

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ



Μηχανική Μάθηση

Συστήματα παραγωγής συστάσεων
Recommender Systems

Λιαπίκος Θεόδωρος ΑΕΜ: 11

Μπαρζόκας Βασίλειος ΑΕΜ: 14

Περιεχόμενα

1. Εισαγωγή	3
2. Δεδομένα εισόδου	5
2.1 Άμεση ανατροφοδότηση (Explicit Feedback)	5
2.2 Έμμεση ανατροφοδότηση (Implicit Feedback)	6
2.3 Συσχέτιση άμεσης και έμμεσης ανατροφοδότησης	7
3. Αλγόριθμοι και τεχνικές	7
3.1 Μοντέλα βασισμένα στη μνήμη (Memory based)	7
3.1.1 Προσέγγιση βασισμένη στο περιεχόμενο (Content based approach)	7
3.1.2 Προσέγγιση συνεργατικού φιλτραρίσματος (Collaborative Filtering)	8
3.1.3 Προσέγγιση παραγοντοποίησης πινάκων (matrix factorization)	10
3.2 Αλγόριθμοι κατασκευής μοντέλου (Model based)	11
3.2.1 Κανόνες συσχέτισης & σειριακοί κανόνες	11
3.2.2 Συσταδοποίηση (Clustering)	11
3.2.3 Δίκτυα Bayes	11
3.2.4 Κατασκευή Γράφων	12
4. Προβλήματα συστημάτων παραγωγής συστάσεων	12
4.1 Χαμηλή πυκνότητα (sparsity) δεδομένων εισόδου	12
4.2 Αδυναμία κλιμάκωσης (scalability)	12
4.3 Πολυσημία (polysemy - synonymy)	13
4.4 Αδυναμία μεταβατικής συσχέτισης	13
4.5 Πρωτοεμφανιζόμενοι χρήστες και προϊόντα	13
4.6 Επικαιροποίηση δεδομένων εισόδου	14
Βιβλιογραφία	15

1. ΕΙΣΑΓΩΓΗ

Η ταχεία ανάπτυξη και εξέλιξη του παγκόσμιου ιστού έδωσε τις βάσεις για την παραγωγή και αποθήκευση τεράστιου όγκου δεδομένων. Παράλληλα δημιούργησε νέες ανάγκες όπως η δυνατότητα διαχείρισης όλης αυτής της πληροφορίας. Η κάλυψη των αναγκών αυτών απαιτήσαν την ανάπτυξη και εισαγωγή νέων τεχνολογιών όπως οι μηχανές αναζήτησης, με πλέον χαρακτηριστικό παράδειγμα την μηχανή αναζήτησης της Google.

Η ανάπτυξη του παγκόσμιου ιστού και του διαδικτύου γενικότερα δημιούργησε ευκαιρίες και σε άλλους τομείς, με πρώτο και σημαντικότερο αυτόν την παροχής αγαθών και υπηρεσιών με τη γέννηση της έννοιας του ηλεκτρονικού εμπορίου (e-Commerce). Το πλέον χαρακτηριστικό και επιτυχημένο παράδειγμα ανάπτυξης ηλεκτρονικού εμπορίου είναι η περίπτωση της Amazon, που από ένα απλό ηλεκτρονικό κατάστημα, διευρύνθηκε και εξελίχθηκε στην πιο επικερδή εταιρεία λιανικής πώλησης στις Η.Π.Α., ξεπερνώντας παραδοσιακές αλυσίδες λιανικής πώλησης όπως είναι η Walmart.

Η διαδικτυακή παρουσία και η δραστηριοποίηση στο ηλεκτρονικό εμπόριο μέσω διαδικτύου είναι πλέον βασικοί παράγοντες ανάπτυξης και επιτυχημένης κερδοφορίας μιας εταιρίας που ασχολείται με την πώληση αγαθών ή υπηρεσιών. Αυτό οφείλεται στα βασικά πλεονεκτήματα που παρέχει η διαδικτυακή σε σχέση με τη παραδοσιακή παρουσία μιας εμπορικής εταιρίας:

- κατάργηση των φυσικών περιορισμών των παραδοσιακών καταστημάτων
- μείωση λειτουργικών εξόδων
- επέκταση της παρουσίας ουσιαστικά σε όλη την υφήλιο
- άμεση δυνατότητα προβολής όλων των προϊόντων-υπηρεσιών ανεξάρτητα πλήθους

Η επιτυχία και ραγδαία ανάπτυξη του ηλεκτρονικού εμπορίου οφείλεται βέβαια όχι μόνο στα πλεονεκτήματα που προσφέρει στις εταιρίες αλλά και στους ίδιους τους καταναλωτές, οι οποίοι μέσα από την οθόνη του υπολογιστή τους ή απλά του κινητού τους έχουν πρόσβαση στις αγορές όλης της υφής. Όπως και στην περίπτωση της διαχείρισης των δεδομένων, η ταχεία ανάπτυξη του συγκεκριμένου κλάδου του ηλεκτρονικού εμπορίου δημιούργησε, πέρα από ευκαιρίες, και μεγάλες ανάγκες οι οποίες πρέπει να καλυφθούν. Η κυριότερη ανάγκη είναι η διαχείριση του τεράστιου ποσού πληροφορίας που κατ' ανάγκη σχετίζεται με το ηλεκτρονικό εμπόριο. Έτσι τα ηλεκτρονικά καταστήματα πρέπει να διαχειριστούν τις ανάγκες εκατομμυρίων, πολλές φορές, πελατών καθημερινά, ενώ οι ίδιοι οι πελάτες καλούνται να επιλέξουν μέσα από μία πλειάδα διαφορετικών προϊόντων για να καλύψουν τις ανάγκες τους. Η ανάγκη κάλυψης των παραπάνω απαιτήσεων οδήγησε στην γέννηση των συστημάτων παραγωγής συστάσεων (Recommender Systems).

Τα συστήματα παραγωγής συστάσεων αποτελούν υλοποιήσεις ειδικών αλγορίθμων που σαν στόχο έχουν να προβλέψουν ποια προϊόντα πιθανό να ικανοποιούν περισσότερο τις ανάγκες ενός συγκεκριμένου πελάτη του καταστήματος. Για να πετύχουν το σκοπό τους τα συστήματα λαμβάνουν υπόψη τους διάφορα δεδομένα που αφορούν τους πελάτες (δημογραφικά στοιχεία, ιστορικό χρήσης κλπ) καθώς και τα προϊόντα (είδος, χαρακτηριστικά, περιγραφή, εμπορική πορεία κλπ) με τελικό ζητούμενο να εξαγάγουν μια μετρική που συσχετίζει έναν συγκεκριμένο

χρήστη με ένα προϊόν ή μια ομάδα προϊόντων, η οποία και τελικά προβάλλεται και συστήνεται στο χρήστη. Με τον τρόπο αυτό το κατάστημα προσπαθεί να προσαρμόσει και να εξατομικεύσει το περιεχόμενό του στις ανάγκες κάθε χρήστη ξεχωριστά, διευκολύνοντας τον στη διαδικασία εντοπισμού του κατάλληλου αγαθού.

Η επιτυχημένη και αξιόπιστη υλοποίηση συστημάτων παραγωγής συστάσεων ενισχύει το ηλεκτρονικό εμπόριο με ποικίλους τρόπους:

- ενισχύουν τον τζίρο των καταστημάτων διευκολύνοντας τους πελάτες στις αγορές τους
- αυξάνουν το πλήθος των ενεργών πελατών προσφέροντας στους απλούς επισκέπτες του ηλεκτρονικού καταστήματος αξιόπιστες λύσεις στις ανάγκες τους
- αυξάνουν την αγοραστική κίνηση των υφιστάμενων πελατών προσφέροντάς τους λύσεις σε διάφορες παράλληλες ανάγκες που πιθανόν έχουν
- διευρύνουν το αγοραστικό προφίλ ενός πελάτη προτείνοντας προϊόντα από διαφορετικές κατηγορίες που σχετίζονται όμως με τις ανάγκες του
- ενισχύουν την αφοσίωση (loyalty) και την εμπιστοσύνη των πελατών προς το κατάστημα ενισχύοντας την πιθανότητα επιστροφής τους για μελλοντικές αγορές
- η επαναλαμβανόμενη αλληλεπίδραση ενός πελάτη με προϊόντα του καταστήματος βοηθά το σύστημα να βελτιώνεται και προσφέρει εξατομικευμένες συστάσεις

Όλα τα παραπάνω βέβαια προϋποθέτουν την ορθή και αξιόπιστη λειτουργία των συστημάτων παραγωγής συστάσεων. Στην πραγματικότητα, η πολυπαραγοντική φύση και η πολυπλοκότητα του προβλήματος οδηγεί πολλές φορές σε εσφαλμένες και ανακριβείς συστάσεις, κάτι που έχει άμεση επίδραση στην αγοραστική συμπεριφορά των χρηστών. Τα σημαντικότερα λάθη που μπορούν να προκύψουν από τα συστήματα παραγωγής συστάσεων είναι:

- εσφαλμένα αρνητικά σφάλματα (false negatives), που αφορά προϊόντα του καταστήματος που ταιριάζουν για την κάλυψη των αναγκών του πελάτη αλλά δεν συστήνονται από το σύστημα
- εσφαλμένα θετικά σφάλματα (false positives), που αφορά προϊόντα του καταστήματος που δεν ταιριάζουν για την κάλυψη των αναγκών του πελάτη και που συστήνονται από το σύστημα

Τα πιο σημαντικά σφάλματα είναι τα εσφαλμένα θετικά, τα οποία οδηγούν στη σύσταση άσχετων προς τις ανάγκες του πελάτη προϊόντων, κάτι που έχει άμεσο αρνητικό αποτέλεσμα στην εμπιστοσύνη του προς το κατάστημα.

Στην πράξη τα προβλήματα που καλούνται να λύσουν τα συστήματα παραγωγής συστάσεων κατατάσσονται σε 2 κατηγορίες:

πρόβλημα πρόβλεψης (prediction problem): αναλύοντας όλα τα στοιχεία που έχει για τον πελάτη και τα προϊόντα του καταστήματος, το σύστημα καλείται να προβλέψει κατά πόσο ο πελάτης θα ικανοποιήσει τις ανάγκες του αγοράζοντας ένα συγκεκριμένο προϊόν, με το οποίο δεν είχε κάποια προηγούμενη επαφή (αγορά, χρήση, αξιολόγηση)

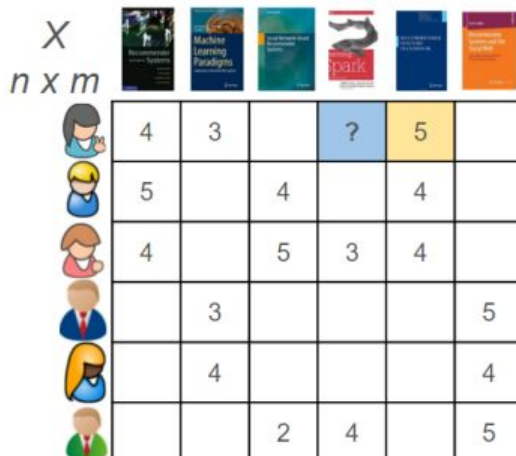
πρόβλημα n -καλύτερων συστάσεων (top- n recommendation problem): παρόμοιο πρόβλημα με το παραπάνω, αλλά το σύστημα βρίσκει, αξιολογεί, κατατάσσει και τελικά προβάλλει στο χρήστη σύνολο από n προϊόντα, με τα οποία δεν είχε κάποια προηγούμενη επαφή, που προβλέπει ότι θα καλύψει επιτυχώς τις ανάγκες του.

2. ΔΕΔΟΜΕΝΑ ΕΙΣΟΔΟΥ

Οι βασικές οντότητες που εμπλέκονται και αξιολογούνται κατά τη διαδικασία παραγωγής συστάσεων είναι ο χρήστης της ηλεκτρονικής υπηρεσίας και τα προσφερόμενα προς αυτόν προϊόντα. Τα δεδομένα που απαιτούνται από τους αλγόριθμους και τις τεχνικές που υλοποιούνται από τα συστήματα συστάσεων είναι αυτά που αφορούν τη σχέση του χρήστη με τα προϊόντα. Είναι βασικό το σύστημα να έχει ένα ικανοποιητικό αρχικό ποσό πληροφορίας, ώστε να είναι σε θέση να το αναλύσει και να μπορεί να εξάγει συστάσεις για προϊόντα άγνωστα μέχρι εκείνη τη στιγμή στο χρήστη.

Στις περισσότερες των περιπτώσεων τα δεδομένα που συλλέγονται για τη σχέση των πελατών με τα προϊόντα του καταστήματος οργανώνονται σε μεγάλους δισδιάστατους πίνακες (πίνακες αλληλεπίδρασης χρήστη-προϊόντος), όπου συνήθως οι γραμμές περιγράφουν τους χρήστες του συστήματος και οι στήλες τα προϊόντα του καταστήματος. Κάθε τιμή του πίνακα αντανακλά τη σχέση (αγορά, επίσκεψη, αξιολόγηση κλπ) ενός συγκεκριμένου χρήστη με ένα συγκεκριμένο προϊόν.

X
 $n \times m$



4	3		?	5	
5		4		4	
4		5	3	4	
	3				5
	4				4
		2	4		5

Στη συνέχεια θα αναφερθούν οι επικρατέστερες μορφές συλλογής στοιχείων που αφορούν τη σχέση των χρηστών με τα προϊόντα ενός καταστήματος.

2.1 Άμεση ανατροφοδότηση (Explicit Feedback)

Η επικρατέστερη μορφή δεδομένων που υπάρχει σήμερα και αφορά αξιολογήσεις (ratings) χρηστών επί προϊόντων που βρίσκονται ήδη στη κατοχή τους. Η αξιολόγηση γίνεται με την ανάθεση μιας αριθμητικής τιμής σε ένα προϊόν. Για την αξιολόγηση χρησιμοποιείται συνήθως:

- Κλίμακα 5 βαθμίδων Likert

- Δυαδική αξιολόγηση (αρέσκεια/δυσαρέσκεια)

Η χρήση της συγκεκριμένης μορφής δεδομένων από τους χρήστες παρουσιάζει τα παρακάτω βασικά μειονεκτήματα:

- Μικρός αριθμός αξιολογήσεων ανά χρήστη σε σχέση με το σύνολο των προϊόντων
- Sparse δεδομένα στους πίνακες αλληλεπίδρασης χρήστη-προϊόντος με μεγάλα κενά (έλλειψη στοιχείων)
- Έντονη επίδραση “θορύβου” από ευκαιριακές μη αντιπροσωπευτικές αξιολογήσεις χρηστών

Όλα αυτά οδηγούν σε χαμηλή αξιοπιστία της παρεχόμενης υπηρεσίας

2.2. Έμμεση ανατροφοδότηση (Implicit Feedback)

Διερευνά τα γενικότερα ενδιαφέροντα ενός χρήστη και αξιολογεί τη “συμπεριφορά” του όταν αλληλεπιδρά με ένα προϊόν (π.χ. η συμπεριφορά του όταν επισκέπτεται την ιστοσελίδα ενός ηλεκτρονικού καταστήματος, όπως ποιες σελίδες προϊόντων ή κατηγοριών προϊόντων επισκέφτηκε, πόσο χρόνο παρέμεινε σε μια σελίδα, προσθήκη προϊόντων στο καλάθι ή στη λίστα επιθυμιών χωρίς απαραίτητα ολοκλήρωση της αγοράς κλπ), η οποία οδηγεί έμμεσα στην εξαγωγή συμπερασμάτων για τη στάση του χρήστη απέναντι στο προϊόν. Πάντως τα παραπάνω είναι η εξαίρεση καθώς σήμερα το κύριο είδος δεδομένων που αξιοποιείται ως έμμεση ανατροφοδότηση είναι η αγορά ενός προϊόντος από το χρήστη.

Η έμμεση ανατροφοδότηση παρουσιάζει ορισμένα σαφέστατα πλεονεκτήματα σε σχέση με την άμεση:

- Δεν απαιτεί τη ρητή αξιολόγηση του χρήστη (έλλειψη χρόνου, διάθεσης κλπ)
- Δεν επιβαρύνει το χρήστη με την αγορά και λήψη του προϊόντος
- Διερευνά νέες κατηγορίες προϊόντων για τις οποίες ο χρήστης εκφράζει απλά το ενδιαφέρον του
- Μεγάλος όγκος διαθέσιμων δεδομένων τα οποία συλλέγονται αυτόματα χωρίς την παρέμβαση του χρήστη

Παράλληλα όμως η έμμεση ανατροφοδότηση συνοδεύεται και από μια σειρά αρνητικών χαρακτηριστικών που δυσχεραίνουν την αξιοποίησή της:

- Απουσία αρνητικής αξιολόγησης. Το γεγονός αυτό καθιστά αδύνατο τον καθορισμό των προϊόντων που δεν αρέσουν στο χρήστη. Απουσία αλληλεπίδρασης με ένα προϊόν μπορεί σημαίνει είτε αρνητική στάση είτε όμως και άγνοια ύπαρξης του προϊόντος.
- Απαιτεί φιλτράρισμα του δημιουργούμενου “θορύβου” ο οποίος μπορεί να παραχθεί κατά την αλληλεπίδραση του χρήστη, που αφορά όμως ανάγκες τρίτου προσώπου (αγορές δώρων κλπ) ή λάθη κατά τη διαδικτυακή πλοήγηση.

2.3 Συσχέτιση άμεσης και έμμεσης ανατροφοδότησης

Η αξιοποίηση της έμμεσης ανατροφοδότησης ως εναλλακτική λύση έναντι της άμεσης ανατροφοδότησης προϋποθέτει την ύπαρξη αποδεδειγμένης ισχυρής συσχέτισης ανάμεσα στις δύο μορφές δεδομένων. Πράγματι διαπιστώθηκε ότι:

- Ο χρόνος προβολής ενός προϊόντος από έναν χρήστη συσχετίζεται σημαντικά με τη θετική αξιολόγησή του και
- Ο αριθμός των επισκέψεων σε ένα προϊόν ή κατηγορία προϊόντος συσχετίζεται επίσης σημαντικά με το ύψος των θετικών αξιολογήσεων ενός χρήστη.

3. ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΤΕΧΝΙΚΕΣ

Υπάρχουν 2 βασικές κατηγορίες

- Μοντέλα βασισμένα στη μνήμη
- Αλγόριθμοι που παράγουν μοντέλα

Η πρώτη κατηγορία παράγει καλύτερα αποτελέσματα κατά την παραγωγή των συστάσεων αλλά χαρακτηρίζεται μεγαλύτερη πολυπλοκότητα.

3.1 Μοντέλα βασισμένα στη μνήμη (Memory based)

3.1.1 Προσέγγιση βασισμένη στο περιεχόμενο (Content based approach)

Εστιάζει στα βασικά χαρακτηριστικά και τις ιδιότητες των προϊόντων. Απαραίτητη προϋπόθεση είναι η δημιουργία ενός προφίλ για κάθε προϊόν ξεχωριστά, το οποίο περιγράφεται από ένα κοινά ορισμένο σύνολο μεταβλητών. Ο αλγόριθμος υλοποιείται σε 2 βήματα:

1. Ο διαχειριστής του συστήματος καθορίζει τις μεταβλητές (features) που χρησιμοποιούνται για τη δημιουργία του προφίλ του προϊόντος (κατηγορία, περιγραφή, μάρκα κλπ). Αποτέλεσμα είναι η αναπαράσταση κάθε προϊόντος ως διάνυσμα σε ένα πεπερασμένο χώρο διαστάσεων
2. Δημιουργία προφίλ κάθε χρήστη με βάση τα στοιχεία εισόδου. Αποτέλεσμα είναι η αναπαράσταση κάθε χρήστη με διάνυσμα σε ένα χώρο διαστάσεων ίσων με το πλήθος των προϊόντων. Προκύπτει από συνδυασμό των διανυσμάτων των προϊόντων που έχει δείξει ενδιαφέρον.

Παραγωγή συστάσεων: Με τη χρήση μετρικών αποστάσεων (π.χ. ομοιότητα συνημιτόνου) μπορούμε να εκτιμήσουμε ποια προϊόντα, άγνωστα προς το χρήστη, μοιάζουν περισσότερο στο προφίλ του χρήστη.

Τα συστήματα βασισμένα στο περιεχόμενο συνοδεύονται από σημαντικούς περιορισμούς που προβληματίζουν στη χρήση τους:

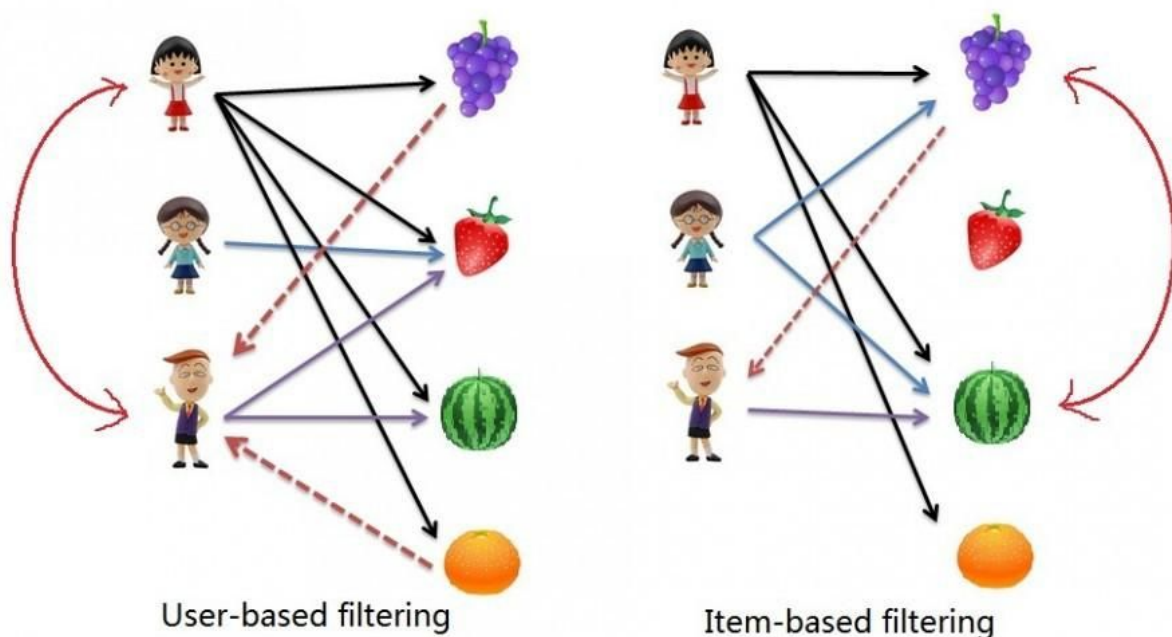
- Δυσκολία στο ορισμό των μεταβλητών (features) περιγραφής των προϊόντων καθώς απαιτείται χρόνος και ενασχόληση ατόμων με εμπειρία
- Χρονοβόρα διαδικασία συλλογής και προ επεξεργασίας των δεδομένων
- Αυξημένη εξειδίκευση συστάσεων. Το σύστημα προτείνει προϊόντα που περιορίζονται στα χαρακτηριστικά που έχει ορίσει εξαρχής ο χρήστης.
- Αδυναμία παροχής συστάσεων σε νέους χρήστες, λόγω έλλειψης δεδομένων εισόδου

Τα παραπάνω καθιστούν τα συγκεκριμένα μοντέλα δύσχρηστα και επιβάλλουν το συνδυασμό τους με άλλες τεχνικές και μεθόδους.

3.1.2 Προσέγγιση συνεργατικού φιλτραρίσματος (Collaborative Filtering)

Υπάρχουν 2 διαφορετικές προσεγγίσεις:

- από τη σκοπιά των χρηστών
- από τη σκοπιά των προϊόντων



A. Συνεργατικό φιλτράρισμα βασισμένο στους χρήστες (User-based filtering)

Αποτελεί μια από τις πιο διαδεδομένες τεχνικές που εφαρμόζονται σήμερα. Εστιάζει στην εύρεση παρόμοιων χρηστών με το χρήστη-στόχο και σύσταση σε αυτόν προϊόντων που έχουν ήδη προμηθευτεί άλλοι χρήστες της ομάδας. Ο αλγόριθμος υλοποιείται σε 2 βήματα:

1. Ο κάθε χρήστης αναπαρίσταται με διάνυσμα σε ένα χώρο διαστάσεων ίσων με το πλήθος των προϊόντων αξιοποιώντας τα δεδομένα εισόδου.
2. Δημιουργία γειτονιάς: Χρησιμοποιώντας μια μετρική απόστασης υπολογίζονται οι αποστάσεις ανάμεσα σε όλους τους χρήστες. Τελικά για κάθε χρήστη προσδιορίζεται ένα σύνολο όμοιων χρηστών (γειτονιά του χρήστη).

Παραγωγή συστάσεων: Για κάθε χρήστη υπολογίζεται μια τιμή σύστασης για κάθε προϊόν που εξετάστηκε από τους υπόλοιπους χρήστες της γειτονιάς τους. Οι συστάσεις ταξινομούνται σε φθίνουσα σειρά ενώ αφαιρούνται όσες αφορούν σε προϊόντα που έχουν ήδη εξεταστεί από το χρήστη.

Η συγκεκριμένη μέθοδος συνοδεύεται από βασικά πλεονεκτήματα που συνέβαλαν στη ευρεία διάδοσή της :

- Δεν απαιτεί τη δημιουργία ξεχωριστού προφίλ για κάθε χρήστη-προϊόν.
- Υλοποίηση σε οποιοδήποτε τύπου περιεχομένου (ταινίες, ρούχα, ηλεκτρονικά κλπ)
- Καλύτερα αποτελέσματα παραγωγής συστάσεων σε σχέση με τα προηγούμενα μοντέλα.
- Μείωση της εξειδίκευση των συστάσεων. Ο χρήστης έρχεται σε επαφή με άλλα προϊόντα που προτίμησαν χρήστες παρόμοιου προφίλ.

Παράλληλα όμως παρουσιάζει και κάποια μειονεκτήματα, όπως:

- Περιορισμένη δυνατότητα κλιμάκωσης (scalability). Αύξηση του αριθμού των χρηστών αυξάνει δραματικά την πολυπλοκότητα εκτέλεσης του υπολογισμού απόστασης χρηστών.

B. Συνεργατικό φιλτράρισμα βασισμένο στα προϊόντα (Item-based filtering)

Εναλλακτική προσέγγιση που θέλει να δώσει λύση σε όποια προβλήματα παρουσιάστηκαν προηγουμένως. Στην υλοποίησή της μοιάζει πολύ με τη προσέγγιση στους χρήστες, απλά ο υπολογισμός της απόστασης γίνεται πλέον ανάμεσα στα προϊόντα:

1. Το κάθε προϊόν αναπαρίσταται με διάνυσμα σε έναν πεπερασμένο χώρο διαστάσεων, που περιλαμβάνει τιμές αξιολόγησης από το σύνολο των χρηστών.
2. Χρησιμοποιώντας μια μετρική απόστασης υπολογίζονται οι αποστάσεις ανάμεσα σε όλα τα προϊόντα.

Παραγωγή συστάσεων: Προβάλλονται προϊόντα που έχουν υψηλές τιμές σύστασης (μεγάλη ομοιότητα) με προϊόντα που έχει ήδη εξετάσει ο χρήστης.

Η συγκεκριμένη προσέγγιση παρουσιάζει τα παρακάτω πλεονεκτήματα:

- Μείωση της πολυπλοκότητας των υπολογισμών:
 - Ο αριθμός των προϊόντων αυξάνει συνήθως πιο αργά σε σχέση με το αριθμό των χρηστών
 - Τα διανύσματα των χρηστών αλλάζουν πιο συχνά από τα διανύσματα των προϊόντων
- Είναι πιο εύκολα να ερμηνευτεί η σύσταση που έχει βασιστεί σε ομοιότητα με άλλα προϊόντα. Αυτό βοηθά σημαντικά στο να δημιουργηθεί ένα αίσθημα εμπιστοσύνης του χρήστη προς την παρεχόμενη υπηρεσία

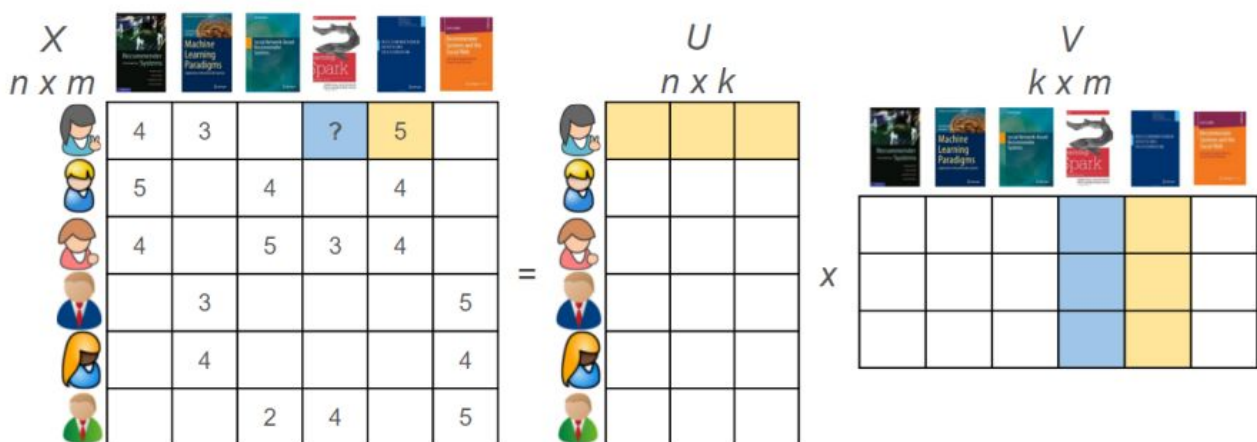
- Μείωση της εξειδίκευση των συστάσεων. Το διάνυσμα κάθε προϊόντος καθορίζεται από τις επιλογές και των υπόλοιπων χρηστών.

Παράλληλα όμως υπάρχει και μια σειρά προβλημάτων που είναι κοινά και στις 2 προσεγγίσεις:

- Υστέρηση απόδοσης κατά την παραγωγή συστάσεων όταν νέοι χρήστες ή νέα προϊόντα εισάγονται στο σύστημα.
- Για την παραγωγή των συστάσεων απαιτείται η είσοδος αρκετών δεδομένων ώστε να δημιουργηθούν οι απαραίτητες επικαλύψεις στις τιμές των προϊόντων από διάφορους χρήστες.
- Χρήση εκτεταμένων πινάκων αλληλεπίδρασης χρήστη-προϊόντος (πλήθος χρηστών \times πλήθος προϊόντων)

3.1.3 Προσέγγιση παραγοντοποίησης πινάκων (matrix factorization)

Τα τελευταία χρόνια, υπό την ενθάρρυνση διαφόρων διαγωνισμών, όπως αυτός της διαδικτυακής πλατφόρμας ταινιών Netflix, αναπτύχθηκαν νέες τεχνικές παραγωγής συστάσεων βασισμένες στην παραγοντοποίηση πινάκων, οι οποίες ανήκουν στην ευρύτερη οικογένεια αλγορίθμων συνεργατικού φιλτραρίσματος. Οι τεχνικές αυτές βασίζονται στην αποσύνθεση του εκτεταμένου πίνακα αλληλεπίδρασης χρήστη-προϊόντος M στο γινόμενο $U \times V$ 2 χαμηλότερων διαστάσεων ορθογώνιων πινάκων. Κάθε γραμμή του πρώτου πίνακα αντιστοιχεί σε έναν χρήστη ενώ κάθε στήλη του δεύτερου πίνακα αντιστοιχεί και σε ένα προϊόν και αναφέρονται γενικά ως λανθάνοντες παράγοντες (latent factors, διάσταση k στους 2 πίνακες).



Είναι δυνατό να προσαρμόσουμε την εκφραστική δύναμη του μοντέλου μεταβάλλοντας κατάλληλα το πλήθος k των λανθανόντων παραγόντων. Έτσι έχει δειχθεί ότι το αποτέλεσμα της παραγοντοποίησης του πίνακα M με τιμή $k=1$ αντιστοιχεί στην παραγωγή της πιο δημοφιλούς σύστασης (most popular recommender, το προϊόν με τις περισσότερες αλληλεπιδράσεις αλλά χωρίς κανένα στοιχείο εξειδίκευσης). Αυξάνοντας την τιμή του k αυξάνεται παράλληλα η εξειδίκευση και κατ' επέκταση η ποιότητα της παραγόμενης σύστασης. Ανεξέλεγκτη όμως

αύξηση της τιμής του k οδηγεί, πέρα από την αύξηση της πολυπλοκότητας, σε υπερμοντελοποίηση (overfit) του μοντέλου και πτώση στην ποιότητα της παραγόμενης σύστασης.

Έχουν προταθεί πολλές προσεγγίσεις για την υλοποίηση της παραγοντοποίησης των πινάκων, όπως η Funk SVD, η SVD++, η Asymmetric SVD κλπ, με την κάθε μία να προσπαθεί να επιλύσει επιμέρους προβλήματα της μεθόδου. Να τονιστεί ότι παρόλο το όνομα των προσεγγίσεων αυτών σε καμία δεν εφαρμόζεται η τεχνική SVD (Singular Value Decomposition)

3.2 Αλγόριθμοι κατασκευής μοντέλου (Model based)

3.2.1 Κανόνες συσχέτισης & σειριακοί κανόνες

A. Κανόνες συσχέτισης. Τα δεδομένα εισάγονται με τη μορφή συναλλαγών (προϊόντα που αγοράστηκαν ταυτόχρονα κατά μια επίσκεψη στο κατάστημα) και η γνώση παράγεται με τη μορφή κανόνων που στη συνέχεια αξιοποιούνται για την παραγωγή συστάσεων.

B. Σειριακοί κανόνες. Τα δεδομένα αφορούν αγορές προϊόντων κατά μια συγκεκριμένη χρονική περίοδο. Μας ενδιαφέρει η σειρά απόκτησης των προϊόντων.

Οι δύο παραπάνω προσεγγίσεις παρουσιάζουν ένα κύριο κοινό μειονέκτημα:

- Αδυναμία παραγωγής συστάσεων για προϊόντα που δεν εμφανίζονται στους κανόνες που εξάγει το σύστημα.

3.2.2 Συσταδοποίηση (Clustering)

Οι χρήστες οργανώνονται σε ένα ορισμένο πλήθος συστάδων βάσει της μεταξύ τους απόστασης (χρήση μετρικών απόστασης) ακολουθώντας μια επαναληπτική διαδικασία. Η συσταδοποίηση μπορεί στη συνέχεια να χρησιμοποιηθεί για τον υπολογισμό της γειτονιάς των χρηστών και τελικά τη σύσταση προϊόντων που εξετάστηκαν από μέλη αυτής της γειτονιάς.

Η συγκεκριμένη προσέγγιση παρουσιάζει το παρακάτω κύριο πλεονέκτημα:

- Γρήγορη εκτέλεση της διαδικασίας ταξινόμησης ενός νέου χρήστη σε μια ήδη δημιουργημένη συστάδα.

Αλλά συνοδεύεται και από μειονεκτήματα:

- Όχι τόσο καλά αποτελέσματα στην παραγωγή συστάσεων σε σχέση με άλλες προσεγγίσεις.

3.2.3 Δίκτυα Bayes

Δημιουργία ενός πιθανοτικού μοντέλου υπό τη μορφή δέντρου απόφασης (Bayes Network). Κάθε κόμβος και ακμή του δέντρου αναπαριστά ένα χαρακτηριστικό του χρήστη.

Η συγκεκριμένη προσέγγιση παρουσιάζει τα παρακάτω κύρια πλεονεκτήματα:

- Παράγονται μικρά σε μέγεθος μοντέλα.
- Γρήγορη εκτέλεση και εξαγωγή συμπερασμάτων.
- Παρουσιάζει ακρίβεια ανάλογη με τις προσεγγίσεις συνεργατικού φιλτραρίσματος.

3.2.4 Κατασκευή Γράφων

Δημιουργία μοντέλων απεικόνισης με τη μορφή γράφων όπου συνήθως οι κόμβοι αντιστοιχούν στους χρήστες του συστήματος και η ακμή αντιστοιχεί σε κάποιο χαρακτηριστικό του χρήστη (π.χ. πλήθος αλληλεπιδράσεων με προϊόντα, βαθμός ομοιότητας με άλλους χρήστες κλπ). Πάνω στους παραγόμενους γράφους μπορούν να εφαρμοστούν εξελιγμένοι αλγόριθμοι γράφων και να εξαχθούν πολύτιμα συμπεράσματα, όπως π.χ. η ανακάλυψη ισχυρών συσχετίσεων μεταξύ χρηστών με χρήση της ιδιότητας της μεταβατικής συσχέτισης. Τα δεδομένα αυτά αξιοποιούνται από τεχνικές, όπως του συνεργατικού φιλτραρίσματος για τον προσδιορισμό της γειτονιάς κάθε χρήστη, οδηγώντας στην αύξηση της ποιότητας των παραγόμενων συστάσεων.

4. ΠΡΟΒΛΗΜΑΤΑ ΣΥΣΤΗΜΑΤΩΝ ΠΑΡΑΓΩΓΗΣ ΣΥΣΤΑΣΕΩΝ

Παρά την εκτεταμένη χρήση και την επιστημονική έρευνα των τελευταίων χρόνων πάνω στα συστήματα παραγωγής συστάσεων, παραμένουν ακόμα σημαντικά προβλήματα τα οποία μένουν να αντιμετωπιστούν.

4.1 Χαμηλή πυκνότητα (sparsity) δεδομένων εισόδου

Τα σύγχρονα μεγάλα ηλεκτρονικά καταστήματα (e-shops) παρέχουν καταλόγους προϊόντων με δεκάδες χιλιάδες κωδικούς, ενώ τα επισκέπτονται καθημερινά επίσης δεκάδες χιλιάδες χρηστών. Ακόμα και οι πιο ενεργοί χρήστες μπορούν να αλληλεπιδράσουν με ένα μικρό μόνο κλάσμα των συνολικών προϊόντων και να αποκτήσουν-αξιολογήσουν ένα ακόμα μικρότερο. Επομένως εγγενές πρόβλημα των συστημάτων παραγωγής συστάσεων είναι η χρήση τεραστίων διαστάσεων πινάκων αλληλεπίδρασης χρήστη-προϊόντος με εξαιρετικά χαμηλή πυκνότητα πραγματικών στοιχείων σε αυτούς.

Το πρόβλημα αυτό επηρεάζει ιδιαίτερα τις τεχνικές συνεργατικού φιλτραρίσματος, όπου γίνεται εκτεταμένη χρήση των παραπάνω πινάκων, π.χ. στη διαδικασία υπολογισμών για τον προσδιορισμό της γειτονιάς κάθε χρήστη. Λύσεις στο παραπάνω πρόβλημα προσφέρουν μεταξύ άλλων:

- η προσεκτική προεπεξεργασία των δεδομένων εισόδου
- η αξιοποίηση της έμμεσης ανατροφοδότησης για τη σύνθεση των τιμών που λείπουν
- ο συνδυασμός πολλών διαφορετικών μεθοδολογιών (υβριδικό σύστημα συστάσεων)

4.2 Αδυναμία κλιμάκωσης (scalability)

Τα σύγχρονα συστήματα παραγωγής συστάσεων καλούνται να παράγουν συστάσεις που αφορούν τεράστια σύνολα χρηστών και προϊόντων τα οποία παράλληλα καθημερινά

αυξάνονται. Το γεγονός αυτό αυξάνει ραγδαία την πολυπλοκότητα των απαιτούμενων υπολογισμών και δημιουργεί σημαντικές επιπτώσεις στην κλιμάκωση των χρησιμοποιούμενων αλγορίθμων. Λύση στο παραπάνω πρόβλημα προσπαθούν να δώσουν:

- η αξιοποίηση της τεχνικής παραγοντοποίησης πινάκων
- η βελτιστοποίηση των χρησιμοποιούμενων αλγορίθμων ώστε:
 - να μειωθεί η πολυπλοκότητά τους
 - να διασπαστεί η αλγοριθμική διαδικασία σε επιμέρους στάδια που επιτρέπουν την αποθήκευση και επαναχρησιμοποίηση των ενδιάμεσων παραγόμενων αποτελεσμάτων

4.3 Πολυσημία (polysemy - synonymy)

Τα κλασικά συστήματα παραγωγής συστάσεων από τη φύση τους αδυνατούν να αναγνωρίσουν και να διαχωρίσουν προϊόντα που μοιάζουν πάρα πολύ, είναι ίδια φύσεως αλλά διαφέρουν σε κάποια χαρακτηριστικά όπως η ονομασία ή η μάρκα του προϊόντος. Αυτό έχει σοβαρές επιπτώσεις στην ποιότητα της παραγόμενης πρόβλεψης. Λύση στο παραπάνω πρόβλημα προσπαθεί να δώσουν:

- η προσεκτική προεπεξεργασία των δεδομένων εισόδου
- η αξιοποίηση της τεχνικής παραγοντοποίησης πινάκων

4.4 Αδυναμία μεταβατικής συσχέτισης

Στα κλασικά συστήματα παραγωγής συστάσεων εμφανίζεται αδυναμία χρήσης της μεταβατικής συσχέτισης. Όταν δηλ. ένας χρήστης συσχετίζεται ισχυρά με κάποιον άλλο χρήστη (βασιζόμενοι στις τιμές κάποιας μετρικής απόστασης) και αυτός με τη σειρά του συσχετίζεται ισχυρά με κάποιον τρίτο χρήστη, τότε ο πρώτος χρήστης δεν θα συσχετιστεί ισχυρά με τον τρίτο χρήστη αν δεν έχουν αλληλεπιδράσει με κοινά προϊόντα. Αυτό έχει σοβαρές επιπτώσεις στην ακρίβεια της παραγόμενης σύστασης, ειδικά σε συστήματα που χρησιμοποιούν τεχνικές συνεργατικού φιλτραρίσματος. Λύση στο παραπάνω πρόβλημα προσπαθεί να δώσει:

- η χρήση προσεγγίσεων που βασίζονται στην ανάλυση γράφων.

4.5 Πρωτοεμφανιζόμενοι χρήστες και προϊόντα

Όταν ένας νέος χρήστης εισέρχεται και εγγράφεται σε ένα ηλεκτρονικό κατάστημα, τότε το σύστημα παραγωγής συστάσεων του καταστήματος λογικά αδυνατεί να παράγει συστάσεις για αυτόν. Πολλά συστήματα μάλιστα απαιτούν την πραγματοποίηση ικανοποιητικού αριθμού αλληλεπιδράσεων του χρήστη με προϊόντα (επίσκεψη σελίδας, αγορά, εισαγωγή στο καλάθι, αξιολόγηση κλπ) πριν ξεκινήσουν να παράγουν αξιόπιστες συστάσεις. Το ίδιο πρόβλημα παρατηρείται και με τα προϊόντα που είτε είναι νεοεισερχόμενα στον κατάλογο είτε είναι παλαιότερα αλλά δεν υπάρχουν αλληλεπιδράσεις των χρηστών με αυτά. Και τα 2 φαινόμενα προκαλούν σημαντικές επιπτώσεις στην ποιότητα των παραγόμενων από το σύστημα προβλέψεων. Λύση στο παραπάνω πρόβλημα προσπαθεί να δώσει:

- η προτροπή στο νέο χρήστη να καταχωρήσει στοιχεία του προφίλ του
- η προσφορά κινήτρων στους νέους χρήστες να πραγματοποιήσουν αξιολογήσεις προϊόντων
- η αξιοποίηση της έμμεσης ανατροφοδότησης

4.6 Επικαιροποίηση δεδομένων εισόδου

Είναι ένα φαινόμενο που παρατηρείται σε παλαιότερους χρήστες του συστήματος, κυρίως αν αυτοί είναι ιδιαίτερα ενεργοί και παράγουν πολλές αλληλεπιδράσεις με προϊόντα. Σαν αποτέλεσμα είναι η συσσώρευση πλήθους δεδομένων για τους συγκεκριμένους χρήστες κατά τη διάρκεια ενός μεγάλου χρονικού διαστήματος χρήσης του συστήματος. Τα κλασικά συστήματα παραγωγής συστάσεων συνήθως αξιοποιούν όλα τα διαθέσιμα δεδομένα που έχουν για ένα χρήστη. Αυτό έχει σαν αποτέλεσμα το σύστημα να αδυνατεί να αντιληφθεί την εξέλιξη του χρήστη ως καταναλωτή, γεγονός που οδηγεί στην παραγωγή συστάσεων χαμηλής γενικά ποιότητας, που δεν ανταποκρίνονται στα πιο πρόσφατα ενδιαφέροντα του χρήστη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Aggarwal, Charu (2016), Recommender System The Textbook. Springer
- Athulya, Remya (2016), A Case Study on Various Recommendation Systems, International Journal of Computer Applications
- Smith, Linden, Two Decades of Recommender Systems at Amazon.com, IEEE Internet Computing
- Wikipedia (visit January 2019), https://en.wikipedia.org/wiki/Recommender_system
- Wikipedia (visit January 2019),
[https://en.wikipedia.org/wiki/Matrix_factorization_\(recommender_systems\)](https://en.wikipedia.org/wiki/Matrix_factorization_(recommender_systems))