησιμοποιήθηκε η μέθοδος ανάλυσης πίνακα σε ιδιάζουσες τιμές.

Οι παραπάνω τεχνικές δοκιμάστηκαν στις παρακάτω τιμές παραμέτρων, όπως φαίνονται στον Πίνακα 8.

Τα πειράματα εκτελέστηκαν και για τα τρία σύνολα δεδομένων που παρουσιάστηκαν στην ενότητα 3.3.

παράμετρος	τιμές
topN	1 → 10 με βήμα 1
topS	1 → 10 με βήμα 1, 10 → 30 με βήμα 5, 1/100 χρηστών, 1/10 χρηστών
k	5 → 100 με βήμα 5

Πίνακας 8: Τιμές παραμέτρων για τις οποίες εκτελέστηκαν τα πειράματα

Τέλος, για την αξιολόγηση των αλγορίθμων χρησιμοποιήθηκε το μέτρο αξιολόγησης F1 όπως αναφέρθηκε στην ενότητα 2.5.2.

### 3.7 Αποτελέσματα πειραμάτων

Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα των πειραμάτων που εκτελέστηκαν. Σε όλες τις περιπτώσεις παρατηρήθηκε ότι τα βέλτιστα αποτελέσματα (βάσει της μετρικής F1) παρουσιάζονται κατά την παραγωγή μόνο μίας σύστασης ανά χρήστη ( $\text{topN} = 1$ ). Το γεγονός αυτό οφείλεται στον τρόπο με τον οποίο δημιουργήθηκαν τα σύνολα ελέγχου (ενότητα 3.6.1). Η παραγωγή δύο και άνω συστάσεων ανά χρήστη μειώνει δραματικά την ακρίβεια του μέτρου F1 εξαιτίας του μικρού συνόλου ελέγχου ανά χρήστη (μία κρυμμένη αγορά ανά χρήστη). Ωστόσο ο πειραματισμός με διάφορες τιμές της παραμέτρου  $\text{topN}$  παραμένει χρήσιμος για την σύγκριση των δύο μεθόδων παραγωγής συστάσεων.

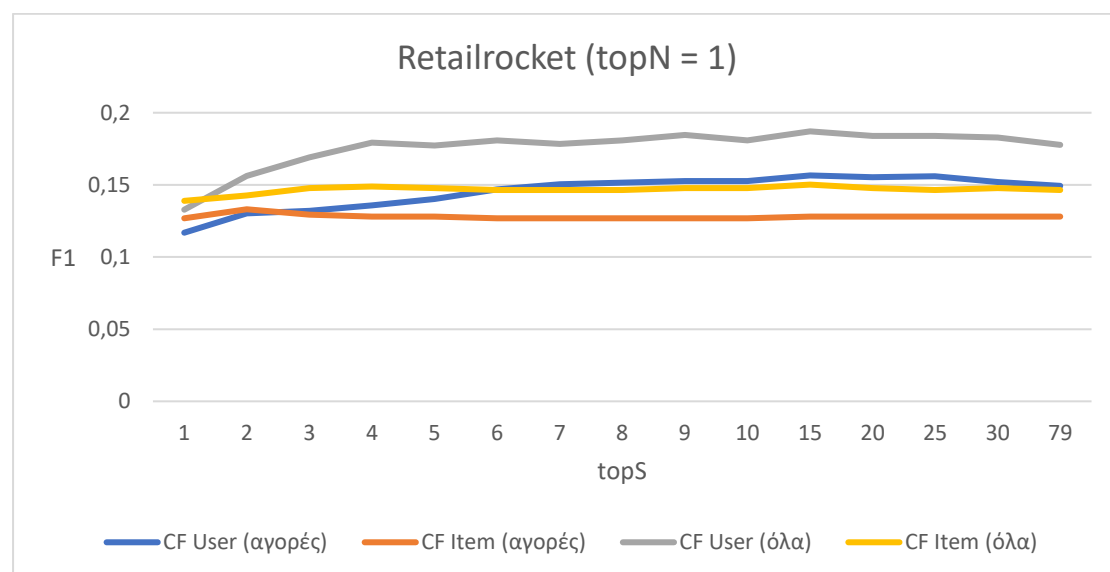
#### 3.7.1 Συνεργατικό φιλτράρισμα

Στις παρακάτω ενότητες παρουσιάζονται τα αποτελέσματα που προέκυψαν μετά την εφαρμογή του συνεργατικού φιλτραρίσματος για κάθε σύνολο δεδομένων εισόδου. Παρατηρήθηκε πως τις χειρότερες επιδόσεις είχε το σύνολο Alibaba ενώ τις καλύτερες το σύνολο Retailrocket.

##### 3.7.1.1 Σύνολο δεδομένων Retailrocket

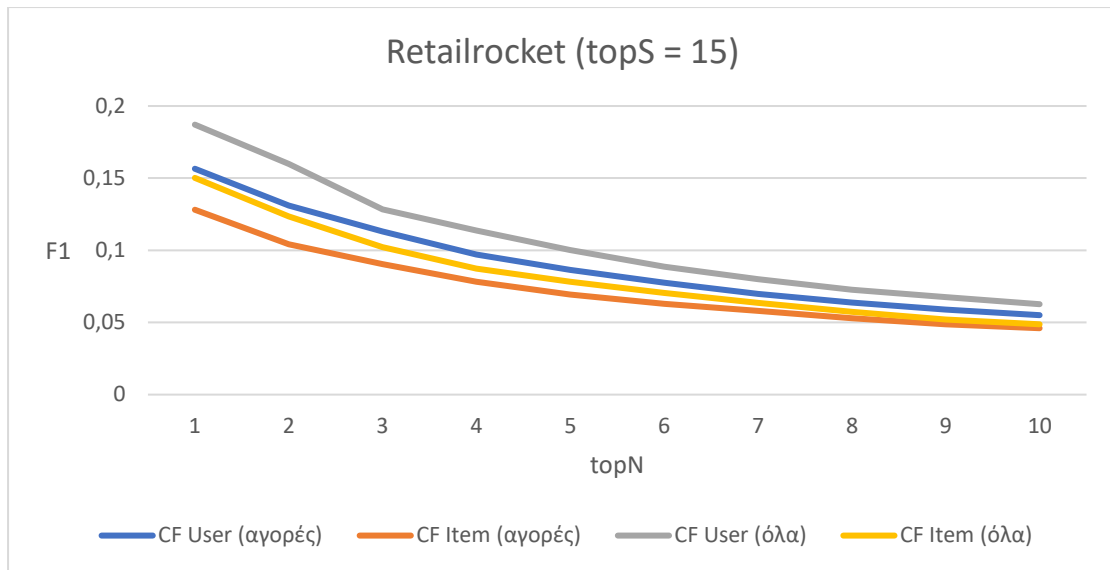
Στο Διάγραμμα 2 παρατηρούμε ότι για όλες τις τιμές της παραμέτρου  $\text{topS}$ , οι τεχνικές που αξιοποιούν όλα τα δεδομένα έμμεσης ανατροφοδότησης παρουσιάζουν καλύτερες επιδόσεις σε σχέση με αυτές που αξιοποιούν μόνο δεδομένα αγοράς.

Επίσης παρατηρείται ότι η τεχνική συνεργατικού φιλτραρίσματος βασισμένη στους χρήστες επικρατεί έναντι της τεχνικής που βασίζεται στα προϊόντα, ιδιαίτερα όταν η τιμή της παραμέτρου  $\text{topS}$  είναι μεγαλύτερη του 6.



Διάγραμμα 2: Συνεργατικό φιλτράρισμα, Retailrocket, σχέση  $\text{topS}$  και F1 με  $\text{topN} = 1$

Διατηρώντας την τιμή της παραμέτρου  $\text{topS}$  ίση με 15 (σημείο όπου οι τεχνικές που αξιοποιούν μόνο δεδομένα αγοράς παρουσιάζουν τις καλύτερες επιδόσεις) παρουσιάζεται το Διάγραμμα 3, στο οποίο επιβεβαιώνονται για άλλη μια φορά οι παραπάνω διαπιστώσεις. Η πτωτική τάση των αποτελεσμάτων σε σχέση με την παράμετρο  $\text{topN}$  οφείλεται στην εξήγηση που παρουσιάστηκε στην εισαγωγή της ενότητας 3.7.

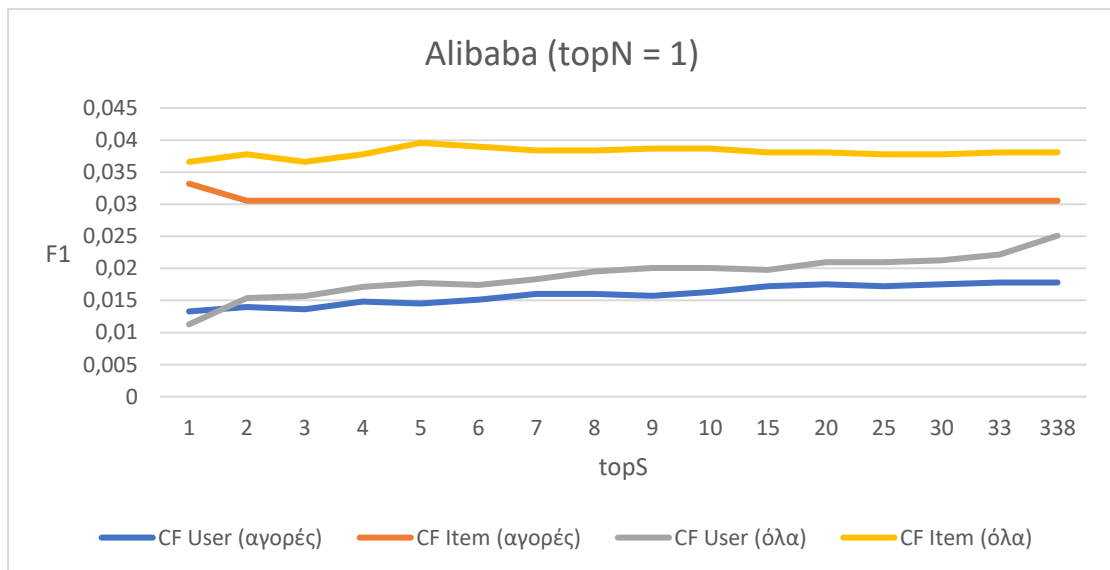


Διάγραμμα 3: Συνεργατικό φιλτράρισμα, Retailrocket, σχέση topN και F1 με topS = 15

### 3.7.1.2 Σύνολο δεδομένων Alibaba

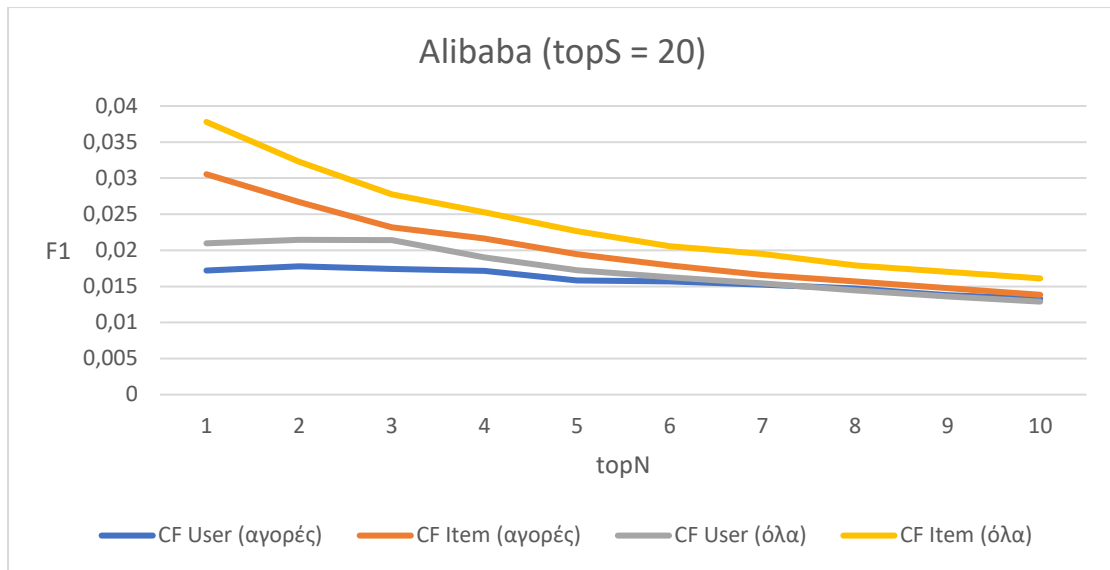
Στο Διάγραμμα 4 παρατηρούμε ότι οι τεχνικές που αξιοποιούν όλα τα δεδομένα έμμεσης ανατροφοδότησης παρουσιάζουν επίσης καλύτερες επιδόσεις σε σχέση με αυτές που αξιοποιούν μόνο δεδομένα αγοράς.

Επίσης παρατηρείται ότι οι τεχνικές συνεργατικού φιλτραρίσματος βασισμένες στα προϊόντα επικρατούν έναντι των τεχνικών που βασίζονται στους χρήστες.



Διάγραμμα 4: Συνεργατικό φιλτράρισμα, Alibaba, σχέση topS και F1 με topN = 1

Διατηρώντας την τιμή της παραμέτρου topS ίση με 20 παρουσιάζεται το Διάγραμμα 5, στο οποίο επιβεβαιώνονται οι παραπάνω διαπιστώσεις. Άξιο αναφοράς είναι το γεγονός ότι σε τιμές μεγαλύτερες του έξι της παραμέτρου topN, όλες οι προσεγγίσεις πλην αυτής που αξιοποιεί όλα τα δεδομένα και βασίζεται στα προϊόντα παρουσιάζουν όμοιες επιδόσεις.

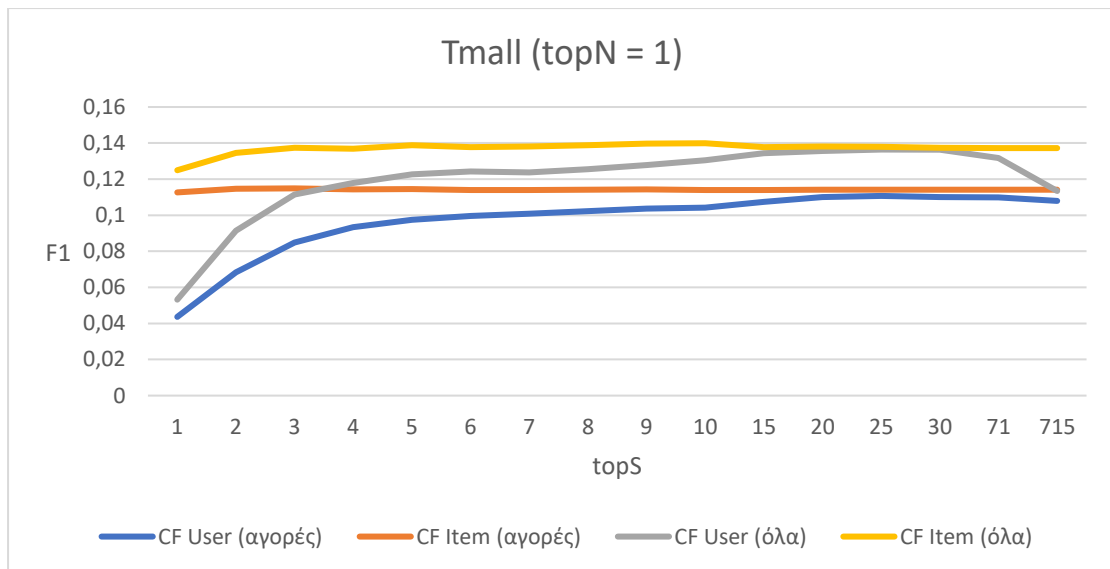


Διάγραμμα 5: Συνεργατικό φιλτράρισμα, Alibaba, σχέση topN και F1 με topS = 20

### 3.7.1.3 Σύνολο δεδομένων Tmall

Στο Διάγραμμα 6 παρατηρούμε ότι για όλες τις τιμές της παραμέτρου topS, οι τεχνικές που αξιοποιούν όλα τα δεδομένα έμμεσης ανατροφοδότησης παρουσιάζουν επίσης καλύτερες επιδόσεις σε σχέση με αυτές που αξιοποιούν μόνο δεδομένα αγοράς.

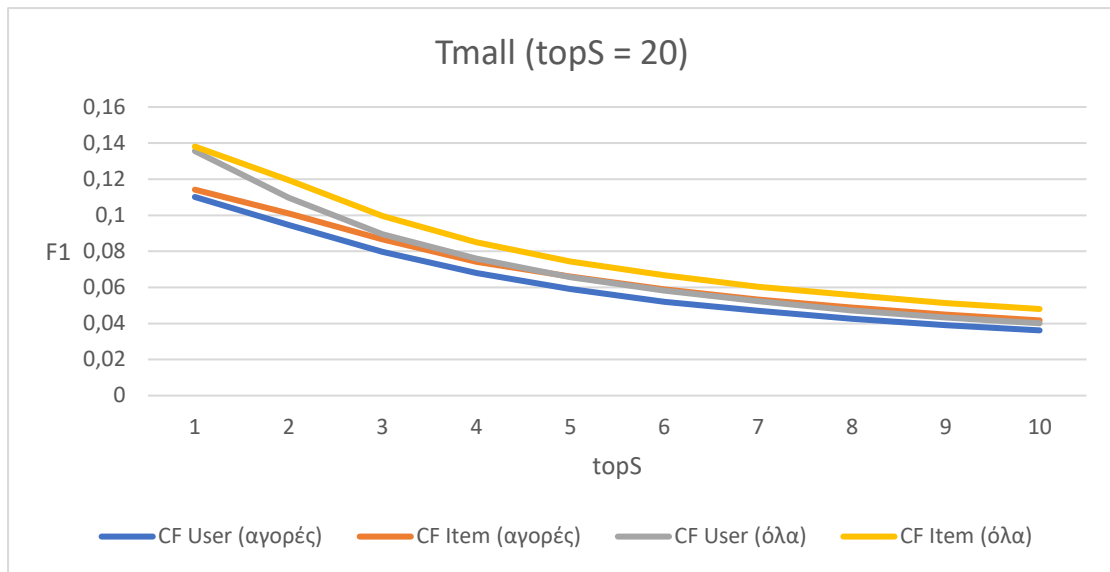
Επίσης παρατηρείται ότι οι τεχνικές συνεργατικού φιλτραρίσματος βασισμένες στα προϊόντα επικρατούν σε γενικές γραμμές έναντι των τεχνικών που βασίζονται στους χρήστες. Τα μόνα σημεία στα οποία οι προσεγγίσεις που βασίζονται στους χρήστες πλησιάζουν τις επιδόσεις των προσεγγίσεων που εστιάζουν στα προϊόντα είναι για τιμές άνω του 15 της παραμέτρου topS.



Διάγραμμα 6: Συνεργατικό φιλτράρισμα, Tmall, σχέση topS και F1 με topN = 1

Διατηρώντας την τιμή της παραμέτρου topS ίση με 20 παρουσιάζεται το Διάγραμμα 7, στο οποίο επιβεβαιώνονται οι παραπάνω διαπιστώσεις. Άξιο αναφοράς είναι το γεγονός ότι σε τιμές μεγαλύτερες του τρία της παραμέτρου topN, όλες οι προσεγγίσεις πλην αυτής που αξιοποιεί όλα τα δεδομένα και βασίζεται στα προϊόντα παρουσιάζουν όμοιες επιδόσεις.





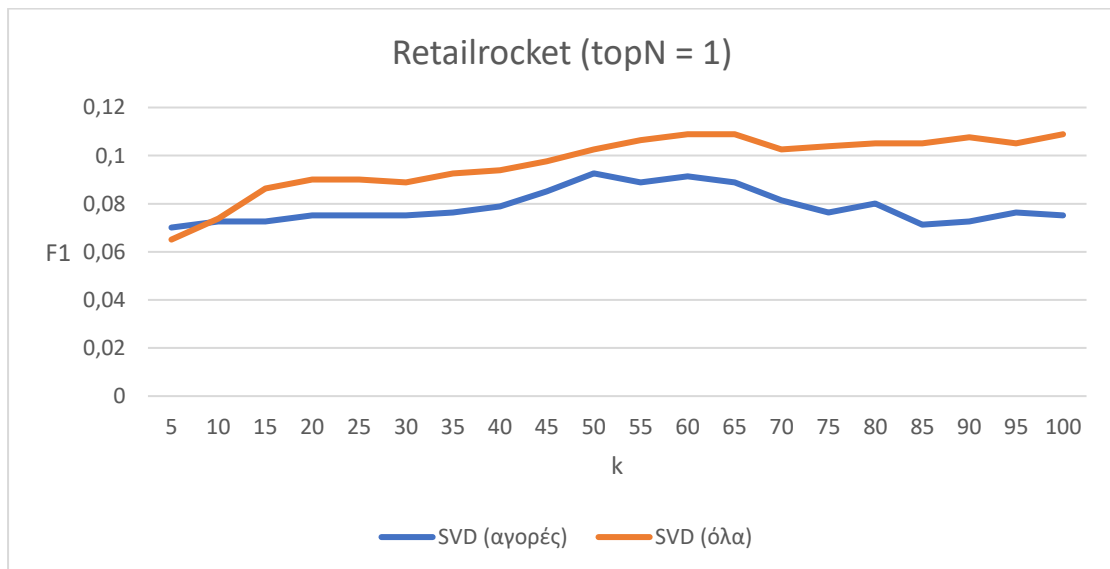
Διάγραμμα 7: Συνεργατικό φιλτράρισμα, Tmall, σχέση topN και F1 με topS = 20

### 3.7.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές

Στις παρακάτω ενότητες παρουσιάζονται τα αποτελέσματα που προέκυψαν μετά την ανάλυση πίνακα σε ιδιάζουσες τιμές για κάθε σύνολο δεδομένων εισόδου. Και εδώ παρατηρήθηκε πως τις χειρότερες επιδόσεις είχε το σύνολο Alibaba.

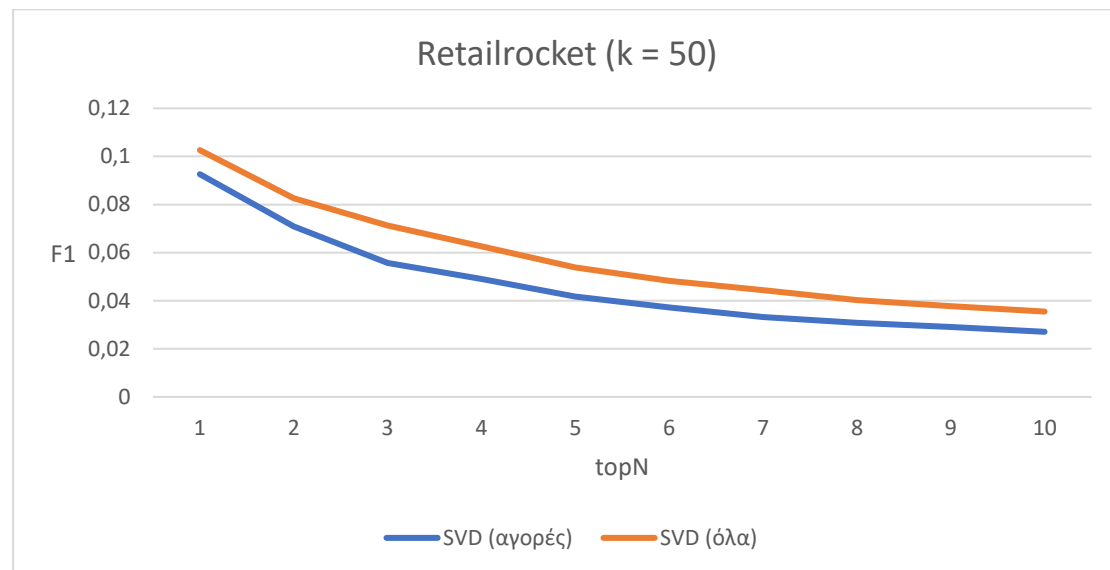
#### 3.7.2.1 Σύνολο δεδομένων Retailrocket

Στο Διάγραμμα 8 παρατηρείται ότι η αξιοποίηση των επιπλέον δεδομένων έμμεσης ανατροφοδότησης ενισχύσει σημαντικά την επίδοση του αλγορίθμου ανάλυσης πίνακα σε ιδιάζουσες τιμές. Τα μόνα σημεία όπου παρουσιάζονται παρόμοιες επιδόσεις είναι στις τιμές 10 και 50 της παραμέτρου  $k$ .



Διάγραμμα 8: Ανάλυση σε ιδιάζουσες τιμές, Retailrocket, σχέση  $k$  και F1 με topN = 1

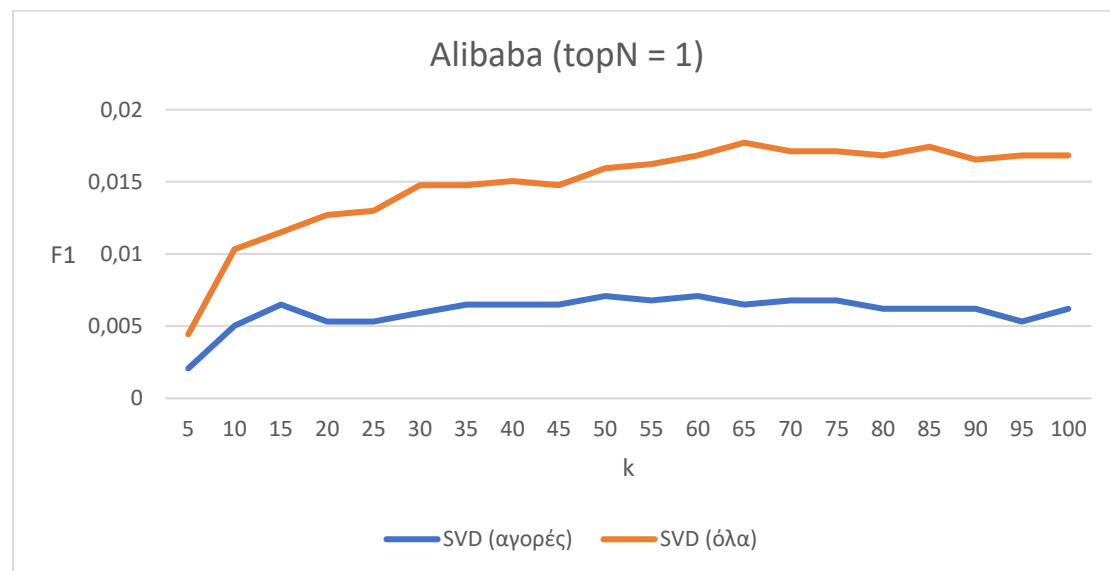
Διατηρώντας την τιμή της παραμέτρου  $k$  στην τιμή 50 παρουσιάζεται το Διάγραμμα 9, στο οποίο επιβεβαιώνεται ότι η αξιοποίηση όλων των ενεργειών έμμεσης ανατροφοδότησης ενισχύει τις επιδόσεις της συγκεκριμένης προσέγγισης.



Διάγραμμα 9: Ανάλυση σε ιδιάζουσες τιμές, Retailrocket, σχέση topN και F1 με  $k = 50$

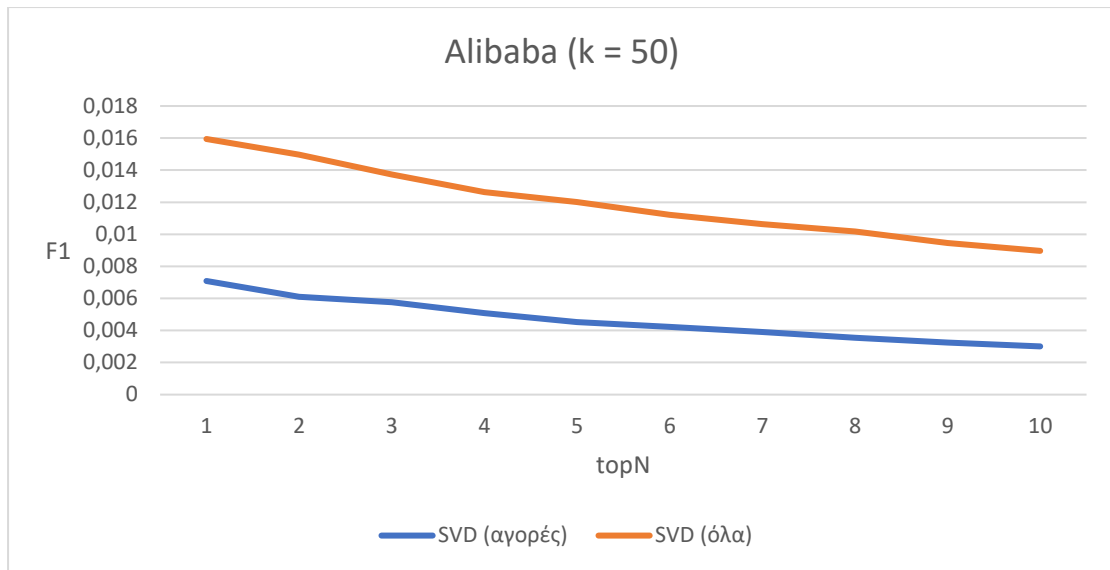
### 3.7.2.2 Σύνολο δεδομένων Alibaba

Στο Διάγραμμα 10 παρατηρείται επίσης ότι η αξιοποίηση των επιπλέον δεδομένων έμμεσης ανατροφοδότησης ενισχύει σημαντικά την επίδοση του αλγορίθμου ανάλυσης πίνακα σε ιδιάζουσες τιμές. Οι διαφορές στο συγκεκριμένο σύνολο δεδομένων είναι σημαντικές.



Διάγραμμα 10: Ανάλυση σε ιδιάζουσες τιμές, Alibaba, σχέση  $k$  και F1 με topN = 1

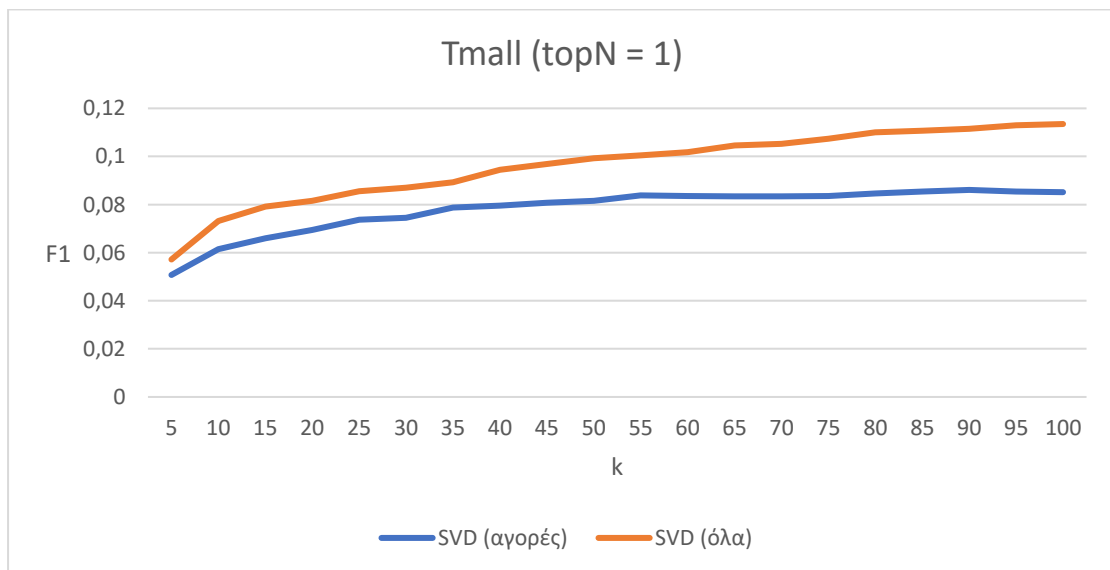
Διατηρώντας την τιμή της παραμέτρου  $k$  στην τιμή 50 παρουσιάζεται το Διάγραμμα 11, στο οποίο επιβεβαιώνεται ότι η αξιοποίηση όλων των ενεργειών έμμεσης ανατροφοδότησης ενισχύει τις επιδόσεις του συγκεκριμένης προσέγγισης.



Διάγραμμα 11: Ανάλυση σε ιδιάζουσες τιμές, Alibaba, σχέση topN και F1 με  $k = 50$

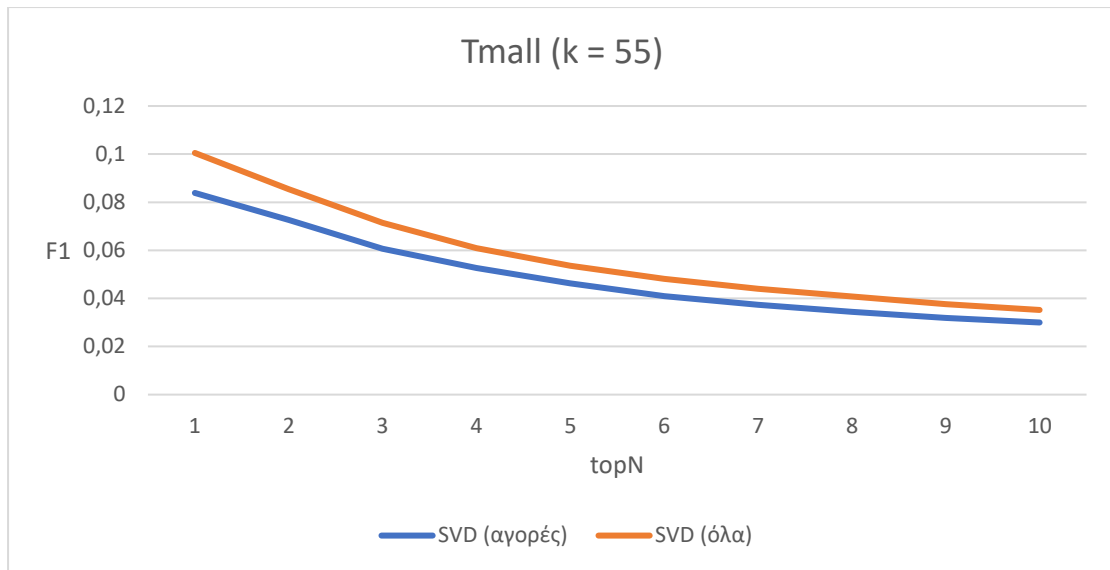
### 3.7.2.3 Σύνολο δεδομένων Tmall

Στο Διάγραμμα 12 παρατηρείται και εδώ ότι η αξιοποίηση των επιπλέον δεδομένων έμμεσης ανατροφοδότησης ενισχύσει σημαντικά την επίδοση του αλγορίθμου ανάλυσης πίνακα σε ιδιάζουσες τιμές.



Διάγραμμα 12: Ανάλυση σε ιδιάζουσες τιμές, Tmall, σχέση  $k$  και F1 με  $\text{topN} = 1$

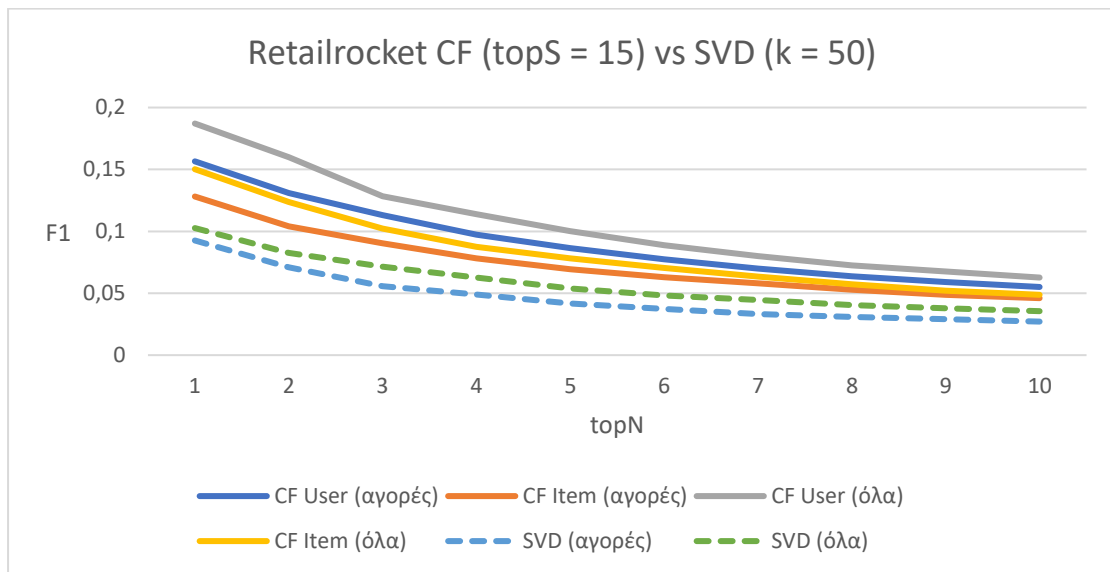
Διατηρώντας την τιμή της παραμέτρου  $k$  στην τιμή 55 παρουσιάζεται το Διάγραμμα 13, στο οποίο επιβεβαιώνεται ότι η αξιοποίηση όλων των ενεργειών έμμεσης ανατροφοδότησης ενισχύει τις επιδόσεις του συγκεκριμένης προσέγγισης.



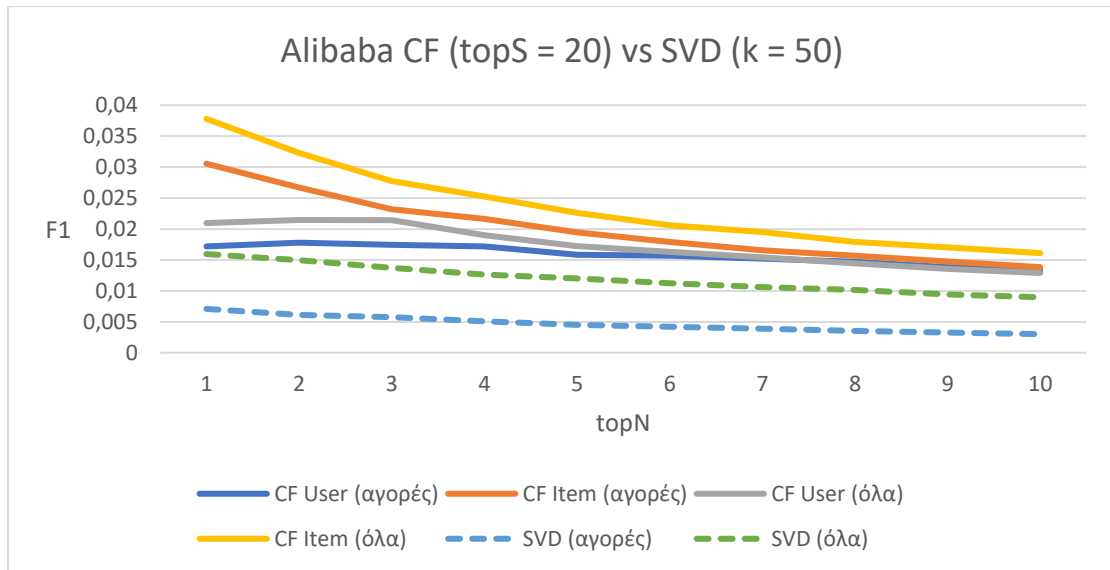
Διάγραμμα 13: Ανάλυση σε ιδιάζουσες τιμές,  $T_{mall}$ , σχέση  $topN$  και  $F1$  με  $k = 55$

### 3.7.3 Σύγκριση των δύο μεθόδων

Στην ενότητα αυτή συγκρίνουμε τις δύο τεχνικές παραγωγής συστάσεων που χρησιμοποιήθηκαν κατά την εκτέλεση των πειραμάτων. Σε όλα τα σύνολα δεδομένων παρατηρήθηκε πως οι τεχνικές συνεργατικού φιλτραρίσματος επικρατούν έναντι της τεχνικής ανάλυσης πίνακα σε ιδιάζουσες τιμές.

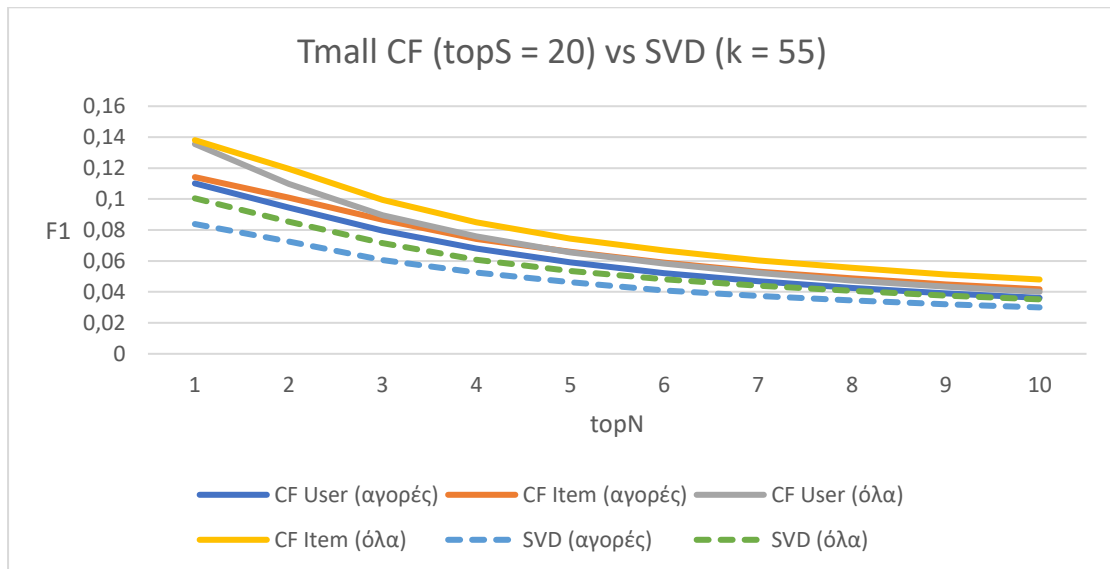


Διάγραμμα 14: Σύγκριση τεχνικών, Retailrocket, σχέση  $topN$  και  $F1$  με  $topS = 15$ ,  $k = 50$



Διάγραμμα 15: Σύγκριση τεχνικών, Alibaba, σχέση topN και F1 με topS = 20, k = 50

Η μόνη περίπτωση στην οποία η τεχνική ανάλυσης πίνακα σε ιδιάζουσες τιμές παραμένει ανταγωνιστική είναι κατά την χρήση του συνόλου δεδομένων Tmall (Διάγραμμα 16).



Διάγραμμα 16: Σύγκριση τεχνικών, Tmall, σχέση topN και F1 με topS = 20, k = 55



## 4 Δοκιμαστική ιστοσελίδα

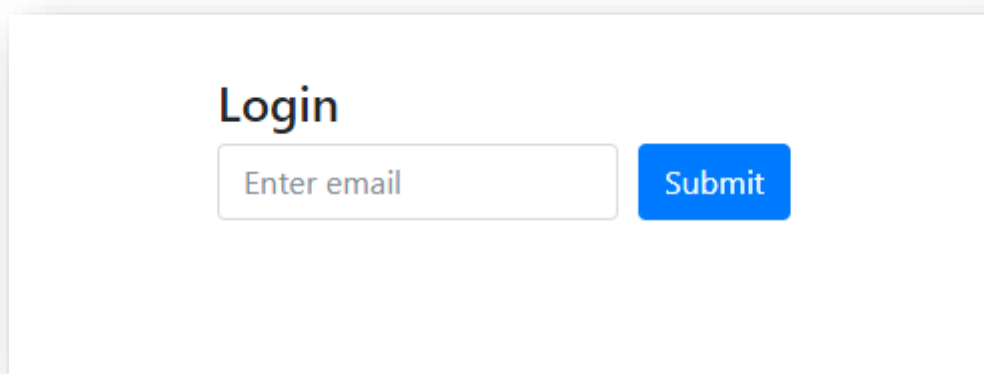
Στο κεφάλαιο αυτό παρουσιάζεται η υλοποίηση ενός δοκιμαστικού ηλεκτρονικού καταστήματος. Η ιστοσελίδα που κατασκευάστηκε μπορεί να χρησιμοποιηθεί μελλοντικά για την συγκέντρωση πειραματικών δεδομένων τα οποία μπορούν να εισαχθούν στις αλγοριθμικές διαδικασίες που αναπτύχθηκαν στο κεφάλαιο 3. Επιπλέον, η παρούσα ιστοσελίδα εξυπηρετεί ως ακράδαντη απόδειξη του γεγονότος ότι τα ηλεκτρονικά καταστήματα αποθηκεύουν εκ φύσεως δεδομένα τα οποία μπορούν να αξιοποιηθούν για την παραγωγή συστάσεων.

### 4.1 Τεχνολογίες υλοποίησης

Για την κατασκευή της ιστοσελίδας χρησιμοποιήθηκε η τεχνολογία Play Framework [32]. Η τεχνολογία Play Framework αποτελεί ένα ελαφρύ (lightweight) πλαίσιο εργασίας το οποίο επιτρέπει γρήγορη κατασκευή ιστοσελίδων αξιοποιώντας τις γλώσσες προγραμματισμού Java ή Scala. Η υλοποίηση πραγματοποιήθηκε στην γλώσσα προγραμματισμού Scala ενώ για τις ανάγκες της αποθήκευσης δεδομένων χρησιμοποιήθηκε η βάση δεδομένων H2 [33]. Τέλος για την μορφοποίηση της ιστοσελίδας αξιοποιήθηκε η τεχνολογία Bootstrap [34].

### 4.2 Παρουσίαση διεπαφής

Στη ενότητα αυτή θα παρουσιαστεί η διεπαφή της ιστοσελίδας του δοκιμαστικού ηλεκτρονικού καταστήματος. Στην Εικόνα 14 φαίνεται η οθόνη εισόδου της εφαρμογής. Ο χρήστης εισάγει μία διεύθυνση ηλεκτρονικού ταχυδρομείου η οποία λειτουργεί ως πρωτεύον κλειδί.



Εικόνα 14: Οθόνη εισόδου, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Αφού ολοκληρωθεί η είσοδος του χρήστη στο σύστημα (Εικόνα 15), παρουσιάζονται τα διαθέσιμα προϊόντα του καταστήματος. Στα αριστερά υπάρχει ένα μενού κατηγοριών με το οποίο ο χρήστης μπορεί να περιορίσει τα εμφανιζόμενα προϊόντα ανά κατηγορία. Στο πάνω μέρος της ιστοσελίδας υπάρχει μενού επιλογών το οποίο επιτρέπει στον χρήστη την πλοήγησή του το σύνολο των αγαπημένων του προϊόντων ή στο καλάθι του. Για τους διαχειριστές του συστήματος παρέχεται και μία επιπλέον διεπιφάνεια για τον έλεγχο του συστήματος.

Οι διαθέσιμες ενέργειες ενός χρήστη παρουσιάζονται δίπλα σε κάθε προϊόν (Εικόνα 15). Δίνεται η δυνατότητα προβολής ενός προϊόντος, προσθήκης του στα αγαπημένα ή και στο καλάθι.

<a href="#">Home</a> <a href="#">My Cart</a> <a href="#">My Favourites</a> <a href="#">Admin</a> <a href="#">Logout</a>							
<b>Categories</b>		<b>Items</b>					
<a href="#">All</a>							
<a href="#">Food</a>							
<a href="#">Clothing</a>							
		<b>ID</b>	<b>Title</b>	<b>Description</b>	<b>Price</b>		
		1	Ruffles Chips	Ride the wave	6.82	<a href="#">View</a>	<a href="#">Add to favourites</a>
		2	Mars Chocolate Bar	Do you feel hungry?	6.86	<a href="#">View</a>	<a href="#">Add to favourites</a>
		3	Coca cola	Are you thirsty?	6.82	<a href="#">View</a>	<a href="#">Add to favourites</a>
		4	Fanta	Be the new PR	5.11	<a href="#">View</a>	<a href="#">Add to favourites</a>
		5	Beans	Red tomato beans	3.05	<a href="#">View</a>	<a href="#">Add to favourites</a>
		6	Potato	My life is potato	2.21	<a href="#">View</a>	<a href="#">Add to favourites</a>
		7	Belt	Black leather belt	33.0	<a href="#">View</a>	<a href="#">Add to favourites</a>
		8	Polo T-Shirt	Available in many colors	12.0	<a href="#">View</a>	<a href="#">Add to favourites</a>
		9	Classic T-Shirt	Available only in red and blue	40.0	<a href="#">View</a>	<a href="#">Add to favourites</a>
		10	Ripped Jeans	New Levis Jeans	11.0	<a href="#">View</a>	<a href="#">Add to favourites</a>
		11	High quality cargo pants	Available in many colors	11.0	<a href="#">View</a>	<a href="#">Add to favourites</a>

Εικόνα 15: Κεντρικό μενού, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Σε περίπτωση που ο χρήστης επιλέξει μία από τις κατηγορίες του μενού που βρίσκεται στα αριστερά της οθόνης, τα προϊόντα αναπροσαρμόζονται όπως στην Εικόνα 16.

<a href="#">Home</a> <a href="#">My Cart</a> <a href="#">My Favourites</a> <a href="#">Admin</a> <a href="#">Logout</a>							
<b>Categories</b>		<b>Items</b>					
<a href="#">All</a>							
<a href="#">Food</a>							
<a href="#">Snacks</a>							
<a href="#">Beverages</a>							
<a href="#">Clothing</a>							
		<b>ID</b>	<b>Title</b>	<b>Description</b>	<b>Price</b>		
		1	Ruffles Chips	Ride the wave	6.82	<a href="#">View</a>	<a href="#">Add to favourites</a>
		2	Mars Chocolate Bar	Do you feel hungry?	6.86	<a href="#">View</a>	<a href="#">Add to favourites</a>

Εικόνα 16: Κεντρικό μενού 2, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Στο μενού των αγαπημένων (Εικόνα 17) ο χρήστης μπορεί να αποκτήσει πιο εύκολη πρόσβαση στα προϊόντα που τον ενδιαφέρουν περισσότερο. Δίνεται η δυνατότητα προβολής του προϊόντος ή προσθήκης του στο καλάθι. Τέλος ο χρήστης μπορεί να διαγράψει αντικείμενα από την λίστα των αγαπημένων του.



<a href="#">Home</a>	<a href="#">My Cart</a>	<a href="#">My Favourites</a>	<a href="#">Admin</a>	<a href="#">Logout</a>
Favourite				
Item ID	Title	Price		
1	<a href="#">Ruffles Chips</a>	6.82	<a href="#">Add to cart</a>	<a href="#">Remove</a>
2	<a href="#">Mars Chocolate Bar</a>	6.86	<a href="#">Add to cart</a>	<a href="#">Remove</a>
9	<a href="#">Classic T-Shirt</a>	40.0	<a href="#">Add to cart</a>	<a href="#">Remove</a>
				<a href="#">Remove all</a>

Εικόνα 17: Οθόνη αγαπημένων, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Στην Εικόνα 18, παρουσιάζεται το ηλεκτρονικό καλάθι του χρήστη. Ο χρήστης μπορεί να εξετάσει για μία τελευταία φορά τα προϊόντα που πρόκειται να αγοράσει. Επιπλέον, του προσφέρεται η δυνατότητα να αλλάξει τις ποσότητες των προϊόντων που βρίσκονται στο καλάθι. Τέλος, το μενού παρουσιάζει το τελικό ποσό της αγοράς και παρέχει στον χρήστη τη δυνατότητα να ολοκληρώσει την αγορά του.

<a href="#">Home</a>	<a href="#">My Cart</a>	<a href="#">My Favourites</a>	<a href="#">Admin</a>	<a href="#">Logout</a>
Cart				
Item ID	Title	Amount	Price	
1	<a href="#">Ruffles Chips</a>	2	13.64	<a href="#">Increase</a> <a href="#">Decrease</a>
2	<a href="#">Mars Chocolate Bar</a>	1	6.86	<a href="#">Increase</a> <a href="#">Decrease</a>
9	<a href="#">Classic T-Shirt</a>	2	80.0	<a href="#">Increase</a> <a href="#">Decrease</a>
4	<a href="#">Fanta</a>	1	5.11	<a href="#">Increase</a> <a href="#">Decrease</a>
8	<a href="#">Polo T-Shirt</a>	1	12.0	<a href="#">Increase</a> <a href="#">Decrease</a>
Total		7	117.61	<a href="#">Issue Order</a> <a href="#">Remove all</a>

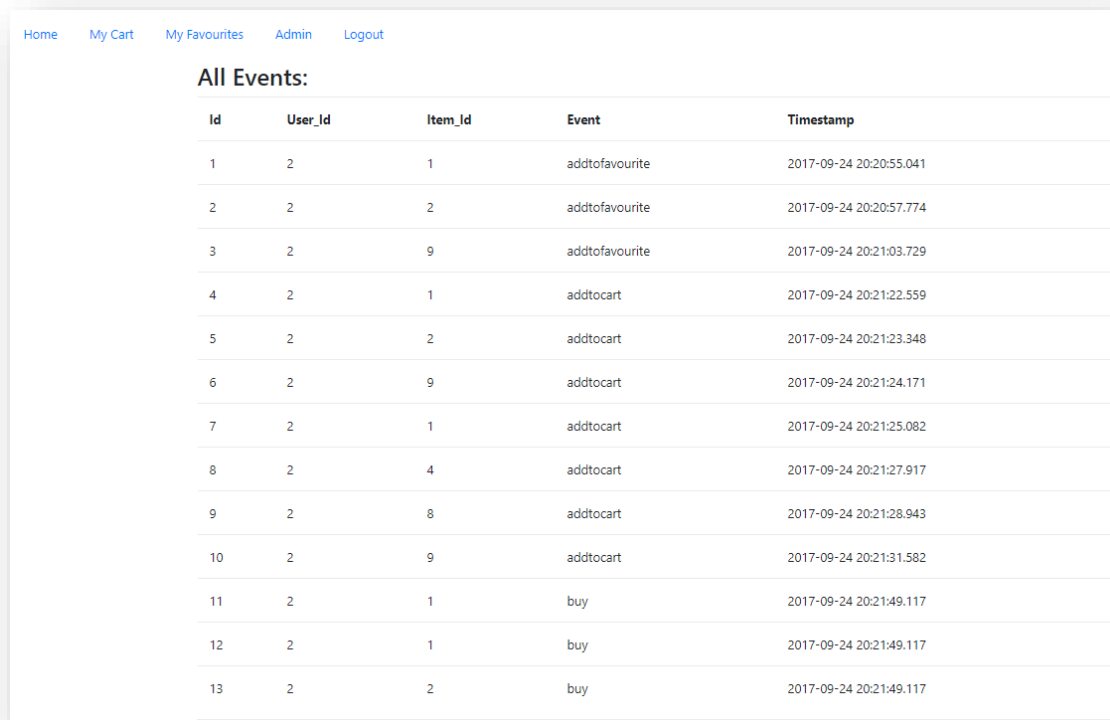
Εικόνα 18: Καλάθι αγορών, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Μετά την ολοκλήρωση μιας αγοράς, εμφανίζεται μήνυμα επιτυχίας στον χρήστη όπως ακριβώς φαίνεται στην Εικόνα 19.

<a href="#">Home</a>	<a href="#">My Cart</a>	<a href="#">My Favourites</a>	<a href="#">Admin</a>	<a href="#">Logout</a>
<div> <b>Success!</b> You have issued your order.           <a href="#">×</a> </div>				

Εικόνα 19: Επιτυχημένη αγορά, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Τέλος, οι διαχειριστές της ιστοσελίδας έχουν πρόσβαση σε ένα ειδικό μενού το οποίο συγκεντρώνει όλα τις ενέργειες που πραγματοποιήθηκαν από τους χρήστες του ηλεκτρονικού καταστήματος (Εικόνα 20).



All Events:				
Id	User_Id	Item_Id	Event	Timestamp
1	2	1	addtofavourite	2017-09-24 20:20:55.041
2	2	2	addtofavourite	2017-09-24 20:20:57.774
3	2	9	addtofavourite	2017-09-24 20:21:03.729
4	2	1	addtocart	2017-09-24 20:21:22.559
5	2	2	addtocart	2017-09-24 20:21:23.348
6	2	9	addtocart	2017-09-24 20:21:24.171
7	2	1	addtocart	2017-09-24 20:21:25.082
8	2	4	addtocart	2017-09-24 20:21:27.917
9	2	8	addtocart	2017-09-24 20:21:28.943
10	2	9	addtocart	2017-09-24 20:21:31.582
11	2	1	buy	2017-09-24 20:21:49.117
12	2	1	buy	2017-09-24 20:21:49.117
13	2	2	buy	2017-09-24 20:21:49.117

Εικόνα 20: Οθόνη διαχείρισης, δοκιμαστική ιστοσελίδα ηλεκτρονικού καταστήματος

Τα δεδομένα που συγκεντρώνονται στην παραπάνω ιστοσελίδα μπορούν να αξιοποιηθούν από τις τεχνικές που αναφέρθηκαν στο κεφάλαιο 3.



## 5 Επίλογος

Στο κεφάλαιο αυτό παρουσιάζονται τα τελικά συμπεράσματα που προέκυψαν μετά την ανάλυση των αποτελεσμάτων από τα πειράματα. Επίσης αναφέρονται μελλοντικές επεκτάσεις που μπορούν να αποτελέσουν νέους ερευνητικούς στόχους.

### 5.1 Τελικά συμπεράσματα

Με το πέρας των πειραμάτων μπορούμε να καταλήξουμε με βεβαιότητα στα παρακάτω συμπεράσματα:

- Η αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης, που αφορούν όχι μόνο τις αγορές των χρηστών ενός ηλεκτρονικού καταστήματος αλλά και τις διάφορες ενέργειες που πραγματοποιούν κατά τη διάρκεια της αγοραστικής διαδικασίας, βελτιώνει την απόδοση των συστημάτων παραγωγής συστάσεων. Το γεγονός αυτό οφείλεται στην αύξηση της πυκνότητας των δεδομένων που ενσωματώνονται στον πίνακα εισόδου. Παρατηρήθηκε ποσοστιαία βελτίωση απόδοσης μεταξύ 17% και 26% στην περίπτωση των τεχνικών συνεργατικού φιλτραρίσματος. Ενώ η απόδοση κατά την εκτέλεση της ανάλυσης πίνακα σε ιδιόζουσες τιμές βελτιώθηκε κατά 11% με 24%.
- Η πυκνότητα του πίνακα εισόδου αποτελεί βασικό παράγοντα που συμβάλλει στην απόδοση ενός συστήματος συστάσεων. Ανάμεσα στα σύνολα δεδομένων που χρησιμοποιήθηκαν, αυτά που παρουσίασαν την χαμηλότερη πυκνότητα στον πίνακα εισόδου εξασφάλισαν παράλληλα και τις χαμηλότερες επιδόσεις.
- Οι τεχνικές συνεργατικού φιλτραρίσματος παρουσίασαν καλύτερες επιδόσεις σε σύγκριση με την τεχνική ανάλυσης πίνακα σε ιδιόζουσες τιμές. Το γεγονός αυτό ίσως οφείλεται στην ύπαρξη μηδενικών τιμών στον πίνακα εισόδου.
- Στις περισσότερες περιπτώσεις, η τεχνική συνεργατικού φιλτραρίσματος βασισμένη στα προϊόντα παρουσίασε καλύτερες επιδόσεις έναντι της εναλλακτικής της, που βασίζεται στους χρήστες. Ωστόσο σε ένα από τα τρία σύνολα δεδομένων, προέκυψε το αντίθετο αποτέλεσμα.

### 5.2 Μελλοντικές επεκτάσεις

Τα παραπάνω συμπεράσματα συντελούν σε μία ενθαρρυντική εικόνα σχετικά με την επίλυση του προβλήματος της έλλειψης πυκνότητας. Το γεγονός αυτό ωστόσο δεν περιορίζει τις μελλοντικές επεκτάσεις που πρέπει να μελετηθούν προκειμένου να εξασφαλιστεί μία πιο ολοκληρωμένη άποψη επί του ζητήματος. Μελλοντικές εργασίες θα μπορούσαν να θίξουν μερικά από τα παρακάτω ζητήματα:

- Δοκιμή διαφορετικών μετρικών ομοιότητας κατά την εκτέλεση της μεθόδου του συνεργατικού φιλτραρίσματος.
- Πειραματισμός με επιπλέον στάδια προεπεξεργασίας, προκειμένου να αντιμετωπιστεί η χαμηλή απόδοση της τεχνικής ανάλυσης πίνακα σε ιδιόζουσες τιμές.
- Εκτέλεση και αξιολόγηση της προσέγγισης με διαφορετικούς αλγορίθμους παραγωγής συστάσεων.
- Αξιοποίηση επιπλέον δεδομένων έμμεσης ανατροφοδότησης που αφορούν την συχνότητα εκτέλεσης των διάφορων ενεργειών, τις κατηγορίες των προϊόντων ή άλλων πληροφοριών που καταγράφονται από τα ηλεκτρονικά καταστήματα.



## Βιβλιογραφία

- [1] J. Bukhari, 'Amazon Is Worth More Than Walmart, Costco, and Target Combined', *Fortune*. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <http://fortune.com/2017/04/05/amazon-walmart-costco-target-market-cap/>. [Ημερομηνία πρόσβασης: 19-Σεπτεμβρίου-2017].
- [2] J. B. Schafer, J. Konstan, και J. Riedl, 'Recommender Systems in e-Commerce', στο *Proceedings of the 1st ACM Conference on Electronic Commerce*, New York, NY, USA, 1999, σσ 158–166.
- [3] E. Vozalis και et al, *Analysis of Recommender Systems' Algorithms*. .
- [4] Y. S. Kim και B.-J. Yum, 'Recommender system based on click stream data using association rule mining', *Expert Syst. Appl.*, τ. 38, τχ. 10, σσ 13320–13327, Σεπτεμβρίου 2011.
- [5] Y. Hu, Y. Koren, και C. Volinsky, 'Collaborative Filtering for Implicit Feedback Datasets', στο *2008 Eighth IEEE International Conference on Data Mining*, 2008, σσ 263–272.
- [6] E. R. Núñez-Valdéz, J. M. Cueva Lovelle, O. Sanjuán Martínez, V. García-Díaz, P. Ordoñez de Pablos, και C. E. Montenegro Marín, 'Implicit feedback techniques on recommender systems applied to electronic books', *Comput. Hum. Behav.*, τ. 28, τχ. 4, σσ 1186–1193, Ιουλίου 2012.
- [7] D. Goldberg, D. Nichols, B. M. Oki, και D. Terry, 'Using Collaborative Filtering to Weave an Information Tapestry', *Commun ACM*, τ. 35, τχ. 12, σσ 61–70, Δεκεμβρίου 1992.
- [8] A. E. V. B. E. V. B. Business, 'How the Netflix Prize Was Won', *WIRED*. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://www.wired.com/2009/09/how-the-netflix-prize-was-won/>. [Ημερομηνία πρόσβασης: 19-Σεπτεμβρίου-2017].
- [9] Y. Koren, R. Bell, και C. Volinsky, 'Matrix Factorization Techniques for Recommender Systems', *Computer*, τ. 42, τχ. 8, σσ 30–37, Αυγούστου 2009.
- [10] G. Linden, B. Smith, και J. York, 'Amazon.com Recommendations: Item-to-Item Collaborative Filtering', *IEEE Internet Comput.*, τ. 7, τχ. 1, σσ 76–80, Ιανουαρίου 2003.
- [11] Z. Huang, D. Zeng, και H. Chen, 'A Comparison of Collaborative-Filtering Recommendation Algorithms for E-commerce', *IEEE Intell. Syst.*, τ. 22, τχ. 5, σσ 68–78, Σεπτεμβρίου 2007.
- [12] B. Sarwar, G. Karypis, J. Konstan, και J. Riedl, 'Item-based Collaborative Filtering Recommendation Algorithms', στο *Proceedings of the 10th International Conference on World Wide Web*, New York, NY, USA, 2001, σσ 285–295.
- [13] G. Karypis, 'Evaluation of Item-Based Top-N Recommendation Algorithms', στο *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, USA, 2001, σσ 247–254.
- [14] D. Billsus και M. J. Pazzani, 'Learning Collaborative Information Filters', στο *Proceedings of the Fifteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1998, σσ 46–54.
- [15] N. Sano, N. Machino, K. Yada, και T. Suzuki, 'Recommendation System for Grocery Store Considering Data Sparsity', *Procedia Comput. Sci.*, τ. 60, σσ 1406–1413, Ιανουαρίου 2015.
- [16] M. G. Vozalis και K. G. Margaritis, 'Using SVD and demographic data for the enhancement of generalized Collaborative Filtering', *Inf. Sci.*, τ. 177, τχ. 15, σσ 3017–3037, Αυγούστου 2007.
- [17] B. M. Sarwar, G. Karypis, J. A. Konstan, και J. T. Riedl, 'Application of Dimensionality Reduction in Recommender System – A Case Study', στο *In Acn Webkdd Workshop*, 2000.
- [18] M. G. Vozalis και K. G. Margaritis, 'Applying SVD on item-based filtering', στο *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 2005, σσ 464–469.
- [19] Y. H. Cho, J. K. Kim, και S. H. Kim, 'A personalized recommender system based on web usage mining and decision tree induction', *Expert Syst. Appl.*, τ. 23, τχ. 3, σσ 329–342, Οκτωβρίου 2002.

- [20] Y. H. Cho και J. K. Kim, 'Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce', *Expert Syst. Appl.*, τ. 26, τχ. 2, σσ 233–246, Φεβρουαρίου 2004.
- [21] T. Q. Lee, Y. Park, και Y.-T. Park, 'A time-based approach to effective recommender systems using implicit feedback', *Expert Syst. Appl.*, τ. 34, τχ. 4, σσ 3055–3062, Μαΐου 2008.
- [22] S. K. Lee, Y. H. Cho, και S. H. Kim, 'Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations', *Inf. Sci.*, τ. 180, τχ. 11, σσ 2142–2155, Ιουνίου 2010.
- [23] K. Choi, D. Yoo, G. Kim, και Y. Suh, 'A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis', *Electron. Commer. Res. Appl.*, τ. 11, τχ. 4, σσ 309–317, Ιουλίου 2012.
- [24] A. Albadvi και M. Shahbazi, 'A hybrid recommendation technique based on product category attributes', *Expert Syst. Appl.*, τ. 36, τχ. 9, σσ 11480–11488, Νοεμβρίου 2009.
- [25] Y. S. Kim, B.-J. Yum, J. Song, και S. M. Kim, 'Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites', *Expert Syst. Appl.*, τ. 28, τχ. 2, σσ 381–393, Φεβρουαρίου 2005.
- [26] 'Retailrocket recommender system dataset'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://www.kaggle.com/retailrocket/ecommerce-dataset>. [Ημερομηνία πρόσβασης: 11-Απριλίου-2017].
- [27] 'DataSet--Data Lab'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.19.MVTahG&id=4>. [Ημερομηνία πρόσβασης: 19-Σεπτεμβρίου-2017].
- [28] 'DataSet--Data Lab'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.21.MVTahG&id=5>. [Ημερομηνία πρόσβασης: 19-Σεπτεμβρίου-2017].
- [29] 'Repeat Buyers Prediction-Challenge the Baseline | Description & Data'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://tianchi.aliyun.com/getStart/information.htm?racId=231576>. [Ημερομηνία πρόσβασης: 19-Σεπτεμβρίου-2017].
- [30] 'IndexedRowMatrix'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.mllib.linalg.distributed.IndexedRowMatrix>. [Ημερομηνία πρόσβασης: 21-Σεπτεμβρίου-2017].
- [31] M. D. Ekstrand, J. T. Riedl, και J. A. Konstan, 'Collaborative Filtering Recommender Systems', *Found Trends Hum-Comput Interact*, τ. 4, τχ. 2, σσ 81–173, Φεβρουαρίου 2011.
- [32] 'Play Framework - Build Modern & Scalable Web Apps with Java and Scala'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <https://www.playframework.com/>. [Ημερομηνία πρόσβασης: 24-Σεπτεμβρίου-2017].
- [33] 'H2 Database Engine'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <http://www.h2database.com/html/main.html>. [Ημερομηνία πρόσβασης: 24-Σεπτεμβρίου-2017].
- [34] M. O. contributors Jacob Thornton, and Bootstrap, 'Bootstrap'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: <http://getbootstrap.com>. [Ημερομηνία πρόσβασης: 24-Σεπτεμβρίου-2017].