Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Βιβλιογραφική Επισκόπηση Συστημάτων Σύστασης** **A Literature Review on Recommendation Systems**
Ονοματεπώνυμο Φοιτητή	**Μικρός Δημήτριος**
Πατρώνυμο	**Γεώργιος**
Αριθμός Μητρώου	**ΜΠΣΠ/15101**
Επιβλέπων	**Τσιχρίτζης Γεωργίος Καθήγητης**

Τριμελής Εξεταστική Επιτροπή

Τσιχριτζής Γεωργίος **Καθήγητης**	**Σωτηρόπουλος Διονύσιος** **Επίκουρος Καθηγητής**	**Σακκόπουλος** **Ευάγγελος** **Επίκουρος** **Καθηγητής**

Ημερομηνία Παράδοσης **11/2019**

Abstract

Στο Διαδίκτυο, όπου ο αριθμός των επιλογών είναι συντριπτικός, για το λόγο αυτό λοιπόν δημιουργείτε η ανάγκη φιλτραρίσματος, ιεράρχησης και αποτελεσματικής παράδοσης σχετικών πληροφοριών προκειμένου να μετριαστεί το πρόβλημα της υπερφόρτωσης των πληροφοριών, γεγονός που δημιούργησε ένα δυνητικό πρόβλημα σε πολλούς χρήστες του Διαδικτύου. Τα συστήματα συστημένων λύσεων επιλύουν αυτό το πρόβλημα αναζητώντας μεγάλο όγκο δυναμικά παραγόμενων πληροφοριών για να παρέχουν στους χρήστες εξατομικευμένο περιεχόμενο και υπηρεσίες. Σκοπός αυτης της διπλώματικης διατριβής είναι να διερευνήσει τα διαφορετικά χαρακτηριστικά και τις δυνατότητες διαφορετικών τεχνικών πρόβλεψης στα συστήματα συστάσεων για να χρησιμεύσει ως πυξίδα για έρευνα και πρακτική στον τομέα των συστημικών συστάσεων.

On the Internet, where the number of choices is overwhelming, there is need to filter, prioritize and efficiently deliver relevant information in order to alleviate the problem of information overload, which has created a potential problem to many Internet users. Recommender systems solve this problem by searching through large volume of dynamically generated information to provide users with personalized content and services. The purpose of this diploma thesis is to explore the different features and capabilities of different technical predictions in the recommendation systems to serve as a compass for research and practice in systemic recommendations.

1. Εισαγωγή

Η ραγδαία αύξηση ψηφιακής πληροφορίας καθώς και ο αριθμός των χρηστών που χρησιμοποιούν το internet έχει δημιουργήσει την ανάγκη ώστε η πληροφορία που παράγεται να κατανέμετε σωστά. Πληροφορικά συστήματα που κάνουν που τραβάνε δεδομένα, ‘όπως το Netflix, YouTube και Twitter έχουν δώσει μια λύση σε αυτό το πρόβλημα αλλά η ιεράρχηση και η εξατομίκευση της πληροφορίας απουσιάζουν. Το γεγονός αυτό λοιπόν έχει δημιουργήσει την ανάγκη των συστημάτων συστάσεων να είναι άκρως σημαντική. Τα συστήματα συστάσεων στην ουσία πρόκειται για συστήματα που φιλτράρουν πληροφορίες ώστε να μην γίνεται καταιγισμός από δεδομένα στον τελικό χρήστη στην περίπτωση που θέλει να πάρει μια απόφαση, οπού αυτή μπορεί να είναι π.χ. επιλογή μιας ταινίας, μια νυχτερινή έξοδος κτλ. Τα συστήματα συστάσεων έχουν την δυνατότητα να προβλέψουν την αρέσκεια ενός χρήστη για μια συγκεκριμένη επιλογή βασιζόμενα στο προφίλ ενός χρήστη.

Γενικά τα συστήματα προτάσεων επωφελούς τόσο για τους παρόδους υπηρεσιών καθώς και για τους χρήστες. Μειώνουν δραματικά τα κόστη συναλλαγών πληροφορίας για να βρουν τα κατάλληλες επιλογές σε 1 online μαγαζί. Επομένως, η ανάγκη χρήσης αποτελεσματικών και ακριβών τεχνικών συστάσεων σε ένα σύστημα που θα παρέχει σχετικές και αξιόπιστες συστάσεις για τους χρήστες δεν μπορεί να υπογραμμιστεί υπερβολικά.

2. Netflix

2.1. Introduction

Το Netflix είναι ένας Αμερικάνικης καταγωγής παροχέας υπηρεσιών μέσων ψυχαγωγίας με έδρα το Los Gatos της Καλιφόρνια. Η εταιρεία ιδρύθηκε το 1997 από τους Reed Hastings και Marc Randolph. Κυρία δραστηριότητα του Netflix είναι η υπηρεσία της online ροής τηλεοπτικών προγραμμάτων με συνδρομητική βάση η οποία προσφέρει στους χρηστές της ανά πάσα στιγμή την ταινία/σειρά που αναζήτα. Από τον Οκτώβριο του 2018, το Netflix έχει συνολικά 137 εκατομμύρια συνδρομητές παγκοσμίως, συμπεριλαμβανομένων 58,46 εκατομμυρίων στις Ηνωμένες Πολιτείες Διατίθεται σε όλο τον κόσμο εκτός της ηπειρωτικής Κίνας, της Συρίας, της Βόρειας Κορέας και της Κριμαίας. Το Netflix έχει επίσης γραφεία στις Κάτω Χώρες, τη Βραζιλία, την Ινδία, την Ιαπωνία και τη Νότια Κορέα. Το αρχικό επιχειρηματικό μοντέλο του Netflix περιελάμβανε πωλήσεις και ενοικίαση DVD μέσω ταχυδρομείου. Hastings έβγαλε DVD πωλήσεις περίπου ένα χρόνο μετά την ίδρυση της Netflix για να επικεντρωθεί στην επιχείρηση ενοικίασης DVD. Το 2007, το Netflix επεκτείνει τις δραστηριότητές της με την εισαγωγή μέσων συνεχούς ροής, διατηρώντας ταυτόχρονα την υπηρεσία ενοικίασης DVD και Blu-ray. Η εταιρεία επεκτάθηκε σε διεθνές επίπεδο, με ροή διαθέσιμη στον Καναδά το 2010 και τη Λατινική Αμερική και την Καραϊβική στη συνέχεια. Η Netflix εισήλθε στη βιομηχανία παραγωγής περιεχομένου το 2012, ξεκινώντας από τη πρώτη της σειρά, τη Lilyhammer. Έχει επεκτείνει σημαντικά την παραγωγή και τη διανομή ταινιών και τηλεοπτικών σειρών από το 2012 και προσφέρει μια ποικιλία περιεχομένου "Netflix Original" μέσω της ηλεκτρονικής βιβλιοθήκης του. Μέχρι τον Ιανουάριο του 2016, οι υπηρεσίες Netflix λειτουργούσαν σε περισσότερες από 190 χώρες. Το Netflix κυκλοφόρησε περίπου 126 πρωτότυπες σειρές και ταινίες το 2016, περισσότερο από οποιοδήποτε άλλο δίκτυο ή καλωδιακό κανάλι. Οι προσπάθειές τους να παράγουν νέο περιεχόμενο, να εξασφαλίσουν τα δικαιώματα για πρόσθετο περιεχόμενο και να διαφοροποιηθούν μέσω 190 χωρών οδήγησαν στην εταιρεία να συγκεντρώσει δισεκατομμύρια χρέη: 21,9 δισ. Δολάρια από τον Σεπτέμβριο του 2017, από 16,8 δισ. Δολάρια από το προηγούμενο έτος. 6.5 δισεκατομμύρια δολάρια αυτού είναι μακροπρόθεσμο χρέος, ενώ τα υπόλοιπα είναι μακροπρόθεσμες υποχρεώσεις.

2.2. Personalised Video Ranker: PVR

Στην αρχική σελίδα του Netflix υπάρχουν περίπου 40 γραμμές για κάθε χρήστη και 75 video ανά γραμμή. Τα βίντεο λοιπόν που παρουσιάζονται σε κάθε γραμμή παράγονται από 1 συγκείμενο αλγόριθμο με όνομα «Personalised Video Ranker». Όπως αναφέρει και το όνομα του αυτός ο αλγόριθμος παρουσιάζει στην αρχική σελίδα ενός χρηστή με συγκεκριμένη σειρά όλα τα video συμφώνα με τις προσωπικές προτιμήσεις ενός χρήστη. Το αποτέλεσμα της σειράς των βίντεο διαφέρει ανά χρήστη διότι κάθε χρήστης έχει διαφορετικές προτιμήσεις. Για καλύτερη λειτουργιά του PVR ο συνδυασμός εξατομικευμένων σημάτων με μια δόση απρόσωπης δημοτικότητας είναι αναγκαία Στην παρακάτω αριστερή εικόνα παρουσιάζετε η αρχική σελίδα μπορεί να παρατηρηθεί πως παρουσιάζετε η αρχική σελίδα για έναν συγκεκριμένο χρήστη ενώ στην δεξιά πρόκειται για μια συνεχή κατάταξη εξατομικευμένων βίντεο Παρακολούθηση συνεδρίας με μια σειρά συνεχούς παρακολούθησης δείτε το Amatriain and Basilico [1] για περισσότερα .

Σε γενικές γραμμές ο σκοπός του αλγορίθμου είναι να ορίσει με συγκεκριμένη σειρά ένα σετ από video που ταιριάζουν καλύτερα στις προτιμήσεις ενός χρήστη από ένα συγκεκριμένο πλαίσιο ώστε ο χρήστης όταν ανοίγει την αρχική σελίδα του προφίλ του να βρίσκει video που θα του αρέσουν περισσότερο. Προκειμένου να επιτευχτεί αυτό εφαρμόζεται η εξής συνάρτηση : $U \times V \times C \rightarrow R$ ‘οπού U: είναι ο χρήστης V:video C:context. Ένας κύριος παράγοντας για την βελτιστοποίηση μια πρότασης είναι η διασημότητα ενός βίντεο. Αυτό γίνεται επειδή αν ληφθεί υπόψη μια μέση λύση πιθανότατα στο χρήστη να αρέσει κάτι το όποιο αρέσει στην πλειοψηφία των χρηστών. Φυσικά όμως η διασημότητα είναι της προσωπικότητας ενός χρήστη έτσι λοιπόν γεννάτε το πρόβλημα στο σύστημα να δημιουργεί τις ίδιες λίστες για κάθε χρήστη. Ένας τρόπος λοιπόν για να αντιμετωπιστεί αυτό το πρόβλημα όταν υπάρχουν διαθέσιμα δεδομένα βαθμολόγησης είναι η ενσωμάτωση της προβλεπόμενης βαθμολογίας κάθε στοιχείου. Αντί λοιπόν να χρησιμοποιηθεί είτε η δημοτικότητα είτε η προβλεπόμενη βαθμολογία μπορεί να υλοποιηθεί μια Τρίτη προσέγγιση η οποία εξισορροπήσει τις αυτές πτυχές. Η προσέγγιση αυτή ονομάζεται παραδοσιακή προσέγγιση Point for Learning (LTR) και αντιμετωπίζει την κατάταξη ως ένα απλό δυαδικό πρόβλημα ταξινόμησης όπου οι μόνες είσοδοι του συστήματος είναι θετικά και αρνητικά παραδείγματα. Τυπικά μοντέλα που χρησιμοποιούνται σε αυτό το πλαίσιο περιλαμβάνουν Logistic Regression ή Gradient Boosted Decision Αυτές οι προσεγγίσεις χρησιμοποιούν μετρήσεις ανάκτησης πληροφοριών συγκεκριμένης κατάταξης για τη μέτρηση της απόδοσης ενός μοντέλου κατάταξης.

2.3. Top-N Video Ranker

Στο σύστημα προτάσεων του Netflix υπάρχει επίσης ο αλγόριθμος με το όνομα «*Top N* video ranker» ο όποιος παράγει συστάσεις στην γραμμή των κορυφαίων επιλογών της αρχικής σελίδας του Netflix. Σκοπός αυτού του αλγορίθμου είναι να βρει τις καλύτερες εξατομικευμένες συστάσεις του κατάλογου για κάθε μέλος εστιάζοντας στο κορυφαίο βίντεο της κάθε κατάταξη. Κάτι το οποίο δεν προσφέρεται ως βαθμός ελευθερίας στον αλγόριθμο μια PVR επειδή είναι φτιαγμένος έτσι ώστε να κατατάσσει αυθαίρετα υποσύνολα του καταλόγου. Συνεπώς, ο Top N ranker βελτιστοποιείται και αξιολογείται με τη χρήση μετρήσεων και αλγορίθμων που φαίνονται μόνο στην κορυφή της κατάταξης του καταλόγου που παράγει ο αλγόριθμος, παρά στην κατάταξη ολόκληρου του καταλόγου (όπως στην περίπτωση του PVR). Το Netflix εκτός από τους αλγορίθμους συστάσεων της αρχικής σελίδας χρησιμοποιεί επίσης τους ακόλουθους αλγόριθμους συστάσεων.

2.4. Random Recommendations

Το πρώτο σύστημα συστάσεων βασίζεται σε τυχαίες συστάσεις. Σε αυτό το σύστημα, πραγματοποιούνται συστάσεις σε χρήστες για τυχαίες ταινίες από το σύνολο δεδομένων ταινιών στους χρήστες του σετ δοκιμών. Με άλλα λόγια, αυτό το σύστημα δεν λαμβάνει υπόψη την ιστορική βαθμολογική συμπεριφορά του χρήστη. Σημειώστε ότι το σετ από το οποίο επιλέχθηκαν οι ταινίες, εξαιρούσε τις ήδη εμφανιζόμενες ταινίες. Αυτό έγινε για να αποφευχθεί η συνδρομή ταινιών που έχουν ήδη προβληθεί.

2.5. Item Base Collaborative Filtering

Η λογική που θα αναπτυχτεί σε αυτή την ενότητα είναι «Item Base Collaborative Filtering». Αφού δημιουργηθεί μια βασική τυχαία σύσταση μπαίνουν σε λειτουργιά πιο προχωρημένες τεχνικές για την δημιουργία περισσότερο σαφών και υπεύθυνων συστάσεων, αυτή η μέθοδος είναι ευρέως γνώστη καθώς και χρησιμοποιείτε σε πολλά συστήματα συστάσεων Η τεχνική αυτή παράγει συστάσεις με βάση τις σχέσεις που υπάρχουν μεταξύ στοιχείων που προέρχονται από έναν πίνακα αξιολογήσεων. Το πρώτο βήμα αυτής της τεχνικής είναι ο υπολογισμός του πίνακα ομοιότητας □ × □ που περιέχει όλες τις ομοιότητες μεταξύ στοιχείων. Σε αυτή τη διαδικασία, χρησιμοποιείται ένα δεδομένο μέτρο ομοιότητας, για παράδειγμα η συσχέτιση Pearson[2] και η ομοιότητα Cosine[3]. Σε αυτή την έρευνα, η ομοιότητα Cosine χρησιμοποιείται όπως προτείνεται από τους Sarwar et al. (2001). Η ομοιότητα Cosine ορίζεται από τον ακόλουθο τύπο, όπου □□ και □□ είναι δύο στοιχεία, □ και □ είναι οι διανύσματα σειράς

$$\square\square\square(\vec{\square}, \square y) = (\vec{\square} * \square y)/||\square x|||\vec{\square}||$$

Σαν δεύτερη πράξη είναι ο υπολογισμός των πραγματικών συστάσεων με βάση το μέγεθος S. Αυτό επιτυγχάνεται υπολογίζοντας το βαρύτιμο άθροισμα της αξιολόγησης ενός χρηστή για κάποια στοιχεία συμφώνα με την παρακάτω formula

$$\square\square\square = \left(\sum \square \in \square(\square) \frac{1}{\square\square\square}\right) * \sum_{\square \in \square(\square)} \square\square\square\square\square$$

Στον τύπο αυτό το Rai είναι η προβλεπόμενη βαθμολογία ενός χρήστη α για ένα στοιχειό i και Sij είναι η ομοιότητα μεταξύ στοιχείων i και j. Επιπλέον το στοιχείο J πρέπει να ανήκει σε W(i) το οποιο είναι ένα υποσύνολο του S(i) και περιέχει όλες τις γνώστες αξιολογήσεις του χρήστη α που βρίσκονται στο S(i). Προκειμένου να αποσαφηνιστεί αυτή η τεχνική υπάρχει το ακόλουθο παράδειγμα. Στον παρακάτω πίνακα παρουσιάζεται ένα παράδειγμα μήτρας ομοιότητας,, το οποίο περιέχει την ομοιότητα Cosine των 6 στοιχείων. Αν υποτεθεί ότι το κ=3 το οποίο σημαίνει ότι μονό οι 3 μεγαλύτερες καταχωρήσεις αποθηκεύονται ανά σειρά. Οι αξιολογήσεις ορισμένων στοιχείων για τον ενεργό χρήστη α είναι γνώστες όποτε στο πίνακα απεικονίζονται αυτές οι αξιολογήσεις. Όποτε ο σκοπός του συστήματος σύστασης είναι να προτείνει ένα στοιχειό το οποίο δεν έχει κάποια αξιολόγηση(στοιχείο i1,i4,i6)

S	I1	I2	I3	I4	I5	I6	Ra
I1	-	0.1	0.4	0.6	0.7	0.2	4.23
I2	0.1	-	0.2	0.1	0.3	0.4	-
I3	0.4	0.2	-	0.4	0.4	0.3	-
I4	0.6	0.1	0.4	-	0.2	0.1	4.29
I5	0.7	0.3	0.4	0.2	-	0.4	-
I6	0.2	0.4	03	0.1	0.4	-	4
Ra	?	4	3	?	5	?	

Όποτε τώρα η πρόβλεψη για τα στοιχειά □1 , □4 and □6 μπορεί να υπολογιστεί ως εξής από τον παρακάτω τύπο:

$$\square\square 1 = \frac{1}{0.4 + 0.7} * (0.4 * 3 + 0.7 * 5) = 4.27$$

$$\square\square 4 = \frac{1}{0.4 + 0.2} * (0.4 * 3 + 0.2 * 5) = 3.67$$

$$\square\square 6 = \frac{1}{0.4 + 0.3 + 0.4} * (0.4 * 4 + 0.3 * 3 + 0.4 * 5) = 4.09$$

Με βάση το παραπάνω το στοιχειό 1 θα προταθεί διότι εχει το υψηλότερη προβλεπόμενη αξιολόγηση

2.6. User Collaborative Filtering

Μια παρόμοια τεχνική με την τεχνική που αναφέραμε στην προηγούμενη ενότητα είναι η «User Collaborative Filtering» Αντί να γίνεται η εύρεση παρόμοιων ταινιών όπως είδαμε προηγουμένως αυτή η τεχνική που βασίζεται στο χρήστη θα ψάξει να βρει παρόμοιους χρήστες για να προτείνει. Το πρώτο βήμα αυτής της τεχνικής είναι να κατηγοριοποιήσει χρήστες μέσω της μεθόδου γειτνίασης και στην συνεχεία να συγκεντρωθούν οι αξιολογήσεις αυτών των χρηστών ώστε να γίνει μια πρόβλεψη. Για να βρούμε τους k πλησιέστερους γείτονες ενός δεδομένου χρήστη □, χρησιμοποιούνται ομοίως μέτρα ομοιότητας όπως ο συντελεστής συσχέτισης Pearson{2} ή η ομοιότητα Cosine{3}. Για φιλτράρισμα με βάση το χρήστη, η ομοιότητα Cosine χρησιμοποιείται ξανά, όπως περιγράφεται στην προηγούμενη ενότητα φιλτραρίσματος που βασίζεται σε στοιχεία. Ωστόσο, τα αντικείμενα αντικαθίστανται από τους χρήστες. Αυτό γίνεται τώρα.

$$\square\square\square(\square,\square) = (\square * \square)/||\square||||\square||$$

Όπου □□ και □□ είναι 2 χρήστες, □ and □ τώρα είναι 2 row vectors που αντιπροσωπεύουν τις αξιολογήσεις των 2 χρηστών. Αφού δημιουργηθεί η γειτονιά του ενεργού χρήστη,, οι κορυφαίοι χρήστες επιλέγονται και θα εκπροσωπούνται ως το σύνολο (□) του ενεργού χρήστη □□. Στη συνέχεια, μια πρόβλεψη μιας συγκεκριμένης ταινίας για ενεργό χρήστη □□ μπορεί να γίνει με τον μέσο όρο των βαθμολογιών της ίδιας ταινίας των χρηστών στο (□). Αυτό μπορεί να απεικονιστεί ως τύπος

$$\square\square\square = \frac{1}{|\square(\square)|}\sum_{\square\in\square(\square)}\square\square\square$$

όπου □□□ είναι η προβλεπόμενη βαθμολογία για τον ενεργό χρήστη □□ της ταινίας □ και □□□ είναι η προβλεπόμενη βαθμολογία για τον χρήστη □, □ ∈ □ (□) της ίδιας ταινίας □. Για να διευκρινιστεί αυτή η τεχνική, ένα απλό παράδειγμα παρουσιάζεται στον παρακάτω Πίνακα. Στον πίνακα αυτό δίδεται ένας πίνακας διαβάθμισης,, ο οποίος περιέχει βαθμολογίες από 6 χρήστες 6 ταινιών. Ας υποθέσουμε πάλι ότι □
= 3, πράγμα που σημαίνει ότι μόνο οι 3 πιο όμοιοι χρήστες βρίσκονται στο □ (□) και συνεπώς θα χρησιμοποιηθούν για τον υπολογισμό των προβλέψεων για τον ενεργό χρήστη. Σε αυτό το παράδειγμα, ο ενεργός χρήστης έχει ήδη δει κάποιες ταινίες και χρησιμοποιείται για τον προσδιορισμό παρόμοιων χρηστών. Οι 3 πιο όμοιοι χρήστες σημειώνονται με έντονους χαρακτήρες (χρήστες □2, □3 και □6).

R	I1	I2	I3	I4	I5	I6
U1	3	2	5	?	3	
U2	?	4	?	4	4	5
U3	2	4	4	?	?	5
U4	3	3	3	5	5	4
U5	4	4	5	3	4	?
U6	2	?	4	4	5	3
Ra	?	4	4	?	5	?
Ra	‘2			4		4.33

2.7. Content Based Filtering

Στις προηγούμενες ενότητες οι αλγόριθμοι που χρησιμοποιήθηκαν επικεντρώνονται στα ενδιαφέρονται ενός χρήστη. Σε αντίθεση λοιπόν με αυτού του είδους τον αλγόριθμο το κέντρο ενδιαφέροντος βασίζετε στα αντικείμενα όπως για παράδειγμα τα είδη των ταινιών. Προκειμένου να υλοποιηθεί σωστά αυτή η τεχνική χρησιμοποιηθεί υλικό από τα δεδομένα του IMDB δεδομένου ότι μόνο οι αξιολογήσεις των ταινιών παρέχονται στο αρχικό σύνολο δεδομένων. Έχοντας αυτές τις αξιολογήσεις το σύστημα μπορεί να πει με πιθανότατα το ενδιαφέρον του χρήστη. Το πρώτο βήμα του αλγορίθμου είναι να δημιουργήσει έναν vector που θα περιέχει την πρόβλεψη κάθε είδους για κάθε χρήστη. Για αυτή την τεχνική οι ταινίες που έχουν βαθμολογηθεί με 3 και πάνω θα συμπεριλήψου στην λειτουργία σύστασης. Όταν ένας χρήστης για παράδειγμα έχει βαθμολογήσει 5 ταινίες τότε μια μήτρα αξιολογήσεων ν επί μ δημιουργείται και περιέχει ν χρήστες και μ ταινίες και γεμίζει με δεδομένα που αφορούν αξιολογήσεις. Η μήτρα αυτή στην συνέχεια πολλαπλασιάζετε με τις πληροφορίες των ειδών ταινιών. Όποτε έτσι δημιουργείτε ένας νέος

πίνακας μ επί κ που περιέχει δεδομένα που αφορούν τα είδη 1 για το αν η ταινία ανήκει σε 1 είδος κ 0 αν δεν ανήκει.
Το αποτέλεσμα μεταξύ της μήτρας βαθμολόγησης και της μήτρας είναι μια νέα μήτρα ν × κ που περιέχει την προδιάθεση κάθε χρήστη προς κάθε είδος, με βάση τις 5 πιο πρόσφατες βαθμολογίες που ήταν 3 και πάνω. Στη συνέχεια, βάσει αυτής της μήτρας προδιάθεσης μπορεί να γίνει μια σύσταση για έναν χρήστη, υπολογίζοντας την απόσταση μεταξύ του διανύσματος προφίλ χρήστη (□-ου σειράς μήτρας για το χρήστη □) και του πίνακα πληροφοριών. Σε γενικές γραμμές, η απόσταση μετρά την ανομοιότητα μεταξύ των διανυσμάτων □ και □ από

$$□(□, □) = (|□ ∪ □| − |□ ∩ □|)/|□ ∪ □|$$

Ο κατάλογος που προκύπτει διανέμεται με τρόπο αύξοντα, πράγμα που σημαίνει ότι η ταινία με τη χαμηλότερη απόσταση Jaccard συνιστάται πρώτα. Ωστόσο, αν υπάρχουν δύο ή περισσότερες ταινίες με την ίδια απόσταση Jaccard, αυτές οι ταινίες θα διαταχθούν σε βαθμολογία IMDb, όπου πρώτα θα σας συστήσει την ταινία με την υψηλότερη βαθμολογία IMDb. Η λίστα που προκύπτει είναι η λίστα των συστάσεων για τον χρήστη και υπάρχουν όλες οι ταινίες που υπάρχουν στο σύνολο δεδομένων. Προκειμένου να παρέχετε καλές συστάσεις, συνιστώνται μόνο οι πρώτες ταινίες της λίστας..

2.8. Belkon Solution to Netflix Prize

Το βραβείο Netflix ήταν ένας ανοικτός διαγωνισμός για τον καλύτερο συνεργάσιμο αλγόριθμο φιλτραρίσματος για την πρόβλεψη αξιολογήσεων χρηστών για ταινίες, με βάση προηγούμενες βαθμολογίες χωρίς άλλες πληροφορίες σχετικά με τους χρήστες ή τις ταινίες,. Στις 21 Σεπτεμβρίου 2009, το μεγάλο βραβείο των 1.000.000 δολαρίων ΗΠΑ δόθηκε στην ομάδα Prakmatic Chaos της BellKor, η οποία κατάφερε να βελτιώσει τον ήδη υπάρχον αλγόριθμο αυξαίνοντας την πιθανότητα πρόβλεψης αξιολογήσεων κατά 10,06%.

Το σύνολο δεδομένων Netflix περιέχει περισσότερες από 100 εκατομμύρια βαθμολογίες ταινιών που έχουν εκτελεστεί από ανώνυμους χρήστες του Netflix μεταξύ 31 Δεκεμβρίου 1999 και 31 Δεκεμβρίου 2005 [4]. Αυτό το σύνολο δεδομένων δίνει αξιολογήσεις για m = 480.189 χρήστες και n = 17.770 ταινίες. Ο διαγωνισμός σχεδιάστηκε σε μορφή εκπαιδευτικού σετ το οποίο σετ περιέχει περίπου 4 εκατομμύρια αξιολογήσεις για τις τελευταίες 9 ταινίες που αξιολογήθηκαν από κάθε χρήστη.
Πιο συγκεκριμένα το αναφερόμενο σετ «σπάει» σε 3 υποσύνολα αυτά είναι

- Σετ Ανίχνευσης
- Σετ Ερωτήματος
- Σετ Δοκιμής

Το σετ ανίχνευσης είναι ένα μέρος του σετ ενώ σύνολα Ερωτήματος και Δοκιμής αποτελούσαν ένα σετ αξιολόγησης, το οποίο είναι γνωστό ως το προκριματικό σετ, στο οποίο οι διαγωνιζόμενοι έπρεπε να προβλέπουν βαθμολογίες. Μόλις ένας διαγωνιζόμενος υποβάλει προβλέψεις, ο prizemaster επιστρέφει το ριζικό μέσο τετραγωνικό σφάλμα (RMSE) που επιτεύχθηκε στο σύνολο Δοκιμής. Έστω λοιπόν μια βαθμολογία rui η οποία υποδεικνύει την προτίμηση ενός χρήστη U για μια ταινία I, το εύρος τιμών για τις βαθμολογήσεις κυμαίνετε ανάμεσα σε 1(αστέρι) έως 5(αστέρια) που υποδηλώνουν έντονο ενδιαφέρον. Προκειμένου να γίνει διάκριση μεταξύ των προβλεπόμενων βαθμολογιών και των γνωστών, η λύση belkon χρησιμοποίει το rui notation για την προβλεπόμενη τιμή του rui.

Στην λύση μπελκον επίσης χρησιμοποιείτε μια ακόμα μεταβλητή. tui η οποία υποδηλώνει το χρόνο βαθμολόγησης rui. Εδώ, ο χρόνος μετράται σε ημέρες, οπότε το tui μετράει τον αριθμό των ημερών που έχουν περάσει από κάποια αρχική χρονική στιγμή. Τα ζευγάρια (u, i) για τα οποία είναι γνωστόο rui αποθηκεύονται στο σύνολο εκπαίδευσης K = {(u, i) | rui είναι γνωστό}. Παρατηρείται λοιπόν ότι το K περιλαμβάνει επίσης το σετ ανίχνευσης. Κάθε χρήστης u συνδέεται με ένα σύνολο αντικειμένων που δηλώνεται από το R (u), το οποίο περιέχει όλα τα στοιχεία για τα οποία υπάρχουν διαθέσιμες αξιολογήσεις από το u. Ομοίως, το R (i) δηλώνει το σύνολο των χρηστών που αξιολόγησαν το στοιχείο i. Έτσι, το N (u) επεκτείνει το R (u) λαμβάνοντας επίσης υπόψη τις βαθμολογίες. Τα μοντέλα για τα δεδομένα βαθμολογίας λαμβάνεται υπόψη ο στόχος αυτής της λύσης είναι η πρόβλεψη μελλοντικών αξιολογήσεων από τους χρήστες. Συστηματικές τάσεις για ορισμένους χρήστες να δίνουν υψηλότερες αξιολογήσεις από άλλες και για ορισμένα στοιχεία να λαμβάνουν υψηλότερες αξιολογήσεις από άλλες. Για την ενθυλάκωση αυτών των αποτελεσμάτων, τα οποία δεν περιλαμβάνουν αλληλεπίδραση χρήστη-

στοιχείου, εντός των προβλεπόμενων βασικών γραμμών. Επειδή αυτοί οι παράγοντες πρόβλεψης τείνουν να συλλάβουν μεγάλο μέρος του παρατηρούμενου σήματος, είναι πολύ σημαντικό να τα διαμορφωθεί με ακρίβεια. Αυτό επιτρέπει την απομόνωση του μέρους του σήματος που αντιπροσωπεύει πραγματικά αλληλεπίδραση χρήστη-στοιχείου και την υποβολή του σε πιο κατάλληλα μοντέλα προτιμήσεων χρήστη. Να σημειωθεί με μ τη συνολική μέση βαθμολογία. Μια πρόβλεψη βασικής γραμμής για μια άγνωστη βαθμολογία rui υποδηλώνεται από το bui και καταγράφει τα αποτελέσματα του χρήστη και του αντικειμένου

$$\square\square\square = \square + \square\square + \square\square$$

Οι παράμετροι bu και bi υποδεικνύουν τις παρατηρούμενες αποκλίσεις του χρήστη u και του στοιχείου i, αντίστοιχα, από τον μέσο όρο. Για παράδειγμα, θεωρώντας ότι ένας χρήστης αξιολογήσει μια ταινία 3.2 έχοντας δεδομένο ότι η μέση βαθμολόγηση σε όλες τις ταινίες είναι 3.7 και πως η ταινία έχει γενική βαθμολόγηση 4.2 δηλαδή 0.5 πάνω από το μέσο όρο και δεδομένου ότι ο χρήστης βαθμολογήσει την ταινία με 3.9. Τότε η εκτίμηση για την αξιολόγηση του χρήστη θα είναι 3.7-0.3+0.5. Ένας τρόπος για την εκτίμηση των παραμέτρων είναι η αποσύνδεση του υπολογισμού των bi από τον υπολογισμό των bu. Πρώτον, για κάθε στοιχείο i έχει οριστεί

$$\square\square = \sum_{\square\epsilon\ \square(} (\square\square\square - \square)\frac{1}{\square 1 + |\square(\square)|}$$

Επομένως για κάθε χρήστη έχει οριστεί

$$\square\square = \sum_{\square\epsilon\ \square(\square)} \frac{\square\square\square - \square - \square}{\square 2} + |\square(\square)|$$

Οι μέσοι όροι συρρικνώνονται προς το μηδέν χρησιμοποιώντας τις παραμέτρους ρύθμισης, λ1, λ2, οι οποίες καθορίζονται με επικύρωση στο σετ ανίχνευσης. Συμφωνά με την λύση belkor θετοντας : λ1 = 25, λ2 = 10. Όποτε η εργασία αυτή αναφέρεται σε πρόβλεψη βασικής γραμμής υπολογιζόμενη με αυτόν τον αποσυνδεδεμένο τρόπο, υποδηλώνεται με b~ ui. Μια ακριβέστερη εκτίμηση του bu και του bi θα τα μεταχειριστεί συμμετρικά, με την επίλυση του προβλήματος των ελάχιστων τετραγώνων

$$\sum_{(\square.\square)\epsilon\ \square} (\square\square\square - \square - \square\square - \square\square)^2 + \square 3(\sum_{\square} \square\square^2 + \sum_{\square} \square\square^2)$$

Στο εξής, το b * υποδηλώνει όλες τις προτιμήσεις των χρηστών και των στοιχείων (bus και bis). Ο πρώτος όρος Σ (u, i) ∈ K (rui - μ + bu + bi) 2 προσπαθεί να βρει bu και bi που ταιριάζουν με τις συγκεκριμένες βαθμολογίες. Ο κανόνας κανονικοποίησης, λ3 (Σu b 2 u + Σi b 2 i), αποφεύγει την υπερφόρτωση με την επιβολή κυρώσεων στα μεγέθη των παραμέτρων. Αυτό το ελάχιστο τετραγωνικό πρόβλημα μπορεί να λυθεί αρκετά αποτελεσματικά με τη μέθοδο της στοχαστικής κλίσης.

Συμφώνα με αυτόν τον αλγόριθμο πιθανά προβλήματα που μπορεί να προκύψουν στις προβλέψεις ασικών γραμμών που αλλάζουν με βάση το χρόνο. Μεγάλο μέρος της χρονικής μεταβλητότητας των δεδομένων περιλαμβάνεται στους προγνωστικούς δείκτες βασικής γραμμής, μέσω δύο σημαντικών χρονικών επιδράσεων. Ο πρώτος αναφέρεται στο γεγονός ότι η δημοτικότητα ενός στοιχείου μπορεί να αλλάξει με την πάροδο του χρόνου Το δεύτερο σημαντικό χρονικό αποτέλεσμα επιτρέπει στους χρήστες να αλλάζουν τις βαθμολογίες τους κατά τη διάρκεια του χρόνου. Για παράδειγμα, ένας χρήστης που τείνει να αξιολογήσει μια μέση ταινία "4 αστέρια", μπορεί τώρα να αξιολογήσει μια τέτοια ταινία "3 αστέρια". Αυτό μπορεί να αντικατοπτρίζει διάφορους παράγοντες, όπως φυσική μετατόπιση στην κλίμακα αξιολόγησης ενός χρήστη, το γεγονός ότι οι βαθμολογίες δίνονται στο πλαίσιο άλλων αξιολογήσεων που δόθηκαν πρόσφατα και επίσης το γεγονός ότι η ταυτότητα του κριτή μέσα σε ένα νοικοκυριό μπορεί να αλλάξει με την πάροδο του χρόνου. Για αυτό και στην λύση Belkon λαμβάνεται μια παράμετρος bu ως συνάρτηση του χρόνου . Αυτό προκαλεί ένα πρότυπο για ένα ευαίσθητο στο χρόνο πρόβλεψης βασικής γραμμής για την βαθμολογία του u κατά την ημέρα tui:

$$\square\square\square = \square + \square\square(\square\square\square) + \square\square(\square\square\square)$$

Στο παραπάνω τύπο τα bu και bi είναι πραγματικές λειτουργίες που αλλάζουν με την πάροδο του χρόνου. Ο ακριβής τρόπος για την κατασκευή αυτών των λειτουργιών θα πρέπει να αντικατοπτρίζει έναν εύλογο τρόπο παραμετροποίησης των συνεπειών των χρονικών αλλαγών. Μια σημαντική διάκριση είναι μεταξύ χρονικών επιδράσεων που εκτείνεται σε παρατεταμένες χρονικές περιόδους και πιο μεταβατικές επιδράσεις. Δεν περιμένει κανείς να κυμαίνεται σε καθημερινή βάση, αλλά να αλλάζει σε μεγαλύτερες χρονικές περιόδους. Η απόφαση για το πώς να χωρίσει το χρονοδιάγραμμα σε Bin θα πρέπει να εξισορροπήσει την επιθυμία για επίτευξη λεπτότερης. Στην εφαρμογή belkon, κάθε Bin αντιστοιχεί σε περίπου δέκα συνεχείς εβδομάδες δεδομένων, που οδηγούν σε 30 Bin που καλύπτουν όλες τις ημέρες στο σύνολο δεδομένων. Μια ημέρα t συνδέεται με ένα ακέραιο Bin (t) (ένας αριθμός μεταξύ 1 και 30 στα δεδομένα μας), έτσι ώστε η κινηματογραφική προκατάληψη να χωρίζεται σε ένα σταθερό τμήμα και ένα τμήμα που αλλάζει την ώρα:

$$□□(□) = □□ + □□$$

Ενώ το binning των παραμέτρων λειτουργεί καλά στα αντικείμενα, είναι περισσότερο μια πρόκληση από την πλευρά των χρηστών. Από τη μία πλευρά, μια λεπτότερη ανάλυση για τους χρήστες να ανιχνεύουν πολύ βραχείες χρονικές επιδράσεις. Από την άλλη πλευρά, δεν αναμένονται αρκετές αξιολογήσεις ανά χρήστη για την παραγωγή αξιόπιστων εκτιμήσεων για απομονωμένους κάδους. Διαφορετικές λειτουργικές μορφές μπορούν να ληφθούν υπόψη για την παραμετροποίηση της συμπεριφοράς του χρήστη, με ποικίλη πολυπλοκότητα και ακρίβεια.

$$□(□) = □□□□(□ − □□) ∗ |□ − □□|^{□}$$

Το μοντέλο των belkon περιλαμβάνει ημερήσιες παραμέτρους. Όποτε μπορεί κάποιος να αναρωτηθεί κατά ποσό αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για προβλέψεις αξιολογήσεων στο μέλλον, σε νέες δηλαδή ημερομηνίες για τις οποίες δεν μπορεί το μοντέλο να εκπαιδευτεί; Η απλή απάντηση είναι ότι για τις μελλοντικές (μη εκπαιδευμένες) ημερομηνίες, οι ημερήσιες παραμέτρους θα πρέπει να λάβουν την προκαθορισμένη τους τιμή. Συγκεκριμένα για το (11), το cu (tui) έχει ρυθμιστεί στο cu, και το bu, το tui είναι μηδενικό. Μετά από όλα, η πρόβλεψη είναι ενδιαφέρουσα μόνο όταν πρόκειται για το μέλλον. Το σετ Netflix Qualifying περιλαμβάνει πολλές αξιολογήσεις σε ημερομηνίες για τις οποίες δεν υπάρχει άλλη αξιολόγηση από τον ίδιο χρήστη και επομένως δεν μπορούν να αξιοποιηθούν οι παράμετροι για την ημέρα. Το μόνο που προσπαθεί να κάνει είναι να καταγράψει παροδικές χρονικές επιδράσεις, οι οποίες είχαν σημαντική επίδραση στην προηγούμενη ανατροφοδότηση των χρηστών. Όταν εντοπίζονται τέτοιες επιδράσεις, πρέπει να συντονιστούν, ώστε να μπορέσουμε να μοντελοποιήσουμε το πιο διαρκή σήμα. Αυτό επιτρέπει στο μοντέλο belkon να καταγράφει καλύτερα τα μακροπρόθεσμα χαρακτηριστικά των δεδομένων, ενώ αφήνοντας τις ειδικές παραμέτρους απορροφούν βραχυπρόθεσμες διακυμάνσεις. Με αυτόν τον τρόπο, οι παράμετροι που αφορούν την ημέρα πραγματοποιούν ένα είδος καθαρισμού δεδομένων, το οποίο βελτιώνει την πρόβλεψη μελλοντικών ημερομηνιών. Το RMSE = 0,9555 αποτέλεσμα του μοντέλου (10) περιλαμβάνεται στο μίγμα. Για να υπάρξουν οι εμπλεκόμενες παραμέτρους, bu, au, αλλά, bi, bi, Bin (t), cu, και κομμένες θα πρέπει να ελαχιστοποιηθεί το κανονικοποιημένο τετράγωνο σφάλμα στο σετ εκπαίδευσης. Η εκμάθηση γίνεται με αλγόριθμο κατάταξης στοχαστικής κλίσης για 30 επαναλήψεις. Χρησιμοποιούμε ξεχωριστό ρυθμό εκμάθησης (μέγεθος βήματος) και τακτοποίηση (αποσύνθεση βάρους) σε κάθε είδος μαθηματικής παραμέτρου, ελαχιστοποιώντας τη συνάρτηση κόστους

$$\sum (□□□ − □ − □□ − □□ ∗ □□□□(□□□) − (□□ + □□□□□(□□□)) ∗ (□□ + □□□□□))^2 + □□□□^2 + □□□□^2$$

$$□,□\epsilon □ \quad + □□□□□□□□^2 + □□□□^2 + □□□□□□□^2 + □□(□□ − 1)^2 + □□□□, □□□^2$$

Οι πραγματικές τιμές των ρυθμών εκμάθησης και των σταθερών τακτοποίησης (λa, λb, ..., λg) έχουν ως εξής:

	Bu	But	Au	Bi	Bibin(t)	Cu	Cut
Irate x 10^3	3	25e-1	1e-2	2	5e-2	8	2
Reg x 10^2	3	5e-1	5000	3	10	1	5e-1

Ο πολλαπλασιαστής cu συρρικνώνεται προς το 1, δηλαδή, (cu - 1) 2, αντί για c 2 u. Παρομοίως, όλες οι μαθησιακές παράμετροι αρχικοποιούνται στο μηδέν, εκτός από το ότι αρχικοποιείται στο 1. Το μείγμα περιλαμβάνει επίσης το αποτέλεσμα του ακριβέστερου προγνωστικού βασικής γραμμής (11). Στην πραγματικότητα, αυτή είναι η μόνη περίπτωση κατά την οποία χρησιμοποιείτε έναν αυτόματο ρυθμιστή παραμέτρων (APT) για να βρεθούν τις καλύτερες. Συγκεκριμένα, χρησιμοποιήθηκε APT1, η οποία περιγράφεται στο [13]. Η βασική συνιστώσα πρόβλεψης είναι ενσωματωμένη στα πιο περιεκτικά μοντέλα. Δεύτερον, αυτό είναι ένα μικρό, γρήγορα εκπαιδευμένο μοντέλο. Έτσι θα μπορεί εύκολα να αντέξει πολλές εκατοντάδες αυτόματες εκτελέσεις που αναζητούν βέλτιστες ρυθμίσεις. Ακόμα, αξίζει να αναφερθεί ότι το πλεονέκτημα του APT ήταν μια μείωση RMSE (μόνο) 0,0016 στις αρχικές χειροκίνητες ρυθμίσεις μας. Οι παράμετροι του αποτελέσματος RMSE = 0,9278 του μοντέλου (11) αποκτήθηκαν με μια διαδικασία κλίσης στοχαστικής κλίσης 40 επαναλήψεων, με τις ακόλουθες σταθερές που διέπουν την εκμάθηση κάθε τύπου παραμέτρου

	Bu	But	Au	Bi	Bibin(t)	Cu	Cut
Irate x 10^3	2.67	2.57	3.11e-3	0.488	0.115	5.64	1.03
Reg x 10^2	2.55	0.231	395	2.55	9.29	4.76	1.90

2.9. Hybrid Filtering

Το υβριδικό μοντέλο είναι μια μέθοδος συνόλου, που σημαίνει ότι είναι ένας συνδυασμός πολλαπλών μοντέλων. Το υβριδικό μοντέλο είναι ένας συνδυασμός ενός συνεργατικού μοντέλου φιλτραρίσματος και του μοντέλου με βάση το περιεχόμενο. Μετά από αυτό, το αντίστοιχο μοντέλο συνεργασίας φιλτραρίσματος χρησιμοποιείται ως μέρος του υβριδικού μοντέλου φιλτραρίσματος. Η μεθοδολογία πίσω από την υβριδική τεχνική φιλτραρίσματος είναι η βέλτιστη χρήση όλων των δεδομένων. Χρησιμοποιώντας μόνο συνεργατικό φιλτράρισμα, κάποιος δεν θεωρεί δεδομένα που σχετίζονται με το περιεχόμενο, όπως τα είδη. Το ίδιο ισχύει και για περιεχόμενο που βασίζεται μόνο στο περιεχόμενο, όπου κανείς δεν θεωρεί σχέσεις με άλλους χρήστες. Προκειμένου να γίνουν πραγματικές συστάσεις, κάθε μοντέλο της μεθόδου του συνόλου θα πρέπει να πάρει βάρος, έτσι ώστε η θέση μιας ταινίας □ για τον χρήστη □ στη σύσταση να είναι

□□□ℎ □ = □□□ ∗ □□□□□ + □□□ ∗ □□□□□

Εάν οι ταινίες έχουν το ίδιο □□□□□, οι ταινίες ταξινομούνται στην αξιολόγηση (πρώτα στην υψηλότερη). Ας δείξουμε αυτή την τεχνική με ένα παράδειγμα. Στον ακόλουθο πίνακα. Παράδειγμα Υβριδικού Φιλτραρίσματος, όπου οι κορυφαίες 8 ταινίες που συνιστώνται για μια, οι κορυφαίες 8 ταινίες για ένα χρήστη υπολογίζονται σύμφωνα με ένα μοντέλο συνεργασίας φιλτραρίσματος (στην περίπτωση αυτή: UBCF) και το μοντέλο με βάση το περιεχόμενο. Επιπλέον, υποθέστε □□□ = 0,4 και □□□ = 0,6. Οι θέσεις σύμφωνα με το υβριδικό μοντέλο φιλτραρίσματος παρουσιάζονται στον σωστό πίνακα.

Pcf	Movie Title
1	Terminator
2	The Apartment
3	Justice League
4	Spartan
5	Deadwood
6	The Prisoner
7	Gladiator
8	Back to the Future

Pcb	Movie Title
1	Justice League
23	The Prisoner
3	Spartan
4	Terminator
5	Back To TheFuture
6	Gladiator
7	The Apartment
8	Deadwood

Phf	Movie Title	Calculation
1	Justice League	0.4*3+0.6*1=1.8
2	The Prisoner	0.4*1+0.6*4=2.8
3	Spartan	0.4*4+0.6*3=3.4
4	Terminator	0.4*6+0.6*2=3.6
5	Back To TheFuture	0.4*2+0.6*7=5.0
6	Gladiator	0.4*8+0.6*5=6.2
7	The Apartment	0.4*7+0.6*6=6.4
8	Deadwood	0.4*5+0.6*8=6.8

2.10. Improving Existing Methods

Σε γενικές γραμμές, η προσπάθεια για βελτιωμένη μοντελοποίηση σύμφωνα με την έρευνα των Robert M. Bell and Yehuda Koren τείνει να ακολουθεί μία από τις τρεις κατευθύνσεις ή έναν συνδυασμό αυτών::

1. Εμβάθυνση των προαναφερόμενων μεθόδων
2. Συνδυασμός διάφορων κλιμάκων δεδομένων
3. Απαγόρευση ρητών πληροφοριών αξιολόγησης με έμμεση συμπεριφορά αξιολόγησης

Ο στόχος είναι να γίνει πρόβλεψη μιας μη βαθμολογημένη αξιολόγηση από το χρήστη u για στοιχείο (ταινία) i, που χαρακτηρίζεται ως rui. Ένα μοντέλο γειτνίασης που οριοθετεί αντικείμενα προσδιορίζει ένα σύνολο γειτονικών στοιχείων N (I, u) που άλλες χρήσεις τείνουν να αποτιμούν την ομοιότητα με την βαθμολογία τους I. Όλα τα αντικείμενα στο N (I, u) πρέπει να έχουν βαθμολογηθεί από u. Η προβλεπόμενη τιμή του Rui λαμβάνεται ως σταθμισμένος μέσος όρος της διαβάθμισης των γειτονικών στοιχείων

$$\square\square\square \leftarrow \square\square\square + \Sigma_{\square\epsilon\ \square(\square;\square)}\ \square\square\square(\square\square\square - \square\square\square) \,/\, \Sigma_{\square\epsilon\ \square(\square;\square)}\ \square\square\square \quad (1)$$

Οι ομοιότητες των αντικειμένων που δηλώνονται από το Si, j παίζουν κεντρικό ρόλο εδώ καθώς χρησιμοποιούνται τόσο για την επιλογή των γειτόνων όσο και για τη στάθμιση του παραπάνω μέσου όρου. Οι κοινές επιλογές είναι η συγγένεια του συντελεστή συσχέτισης Pearson. Προκειμένου να οδοντωθεί κάποιος βασικός παράγοντας πρόβλεψης για τη rui, η χρήση του buii ήταν απαραίτητη. Είναι σημαντικό να χρησιμοποιηθούν αυτές οι βασικές τιμές για την εξάλειψη των ειδικών προκαταλήψεων του αντικειμένου και του χρήστη, οι οποίες ενδέχεται να εμποδίσουν το μοντέλο να αποκαλύψει τις περισσότερες σχέσεις θεμελίων. Οι προηγούμενες μέθοδοι συχνά λαμβάνουν το bui ως μέση βαθμολογία του χρήστη u ή του στοιχείου I.
Οι μέθοδοι που βασίζονται στην γειτονιά έγιναν πολύ δημοφιλείς επειδή είναι διαισθητικές και σχετικές με την σχετικότητα. Ειδικότερα, δεν απαιτούν συντονισμό πολλών παραμέτρων ή εκτεταμένο στάδιο εκπαίδευσης. Παρέχουν επίσης μια συνοπτική και διαισθητική αιτιολόγηση για τις υπολογισμένες προβλέψεις.

Μια δεύτερη μέθοδος που μπορεί να βελτιώσει τις προβλέψεις είναι τα μοντέλα Latent Factor[6]
Τα μοντέλα λανθασμένου παράγοντα μετρούν τη συμφωνία των χρηστών και των ταινιών σε μια σειρά χαρακτηριστικών που αντλούνται από τα δεδομένα. Με αυτό τον τρόπο κάθε χρήστης μπορεί να συσχετιστεί με ένα διάνυσμα παράγοντα χρήστη Pu ∈ R, και κάθε ταινία με ένα διάνυσμα παράγοντα ταινίας Qi ∈ R. Το πιο εμπλεκόμενο μέρος είναι η εκτίμηση των παραγόντων. Μια απλή προσέγγιση θα ελαχιστοποιούσε τη λειτουργία τακτοποιημένου κόστους:

$$\Sigma_{|\square,\square|\epsilon\ \square}(\square\square\square - \square\square^{\square}{}_{*}\ \square\square)^2 + \square(\square\square^2 + \square\square^2) \qquad (3)$$

Το σύνολο K περιέχει όλα τα ζευγάρια (u, i) για τα οποία ο Rui είναι γνωστός ως η παράμετρος αποκαταστάσεως λ εμποδίζει την υπερφόρτωση μιας τυπικής τιμής είναι λ = 0,05. Αυτό το μοντέλο μπορεί να ωφεληθεί από την κίνηση στις τρεις κατευθύνσεις
Η τετραγωνισμένη ποινή που χρησιμοποιείται στην (3) υποθέτει ότι όλοι οι συντελεστές προέρχονται από ανεξάρτητη κανονική κατανομή με μηδενική μέση και ίδια διακύμανση. Ένα πιο πλούσιο μοντέλο υιοθετεί μια γενική πολυμεταβλητή κανονική κατανομή για τους παράγοντες. Κατά συνέπεια, η βαθμολογία μπορεί να διαμορφωθεί από τα ακόλουθα. Αρχικά σχεδιάζουμε παράγοντες χρήστη και ταινία από την κανονική κοινή κατανομή

$$\square\square \sim \square(\square, \square) \qquad \square\square \sim \square(\square, \square) \quad (4)$$

Η κάθε βαθμολογία Rui λαμβάνεται από την κανονική κατανομή

$$\square\square\square = \square(\square\square^{\square}\square\square, \square) \quad (5)$$

Οι πριμοδοτήσεις που δίδονται στο (4) αποφεύγουν την υπερφόρτωση με τη συρρίκνωση των παραγόντων προς τις αρχικές τιμές όταν δεν υπάρχουν αρκετά. Η αφαίρεση των αντισταθμίσεων παραμέτρων που υποδηλώνει το (3) παράγει σημαντικά κέρδη ακρίβειας. Για παράδειγμα 60 παραγόντων, το RMSE στα δεδομένα δοκιμής έπεσε στο 0,899. Φυσικά προστιθέμενη εκφραστική δύναμη του Gausian προηγείται με πρόσθετη πολυπλοκότητα των αλγορίθμων βελτιστοποίησης. Υπάρχουν δύο τεχνικές. Το ένα βασίζεται στη δειγματοληψία Gibbs και το άλλο βασίζεται στη μεγιστοποίηση των προσδοκιών, όπου εναλλάσσονται μεταξύ των συντελεστών σταθεροποίησης και

καθορίζουν τις παραμέτρους [12]. Επίσης, ο τύπος (3) μπορεί να βελτιώσει την ακρίβεια της πρόβλεψης με την ενσωμάτωση είτε της τοπικής θέσης της δυαδικής θέσης στο μοντέλο. Η ενσωμάτωση της τοπικής άποψης παρέχει πιο ολοκληρωμένη προβολή των δεδομένων σε πολλά επίπεδα. Για το σκοπό αυτό, η προσαρμογή του παράγοντα χρήστη ότι η συμπεριφορά του μοντέλου ενός χρήστη στη γειτονιά μιας δεδομένης ταινίας.

$$\sum_{\square\epsilon\ \square(\square)} \square\square\square(\square\square\square - \square\square(\square)^{\square}\square\square)^2 + \square|\square\square(\square)|^2$$

Εδώ, N (u) είναι το σύνολο των ταινιών που αξιολογούνται από το χρήστη u. Το σταθερό Sij είναι ένα μέτρο ομοιότητας μεταξύ των ταινιών i και j. Με την ενσωμάτωση της τοπικής άποψης, η ακρίβεια της πρόβλεψης βελτιώνεται σημαντικά. Για την περίπτωση των 60 παραγόντων, το σφάλμα μειώνεται στο RMSE = 0,897

Αντί να ενσωματώσουμε την τοπική άποψη, μπορούμε να ενσωματώσουμε τη δυαδική άποψη. Χρησιμοποιείται μια αρχή από τη μέθοδο Paterek NVSD [12333333]. Η μέθοδος NVSD επαναλαμβάνεται από την αυθαίρετη παραμετροποίηση κάθε χρήστη, αλλά μάλλον διαμορφώνει κάθε χρήστη με βάση τις ταινίες που έχει αξιολογήσει. Με αυτό τον τρόπο κάθε ταινία εγώ συνοδεύεται με δύο παράγοντες ταινίας φορέα Qi και Zi. Η φαντασία ενός χρήστη u είναι μέσα από το σταθμισμένο άθροισμα έτσι Rui προβλέπεται ως. Είναι σημαντικό ότι η NVSD μοντελοποιεί τη συμπεριφορά του χρήστη με βάση μόνο τη δυαδική άποψη. Αυτό επιτρέπει μια απλή ενσωμάτωση της δυαδικής παρατήρησης στο μοντέλο παραγοντοποίησης

Οι παράμετροι υπολογίζονται με ελαχιστοποίηση της κλίσης του συσχετιζόμενου κανονικοποιημένου τετράγωνου σφάλματος. Το μοντέλο που χρησιμοποιήθηκε με 60 παράγοντες προβλέπει το σύνολο δοκιμών με RMSE μικρότερο από 0,897. Για να συνοψίσουμε την ακρίβεια του μοντέλου των λανθάνων παραγόντων, μπορούμε να εννοήσουμε ακολουθώντας τρεις διαφορετικές κατευθύνσεις. Η πρώτη μπορεί να εμβαθύνει τα θεμέλια του μοντέλου. Εναλλακτικά, μπορεί κανείς να διατηρήσει την απλή δομή του αρχικού μοντέλου, αλλά να αποκτήσει ακόμα μεγαλύτερη ακρίβεια, εισάγοντας συμπληρωματικές προοπτικές των δεδομένων στο μοντέλο είτε της τοπικής προβολής είτε της δυαδικής προβολής.

3. Youtube

3.1. Introduction

Το YouTube είναι η μεγαλύτερη πλατφόρμα στον κόσμο για τη δημιουργία, την κοινή χρήση και την εύρεση βίντεο. Οι συστάσεις του YouTube έχουν την ευθύνη να βοηθήσουν περισσότερους από ένα δισεκατομμύριο χρήστες να ανακαλύψουν εξατομικευμένο περιεχόμενο από ένα συνεχώς αυξανόμενο αριθμό βίντεο. Σε αυτό το κεφάλαιο θα συζητηθεί πώς το YouTube προτείνει βίντεο και επίσης μια ανάλυση στο background interface"Tensorflow" πρόκειται να γίνει. Δεδομένου του μεγάλου αριθμού πληροφορίας απλοί αλγόριθμοι συστάσεων που δουλεύουν αξιόπιστα σε μικρά προβλήματα δεν είναι ιδανικοί για το σύστημα το οποίο το Youtube προτείνει βίντεο λόγου σαν κάνοντας έτσι την σύσταση των βίντεο δύσκολή. Πέρα όμως αυτό το γεγονός ένας παράγοντας που κάνει δύσκολη την σύσταση βίντεο στο Youtube η σωστή διαμόρφωση του συστήματος από νέες μεταμορφώσεις ανά δευτερόλεπτο για τον λόγο αυτό το σύστημα θα πρέπει να ανταποκρίνεται αρκετά ώστε να διαμορφώνει το περιεχόμενο που έχει μεταφορτωθεί πρόσφατα καθώς και τις τελευταίες ενέργειες του χρήστη. Ένας ακόμα παράγοντας ο οποίος καθιστά δύσκολο το έργο για την σύσταση των βίντεο είναι η ιστορική συμπεριφορά των χρηστών λόγο της ασυμφωνίας που υπάρχει στο των βίντεο που υπάρχουν ανά καιρούς από κάθε χρήστη για αυτό και οι αλγόριθμοι πρέπει είναι ισχυροί σε αυτούς τους παράγοντες προκειμένου μια σύσταση να θεωρηθεί αξιόλογη.

3.2. System Overview

Το σύστημα του Youtube απαρτίζεται από 2 ευρέως γνωστά νευρωνικά δίκτυα αυτά είναι [7]:

- Candidate Generation network
- Ranking network

Candidate generation network: Το είδος αυτού του νευρωνικού δικτύου δέχεται ως είσοδο το ιστορικό παρακολούθησης ενός χρήστη και ανακτά ένα μικρό υποσύνολο από βίντεο από μια τεράστια γκάμα. Τα βίντεο που λοιπόν που παράγονται από το σύστημα πρέπει σε μεγάλο βαθμό να είναι σχετικά με τον χρήστη διαφορετικά ο κίνδυνος να προταθούν βίντεο που δεν έχουν σχέση με τα ενδιαφέροντα του χρήστη είναι μεγάλος. Το δίκτυο υποψήφιων γενεών παρέχει μόνο ευρεία εξατομίκευση μέσω συνεργατικού φιλτραρίσματος. Η ομοιότητα μεταξύ των χρηστών εκφράζεται σύμφωνα με την παρακολούθηση κοινών βίντεο καθώς και ανάλογα με το ιστορικό των αναζητήσεων τους. Παρουσιάζοντας λοιπόν τις καλύτερες συμφώνα με το σύστημα συστάσεις σε μια λίστα προαπαιτεί την διάκριση σχετικών βίντεο. Κατά την δημιουργία λοιπόν των βίντεο προς σύσταση το τεράστιο περιεχόμενο του YouTube συρρικνώνετε σε εκατοντάδες βίντεο τα οποία μπορεί να είναι σχετικά ως προς τον χρήστη. Ο προκάτοχος του συνιστώμενου που περιγράφηκε εδώ ήταν μια προσέγγιση παραγοντοποίησης μήτρας που εκπαιδεύτηκε κάτω από την απώλεια βαθμού [8]. Οι πρώτες επαναλήψεις του μοντέλου νευρωνικών δικτύων μιμούνται αυτή τη συμπεριφορά παραγοντοποίησης με ρηχά δίκτυα που ενσωματώνουν μόνο τα προηγούμενα ρολόγια του χρήστη. Από αυτή την άποψη, η προσέγγιση μπορεί να θεωρηθεί ως μια μη γραμμική γενίκευση των τεχνικών παραγοντοποίησης.

Ranking network: Αυτό το δίκτυο λειτουργεί ως εξής: αναθέτοντας μια βαθμολογία σε κάθε βίντεο σύμφωνα με μια επιθυμητή αντικειμενική λειτουργία χρησιμοποιώντας ένα πλούσιο σύνολο χαρακτηριστικών που περιγράφουν το βίντεο και το χρήστη. Τα βίντεο με τα υψηλότερα σκορ παρουσιάζονται στον χρήστη, ταξινομημένα με βάση το σκορ τους. Η προσέγγιση σε δύο στάδια της σύστασης επιτρέπει τη διατύπωση συστάσεων από ένα πολύ μεγάλο αριθμό (εκατομμυρίων) βίντεο ενώ ταυτόχρονα είναι βέβαιο ότι ο μικρός αριθμός βίντεο που εμφανίζονται στη συσκευή είναι εξατομικευμένος και αφοσιωμένος για τον χρήστη.

3.3. Candidate Generation

3.3.1. Recommendation as Classification

Σε αυτή την ενότητα θα γίνει μια ανάλυση του τρόπου με τον οποίο λειτουργεί η τεχνική candidate generation. Έστω ότι θεωρούμε μια σύσταση που γίνεται σε 1 συγκεκριμένο χρόνο παρακολούθησης

βίντεο « Wt »˝ σε χρόνο "T" μεταξύ εκατομμυρίων βίντεο "i" (τάξεις) από ένα σώμα "V" βασισμένο σε χρήστη "U" και περιεχόμενο C.

$$P(Wt = i|U,C) = e^{Uiu} / \sum_{j \in V} e^{Uju}$$

όπου U ∈ RN αντιπροσωπεύει μια ενσωμάτωση του χρήστη, το ζεύγος συμφραζομένων και το vj ∈ RN αντιπροσωπεύουν τις ενσωματωμένες για κάθε υποψήφιο βίντεο. Σε αυτή τη ρύθμιση, μια ενσωμάτωση είναι απλά μια χαρτογράφηση των αραιών οντοτήτων (μεμονωμένα βίντεο, χρήστες κλπ.) Σε ένα πυκνό διάνυσμα στο R N. Σκοπός αυτού του νευρονικού δικτύου είναι να μάθει τις ενσωματωμένες στο χρήστη "U" ως συνάρτηση του ιστορικού και του πλαισίου του χρήστη που είναι χρήσιμες για τη διάκριση μεταξύ των βίντεο με έναν ταξινομητή softmax. Αν και στο YouTube υπάρχουν σαφείς μηχανισμοί ανάδρασης (thumb up / down, έρευνες εντός προϊόντος κ.λπ.), χρησιμοποιείται το σιωπηλό feedback των ρολογιών για την εκπαίδευση του μοντέλου, όπου ένας χρήστης που ολοκληρώνει ένα βίντεο είναι ένα θετικό παράδειγμα. Η επιλογή αυτή βασίζεται στις πιο πολυσυζητημένες ιστορικές προδιαγραφές των χρηστών, οι οποίες μας επιτρέπουν να παράγουμε συστάσεις βαθιά στην ουρά όπου η ρητή ανατροφοδότηση είναι εξαιρετικά αραιή.

3.3.2. Efficient Extreme Multiclass

Προκειμένου να εκπαιδευτεί αποτελεσματικά ένα τέτοιο μοντέλο με εκατομμύρια τάξεις, απαιτείται η χρήση μιας συγκεκριμένης τεχνικής. Η τεχνική αυτή δειγματολήπτη τα δεδομένα αρνητικών κλάσεων από την κατανομή του υποβάθρου και στη συνέχεια διορθώνει αυτή τη δειγματοληψία μέσω της βαρύνσας σημασίας [9]. Για κάθε παράδειγμα, ελαχιστοποιείται η απώλεια διασταυρούμενης εντροπίας για την αληθινή ετικέτα και τις αρνητικές κλάσεις του δείγματος. Στην πράξη, δειγματίζονται αρκετές χιλιάδες δεδομένα αρνητικών κλάσεων, που αντιστοιχούν σε περισσότερες από 100 φορές ταχύτερη μετάδοση από την παραδοσιακή softmax. Μια δημοφιλής εναλλακτική προσέγγιση είναι η ιεραρχική softmax [10] . Στην ιεραρχική softmax, η διέλευση κάθε κόμβου στο δέντρο συνεπάγεται διάκριση μεταξύ ομάδων τάξεων που συχνά δεν σχετίζονται, καθιστώντας το πρόβλημα ταξινόμησης πολύ δυσκολότερο και υποβαθμιστικό. Η βαθμολόγηση εκατομμυρίων αντικειμένων υπό αυστηρή εξυπηρέτηση λανθάνουσας διάρκειας δεκάδων χιλιοστών του δευτερολέπτου απαιτεί ένα κατά προσέγγιση σύστημα βαθμολόγησης υπογραμμισμένο στον αριθμό των κατηγοριών. Τα προηγούμενα συστήματα στο YouTube βασίστηκαν στον κατακερματισμό [11] και ο ταξινομητής που περιγράφεται εδώ χρησιμοποιεί μια παρόμοια προσέγγιση. Δεδομένου ότι οι βαθμονομημένες πιθανότητες από τη στρώση εξόδου softmax δεν χρειάζονται κατά το χρόνο εξυπηρέτησης, το πρόβλημα βαθμολόγησης μειώνεται σε μια πλησιέστερη αναζήτηση γειτονικού χώρου για το οποίο μπορούν να χρησιμοποιηθούν βιβλιοθήκες γενικής χρήσης [12].

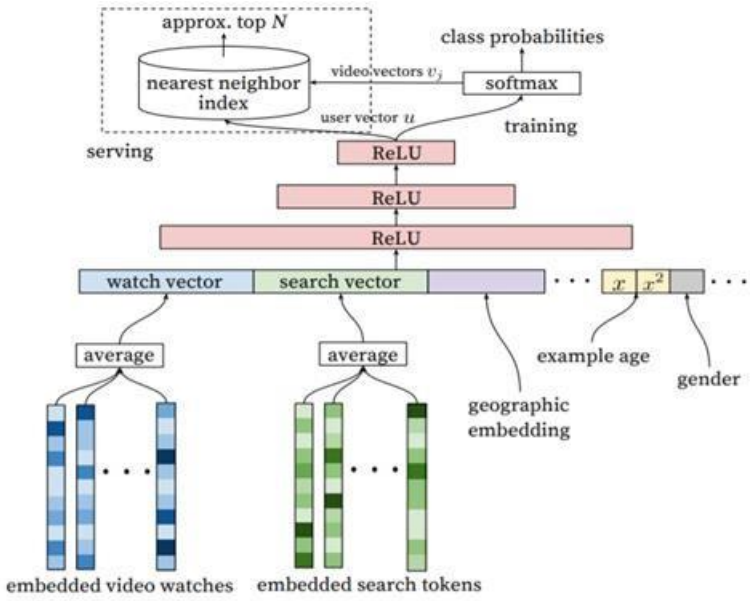
3.3.3. Heterogenous Signals dd

Ένα βασικό πλεονέκτημα της χρήσης των βαθιών νευρωνικών δικτύων ως γενίκευσης της παραγοντοποίησης των μήτρων είναι ότι αυθαίρετα συνεχή και κατηγορηματικά χαρακτηριστικά μπορούν εύκολα να προστεθούν στο μοντέλο. Το ιστορικό αναζήτησης αντιμετωπίζεται παρόμοια με το ιστορικό παρακολούθησης - κάθε ερώτημα μετατρέπεται σε unigrams και bigrams και κάθε στοιχείο είναι ενσωματωμένο. Μόλις υπολογιστεί ο μέσος όρος, τα ενοποιημένα ερωτήματα των ενοποιημένων ερωτημάτων του χρήστη αντιπροσωπεύουν ένα συνοπτικό πυκνό ιστορικό αναζήτησης. Τα δημογραφικά χαρακτηριστικά είναι σημαντικά για την παροχή προνομίων, ώστε οι συστάσεις να συμπεριφέρονται εύλογα για τους νέους χρήστες. Η γεωγραφική περιοχή και η συσκευή του χρήστη είναι ενσωματωμένες και συνεκτικοποιημένες. Απλές δυαδικές και συνεχείς λειτουργίες όπως το φύλο του χρήστη, η καταγεγραμμένη κατάσταση και η ηλικία εισάγονται απευθείας στο δίκτυο ως πραγματικές τιμές ομαλοποιημένες στο [0, 1].

3.3.4. Label and Context Selection

Πριν γίνει η μεταφορά ενός αποτελέσματος σε 1 συγκεκριμένο πλαίσιο πρέπει να λάβει χώρα η επίλυση ενός δευτερεύοντος προβλήματος. Ένα κλασικό παράδειγμα είναι η υπόθεση ότι η ακριβής πρόβλεψη αξιολογήσεων οδηγεί σε αποτελεσματικές συστάσεις ταινιών [13]. Η επιλογή αυτού του

αντικατασταθέντος μαθησιακού προβλήματος έχει μεγάλη σημασία για την απόδοση στις δοκιμές A / B, αλλά είναι πολύ δύσκολο να μετρηθεί με τα πειράματα εκτός σύνδεσης. Τα παραδείγματα κατάρτισης δημιουργούνται από όλα τα ρολόγια του YouTube (ακόμη και εκείνα που είναι ενσωματωμένα σε άλλους ιστότοπους) αντί να παρακολουθούν μόνο τις συστάσεις που παράγονται. Διαφορετικά, θα ήταν πολύ δύσκολο να εμφανιστεί το νέο περιεχόμενο και ο συνήγορος θα ήταν υπερβολικά προκατειλημμένος προς εκμετάλλευση. Εάν οι χρήστες ανακαλύπτουν βίντεο με άλλα μέσα από τις συστάσεις του YouTube, μια καλή λύση είναι να είναι δυνατή η ταχεία διάδοση αυτής της ανακάλυψης σε άλλους μέσω συνεργατικού φιλτραρίσματος. Μια άλλη βασική ιδέα ότι οι βελτιωμένες ζωντανές μετρήσεις ήταν να δημιουργήσουν ένα σταθερό αριθμό εκπαιδευτικών παραδειγμάτων ανά χρήστη, σταθμίζοντας αποτελεσματικά τους χρήστες μας στην λειτουργία απώλειας. Αυτό εμπόδισε μια μικρή ομάδα πολύ ενεργών χρηστών να κυριαρχήσουν στην απώλεια. Κάπως αντίθετα από τον τρόπο με τον οποίο θα πρέπει να ληφθεί μέριμνα, πρέπει να ληφθεί μέριμνα να μην παρέχονται πληροφορίες από τον ταξινομητή, προκειμένου να εμποδιστεί το μοντέλο να αξιοποιήσει τη δομή του ιστότοπου και να υπερκεράσει το υποκατάστατορόβλημα. Σκεφθείτε ως παράδειγμα



Παράδειγμα αρχιτεκτονίκης μοντέλου «deep candidate generation»

περίπτωση κατά την οποία ο χρήστης μόλις εξέδωσε ένα ερώτημα αναζήτησης για "taylor swift". Δεδομένου ότι το πρόβλημα τίθεται ως πρόβλεψη του επόμενου βίντεο που παρακολουθήσατε, ένας ταξινομητής που έχει δώσει αυτές τις πληροφορίες θα προβλέψει ότι τα πιο πιθανά βίντεο που θα παρακολουθούνται είναι εκείνα που εμφανίζονται στην αντίστοιχη σελίδα αποτελεσμάτων αναζήτησης για "taylor swift". Αδιαμφισβήτητα, η αναπαραγωγή της τελευταίας σελίδας αναζήτησης του χρήστη, όπως οι συστάσεις της αρχικής σελίδας, εκτελείται ελάχιστα. Με την απόρριψη πληροφοριών αλληλουχίας και την αντιπροσώπευση ερωτημάτων αναζήτησης με μια μη προσαρμοσμένη σακούλα των μαρκών, ο ταξινομητής δεν γνωρίζει πλέον άμεσα την προέλευση της ετικέτας. Τα φυσικά μοντέλα κατανάλωσης βίντεο οδηγούν συνήθως σε πολύ ασύμμετρες πιθανότητες συν-παρακολούθησης. Οι επεισοδιακές σειρές συνήθως παρακολουθούνται διαδοχικά και οι χρήστες συχνά ανακαλύπτουν καλλιτέχνες σε ένα είδος που αρχίζει με το πιο ευρέως δημοφιλές πριν επικεντρωθούν σε μικρότερες κόγχες. Επομένως, διαπιστώνεται ότι πολύ καλύτερη απόδοση που προβλέπει την επόμενη παρακολούθηση του χρήστη, παρά την πρόβλεψη ενός τυχαία αναστημένου ρολογιού (Εικόνα 5). Πολλά συνεργατικά συστήματα φιλτραρίσματος επιλέγουν σιωπηρά τις ετικέτες και το πλαίσιο κρατώντας ένα τυχαίο στοιχείο και προβλέποντάς το από άλλα στοιχεία του ιστορικού του χρήστη (5α). Αυτό διαρρέει τις μελλοντικές πληροφορίες και αγνοεί οποιαδήποτε ασύμμετρα πρότυπα κατανάλωσης.

3.4. Ranking

Ο πρωταρχικός ρόλος της κατάταξης είναι η χρήση δεδομένων εμφάνισης για την εξειδίκευση και τη βαθμονόμηση των υποψήφιων προβλέψεων για τη συγκεκριμένη διεπαφή χρήστη. Για παράδειγμα, ένας χρήστης μπορεί να παρακολουθήσει ένα δεδομένο βίντεο με μεγάλη πιθανότητα γενικά, αλλά είναι απίθανο να κάνει κλικ στην συγκεκριμένη εμφάνιση αρχικής σελίδας λόγω της επιλογής μικρογραφίας. Κατά τη διάρκεια της κατάταξης, η πρόσβαση σε πολλά άλλα χαρακτηριστικά που περιγράφουν το βίντεο και τη σχέση του χρήστη με το βίντεο, επειδή βαθμολογούνται μόνο μερικές εκατοντάδες βίντεο και όχι τα εκατομμύρια που σημειώθηκαν στην υποψήφια γενιά. Η κατάταξη είναι επίσης ζωτικής σημασίας για τη συγκέντρωση διαφορετικών υποψηφίων πηγών, των οποίων οι βαθμολογίες δεν είναι άμεσα συγκρίσιμες. Χρησιμοποιείται ένα βαθύ νευρωνικό δίκτυο με παρόμοια αρχιτεκτονική με την υποψήφια γενιά για να εκχωρήσετε ένα ανεξάρτητο σκορ σε κάθε εμφάνιση βίντεο χρησιμοποιώντας τη λογική παλινδρόμηση. Στη συνέχεια, ο κατάλογος των βίντεο ταξινομείται βάσει αυτού του σκορ και επιστρέφεται στον χρήστη. Ο τελικός στόχος κατάταξης ρυθμίζεται συνεχώς με βάση τα αποτελέσματα δοκιμών A / B, αλλά είναι γενικά απλή λειτουργία του αναμενόμενου χρόνου παρακολούθησης ανά εντύπωση. Η κατάταξη βάσει του ποσοστού κλικ αυξάνει συχνά τα παραπλανητικά βίντεο που ο χρήστης δεν ολοκληρώνει ("clickbait") ενώ ο χρόνος παρακολούθησης καταγράφει καλύτερα την αφοσίωση [14, 15].

Αντιπροσωπευτικό χαρακτηριστικό: Τα χαρακτηριστικά διαχωρίζονται με την παραδοσιακή ταξινόμηση κατηγορικών και συνεχών / σειριακών χαρακτηριστικών. Τα κατηγορικά χαρακτηριστικά που χρησιμοποιούμε ποικίλλουν ευρέως στην καρδιανή τους - μερικά είναι δυαδικά (π.χ. αν ο χρήστης είναι συνδεδεμένος) ενώ άλλοι έχουν εκατομμύρια πιθανές τιμές (π.χ. το τελευταίο ερώτημα αναζήτησης του χρήστη). Τα χαρακτηριστικά χωρίζονται περαιτέρω ανάλογα με το εάν συνεισφέρουν μόνο μία μόνο τιμή ("μονοδύναμη") ή ένα σύνολο τιμών ("πολυσθενή"). Ένα παράδειγμα μιας μονοβαθμικής κατηγορικής συνάρτησης είναι το αναγνωριστικό βίντεο της εμφανίσεως που βαθμολογείται, ενώ ένα αντίστοιχο πολυσθενές χαρακτηριστικό μπορεί να είναι μια τσάντα με τα τελευταία αναγνωριστικά βίντεο N που έχει παρακολουθήσει ο χρήστης. Επίσης, ταξινομούμε χαρακτηριστικά σύμφωνα με το αν περιγράφουν ιδιότητες του στοιχείου ("εμφάνιση") ή ιδιότητες του χρήστη / πλαισίου ("ερώτημα"). Οι λειτουργίες ερωτήματος υπολογίζονται μία φορά ανά αίτηση και υπολογίζονται οι λειτουργίες εμφάνισης για κάθε στοιχείο που έχει βαθμολογηθεί.

3.5. TensofrFlow

3.5.1. System Overview

Το TensorFlow είναι μια Python βιβλιοθήκη ανοιχτού κώδικα για αριθμητικούς υπολογισμούς που κάνει την μηχανική μάθηση γρηγορότερη και ευκολότερη. Στις μέρες μας η δημιουργία μοντέλων μηχανικής μάθησης είναι πλέον πολύ πιο εύκολη όπως συνήθιζε να ήταν λόγω των frameworks που έχουν αναπτυχτεί για αυτό το σκοπό., όπως το μοντέλο μηχανικής μάθησης της Google που διευκολύνουν τη διαδικασία απόκτησης δεδομένων, εκπαιδευτικών μοντέλων, προγνωστικών και βελτίωσης των μελλοντικών αποτελεσμάτων.

Το TensorFlow συνδυάζει μια σειρά από μοντέλα μηχανικής μάθησης και βαθιάς μάθησης (γνωστά και ως νευρωνικά δίκτυα) και αλγορίθμους και τα καθιστά χρήσιμα μέσω μιας κοινής μεταφοράς. Χρησιμοποιεί την Python για να παρέχει ένα βολικό API front-end για την κατασκευή εφαρμογών με το πλαίσιο, ενώ εκτελεί αυτές τις εφαρμογές σε C ++ υψηλής απόδοσης.

Το TensorFlow μπορεί να εκπαιδεύσει και να εκτελέσει βαθιά νευρωνικά δίκτυα για χειρόγραφη ταξινόμηση ψηφίων, αναγνώριση εικόνας, ενσωματωμένες λέξεις, επαναλαμβανόμενα νευρωνικά δίκτυα, μοντέλα ακολουθίας για μηχανική μετάφραση, επεξεργασία φυσικής γλώσσας και προσομοιώσεις βασισμένες σε PDE (μερική διαφορική εξίσωση). Το καλύτερο από όλα, το TensorFlow υποστηρίζει την πρόβλεψη της παραγωγής σε κλίμακα, με τα ίδια μοντέλα που χρησιμοποιούνται για την εκπαίδευση.

3.5.2. How TensorFlow works

Το TensorFlow επιτρέπει στους προγραμματιστές να δημιουργούν γραφικές παραστάσεις ροών δεδομένων που περιγράφουν τον τρόπο με τον οποίο τα δεδομένα μετακινούνται μέσω ενός γραφήματος ή μιας σειράς κόμβων επεξεργασίας. Κάθε κόμβος στο γράφημα αντιπροσωπεύει μια μαθηματική

λειτουργία και κάθε σύνδεση ή άκρη μεταξύ κόμβων είναι μια πολυδιάστατη διάταξη δεδομένων. Οι κόμβοι και τα tensors στο TensorFlow είναι αντικείμενα Python και οι εφαρμογές TensorFlow είναι εφαρμογές Python. Ωστόσο, οι πραγματικές εργασίες μαθηματικών δεν εκτελούνται στην Python. Οι βιβλιοθήκες μετασχηματισμών που είναι διαθέσιμες μέσω του TensorFlow γράφονται ως δυαδικά αρχεία C ++ υψηλής απόδοσης. Η Python κατευθύνει μόνο την κυκλοφορία μεταξύ των κομματιών και παρέχει αφαίρεση προγραμματισμού υψηλού επιπέδου για να τους συνδέει μαζί.

Οι εφαρμογές TensorFlow μπορούν να εκτελεστούν σε κάθε σύστημα το οποίο υποστηρίζεται από ένα τοπικό μηχάνημα, ένα cluster στο cloud, συσκευές iOS και Android, CPU ή GPU. Για πιο γρήγορη εκτέλεση των εφαρμογών TensorFlow η Google συνιστά να χρησιμοποιηθεί το cloud που παρέχει ένα προσαρμοσμένο μοντέλο TensorFlow Processing Unit (TPU). Ωστόσο, τα μοντέλα που προκύπτουν από το TensorFlow μπορούν να αναπτυχθούν στις περισσότερες συσκευές όπου θα χρησιμοποιηθούν για την προβολή των προβλέψεων. Γενικά τα συστήματα τεχνητής νοημοσύνης και οι αλγόριθμοι μηχανικής μάθησης έχουν δημιουργήσει θεματικά αποτελέσματα σε τομείς όπως

- Επεξεργασίας γλώσσας
- Αναγνώρισης ομιλίας
- Όρασης

Πέρα από αυτούς του τομείς λοιπόν η μηχανική μάθηση έχει δημιουργήσει ευκολίες στην καθημερινή ζωή με την δημιουργία συστάσεων από τα διάφορα συστήματα από πλατφόρμες ηλεκτρονικού εμπορίου, ανίχνευση οικονομικών απατών τα προσαρμοσμένα αποτελέσματα αναζήτησης ιστού και οι τροφές κοινωνικού δικτύου, καθώς και νέες ανακαλύψεις στη γονιδιωματική. Ένας ιδιαίτερος κλάδος της μηχανικής μάθησης, «Deep Learning», έχει αποδειχθεί ιδιαίτερα αποτελεσματικός τα τελευταία χρόνια. Η τεχνική «Deep Learning» είναι μια οικογένεια αλγορίθμων εκμάθησης αναπαράστασης που χρησιμοποιούν σύνθετες αρχιτεκτονικές νευρωνικών δικτύων με μεγάλο αριθμό κρυφών επιπέδων, το καθένα από τα οποία αποτελείται απλούς αλλά μη γραμμικούς μετασχηματισμούς στα δεδομένα εισόδου. Τα τελευταία χρόνια η άνοδος της τεχνικής αυτής επέτρεψε κυρίως τη μεγαλύτερη διαθεσιμότητα μεγάλων συνόλων δεδομένων που περιέχουν περισσότερα παραδείγματα κατάρτισης. Οι αλγόριθμοι Deep Learning και τα μεμονωμένα αρχιτεκτονικά συστατικά, όπως οι μετασχηματισμοί αναπαράστασης, οι λειτουργίες ενεργοποίησης ή οι μέθοδοι νομιμοποίησης μπορούν αρχικά να εκφραστούν σε μαθηματική σχέσεις, πρέπει τελικά να μεταγραφούν σε ένα πρόγραμμα υπολογιστή για πραγματική χρήση στον κόσμο. Για το σκοπό αυτό, υπάρχουν ορισμένες βιβλιοθήκες και πλαίσια λογισμικού ανοιχτού κώδικα καθώς και εμπορικών μηχανών μάθησης. Μεταξύ αυτών είναι οι Theano [5], Torch [6], scikitlearn [7] και πολλά άλλα. Τον Νοέμβριο του 2015, αυτή η λίστα επεκτάθηκε από την TensorFlow, μια νέα βιβλιοθήκη λογισμικού εκμάθησης μηχανών που κυκλοφόρησε από την Google [8].

3.5.3. TensorFlow benefits

Το μεγαλύτερο πλεονέκτημα που προσφέρει το TensorFlow για την ανάπτυξη της μηχανικής μάθησης είναι η αφαίρεση. Αντί να ασχολείται με τις ιδιαιτερότητες των αλγορίθμων υλοποίησης ή να ανακαλύπτει τους κατάλληλους τρόπους για να αναστέλλει την έξοδο μιας λειτουργίας στην είσοδο άλλου, ο προγραμματιστής μπορεί να επικεντρωθεί στη συνολική λογική της εφαρμογής. Επίσης προσφέρει δυνατότητες στον εντοπισμό σφαλμάτων και ενδοσκόπησης στις εφαρμογές του. Η σουίτα απεικόνισης TensorBoard επιτρέπει την επιθεώρηση και την ελάφρυνση του τρόπου με τον οποίο λειτουργούν τα γραφήματα μέσω ενός διαδραστικού πίνακα ελέγχου.

3.5.4. General Machine Learining

Σε αυτή την παράγραφο θα γίνει μια σύντομη ανασκόπηση από μια μικρή σειρά γενικών βιβλιοθηκών μηχανικής μάθησης με χρονολογική σειρά αυτές είναι:

- MLC ++
- OpenCV2
- scikit-learn3
- Accord.NET
- Massive Online Analysis 6

- Mahout
- Spark MLlib

MLC ++ είναι μια βιβλιοθήκη λογισμικού που αναπτύχθηκε στη γλώσσα προγραμματισμού C ++ παρέχοντας αλγόριθμους μαζί με ένα framework σύγκρισης για μια σειρά εξόρυξης δεδομένων, στατιστική ανάλυση καθώς και τεχνικές αναγνώρισης προτύπων. Αρχικά αναπτύχθηκε στο Πανεπιστήμιο του Στάνφορντ το 1994 και πλέον ανήκει και συντηρείται από τη Silicon Graphics, Inc (SGI1).

OpenCV2 (Open Computer Vision) απελευθερώθηκε το 2000 από τους Bradski et al. . Στοχεύει πρωτίστως στην επίλυση μαθησιακών εργασιών στον τομέα της ορατότητας υπολογιστών και της αναγνώρισης εικόνων, συμπεριλαμβανομένης μιας συλλογής αλγορίθμων αναγνώρισης προσώπου, αναγνώρισης αντικειμένων, εξαγωγής μοντέλου 3D και άλλων σκοπών. Απελευθερώνεται με άδεια BSD και παρέχει διασυνδέσεις σε πολλές γλώσσες προγραμματισμού όπως C ++, Python και MATLAB.

Scikit-learn3 [7]. Το έργο scikit-learn αναπτύχθηκε αρχικά από τον David Cournapeu ως μέρος του προγράμματος Google Summer of Code4 το 2008. Πρόκειται για μια βιβλιοθήκη ανοιχτού κώδικα μάθησης που γράφεται στην Python, πάνω από τα πλαίσια NumPy, SciPy και matplotlib. Είναι χρήσιμο για μια μεγάλη τάξη μαθητειακών και μη εποπτευόμενων μαθησιακών προβλημάτων.

Accord.NET ξεχωρίζει από τα προαναφερθέντα παραδείγματα, καθώς γράφεται στη γλώσσα προγραμματισμού C # ("C Sharp"). Κυκλοφόρησε το 2008, αποτελείται όχι μόνο από μια ποικιλία αλγορίθμων μηχανικής μάθησης, αλλά και από μονάδες επεξεργασίας σήματος για αναγνώριση ομιλίας και εικόνας .

Massive Online Analysis (MOA) είναι ένα πλαίσιο ανοιχτού κώδικα για την online και offline ανάλυση των μαζικών, δυνητικά απεριόριστων ροών δεδομένων. Το MOA περιλαμβάνει μια ποικιλία εργαλείων για την ταξινόμηση, την παλινδρόμηση, τα συστήματα συνημμένων και άλλους κλάδους. Είναι γραμμένο στη γλώσσα προγραμματισμού Java και διατηρείται από το προσωπικό του Πανεπιστημίου Waikato της Νέας Ζηλανδίας. Σχεδιάστηκε το 2010.

Mahout, μέρος του Apache Software Foundation, είναι ένα περιβάλλον προγραμματισμού Java για εφαρμογές κλιμακούμενης μηχανικής μάθησης, που χτίζονται πάνω από την πλατφόρμα Apache Hadoop. Επιτρέπει την ανάλυση μεγάλων συνόλων δεδομένων που διανέμονται στο Hadoop Distributed File System (HDFS) χρησιμοποιώντας το παράδειγμα προγραμματισμού MapReduce. Το Mahout παρέχει αλγόριθμους μηχανικής μάθησης για ταξινόμηση, ομαδοποίηση και φιλτράρισμα. Το Pattern10 είναι μια ενότητα μάθησης μηχανών Python που συμπεριλαμβάνουμε στη λίστα μας εξαιτίας της πλούσιας σειράς εγκαταστάσεων εξόρυξης ιστού. Περιλαμβάνει όχι μόνο γενικούς αλγόριθμους μηχανικής μάθησης (π.χ. ομαδοποίηση, ταξινόμηση ή αναζήτηση πλησιέστερων γειτόνων) και μεθόδους επεξεργασίας φυσικών γλωσσών (π.χ. n-gram αναζήτηση ή ανάλυση συναισθημάτων), αλλά και ένα crawler ιστού που μπορεί για παράδειγμα να παραλάβει καταχωρήσεις Tweets ή Wikipedia , διευκολύνοντας την ταχεία ανάλυση των δεδομένων αυτών των πηγών. Δημοσιεύθηκε από το Πανεπιστήμιο της Αμβέρσας το 2012 και είναι ανοικτού κώδικα.

Spark MLlib είναι μια πλατφόρμα ανοιχτού κώδικα μάθησης και ανάλυσης δεδομένων που κυκλοφόρησε το 2015 και κατασκευάστηκε στην κορυφή του έργου Apache Spark, ενός συστήματος ταχείας συστοιχίας υπολογιστών. Παρόμοια με το Apache Mahout, υποστηρίζει την επεξεργασία κατανεμημένων συνόλων δεδομένων μεγάλης κλίμακας και την κατάρτιση μοντέλων μηχανικής μάθησης σε ένα σύμπλεγμα υλικών βασικών προϊόντων. Για αυτό, περιλαμβάνει ταξινόμηση, παλινδρόμηση, συστοιχία και άλλους αλγόριθμους μηχανικής μάθησης .

3.5.5. Deep Learning

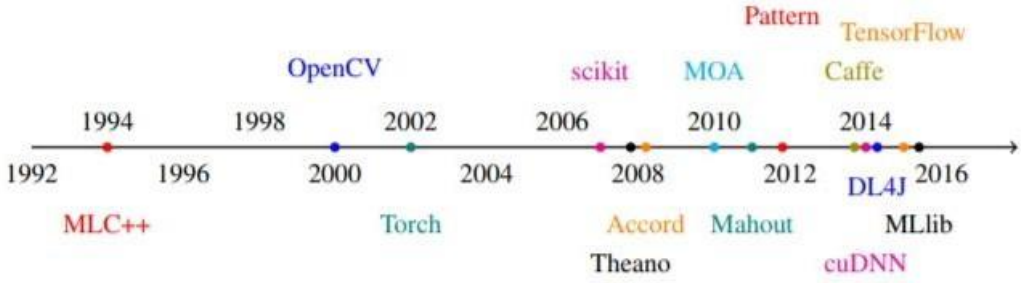
Σε αυτή την παράγραφο θα γίνει μια σύντομη περιγραφή των frameworks χρησιμοποιούνται και είναι ειδικά για την κατάρτιση μοντέλων «Deep Learning» αυτά είναι

- Torch
- Theano
- Caffe

- Deeplearning4J (DL4J)

Torch Το πρώτο και το παλαιότερο πλαίσιο στον κατάλογο που ενδείκνυται για την ανάπτυξη και την κατάρτιση βαθιών νευρωνικών δικτύων είναι το Torch, που κυκλοφόρησε ήδη το 2002 . Το torch αποτελούσε αρχικά μια καθαρή εφαρμογή και διεπαφή C ++. Σήμερα, ο πυρήνας του υλοποιείται στο C / CUDA ενώ εκθέτει μια διεπαφή στη γλώσσα scripting Lua. Γι 'αυτό, ο Torch χρησιμοποιεί έναν μεταγλωττιστή LuaJIT (just-in-time) για να συνδέσει τις ρουτίνες Lua με τις υποκείμενες υλοποιήσεις της C. Περιλαμβάνει, μεταξύ άλλων, αριθμητικές ρουτίνες βελτιστοποίησης, μοντέλα νευρωνικού δικτύου καθώς και αντικείμενα n-διάστατης συστοιχίας (tensor) γενικής χρήσης

Theano, που κυκλοφόρησε το 2008, είναι μια άλλη αξιοσημείωτη βιβλιοθήκη βαθιάς μάθησης. Κανονικά δεν πρόκειται για βιβλιοθήκη μηχανικής μάθησης. Αντίθετα είναι ένα framework το οποιο επιτρέπει στους χρήστες να δηλώσουν μαθηματικές εκφράσεις. Αυτά βελτιστοποιούνται στη συνέχεια, και τελικώς καταρτίζονται και τελικά εκτελούνται σε συσκευές CPU ή GPU.

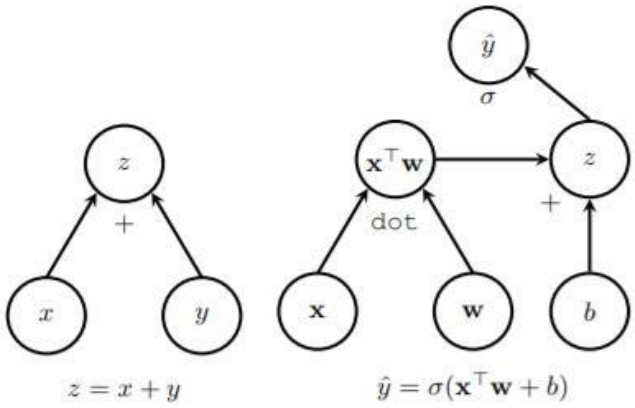


Το Caffe είναι μια βιβλιοθήκη βαθιάς μάθησης ανοιχτού κώδικα που διατηρείται από το Κέντρο Vision and Learning του Berkeley (BVLC). Το Caffe υλοποιείται στην C ++ και χρησιμοποιεί τα στρώματα νευρωνικών δικτύων ως βασικά υπολογιστικά δομικά στοιχεία (σε αντίθεση με το Theano και άλλα, όπου ο χρήστης πρέπει να καθορίσει μεμονωμένες μαθηματικές πράξεις που συνθέτουν στρώματα) Το Caffe είναι ιδιαίτερα κατάλληλο για την ανάπτυξη και την κατάρτιση συνελικτικών νευρωνικών δικτύων (CNNs ή ConvNets), που χρησιμοποιούνται εκτεταμένα στον τομέα της αναγνώρισης εικόνας. Ενώ τα παραπάνω πλαίσια μηχανικής μάθησης επέτρεψαν τον ορισμό μοντέλων βαθιάς μάθησης σε Python, MATLAB και Lua,

Deeplearning4J17 (DL4J) επιτρέπει επίσης στον προγραμματιστή της Java να δημιουργεί βαθιά νευρωνικά δίκτυα. Το DL4J περιλαμβάνει λειτουργίες για τη δημιουργία περιορισμένων μηχανών Boltzmann, συνθετικών και επαναλαμβανόμενων νευρωνικών δικτύων, δικτύων βαθιάς πίστης και άλλων τύπων μοντέλων βαθιάς μάθησης. Επιπλέον, το DL4J επιτρέπει την οριζόντια δυνατότητα κλιμάκωσης χρησιμοποιώντας κατανεμημένες πλατφόρμες υπολογιστών όπως Apache Hadoop ή Spark. Απελευθερώθηκε το 2014 από τον Adam Gibson με άδεια ανοικτού κώδικα Apache 2.0.

3.5.6. Computational Graph Architecture

Σε αυτή την ενότητα θα γίνει μια αναλυτική της αρχιτεκτονικής του TensorFlow Στο TensorFlow, οι αλγόριθμοι μηχανικής μάθησης αντιπροσωπεύονται ως υπολογιστικά γραφήματα. Στην παρακάτω εικόνα είναι παράδειγμα υπολογιστικού γραφήματος



Παραδείγματα υπολογιστικών γραφημάτων. Το αριστερό γράφημα εμφανίζει έναν πολύ απλό υπολογισμό, που αποτελείται από μια μόνο προσθήκη των δύο μεταβλητών εισόδου x και y. Σε αυτή την περίπτωση, το z είναι το αποτέλεσμα της λειτουργίας +. Το δεξί γράφημα δίδει ένα πιο πολύπλοκο παράδειγμα υπολογισμού μιας μεταβλητής logistic regression y σε για παράδειγμα το διάνυσμα x, το vector w καθώς επίσης και μια κλίμακα skalar b. Όπως φαίνεται στο γράφημα, το y είναι το αποτέλεσμα της σιγμοειδούς ή logistic συνάρτησης σ.

Ένα υπολογιστικό γράφημα απαρτίζεται από κόμβους οι όποιοι αντιπροσωπεύουν τις λειτουργιές που γίνονται και τις γραμμές οι οποίες αντιπροσωπεύουν τα δεδομένα που ρέουν από την μια λειτουργία στην άλλη. Αν μια μεταβλητή εξόδου z είναι το αποτέλεσμα της εφαρμογής μιας δυαδικής λειτουργίας σε δύο εισόδους x και y, τότε σχεδιάζετε κατευθυνόμενες γραμμές από τα x και y σε έναν κόμβο εξόδου που αντιπροσωπεύει το z και προσθέτουμε την κορυφή με μια ετικέτα που περιγράφει τον πραγματοποιημένο υπολογισμό. Τρείς είναι τα βασικά χαρακτηριστικά ενός TensorFlow μοντέλου

1. Operations
2. Tensors
3. Variables

1) Operations: Στο TensorFlow, οι κόμβοι αντιπροσωπεύουν λειτουργίες, οι οποίες με τη σειρά τους εκφράζουν το συνδυασμό ή μετασχηματισμό των δεδομένων που ρέουν μέσω του γραφήματος. Ενα operation μπορεί να έχει μηδενικές ή περισσότερες εισόδους και να παράγει μηδέν ή περισσότερες εξόδους. Ως εκ τούτου, μπορεί να αντιπροσωπεύει μια μαθηματική εξίσωση, μια μεταβλητή ή σταθερά, μια οδηγία ροής ελέγχου, μια λειτουργία εισόδου / εξόδου αρχείου ή ακόμα και μια θύρα επικοινωνίας δικτύου..

2) Tensors: Στο TensorFlow, τα άκρα αντιπροσωπεύουν τα δεδομένα που ρέουν από τη μία λειτουργία στην άλλη και αναφέρονται ως tensors. Ένα tensor είναι μια πολυδιάστατη συλλογή ομοιογενών τιμών με σταθερό, στατικό τύπο. Ο αριθμός των διαστάσεων ενός tensor ονομάζεται τάξη. Το σχήμα ενός tensor είναι η πλειάδα που περιγράφει το μέγεθός του, δηλαδή τον αριθμό των εξαρτημάτων σε κάθε διάσταση. Από την μαθηματική έννοια, ένα tensor είναι η γενίκευση των δισδιάστατων πινάκων, των μονοδιάστατων διανυσμάτων και επίσης των βαθμίδων, οι οποίοι είναι απλά οι tensors της τάξης μηδέν. Από την άποψη της υπολογιστικής γραφικής παράστασης, ένας τάσης μπορεί να θεωρηθεί ως συμβολική λαβή σε μία από τις εξόδους μιας λειτουργίας. Ο ίδιος ο tensor δεν διατηρεί ή αποθηκεύει τιμές στη μνήμη, αλλά παρέχει μόνο μια διεπαφή για την ανάκτηση της τιμής που αναφέρεται από τον tensor. Κατά τη δημιουργία μιας λειτουργίας στο περιβάλλον προγραμματισμού TensorFlow, όπως για την έκφραση x + y, επιστρέφει ένα αντικείμενο tensor. Αυτός ο tensor μπορεί στη συνέχεια να παρέχεται ως είσοδος σε άλλους υπολογισμούς, συνδέοντας έτσι την πηγή και τις.

3) Variables: Variables στο TensorFlow, είναι απλώς ειδικές λειτουργίες που μπορούν να προστεθούν σε ένα υπολογιστικό γράφημα. Οι Variables μπορούν να περιγραφούν ως επίμονες, μεταβλητές λαβές σε μνήμες μνήμης που αποθηκεύουν tensors. Ως εκ τούτου, οι μεταβλητές χαρακτηρίζονται από ένα ορισμένο σχήμα και έναν σταθερό τύπο. Κατά τη δημιουργία ενός μεταβλητού κόμβου για ένα γράφημα TensorFlow, είναι απαραίτητο να παρέχεται ένας tensorμε τον οποίο η μεταβλητή αρχικοποιείται κατά

την εκτέλεση γραφήματος. Η κατασκευή μιας μεταβλητής έχει ως αποτέλεσμα την προσθήκη τριών διακριτών κόμβων στο γράφημα

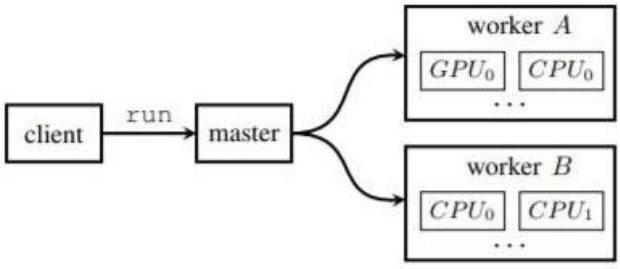
1) Ο πραγματικός μεταβλητός κόμβος, που κρατάει τη μεταβλητή κατάσταση
2) Μια πράξη που παράγει την αρχική τιμή, συχνά μια σταθερά
3) Μια λειτουργία αρχικοποίησης, η οποία εκχωρεί την αρχική τιμή στον μεταβλητό τανυστή μετά την αξιολόγηση του γραφήματος
4) Περίοδοι: Στο TensorFlow, η εκτέλεση των λειτουργιών και η αξιολόγηση των τανυστή μπορεί να γίνει μόνο σε ένα ειδικό περιβάλλον που αναφέρεται ως συνεδρία.

3.5.7. Execution Model

Το Προκειμένου ένα υπολογιστικό γράφημα να εκτελεστεί με βάση τα διαφορά στοιχεία που αναφέρθηκαν στην προηγούμενη ενότητα, το TensorFlow διαιρεί την διεργασία της εφαρμογής σε 4 φάσεις αυτές είναι:

- Client
- Master
- Workers
- Devices

Όταν ο Client ζητήσει την αξιολόγηση ενός γραφήματος TensorFlow μέσω ρουτίνας εκτέλεσης μιας συνόδου, αυτό το ερώτημα αποστέλλεται στη Master διαδικασία η οποία με τη σειρά της μεταβιβάζει την εργασία σε μία ή περισσότερες διεργασίες Workers και συντονίζει την εκτέλεση τους. Στην συνέχεια κάθε διεργασία Worker είναι υπεύθυνη για την επίβλεψη μιας ή περισσότερων devices, οι οποίες είναι οι φυσικές μονάδες επεξεργασίας για τις οποίες υλοποιούνται οι πυρήνες μιας λειτουργίας. Μέσα σε αυτό το μοντέλο, υπάρχουν δύο βαθμοί κλιμακώσεως. Ο πρώτος βαθμός αφορά την κλιμάκωση του αριθμού των μηχανών στις οποίες εκτελείται ένα γράφημα. Ο δεύτερος βαθμός αναφέρεται στο γεγονός ότι σε κάθε μηχανή μπορεί να υπάρχουν περισσότερες από μία devices, όπως για παράδειγμα πέντε ανεξάρτητες μονάδες GPU ή / και τρεις CPU. Στο παρακάτω σχήμα παρουσιάζετε μια εικόνα αυτό του μοντέλου



Οπτικοποίηση των διαφορετικών παραγόντων εκτέλεσης σε διαμόρφωση υλικού πολλαπλών συσκευών, πολλαπλών συσκευών..

1) Devices: Πρόκειται για τις μικρότερες και πιο βασικές οντότητες στο μοντέλο εκτέλεσης TensorFlow. Όλοι οι κόμβοι στο γράφημα, δηλαδή ο πυρήνας κάθε λειτουργίας, πρέπει τελικά να αντιστοιχιστούν σε μια διαθέσιμη device που θα εκτελεστεί. Στην πράξη, ένα device θα είναι συχνότερα είτε CPU είτε GPU. Ωστόσο, το TensorFlow υποστηρίζει την εγγραφή άλλων τύπων φυσικών μονάδων εκτέλεσης από το χρήστη. Επομένως, είναι ευνόητα εύκολο να ενσωματωθούν νέες τάξεις συσκευών, καθώς αναδύεται νέο υλικό.

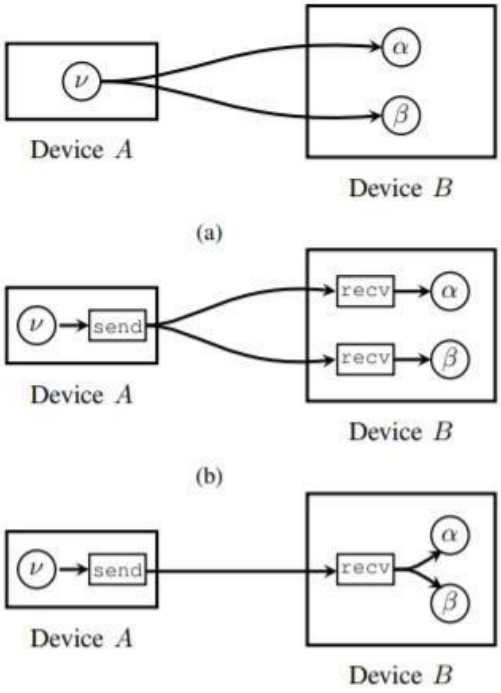
2) Placement Algorithm: Προκειμένου να προσδιοριστεί για το ποιοι κόμβοι θα τοποθετηθούν σε ποια συσκευή χρησιμοποιείτε ένας αλγόριθμος τοποθέτησης ο οποίος προσομοιώνει την εκτέλεση του υπολογιστικού γραφήματος και διασχίζει τους κόμβους από τους tensor εισόδου και tensor εξόδου. Αυτό το μοντέλο κόστους λαμβάνει υπόψη τέσσερα τεμάχια πληροφοριών για να προσδιορίσει τη βέλτιστη συσκευή d = arg mind ∈ D Cv (d) στην οποία θα τοποθετηθεί ο κόμβος κατά την εκτέλεση:

1) Το αν υπάρχει ή όχι κάποια εφαρμογή (πυρήνας) για έναν κόμβο στη συγκεκριμένη συσκευή.
2) Εκτιμήσεις του μεγέθους (σε bytes) για τους tensors εισόδου και εξόδου ενός κόμβου.
3) Ο αναμενόμενος χρόνος εκτέλεσης του πυρήνα στη συσκευή.
4) Ένας ευρετικός για το κόστος της μετάδοσης.

Για να εξασφαλίστει η μέγιστη απόδοση ενός μοντέλου εκτέλεσης TensorFlow προστίθεται ενας αριθμός βελτιστοποιήσεων αυτού του είδους οι βελτιστοποιήσεις αναφέρονται ως εξής:

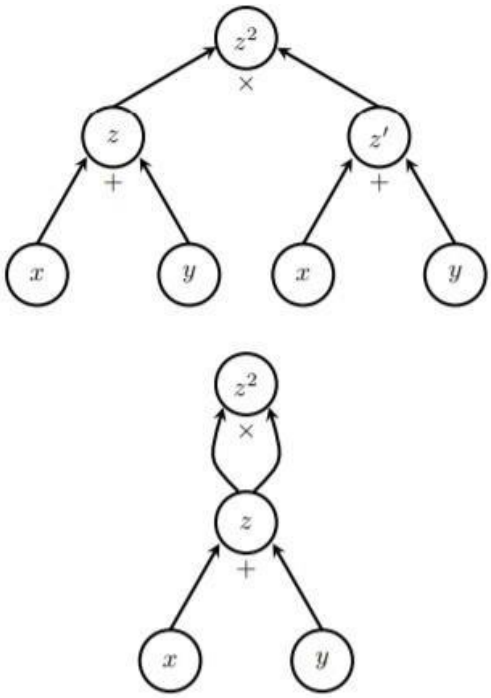
1) Common Subgraph Elimination
2) Scheduling
3) Lossy Compression

Common Subgraph Elimination: Μια βελτιστοποίηση που εκτελείται από πολλούς σύγχρονους compliers είναι η κοινή εξάλειψη υποεπιλογής, με την οποία ένας μεταγλωττιστής μπορεί να αντικαταστήσει τον υπολογισμό μιας ίδιας αξίας δύο ή περισσότερες φορές από μία μόνο περίπτωση αυτού του υπολογισμού ο αποτέλεσμα αποθηκεύεται τότε σε μια προσωρινή μεταβλητή και επαναχρησιμοποιείται όπου προηγουμένως υπολογίστηκε εκ νέου. Ομοίως, σε ένα γράφημα TensorFlow, μπορεί να προκύψει ότι η ίδια λειτουργία εκτελείται ταυτόχρονα με πανομοιότυπες εισόδους



Scheduling: Μια απλή αλλά ισχυρή βελτιστοποίηση είναι η προγραμματιζόμενη εκτέλεση κόμβων όσο το δυνατόν αργότερα. Η εξασφάλιση ότι τα αποτελέσματα των λειτουργιών παραμένουν στη μνήμη μόνο για το ελάχιστο απαιτούμενο χρονικό διάστημα μειώνει την κατανάλωση μέγιστης μνήμης και έτσι μπορεί να βελτιώσει σημαντικά τη συνολική απόδοση του συστήματος

Lossy Compression: Ένας από τους πρωταρχικούς στόχους πολλών αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται για ταξινόμηση, αναγνώριση ή άλλες εργασίες είναι η δημιουργία ισχυρών μοντέλων. Με ισχυρή εννοούμε ότι ένα βέλτιστα καταρτισμένο μοντέλο δεν θα πρέπει να αλλάξει ιδανικά την απόκριση του εάν τροφοδοτηθεί πρώτα ένα σήμα και στη συνέχεια να τροποποιήσει αυτό το σήμα. Για το λόγο αυτό, μια άλλη βελτιστοποίηση που εκτελείται από το TensorFlow είναι η εσωτερική προσθήκη μετατροπικών κόμβων στο υπολογιστικό γράφημα, το οποίο μετατρέπει τέτοιες υψηλές τιμές ακριβείας 32-bit κινητής υποδιαστολής σε παραμορφωμένες αναπαραστάσεις 16 bit όταν επικοινωνούν μεταξύ συσκευών και σε μηχανές



Ένα παράδειγμα για το πώς χρησιμοποιείται η κοινή εξάλειψη υπογράμματος για να μετασχηματιστούν οι εξισώσεις z = x + y, z 0 = x + y, z 2 = z • z 0 σε δύο μόνο εξισώσεις z = x + y και z 2 = z • z. Αυτός ο υπολογισμός θα μπορούσε θεωρητικά να βελτιστοποιηθεί περαιτέρω σε μια τετραγωνική λειτουργία που απαιτεί μόνο μία είσοδο (μειώνοντας έτσι το κόστος της κίνησης των δεδομένων), αν και δεν είναι γνωστό εάν το TensorFlow χρησιμοποιεί μια τέτοια δευτερογενή κανονικοποίηση.

4. Twitter`

4.1. Introduction

Το Twitter είναι ένα online δίκτυο κοινωνικής πληροφόρησης που ξεκίνησε τον Ιούλιο του 2006. Έως το 2012, ο αριθμός των χρηστών του Twitter έχει αυξηθεί σε πάνω από 140 εκατομμύρια. Σε αντίθεση με πολλά άλλα online κοινωνικά δίκτυα, οι σχέσεις χρηστών στο δίκτυο του Twitter μπορούν να είναι κοινωνικές ή ενημερωτικές ή και οι δύο. Καθώς οι χρήστες Twitter παράγουν περισσότερα από 300M tweets κάθε μέρα, αυτοί οι χρήστες είναι επίσης συγκλονισμένοι από το τεράστιο όγκο των διαθέσιμων πληροφοριών και τον τεράστιο αριθμό των ατόμων με τα οποία μπορούν να αλληλεπιδρούν. Για να ξεπεραστεί το παραπάνω πρόβλημα υπερφόρτωσης πληροφοριών, μπορούν να εισαχθούν συστήματα συστημένων για να βοηθήσουν τους χρήστες να κάνουν την κατάλληλη επιλογή. Ενώ ορισμένα από αυτά έχουν ήδη αναπτυχθεί μέχρι σήμερα, τα περισσότερα από αυτά εξακολουθούν να μελετώνται ως ερευνητικά προγράμματα σε πανεπιστήμια και βιομηχανικά εργαστήρια. Αυτά τα ερευνητικά προγράμματα συνήθως απευθύνονται σε μεμονωμένες εργασίες συστάσεων. Επί του παρόντος, δεν υπάρχει ολοκληρωμένη έρευνα για τη σφαίρα της σύστασης στο Twitter για την κατηγοριοποίηση των υφιστάμενων έργων καθώς και για τον εντοπισμό περιοχών που πρέπει να μελετηθούν περαιτέρω.

4.2. Collaborative Filtering

Συνεργασία Φιλτράρισμα είναι η διαδικασία του φιλτραρίσματος ή της αξιολόγησης των αντικειμένων χρησιμοποιώντας τις απόψεις των άλλων ανθρώπων και είναι η κύρια μέθοδος σύστασης του Twitter που θα αναλυθεί. Αυτή η τεχνική εχει 2 προσεγγίσεις

- Ειδική
- Γενική

Ειδική: Αυτού του είδους η προσέγγιση είναι μια μέθοδος αυτόματης πρόβλεψης σχετικά με πιθανά ενδιαφέροντα ενός χρήστη συλλέγοντας πληροφορίες και προτιμήσεις από πολλαπλούς χρήστες . Για παραδείγματα αν ένας χρήστης Α έχει την ίδια άποψη επί ενός θέματος με έναν χρήστη Β τότε είναι πιθανό ο χρήστης Α και ο χρήστης Β να μοιράζονται την ίδια άποψη σε 1 διαφορετικό ζήτημα. Όποτε αυτό οδηγεί στο εξής: Αν ο χρήστης Β ορίσει μια λίστα με ταινίες, προτιμήσεις σε φαγητό κ.τ.λ. πιθανόν αυτά τα ενδιαφέροντα να ανταποκρίνονται και στον Α.

Γενική: Αυτού του είδους η προσέγγιση είναι η διαδικασία φιλτραρίσματος για πληροφορίες ή πρότυπα με τη χρήση τεχνικών που συνεπάγονται τη συνεργασία μεταξύ πολλών παραγόντων, σημείων θέασης, πηγών δεδομένων κλπ. Οι εφαρμογές φιλτραρίσματος συνεργασίας συνεπάγονται τυπικά πολύ μεγάλα σύνολα δεδομένων. Οι μέθοδοι φιλικής φιλτραρίσματος έχουν εφαρμοστεί σε πολλά διαφορετικά είδη δεδομένων, όπως: δεδομένα ανίχνευσης και παρακολούθησης, όπως στην εξερεύνηση ορυκτών, στην περιβαλλοντικαή ανίχνευση σε μεγάλες περιοχές ή σε πολλαπλούς αισθητήρες. οικονομικά δεδομένα, όπως τα ιδρύματα χρηματοπιστωτικών υπηρεσιών που ενσωματώνουν πολλές χρηματοοικονομικές πηγές · ή στο ηλεκτρονικό εμπόριο και τις εφαρμογές ιστού όπου επικεντρώνεται στα δεδομένα των χρηστών κλπ

4.3. Uses For Collaborative Filtering

Μέχρι στιγμής έχει αναφερθεί σε γενικές γραμμές τι είναι το συνεργατικό Φιλτράρισμα. Σε αυτή την ενότητα θα γίνει η αναφορά στο τι εξυπηρετεί και ποιες οι εφαρμογές αυτή της τεχνικής καθώς και ειδή αλγορίθμων συνεργατικού φιλτραρίσματος. Οπότε πιθανές ανάγκες αυτής της τεχνικής είναι η εξής:

Εύρεση νέων στοιχείων που αρέσουν σε 1 χρήστη. Διότι ως γνωστόν σε 1 κόσμο υπερφόρτωσης πληροφοριών δεν είναι σε θέση ο χρήστης να δει όλες τι επιλογές λόγο πιθανού περιορισμένου χρόνου για αυτό και ζήτα παρουσίαση πιθανών επίλογων

- Συμβουλές σχετικά με ένα συγκεκριμένο θέμα.
- Δυνατότητα σύστασης για την εύρεση ενός χρήστη ή μερικών χρηστών.
- Βοηθήστε την ομάδα μας να βρει κάτι νέο που θα θέλαμε.

Collaborative Filtering Algorythms

Σε αυτή την ενότητα θα εξεταστούν ορισμένοι από τους πιο γνωστούς αλγορίθμους συνεργατικού φιλτραρίσματος. Οι αλγόριθμοι χωρίζονται σε 2 κατηγορίες

- Αλγόριθμοι μνήμης που απαιτούν όλες τις αξιολογήσεις, τα στοιχεία και τους χρήστες να αποθηκεύονται σε αλγόριθμους μνήμης
- Μοντέλου που περιοδικά δημιουργούν μια σύνοψη των υποδειγμάτων βαθμολόγησης εκτός σύνδεσης
- Τα μοντέλα που βασίζονται στη μνήμη δεν κλιμακώνονται καλά για εφαρμογές σε πραγματικό κόσμο. Έτσι, σχεδόν όλοι οι πρακτικοί αλγόριθμοι χρησιμοποιούν κάποια μορφή προ-υπολογισμού για να μειώσουν την πολυπλοκότητα του χρόνου

Στην συνέχεια αυτής της ενότητας η αλγόριθμοι που θα αναλυθούν είναι οι

- Non Probalistic Algorythms
- Probalistic Algorythms

Non Probalistic Algorythms Οι πιο γνωστοί αλγόριθμοι CF είναι οι πλησιέστεροι αλγόριθμοι πλησίον και αφορούν με βάση το χρήστη και με βάση το αντικείμενο.

User-Based Nearest Neighbor Algorithms Οι πρώτοι αλγόριθμοι δημιούργησαν προβλέψεις για χρήστες βάσει αξιολογήσεων από παρόμοιους χρήστες. Ονομάζουμε αυτούς τους παρόμοιους χρήστες γείτονες. Αν ένας χρήστης n είναι παρόμοιος με έναν χρήστη u, λέμε ότι n είναι ένας γείτονας του u. Οι αλγόριθμοι με βάση το χρήστη δημιουργούν μια πρόβλεψη για ένα στοιχείο i αναλύοντας τις αξιολογήσεις για το i από τους χρήστες της γειτονιάς του u. Ο υπολογισμός των βαθμολογιών των γειτόνων γίνεται με βάση τον ακόλουθο τύπο:

$$pred(u, i) = \frac{\sum_{n \in neighbors(u)} Rn}{number\ of\ neighbors}$$

Η εξίσωση δεν θεωρείτε αξιόπιστη διότι δεν λαμβάνει υπόψη το γεγονός ότι ορισμένα μέλη της γειτονιάς u έχουν ένα υψηλότερο επίπεδο ομοιότητας με το u από άλλα. Έτσι, αν το userSim (u, n) είναι ένα μέτρο της ομοιότητας μεταξύ του χρήστη στόχου u και ενός γειτονικού n, μια πρόβλεψη μπορεί να δοθεί από την εξίσωση .

$$pred(u,i) = \sum_{n \in neighbors(u)} UserSim(u,n) * Rni$$

Δυστυχώς, αν οι ομοιότητες των γειτόνων δεν προσθέτουν μέχρι ένα, αυτή η πρόβλεψη θα είναι εσφαλμένη. Σύμφωνα με την εξίσωση, ομαλοποιεί την πρόβλεψη διαιρώντας με το άθροισμα των ομοιότητες των γειτόνων.

$$\square\square\square\square(\square,\square) = \frac{\sum_{\square \in \square\square\square\square\square} \square\square\square\square(\square)}{\square\square\square\square\square\square\square(\square,\square) * \square\square\square}$$
$$\sum_{\square \in \square\square\square\square\square} \square\square\square\square\square\square\square(\square,\square)$$

Βεβαια σημαντικό ρολο για μια αξιολόγηση οσο και αν φαινεται περιεργο είναι η ψυχολογική κατάσταση ενός χρήστη δηλαδη καποιος «αισιοδοξος» χρήστης εχει την τάση να βαθμολογει με 4-5 αστερια σε σχέση με καποιο λιγοτερο ο οποιος θα βαθμολογεί 3-5 αστέρια Για να αντισταθμιστούν οι μεταβολές της κλίμακας αξιολόγησης, ο μέσος όρος της εξίσωσης προσαρμόζεται στις μέσες βαθμολογίες των χρηστών.

$$\square\square\square\square(\square,\square) = \square\square + \frac{\sum_{\square \in \square\square\square\square\square} \square\square\square\square\square\square\square(\square,\square) * (\square\square\square - \square\square)}{\sum_{\square \in \square\square\square\square\square} \square\square\square\square\square\square\square(\square,\square)}$$

Το σύστημα GroupLens για ομάδες συζήτησης Usenet, ένα από τα πρώτα συστήματα CF, ορίζει το userSim () στην εξίσωση 4 χρησιμοποιώντας τη συσχέτιση Pearson [51]. Ο συντελεστής συσχέτισης Pearson υπολογίζεται συγκρίνοντας τις βαθμολογίες για όλα τα στοιχεία που αξιολογούνται τόσο από τον χρήστη-στόχο όσο και από τον γείτονα (π.χ. Η εξίσωση 5 δίνει τον τύπο για τη συσχέτιση Pearson μεταξύ του χρήστη u και του γειτονικού n, όπου CRu, n. υποδηλώνει το σύνολο των αποκομμένων στοιχείων μεταξύ u και n.

$$UserSim(u,n) = \frac{\sum_{i \in CRu,n}(Rui - Ru)(Rni - Rn)}{\sqrt{\sum_{i \in CRu,n}(Rui - Ru)^2}\ \sqrt{\sum_{i \in CRu,n}(Rni - Rn)^2}}$$

Practical Challenges of User-Based Algorithms Ο αλγόριθμος πλησιέστερων γειτόνων που βασίζεται στον χρήστη καταγράφει τον τρόπο με τον οποίο λειτουργεί η ανταλλαγή συστάσεων από στόμα σε στόμα και μπορεί να ανιχνεύσει σύνθετα μοτίβα που έχουν αρκετούς χρήστες; Ωστόσο καλείται να αντιμετωπίσει διάφορα προβλήματα που αφορούν στην αλίωση των προβλέψεων. Αυτά τα προβλήματα μπορει να είναι αραία δηλαδη να μην υπάρχει αρκετή βαθμολόγηση από χρήστες ή λάθος συσχετίσεις. Ένα άλλο πρόβλημα με τη συσχέτιση του Pearson είναι ότι δεν καταφέρνει να ενσωματώσει συμφωνία για μια ταινία στο σύνολο του πληθυσμού. Ο συσχετισμός Pearson δεν καταγράφει αυτή τη διάκριση. Ορισμένοι αλγόριθμοι με βάση το χρήστη συμβάλλουν στη συμφωνία παγκόσμιου στοιχείου.

συμπεριλαμβάνοντας τα βάρη αντιστρόφως ανάλογα με τη δημοτικότητα ενός στοιχείου κατά τον υπολογισμό των συσχετισμών των χρηστών. Σε γενικές γραμμές τα δεδομένα βαθμολογίας είναι

επιρρεπή σε λανθασμένες συσχετίσεις. Για παράδειγμα αν κάποιοι χρήστες έχουν ελάχιστα κοινά σημεία τότε οι αξιολογήσεις να ταιριάζουν ακριβώς είναι πολύ σύνηθες φαινόμενο. Έστω κάποιο τρόπο δεν αντιμετωπιστούν αυτές οι ομοιότητες τότε ένας χρήστης μπορεί να είναι κυρίαρχος στο κομμάτι της γειτονιάς του. Επίσης 2 χρήστες οι όποιοι συμφωνούν για μια ταινία που γενικώς είναι αγαπητή στο εύρη κοινό σαν γεγονός είναι λιγότερο σημαντικό από μια αμφιλεγόμενη ταινία. Ορισμένοι αλγόριθμοι με βάση το χρήστη συμβάλλουν στη συμφωνία παγκόσμιου στοιχείου, συμπεριλαμβάνοντας τα βάρη αντιστρόφως ανάλογα με τη δημοτικότητα ενός στοιχείου κατά τον υπολογισμό των συσχετισμών των χρηστών. Υπάρχουν 2 τρόποι συσχέτισης δεδομένων αυτοί είναι

• Υποδειγματοληψία - Στη δειγματοληψία, επιλέγεται ένα υποσύνολο χρηστών πριν από τον υπολογισμό πρόβλεψης. Ο χρόνος υπολογισμού της γειτονιάς παραμένει σταθερός και έχουν προταθεί προγράμματα για έξυπνη επιλογή των γειτόνων προκειμένου να επιτευχθεί σχεδόν η ίδια ακρίβεια.

• Clustering - Οι αλγόριθμοι clustering είναι υπεύθυνοι για την ομαδοποίηση ενός συνόλου αντικείμενων με τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (που ονομάζεται σύμπλεγμα) να είναι περισσότερο παρόμοια μεταξύ τους παρά με εκείνα σε άλλες ομάδες. Πρόκειται για ένα βασικό καθήκον διερευνητικής εξόρυξης δεδομένων και μιας κοινής τεχνικής για την ανάλυση στατιστικών δεδομένων που χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης, της αναγνώρισης προτύπων, της ανάλυσης εικόνας, της ανάκτησης πληροφοριών, της βιοπληροφορικής, της συμπίεσης δεδομένων και των γραφικών υπολογιστών.

.

Item-Based Nearest Neighbor Algorithms

Στην προηγούμενη ενότητα έγινε αναφορά στους αλγορίθμους με βάση το χρήστη εδώ θα γίνει ανάλυση με βάση τα αντικείμενα. Ενώ οι αλγόριθμοι που βασίζονται σε χρήστες παράγουν προβλέψεις βασισμένες σε ομοιότητες μεταξύ χρηστών, οι αλγόριθμοι που βασίζονται σε στοιχεία δημιουργούν προβλέψεις βασισμένες σε ομοιότητες μεταξύ των στοιχείων. Προκειμένου να δουλέψει σωστά αύτη η τεχνική πρέπει να βασιστεί σε δεδομένα αξιολογήσεων ενός χρήστη για παρόμοιε αντικείμενα. Με βάση τον πίνακα στην προσπάθεια πρόβλεψης της βαθμολογίας για το Diary για τον χρήστη 4 παρατηρείτε ότι οι βαθμολογίες για την Diary είναι πολύ παρόμοιες με τις βαθμολογίες για το "Titanic", αλλά όχι τόσο παρόμοιες με τις βαθμολογίες για το "Kill Bill v.2." Επειδή το "Titanic" είναι παρόμοια με την "Diary", ίσως υποθέσουμε ότι η βαθμολογία για το "Titanic" είναι πιο σημαντική. Όποτε ίσως μια θεωρητικά σωστή σχέση που περιέγραφε μια πρόβλεψη είναι η εξής: 0,25 * 3 + 0,75 * 4 = 3,75

	Kill Bill v.2	Diary	Titanic
	5	4	4
	4	2	4
	4	3	4
	3	X	5

Η παρακάτω εξίσωση περιγράφει την πρόβλεψη για έναν χρήστη U για αντικείμενα I.

$$\square\square\square\square(\square,\square) = \frac{\sum_{\square\in\square\square\square\square\square\square\square\square\square(\square)}\square\square\square\square\square\square(\square,\square)*\ \square\square\square}{\sum_{\square\in\square\square\square\square\square\square\square\square\square(\square)}\square\square\square\square\square\square(\square,\square)}$$

Η εξίσωση 7 δίνει τον τύπο για ομοιότητα προσαρμοσμένου-συνημιτόνου, όπου RBi, j δηλώνει το σύνολο των χρηστών που έχουν αξιολογήσει τόσο το στοιχείο i όσο και το στοιχείο j.

$$\text{Pearson}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}}$$

Η μόνη διαφορά από τη συσχέτιση Pearson είναι ότι η μέση προσαρμογή γίνεται σε σχέση με τον χρήστη και όχι με το στοιχείο. Όπως και στην συσχέτιση χρήστη Pearson, η τιμή συσχέτισης κυμαίνεται από -1,0 έως 1,0. Υπάρχουν στοιχεία που δείχνουν ότι οι πλησιέστεροι αλγόριθμοι πλησίον γειτονικών αντικειμένων είναι ακριβέστεροι στην πρόβλεψη των αξιολογήσεων από τους συναδέλφους με βάση το χρήστη.

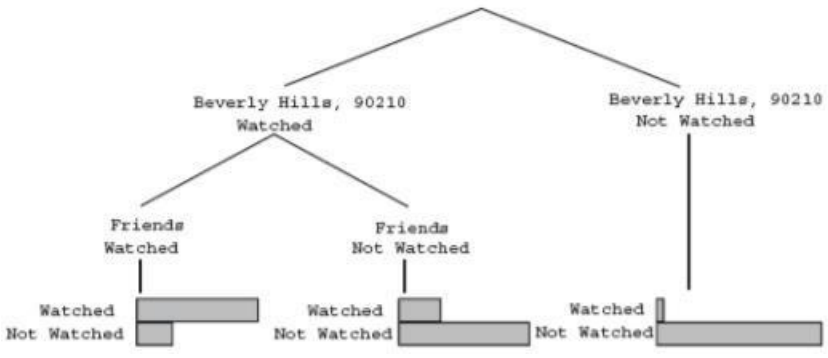
Practical Challenges in Item-Based Algorithms Challenges in Item-Based Algorithms Θεωρητικά, ο μέγεθος του μοντέλου θα μπορούσε να είναι τόσο μεγάλο όσο το τετράγωνο του αριθμού των στοιχείων. Αν με καποιο τροπο μειωθεί το μέγεθος του μοντέλου τοτε επιτυχνανεται λιγοτερη χρηση μνήμης καθώς και αυξηση της αποδοτικότητας του CPU. Δυστηχώς όμως το μοντέλο «χάνει» σε αξιοπιστια οσον αφορα το γεγονός της σωστής προβλεψης. καθώς τα στοιχεία που συσχετίζονται με τις αξιολογήσεις του χρήστη ενδέχεται να μην περιέχουν το στοιχείο στόχο

4.4. Probalistic Algorithms

Σε γενικές γραμμές, οι πιθανολογικοί αλγόριθμοι προσπαθούν να εκμεταλλευτούν καλά κατανοητούς φορμαλισμούς πιθανότητας. Οι περισσότεροι πιθανολογικοί αλγόριθμοι υπολογίζουν την πιθανότητα με δεδομένο ένα χρήστη u και ένα εκτιμώμενο στοιχείο i, έχει αποδώσει το στοιχείο σε r: p (r | u, i). Η παρακάτω εξίσωση δίνει τον τύπο για την βαθμολογία ενός χρήστη U για 1 αντικείμενο I

$$E(r|u, i) = \sum_r r * p(r|u, i)$$

το πιο δημοφιλές πιθανολογικό πλαίσιο περιλαμβάνει μοντέλα Bayesian δικτύου που εξάγουν πιθανολογικές εξαρτήσεις μεταξύ χρηστών ή αντικειμένων. Μερικοί από τους παλαιότερους πιθανολογικούς αλγορίθμους προτάθηκαν από τους Breese et al., Οι οποίοι περιγράφουν μια μέθοδο για την παραγωγή και εφαρμογή Bayesian δικτύων χρησιμοποιώντας δέντρα αποφάσεων για να αντιπροσωπεύουν συμπαγώς πίνακες πιθανοτήτων [9]. Το παρακάτω σχήμα Για παράδειγμα, το Σχ. 3 δείχνει ότι οι χρήστες που δεν παρακολουθούν "Beverly Hills, 90210" είναι πολύ πιθανό να μην παρακολουθήσουν το Melrose Place. Για κάθε συνιστώμενο στοιχείο κατασκευάζεται ξεχωριστό δέντρο. Ο κλάδος που επιλέγεται σε έναν κόμβο του δέντρου εξαρτάται από την αξιολόγηση του χρήστη (ή την έλλειψη βαθμολογίας) για ένα συγκεκριμένο στοιχείο.



Ένα δέντρο απόφασης για έναν χρήστη ο οποίος βλέπει την σειρά « Melrose Place » σύμφωνα με το αν βλέπει ή όχι την σειρά «Friends» και/ή την σειρά «Beverly hills

Οι πιθανότητες για το Watched vs. No Watched εμφανίζονται στα φύλλα του δέντρου και εξαρτώνται από την κατάσταση προβολής των προγραμμάτων στους γονικούς κόμβους. Έχει επίσης υπάρξει μια καλή εργασία για την ανάπτυξη πιθανοτήτων τεχνικών συσσωμάτωσης / μείωσης διαστάσεωνProbabilities Η εξίσωση 9 δίνει τον τύπο για τον υπολογισμό της πιθανότητας του χρήστη u τιμή στοιχείου ι τιμής r.

$$p(r|u,i) = \sum_{z} p(r|i,z)p(z|u)$$

Η αντίστοιχη πρόβλεψη είναι η προσδοκία της τιμής αξιολόγησης.

$$E(r|u,i) = \sum_{r}(r\sum_{z} p(r|z,i)p(z|u))$$

4.5. Over Arching Practical Concerns

Ανεξάρτητα από την επιλογή του αλγορίθμου, τα συστήματα CF του πραγματικού κόσμου πρέπει να αντιμετωπίσουν αρκετά προβλήματα:

Απόρριψη οντοτήτων με σπάνια αξιοπιστία - Οι αλγόριθμοι συχνά ενσωματώνουν μόνο δεδομένα με τιμές μεγαλύτερες από k. Σε έναν αλγόριθμο που βασίζεται στον χρήστη, για παράδειγμα, θα απορρίψαμε τους γείτονες με λιγότερες από τις συν-βαθμολογίες με τον χρήστη-στόχο..

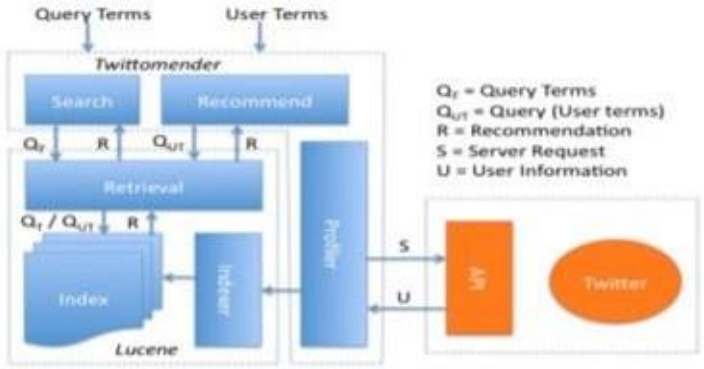
Προσαρμογή των υπολογισμών για σπάνια εκτιμημένες οντότητες. Αυτή η τεχνική προσαρμόζει τους υπολογισμούς για οντότητες με σπάνια βαθμολογία τραβώντας τις πιο κοντά σε έναν αναμενόμενο μέσο. Για παράδειγμα, οι ομοιότητες Pearson για χρήστες με λίγες συν-βαθμολογίες μπορεί να προσαρμοστούν πιο κοντά στο 0.

Ενσωμάτωση μιας προηγούμενης πεποίθησης – Σε αυτή την περίπτωση η αποφυγή της να παραμόρφωση είναι δυνατή με την ενσωμάτωση τεχνητών σημείων δεδομένων που ταιριάζουν με την αναμενόμενη κατανομή.

4.6. Twittomender

Το Twitter προσφέρει μόνο πολύ στοιχειώδεις υπηρεσίες αναζήτησης για να βοηθήσει τους χρήστες να βρουν νέους χρήστες που θα ακολουθήσουν. Αυτό αντιπροσωπεύει μια σημαντική ευκαιρία για τα συστήματά τους. Σε αυτή την ενότητα θα περιγραφεί το σύστημα Twittomender recommender, εστιάζοντας στην αρχιτεκτονική του συστήματος, τον τρόπο με τον οποίο οι χρήστες είναι διαμορφωμένοι και πώς αυτά τα προφίλ μπορούν να χρησιμοποιηθούν για να υποδείξουν ενδιαφέροντες χρήστες που θα ακολουθήσουν.

System Architecture Το σύστημα Twittomender έχει αναπτυχθεί ως υπηρεσία Web service. βλ. Εικόνα. Χρησιμοποιεί το Twitter API 3 για να δημιουργήσει και να διατηρήσει μια βάση δεδομένων των χρηστών του Twitter, των tweets τους, των followers τους και των ακολούθων.



Σχήμα αρχιτεκτονικής του Twittomender

Το twittomender εχει αναπτυχθεί με τις εξής λειτουργίες:

1. User Search — Σε αυτή την λειτουργία ένας χρήστης μπορεί μέσω μιας αναζήτησης να ανακτήσει λίστες με σχετικούς χρήστες του Twitter . Το σχήμα 2 δείχνει το αποτέλεσμα μιας αναζήτησης ατόμων που σχετίζονται με το ερώτημα "κοινωνική αναζήτηση". Κάθε αποτέλεσμα είναι ένας χρήστης του Twitter που έχει καταχωριστεί από το Twittomender, μαζί με σχετικές πληροφορίες, όπως το όνομα χρήστη, η περιγραφή, οι δημοφιλείς όροι από τα πρόσφατα tweets τους και τα πιο πρόσφατα tweets τους
2. User Recommendation — Σε αυτόν τον τρόπο, το προφίλ Twitter του χρήστη ενεργεί ως μια μορφή ερωτήματος για τη δημιουργία προληπτικών συστάσεων χρηστών που θα ακολουθήσουν. Στο σχήμα 3 βλέπουμε τη λίστα συστάσεων που δημιουργήθηκε για έναν συγκεκριμένο χρήστη του Twitter. Το πλαίσιο ερωτήματος εμφανίζει τους όρους που εξάγονται από το προφίλ του χρήστη ως ερώτημα και η λίστα αποτελεσμάτων είναι ένα σύνολο χρηστών που κρίνεται ότι είναι σχετικοί με τον χρήστη-στόχο. Για να χρησιμοποιήσετε το Twittomender, οι χρήστες πρέπει να συγχρονίσουν τον υπάρχοντα λογαριασμό τους Twitter με το Twittomender, ώστε να μπορεί να δημιουργηθεί ένα κατάλληλο προφίλ.

Profiling Users on Twitter Συμφώνα με την ερευνά των John Hannon, Mike Bennett, Barry Smyth.Η λειτουργία της αναζήτησης και του συστήματος που παρέχονται από το twittomender βασίζεται στης διαθεσημότητα ενός χρήστη που αντικατοπτρίζουν τα ενδιαφέροντα των χρήστων. Σε αυτή την ενότητα θα γίνει αναφορα στις διάφορες εξισώσεις που απαρτίζουν το σύστημα του twittomender. Έστω ότι έχουμε σαν μια οντότητα τα πιο πρόσφατα tweets που έχουν γίνει από 1 χρήστη από την εξίσωση που χρησιμοποιούν οι John Hannon, Mike Bennett, Barry Smyth θεωρούν όπως φαίνεται από την Εξίσωση , ένα χρήστη-στόχο, UT, tweets (UT) να είναι το σύνολο των πρόσφατων tweets για UT. υποθέτοντας ότι τα tweets (UT) είναι τα 100 πιο πρόσφατα tweets του χρήστη. Με αυτό τον τρόπο τα tweets (UT) παρέχουν τη βάση για μια προσέγγιση βασισμένη στο περιεχόμενο για τη δημιουργία προφίλ χρηστών, προφανώς υπό την προϋπόθεση ότι οι χρήστες είναι πιθανό να τιτλοφορούν για πράγματα που τους ενδιαφέρουν.

□□□□□□(□□) = {□1,… . , □□}

Αυτό μπορεί να επεκταθεί περαιτέρω. Κάθε χρήστης Twitter ακολουθεί ένα σύνολο άλλων χρηστών, τους επόμενους, και κάθε χρήστης ακολουθείται από ένα σύνολο χρηστών που ονομάζονται οι followers τους. Βλ. επόμενες 2 εξισώσεις. Μπορεί λογικά να υποθέσει ότι τα tweets των ακολούθων και των οπαδών τους μπορούν να παρέχουν περαιτέρω πληροφορίες για τα ενδιαφέροντα ενός χρήστη.

$$\square\square\square\square\square\square\square\square\square(\square\square) = \{\square 1, \ldots, \square\square\}$$

$$\square\square\square\square\square\square\square\square\square(\square\square) = \{\square 1, \ldots, \square\square\}$$

Ο χρήστης επιλέγει ενεργά τους επόμενους, πιθανώς επειδή περιμένουν ότι τα tweets τους θα ενδιαφέρουν και έτσι μπορεί να χρησιμοποιήσει τα tweets τους με τον ίδιο τρόπο όπως τα tweets του χρήστη, ως συμπληρωματική πηγή πληροφοριών προφίλ. Έτσι, οι followeetweets (UT) είναι το σύνολο των tweets των ακολούθων του UT .

$$\square\square\square\square\square\square\square\square\square\square\square\square(\square\square) = \bigcup_{\forall\, \square\square \in \square\square\square\square\square\square\square\square(\square\square)} (\square\square\square\square\square\square(\square\square))$$

$$\square\square\square\square\square\square\square\square\square\square\square\square\square(\square\square) = \bigcup_{\forall\, \square\square \in \square\square\square\square\square\square\square\square(\square\square)} (\square\square\square\square\square\square(\square\square))$$

Με τη σειρά του, ο follower του UT κάνει την ενεργό απόφαση να ακολουθήσει την UT, πιθανώς επειδή (οι οπαδοί της UT) αναμένουν ότι τα tweets της UT θα παρουσιάσουν ενδιαφέρον. Αλλά τα tweets αυτών των οπαδών θα ενδιαφέρουν την UT; Και μπορούν να χρησιμοποιηθούν ως μια βιώσιμη πηγή πληροφόρησης για τη δημιουργία προφίλ σύμφωνα με την Εξίσωση 5; Ούτε οι ερωτήσεις μπορούν να επιβεβαιωθούν με εμπιστοσύνη, αφού στην περίπτωση της πλειοψηφίας των χρηστών Twitter τουλάχιστον οι χρήστες ασκούν λίγο έλεγχο στους οπαδούς τους. οι χρήστες σπάνια κλαδεύουν τους οπαδούς τους που δεν τους ενδιαφέρουν και υπάρχουν πολλές περιπτώσεις οπαδών που δημιουργούν πολύ λίγα tweets τους. Ωστόσο, τα tweets των οπαδών παρέχουν σίγουρα μια ενδιαφέρουσα πηγή πληροφοριών προφίλ που αξίζει να εξερευνήσετε. Συνοπτικά, τα παραπάνω υποδηλώνουν 5 βασικές στρατηγικές δημιουργίας προφίλ:

(1) που εκπροσωπούν τους χρήστες με δικά τους tweets (tweets (UT)),

(2) από τα tweets των ακολούθων (followeetweets (UT)).

(3), από τα tweets των οπαδών τους (followertweets (UT))?

(4) από τα ονόματα των ακολούθων (ακολουθούμενοι (UT)).

(5) ή από τις ταυτότητες των οπαδών τους (οπαδοί (UT)).

Και φυσικά, όπως θα συζητήσουμε, αυτές οι διαφορετικές πηγές δεδομένων προφίλ μπορούν να συνδυαστούν με διάφορους τρόπους, έτσι ώστε, για παράδειγμα, να μπορούμε να αντιπροσωπεύουμε ένα χρήστη από τα δικά του tweets, τα tweets των followees τους και τα tweets των οπαδών

Indexing & Recommendation Τώρα που η βάση για τη δημιουργία προφίλ των χρηστών του Twitter είναι γνωστή, με βάση τα tweets και / ή τις κοινωνικές συνδέσεις, μπορεί να αναπροσαρμόσει αυτά τα προφίλ και να αναπτύξει το πλαίσιο συστάσεων για την απόδοση αποτελεσμάτων βάσει ενός προφίλ χρήστη στόχου ή μάλιστα ενός συγκεκριμένου συνόλου ερωτημάτων. Υπάρχουν πολλά πλεονεκτήματα για να προχωρήσουμε με αυτόν τον τρόπο, σε αντίθεση με την ανάπτυξη ενός πλαισίου σύστασης βάσει παραγγελίας. Αρχικά, η Lucene παρέχει μια αποδεδειγμένη, ισχυρή και κλιμακούμενη πλατφόρμα ευρετηρίασης και ανάκτησης που σχεδιάστηκε για να αντιμετωπίσει τα δεδομένα και τη χρήση της κλίμακας Web. Επιπλέον, παρέχει πρόσβαση σε ισχυρές λειτουργίες ευρετηρίασης και αντιστάθμισης χρόνου που θα φιλοξενήσουν μια πιο εξελιγμένη προσέγγιση στο σχεδιασμό των χρηστών απ 'ό, τι ένα απλό σύστημα σταθμίσεων με βάση τη συχνότητα. Τέλος, οι λειτουργίες ανάκτησης της Lucene μπορούν

να χρησιμοποιηθούν απευθείας για την ανάκτηση των προφίλ βάσει ερωτήματος και μπορούν εύκολα να προσαρμοστούν για σύσταση. Δεδομένου ότι η Lucene είναι μια μηχανή αναζήτησης που βασίζεται σε κείμενο, οι βασικές της μονάδες πληροφοριών είναι έγγραφα που πρέπει να αναπροσαρμόζονται και να αποθηκεύονται για ανάκτηση. Χρησιμοποιώντας τα χαρακτηριστικά ευρετηρίασης του Lucene, μπορεί να αντιπροσωπεύει κάθε UT ως σταθμισμένο όρο-φορέα, profile (UT, source) (βλ. Εξίσωση 6), έτσι ώστε το i th στοιχείο αυτού του φορέα να αντιπροσωπεύει τον i μοναδικό όρο στην πηγή, και το βάρος αυτού του i όρου (wi) αντιπροσωπεύει τη σημασία αυτού του όρου για το UT. Στην περίπτωση που η πηγή είναι μία από τις πηγές περιεχομένου (tweets (UT), followeetweets (UT) ή followertweets (UT), τότε αυτοί οι όροι θα είναι οι λέξεις που χρησιμοποιούνται στα tweets των σχετικών χρηστών, ενώ όταν η πηγή είναι μία από τις πηγές κοινωνικές / συνεργατικές πηγές (follow-ups (UT) ή followers (UT)) τότε αυτοί οι όροι θα είναι IDs χρηστών. Στη συνέχεια, θα χρησιμοποιήσουμε το profile (UT) αντί για το profile (UT, source) χωρίς απώλεια της γενικότητας σε περιπτώσεις όπου η παράμετρος πηγής είναι σαφής

$$\square\square\square\square\square\square\square(\square\square, \square\square\square\square\square\square) = \{\square 1, \ldots, \square\square\}$$

Θα μπορούσε να χρησιμοποιηθεί ένας απλός μετρητής συχνότητας ως συνάρτηση σταθμίσεως όρου, έτσι ώστε το διάνυσμα προφίλ να μπορεί να αποτελείται από τις μετρήσεις συχνότητας των διαφόρων λέξεων που χρησιμοποιούνται στα tweets της UT, για παράδειγμα. Ωστόσο, σε αυτή την περίπτωση μπορεί να χρησιμοποιηθεί η μετρική βαρύτητας TF-IDF της Lucene. Έτσι, η βαθμολογία TF-IDF του όρου ti στο UT είναι ανάλογη με τη συχνότητα εμφάνισής του στο προφίλ (UT) και αντιστρόφως ανάλογη με τη συχνότητά του στα άλλα προφίλ U, όπως φαίνεται στην εξίσωση 7 έως 9. Αυτό έχει ως αποτέλεσμα υψηλότερο βάρους για τους όρους προφίλ που είναι συχνές σε ένα δεδομένο προφίλ αλλά σπάνια σε όλη τη βάση προφίλ, γεγονός που συμβάλλει στη διάκριση των προφίλ κατά τη διάρκεια της ανάκτησης με την έκπτωση των αγώνων στους κοινούς όρους (λέξεις ή χρήστες). Για παράδειγμα, εάν εκπροσωπούμε κάθε χρήστη μόνο με τα δικά του ακατέργαστα tweets, τότε η στάθμιση TF-IDF θα δώσει μεγαλύτερο βάρος στους όρους που είναι κοινά για το UT, αλλά ασυνήθιστο για τον υπόλοιπο πληθυσμό χρηστών. Αυτοί οι όροι υψηλής βαθμολογίας χρησιμεύουν για την καλύτερη διάκριση των συμφερόντων της UT σε σχέση με τους άλλους χρήστες κατά τη διάρκεια της ανάκτησης.

$$\square\square - \square\square\square(\square\square, \square\square, \square) = \square\square(\square\square, \square\square) * \square\square\square(\square\square, \square)$$

$$\square\square(\square\square, \square\square) = \square\square, \square / \sum_{\square} \square\square, \square$$

$$\square\square\square(\square\square, \square) = \log \frac{|\square|}{|\{\square : \square\square \in \square\}|}$$

Η ανάκτηση βάσει ερωτήματος και η σύσταση βάσει προφίλ εφαρμόζονται στη συνέχεια με τη συνήθη λειτουργία ανάκτησης Lucene, με το έγγραφο προφίλ του χρήστη του στόχου να χρησιμεύει ως ερώτημα αναζήτησης στην περίπτωση του τελευταίου. Όλα αυτά παρέχουν ένα πολύ ισχυρό και ευέλικτο πλαίσιο ανάκτησης και σύστασης, καθώς τα προφίλ μπορούν να εκπροσωπούνται και να αναπροσαρμόζονται από έναν συνδυασμό όρων πηγής, αξιοποιώντας αποτελεσματικά μια ποικιλία διαφορετικών στρατηγικών συστάσεων, απλές στρατηγικές βασισμένες στο περιεχόμενο ή κοινωνικές συστάσεις σε πιο εξελιγμένα υβρίδια . Για παράδειγμα, χρησιμοποιώντας τις πηγές περιεχομένου μπορούμε να δημιουργήσουμε ένα χώρο συστημένων συστημένων με βάση το περιεχόμενο. Αντιστρόφως, με την ευρετηρίαση των χρηστών απλώς από τους επόμενους ή τους οπαδούς τους ή και από τους δύο, μπορούμε να δημιουργήσουμε συνηθισμένο στυλ φιλτραρίσματος [11, 16] συνιστώντες.

4.7. TadVise

Το Tadvise είναι επίσης μια λειτουργιά του twitter και είναι σε θέση να συστήσει νέους followers με βάση τα hashtag τους. Ο σκοπός του Tadvise είναι να βοηθήσει τους χρήστες να γνωρίζουν καλύτερα τους followers τους. Ένα σύνολο από hashtags συνδέεται με το προφίλ κάθε χρήστη, καθώς τα hashtags εμφανίζονται στα tweets του χρήστη. Το βάρος κάθε hashtag στο προφίλ του χρήστη καθορίζεται από το συνολικό PageRank των χρηστών που αναφέρουν τον ιδιοκτήτη του προφίλ με το αντίστοιχο hashtag. Η λογική πίσω από αυτό είναι ότι ένα hashtag είναι πολύ σχετικό με ένα χρήστη, αν χρησιμοποιείται συχνά στα εισερχόμενα tweets του χρήστη από εξαιρετικά έγκυρους χρήστες. Το Tadvise στη συνέχεια συνιστά καλά συνδεδεμένους θεματικούς χρήστες ως ακόλουθους. Αυτοί οι χρήστες μπορούν να χρησιμεύσουν

ως κόμβοι για τη μετάδοση ενός μηνύματος σε ένα ευρύτερο κοινό. Οι υποψήφιοι followers κατατάσσονται σύμφωνα με τα αποτελέσματα των κόμβων τους που αντιπροσωπεύουν τον αριθμό των ενδιαφερομένων χρηστών που θα μπορούσαν ενδεχομένως να λαμβάνουν tweets από το πρώτο. Με δεδομένο ένα χρήστη και ένα tweet με τουλάχιστον ένα hashtag, ο Tadvise καθορίζει εάν το tweet πιθανόν να διαχέεται από το χρήστη. Πρώτον, ο Tadvise προσδιορίζει εάν η ετικέτα (ες) που χρησιμοποιείται στο tweet είναι σχετικές με τους οπαδούς και τους οπαδούς των ακόλουθων. Εάν υπάρχει ένας μεγάλος αριθμός σχετικών οπαδών και οπαδών των ακολούθων που έχουν υψηλά προφίλ βάρους για το δεδομένο hashtag, το tweet αναμένεται να προσελκύσει μεγάλη προσοχή. Διαφορετικά, οι οπαδοί και οι οπαδοί των οπαδών μπορούν να επιλέξουν να αγνοήσουν το tweet.`

Tadvise Overview and Components

Το Tadvise δημιουργεί προφίλ χρηστών για twitterers προκειμένου να συστήσει tweets ή retweets που θα μπορούσαν ενδεχομένως να είναι σχετικές με μια κοινότητα των οπαδών τους. Προκειμένου να επιτευχτεί αυτούς η εγγραφή στο σύστημα του tadvise είναι αναγκαία μόλις η εγγραφή έρθει εις πέρας τότε το tadvise σέρνει το κοινωνικό δίκτυο του u και δημιουργεί προφίλ χρηστών των οπαδών του. Στην περίπτωση που ο χρήστης εισέλθει στην αρχική σελίδα τότε θα δει 1 σύνολο από συμβουλών . Το κόκκινο φως σημαίνει ότι κανένας από τους οπαδούς του u δεν έχει επισημανθεί με τις ετικέτες (hash) στο τιτίβισμα. Τέλος, το πορτοκαλί φως σημαίνει ότι ορισμένοι από τους οπαδούς του u έχουν επισημανθεί με τις ετικέτες (hash) στο tweet, αλλά δεν είναι η πλειοψηφία των οπαδών του u.

Το tadvise έχει τις εξής επισημάνσεις

- Κόκκινη σημαίνει ότι κανένας από τους οπαδούς του u δεν έχει επισημανθεί με τις ετικέτες
- Πορτοκαλί σημαίνει ότι ορισμένοι από τους οπαδούς του u έχουν επισημανθεί με τις ετικέτες\
- Πράσινη σημαίνει ότι η πλειοψηφία των οπαδών του u έχουν επισημανθεί με μία ή περισσότερες ετικέτες

Επίσης το tadvise αποτελείται από 3 βασικά στοιχεία

- Crawlet of the tadvise
- User Profile Builder of Tadvise
- Advice Engine of Tadvise

Crawler of Tadvise

Η συνιστώσα ανιχνεσης του Tadvise παιρνει σημείο ως Input και χρησιμοποιεί το API Twitter για την ανίχνευση twitterers. Το στοιχείο ανίχνευσης εκτελεί τη δουλειά του σε δύο βήματα. Αρχικά ανιχνεύει το δίκτυο των οπαδών σε απόσταση ενός και δύο από ένα σημειο Το δεύτερο βήμα της ανίχνευσης είναι η ανίχνευση λιστών Twitter

User Profile Builder of Tadvise

Σύμφωνα με την ερευνα των Peyman Nasirifard και Conor Hayes προκειμένου να γίνει η αξιολόγηση για το ποσο συναφή είναι 1 tweet για έναν χρήστη uj, δημιουργούν ένα σταθμισμένο προφίλ χρήστη uj που περιέχει μεταδεδομένα για τις κοινότητες, τα ενδιαφέροντα κτλ. Συνοπτικά, κάθε προφίλ χρήστη αποτελείται από μεταδεδομένα που προέρχονται από λίστες Twitter ετικέτες) που σχετίζονται με το χρήστη από άλλους χρήστες. Για την δημιουργία λοιπόν αυτού του σταθμισμένου προφίλ, πρέπει να ταξινομηθούν οι ετικέτες που έχουν συσχετιστεί με έναν χρήστη.Για να το πετύχουν αυτό οι Peyman Nasirifard και Conor Hayes χρησημοποιούν το μοντέλο της εξίσωσης από την έρευνα Kwak et al

$$\square\square\square\square(\square\square) = \log(\#\square\square^{\square 0})$$

Advice Engine of Tadvise Η συνιστώσα μηχανής συμβουλών λαμβάνει προφίλ χρηστών και ένα tweet ως εισροές και παρέχει δύο είδη συμβουλών διάχυσης σε πραγματικό χρόνο:

α) Προφίλ κοινού που επιτρέπει στους χρήστες να εντοπίζουν το υποσύνολο των οπαδών τους που έχουν επισημανθεί με έναν όρο που χρησιμοποιείται στο τιτίβισμα.

β) να συστήσει καλά συνδεδεμένους θεματικούς χρήστες για ένα tweet, ο οποίος μπορεί να επαναλάβει το tweet. Με δεδομένο ένα tweet και ένα χρήστη ui, πρώτα εξάγουμε ετικέτες από το tweet.

Για την ορθή λειτουργία του συστήματος ο εμπλουτισμός από hashtags είναι αρκετά συματικος διότι παρέχουν στο εν λόγω μοντέλο ένα σύνολο από hashtags που είναι σχετικά με τα αρχικά hashtag. Το σύστημα λοιπόν περιλαμβάνει 2 μέρη για να είναι σε θέση να προτείνει στον τελικό χρήστη. Αρχικά οι Peyman Nasirifard και Conor Hayes δημιουργούν μια συγκεντρωτική σε προφίλ χρηστών περιλαμβάνοντας τους ακόλουθους που βρίσκονται σε απόσταση 1και 2 από το δίκτυο τους. Παρουσιάζοντας λοιπόν αυτά τα συγκεντρωτικά προγράμματα με ονόματα followersP rofile1(ui) και followersP rofile2(ui) αντίστοιχα. Αυτά τα προγράμματα περιέχουν ταξινομημένα βάρη τα οποία βάρη συγκεντρώνονται από followersP rofile1(ui) και followersP rofile2(ui)Στο δεύτερο μέρος το Tadvise δείχνει λοιπον για το αν είναι σε θέση να βρει αντιπροσωπευτικά tags συμφώνα με την σήμανση που αναφέραμε σε προηγούμενη παράγραφο. Στην περίπτωση που το Tadvise δεν είναι σε θέση να βρει αντιπροσωπευτικές ετικέτες ενός tweet τότε θα παρουσιάσει κόκκινη σήμανση . Αντί να εφαρμοστεί ένα σταθερό όριο σε κάθε προφίλ, οι Peyman Nasirifard και Conor Hayes ορίζουν ένα σημείο μεταξύ των δύο διαμερισμάτων εφαρμόζοντας τον αλγόριθμο ομαδοποίησης k-mean με k = 2. Το πρώτο διαμέρισμα, το οποίο ομαδοποιεί ετικέτες υψηλής κατάταξης, αντιπροσωπεύει την πηγή του πράσινου φωτός. Το δεύτερο διαμέρισμα αντιπροσωπεύει την πηγή της συμβολής του πορτοκαλί φωτός. Αν το μοντέλου Tadvise δεν είναι σε θέση να βρει hashtags σε κάθε διαμέρισμα τότε αυτό παρουσιάζει την ένδειξη κόκκινης σήμανσης, αν δεν είναι σε θέση να βρει αντιπροσωπευτικές ετικέτες ενός τιτλοστομή σε κάθε διαμέρισμα. Η συμπεριφορά όμως αυτή δεν είναι απόλυτη το οποίο σημαίνει ότι ο ένας χρήστης μπορεί να εισάγει νέα περιεχόμενα. Ο παρακάτω αλγόριθμος δείχνει τον ψευδοκώδικα του δεύτερου μέρους της συμβουλής. Η είσοδος αυτού του αλγορίθμου είναι ένα κατευθυνόμενο γράφημα g το οποίο είναι κατασκευασμένο ως εξής: Η ρίζα του g είναι ο σπόρος ui.Γινετε επίσης η προσθήκη σε όλα τα μέση του UF Προσθέτουμε επίσης όλα τα μέλη του Uf o i στο g (uj → ui). Ο λόγος είναι ότι όταν ενας χρήστης ui δημιουργεί ένα tweets, όλοι οι followers λαμβάνουν αυτό το tweet και έτσι μπορούν να ενεργούν ως πιθανοί κόμβοι. Στη συνέχεια, οι ακόλουθοι του κάθε οπαδού του ui, που έχουν επισημανθεί με μία ή περισσότερες ετικέτες (hash) στο tweet, θα προστεθούν στο g (χρησιμοποιώντας followersP rof ile2 (ui)). Ο αλγόριθμος βρίσκει k κόμβους σε g χρησιμοποιώντας In-degree έτσι ώστε οι κόμβοι να καλύπτουν όσο το δυνατόν περισσότερους ενδιαφερόμενους οπαδούς (σε απόσταση 2 ui) και να έχουν όσο το δυνατόν περισσότερους αλληλοκαλυπτόμενους οπαδούς.

4.8. Hastag Recommendations

Γενικά η έννοια του hashtag υπάρχει ώστε να εξυπηρετεί του χρήστες του twitter στην περίπτωση που είτε θέλουν να κατηγοριοποιήσουν τα tweet τους. Είτε πρόκειται για την φόρτωση φωτογραφιών. Ακόμα και live μεταδόσεις για οποιοδήποτε θέμα. Τα συστήματα προτάσεων . Επομένως, προτείνονται συστήματα συνιστώμενων για την υποβολή κατάλληλων εκθέσεων για τους χρήστες

Recommending Hashtags in Twitter with TF-IDF Scheme

Συμφώνα με την έρευνα του Zangerle et al. ο πρωταρχικός σκοπός των hashtag είναι να κατηγοριοποιήσουν τα tweets και να διευκολύνουν την αναζήτηση. Το paper συνιστά κατάλληλα hashtags στον χρήστη, ανάλογα με το περιεχόμενο που εισάγει ο χρήστης χωρίς να λαμβάνει υπόψη τις προτιμήσεις του χρήστη για συγκεκριμένα hashtags.

Όταν ένας χρήστης γράφει ένα tweet, το σύστημα προτάσεων ανακτά ένα σύνολο tweets παρόμοιο με το εκάστοτε tweet. Στη συνέχεια, τα hashtags εξάγονται από τα ανακτημένα παρόμοια tweets και ταξινομούνται με βάση τον αριθμό των εμφανίσεών τους σε ολόκληρο το σύνολο δεδομένων, τον αριθμό των εμφανίσεών τους στο σύνολο δεδομένων ή τα αποτελέσματα ομοιότητας των tweets . Τα μέτρα ακρίβειας και ανάκλησης αυτών των τριών βαθμών κατάταξης δείχνουν ότι η βαθμολογία LikeityRank

είναι η καλύτερη από αυτές και η απόδοση του συστήματος συνιστώμενων είναι η καλύτερη όταν συνιστώνται μόνο πέντε hashtags.

Suggesting Hashtags on Twitter using Bayes Model

Ένας διαφορετικός τρόπος για την σύσταση hashtags είναι το μοντέλο που προτείνεται από Mazzia et al. [17]. Παρόμοιος με την προηγούμενη ενότητα αυτό το μοντέλο κάνει σύσταση hashtags ανάλογα με το περιεχόμενο του χρήστη. Η διαφορά σε αυτό το μοντέλο είναι ότι σε αντίθεση με το προηγούμενο μοντέλο κάνει την χρήση του μοντέλου Bayes, στο οποίο ο υπολογισμός γίνεται με βάση την πιθανότητα χρήσης hashtags. Προκειμένου λοιπόν να λειτουργήσει σωστά ο αλγόριθμος πρέπει να γίνει ο καθαρισμός των δεδομένων από διάφορους παράγοντες όπως π.χ. spam. Τα spam φιλτράρονται περιορίζοντας των αριθμό των tweets. Στη παρακάτω εξίσωση που χρησιμοποιείτε από τον λόγω μοντέλο Bayes

$$p(Ci|x1,\dots , xn) = p(Ci)p(x1|Ci) \dots p(Ci)p(xn|Ci)/p(x1 \dots xn)$$

εδώ Ci αντιπροσωπεύει την i th hashtag και x1, ..., xn αντιπροσωπεύει τις λέξεις. p (Ci | x1, ..., xn) είναι η πιθανότητα χρήσης του hashtag Ci με δεδομένες τις λέξεις που ο χρήστης παρέχει και συνιστώνται στον χρήστη οι χάρτες με τις υψηλότερες πιθανότητες. Το p (Ci) είναι η αναλογία του αριθμού των χρόνων που έχει χρησιμοποιηθεί το Ci ως προς τον συνολικό αριθμό των tweets με hashtags. p (x1 | Ci) ... p (xn | Ci) υπολογίζεται από τα υπάρχοντα δεδομένα των tweets. Το χαρτί προτείνει επίσης ένα άλλο μοντέλο που χρησιμοποιεί την αντίστροφη συχνότητα εγγράφων (IDF) για να υπολογίσει την πιθανότητα

$$p(x1,\dots , xn|Ci) = p(x1|Ci)^{1-t1} \dots p(x1|Ci)^{1-tn}$$

όπου tj είναι το βάρος IDF της λέξης xj.

Υψηλής διαστάσεων Euclidean διαστημικό μοντέλο Το χαρτί που προτείνεται από Li et al. [16] συνιστά επίσης τα hashtags που βασίζονται στις πληροφορίες που παρέχονται από τα προηγούμενα παρόμοια tweets. Κατασκευάζει υψηλού διαστάσεων Euclidean χώρο με τα λόγια του tweets. Οι συμβολισμοί των tweets που έχουν τις ελάχιστες αποστάσεις συνιστώνται. Η απόσταση των tweets σε αυτήν την προσέγγιση μετράται ως 1) Ευκλείδεια Απόσταση, 2) Οντολογική Βάση Απόσταση (OBD), ή 3) Κεντρική Απόσταση Βάσης Οντολογίας (COBD). Η σύγκριση των ποσοστών σφάλματος για αυτές τις τρεις μεθόδους δείχνει ότι η μέθοδος OBD εκτελεί τις καλύτερες.

where tj is the IDF weight of the word xj .

4.9. Tweet Recommendation

Όλα τα tweets από τους επόμενους χρήστες εμφανίζονται στην αρχική σελίδα του χρήστη. Όταν ο χρήστης παρακολουθεί πολλούς ενεργούς χρήστες, υπάρχουν πιθανότητες να χάσει ο χρήστης την ανάγνωση ορισμένων ενδιαφερόντων tweets. Με το προσεκτικό φιλτράρισμα πληροφοριών, μπορούν να επιλεγούν και να τονιστούν σημαντικά tweets σύμφωνα με τις προτιμήσεις του χρήστη. Όταν ένας χρήστης παρακολουθεί 1 συγκριμένο αριθμό από χρήστες τότε είναι πολύ πιθανό να μην προσέξει διαφορά tweets τα οποία θα είναι σχετικά με τα ενδιαφέροντα του. Οπότε η ανάγκη για φιλτράρισμα των tweets με βάση τα ενδιαφέροντα tweets είναι αναγκαίο. Συμφώνα με την έρευνα των Uysal, I., Croft, B.W[24] εισάγουν 2 μεθόδους.

1) Κατάταξη εισερχόμενων tweets

2) Κατάταξη των στοχοθετημένων χρηστών.

Στην πρώτη μέθοδο, για κάθε χρήστη, τα tweets κατατάσσονται ανάλογα με την πιθανότητα να τους επαναληφθεί από τον χρήστη. Στη δεύτερη μέθοδο, για κάθε tweet, οι χρήστες κατατάσσονται ανάλογα με τις πιθανότητες επανάληψης του tweet. Η υποκείμενη υπόθεση είναι ότι ένα tweet θεωρείται σχετικό και συνιστάται σε ένα χρήστη, εάν ο χρήστης είναι πιθανό να επαναλάβει το tweet. Στην έρευνα των

Uysal, I., Croft, B.W αντιμετωπίζουν την κατάταξη ως πρόβλημα ταξινόμησης. Η λειτουργιά αυτή έχει ως σκοπό την εκπαίδευση ενός ταξινομιτη με βαση συγκεκριμένα χαρακτηριστικά τα οποία είναι:

- Χαρακτηριστικά συντακτών
- Χαρακτηριστικά tweet
- Χαρακτηριστικά χρηστών

Οι λειτουργίες που βασίζονται στο Tweet είναι οι συντακτικές δυνατότητες του tweet, όπως οι hashtags, οι διευθύνσεις URL , κλπ. - - Οι λειτουργίες που βασίζονται στον χρήστη σχετίζονται με τον χρήστη του οποίου το tweet κατατάσσεται Ο εκπαιδευμένος ταξινομητής θα προβλέψει εάν ένα δεδομένο tweet ενδέχεται να επαναληφθεί από έναν συγκεκριμένο χρήστη, ανάλογα με τα παραπάνω χαρακτηριστικά.

Personalized News Recommendation by analyzing Tweet Contents

Το εξατομικευμένο σύστημα συνημμένων ειδήσεων από τον Morales χρησιμοποιεί tweets για να δημιουργήσει προφίλ χρηστών και να συστήσει ενδιαφέροντα άρθρα ειδήσεων του Yahoo σε χρήστες με βάση την εποπτευόμενη μέθοδο εκμάθησης. Ο αλγόριθμος κατάταξης προτάσεων δίνεται από τον ακόλουθο τύπο

$$\square\square(\square, \square) = \underset{\square}{\square} \sum (\square, \square) + \square\square\square(\underset{\square}{\square}, \square) + \square \prod (\square)$$

όπου P RT (u, n) = Κατάταξη των ειδήσεων n για το χρήστη u; T (u, n) = Συνάφεια που βασίζεται στοπεριεχόμενο μεταξύ του χρήστη u και των ειδήσεων n κατά το χρόνο T. Q ΓT (u, n) = Συνδεδεμένη κοινωνική σχέση μεταξύ του χρήστη u και των ειδήσεων n στην ώρα T. T (n) = Δημοτικότητα των ειδήσεων n την ώρα T; α, β, γ = Συντελεστές που καθορίζουν τα σχετικά βάρη των εξαρτημάτων. Το χαρτί χρησιμοποιεί το σύστημα εξαγωγής οντοτήτων φάσματος [20] και εφαρμόζει την έννοια της οντότητας για να βρει τη συγγένεια μεταξύ των tweets και των ειδησεογραφικών άρθρων Η συγγένεια περιεχομένου (P T (u, n)) καταγράφει τη διαίσθηση ότι αν τα άρθρα ειδήσεων και τα tweets του χρήστη βρίσκονται κάτω από κοινές οντότητες, τότε οι ειδήσεις αφορούν τον χρήστη. Η συγγένεια με βάση την κοινωνία ΓT (u, n) υπολογίζει τις σχετικές βαθμολογίες λαμβάνοντας υπόψη τα tweets που συντάσσονται από τους γειτονικούς χρήστες. Άλλα χαρακτηριστικά, όπως η ηλικία, η ζεστασιά και ο αριθμός κλικ στα άρθρα ειδήσεων, εφαρμόζονται επίσης στον αλγόριθμο μάθησης

5. Conclusions

Το συστήματα που περιγράφηκαν για τη σύνταξη συστάσεων σε πραγματικό χρόνο, όπως συζητείται, συνίσταται σε μια ποικιλία αλγορίθμων, προκειμένου να συστήσει στους χρήστες τους την καλύτερη επιλογή είτε για να είναι μια ταινία, ένα βίντεο ή ένα άρθρο φυσικά εκεί πολλά που δεν είναι που καλύπτονται από αυτό το Έργο σχετικά με την πλήρη λογική που συζητήσαμε αυτά τα 3 συστήματα. Τα συστήματα συστάσεων είναι βέβαιο ότι θα έχουν ένα λαμπρό μέλλον στην τεχνολογία προγραμματισμού και σίγουρα θα πρέπει να συνεχίσουν να βελτιώνονται ώστε οι χρήστες να έχουν τα καλύτερα συνιστώμενα στοιχεία για μια καλύτερη εμπειρία χρήστη

6. References

1)Xavier Amatriain and Justin Basilico. 2012. Netflix Recommendations: Beyond the 5 stars (Part 2). Retrieved December 6, 2015 from http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5- stars
2)Berry_et_al-2010-
3)Proceedings_of_the_American_Society_for_Information_Science_and_Technology
4)Lessons from the Netflix Prize Challenge Robert M. Bell and Yehuda Koren
5)A Social Network-Based Recommender System (SNRS) Jianming He and Wesley W. Chu
6)The Netflix Recommender System: Algorithms, Business Value, and Innovation
7) Deep Neural Networks for YouTube Recommendations Paul Covington, Jay Adams, Emre Sargin Google Mountain View, CA

8)S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On using very large target vocabulary for neural machine translation. CoRR, abs/1412.2007, 2014.

9)S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On using very large target vocabulary for neural machine translation. CoRR, abs/1412.2007, 2014.
10) F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In AISTATS^a_A _Z05
11) J. Weston, A. Makadia, and H. Yee. Label partitioning for sublinear ranking. In S. Dasgupta and D. Mcallester, editors, Proceedings of the 30th International Conference on Machine Learning (ICML-13), volume 28 Workshop and Conference Proceedings, May 2013.
12) T. Liu, A. W. Moore, A. Gray, and K. Yang. An investigation of practical approximate nearest neighbor algorithms.
13) E. Meyerson. Youtube now: Why we focus on watch time. http://youtubecreator.blogspot.com/2012/08/ youtube-now-why-we-focus-on-watch-time.html. Accessed: 2016-04-20.
14) X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: Dwell time for personalization. In Proceedings of the 8th ACM Conference onRecommender Systems, RecSys '14

16) Recommending #-Tags in Twitter
17) Using Twitter to Recommend Real-Time Topical News Owen Phelan, Kevin McCarthy & Barry Smyth
18) Real-Time Twitter Recommendation: Online Motif Detection in Large Dynamic Graphs
19)A Survey of Recommender Systems in Twitter

7. Βιβλιογραφία

Lessons from NetFlix Prize Challenge

A Social Network-Based Recommender System

The Netflix Recommender System: Algorithms, Business Value,

The Evolution of Recommendation Systems

The BellKor Solution to the Netflix Grand Prize

A Tour of TensorFlow

Early Stabilizing Feature Importance for TensorFlow Deep Neural Networks

Quantization and Training of Neural Networks for Efficient

Deep Neural Networks for YouTube Recommendations

2007 Book The Adaptive Web

Large Scale Analytics on Factors Impacting Retweet in Twitter Network

Recommending #-Tags in Twitter

Using Twitter to Recommend Real-Time Topical News

Real-Time Twitter Recommendation: Online Motif Detection in Large Dynamic Graphs ∗

A Survey of Recommender Systems in Twitter

Tadvise: A Twitter Assistant Based on Twitter Lists

https://www.wikipedia.org/