4. Context-aware Συστήματα Συστάσεων

Σύνοψη: *Στο παρόν κεφάλαιο θα μελετήσουμε συστήματα συστάσεων που λαμβάνουν υπόψη τους* ***εξωτερικούς (contextual) παράγοντες*** *της αλληλεπίδρασης του χρήστη μ' ένα προϊόν ή στοιχείο (π.χ. τη* ***χρονική στιγμή*** *της διάδρασης, τη* ***γεωγραφική θέση*** *του χρήστη, τη* ***συναισθηματική του διάθεση****, τη* ***συσκευή επικοινωνίας*** *που χρησιμοποιεί κ.λπ.). Επίσης, θα περιγράψουμε* ***πρωτόκολλα αποτίμησης της ποιότητας των αλγορίθμων*** *που συνυπολογίζουν την προαναφερθείσα* ***χρονική στιγμή****. Στη συνέχεια, ως ένα υπόδειγμα των* ***context-aware συστημάτων συστάσεων*** *θα αναλύσουμε αλγορίθμους οι οποίοι προτείνουν είτε* ***μεμονωμένα σημεία ενδιαφέροντος*** *(μουσεία, αρχαιολογικούς χώρους, αξιοθέατα κ.λπ.) είτε ένα* ***ανάλογο σύνολο όπου μπορεί να περιηγηθεί*** *διαδοχικά ένας χρήστης κατά τη διάρκεια της επίσκεψής του σε μια πόλη (****Point of Interest Recommendation*** *versus* ***Itinerary Recommendation****). Τέλος, θα αναφερθούμε σε* ***υβριδικά*** *συστήματα συστάσεων, που, ως εκ τούτου, συνδυάζουν τα* ***δεδομένα της αλληλεπίδρασης χρηστών και στοιχείων (collaborative data)*** *με τα* ***δεδομένα περιεχομένου (content-based data)****, τα οποία με τη σειρά τους περιγράφουν τα στοιχεία και τους χρήστες (π.χ. τη* ***συναισθηματική διάθεση*** *ενός χρήστη την ώρα της διάδρασής του με ένα στοιχείο,* ***το μέρος*** *στο οποίο βρίσκεται κ.λπ.).*

Προαπαιτούμενη γνώση: *Ο αναγνώστης καλό θα ήταν να έχει προηγουμένως μελετήσει το δεύτερο και το τρίτο κεφάλαιο.*

4.1 Time-aware Συστήματα Συστάσεων

4.1.1 Εισαγωγή

Οι αλγόριθμοι παροχής συστάσεων που περιγράψαμε στα προηγούμενα κεφάλαια λαμβάνουν υπόψη μόνο πληροφορίες για τους χρήστες και για τα στοιχεία με τα οποία αυτοί αλληλεπιδρούν. Παρόλα αυτά, υπάρχουν *επιπρόσθετες πληροφορίες* που μπορούμε να εισάγουμε στους αλγορίθμους

παροχής συστάσεων, και οι οποίες αφορούν στο γενικότερο πλαίσιο (π.χ. *τη χρονική στιγμή* και *το μέρος* όπου βρίσκεται ο χρήστης) της αλληλεπίδρασης ενός χρήστη μ' ένα στοιχείο/προϊόν. Παραδείγματος χάρη, λαμβάνοντας υπόψη τη *διάσταση του χρόνου* μπορούμε να εκπαιδεύσουμε τα μοντέλα πρόβλεψης, ώστε να αντιλαμβάνονται και τις *βραχυχρόνιες* και όχι μόνο τις *μακροχρόνιες προτιμήσεις* των χρηστών.

Στο παρόν, λοιπόν, εξετάζουμε τη δυνατότητα αξιοποίησης της *διάστασης του χρόνου* από τα *αρχεία καταγραφής των εξυπηρετητών* (*server log files*), στα οποία τηρούνται πληροφορίες για την αλληλεπίδραση των χρηστών με τα στοιχεία (π.χ. τη *χρονική στιγμή έναρξης* της διάδρασης, τη *διάρκειά της* κ.λπ.). Τα παραδοσιακά συστήματα συστάσεων δεν αξιοποιούν τη διάσταση του χρόνου στα μοντέλα εκπαίδευσης των δεδομένων τους, με αποτέλεσμα να δημιουργούν μοντέλα πρόβλεψης που εστιάζουν κυρίως στις *μακροχρόνιες προβλέψεις* (δηλαδή στο *τι γενικώς αρέσει* στον υπό εξέταση χρήστη). Ωστόσο, όταν ένας αλγόριθμος επεξεργαστεί το σύνολο των δεδομένων μ' έναν συνεχή και γραμμικά εξελισσόμενο (ως προς τον χρόνο) τρόπο, ως μια *χρονοσειρά* (*time series*), τότε αυτή η προσέγγιση είναι κατάλληλη για *βραχυχρόνιες προβλέψεις*, δεδομένου ότι οι προτιμήσεις των χρηστών διαφοροποιούνται βάσει της *διάστασης του χρόνου* (π.χ. τα πρωινά ο χρήστης ενδιαφέρεται για συγκεκριμένη κατηγορία ειδησεογραφικών άρθρων, όπως για παράδειγμα είναι αυτά για τη μόδα, αν και γενικώς του αρέσουν τα άρθρα για την πολιτική και τα σπορ).

Όσον αφορά, τώρα, στα *πρωτόκολλα αποτίμησης της ποιότητας των αλγορίθμων*, αυτά θα πρέπει να διαφοροποιούνται καταλλήλως όταν συνυπολογίζεται από τον αλγόριθμο η *διάσταση του χρόνου*. Σε αντίθεση με τον παραδοσιακό τρόπο αξιολόγησης των συστημάτων συστάσεων που *διαχωρίζουν μόνο μία φορά* το σύνολο των δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου (*one time split*), στα *time-aware συστήματα* επιβάλλεται να γίνονται περισσότεροι του ενός διαχωρισμοί. Επίσης, θα πρέπει πάντα να εκπαιδεύουμε τα μοντέλα μας με παρελθοντικά δεδομένα, για να μπορέσουμε να προβλέψουμε τα μελλοντικά **(prequential evaluation)**. Το γεγονός αυτό μειώνει τις περιπτώσεις όπου μπορεί εσφαλμένα να παρατηρηθούν καλά αποτελέσματα μ' έναν και μοναδικό διαχωρισμό των δεδομένων εκπαίδευσης-δοκιμής. Σε αντίθετη περίπτωση, όταν οι διαχωρισμοί είναι περισσότεροι, μπορεί τα καλά αποτελέσματα να μην επαληθεύονται, κάτι που θα αναλυθεί παρακάτω εκτενώς.

Τέλος, το *prequential evaluation* επιτρέπει στον ερευνητή να παρατηρήσει και άλλα μεταβαλλόμενα χαρακτηριστικά των αλγορίθμων, όπως η *χρονική ποικιλομορφία* των συστάσεων έναντι των ίδιων επαναλαμβανόμενων προτεινόμενων στοιχείων. Ως παράδειγμα μπορούμε να αναφέρουμε την περίπτωση του να συστήνεται επαναλαμβανόμενα σ' έναν χρήστη το ίδιο στοιχείο, σε αντίθεση με την περίπτωση κατά την οποία του προτείνονται νέα στοιχεία, με αποτέλεσμα να διατηρείται αμείωτο το ενδιαφέρον του για την υπηρεσία συστάσεων.

4.1.2 Session-based Συστήματα Συστάσεων

Για προτεινόμενα στοιχεία με μικρή διάρκεια "ζωής", τα οποία δηλαδή παραμένουν για λίγο χρόνο επίκαιρα, όπως λόγου χάρη τα ειδησεογραφικά άρθρα, μας ενδιαφέρει η δημιουργία ενός *session-based συστήματος* που θα παρέχει φρέσκες, επίκαιρες -και ως εκ τούτου βραχύβιες- συστάσεις, οι οποίες θα βασίζονται κάθε φορά αποκλειστικά στην *εκάστοτε συνεδρία του χρήστη* (*user session*).

Ορισμός 4.1 **Συνεδρία Χρήστη (User Session)** είναι *το σύνολο των αλληλεπιδράσεων του χρήστη με τα στοιχεία μιας online υπηρεσίας*, με την προϋπόθεση ότι η διάρκεια αυτών των αλληλεπιδράσεων είναι σχετικά βραχεία (π.χ. 30 λεπτά της ώρας). Ένας χρήστης μπορεί να έχει περισσότερες από μία συνεδρίες με την ίδια online υπηρεσία κατά τη διάρκεια της ημέρας. Με τον όρο *session-based* χαρακτηρίζονται τα συστήματα συστάσεων όπου όλη η δραστηριότητα ενός χρήστη χωρίζεται σε υποσύνολα επιμέρους συνεδριών. Αυτό σημαίνει ότι ένας χρήστης μπορεί να έχει πολλές συνεδρίες όπου κατά τη διάρκειά τους να αλληλεπιδρά μ' ένα ή περισσότερα διαφορετικά στοιχεία [Epure et al., 2017, Liu et al., 2010, Shani et al., 2005].

Στη συνέχεια θα χρησιμοποιήσουμε ένα παράδειγμα για να διευκολύνουμε τον αναγνώστη στην κατανόηση την έννοιας *συνεδρία χρήστη*.

Παράδειγμα 4.1 Έστω, λοιπόν, ότι είμαστε οι διαχειριστές μιας διαδικτυακής εφημερίδας. Με $\mathcal{U}$ συμβολίζουμε το σύνολο των χρηστών που επισκέπτονται την εφημερίδα, ενώ με $\mathcal{I}$ το σύνολο των ειδησεογραφικών της άρθρων. Επίσης, έστω ότι παρακολουθούμε την αλληλεπίδραση των χρηστών με τα άρθρα της διαδικτυακής εφημερίδας. Συγκεκριμένα, κάθε φορά που ένας χρήστης διαβάζει ένα ή περισσότερα άρθρα κατά τη διάρκεια ενός σύντομου χρονικού διαστήματος, εμείς αποθηκεύουμε αυτήν τη διάδρασή του στη βάση δεδομένων ως μια *συνεδρία χρήστη (user session)*. Αυτές οι αλληλεπιδράσεις του με τα άρθρα της εφημερίδας έχουν μια αλληλουχία. Μ' άλλα λόγια, γνωρίζουμε για κάθε στοιχείο που ανήκει σε μια συνεδρία ενός χρήστη το κατά πόσο αυτό έχει επιλεγεί πρώτο, δεύτερο ή τελευταίο, καθώς και τη χρονική διάρκεια της διάδρασης του χρήστη μ' αυτό. Παραδείγματος χάρη, η *συνεδρία* $S_1(user = u_1, TimeStarted = t_1|\{i_1, 20sec\}, \{i_2, 145sec\})$ δείχνει ότι κατά τη διάρκειά της (S_1), εκκινώντας από το χρονικό σημείο t_1 από τον χρήστη u_1, το στοιχείο i_1 επιλέχθηκε πρώτο και διαβάστηκε για 20 δευτερόλεπτα, ενώ το i_2 επιλέχθηκε δεύτερο και τράβηξε την προσοχή του χρήστη για 145 δευτερόλεπτα.
Έστω, λοιπόν, ότι ενδιαφερόμαστε να παράσχουμε αποτελεσματικότερη ενημέρωση στους αναγνώστες μας μέσω μιας *μικρής εξατομικευμένης λίστας συστάσεων* που θα περιέχει τα N *κορυφαία άρθρα για τον κάθε υπό εξέταση χρήστη*. Αυτό συνεπάγεται ότι μια τέτοια λίστα θα

πρέπει να παράγεται εκ νέου κάθε φορά που θα επιλέγεται κάποιο καινούργιο άρθρο από έναν χρήστη. Συνεπώς, το σύστημα συστάσεων θα παρακολουθεί τον τρόπο με τον οποίο οι επισκέπτες αντιδρούν (εάν δείχνουν ενδιαφέρον ή όχι) στις συστάσεις των άρθρων, προκειμένου να προσπαθήσει να προβλέψει σωστά το *επόμενο κλικ του χρήστη εντός της συνεδρίας του (user session).* Η εξέταση της εκάστοτε συμπεριφοράς του χρήστη θα βασίζεται πάντα στην πιο πρόσφατη συνεδρία του με τη βοήθεια ενός κυλιόμενου χρονικού πλαισίου που κάθε φορά θα περιορίζεται σ' ένα συγκεκριμένο μόνο χρονικό διάστημα (π.χ. 30 λεπτά της ώρας), κατά το οποίο θα έχει πραγματοποιηθεί η διάδραση του χρήστη με τα στοιχεία. Τελικός στόχος μας θα είναι να προτείνουμε συναφή με τα ενδιαφέροντα του χρήστη άρθρα ειδήσεων.

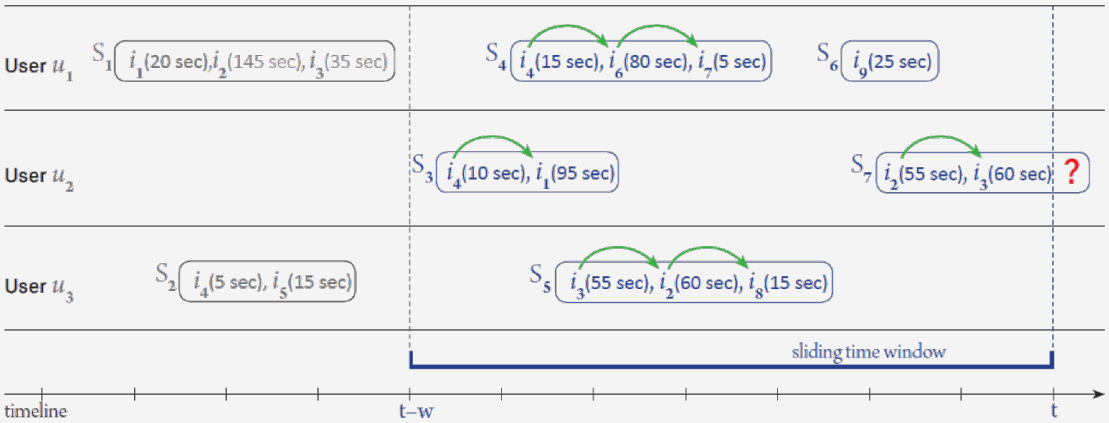
Οι *αλγόριθμοι συστάσεων με βάση τη συνεδρία του χρήστη (session-based recommendations)* αποτελούνται συνήθως από δύο *διαδικασίες (procedures).* Η πρώτη πραγματοποιεί τη συνεχή *επικαιροποίηση του προφίλ χρήστη.* Συγκεκριμένα, διαβάζει *στιγμιότυπα δεδομένων* από τη ροή των συνεδριών του χρήστη συνδυάζοντάς τα μεταξύ τους. Έτσι, το *προφίλ χρήστη* ενημερώνεται συνεχώς βάσει των άρθρων που αυτός επέλεξε και προκύπτουν από τη ροή συνεδριών S. Στη συνέχεια, ένα *κυλιόμενο χρονικό παράθυρο* μεγέθους w δηλώνει ότι η επεξεργασία των δεδομένων θα αφορά μόνο στα δεδομένα της χρονικής στιγμής t καθώς θα συμπεριλαμβάνει και όλα τα γεγονότα που δεν είναι παλαιότερα από $t - w$. Επομένως, η διαδικασία ενημέρωσης του *προφίλ χρήστη* θέτει ένα χρονικό διάστημα $(t - w, t]$ μέσα στο οποίο γίνεται αποδεκτή κάθε αλληλεπίδραση του χρήστη μ' ένα άρθρο. Ακολουθεί, τώρα, η δεύτερη διαδικασία: πρόκειται για τη μηχανή συστάσεων που αξιοποιεί το συνεχώς επικαιροποιούμενο προφίλ χρήστη για να προτείνει σ' αυτόν τα κορυφαία N άρθρα. Ο ακόλουθος πίνακας, 4.1, συνοψίζει ορισμένα βασικά σύμβολα που θα χρησιμοποιηθούν στη συνέχεια.

Σύμβολο	**Περιγραφή**
U	*Σύνολο χρηστών*, $U = \{U_1, U_2, ..., U_n\}$
I	*Σύνολο στοιχείων*, $I = \{i_1, i_2, ..., i_m\}$
S	*Σύνολο συνεδριών*, $S = \{S_1, S_2, ..., S_N\}$
S_u	*Σύνολο συνεδριών του χρήστη u*
S_i	*Σύνολο συνεδριών* στις οποίες ανήκει *το στοιχείο i*
S^{t_p}, S^{t_y}	*Σύνολο συνεδριών στο χρονικό παράθυρο* t_p ή t_y (έτος)
w	Το *μέγεθος του κυλιόμενου χρονικού παραθύρου*

Πίνακας 4.1: Τα σύμβολα του παραδείγματος και η περιγραφή τους.

Για να εξηγήσουμε καλύτερα τον *αλγόριθμο συστάσεων βάσει συνεδρίας χρήστη (session based recommendation),* θα χρησιμοποιήσουμε μια γραφική απεικόνιση των δεδομένων του παραδείγματός μας.

Παράδειγμα 4.2 Στο Σχήμα 4.1 εμφανίζονται 3 χρήστες και θέλουμε να προβλέψουμε το άρθρο που θα επιλέξει ο χρήστης u_2 στη συνεδρία S_7, η οποία όμως δεν έχει ακόμη ολοκληρωθεί. Επιπλέον, για τον υπολογισμό της *ομοιότητας* μεταξύ του υπό εξέταση χρήστη u_2 με τους δύο άλλους χρήστες, θα πρέπει να μη λάβουμε υπόψη τις συνεδρίες S_1 και S_2 λόγω του ότι βρίσκονται εκτός του *έγκυρου χρονικού διαστήματος* που έχει οριστεί ως $(t-w, t]$. Αυτό το *κυλιόμενο χρονικό παράθυρο* εξασφαλίζει το επίκαιρο των ειδησεογραφικών άρθρων, τα οποία, όπως προαναφέρθηκε, μπορεί να έχουν μικρή διάρκεια ζωής (π.χ. τα ειδησεογραφικά άρθρα της προηγούμενης ημέρας συνήθως τα θεωρούμε παλιά και ξεπερασμένα). Συνεπώς, θα πρέπει να προσπαθήσουμε να μην προτείνουμε στους χρήστες άρθρα που δεν είναι πολύ πρόσφατα. Από την άλλη μεριά, όταν δύο ή περισσότερα στοιχεία (εδώ, άρθρα) επιλέγονται από τον ίδιο χρήστη κατά την ίδια συνεδρία, αυτά πρέπει να θεωρηθούν περισσότερο συναφή σε σύγκριση με στοιχεία που επιλέχθηκαν από τον ίδιο σε προηγούμενες, διαφορετικές συνεδρίες. Για παράδειγμα, λαμβάνοντας υπόψη τη διάδραση του χρήστη u_1 με τα στοιχεία μπορούμε να συμπεράνουμε ότι το στοιχείο i_4 είναι πιο *συναφές* με τα στοιχεία i_6 και i_7 από ό,τι το στοιχείο i_9, καθώς αυτά επιλέχθηκαν στην ίδια συνεδρία (S_4) μαζί με το στοιχείο i_4.



Σχήμα 4.1: Γραφική απεικόνιση του παραδείγματός μας (*User-Items-Sessions*).

Τώρα, παρατηρήστε ξανά στο τρέχον παράδειγμά μας, το οποίο απεικονίζεται στο Σχήμα 4.1, ότι η συνεδρία S_7 του χρήστη u_2 είναι ενεργή και εξελίσσεται. Έτσι, τα στοιχεία i_2 και i_3 της συνεδρίας S_7 μπορούν να χρησιμοποιηθούν προκειμένου να γίνουν μελλοντικές συστάσεις στοιχείων στον χρήστη u_2. Όπως φαίνεται, ο χρήστης u_3 έχει επίσης επιλέξει τα ίδια στοιχεία (εντός του χρονικού διαστήματος του *έγκυρου χρονικού παραθύρου*) με τον u_2. Επιπρόσθετα, έχει επιλέξει το i_8, το οποίο θα μπορούσε να αποτελέσει μια καλή σύσταση για τον u_2. Σημειώστε ότι και ο χρήστης u_1 έχει επιλέξει ακριβώς τα ίδια στοιχεία με αυτά της συνεδρίας S_7 του u_2. Ωστόσο, δεν μπορούμε να χρησιμοποιήσουμε τη συνεδρία S_1 για την παραγωγή συστάσεων, καθώς δεν βρίσκεται εντός του *έγκυρου χρονικού παραθύρου*. Συ-

μπεραίνουμε, λοιπόν, ότι τα στοιχεία που επιλέγονται εντός της ίδιας συνεδρίας θεωρούνται πιο *όμοια* από αυτά που επιλέγονται από τον ίδιο χρήστη σε διαφορετικές συνεδρίες.

4.1.3 Πρόβλεψη Επόμενου Click Χρήστη (Next Click Prediction)

Όπως ήδη αναφέρθηκε στην Ενότητα 3.6 του αντίστοιχου κεφαλαίου, δοθέντος ενός στοιχείου με τιμές $\mathbf{X} = \langle x_1, \ldots, x_d \rangle$ μπορούμε κατά το *Θεώρημα του Bayes* να το κατατάξουμε σε μια *κλάση* (π.χ. *Like* ή *Dislike*) σύμφωνα με την Εξίσωση 4.1:

$$\arg\max_{\forall 1 \leq j \leq m} \Big\{ P(X = \langle x_1 \ldots x_d \rangle | C = c_j) \cdot P(C = c_j) \Big\} \tag{4.1}$$

Η Εξίσωση 4.1 υπολογίζει τις *δεσμευμένες πιθανότητες* της μορφής $P(X = \langle x_1 \ldots x_d \rangle | C = c_j)$ βάσει των d *χαρακτηριστικών* (*attributes*) ενός στοιχείου $\mathbf{X}$. Με βάση λοιπόν την *αφελή* (απλουστευτική) *παραδοχή του Bayes* (*naive Bayes inference*), η οποία θεωρεί ότι υπάρχει ανεξαρτησία μεταξύ των χαρακτηριστικών d ενός στοιχείου, μπορούμε να προβλέψουμε τα στοιχεία που θα συμπεριληφθούν σε μια συνεδρία S_N ενός χρήστη u βάσει των στοιχείων που έχουν ήδη συμπεριληφθεί στην S_N. Μ' άλλα λόγια, θεωρούμε ότι τα d χαρακτηριστικά του κάθε υπό εξέταση στοιχείου (στο παράδειγμά μας, άρθρων) είναι ανεξάρτητα μεταξύ τους. Οπότε, σ' αυτήν την περίπτωση ισχύει ότι:

$$P(X = \langle x_1 \ldots x_d \rangle | C = c_j) \;=\; \prod_{i=1}^{d} P(X = x_i | C = c_j)$$

Συνεπώς, η αρχική Εξίσωση 4.1 μετατρέπεται ως εξής:

$$\arg\max_{\forall 1 \leq j \leq m} \left\{ \prod_{i=1}^{d} P(X = x_i | C = c_j) \cdot P(C = c_j) \right\} \tag{4.2}$$

Στη συνέχεια, προσαρμόζουμε τις μεταβλητές του παραπάνω μαθηματικού τύπου στα σύμβολα που χρησιμοποιούμε στο τρέχον παράδειγμα, ώστε να υπολογίσουμε τις *πιθανότητες μετάβασης* (**transition probabilities**) μεταξύ δύο οποιωνδήποτε επόμενων στοιχείων σε κάθε ξεχωριστή συνεδρία και πάντα εντός του *χρονικού παραθύρου* t_p ως ακολούθως:

$$p\Big(j \in S_N | i_{1:m} \in S_N)\Big) \propto \prod_{\substack{i_k \in S_N \\ k=1..m}} p(j \in S_N | i_k \in S_N) \tag{4.3}$$

όπου i_k το σύνολο των στοιχείων στα οποία ο χρήστης u έχει ήδη κάνει κλικ στην τρέχουσα συνεδρία S_N, και j το στοιχείο για το οποίο θέλουμε να προβλέψουμε εάν θα πρέπει να είναι το επόμενο προτεινόμενο σ' αυτήν (S_N). Τονίζεται βέβαια ότι η *ακρίβεια πρόβλεψης βάσει του naive Bayes inference* επηρεάζεται αρνητικά από το γεγονός ότι σε πραγματικά δεδομένα σχεδόν πάντοτε υπάρχουν εξαρτήσεις μεταξύ των χαρακτηριστικών d ενός στοιχείου i.

Με χρήση της Εξίσωσης 4.3, λοιπόν, μπορούμε να φτιάξουμε έναν **"Πίνακα Πιθανοτήτων Μετάβασης" (Transition Probability Matrix)**, ο οποίος θα αποτυπώνει την *πιθανότητα μετάβασης από ένα στοιχείο προς τα υπόλοιπα* κατά τη συνεδρία ενός χρήστη. Εναλλακτικά, μπορούμε απλώς να μετρήσουμε πόσο συχνά είδαν οι χρήστες το στοιχείο i_b αμέσως μετά την προβολή του i_a.

Έστω, επομένως, ότι η συνεδρία S_n του υπό εξέταση χρήστη περιέχει ένα σύνολο $S_n = (i_1, i_2, ..., i_m)$ χρονικά διατεταγμένων στοιχείων με τα οποία αυτός αλληλεπίδρασε, και ένα σύνολο S όλων των συνεδριών $S = \{S_1, S_2, ..., S_N\}$. Με δεδομένο ότι η τρέχουσα συνεδρία S_N ενός χρήστη με i_m είναι το τελευταίο στοιχείο στο S_N, μπορούμε να ορίσουμε τη βαθμολογία που προβλέπουμε για ένα υποψήφιο προς σύσταση στοιχείο j ως εξής:

$$p(j \in S_N | i_m \in S_N) = score(j, i_m) = \frac{1}{\sum_{\substack{S_n \in S \\ n=1..N}} \sum_{\substack{i_k \; in S_n \\ k=1..m-1}} isSame(i_m, i_k)} \cdot \sum_{\substack{S_n \in S \\ n=1..N}} \sum_{\substack{i_k \in S_n \\ k=1..m-1}} isSame(i_m, i_k) \cdot isSame(j, i_{k+1}) \quad (4.4)$$

όπου η *συνάρτηση* $isSame(i_a, i_b)$ υποδεικνύει ότι τα i_a και i_b αφορούν στο ίδιο στοιχείο όπως φαίνεται παρακάτω:

$$isSame(i_a, i_b) = \begin{cases} 1, & \text{if } i_a = i_b; \\ 0, & \text{if } i_a \neq i_b. \end{cases}$$

Εφαρμόζοντας ως εκ τούτου την Εξίσωση 4.4 στα δεδομένα του Σχήματος 4.1, η *πιθανότητα μετάβασης* από το στοιχείο i_4 στο στοιχείο i_6 είναι ίση με $\frac{1}{2}$ όπως προκύπτει και από τον Πίνακα 4.2. Και αυτό συμβαίνει επειδή σε όλες τις συνεδρίες του *χρονικού παραθύρου* t_p υπάρχει μόνο μία περίπτωση όπου το i_4 ακολουθείται από το i_6 (συνεδρία S_4), ενώ ο παρονομαστής είναι ίσος με δύο αφού υπάρχουν δύο συνεδρίες στις οποίες το i_4 ακολουθείται από οποιοδήποτε άλλο στοιχείο (συνεδρίες S_3 και S_4). Ο *πίνακας πιθανοτήτων μετάβασης* του τρέχοντος παραδείγματός μας παρουσιάζεται στη συνέχεια (Πίνακας 4.2) - να σημειωθεί ότι οι γραμμές και οι στήλες, οι οποίες έχουν μηδενικές τιμές για όλες τις καταχωρίσεις τους, δεν εμφανίζονται.

	i_1	i_2	i_3	i_6	i_7	i_8
i_2	0	0	0.5	0	0	0.5
i_3	0	1	0	0	0	0
i_4	0.5	0	0	0.5	0	0
i_6	0	0	0	0	1	0

Πίνακας 4.2: *Πίνακας πιθανοτήτων μετάβασης* για το τρέχον παράδειγμα.

Συνοψίζοντας, με βάση την *απλουστευτική παραδοχή του Bayes (naive Bayes Inference)* μπορούμε να προβλέψουμε την *πιθανότητα μετάβασης* μεταξύ οποιουδήποτε ζεύγους στοιχείων εντός κάθε συνεδρίας χρήστη ανεξάρτητα από άλλες συνεδρίες.

4.1.4 Δημιουργία Λίστας με Προτεινόμενα Στοιχεία

Σε συνέχεια των παραπάνω, η μηχανή συστάσεων θα παράξει μια *λίστα με προτεινόμενα άρθρα για τον υπό εξέταση χρήστη* με βάση τον ήδη δημιουργηθέντα *Transition Probability Matrix - TPM* (για τον *πίνακα πιθανοτήτων μετάβασης* βλ. προηγούμενη υποενότητα). Ως προς τον υπό εξέταση χρήστη u, επομένως, η μηχανή συστάσεων ελέγχει το σύνολο των στοιχείων που αυτός είδε πρόσφατα $I_{t_p,u}$ (δηλαδή αυτά με τα οποία αλληλεπίδρασε κατά την τρέχουσα χρονική περίοδο t_p), και υπολογίζει το K_i, που είναι το σύνολο των k *πλησιέστερων στοιχείων* για κάθε στοιχείο i που ανήκει στο $I_{t_p,u}$. Έπειτα, για τον χρήστη u του *χρονικού διαστήματος* t_p και για κάθε στοιχείο j που είναι υποψήφιο να προταθεί υπολογίζουμε ένα *σκορ κατάταξής* του, $score(t_p, u, j)$, ως εξής:

$$score(t_p, u, j) = \sum_{i \in I_{t_p,u}} TPM(i, j) * \mathbf{1}_{(j,K_i)} \tag{4.5}$$

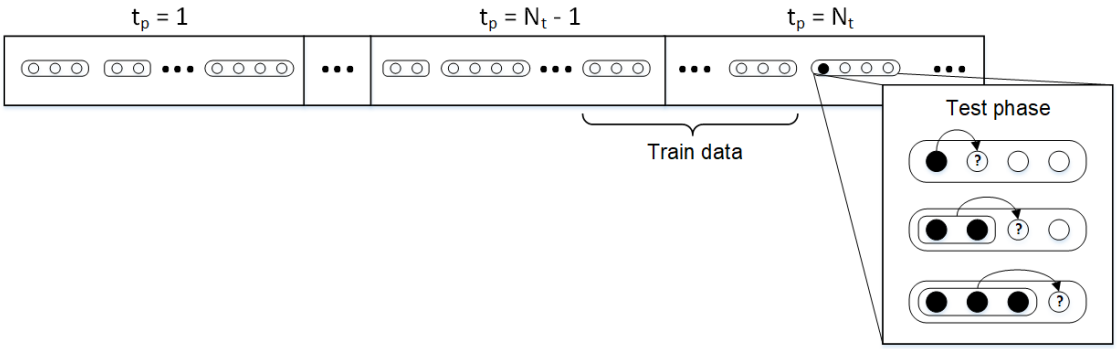
όπου $\mathbf{1}_{(j,K_i)}$ μια *συνάρτηση-δείκτης* που ισούται με 1 εάν το στοιχείο j βρίσκεται εντός των k *πλησιέστερων γειτόνων* του στοιχείου i, και με 0 όταν δεν ισχύει η παραπάνω συνθήκη. Επιπλέον, η TPM(i,j) είναι μια συνάρτηση που επιστρέφει ως αποτέλεσμα την *πιθανότητα μετάβασης* από ένα στοιχείο i προς ένα στοιχείο j με βάση τον ανάλογο *πίνακα πιθανοτήτων μετάβασης*. Τέλος, για τον υπό εξέταση χρήστη u κατατάσσουμε τα στοιχεία κατά φθίνουσα βαθμολογία και του προτείνουμε τα κορυφαία-N στοιχεία.

4.1.5 Πρωτόκολλο Αξιολόγησης Συστημάτων Συστάσεων βάσει Συνεδρίας Χρήστη

Στην παρούσα υποενότητα παρουσιάζουμε ένα *πρωτόκολλο αποτίμησης της ποιότητας των αλγορίθμων* που λαμβάνουν υπόψη τη *χρονική στιγμή της αλληλεπίδρασης* ενός χρήστη μ' ένα στοιχείο. Το πρωτόκολλο αυτό είναι παρόμοιο μ' αυτό που εισήγαγαν ο Jannach και οι συνεργάτες του ([Jannach et al., 2017], [Ludewig and Jannach, 2018]) για την *πρόβλεψη του επόμενου στοιχείου εντός μιας συνεδρίας χρήστη*, το οποίο είναι γνωστό και ως *Πρωτόκολλο Αξιολόγησης Διαδοχικών Προβλέψεων (Prequential Evaluation)* ([Quadrana et al., 2018], [Vinagre et al., 2014]).

Αναλυτικότερα, όπως φαίνεται στο Σχήμα 4.2 η διάσταση του χρόνου χωρίζεται σε N_t ίσες χρονικές περιόδους t_p, και στη συνέχεια τα δεδομένα διαχωρίζονται με τέτοιον τρόπο ώστε σε κάθε χρονική περίοδο να υπάρχουν μόνο οι *συνεδρίες χρήστη* που πραγματοποιήθηκαν κατά τη διάρκειά της. Εφαρμόζουμε αυτόν τον *διαχωρισμό των δεδομένων*, για να συγκεντρώσουμε αργότερα τα αποτελέσματα της αξιολόγησης της επίδοσης του αλγορίθμου συστάσεων για κάθε χρονική περίοδο ξεχωριστά, προκειμένου να μπορέσουμε να αντιληφθούμε τυχόν αλλαγές στη συμπεριφορά των

χρηστών στο *προκαθορισμένο χρονικό πλαίσιο*. Προσέχουμε εδώ ότι t_p είναι το μέγεθος του *χρονικού ορίζοντα* για τον οποίο θέλουμε να μπορούμε να κάνουμε προβλέψεις συστάσεων. Άρα, το μέγεθος t_p αφορά στο πόσο μακρυά στο μέλλον θα προσπαθήσουμε να προβλέψουμε αλληλεπιδράσεις μεταξύ χρηστών και στοιχείων (π.χ. μερικές ώρες ή μερικές ημέρες στο μέλλον). Ακολούθως, χρη-



Σχήμα 4.2: *Πρωτόκολλο αξιολόγησης διαδοχικών προβλέψεων (Prequential Evaluation).*

σιμοποιούμε επίσης την παράμετρο w για να καθορίσουμε το μέγεθος του *χρονικού παραθύρου* στο οποίο εκπαιδεύεται το μοντέλο προβλέψεών μας. Παρατηρήστε στο Σχήμα 4.2 ότι τα δεδομένα που ανήκουν εντός του w αναφέρονται ως *Train data*. Η παράμετρος w ελέγχει πόσο πίσω στο παρελθόν πηγαίνουμε για να αξιοποιήσουμε πληροφορίες που αφορούν στη διάδραση χρηστών με στοιχεία. Σημειώστε ότι, εάν το w είναι πολύ μεγάλο, τότε το σύστημα συστάσεων δεν θα είναι ευαίσθητο στις *ξαφνικές αλλαγές συμπεριφοράς των χρηστών* (*concept drifts*). Από την άλλη μεριά, εάν είναι πολύ μικρό, τότε δεν θα υπάρχουν αρκετά δεδομένα για να φτιαχτεί ένα μοντέλο που να *προβλέπει* αποτελεσματικά τα *επόμενα στοιχεία της συνεδρίας ενός χρήστη*. Επιπρόσθετα, στη φάση αποτίμησης του μοντέλου για την αξιολόγηση των *δεδομένων ελέγχου (test data)* χρησιμοποιούμε την παράμετρο v προκειμένου να καθορίσουμε πόσες αλληλεπιδράσεις ενός χρήστη με στοιχεία από την (τρέχουσα) υπό αξιολόγηση συνεδρία θα χρησιμοποιήσουμε, ώστε να επιχειρήσουμε στη συνέχεια να προτείνουμε ένα στοιχείο. Όπως φαίνεται στο Σχήμα 4.2, και συγκεκριμένα στο πλαίσιο *Test Phase*, αφού ελέγξουμε την πρώτη πρόβλεψη, το στιγμιότυπο των δεδομένων προστίθεται στα δεδομένα εκπαίδευσης του μοντέλου και η διαδικασία επαναλαμβάνεται μέχρι να σταματήσουν να υπάρχουν άλλα στιγμιότυπα δεδομένων στη συνεδρία. Τέλος, αξιολογούμε την *ακρίβεια* που πετυχαίνουμε όταν προτείνουμε τα *κορυφαία-N στοιχεία* για την πρόβλεψη του επόμενου στοιχείου κατά τη συνεδρία του υπό εξέταση χρήστη. Η *ακρίβεια (precision)* προκύπτει από τον λόγο του αριθμού των επιτυχημένων προβλέψεων προς τον συνολικό αριθμό των προτεινόμενων στοιχείων.

4.1.6 Σύγκριση Πρωτοκόλλων Αξιολόγησης ως προς την Εξάρτησή τους ή μη από την Παράμετρο του Χρόνου

Στο παρόν θα περιγράψουμε το πλαίσιο για τη δημιουργία ενός *χρονικά εξαρτώμενου* (*time dependent*) πρωτοκόλλου αξιολόγησης των συστημάτων συστάσεων. Στο σημείο αυτό τονίζουμε

ότι τα παραδοσιακά συστήματα συστάσεων συνήθως χρησιμοποιούν ένα πρωτόκολλο αξιολόγησης που είναι *ανεξάρτητο (time independent) από τη διάσταση του χρόνου*. Παραδείγματος χάρη, οι περισσότερες online βιβλιοθήκες συστημάτων συστάσεων ανοικτού κώδικα (*MyMediaLite, LibRec* κ.λπ.) ακολουθούν την παραδοσιακή μεθοδολογία αξιολόγησης, η οποία αγνοεί εντελώς την *παράμετρο του χρόνου*, όπως προαναφέραμε. Σύμφωνα με την παραδοσιακή μεθοδολογία τα στοιχεία με τα οποία αλληλεπίδρασαν οι χρήστες χωρίζονται σε δύο σύνολα δεδομένων: (i) αυτά που χρησιμοποιούμε για την εκπαίδευση ενός μοντέλου πρόβλεψης, δηλαδή το *σύνολο των δεδομένων εκπαίδευσης* (*train data*), και (ii) αυτά που χρησιμοποιούμε για τον έλεγχο της ποιότητας του μοντέλου πρόβλεψης, δηλαδή το *σύνολο των δεδομένων ελέγχου* (*test data*). Επομένως τα πρώτα χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου συστάσεων, ενώ τα δεύτερα χρησιμεύουν ως *σημείο αναφοράς* (*ground truth*) για την αξιολόγηση της *ακρίβειας* του αλγορίθμου συστάσεων.

Παρακάτω λοιπόν θα αναφέρουμε δύο τεχνικές που χρησιμοποιούνται για τη δειγματοληψία των *δεδομένων εκπαίδευσης και ελέγχου*:

1. **Χρονικά ανεξάρτητη δειγματοληψία**: Εδώ, η *χρονική στιγμή* της αλληλεπίδρασης του χρήστη με το στοιχείο δεν λαμβάνεται καθόλου υπόψη. Μπορεί να χρησιμοποιηθούν κάποια κριτήρια επιλογής υποψηφίων στοιχείων, αλλά γενικά η εκλογή αυτών που θα τοποθετηθούν στα σύνολα δεδομένων εκπαίδευσης και δοκιμής γίνεται *βάσει τυχαιότητας* και *ανεξαρτήτως χρόνου*.

2. **Χρονικά εξαρτώμενη δειγματοληψία**: Στην περίπτωση αυτή τα στοιχεία κατατάσσονται με βάση τη *χρονική στιγμή* της αλληλεπίδρασης τους με τον χρήστη, καθώς και *βάσει της διαδοχής* τους. Έτσι, εκκινώντας από ένα σύνολο δεδομένων και μια χρονική στιγμή t, ορίζουμε το *χρονικό παράθυρο* του συνόλου των *δεδομένων εκπαίδευσης (train data)* ως W_{tr}, και το *χρονικό παράθυρο* του συνόλου *δεδομένων ελέγχου (test data)* ως W_{ts}. Το πρώτο (W_{tr}) υποδεικνύει πόσο συχνά το σύστημα αποτίμησης θα επικαιροποιείται με επιπρόσθετα *δεδομένα εκπαίδευσης* που θα προστίθενται σωρευτικά στο ήδη υπάρχον σύνολο εκπαίδευσης. Ενώ το δεύτερο (W_{ts}), το σύνολο δηλαδή των *δεδομένων ελέγχου*, θα εμπεριέχει τα «μελλοντικά» στοιχεία, τα οποία θα χρησιμοποιηθούν μόνο για την αξιολόγηση του μοντέλου πρόβλεψης.

Εν συνεχεία, ο Πίνακας 4.3 της ακόλουθης υποενότητας συνοψίζει τα πλεονεκτήματα και τα μειονεκτήματα των δύο πρωτοκόλλων αξιολόγησης.

4.1.7 Οι Προκλήσεις των Χρονικά Εξαρτώμενων Πρωτοκόλλων Αξιολόγησης

Όταν λαμβάνουμε υπόψη τη *διάσταση του χρόνου*, οι συστάσεις στοιχείων μπορούν να αξιολογηθούν με δύο τρόπους: (α) Βάσει *μίας μόνο χρονικής στιγμής* (*single-time split evaluation*) ή (β) με διαχωρισμό των δεδομένων βάσει *πολλαπλών χρονικών στιγμών* (*multiple-time splits evaluation*). Αναλυτικότερα:

Μέθοδος Δειγματοληψίας	Πλεονεκτήματα	Μειονεκτήματα
Χρονικά Ανεξάρτητη	• Απαιτούνται λιγότερα δεδομένα. • Λιγότεροι υπολογισμοί.	• Μη αντιπροσωπευτική της πραγματικότητας. • Μη χρονικά *train-test splits.*
Χρονικά Εξαρτώμενη	• Αντιπροσωπευτική της πραγματικότητας. • Χρονικά *train-test splits.*	• Απαιτούνται *Χρονοσφραγίδες* • Περισσότεροι υπολογισμοί.

Πίνακας 4.3: Σύγκριση μεταξύ *Χρονικά εξαρτώμενων* και *Χρονικά ανεξάρτητων* πρωτοκόλλων αξιολόγησης.

(α) Η αξιολόγηση με διαχωρισμό δεδομένων βάσει *μίας μόνο χρονικής στιγμής* εξετάζει (χρονικά) μία μόνο διάσπασή τους, και χρησιμοποιώντας τα δεδομένα του παρελθόντος επιδιώκει να προβλέψει εκείνα που βρίσκονται μετά από αυτό το χρονικό σημείο. Ειδικότερα, ορίζουμε ένα χρονικό σημείο και χρησιμοποιούμε όλα τα στοιχεία με τα οποία ήδη αλληλεπίδρασε ο χρήστης νωρίτερα για την εκπαίδευση του μοντέλου μας καθώς και όλα τα στοιχεία που βρίσκονται μετά από αυτό το χρονικό σημείο για να ελέγξουμε το μοντέλο πρόβλεψής μας. Αυτή η προσέγγιση βασίζεται στην παραδοχή ότι τα προγενέστερα χρονικά δεδομένα επαρκούν για την πρόβλεψη όλων των δεδομένων που θα προκύψουν στο μέλλον.

(β) Στην αξιολόγηση όμως με διαχωρισμό των δεδομένων βάσει *πολλαπλών χρονικών στιγμών*, χωρίζουμε τα δεδομένα σε διαφορετικές *χρονικές περιόδους (χρονικές θυρίδες)* ίσου μεγέθους. Στη συνέχεια, χρησιμοποιούμε τα δεδομένα από την προηγούμενη *χρονική θυρίδα* για να προβλέψουμε τα δεδομένα της επόμενης. Το δεύτερο αυτό πρωτόκολλο αξιολόγησης προσομοιώνει καλύτερα τα πραγματικά συστήματα συστάσεων, επειδή είναι πιο ευαίσθητο στην αποτύπωση των αλλαγών στη συμπεριφορά των χρηστών σε βάθος χρόνου. Ωστόσο, η *αξιολόγηση πολλαπλών χρονικών διαχωρισμών* μπορεί να αντιμετωπίσει τις ακόλουθες προκλήσεις και ως εκ τούτου δυσκολίες:

- **1η Πρόκληση: Ο καθορισμός του μεγέθους του χρονικού διαστήματος που θα ορίσουμε ως "μέλλον" και όπου θα βρίσκονται τα δεδομένα ελέγχου (test data) που προσπαθούμε να προβλέψουμε.** Στην αξιολόγηση με διαχωρισμό των δεδομένων βάσει *πολλαπλών χρονικών στιγμών*, τα δεδομένα χωρίζονται σε πολλές *χρονικές θυρίδες*, όπως ήδη εξηγήσαμε. Έτσι, οι χρήστες που εμφανίζονται στην τρέχουσα *χρονική θυρίδα* t_i μπορεί να μην εμφανιστούν στην επόμενη *χρονική θυρίδα* $t_i + 1$, πράγμα που σημαίνει ότι δεν θα είμαστε σε θέση να συμπεριλάβουμε αυτούς τους χρήστες στην αξιολόγησή μας. Ωστόσο, οι ίδιοι χρήστες μπορεί να εμφανιστούν σε μεταγενέστερες από την $t_i + 1$ *χρονικές θυρίδες*. Άρα, ένας απλός τρόπος για να ξεπεράσουμε αυτό το πρόβλημα είναι να θεωρήσουμε ως "μέλλον" (*test data*) για την αξιολόγηση του συστήματος συστάσεων όλες τις μελλοντικές αλληλεπιδράσεις των χρηστών.

- **2η Πρόκληση: Οι "χρήστες μιας μεμονωμένης επίσκεψης" ή αυτοί που επισκέπτονται πολύ σπάνια μια ιστοσελίδα αλληλεπιδρώντας επομένως και με την εκεί υπηρεσία συστάσεων**. Αναλυτικότερα, σε σημαντικό αριθμό ιστοσελίδων πολλοί χρήστες εμφανίζονται εξαιρετικά σπάνια ή ακόμη και μία μόνο φορά. Στο σύνολο δεδομένων *MovieLens Hetrec 2011* για παράδειγμα, τριακόσιοι χρήστες (περίπου το 14% δηλαδή του συνόλου) εμφανίζονται μόνο μία φορά. Ως εκ τούτου, είναι δυνατόν να υπάρξει περίπτωση όπου η αξιολόγηση να απαιτεί μία σύσταση στοιχείου για κάθε χρήστη που εμφανίζεται στο t_i, αλλά η σύσταση να μην μπορεί να αποτιμηθεί δεδομένου ότι ορισμένοι χρήστες ενδέχεται να μην εμφανιστούν ξανά. Αυτοί, επομένως, θα πρέπει να εξαιρούνται από τη διαδικασία αξιολόγησης.

- **3η Πρόκληση: Ο αποκαλούμενος "χρήστης ψυχρής εκκίνησης" (cold start user)**. Με τον όρο *ψυχρή εκκίνηση* αναφερόμαστε στην αδυναμία της μηχανής συστάσεων να ξεκινήσει να προτείνει στοιχεία εάν δεν υπάρχουν επαρκείς αλληλεπιδράσεις ενός χρήστη με τα στοιχεία του συστήματος συστάσεων. Το πρόβλημα λοιπόν είναι ότι πολλοί χρήστες επανεμφανίζονται ακανόνιστα και σε χρονικές στιγμές που απέχουν πολύ μεταξύ τους. Έτσι, εάν θέσουμε έναν χρονικό ορίζοντα αξιολόγησης που να λήγει πριν από την επόμενη επανεμφάνιση του χρήστη, τότε η αξιολόγησή μας δεν μπορεί να λάβει υπόψη της τον δεύτερο. Εάν πάλι ορίσουμε έναν εξαιρετικά ευρύ ορίζοντα αξιολόγησης για να εξασφαλίσουμε την επόμενη επανεμφάνιση τέτοιων χρηστών, τότε το σύστημα συστάσεων δεν θα είναι ευαίσθητο στον εντοπισμό των *ξαφνικών αλλαγών συμπεριφοράς* των χρηστών (*concept drifts*).

- **4η Πρόκληση: Ο αντίκτυπος των παρελθόντων αλληλεπιδράσεων**. Είναι σημαντικό να αποφασίσουμε πόσο πίσω στο παρελθόν θα πάμε για να αξιοποιήσουμε πληροφορίες που αφορούν στη διάδραση χρηστών με στοιχεία. Άρα, ή θα πρέπει να προσπαθήσουμε να αγνοήσουμε κάποια δεδομένα του παρελθόντος ή, εναλλακτικά, θα πρέπει πάντα να συσσωρεύουμε τα δεδομένα από το παρελθόν και να τα χρησιμοποιούμε όλα. Επιπλέον, μπορούμε να χρησιμοποιήσουμε έναν *συντελεστή εξασθένισης* για να μειώσουμε τη βαρύτητα των πολύ παλιών αλληλεπιδράσεων των χρηστών με τα στοιχεία. Για παράδειγμα, στη σύσταση ειδησεογραφίας σε χρήστες τα ανάλογα άρθρα μένουν στην επικαιρότητα για εξαιρετικά βραχύ χρονικό διάστημα. Άρα, ένας αλγόριθμος θα πρέπει είτε να παρακολουθεί όλες τις προηγούμενες αλληλεπιδράσεις του χρήστη είτε να προσπαθεί να μειώσει τις αλληλεπιδράσεις του με τα στοιχεία που βρίσκονται πολύ πίσω στο παρελθόν.

4.2 Συστήματα Συστάσεων βάσει Γεωγραφικής Θέσης

4.2.1 Εισαγωγή

Οι γρήγορες ταχύτητες που προσφέρουν τα *ευρυζωνικά ασύρματα δίκτυα* (*4G* και *5G*) και η δυνατότητα *εντοπισμού της γεωγραφικής θέσης του χρήστη (γεωγραφικές συντεταγμένες GPS)* οδήγησαν

στην εμφάνιση "έξυπνων" κινητών τηλεφώνων, ταμπλετών και ανάλογων συσκευών που επέτρεψαν την *πρόσβαση στον Παγκόσμιο Ιστό από οποιοδήποτε σημείο* (*ubiquitous access*). Σ' αυτήν τη νέα εποχή λοιπόν, οι χρήστες μπορούν να επωφεληθούν από την πρόσβαση σε υπηρεσίες μέσω των *κινητών συσκευών* (*mobile devices*) τους από οπουδήποτε όπως ήδη τονίσαμε.

Τα *συστήματα συστάσεων βάσει γεωγραφικής θέσης* (*Location aware Recommender Systems - LARS*), επομένως, μπορούν να θεωρηθούν ως ειδικές περιπτώσεις των *συστημάτων συστάσεων με βάση το πλαίσιο* (*Context Aware Recommender Systems - CARS*) εντός του οποίου κινείται ο χρήστης, που στη συγκεκριμένη περίπτωση είναι *ο τόπος όπου βρίσκεται. Το (γεωγραφικό) σημείο* λοιπόν όπου εντοπίζεται ο χρήστης μπορεί να επηρεάσει τη διαδικασία συστάσεων με τρεις τρόπους:

1. *Η γεωγραφική θέση ενός χρήστη* μπορεί να επηρεάζει σημαντικά τις προτιμήσεις του όσον αφορά στην κουλτούρα, στην ένδυση, στις διατροφικές του συνήθειες κ.λπ. Για παράδειγμα, μια ανάλυση του συνόλου δεδομένων *MovieLens* [Levandoski et al., 2012] έδειξε ότι η κορυφαία προτίμηση των χρηστών από το *Wisconsin* είναι οι *Πολεμικές* ταινίες, ενώ η κορυφαία προτίμηση των χρηστών από τη *Florida* είναι οι ταινίες *Επιστημονικής Φαντασίας*. Αυτού του είδους η συσχέτιση μεταξύ των προτιμήσεων του χρήστη και της περιοχής που κατοικεί είναι γνωστή και ως *προτίμηση χρήστη λόγω γεωγραφικής θέσης (preference locality)*. Στην περίπτωση αυτή *ο τόπος συνδέεται εγγενώς με τον χρήστη* αλλά όχι απαραίτητα και με τα στοιχεία με τα οποία αυτός αλληλεπιδρά.

2. Οι χρήστες "έξυπνων" κινητών τηλεφώνων συχνά θέλουν να επισκεφθούν π.χ. εστιατόρια ή χώρους ψυχαγωγίας που πιθανόν βρίσκονται κοντά στην *τρέχουσα γεωγραφική θέση* τους. Στην παραπάνω περίπτωση, όμως, είναι τα προτεινόμενα στοιχεία που συνδέονται με την υπό εξέταση τοποθεσία και όχι απαραίτητα και ο χρήστης. Επομένως, η σύνδεση μεταξύ των *σημείων ενδιαφέροντος* (εστιατόρια, χώροι αναψυχής κ.λπ.) και της *περιοχής* που αυτά ανήκουν είναι γνωστή ως **"συσχέτιση ταξιδιωτικών σημείων ενδιαφέροντος με γεωγραφική θέση χρήστη" (travel locality)**. Σχετική με τον όρο που προαναφέραμε είναι μια ερευνητική εργασία που έγινε για το σύνολο των δεδομένων του *Foursquare* [Levandoski et al., 2012] και έδειξε ότι το 45% των χρηστών ταξιδεύουν λιγότερο από 16 χιλιόμετρα για να επισκεφθούν ένα εστιατόριο στην περιοχή τους, ενώ το 75% των χρηστών διανύουν για τον ίδιο σκοπό λιγότερο από 80 χιλιόμετρα. Άρα, *ο τόπος συνδέεται άρρηκτα με το σημείο ενδιαφέροντος* (π.χ. εστιατόριο), όχι όμως υποχρεωτικά και ο χρήστης. Ως εκ τούτου, *είναι επιθυμητό να παρέχονται συστάσεις σημείων ενδιαφέροντος που θα βρίσκονται κοντά στη γεωγραφική θέση του υπό εξέταση χρήστη*. Αυτό επιτυγχάνεται με την επιβολή "ποινής" -μείωσης δηλαδή του σκορ- επί ενός συγκριτικά πιο απομακρυσμένου σημείου προς επίσκεψη, και η οποία αυξάνεται ανάλογα με την απόσταση που πρέπει να διανύσει κάποιος για να φτάσει σ' αυτό το *υποψήφιο προς σύσταση σημείο ενδιαφέροντος*. Πιο συγκεκριμένα, υπολογίζεται η *απόσταση* $d(i, j)$ μεταξύ της γεωγραφικής θέσης του υπό εξέταση χρήστη i και της γεωγραφικής θέσης του υποψήφιου

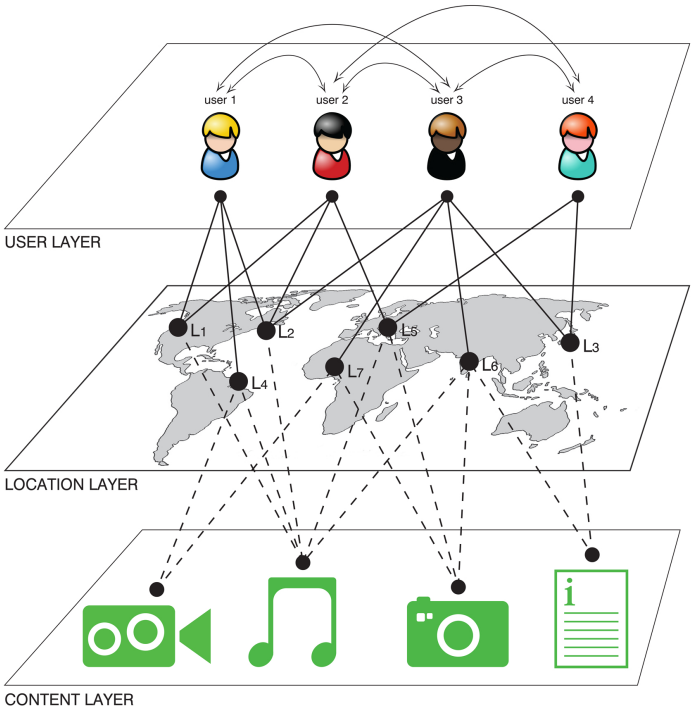
προς σύσταση στοιχείου j. Η βαθμολογία $\hat{r}_{ij}$ του χρήστη i για το στοιχείο j προβλέπεται αρχικά μ' έναν παραδοσιακό αλγόριθμο συστάσεων (π.χ. *συνεργατικού φιλτραρίσματος* ή *φιλτραρίσματος βάσει περιεχομένου*). Έπειτα, η αρχική βαθμολογία, που προβλέφθηκε από τον παραδοσιακό αλγόριθμο, σταθμίζεται με μια *συνάρτηση* $F(\cdot)$ της *απόστασης* $d(i,j)$. Η νέα αυτή βαθμολογία $\hat{r}^d_{ij}$ υπολογίζεται ως εξής:

$$\hat{r}^d_{ij} = \hat{r}_{ij} - F(d(i,j))$$

όπου $F(\cdot)$ μία μη φθίνουσα συνάρτηση της απόστασης $d(i,j)$, έτσι ώστε η *στάθμιση* να είναι κανονικοποιημένη ως προς την κλίμακα βαθμολογίας (π.χ. 1 έως 5 αστέρια) που δίνουν οι χρήστες προς τα *υποψήφια προς σύσταση στοιχεία* [Levandoski et al., 2012].

3. Τέλος, υπάρχει και η περίπτωση να συνδέονται *εγγενώς τόσο οι χρήστες όσο και τα στοιχεία με τις τοποθεσίες*. Ένας χρήστης που ταξιδεύει, λόγου χάρη, μπορεί να δημιουργήσει ένα προφίλ το οποίο θα υποδεικνύει τη διεύθυνση μόνιμης κατοικίας του. Ταυτόχρονα, μπορεί να καταγράφει τις βαθμολογίες του για τα υπό εξέταση εστιατόρια των περιοχών που επισκέπτεται. Έτσι, όταν άλλοι χρήστες θα πραγματοποιούν ερωτήματα από συγκεκριμένες γεωγραφικές θέσεις κατά τη διάρκεια των διακοπών τους, **τόσο η preference όσο και η travel locality θα διαδραματίζουν ταυτόχρονα ρόλο** στην επιλογή του προτεινόμενου σημείου ενδιαφέροντος.

Μια επιπρόσθετη δυνατότητα που προσφέρεται στα *location-aware συστήματα* είναι ότι οι χρήστες μπορούν να μοιράζονται πληροφορίες σχετικά και με την τοποθεσία τους καθώς και να ανταλλάσσουν πληροφορίες και μεταξύ τους, για να αξιοποιήσουν *συνεργατικά* τις εμπειρίες τους με τη χρήση των *κοινωνικών δικτύων βάσει γεωγραφικής θέσης* (*Location-based Social Networks - LBSNs*). Όπως φαίνεται και στο Σχήμα 4.3, οι χρήστες μπορούν να επισκέπτονται τοποθεσίες στον πραγματικό κόσμο και να αναρτούν περιεχόμενο (π.χ. σχόλια, φωτογραφίες, βίντεο) με ταυτόχρονη επισήμανση της γεωγραφικής τους θέσης. Αναλυτικότερα, στο Σχήμα 4.3 παρουσιάζονται 3 επίπεδα ανάλυσης (*επίπεδο χρήστη, τοποθεσίας* και *περιεχομένου*). Είναι προφανές ότι κάποιος μπορεί να εκμεταλλευτεί πληροφορίες από το κάθε επίπεδο ξεχωριστά, για να υπολογίσει πιθανές *ομοιότητες* μεταξύ των οντοτήτων του κάθε επιπέδου και βάσει αυτών να μπορεί να κάνει συστάσεις. Παραδείγματος χάρη, μπορούμε να υπολογίσουμε τη *γεωγραφική απόσταση* (δηλαδή την *ευκλείδεια απόσταση*) μεταξύ κάθε ζεύγους *σημείων ενδιαφέροντος (Points of Interest - POIs)* στο *επίπεδο γεωγραφικής θέσης (location layer)* του Σχήματος 4.3. Επιπλέον, μπορούμε να υπολογίσουμε την *ομοιότητα* μεταξύ των χρηστών με βάση το κοινωνικό δίκτυο που τους συνδέει στο *επίπεδο χρήστη (user layer)* του Σχήματος 4.3. Όσον αφορά τώρα στο *επίπεδο περιεχομένου (content layer)* του Σχήματος 4.3, μπορούμε να υπολογίσουμε την *ομοιότητα* μεταξύ των τύπων πληροφορίας (π.χ. βίντεο, φωτογραφίες κ.λπ.) με βάση τα *μεταδεδομένα* τους (π.χ. *tags*). Παρατηρήστε, επίσης, την τριμερή σχέση μεταξύ



Σχήμα 4.3: Οπτική αναπαράσταση των χρηστών, των τοποθεσιών και του περιεχομένου [π.χ. βίντεο, μουσική, φωτογραφίες, "ετικέτες"/επισημάνσεις (tags) κ.λπ.].

των οντοτήτων (π.χ. χρήστης, τοποθεσία, περιεχόμενο), η οποία διατρέχει όλα τα επίπεδα και μπορεί να χρησιμοποιηθεί και αυτή για τον υπολογισμό της *συνάφειας* μεταξύ των διαφορετικών τύπων των κόμβων του γράφου του Σχήματος 4.3.

Τέλος, τα *LBSNs* μπορούν να προσφέρουν 2 κατηγορίες συστάσεων: (α) *Γενικές* (μη εξατομικευμένες), και (β) *Εξατομικευμένες*.

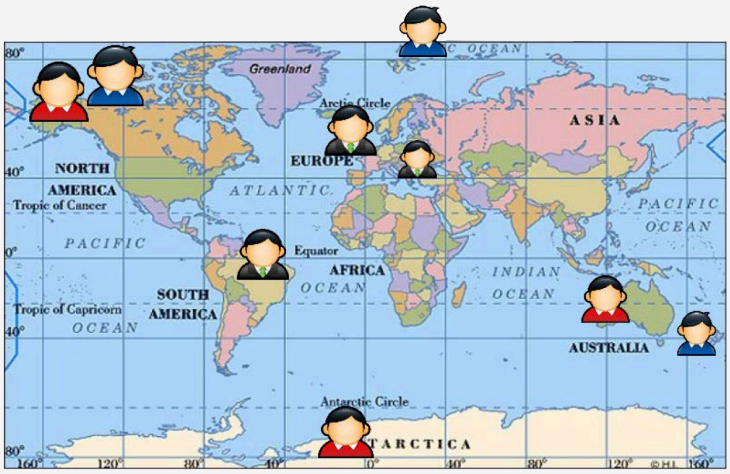
- Οι **Γενικές (μη-εξατομικευμένες) συστάσεις** βασίζονται σε *δημοφιλείς τοποθεσίες (POIs), εκδηλώσεις (events)* και *δραστηριότητες (activities)*. Αξίζει να αναφερθεί ότι για αυτό το είδος σύστασης δεν απαιτείται το *ιστορικό* των κινήσεων των χρηστών ή επιπρόσθετες πληροφορίες. Δηλαδή, σε κάθε χρήστη παρέχεται η ίδια σύσταση με βάση τις πιο δημοφιλείς τοποθεσίες οι οποίες προκύπτουν από τη συχνότητα που τις επισκέπτονται οι ίδιοι.

- Οι **Εξατομικευμένες συστάσεις** από την άλλη, χρησιμοποιούν πληροφορίες από τις *προτιμήσεις των χρηστών* και από το *ιστορικό* των επισκέψεών τους σε τοποθεσίες, και επομένως προτείνουν *σημεία ενδιαφέροντος* στους χρήστες με βάση το προφίλ του καθενός, κυρίως με τη μέθοδο του *Συνεργατικού Φιλτραρίσματος (CF)*. Συνάγουμε έτσι το συμπέρασμα ότι οι περισσότεροι αλγόριθμοι συστάσεων επιλέγουν την *τεχνική CF* για να προτείνουν τόπους, δραστηριότητες, προϊόντα ή ακόμα και φίλους.

4.3 Συστήματα Συστάσεων για LBSNs

Τα συστήματα συστάσεων για LBSNs μπορούν να βελτιώσουν τις συστάσεις τους για τοποθεσίες ή δραστηριότητες που πιθανόν να ενδιαφέρουν τον υπό εξέταση χρήστη αξιοποιώντας την τρέχουσα γεωγραφική του θέση ως βοηθητική πηγή πληροφόρησης. Ο αλγόριθμος σύστασης λοιπόν προσπαθεί να εξάγει την *ομοιότητα* μεταξύ χρηστών λαμβάνοντας υπόψη τις τοποθεσίες που έχουν ήδη επισκεφθεί.

Παράδειγμα 4.3 Στη συνέχεια θα χρησιμοποιήσουμε ένα απλό παράδειγμα που περιγράφει τη βασική διαδικασία παραγωγής συστάσεων στα *LBSNs*. Στο Σχήμα 4.4, λοιπόν, απεικονίζεται ένας παγκόσμιος χάρτης με τις τοποθεσίες που έχουν επισκεφθεί κατά το παρελθόν 3 χρήστες (Κόκκινος, Μαύρος και Μπλε). η Αρκτική (Βόρειος Πόλος) είτε η Νέα Ζηλανδία, τα οποία είναι μέρη που είχε ήδη επισκεφθεί ο Μπλε χρήστης.



Σχήμα 4.4: Οι τρεις χρήστες (με κόκκινο, μαύρο και μπλε χρώμα) και τα διάφορα μέρη που επισκέφθηκαν στο παρελθόν.

Ας υποθέσουμε ότι θέλουμε να συστήσουμε στον Κόκκινο χρήστη νέους προορισμούς για ταξίδια. Παρακάτω, ο Πίνακας 4.4 παρουσιάζει τις επισκέψεις των χρηστών σε διάφορα μέρη του πλανήτη μας. Όπως φαίνεται, τόσο ο Κόκκινος όσο και ο Μπλε χρήστης έχουν επισκεφθεί την Αλάσκα στη Βόρεια Αμερική. Επόμενο είναι λοιπόν, ότι ένας αλγόριθμος σύστασης θα τους θεωρούσε περισσότερο *συναφείς* (λόγω του ενός κοινού τους ταξιδιού) από ό,τι τον Κόκκινο με τον Μαύρο, δεδομένου ότι το δεύτερο ζευγάρι δεν έχει κανένα κοινό ταξίδι στο παρελθόν. Άρα, εάν έπρεπε να συστήσουμε έναν προορισμό στον Κόκκινο χρήστη, αυτός θα ήταν είτε η Αρκτική (Βόρειος Πόλος) είτε η Νέα Ζηλανδία, τα οποία είναι μέρη που είχε ήδη επισκεφθεί ο Μπλε χρήστης.

Τοποθεσίες	Κόκκινος Χρήστης	Μπλε Χρήστης	Μαύρος Χρήστης
Αλάσκα (Βόρεια Αμερική)	Ναι	Ναι	-
Ανταρκτική (Νότιος Πόλος)	Ναι	-	-
Αρκτική (Βόρειος Πόλος)	-	Ναι	-
Αυστραλία	Ναι	-	-
Νέα Ζηλανδία	-	Ναι	-
Βόρεια Ευρώπη	-	-	Ναι
Νότια Αμερική	-	-	Ναι
Νότια Ευρώπη	-	-	Ναι

Πίνακας 4.4: Οι επισκέψεις των τριών χρηστών σε διάφορα μέρη του πλανήτη. Παρατηρήστε ότι μόνο ο Κόκκινος και ο Μπλε χρήστης έχουν ένα κοινό ταξίδι - στην Αλάσκα της Βόρειας Αμερικής.

Σημειώστε, ακόμη, ότι υπάρχει μια σημασιολογική συσχέτιση μεταξύ των τοποθεσιών τις οποίες επισκέφτηκαν ο Κόκκινος και ο Μπλε χρήστης: μπορούμε να εξαγάγουμε το συμπέρασμα ότι οι δύο χρήστες ενδιαφέρονται είτε για κάποιον από τους 2 Πόλους της Γης (Αρκτική-Ανταρκτική) είτε για την Ωκεανία (Αυστραλία-Νέα Ζηλανδία). Επιπρόσθετα, είναι σημαντικό να τονίσουμε ότι τα υψηλά επίπεδα ιεράρχησης της γεωγραφικής διάστασης (τοποθεσία, πόλη, χώρα, ήπειρος κ.λπ.) θα μπορούσαν επίσης να διαδραματίσουν σημαντικό ρόλο σε μια μηχανή συστάσεων για *LBSNs*, δεδομένου ότι μπορούν να συσχετίσουν σημασιολογικά τοποθεσίες που σε χαμηλότερο επίπεδο ανάλυσης δεν φαίνονται *συναφείς*.

4.4 Τύποι Συστάσεων στα LBSNs

Εμβαθύνοντας, τώρα, στα *συστήματα συστάσεων για LBSNs* μπορούμε να τα κατηγοριοποιήσουμε με βάση τον τύπο των συστάσεων που παρέχουν σε 5 βασικές κατηγορίες: (α) *Σύσταση Φίλου* (*Friend Recommendation*), (β) *Σύσταση Τοποθεσίας* [*Point of Interest (POI) Recommendation*], (γ) *Σύσταση Διαδρομής/Δρομολογίου* (*Route/Itinerary Recommendation*) - συνίσταται από πολλά *διατεταγμένα POIs*, (δ) *Σύσταση Δραστηριότητας* (*Activity Recommendation*) και (ε) *Σύσταση Εκδήλωσης* (*Event Recommendation*). Στη συνέχεια θα παρουσιάσουμε καθέναν από αυτούς τους τύπους συστάσεων ξεχωριστά.

4.4.1 Σύσταση Φίλου

Στα *LBSNs* δύο χρήστες μπορούν να είναι διαδικτυακοί φίλοι, να μοιράζονται πληροφορίες και φωτογραφίες για μέρη που επισκέπτονται, διαδρομές ταξιδιού σε μια πόλη κ.λπ. Όσον αφορά στη *σύσταση φίλου*, ο υπό εξέταση χρήστης λαμβάνει μια λίστα με άλλους χρήστες που μπορεί να

γνωρίζει στην πραγματική του ζωή. Ωστόσο, τα *LBSNs* συνήθως εξετάζουν μόνο μονοπάτια μέγιστου μήκους 2 *βημάτων (hops)* -όπου ως βήμα λογίζεται η μία ακμή που συνδέει δύο κόμβους ενός γράφου- μεταξύ ενός χρήστη και των υποψηφίων φίλων του. Για παράδειγμα, στο *Facebook* προτείνονται μόνο οι φίλοι των φίλων μας. Από την άλλη πλευρά, άλλα ερευνητικά άρθρα εξετάζουν τη συνολική δομή των μονοπατιών ενός γράφου. Λόγου χάρη, οι Παπαδημητρίου, Συμεωνίδης και λοιποί συν-συγγραφείς πρότειναν τον *αλγόριθμο Friendlink* [Papadimitriou et al., 2011], ο οποίος παρέχει συστάσεις φίλων διασχίζοντας μονοπάτια τριών ή τεσσάρων *βημάτων (hops)* του κοινωνικού γράφου ενός χρήστη. Στην ίδια κατεύθυνση, η ακόλουθη ερευνητική εργασία, [Scellato et al., 2011], μελέτησε το *πρόβλημα της πρόβλεψης συνδέσμων* (*link prediction*) σ' έναν κοινωνικό γράφο και κατέληξε στο συμπέρασμα ότι περίπου το 30% των νέων συνδέσμων στον γράφο είναι μεταξύ των δύο κόμβων του, αυτού των "χρηστών" και αυτού των "τοποθεσιών", δηλαδή μεταξύ χρηστών που επισκέπτονται τα ίδια μέρη.

4.4.2 Σύσταση Τοποθεσίας ή Σημείου Ενδιαφέροντος (POI)

Όταν ένας χρήστης ταξιδεύει για πρώτη φορά σ' έναν τόπο, πιθανότατα θα θέλει να μάθει πού να πάει για καλό φαγητό ή ποια είναι τα σημαντικά αξιοθέατα. *Η σύσταση τοποθεσίας (POI)*, επομένως, μπορεί να παρέχει χρήσιμη καθοδήγηση σ' έναν χρήστη για την επίσκεψη σε ενδιαφέροντα μέρη. Η ερευνητική εργασία που ακολουθεί, [Ye et al., 2011], κατέδειξε ότι η *γεωγραφική συνάφεια μεταξύ των POIs* παίζει σημαντικό ρόλο στη συμπεριφορά του χρήστη ως προς τα μέρη που επισκέπτεται. Επιπλέον, η ίδια εργασία πρότεινε ένα ενοποιημένο πλαίσιο σύστασης σημείων ενδιαφέροντος, το οποίο συνδυάζει την προτίμηση του χρήστη σ' ένα *POI* με τη στενή γεωγραφική συσχέτισή του μ' αυτό. Μια δεύτερη ερευνητική εργασία, [Saez-Trumper et al., 2012], κατέδειξε ότι η σύσταση τοποθεσιών που βρίσκονται πλησιέστερα στο *γεωγραφικό σημείο ενδιαφέροντος* ενός χρήστη (π.χ. το σπίτι που αυτός διαμένει) δίνει αποτελέσματα που είναι εξίσου ακριβή με τον *αλγόριθμο συνεργατικού φιλτραρίσματος* που βασίζεται στην αλληλεπίδραση χρηστών με τοποθεσίες.

4.4.3 Σύσταση Διαδρομής/Δρομολογίου

Σ' αυτόν τον τύπο συστάσεων προτείνεται μια διαδρομή που συνίσταται από περισσότερες της μίας *διατεταγμένες τοποθεσίες*, τις οποίες δηλαδή μπορεί κάποιος να επισκεφτεί διαδοχικά σ' ένα συνήθως προκαθορισμένο χρονικό διάστημα. Για παράδειγμα, ένας χρήστης σε μια ξένη χώρα μπορεί να θέλει να επισκεφθεί τα πιο σημαντικά *POIs*, ενδεικτικά, αρχαιολογικούς χώρους ή μουσεία. Σε μια τέτοια περίπτωση ο καθορισμός ενός δρομολογίου με διατεταγμένες τοποθεσίες είναι πολύ χρήσιμος για την καθοδήγηση του προαναφερθέντος ενδιαφερομένου. Αυτό το είδος σύστασης χρησιμοποιεί πληροφορίες σχετικά με την παρούσα θέση του τρέχοντος χρήστη, το ιστορικό της γεωγραφικής του θέσης, το ιστορικό των φίλων του, καθώς και τον χρόνο που αυτός έχει στη διάθεσή του για να επισκεφθεί μια περιοχή ή πόλη.

4.4.4 Σύσταση Δραστηριότητας

Συνήθως ο τύπος μιας δραστηριότητας που ασκούμε σχετίζεται με συγκεκριμένα μέρη τα οποία πρέπει να επισκεφθούμε. Παραδείγματος χάρη, όταν ένας χρήστης κάνει check-in σ' ένα γήπεδο μπάσκετ αυτό μπορεί να σημαίνει ότι θα παίξει ο ίδιος ή ότι θα παρακολουθήσει έναν αγώνα. Με την ίδια λογική, όταν ο ίδιος χρήστης βρίσκεται κοντά σε γήπεδο μπάσκετ, η μηχανή συστάσεων μπορεί να προτείνει σ' αυτόν δραστηριότητες σχετικές με το άθλημα της καλαθοσφαίρισης. Σημειώστε ότι υπάρχουν ερευνητικές εργασίες που αξιοποιούν και άλλες *contextual πληροφορίες* (εκτός από την τοποθεσία) για να παρέχουν συστάσεις δραστηριοτήτων στους χρήστες. Η ακόλουθη ερευνητική εργασία, [Bellotti et al., 2008], λόγου χάρη, πρότεινε έναν οδηγό αναψυχής με την ονομασία *Magitti*. Η προτεινόμενη μέθοδος προβλέπει μια δραστηριότητα για έναν χρήστη βάσει του ευρύτερου πλαισίου στο οποίο αυτός βρίσκεται και με το οποίο αλληλεπιδρά (χρονική στιγμή, συσκευή, γεωγραφική θέση κ.λπ.), ενώ λαμβάνει υπόψη και τη συναισθηματική του διάθεση. Χρησιμοποιώντας αυτές τις πληροφορίες, ο οδηγός αναψυχής *Magitti* παρέχει αυτόματα διάφορες *συστάσεις δραστηριοτήτων* (π.χ. φαγητό, ψώνια, κινηματογράφο ή ανάγνωση ενός βιβλίου).

4.4.5 Σύσταση Εκδήλωσης

Πρόκειται για μια υποπερίπτωση της *σύστασης δραστηριότητας* και προτείνει στον υπό εξέταση χρήστη διάφορες εκδηλώσεις (συναυλίες, πάρτι, αθλητικές διοργανώσεις κ.λπ.) που λαμβάνουν χώρα είτε στον ψηφιακό κόσμο είτε σε μια συγκεκριμένη πόλη. Έτσι, πολλές ερευνητικές εργασίες στοχεύουν στον εντοπισμό εκδηλώσεων, όπως συναυλίες και φεστιβάλ, χρησιμοποιώντας πληροφορίες από *γεωγραφικές ετικέτες (tags)* που έχουν αναρτηθεί από μεγάλο αριθμό χρηστών [Lee et al., 2011]. Είναι παρατηρημένο άλλωστε ότι οι άνθρωποι που συμμετέχουν σ' ένα κοινωνικό γεγονός έχουν την τάση να αναρτούν επί τόπου αυξημένο όγκο πληροφοριών σχετικά μ' αυτό.

Σε άμεση σχέση με τα προαναφερθέντα η ακόλουθη ερευνητική εργασία, [Quercia et al., 2010], μελέτησε διαφορετικούς τρόπους *σύστασης εκδηλώσεων* στην ευρύτερη περιοχή της Βοστώνης. Συγκεκριμένα, συνέλεξε πληροφορίες για εκδηλώσεις που έλαβαν χώρα κατά τη διάρκεια μιας συγκεκριμένης περιόδου και τις συνδύασε με τις ληφθείσες συντεταγμένες *GPS* από ένα εκατομμύριο χρήστες κινητών τηλεφώνων. Οι συγγραφείς, αφού ανέλυσαν το παραπάνω σύνολο δεδομένων, μπόρεσαν να προβλέψουν με υψηλή *ακρίβεια* σε ποια κοινωνική εκδήλωση επρόκειτο να συμμετάσχουν οι κάτοικοι μιας περιοχής. Με βάση αυτά τα δεδομένα προσπάθησαν να παράσχουν γενικές *συστάσεις εκδηλώσεων* σε χρήστες χρησιμοποιώντας μόνο τη γεωγραφική θέση του σπιτιού τους. Δοκίμασαν για τον σκοπό αυτόν μια ποικιλία αλγορίθμων για τη σύσταση κοινωνικών εκδηλώσεων (π.χ. *δημοφιλείς εκδηλώσεις* γενικά, *δημοφιλείς εκδηλώσεις για μια συγκεκριμένη περιοχή* της πόλης, *γεωγραφικά κοντινές ως προς τον χρήστη εκδηλώσεις* κ.λπ.). Τα αποτελέσματα αυτής της μελέτης κατέδειξαν ότι ο πιο αποτελεσματικός αλγόριθμος είναι αυτός που προτείνει απλά δημοφιλείς εκδηλώσεις, ενώ ο λιγότερο αποτελεσματικός προτείνει εκδηλώσεις που είναι γεωγραφικά πολύ κο-

ντά στην περιοχή του υπό εξέταση χρήστη. Τέλος, για *εξατομικευμένες συστάσεις εκδηλώσεων* στους χρήστες του *Facebook*, η παρακάτω ερευνητική εργασία, [Kayaalp et al., 2009], πρότεινε μια *υβριδική* μέθοδο, η οποία συνδυάζει το *συνεργατικό φιλτράρισμα* με το *φιλτράρισμα βάσει περιεχομένου.*

4.5 Hybrid και Ensemble Μέθοδοι

Οι *Hybrid* και οι *Ensemble* μέθοδοι ακολουθούν τη διαδικασία του συνδυασμού αλγορίθμων αλλά με ελαφρώς διαφορετικό η καθεμία τρόπο. Οι *ensemble μέθοδοι* συνδυάζουν πολλαπλούς αλγορίθμους οι οποίοι προσπελαύνουν το ίδιο σύνολο *ομοιογενών* δεδομένων (π.χ. *δεδομένα περιεχομένου*). Αντίθετα, οι *υβριδικές* μέθοδοι συνδυάζουν πολλαπλούς αλγορίθμους οι οποίοι όμως προσπελαύνουν εντελώς *ετερογενή* δεδομένα που ως εκ τούτου προέρχονται από διαφορετικές πηγές πληροφορίας (π.χ. βίντεο, εικόνα, κείμενο κ.λπ.). Στις επόμενες υποενότητες θα παρουσιάσουμε την κάθε κατηγορία ξεχωριστά.

4.5.1 Υβριδικές Μέθοδοι

Τα *υβριδικά συστήματα συστάσεων* (*hybrid recommender systems*) χρησιμοποιούνται για να αξιοποιήσουν συνδυαστικά πολλαπλές πηγές δεδομένων (*δεδομένα χρήσης, δεδομένα περιεχομένου* κ.λπ.) αλλά και διαφορετικούς τύπους αλγορίθμων για την επεξεργασία τους. Ένα σημαντικό κίνητρο για την κατασκευή *υβριδικών* συστημάτων συστάσεων είναι ότι οι μεμονωμένοι τύποι συστημάτων συστάσεων (όπως αυτοί του *συνεργατικού φιλτραρίσματος*, ή του *φιλτραρίσματος βάσει περιεχόμενου* κ.λπ.) έχουν τα δικά τους πλεονεκτήματα και αδυναμίες. Για παράδειγμα, κατά την *ψυχρή εκκίνηση (cold start)* το σύστημα συστάσεων βάσει περιεχομένου λειτουργεί πιο αποτελεσματικά σε σχέση μ' αυτό του *συνεργατικού φιλτραρίσματος*. Και αυτό συμβαίνει επειδή κατά την εκκίνηση της μηχανής συστάσεων δεν έχουμε αρκετές αλληλεπιδράσεις των χρηστών με τα *στοιχεία (items)*, προκειμένου να συσχετίσουμε επαρκώς τους χρήστες μεταξύ τους και άρα να κάνουμε αποτελεσματικές προβλέψεις. Από την άλλη μεριά, η μηχανή *συστάσεων συνεργατικού φιλτραρίσματος* λειτουργεί πιο αποτελεσματικά όταν υπάρχουν επαρκή δεδομένα. Συμπερασματικά, τα *υβριδικά συστήματα συστάσεων* προσπαθούν να αξιοποιήσουν συνδυαστικά τα πλεονεκτήματα αυτών των επιμέρους συστημάτων με στόχο τη δημιουργία ενός ενοποιημένου συστήματος με μεγαλύτερη συνολικά αποτελεσματικότητα.

4.5.2 Ensemble Μέθοδοι

Σε αντίθεση με τα *υβριδικά (hybrid) συστήματα* οι *ensemble μέθοδοι* χρησιμοποιούνται για τη βελτίωση της αποτελεσματικότητας των συστημάτων συστάσεων με τέτοιον τρόπο ώστε οι επιμέρους μέθοδοι να αντλούν πάντα από τον ίδιο πίνακα βαθμολογίας χρηστών-στοιχείων και όχι από διαφορετικές πηγές δεδομένων. Άρα, οι επιμέρους μέθοδοι που συμμετέχουν για να συγκροτήσουν

την *ensemble μέθοδο* χρησιμοποιούν την ίδια πηγή δεδομένων, και όχι διαφορετικές. Ο βασικός στόχος πίσω από τη χρήση πολλών διαφορετικών μοντέλων είναι η ποικιλομορφία και συνακόλουθα η απομείωση της *μεροληψίας* που μπορεί να ενυπάρχει σε καθένα από τα παραπάνω ξεχωριστά. Ως εκ τούτου, οι διαφορετικοί τρόποι συνδυασμού των μοντέλων, όπως η *τεχνική bagging* ή η *τεχνική boosting*, μπορούν να προσαρμοστούν με σχετικά μικρές τροποποιήσεις για τη συγκρότηση μιας *ensemble μεθόδου*. Τονίζεται ότι με τον όρο *bagging* αναφερόμαστε στην ισότιμη συμμετοχή κάθε επιμέρους μεθόδου στη λήψη απόφασης, ενώ με τον όρο *boosting* ονομάζουμε την απόδοση σε κάθε επιμέρους μέθοδο διαφορετικών *σταθμισμένων συντελεστών σημαντικότητας* ως προς τη λήψη της τελικής απόφασης. Τέλος, αξίζει να σημειωθεί ότι οι κορυφαίες συμμετοχές στον διαγωνισμό *Netflix Prize 2008* [Koren et al., 2009] ήταν όλες *ensemble μέθοδοι*, γεγονός που συνηγορεί υπέρ της αυξημένης αποτελεσματικότητας αυτών των μεθόδων.

4.6 Τύποι Υβριδικών Συστημάτων

Τα *υβριδικά συστήματα* χωρίζονται βάσει του τρόπου σχεδιασμού τους (α) σε *μονολιθικά*, (β) σε *σειριακής διάταξης*, και (γ) σε *παράλληλης διάταξης*. Ακολουθεί αναλυτική παρουσίασή τους.

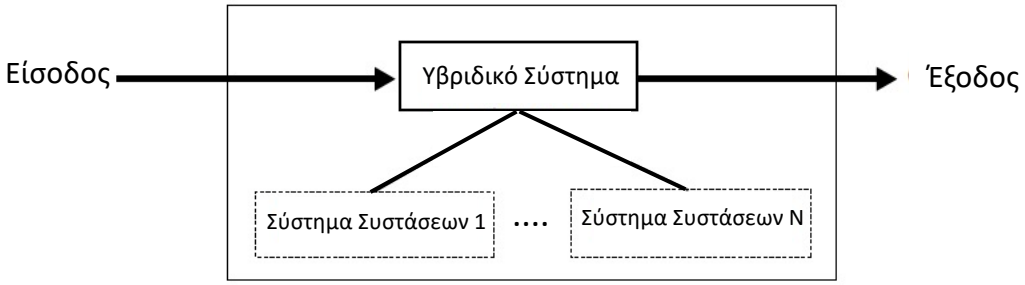
4.6.1 Μονολιθικά Υβριδικά Συστήματα

Τα *μονολιθικά υβριδικά συστήματα* ενσωματώνουν στον σχεδιασμό τους διαφορετικές τεχνικές σύστασης (π.χ. *φιλτάρισμα βάσει περιεχομένου* σε συνδυασμό με *συνεργατικό φιλτράρισμα*) με τέτοιον τρόπο, ώστε να μην μπορεί να υπάρξει σαφής διάκριση μεταξύ τους (Σχήμα 4.5). Έτσι, οι αλγόριθμοι συστάσεων που βασίζονται στο *συνεργατικό φιλτράρισμα* ή στο *φιλτράρισμα βάσει περιεχομένου* τροποποιούνται κατάλληλα, για να χρησιμοποιηθούν στο πλαίσιο του *μονολιθικού υβριδικού συστήματος*. Επιπροσθέτως, γίνονται μετατροπές και συνδυάζονται κατάλληλα οι διαφορετικές πηγές δεδομένων, για να χρησιμοποιηθούν στο πλαίσιο ενός ενιαίου συστήματος συστάσεων. Η προαναφερθείσα διαδικασία *συνδυασμού και ομογενοποίησης* των δεδομένων εισόδου (όποιου είδους, π.χ. *συνεργατικά* ή *περιεχομένου*) είναι γνωστή ως **"συνδυασμός χαρακτηριστικών" (feature combination)**. Ως εκ τούτου, σ' ένα *μονολιθικό υβριδικό σύστημα* δεν μπορεί να ξεχωρίσει εύκολα κάποιος τα επιμέρους συστήματα συστάσεων που το απαρτίζουν, δεδομένου ότι τα αρχικά ετερογενή δεδομένα *συγχωνεύονται και ομογενοποιούνται* όπως τονίσαμε σ' ένα σύνολο δεδομένων.

Για παράδειγμα, ένα *μονολιθικό σύστημα συστάσεων* μπορεί να απαρτίζεται από (α) ένα σύστημα συστάσεων που εδράζεται στο *φιλτράρισμα βάσει περιεχόμενου* και (β) από ένα σύστημα συστάσεων *συνεργατικού φιλτραρίσματος*, το οποίο, όπως έχουμε ήδη εξηγήσει, αξιοποιεί τα *συνεργατικά δεδομένα* της αλληλεπίδρασης των χρηστών με την υπηρεσία για τον προσδιορισμό της *ομοιότητας* μεταξύ των υποψήφιων προς σύσταση στοιχείων.

Συνοψίζοντας, τα *μονολιθικά υβριδικά συστήματα* αποτελούνται από μια ενιαία μηχανή συστάσεων που ενσωματώνει πολλαπλές διαφορετικές μεθόδους μετά από κατάλληλη προεπεξεργασία

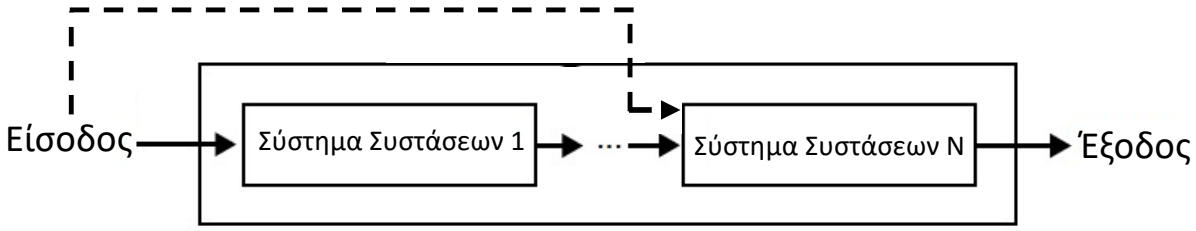
και με συνδυασμό διαφορετικών τύπων δεδομένων εισόδου. Η *υβριδοποίηση* αυτή επιτυγχάνεται με κατάλληλη τροποποίηση των *βημάτων* κάθε αλγορίθμου. Συνήθως χρησιμοποιούνται ειδικά *βήματα* για μια πρώτη επεξεργασία των δεδομένων και την ακόλουθη μετατροπή των δεδομένων εισόδου σε μια *αναπαράσταση* που μπορεί να αξιοποιηθεί από κάθε αλγόριθμο που συμμετέχει στο *μονολιθικό υβριδικό σύστημα.*



Σχήμα 4.5: Αρχιτεκτονική *Μονολιθικού Υβριδικού Συστήματος Συστάσεων.*

4.6.2 Υβριδικά Συστήματα Σειριακής Διάταξης

Στα *υβριδικά συστήματα σειριακής διάταξης* τα επιμέρους συστήματα συστάσεων τοποθετούνται σε σειρά το ένα μετά το άλλο, ώστε η έξοδος του ενός χρησιμοποιείται ως είσοδος στο επόμενο (Σχήμα 4.6). Το σύστημα συστάσεων που προηγείται στη *σειριακή διάταξη* μπορεί είτε (α) να προεπεξεργάζεται τα δεδομένα εισόδου για τη δημιουργία ενός μοντέλου πρόβλεψης βαθμολογιών χρήστη προς στοιχεία που αξιοποιείται από το επόμενο στάδιο είτε (β) να παραδίδει απευθείας στο επόμενο στάδιο μια λίστα συστάσεων με προτεινόμενα στοιχεία για περαιτέρω βελτίωσή της. Έτσι, η λίστα συστάσεων περιορίζεται μόνο σ' εκείνα τα στοιχεία που είχαν επίσης προταθεί στο προηγούμενο στάδιο. Εναλλακτικά, το *υβριδικό σύστημα σειριακής διάταξης* μπορεί να βελτιώνει διαδοχικά τις συστάσεις του προηγούμενου συστήματος με τη βοήθεια της *τεχνικής boosting* (π.χ. *αλγόριθμος AdaBoost*), με την οποία γίνεται η απόδοση σε κάθε επιμέρους μέθοδο διαφορετικών *σταθμισμένων συντελεστών σημαντικότητας* ως προς τη λήψη της τελικής απόφασης.



Σχήμα 4.6: Αρχιτεκτονική *Υβριδικού Συστήματος Σειριακής Διάταξης.*

Παράδειγμα 4.4 Έστω ότι δίνεται ένα *υβριδικό σύστημα σειριακής διάταξης* το οποίο αποτελείται από n επιμέρους συστήματα συστάσεων {**Σύστημα**$_1$ ...**Σύστημα**$_n$} και η τελική πρόβλεψη της βαθμολογίας του χρήστη u προς ένα στοιχείο j υπολογίζεται από το n-οστό σύστημα συστάσεων. Ωστόσο, ένα στοιχείο j θα μπορεί να είναι υποψήφιο προς σύσταση από το k-οστό σύστημα μόνο εάν το $(k-1)$-οστό σύστημά του έχει επίσης αποδώσει μη μηδενική βαθμολογία. Συνεπώς, για $\forall k \geq 2$ επιμέρους συνεχόμενα συστήματα συστάσεων που βρίσκονται σε σειριακή διάταξη θα πρέπει να ισχύει:

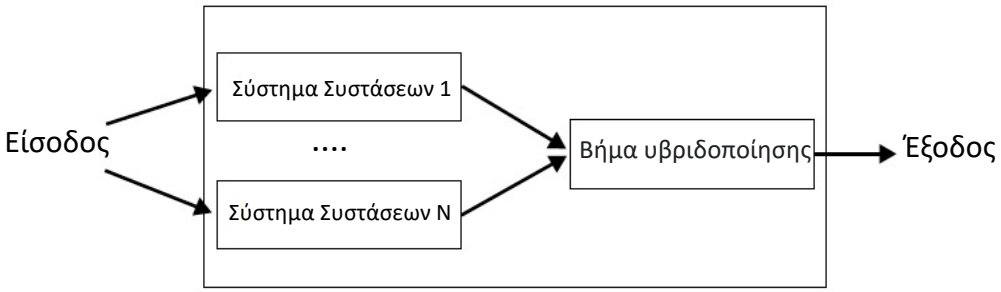
$$\hat{r}_{uj}^{k} = \begin{cases} \hat{r}_{uj}^{k} & : \hat{r}_{uj}^{k-1} \neq 0 \\ 0 & : \text{Διαφορετικά} \end{cases}$$

όπου $\hat{r}_{uj}^{k}$ η βαθμολογία του χρήστη u για το στοιχείο j του συστήματος συστάσεων k.

Έτσι, σ' ένα *υβριδικό σύστημα σειριακής διάταξης* όλα τα επιμέρους συστήματα συστάσεων, εκτός από το πρώτο, μπορούν να αλλάξουν τη σειρά της λίστας προτεινόμενων στοιχείων ή να αποκλείσουν ένα στοιχείο εφόσον προβλέπουν ότι η βαθμολογία του θα ισούται με 0. Δεν μπορούν ωστόσο να εισάγουν νέα στοιχεία που θα έχουν ήδη αποκλειστεί από ένα επιμέρους σύστημα συστάσεων σε προηγούμενο στάδιο. Κατά συνέπεια μπορεί να προκύψουν καταστάσεις στις οποίες το *υβριδικό σύστημα σειριακής διάταξης* δεν θα παράγει τον απαιτούμενο αριθμό προτεινόμενων στοιχείων (*χαμηλό item coverage*), μειώνοντας έτσι την αποτελεσματικότητα του συστήματος.

4.6.3 Υβριδικά Συστήματα Παράλληλης Διάταξης

Τα *υβριδικά συστήματα παράλληλης διάταξης* απαιτούν την ύπαρξη τουλάχιστον δύο ή περισσότερων επιμέρους συστημάτων συστάσεων τα οποία εφαρμόζονται παράλληλα (Σχήμα 4.7). Με βάση τις εισόδους τους, τα δύο ή περισσότερα επιμέρους συστήματα συστάσεων λειτουργούν ανεξάρτητα το ένα από το άλλο και παράγουν ξεχωριστές λίστες συστάσεων. Στο επόμενο βήμα γίνεται η *υβριδοποίηση*, κατά την οποία οι προαναφερθείσες λίστες συνδυάζονται σε μία έξοδο, άρα και σε μία τελική λίστα συστάσεων.



Σχήμα 4.7: Αρχιτεκτονική *Υβριδικού Συστήματος Παράλληλης Διάταξης*.

Στο σημείο αυτό είναι αναγκαίο να τονιστεί ότι η *υβριδοποίηση* γίνεται συνήθως με τη χρήση *συντελεστών βαρύτητας* που εφαρμόζονται σε κάθε επιμέρους σύστημα συστάσεων, προκειμένου να βελτιστοποιηθεί η προβλεπτική ικανότητά του. Σ' αυτήν την περίπτωση το *υβριδικό σύστημα* ονομάζεται *σταθμισμένο* (*weighted hybrid*). Η *υβριδοποίηση* μπορεί να υλοποιηθεί είτε (α) στο επίπεδο της πρόβλεψης της βαθμολογίας του χρήστη για τα στοιχεία και κατόπιν της κατάταξης των στοιχείων βάσει της προβλεπόμενης βαθμολογίας τους είτε (β) στο επίπεδο της απευθείας συγχώνευσης των επιμέρους διαφορετικών λιστών συστάσεων που δημιουργεί το κάθε σύστημα συστάσεων. Στη συνέχεια θα παραθέσουμε σχετικό παράδειγμα ανά κατηγορία.

Παράδειγμα 4.5 Έστω, λοιπόν, ότι μας δίνεται ένας δισδιάστατος πίνακας R μεγέθους $m \times n$ της βαθμολογίας χρηστών επί των στοιχείων. Στα *σταθμισμένα υβριδικά συστήματα*, όπως αναφέρθηκε προηγουμένως, τα επιμέρους συστήματα συστάσεων συνδυάζονται χρησιμοποιώντας ένα σύνολο *συντελεστών βαρύτητας*. Έστω, ακόμη, ότι $\hat{R}_1 \ldots \hat{R}_q$ είναι οι διαφορετικοί πίνακες βαθμολογίας χρηστών-στοιχείων που προβλέπει το κάθε επιμέρους σύστημα συστάσεων. Στη συνέχεια, για ένα σύνολο *συντελεστών βαρύτητας* $\beta_1 \ldots \beta_q$ το *σταθμισμένο υβριδικό σύστημα* προκύπτει ως ακολούθως:

$$\hat{R} = \sum_{i=1}^{q} \beta_i \times \hat{R}_i$$

Μπορούμε όμως να επαναδιατυπώσουμε την παραπάνω εξίσωση, για να υπολογίσουμε κάθε στοιχείο του πίνακα $\hat{R}$ ξεχωριστά:

$$\hat{r}_{uj} = \sum_{i=1}^{q} \beta_i \times \hat{r}^i_{uj}$$

όπου $\hat{r}^i_{uj}$ η βαθμολογία του χρήστη u για το στοιχείο j του συστήματος συστάσεων i, ενώ $\hat{r}_{uj}$ η τελική πρόβλεψη του *υβριδικού συστήματος*. Στην απλούστερη εκδοχή του *υβριδικού συστήματος* είναι δυνατόν να επιλέξουμε όλοι οι *συντελεστές βαρύτητας* να είναι ίσοι με $\beta_1 = \beta_2 = \ldots = \beta_q = 1/q$ και όπου q ο αριθμός των συστημάτων συστάσεων που συμμετέχουν σ' αυτήν.

Έστω, τώρα, ότι δύο συστήματα συστάσεων χρησιμοποιούνται για να προτείνουν μια *διατεταγμένη λίστα* από τρεις ταινίες σ' έναν υπό εξέταση χρήστη. Όπως γίνεται εύκολα αντιληπτό από τον Πίνακα 4.5 οι διαφορετικές λίστες συστάσεων του κάθε συστήματος *υβριδοποιούνται* με τη χρήση ενός *ομοιόμορφου σχήματος στάθμισης* με τους *συντελεστές βαρύτητας* να είναι ίσοι με $\beta_1 = \beta_2 = 0.5$. Όπως μπορείτε επίσης να παρατηρήσετε στον Πίνακα 4.5, στην τελευταία στήλη, η *σταθμισμένη υβριδική λίστα σύστασης* του Συστήματος$_{1+2}$ υπολογίζει μια διαφορετική κατάταξη των ταινιών (συνδυάζοντας τις προβλέψεις βαθμολογιών από το Σύστημα$_1$ και το Σύστημα$_2$) από ό,τι το κάθε σύστημα συστάσεων ξεχωριστά.

Σύστημα ⇒ Ταινία ⇓	Σύστημα$_1$ Πρόβλεψη	Σύστημα$_1$ Κατάταξη	Σύστημα$_2$ Πρόβλεψη	Σύστηματα$_2$ Κατάταξη	Σύστημα$_{1+2}$ Πρόβλεψη	Σύστημα$_{1+2}$ Κατάταξη
Ταινία1	0.5	1	0.8	2	0.65	1
Ταινία2	0.1	3	0.9	1	0.5	2
Ταινία3	0.3	2	0.4	3	0.35	3

Πίνακας 4.5: Πρόβλεψη βαθμολογίας και κατάταξη ταινιών με τη χρήση αφενός *μεμονωμένων* συστημάτων, και αφετέρου ενός *υβριδικού συστήματος συστάσεων*.

Η σχετική εμπειρία έχει δείξει παρόλα αυτά ότι στην πράξη είναι καλύτερο να *σταθμίζονται* τα συστήματα συστάσεων ως προς τη *σημαντικότητά* τους, ώστε να δίνεται μεγαλύτερη βαρύτητα σ' εκείνα που είναι αποτελεσματικότερα ως προς την *ακρίβεια* των προβλέψεών τους. Στη συνέχεια θα δούμε ένα παράδειγμα που αφορά στον *συγκερασμό λιστών κατάταξης* δύο διαφορετικών συστημάτων συστάσεων. Περισσότερες πληροφορίες για τους τρόπους *συγκερασμού* των βαθμολογιών αλλά και των λιστών κατάταξης δίνονται στην Ενότητα 10.1 του αντίστοιχου κεφαλαίου.

Παράδειγμα 4.6 Ας υποθέσουμε ότι ο υπό εξέταση χρήστης του προηγούμενου παραδείγματός μας είχε δει και τις τρεις ταινίες που του προτάθηκαν και του είχαν αρέσει και οι τρεις, και συνεπώς τις αξιολογεί όλες αναλόγως, δηλαδή με τη μέγιστη τιμή της κλίμακας βαθμολογίας που είναι ίση με τη μονάδα (1). Για να μειώσουμε τη διαφορά μεταξύ της πραγματικής βαθμολογίας του χρήστη και της προβλεπόμενης βαθμολογίας του *υβριδικού συστήματος* θα πρέπει να προσδιοριστούν οι βέλτιστοι *συντελεστές βαρύτητας*. Συνεπώς είναι απαραίτητο να μπορούμε να αξιολογήσουμε την αποτελεσματικότητα του κάθε δυνατού συνδυασμού *συντελεστών βαρύτητας* β_1 και β_2, προκειμένου να εντοπίσουμε τον καλύτερο συνδυασμό τιμών τους. Αυτό γίνεται με μια διαδικασία που είναι γνωστή ως *Grid Search* (*εξαντλητική αναζήτηση*). Αναλυτικότερα, επιλέγουμε εκείνον τον συνδυασμό των *συντελεστών βαρύτητας* που ελαχιστοποιεί μια επιλεγμένη μετρική, όπως το *μέσο απόλυτο σφάλμα* (*Mean Absolute Error - MAE*), το οποίο και υπολογίζεται ως ακολούθως:

$$MAE = \frac{1}{|R|} \sum_{r_{uj} \in R} \sum_{i=1}^{q} \beta_i \times \left| r_{uj} - \hat{r}^i_{uj} \right| \tag{4.6}$$

όπου $|R|$ ο αριθμός των βαθμολογιών που υπάρχουν στον αρχικό πίνακα βαθμολογίας χρήστη-στοιχείου, r_{uj} η πραγματική βαθμολογία, και $\hat{r}^i_{uj}$ η βαθμολογία που προβλέπει το σύστημα συστάσεων i. Ο Πίνακας 4.6 εμφανίζει το *μέσο απόλυτο σφάλμα (mean absolute error - MAE)* των προβλέψεων όταν συνδυάζουμε το Σύστημα$_1$ και το Σύστημα$_2$ για διαφορετικές τιμές των *συντελεστών βαρύτητας* β_1 και β_2 στο εύρος τιμών του διαστήματος [0.1-0.9]. Όπως φαίνεται στην τελευταία στήλη του Πίνακα 4.6, το *MAE (Mean Absolute Error)* προοδευτικά μειώνεται καθώς αυξάνουμε τον *συντελεστή βαρύτητας* β_2 για το Σύστημα$_2$ ενώ ταυτόχρονα μειώνουμε τον *συντελεστή βαρύτητας* β_1 για το Σύστημα$_1$.

Σύστημα$_1$ Συντελ. β_1	Σύστημα$_2$ Συντελ. β_2	Ταινία	Πραγματική Τιμή r_{uj}	Σύστημα$_1$ Πρόβλεψη Τιμής	Σύστημα$_2$ Πρόβλεψη Τιμής	Διαφορά μεταξύ Πραγματικής και Προβλεπόμενης Τιμής	MAE
0.9	0.1	Ταινία1	1	0.5	0.8	$1-(0.9*0.5$ $+0.1*0.8)=0.47$	0.66
		Ταινία2	1	0.1	0.9	0.82	
		Ταινία3	1	0.3	0.4	0.69	
0.7	0.3	Ταινία1	1	0.5	0.8	0.41	0.58
		Ταινία2	1	0.1	0.9	0.66	
		Ταινία3	1	0.3	0.4	0.67	
0.5	0.5	Ταινία1	1	0.5	0.8	0.35	0.5
		Ταινία2	1	0.1	0.9	0.5	
		Ταινία3	1	0.3	0.4	0.65	
0.3	0.7	Ταινία1	1	0.5	0.8	0.29	0.42
		Ταινία2	1	0.1	0.9	0.34	
		Ταινία3	1	0.3	0.4	0.63	
0.1	0.9	Ταινία1	1	0.5	0.8	0.23	0.34
		Ταινία2	1	0.1	0.9	0.18	
		Ταινία3	1	0.3	0.4	0.61	

Πίνακας 4.6: Πρόβλεψη βαθμολογιών χρήστη μέσω *δυναμικά εξελισσόμενης στάθμισης* (από 0.1 έως 0.9) των *συντελεστών βαρύτητας* (β_1 και β_2) που αποδίδουμε για τα δύο συστήματα συστάσεων και υπολογισμός του *Mean Absolute Error (MAE)*.

4.7 Ερωτήσεις Κεφαλαίου

1. Τί ονομάζουμε *συνεδρία χρήστη (user session)*; Να αναφέρετε μια τεχνική με την οποία να μπορεί να *προβλέψει κάποιος το επόμενο click του χρήστη* κατά τη διάρκεια μίας συνεδρίας.

2. Περιγράψτε το *πρωτόκολλο αξιολόγησης* των συστημάτων συστάσεων που *βασίζεται σε συνεδρίες χρηστών* και εξαρτάται από τη *διάσταση του χρόνου*.

3. Συγκρίνετε το *πρωτόκολλο αξιολόγησης* συστημάτων συστάσεων που *εξαρτάται απο τη διάσταση του χρόνου* μ' αυτό που είναι *χρονικά ανεξάρτητο*. Ποιο από τα δύο αποτυπώνει καλύτερα την πραγματικότητα;

4. Ποιες οι προκλήσεις που αντιμετωπίζει το *πρωτόκολλο αξιολόγησης συστημάτων συστάσεων βάσει χρόνου*;

5. *Η γεωγραφική θέση του χρήστη* είναι σημαντική για τη βελτίωση της *αποτελεσματικότητας* των συστημάτων συστάσεων. Να τεκμηριώσετε την παραπάνω επισήμανση.

6. Να αναφέρετε τους τύπους συστάσεων στα κοινωνικά δίκτυα *που βασίζονται στη γεωγραφική θέση του χρήστη*. Ποια η διαφορά μεταξύ της *σύστασης σημείου ενδιαφέροντος* από τη *σύσταση δρομολογίου* ή *διαδρομής*;

7. Ποια η διαφορά μεταξύ των *hybrid* και των *ensemble* συστημάτων συστάσεων;

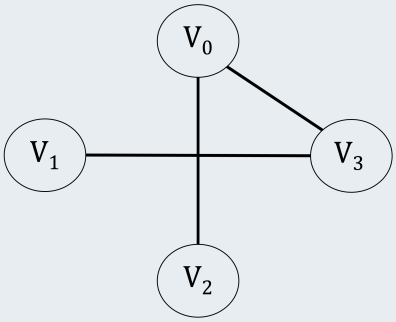
8. Σε τί διαφέρουν τα *συστήματα συστάσεων παράλληλης διάταξης* από αυτά της *σειριακής*; Ποιο εκ των δύο υπερέχει και γιατί;

9. Σε τί διαφέρουν τα *μονολιθικά συστήματα συστάσεων* από αυτά της *σειριακής διάταξης*; Ποιο εκ των δύο υπερέχει και γιατί;

4.8 Ασκήσεις Προγραμματισμού

Άσκηση 1. Δημιουργία ενός συστήματος συστάσεων υποψηφίων φίλων βάσει δεδομένων δομής ενός γράφου. Ο αντικειμενικός στόχος είναι να προταθούν νέοι φίλοι στον υπό εξέταση θεατή (viewer) βάσει των δεσμών φιλίας που αυτός έχει με άλλους θεατές (viewers).

Μας δίνεται η παρακάτω λίστα με το id των τεσσάρων θεατών (*viewers*) του Σχήματος 4.8:



Σχήμα 4.8: Παράδειγμα ενός γράφου όπου οι κόμβοι είναι οι *θεατές*, ενώ οι ακμές είναι η "φιλική" σχέση μεταξύ τους.

```
viewers = [
    { "viewer_id": 0 },
    { "viewer_id": 1 },
    { "viewer_id": 2 },
    { "viewer_id": 3 }]
```

Επίσης, μας δίνεται μια λίστα ζευγαριών που αντιπροσωπεύουν τις "φιλίες" μεταξύ των τεσσάρων *θεατών* του Σχήματος 4.8:

```
links_between_viewers = [(0, 2), (0, 3), (1,2)]
```

Στη συνέχεια μπορούμε να δημιουργήσουμε ένα *λεξικό* όπου τα *κλειδιά* του θα είναι τα ids των *θεατών* και οι τιμές του θα είναι λίστες που θα περιλαμβάνουν τα ids όλων των αντίστοιχων φίλων του κάθε *υπό εξέταση θεατή* με τη βοήθεια της *βιβλιοθήκης defaultdict*.

```
from collections import defaultdict
viewer_and_a_list_of_his_friends = defaultdict(list)
for i, j in links_between_viewers:
    viewer_and_a_list_of_his_friends[i].append(j)
    viewer_and_a_list_of_his_friends[j].append(i)
print(viewer_and_a_list_of_his_friends)
```

Ο προηγούμενος κώδικας δημιουργεί το ακόλουθο *λεξικό*:

defaultdict(<class 'list'>, 0: [2, 3], 2: [0, 1], 3: [0], 1: [2])

Για παράδειγμα, ο *θεατής* με id 0 είναι φίλος με τον *θεατή* με id 2 και τον *θεατή* με id 3.

Έπειτα θα δημιουργήσουμε μια συνάρτηση που θα βρίσκει τους φίλους των φίλων του *υπό εξέταση θεατή* με τη χρήση του παρακάτω κώδικα:

```
def friends_of_friends(viewer):
    viewer_id = viewer["viewer_id"]
    dictionary = {}
    for ids in viewers:
        counter = 0
        for my_viewer_id in viewer_and_a_list_of_his_friends[viewer_id]:
            for my_viewer_id in viewer_and_a_list_of_his_friends[my_viewer_id]:
                if ids['id'] == my_viewer_id:
                    if my_viewer_id != viewer["viewer_id"]:
                        counter +=1
                        dictionary[my_viewer_id] = counter
    return dict(sorted(dictionary.items(), key=lambda x: x[1], reverse=True))

friends_of_friends(viewers[3])
```

Συνεπώς, ο φίλος των φίλων που προτείνεται από τον παραπάνω κώδικα για τον *θεατή* με id 3 είναι ο *θεατής* με id 2 όπως φαίνεται στη συνέχεια:

{2: 1}.

Αυτό σημαίνει ότι ο *θεατής* με id 2 είναι εκείνος που θα προταθεί στον *υπό εξέταση θεατή* με id 3. Ακόμα, η μονάδα μετά την άνω και κάτω τελεία σημαίνει ότι υπάρχει ένας κοινός φίλος (*θεατής* με id 0) που συνδέει τους *θεατές* με id 3 και id 2.

Άσκηση 2: Σύσταση υποψηφίων φίλων βάσει ενός υβριδικού αλγορίθμου συστάσεων που συνδυάζει το σύστημα συστάσεων βάσει περιεχομένου που αναπτύχθηκε στην Ενότητα 3.10 (βλ. Ασκήσεις Προγραμματισμού) του αντίστοιχου κεφαλαίου με το σύστημα συστάσεων που αναπτύχθηκε στην 1η προγραμματιστική άσκηση του παρόντος κεφαλαίου. Σκοπός αυτού του υβριδικού συστήματος είναι να επανακατατάξει τους προτεινόμενους φίλους ανάλογα με: (α) Πόσα κοινά είδη ταινιών έχει ένας θεατής με τους άλλους θεατές (1ο Σύστημα Συστάσεων) και (β) Πόσους κοινούς φίλους έχει ένας θεατής με τους άλλους θεατές φίλους του (2ο Σύστημα Συστάσεων). Επιπλέον, κάνουμε την υπόθεση ότι τα κοινά ενδιαφέροντα ως προς τα είδη ταινιών πρέπει να έχουν παρόμοια βαρύτητα με τους κοινούς φίλους. Επομένως δίνουμε συντελεστή βαρύτητας 0.5 στον αριθμό των κοινών φίλων και συντελεστή βαρύτητας 0.5 στον αριθμό των κοινών ειδών ταινιών.

Η τελική κατάταξη των *υποψηφίων προς σύσταση φίλων* βάσει του *υβριδικού αλγορίθμου συστάσεων* υλοποιείται με τον παρακάτω κώδικα:

```
1  def closest_friends_ranking(viewer):
2      interesting_friendrec=most_common_genres(viewer)
3      closest_friendrec=friends_of_friends(viewer)
4      friendrec_dict={}
5      common_friend_coefficient = 0.5
6
7     #Κατάταξη των υποψηφίων φίλων βάσει των κοινών φίλων που έχουν με τον
       υπό εξέταση θεατή
8      for closest in closest_friendrec:
9          for interesting in interesting_friendrec:
10             if closest==interesting:
11                 friendrec_dict[closest]=
       common_friend_coefficient*(closest_friendrec[closest]+
       interesting_friendrec[interesting])
12         if closest not in friendrec_dict:
13             friendrec_dict[closest]=common_friend_coefficient*(
       closest_friendrec[closest])
14
```

```
    #Κατάταξη των υποψηφίων φίλων βάσει των κοινών ενδιαφερόντων τους ως προς τα είδη ταινιών ( ρομαντική, πολεμική, κλπ. ) που έχουν με τον υπό εξέταση θεατή
    for interesting in interesting_friendrec:
        if interesting not in friendrec_dict:
          friendrec_dict[interesting]=(1-common_friend_coefficient)*interesting_friendrec[interesting]

     friendrec_dict_sort= dict(sorted( friendrec_dict.items(), key=lambda item:item[1], reverse=True))
     print(friendrec_dict_sort)

#Δοκιμαστική εκτύπωση για τον θεατή με id 3:
print(most_common_genres(viewers[3]))
print(friends_of_friends(viewers[3]))

closest_friends_ranking(viewers[3])
```

Τα αποτελέσματα του προηγούμενου κώδικα είναι οι τρεις ακόλουθες προτεινόμενες λίστες φίλων βάσει των τριών αλγορίθμων (*φιλτραρίσματος βάσει περιεχομένου*, *βάσει συνδέσμων φιλίας μεταξύ των θεατών*, και του *υβριδικού*):
{2: 3, 0: 2, 1: 1}

{2: 1}

{2: 2.0, 0: 1.0, 1: 0.5}.

Επομένως, βάσει του *υβριδικού αλγορίθμου* (βλ. την τελευταία προτεινόμενη λίστα φίλων) για τον *υπό εξέταση θεατή* με id 3 προτείνεται ως κορυφαίος φίλος ο *θεατής* id με 2, επειδή συγκεντρώνει τη μεγαλύτερη προβλεπόμενη βαθμολογία, 2 βαθμούς. Εν συνέχεια, ως δεύτερος κορυφαίος φίλος προτείνεται ο *θεατής* με id 0 που έχει σκορ ίσο με 1. Τέλος, ο τρίτος στη λίστα των προτεινόμενων φίλων είναι ο *θεατής* με id 1 συγκεντρώνοντας 0.5 βαθμούς.

Βιβλιογραφία

V. Bellotti, B. Begole, E. Chi, N. Ducheneaut, J. Fang, E. Isaacs, T. King, M. Newman, K. Partridge, B. Price, P. Rasmussen, M. Roberts, D. Schiano, and A. Walendowski. Activity-based serendipitous recommendations with the magitti mobile leisure guide. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1157--1166, Florence, Italy, 2008.

E. V. Epure, B. Kille, J. E. Ingvaldsen, R. Deneckère, C. Salinesi, and S. Albayrak. Recommending personalized news in short user sessions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys'17, Como, Italy*, pages 121--129. ACM, 2017. 10.1145/3109859.3109894. URL `http://doi.acm.org/10.1145/3109859.3109894`.

D. Jannach, L. Lerche, and M. Jugovac. Adaptation and evaluation of recommendations for short-term shopping goals. In *Proceedings of the ninth ACM Conference on Recommender Systems*, RecSys '15, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4652-8.

M. Kayaalp, T. Ozyer, and S. Ozyer. A collaborative and content based event recommendation system integrated with data collection scrapers and services at a social networking site. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pages 113--118, Athens, Greece, 2009.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems, 2009.

R. Lee, S. Wakamiya, and K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321--349, 2011.

J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *2012 IEEE 28th international conference on data engineering*, pages 450--461. IEEE, 2012.

J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces, IUI'10, Hong Kong, China*, pages 31--40. ACM, 2010. 10.1145/1719970.1719976. URL `http://doi.acm.org/10.1145/1719970.1719976`.

M. Ludewig and D. Jannach. Evaluation of session-based recommendation algorithms. In *Journal Computing Report Repository (2018)*, volume 28, pages 331--390, 2018.

A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. Friendlink: Link prediction in social networks via bounded local path traversal. In *Proceedings of the 3rd Conference on Computational Aspects of Social Networks (CASON)*, pages 66--71, Salamanca, Spain, 2011.

M. Quadrana, P. Cremonesi, and D. Jannach. Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, 51(4):1--36, 2018.

D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 971--976, Sydney, Australia, 2010.

D. Saez-Trumper, D. Quercia, and J. Crowcroft. Ads and the city: considering geographic distance goes a long way. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys)*, pages 187--194, Dublin, Ireland, 2012.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1046--1054, San Diego, CA, 2011.

G. Shani, D. Heckerman, and R. I. Brafman. An mdp-based recommender system. In *Journal of Machine Learning Research (2005)*, volume 6, pages 1265--1295. JMLR, 2005. URL `http://www.jmlr.org/papers/v6/shani05a.html`.

J. Vinagre, A. M. Jorge, and J. Gama. Evaluation of recommender systems in streaming environments. In *Workshop on Recommender Systems Evaluation: Dimensions and Design (REDD 2014), held in conjunction with RecSys 2014*, 2014.

M. Ye, P. Yin, W. Lee, and D. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th ACM SIGIR International Conference on Research and development in Information Retrieval (SIGIR)*, pages 325--334, Beijing, China, 2011.