ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEB INTELLIGENCE

Ανάλυση και πρόβλεψη της συμπεριφοράς πελατών διαδικτυακών ξενοδοχειακών υπηρεσιών με χρήση της γραφικής βάσης δεδομένων Neo4j

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΙΡΑΝΤΑΣ ΑΝΤΩΝΙΟΥ

Επιβλέπων : Κωνσταντίνος Διαμαντάρας
Καθηγητής, ΔΙ.ΠΑ.Ε

Θεσσαλονίκη, Φεβρουάριος 2022

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Ανάλυση και πρόβλεψη της συμπεριφοράς πελατών διαδικτυακών ξενοδοχειακών υπηρεσιών με χρήση της γραφικής βάσης δεδομένων Neo4j

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΙΡΑΝΤΑΣ ΑΝΤΩΝΙΟΥ

Επιβλέπων : Κωνσταντίνος Διαμαντάρας
Καθηγητής ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 5 Μαρτίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
……		……
Όνομα Επώνυμο	Όνομα Επώνυμο	'Ονομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Φεβρουάριος 2022

(Υπογραφή)

……

ΜΙΡΑΝΤΑ ΑΝΤΩΝΙΟΥ

Πληροφορικός ΠΑ.ΜΑΚ.

Περίληψη

Τα συστήματα συστάσεων είναι ένα βασικό εργαλείο για την πρόβλεψη της συμπεριφοράς των χρηστών και χρησιμοποιείται ευρέως για τη βελτίωση των παρεχόμενων υπηρεσιών. Η παρούσα διπλωματική εργασία επικεντρώνεται στη μελέτη, ανάπτυξη αλγορίθμων και την εφαρμογή τους στην πρόβλεψη της συμπεριφοράς του χρήστη κατά την επιλογή ξενοδοχείων στη διαδικτυακή πλατφόρμα trivago. Αρχικά, προσεγγίζεται βιβλιογραφικά το πεδίο της παραγωγής συστάσεων, της αρχιτεκτονικής τους και των σημαντικότερων ζητημάτων που αναδύονται.

Η συμπεριφορά των χρηστών που καταγράφηκε κατά τη διάδρασή τους στην πλατφόρμα, αποτυπώνεται σε ένα διάγραμμα στην πλατφόρμα Neo4j. Προκειμένου να εξαχθούν συμπεράσματα για τα ξενοδοχεία που είναι πιθανό να επιλέξει ο χρήστης, χρησιμοποιήθηκε ο αλγόριθμος node2vec, ενώ για να εντοπιστούν τα πιο «όμοια» ξενοδοχεία, χρησιμοποιήθηκε η ευκλείδεια απόσταση. Προκειμένου να αξιολογηθεί το σύστημα ως προς την αποτελεσματικότητά του χρησιμοποιήθηκε η μέση αντίστροφη κατάταξη (MRR). Τέλος, το πλαίσιο της διπλωματικής ολοκληρώνεται μέσω συγκριτικών πειραμάτων με πραγματικά δεδομένα.

Λέξεις Κλειδιά: <<συστήμα σύστασης, αλγόριθμοι βασισμένοι σε γράφο, γραφική βάση δεδομένων, πρόβλεψη συνδέσμων, top-N συστάσεις >>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Recommender systems are a key tool for predicting user behavior and are widely used to improve services. This dissertation focuses on the research, development of algorithms and their application in predicting user behavior when selecting hotels on the online platform trivago. First, a bibliographic approach of the field of production of recommendations, their architecture and the most important issues that arise is presented.

The behavior of users, that was recorded during their interaction on the platform, is reflected in a diagram on the Neo4j platform. In order to draw conclusions about the hotels that the user is likely to choose, the node2vec algorithm was used, while in order to identify the most "similar" hotels, the Euclidean distance was used. To evaluate the system in terms of its effectiveness, the mean reciprocal rank (MRR) was used. Finally, the dissertation framework is completed through comparative experiments with real data.

Keywords:<<recommender system, graph-based algorithms, graph database, link prediction, top-N recommendation>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Φτάνοντας στην ολοκλήρωση του μεταπτυχιακού «Ευφυείς Τεχνολογίες Διαδικτύου» θέλω να ευχαριστήσω την οικογένεια μου και τους δικούς μου ανθρώπους για τη στήριξη, την υπομονή και την ενθάρρυνση που μου προσέφεραν απλόχερα καθ' όλη τη διάρκεια της φοίτησης μου στο Διεθνές Πανεπιστήμιο.

Τις θερμές ευχαριστίες μου οφείλω στον Α' επιβλέποντα καθηγητή μου κ. Διαμαντάρα Κωνσταντίνο για την επιστημονική καθοδήγηση, τις εύστοχες παρατηρήσεις, την αμέριστη συμπαράσταση και την προθυμία του να με βοηθήσει καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Επίσης να ευχαριστήσω την κ. Ασδρέ Αικατερίνη για τις υποδείξεις της.

Κατάλογος Σχημάτων

Σχήμα 4.1: Διαχωρισμός σετ δεδομένων 37
Σχήμα 5.1: Γράφος ιδιοτήτων 41
Σχήμα 6.1: Αρχιτεκτονική των μεθόδων CBOW/ SkipGram 58
Σχήμα 6.2: Mean Reciprocal Rank 62

Κατάλογος Πινάκων

Πίνακας 2.1: Κατηγορίες υβριδικών συστημάτων 12
Πίνακας 5.1: Σύγκριση ΣΒΔ /Neo4j 46
Πίνακας 6.1: Παράμετροι συνάρτησης gds.beta.node2vec.write 59
Πίνακας 6.2: Αποτελέσματα εκτέλεσης πειραμάτων 63

Κατάλογος Γραφημάτων

Γράφημα 6.1: Εκτέλεση 1ου πειράματος 64
Γράφημα 6.2: Εκτέλεση 2ου πειράματος 65
Γράφημα 6.3: Εκτέλεση 3ου πειράματος 65
Γράφημα 6.4: Εκτέλεση 4ου πειράματος 66

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
ΑΝΔ	Αναδρομικά Νευρωνικά Δίκτυα
MF	Matrix Factorization
BPR	Bayesian Personalized Ranking
OTB	Outside-The-Box
GCN	Graph Convolutional Network
ΣΒΔ	Σχεσιακές Βάσεις Δεδομένων
MRR	Mean Reciprocal Rank
ΑΕΠ	Ακαθάριστο Εγχώριο Προϊόν

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Περιγραφή του προβλήματος 1

1.2 Στόχοι της διπλωματικής εργασίας 2

1.3 Διάρθρωση 3

2 Συστήματα Συστάσεων 4

2.1 Εισαγωγή 4

2.2 Γενιές ενός συστήματος σύστασης 5

2.2.1 Συστήματα πρώτης γενιάς 5

2.2.2 Συστήματα δεύτερης γενιάς 6

2.2.3 Συστήματα τρίτης γενιάς 6

2.3 Τύποι Συστημάτων Σύστασης 7

2.3.1 Συνεργατικό Φιλτράρισμα 7

2.3.2 Φιλτράρισμα με βάση το περιεχόμενο 8

2.3.3 Δημογραφικό Φιλτράρισμα 10

2.3.4 Υβριδικά Συστήματα Συστάσεων 11

2.3.5 Πολυκριτηριακά Συστήματα Συστάσεων 13

2.3.6 Φιλτράρισμα με βάση τη γνώση 14

2.3.7 Συστήματα βασισμένα σε περιόδους σύνδεσης 15

2.4 Περιορισμοί στα συστήματα συστάσεων 18

2.4.1 Το πρόβλημα της «ψυχρής εκκίνησης» 18

2.4.2 Το πρόβλημα της πλήρους εξειδίκευσης 20

2.4.3 Το πρόβλημα της συνωνυμίας 21

2.4.4 Το πρόβλημα της ένεσης προφίλ 21

2.4.5 Το πρόβλημα των απρόβλεπτων αντικειμένων 22

2.4.6 Το πρόβλημα κινδύνου 22

2.5 Ιδιωτικότητα ενός συστήματος σύστασης 23

2.6 Νευρωνικά Δίκτυα 24

3 Συστήματα συστάσεων ξενοδοχειακών επιχειρήσεων 26

3.1 Συστάσεις σε διαδικτυακές υπηρεσίες για εύρεση καταλύματος 26
3.2 Τεχνολογίες που εμφανίζονται στα υφιστάμενα συστήματα σύστασης 28
3.3 Αρχιτεκτονική 28
3.3.1 Αρχιτεκτονική βασισμένη σε τεχνικές συνεργατικού φιλτραρίσματος 29
3.4 Συνήθη προβλήματα που εμφανίζονται στα τουριστικά συστήματα σύστασης 30
3.5 Προκλήσεις 31
4 Ορισμός προβλήματος 33
4.1 Εισαγωγή 33
4.2 Συνέδριο RecSys 34
4.2.1 RecSys Challenge 2019 35
4.3 Επιλογή δεδομένων 35
4.4 Περιορισμοί 36
4.5 Περιγραφή του προβλήματος 36
4.6 Μεταβλητές Trivago 37
5 Τεχνολογικό πλαίσιο 40
5.1 Γράφος ιδιοτήτων (Property Graph) 41
5.2 Neo4j 42
5.3 Cypher 47
5.3.1 Η Cypher ως παράδειγμα εντολών 48
5.4 R 52
5.5 VBA 52
5.6 Microsoft Excel 53
6 Μεθοδολογία και πειραματικά αποτελέσματα 54
6.1 Οντότητες και σχέσεις 54
6.2 Περιορισμοί 56
6.3 Φιλτράρισμα δεδομένων 56
6.4 Πειράματα 56
6.5 Αλγόριθμοι εύρεσης πλησιέστερων ξενοδοχείων 57
6.5.1 Word2vec 57
6.5.2 Node2vec 58
6.5.3 Ευκλείδεια απόσταση 61

6.6 Πρωτόκολλο αξιολόγησης.....62

7 Επίλογος.....67

7.1 Σύνοψη και συμπεράσματα.....67

7.2 Μελλοντικές επεκτάσεις.....69

8 Βιβλιογραφία.....72

1

Εισαγωγή

1.1 Περιγραφή του προβλήματος

Ο ανθρώπινος νους, σύμφωνα με τη φιλοσοφία οφείλει να αναπροσαρμόζεται συνεχώς στις αλλαγές του περιβάλλοντος βρίσκοντας συνεχώς τη βέλτιστη λύση. Για την εύρεση της βέλτιστης λύσης στα προβλήματα απαιτείται κριτική σκέψη και πειθαρχία. Εκτός από τα δύο αυτά ανθρώπινα χαρακτηριστικά, κυρίαρχο ρόλο διαδραματίζουν και τα αναπτυγμένα συστήματα διαχείρισης γνώσης, τα οποία βρίσκουν λύση σε ένα πρόβλημα, με βάση τη διαθέσιμη πληροφορία.

Η εξέλιξη της επιστήμης της πληροφορικής σε συνδυασμό με την τεχνολογική πρόοδο επέφερε αλλαγές στην καθημερινή ζωή των ανθρώπων, βελτιώνοντας παράλληλα το βιοτικό τους επίπεδο. Σταθμός στην εξέλιξη της τεχνολογίας ήταν η υπηρεσία του παγκόσμιου ιστού, με αποτέλεσμα η διαθέσιμη πληροφορία να πολλαπλασιαστεί. Η συνεχής αύξηση της διαθέσιμης πληροφορίας, καθιστά ακόμη πιο δύσκολη την ανεύρεση αποτελεσματικών αντικειμένων αναζήτησης, τη στιγμή που ο χρήστης το επιθυμεί. Το να διαχειριστεί αποτελεσματικά η αυξανόμενη γνώση, συνεπώς, αποτέλεσε μια τεράστια πρόκληση για τους ερευνητές οι οποίοι προσπάθησαν μέσω ανάπτυξης κατάλληλων αλγορίθμων σύστασης να «προσδιορίσουν» την καταλληλότερη πληροφορία την στιγμή που ο χρήστης αναζητά έναν όρο.

Η ιδέα της ανάπτυξης συστημάτων σύστασης προήλθε από την καθημερινή μας ζωή. Όταν έχουμε να επιλέξουμε την αγορά ενός προϊόντος, για παράδειγμα, είναι αρκετά σύνηθες για

εμάς να ρωτήσουμε έναν φίλο, ένα γνωστό ή ένα συγγενή μας για την επιλογή του. Ένα σύστημα σύστασης κατά βάση λειτουργεί με παρόμοιο τρόπο: συγκεντρώνει τα χαρακτηριστικά ενός χρήστη και καταγράφει την πλοήγησή του σε μια ιστοσελίδα, φτάνοντας μέχρι την τελική επιλογή προϊόντος ή υπηρεσίας. Η εξατομικευμένη καταγραφή πλοήγησης για κάθε χρήστη, οδηγεί σε εξατομικευμένη προβολή προτεινόμενων προϊόντων στους χρήστες. Αυτό σημαίνει, ότι τα προτεινόμενα προϊόντα βασίζονται στις καταχωρημένες πληροφορίες είτε χαρακτηριστικών είτε πλοήγησης του χρήστη. Σε αυτή τη διαδικασία, ιδιαίτερα βοηθητικές στην καταγραφή και διαρκή ενημέρωση αυτών των προτιμήσεων είναι και οι βάσεις δεδομένων. Οι βάσεις δεδομένων αποτυπώνουν διαρκώς τη κίνηση του χρήστη μέσα στην ιστοσελίδα και ενημερώνονται με τις νέες κινήσεις του. Προσφέρουν μια αποθήκη βάση της οποίας μπορεί να υπολογίζεται συνεχώς ο βαθμός ομοιότητας των χρηστών, βοηθώντας στην παροχή αποτελεσματικότερων συστάσεων προϊόντων. Σημαντικά πλεονεκτήματα, ειδικότερα, παρουσιάζονται στις βάσεις δεδομένων γράφων καθώς με τη βοήθειά τους οι οντότητες απεικονίζονται ως κόμβους και αντικατοπτρίζει απευθείας τις σχέσεις με την πραγματική τους μορφή.

Στην παρούσα διπλωματική εργασία σχεδιάζεται ένα σύστημα το οποίο λαμβάνει υπόψη κατά βάση την πλοήγηση του χρήστη μέσα σε μια ιστοσελίδα τουριστικών προϊόντων. Υπολογίζεται ένας δείκτης ομοιότητας μεταξύ των συνεδριών της υπηρεσίας αυτής και προτείνονται στους χρήστες πιθανές επιλογές βάσει αυτού του δείκτη ομοιότητας. Οι πιθανές επιλογές του χρήστη, επιστρέφονται ως μια λίστα ενώ στη συνέχεια με βάση την τελική του επιλογή υπολογίζεται το πόσο σωστά αυτή προβλέπεται.

1.2 Στόχοι της διπλωματικής εργασίας

Η ανάγκη ανάπτυξης κατάλληλων προτάσεων για τους χρήστες, αποτελεί μια πρόκληση σε αρκετούς τομείς της καθημερινότητας και σκοπός πολλών ερευνητών αλλά και εταιρειών, με σκοπό τη βελτίωση των παρεχόμενων υπηρεσιών αλλά και την αύξηση του κέρδους. Ο κύριος στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός συστήματος σύστασης για χρήστες που επισκέπτονται τη διαδικτυακή σελίδα της trivago, προκειμένου να βρουν ένα κατάλυμα το οποίο ανταποκρίνεται στις ανάγκες τους. Οι συστάσεις αυτές πρέπει να είναι σχετικές με την πρώτη αναζήτηση του χρήστη, προκειμένου η επιλογή του κατάλληλου ξενοδοχείου να είναι μια εύκολη, γρήγορη και ευχάριστη διαδικασία. Επειδή κατά την πλοήγηση σε αυτήν την ιστοσελίδα, δε σκιαγραφείται το προφίλ του χρήστη, το πρώτο κλικ σε ένα ξενοδοχείο θεωρείται καθοριστικό.

Για την επίτευξη αυτού του στόχου, αρχικά μελετήθηκαν τα συστήματα σύστασης που υπάρχουν στη βιβλιογραφία και στη συνέχεια επιλέχθηκαν τα στοιχεία εκείνα των συστημάτων

που ταίριαζαν καλύτερα στις ανάγκες του προβλήματος που πρέπει να αντιμετωπιστούν. Προκειμένου να επιστραφούν τα πιο όμοια ξενοδοχεία χρησιμοποιείται η Ευκλείδεια απόσταση, ενώ για την εξαγωγή των συμπερασμάτων και για την αξιολόγηση τους χρησιμοποιήθηκε σαν μετρική η μέση αμοιβαία κατάταξη, ενώ για την παραγωγή συστάσεων χρησιμοποιήθηκε ο αλγόριθμος Node2vec.

1.3 Διάρθρωση

Στο πρώτο κεφάλαιο, επεξηγούνται τα συστήματα συστάσεων ως απαραίτητο μέρος της καθημερινότητάς μας, ενώ τονίζεται και η αναγκαιότητά τους.

Στο δεύτερο κεφάλαιο, γίνεται βιβλιογραφική ανασκόπηση στα πρώιμα και τα νεότερα συστήματα συστάσεων ανά γενιές και τύπους. Αναλύονται συστήματα σύστασης ευρέως διαδεδομένα, η αρχιτεκτονική, τα πλεονεκτήματα καθώς και τα μειονεκτήματα τους. Επιπλέον, κατονομάζονται οι περιορισμοί, καθώς και ένα σημαντικό θέμα το οποίο αναδύθηκε κατά τα τελευταία χρόνια, η ιδιωτικότητα των συστημάτων σύστασης.

Στο τρίτο κεφάλαιο, διατυπώνεται το πρόβλημα της εύρεσης κατάλληλων καταλυμάτων για τους χρήστες της διαδικτυακής υπηρεσίας trivago. Αναλύεται η μεθοδολογία, η αρχιτεκτονική καθώς και το τεχνολογικό πλαίσιο ανάπτυξης του συστήματος. Αναφέρονται συνήθη προβλήματα που εμφανίζονται στα τουριστικά συστήματα σύστασης, οι προκλήσεις ενώ γίνεται και εκτενής αναφορά στη διαδικασία φιλτραρίσματος δεδομένων που ακολουθήθηκε.

Στο τέταρτο κεφάλαιο, παρουσιάζονται τα συμπεράσματα οι προκλήσεις καθώς και οι πιθανές βελτιώσεις που μπορεί να έχει το σύστημα σύστασης.

2

Συστήματα Συστάσεων

2.1 Εισαγωγή

Ο τεράστιος και αυξανόμενος όγκος πληροφοριών που διατίθεται σήμερα θέτει πολλές προκλήσεις στα συστήματα πληροφοριών. Ως σύστημα συστάσεων ή προτάσεων, μπορούν να χαρακτηριστούν διάφορα εργαλεία λογισμικού ή τεχνικές τα οποία παρέχουν προτάσεις σε έναν χρήστη [1] . Οι προτάσεις αυτές σχετίζονται με τη συμπεριφορά του χρήστη κατά τη διάδρασή του σε μια πλατφόρμα όπως για παράδειγμα τι μουσική ακούει, τι εφημερίδες διαβάζει ή τι αντικείμενα προτιμά να αγοράζει.

Τα συστήματα προτάσεων χρησιμοποιούνται σε μια σωρεία πλατφορμών διαφορετικής εμβέλειας αλλά κοινής στοχοθεσίας, δηλαδή τη σωστή παροχή προτάσεων στον χρήστη. Χαρακτηριστικά παραδείγματα είναι μερικές ευρέως γνωστές πλατφόρμες όπως Youtube [2], Netflix [3], Spotify [4] (γίνεται παροχή οπτικοακουστικών προτάσεων βάσει της συμπεριφοράς του χρήστη), Amazon [5] (παροχή προτάσεων σε προϊόντα) και φυσικά σε πλατφόρμες κοινωνικής δικτύωσης όπως Twitter [4], Instagram [6] (παροχή προτάσεων φιλίας, ή προβολή διαφημίσεων προϊόντων).

Σε μια πιο ελεύθερη ερμηνεία του όρου, θα χαρακτηρίζαμε τα συστήματα προτάσεων ως αλγόριθμους οι οποίοι προτείνουν στον χρήστη αντικείμενα παρόμοια με αυτά που έχει αλληλεπιδράσει στο παρελθόν. Είναι εύλογο, πως αυτοί οι αλγόριθμοι παράγουν τεράστιο κέρδος στις εταιρείες που χρησιμοποιούνται, με χαρακτηριστικό παράδειγμα έναν διαγωνισμό που διεξήγαγε η εταιρεία Netflix με χρηματικό έπαθλο ενός δισεκατομμυρίου ευρώ [7]. Ο σκοπός του διαγωνισμού ήταν να παραχθεί αλγόριθμος που παρήγαγε κατά το δυνατό πιο αποτελεσματικές συστάσεις από το Cinematech, που χρησιμοποιούσε τότε η Netflix [8] .

Η απαρχή των συστημάτων σύστασης, συναντάται στα αρχαιότερα χρόνια. Μεταξύ της περιόδου 4000 έως 1200 π.Χ. οι αρχαίοι πολιτισμοί άνθισαν, μαζί με τις πρώτες συστάσεις,

οι οποίες είχαν τη μορφή ερωτημάτων. Η εμφάνιση των πρώτων συστάσεων εκείνα τα χρόνια, ενδεχομένως να απαντούσαν σε απλά ερωτήματα όπως ποια καλλιέργεια να καλλιεργηθεί και τι ώρα, τι θρησκεία να ακολουθηθεί από μια κοινότητα κ.λπ. Κατά την περίοδο του αποικισμού (μεταξύ του 11ου και του 18ου αιώνα), οι συστάσεις άρχισαν να εμπλουτίζονται με μεγαλύτερη ποικιλία ερωτημάτων όπως η γονιμότητα της γης (σε ποια περιοχή συμφέρει να καλλιεργηθούν ποιοι καρποί ανάλογα με τη σύστασή της), το ανθρώπινο δυναμικό (ποιοι θα επιλεγούν ως δούλοι) [9]. Τα συστήματα σύστασης με μορφή εγγύτερη στη σημερινή εντοπίζεται το 1992 με το σύστημα Tapestry, όπου ακούστηκε και πρώτη φορά ο όρος «συνεργατικό φιλτράρισμα», ο ακρογωνιαίος λίθος των συστημάτων σύστασης [11]. Ο λόγος ανάπτυξης του Tapestry ήταν η ανάγκη διαχείρισης ενός μεγάλου όγκου δεδομένων ηλεκτρονικής αλληλογραφίας. Τα δεδομένα αυτά μπορούσαν να κατηγοριοποιηθούν με βάσει το περιεχόμενό τους, ώστε ο παραλήπτης του να έφτανε γρήγορα σε αυτά αν επιθυμούσε να τα ανακαλέσει κάποιο από τα μηνύματά του [12].

2.2 Γενιές ενός συστήματος σύστασης

Τα συστήματα συστάσεων έχουν αναπτυχθεί ως ένα εργαλείο που διαχειρίζεται έναν τεράστιο όγκο πληροφοριών. Χρησιμοποιήθηκαν αρχικά στο ηλεκτρονικό εμπόριο για τη βελτίωση των πωλήσεων ενώ στη συνέχεια, αξιοποιήθηκαν και πληροφορίες των κοινωνικών δικτύων προκειμένου να υπερνικηθούν τα προβλήματα που εμφανιζόταν και να βελτιωθεί η ποιότητα των παρεχόμενων συστάσεων. Τα πιο σύγχρονα συστήματα συστάσεων βασίζονται σε ετικέτες RFID. Είναι δυνατή η κατηγοριοποίησή τους, ανάλογα με το επίπεδο εξέλιξης στο οποίο βρίσκονται σε συστήματα πρώτης γενιάς, συστήματα δεύτερης γενιάς και συστήματα τρίτης γενιάς. Η αξιολόγησή τους καθώς και οι μετρήσεις που χρησιμοποιούνται σε αυτές τις γενιές διαφοροποιούνται επίσης [12].

2.2.1 Συστήματα πρώτης γενιάς

Τα συστήματα προτάσεων πρώτης γενιάς χρησιμοποιούν κυρίως τις ακόλουθες τεχνικές: φιλτράρισμα με βάση το περιεχόμενο, συνεργατικό φιλτράρισμα και υβριδικές τεχνικές (συνδυασμός των δύο προαναφερθέντων). Πιο συγκεκριμένα, οι μετρικές ομοιότητας που κατεύθυναν τις συγκεκριμένες τεχνικές είναι η ομοιότητα συνημιτόνου, η Ευκλείδεια απόσταση και ο συντελεστής συσχέτισης του Pearson[1]. Αξίζει να αναφερθεί πως ο συντελεστής συσχέτισης του Pearson παρουσιάζει ένα μειονέκτημα στη μέτρηση ομοιότητας, ειδικά όταν ο χρήστης έχει αξιολογήσει ένα μικρό αριθμό αντικειμένων, κάνοντας πιο δύσκολη την παραγωγή σωστών αποτελεσμάτων[13].

2.2.2 Συστήματα δεύτερης γενιάς

Η δεύτερη γενιά των συστημάτων σύστασης, βασίστηκε κυρίως στη γνώση των ενδιαφερόντων των χρηστών, βάσει της οποίας συστάθηκε ένα αρχικό προφίλ χρήστη. Όλες οι συστάσεις αργότερα, βασίστηκαν σε αυτό το προσωπικό «χαρτοφυλάκιο» του χρήστη. Το χαρτοφυλάκιο του χρήστη εμπλουτίζεται και από άλλες πηγές, με την κυριότερη να αποτελεί τα κοινωνικά δίκτυα. Ο εμπλουτισμός και η χρήση πληροφοριών από διαφορετικές πηγές χρησιμοποιούνται για τη βελτίωση των αποτελεσμάτων πρόβλεψης και την υπέρβαση των περιορισμών των συστημάτων συστάσεων πρώτης γενιάς.

2.2.3 Συστήματα τρίτης γενιάς

Η τρίτη γενιά των συστημάτων σύστασης, εμφανίστηκε με την ανάπτυξη των κινητών συσκευών. Ένα τέτοιο σύστημα στηρίζεται κυρίως στο διαδίκτυο των πραγμάτων και στην τοποθεσία του χρήστη. Δύο παραδείγματα [13] τέτοιων συστημάτων σύστασης είναι η ανάπτυξη συνεργατικών αλγόριθμων φιλτραρίσματος που χρησιμοποιούνται από συσκευές διαδικτύου των πραγμάτων που αναπτύχθηκε από τους Organero et al [14] και ενός συστήματος συνιστάμενης θέσης που αναπτύχθηκε από τους Levandoski et al [15]. Το πρώτο, αναπτύχθηκε στο πανεπιστήμιο της Μαδρίτης όπου χρησιμοποιήθηκαν περισσότερες από 75 συσκευές με ετικέτες RFID, οι οποίες μπορούσαν να αναγνωσθούν από NFC. Οι ετικέτες παρείχαν συνδέσμους προς εξωτερικούς πόρους πληροφοριών στους οποίους θα μπορούσαν να έχουν πρόσβαση οι συμμετέχοντες χρησιμοποιώντας ένα κινητό τηλέφωνο Nokia 6131 NFC[14]. Το σύστημα σύστασης τοποθεσίας ή LARS όπως ονομάστηκε από τους δημιουργούς του χρησιμοποιούσε βαθμολογίες για να προτείνει τοποθεσίες. Οι προκάτοχοί του, δηλαδή τα παραδοσιακά συστήματα σύστασης δε λάμβαναν υπόψη τις χωρικές διαστάσεις των δεδομένων. Το LARS, από την άλλη πλευρά, υποστηρίζει μια τριπλέτα ταξινόμησης ανάλογα με το αν ένα αντικείμενο είναι χωρικό ή όχι δηλαδή: χωρικές αξιολογήσεις για μη χωρικά αντικείμενα, μη χωρικές αξιολογήσεις για χωρικά αντικείμενα και χωρικές αξιολογήσεις για χωρικά αντικείμενα [15]. Τα πειραματικά δεδομένα που παρουσιάζουν το συγκριτικό πλεονέκτημα της τεχνικής αυτής σε μια εφαρμογή μεγάλης κλίμακας, υπάρχουν στην εφαρμογή Foursquare.

2.3 Τύποι Συστημάτων Σύστασης

2.3.1 Συνεργατικό Φιλτράρισμα

Μια βασική τεχνική που υπάρχει στα συστήματα συστάσεων είναι αυτή του συνεργατικού φιλτραρίσματος. Υπάρχουν δύο βασικές μέθοδοι που χρησιμοποιούνται από τα συστήματα συστάσεων, οι οποίες βασίζονται στη μνήμη ή στο μοντέλο και αναλύονται παρακάτω.

- Αλγόριθμοι με βάση τη μνήμη

Οι αλγόριθμοι με βάση τη μνήμη αξιολογούν ολόκληρο τον πίνακα των αντικειμένων που προτιμά ο χρήστης κατατάσσοντάς τον σε μια «γειτονιά» ανάλογα με τις προτιμήσεις που είχε [16]. Τα αντικείμενα αυτά, μπορεί να ταξινομηθούν με ακόμη μεγαλύτερη ακρίβεια ακολουθώντας τις εξής προσεγγίσεις αλγορίθμων με βάση τη μνήμη: βασισμένα στο χρήστη ή βασισμένα σε αντικείμενα. Η πρώτη προσέγγιση, ακολουθείται σε περίπτωση που αξιοποιούνται μοτίβα στις σχέσεις χρήστη-χρήστη ενώ η δεύτερη σε περίπτωση που αξιοποιούνται μοτίβα μεταξύ αντικειμένων-αντικειμένων [17]. Τα συνεργατικά συστήματα σύστασης αν και παρέχουν πολύ χρήσιμες πληροφορίες, είναι ιδιαίτερα ευάλωτα σε επιθέσεις, διότι οι εισβολείς μπορούν να χρησιμοποιήσουν αυτοματοποιημένα μέσα για να εισάγουν μεγάλο αριθμό προφίλ με συγκεκριμένα μοτίβα προτιμήσεων, γεγονός που έχει ως αποτέλεσμα συστάσεις που ευνοούν ή όχι συγκεκριμένα αντικείμενα [18].

- Αλγόριθμοι με βάση το μοντέλο

Σε αντίθεση με τους αλγορίθμους με βάση τη μνήμη, οι αλγόριθμοι με βάση το μοντέλο ομαδοποιούν τους χρήστες από το σύνολο εκπαίδευσης σε μικρότερες ομάδες, με βάση τα μοτίβα των βαθμολογιών τους. Ουσιαστικά, οι τεχνικές μοντέλου, χρησιμοποιούν τα δεδομένα αξιολόγησης για να εκπαιδεύσουν το μοντέλο και στη συνέχεια αυτό το μοντέλο χρησιμοποιείται προκειμένου να αντληθούν οι συστάσεις [19].

- Σύγκριση προσέγγισης με βάση το μοντέλο και τη μνήμη

Όπως αναφέρθηκε, οι προσεγγίσεις με βάση τη μνήμη χρησιμοποιούν ολόκληρη τη βάση δεδομένων (που περιλαμβάνει αξιολογήσεις και προτιμήσεις των χρηστών), ενώ οι προσεγγίσεις βάσει μοντέλου συγκεντρώνουν ολόκληρη τη βάση δεδομένων σε μια μικρότερη δομή δεδομένων, που ονομάζεται μοντέλο. Στη βιβλιογραφία, αναφέρεται πως με βάση θεωρητικές μελέτες και πρακτικές μεθόδους πως οι αλγόριθμοι μνήμης (όπως π.χ. αλγόριθμοι πλησιέστερου γείτονα), έχουν εξαιρετική απόδοση, από την άποψη της ακρίβειας, για δεδομένα αξιολόγησης πολλαπλών τιμών ενώ, από την άλλη πλευρά, οι

αλγόριθμοι βάσει μοντέλου χειρίζονται αποτελεσματικά την επεκτασιμότητα σε μεγάλα σύνολα δεδομένων[20]. Ωστόσο, μια σημαντική δυσκολία στο σχεδιασμό συστημάτων συνεργατικού φιλτραρίσματος έγκειται στο πρόβλημα της τυποποίησης της ανθρώπινης αντίληψης και προτιμήσεων[21]. Για παράδειγμα, δε μπορεί να τυποποιηθεί με μια μετρική το γιατί ένας χρήστης προτιμά μια συγκεκριμένη μάρκα αυτοκινήτου, ή προτιμά να ακούει ένα συγκεκριμένο είδος μουσικής. Αντίστοιχα δε μπορούν να τυποποιηθούν και οι προσωπικές προτιμήσεις π.χ. στο προηγούμενο παράδειγμα μπορεί για έναν χρήστη η κλασσική μουσική να θεωρείται καλύτερης ποιότητας από την ποπ μουσική κοκ. Μια πολύ ενδιαφέρουσα πρόταση είναι ο συνδυασμός αυτών τον δύο τεχνικών, όπως πραγματοποίησαν οι SongJie Gong et al, σε ένα σύνολο δεδομένων της MovieLens για συστάσεις ταινιών.

2.3.2 Φιλτράρισμα με βάση το περιεχόμενο

Το φιλτράρισμα βάσει περιεχομένου χρησιμοποιεί μεταδεδομένα για να προτείνει άλλα στοιχεία παρόμοια με αυτά που αρέσουν στον χρήστη, βάσει των προηγούμενων ενεργειών ή των σχολίων του [22]. Βασίζεται, συνεπώς, στη διαθεσιμότητα των μεταδεδομένων των αντικειμένων (πληροφορίες που είτε έχουν εξαχθεί είτε έχουν δημιουργηθεί) και ενός προφίλ το οποίο δίνει σημαντικότητα σε αυτά τα χαρακτηριστικά [23]. Για παράδειγμα, στην περίπτωση ενός μουσικού κομματιού, ως μεταδεδομένα θεωρούνται ο καλλιτέχνης, το άλμπουμ και το είδος της μουσικής στο οποίο ανήκει. Ακριβώς όπως στην περίπτωση των μουσικών κομματιών, μπορούν να συγκεντρωθούν τέτοια δεδομένα για τη συμπεριφορά των χρηστών, είτε αναλύοντας την συμπεριφορά τους είτε ρωτώντας επακριβώς τους χρήστες για τις προτιμήσεις τους.

Σύμφωνα με τους Van Meteren & Van Someren [24] τα μεταδεδομένα χρησιμοποιούνται για να συσταθεί ένα μοντέλο χρήστη, το οποίο επιτρέπει να ταξινομηθούν αντικείμενα τα οποία δεν έχουν ταξινομηθεί σε μία θετική κλάση c (σχετική με τον χρήστη) ή αρνητική κλάση ‗c (άσχετη με τον χρήστη). Το σετ εκπαίδευσης αποτελείται από τα στοιχεία που βρήκε ο χρήστης ενδιαφέροντα. Αυτά τα στοιχεία σχηματίζουν ένα σετ με ένα κοινό χαρακτηριστικό, που καθορίζει την κλάση του αντικειμένου βάσει είτε της βαθμολογίας του χρήστη είτε των «σιωπηρών αποδείξεων». Επισήμως, ένα αντικείμενο περιγράφεται ως ένα διάνυσμα X=(x_1,x_2,…,x_n) με n συνιστώσες. Αυτό που πρέπει να κάνει η μέθοδος μάθησης είναι να επιλέξει μια συνάρτηση που βασίζεται σε ένα εκπαιδευτικό σύνολο διανυσμάτων εισόδου m που μπορούν να ταξινομήσουν οποιοδήποτε στοιχείο στη συλλογή. Η συνάρτηση H(x) θα μπορεί να ταξινομήσει ένα αταξινόμητο στοιχείο ως θετικό είτε αρνητικό στο ενώ μπορεί να χρησιμοποιηθεί ένα κατώφλι για να προσδιορίσει εάν το αντικείμενο είναι σχετικό ή άσχετο με τον χρήστη.

Δύο βασικά πλεονεκτήματα του συνεργατικού φιλτραρίσματος είναι τα ακόλουθα [10],[21]: Πρώτον δεν απαιτεί μεγάλο πλήθος δεδομένων για να εξάγει συμπεράσματα με ακρίβεια, ενώ δεύτερον, οι νέοι χρήστες που προστίθενται μπορούν να συσταθούν απευθείας με το που είναι διαθέσιμα τα μεταδεδομένα τους. Τα μεταδεδομένα μπορεί να προέλθουν αυτόματα από ηλεκτρονικούς καταλόγους, αλλά σε περίπτωση που είναι υποκειμενικά είναι καλύτερο να προέλθουν από χειροκίνητη εισαγωγή. Το μοντέλο που θα συσταθεί μπορεί να συλλάβει τα συγκεκριμένα ενδιαφέροντα ενός χρήστη και μπορεί να προτείνει εξειδικευμένα αντικείμενα που ενδιαφέρονται πολύ λίγοι άλλοι χρήστες. Με αυτόν τον τρόπο, διατηρείται η υποκειμενικότητα στις συστάσεις.

Στην περίπτωση που έχουμε υποκειμενικότητα στα μεταδεδομένα, δηλαδή ως ένα βαθμό χειροκίνητη εισαγωγή, η τεχνική αυτή απαιτεί πολλά από αυτά, γεγονός που καθιστά αυτό τον τρόπο σύστασης ως λιγότερο αυτοματοποιημένο. Επιπρόσθετα , το μοντέλο μπορεί να κάνει προτάσεις μόνο με βάση τα υπάρχοντα ενδιαφέροντα του χρήστη, με άλλα λόγια, έχει περιορισμένη δυνατότητα επέκτασης στα υπάρχοντα ενδιαφέροντα των χρηστών. Η έλλειψη αυτοματοποίησης και επεκτασιμότητας [22] αποτελούν δύο από τα βασικότερα μειονεκτήματα που παρουσιάζει ένα μοντέλο που έχει δημιουργηθεί με προσέγγιση συνεργατικού φιλτραρίσματος.

2.3.2.1.1 Αρχιτεκτονική

Η διαδικασία προτάσεων εκτελείται σε τρία βήματα, καθένα από τα οποία αντιμετωπίζεται από ένα ξεχωριστό δομικό στοιχείο της αρχιτεκτονικής [25]:

- Αναλυτής Περιεχομένου: Όταν οι πληροφορίες δεν έχουν δομή (π.χ. κείμενο), απαιτείται κάποιο είδος προεπεξεργασίας προκειμένου να εξαχθούν δομημένες πληροφορίες. Η κύρια ευθύνη του αναλυτή περιεχομένου είναι να αναπαριστά το περιεχόμενο των αντικειμένων (π.χ. έγγραφα, ιστοσελίδες κ.λπ.) που εισέρχονται, σε μορφή κατάλληλη για τα επόμενα στάδια επεξεργασίας. Τα μεταδεδομένα αναλύονται με τεχνικές εξαγωγής χαρακτηριστικών προκειμένου να μετατοπιστεί η αναπαράσταση στοιχείων από τον αρχικό χώρο πληροφοριών στον στόχο (π.χ. ιστοσελίδες που εκπροσωπούνται ως διανύσματα λέξεων-κλειδιών). Οι αναλύσεις που προκύπτουν, εισάγονται στα επόμενα δύο βήματα: την εκμάθηση προφίλ και το στοιχείο φιλτραρίσματος.

- Εκμάθηση προφίλ: Το στάδιο αυτό συλλέγει αντιπροσωπευτικά δεδομένα των προτιμήσεων του χρήστη και προσπαθεί να τα γενικεύσει, προκειμένου να συσταθεί το προφίλ του χρήστη. Συνήθως, η στρατηγική γενίκευσης υλοποιείται μέσω τεχνικών μηχανικής εκμάθησης οι οποίες είναι σε θέση να δημιουργήσουν ένα μοντέλο πρόβλεψης το οποίο στηρίζεται σε παλαιότερες προτιμήσεις του εκάστοτε χρήστη. Παραδείγματα

εκπαίδευσης μπορεί να είναι για παράδειγμα ένα βίντεο σε μια πλατφόρμα κοινωνικής δικτύωσης όπως το YouTube που έχει δοθεί θετικό ή αρνητικό σχόλιο από τον χρήστη.

- Στοιχείο Φιλτραρίσματος: Αυτό το βήμα εκμεταλλεύεται το προηγούμενο για να προτείνει σχετικά αντικείμενα, ταιριάζοντας το προφίλ του χρήστη με τα αντικείμενα που πρόκειται να προταθούν. Τόσο τα προφίλ χρηστών, όσο και, τα ίδια τα στοιχεία, που αντιπροσωπεύονται ως διανύσματα σταθμισμένου όρου (π.χ. με βάση το μοντέλο στάθμισης όρου TF.IDF). Οι προβλέψεις για το ενδιαφέρον των χρηστών για ένα συγκεκριμένο στοιχείο μπορούν να ληφθούν βάσει του υπολογισμού ομοιότητας διανυσμάτων (π.χ. χρησιμοποιώντας το μέτρο ομοιότητας συνημιτόνου) ή χρησιμοποιώντας πιθανοτικές προσεγγίσεις όπως η ταξινόμηση Bayesian [25].

2.3.3 Δημογραφικό Φιλτράρισμα

Το δημογραφικό φιλτράρισμα κατηγοριοποιεί τους χρήστες σε ομάδες με βάση τα δημογραφικά τους χαρακτηριστικά [26]. Τα δημογραφικά χαρακτηριστικά περιλαμβάνουν πληροφορίες σχετικά με το φύλο, την ηλικία, την εκπαίδευση καθώς και τα ενδιαφέροντα του κάθε χρήστη [27]. Προφανέστατα, οι ομάδες χρηστών συγκροτούνται από χρήστες με παρόμοια χαρακτηριστικά και συμπεριφορά. Σε αυτήν την περίπτωση, προκειμένου να γίνει η σύσταση, λαμβάνεται υπ' όψιν σε ποια ομάδα ανήκει ο χρήστης και στη συνέχεια εντοπίζονται οι προτιμήσεις των υπόλοιπων χρηστών της ίδιας ομάδας [26].

Τα συστήματα σύστασης δημογραφικού φιλτραρίσματος εμφανίζουν πολλές ομοιότητες με τα συστήματα συνεργατικού φιλτραρίσματος, καθώς αναμένεται ένα μεγάλο πλήθος χρηστών να μοιράζεται κοινά ενδιαφέροντα.

Υπάρχουν περιπτώσεις στη βιβλιογραφία που η έρευνα έδειξε ότι τα δημογραφικά στοιχεία και τα χαρακτηριστικά των χρηστών μπορεί να έχουν σημαντικό αντίκτυπο στις αναλογίες κλικ προς αριθμό εμφανίσεων (ερευνητική εργασία) στα συστήματα σύστασης [28]. Για παράδειγμα, στην προαναφερθείσα περίπτωση το φύλο είχε μόνο οριακό αντίκτυπο, όμως η ηλικία επηρέασε έντονα την αναλογία κλικ ως προς τον αριθμό εμφανίσεων [28]. Αυτό σημαίνει ότι τα δημογραφικά χαρακτηριστικά, διαδραματίζουν μεγάλο ρόλο στα συστήματα σύστασης, εντούτοις πρέπει να εξετάζονται ενδελεχώς ανάλογα με τη περίπτωση του συστήματος που θέλουμε να μελετηθεί. Η περίπτωση του METIS [29], ενός συστήματος ηλεκτρονικού εμπορίου που βασίζεται σε μια πλατφόρμα υπηρεσιών μικρο-ιστολογίων, είναι ένα ακόμη ένα επιτυχημένο παράδειγμα εφαρμογής ενός δημογραφικού συστήματος σύστασης. Στο METIS τα χαρακτηριστικά των χρηστών εξάγονται από το δημόσιο προφίλ τους και τα δημογραφικά στοιχεία των προϊόντων που μαθαίνουν τόσο από τις διαδικτυακές κριτικές προϊόντων όσο και από τα μικρο-ιστολόγια. Μία ακόμη περίπτωση ενός πετυχημένου

συστήματος δημογραφικού φιλτραρίσματος είναι ένα σύστημα που αναπτύχθηκε για την πλατφόρμα TripAdvisor [30]. Συγκεκριμένα, αυτό το σύστημα κατηγοριοποιεί τους τουρίστες χρησιμοποιώντας τις δημογραφικές τους πληροφορίες και μετά τους κάνει προτάσεις βάσει των δημογραφικών δεδομένων τους. Τα αποτελέσματα δείχνουν ότι οι μέθοδοι και δημογραφικές πληροφορίες μπορούν να χρησιμοποιηθούν για την πρόβλεψη βαθμολογιών σε αξιοθέατα.

Ένα βασικό πλεονέκτημα των δημογραφικών συστημάτων σύστασης είναι ότι όπως για παράδειγμα στην προηγούμενη περίπτωση του TripAdvisor, δεν απαιτείται ιστορικό αξιολογήσεων ή κάποια επιπλέον γνώση, οπότε ένας νεοεισερχόμενος τουρίστας στην πλατφόρμα, μπορεί να λάβει αποτελεσματική σύσταση. Αντίστοιχα στα μειονεκτήματα των δημογραφικών συστημάτων σύστασης συγκαταλέγεται η δυσκολία συγκέντρωσης δημογραφικών στοιχείων, καθώς τα δημογραφικά δεν είναι ένα από τα στοιχεία που συγκεντρώνονται εύκολα στους χρήστες [31]. Πολλοί από αυτούς, μπορεί να χρησιμοποιήσουν ψεύτικα στοιχεία εγγραφής σε μία πλατφόρμα, ή να μην επιθυμούν να αποκαλύψουν τέτοια στοιχεία. Σε τέτοιου είδους συστήματα, επίσης, είναι πολύ πιθανό να μη συμπεριληφθεί μια πιθανή αλλαγή προτιμήσεων χρηστών που συμβαίνει με την πάροδο του χρόνου [32]. Εκτός από τις προτιμήσεις, είναι δύσκολο να εξασφαλιστεί και ακρίβεια των αποτελεσμάτων. Προκειμένου ωστόσο, να εξασφαλιστεί η μέγιστη δυνατή ακρίβεια, είναι ιδανικό τα δημογραφικά συστήματα σύστασης να συνδυάζονται με άλλες τεχνικές, δημιουργώντας υβριδικά συστήματα τα οποία θα αναλυθούν παρακάτω [33].

2.3.4 Υβριδικά Συστήματα Συστάσεων

Από γλωσσική άποψη, ο όρος υβριδικός προέρχεται από το λατινικό ουσιαστικό hybrida (μικτής προέλευσης) και δηλώνει ένα αντικείμενο που είναι φτιαγμένο συνδυάζοντας δύο διαφορετικά στοιχεία [34]. Ένας κοινός τόπος στα συστήματα συστάσεων είναι η ανάγκη συνδυασμού δύο ή ακόμη και περισσότερων τεχνικών ώστε να επιτευχθεί μέγιστη ακρίβεια [35]. Αυτή είναι η φιλοσοφία στην οποία κινούνται τα υβριδικά συστήματα συστάσεων, εφόσον όλες οι υπάρχουσες τεχνικές προτάσεων έχουν πλεονεκτήματα και αδυναμίες [36]. Το πιο λογικό, λοιπόν, είναι αυτές να συνδυαστούν και να χρησιμοποιηθούν με διαφορετικούς τρόπους, ώστε να αντιμετωπίσουν μια διαφορετική πρόκληση κάθε φορά. Είναι πρόδηλο, πως η τεχνική ή το σύστημα που θα ακολουθηθεί είναι εξατομικευμένη ανάλογα με το τι αντικείμενο πρόκειται να συστήσουμε και σε ποιους. Ωστόσο, οι υβριδικές μέθοδοι βρίσκονται ακόμη σε αρχικό στάδιο, οπότε υπάρχει περίπτωση μεγαλύτερης εξέλιξης προς τέτοια συστήματα συστάσεων με την πάροδο του χρόνου. Οι πιθανότητες ροπής προς υβριδικά συστήματα συστάσεων, είναι σαφώς πολύ μεγαλύτερες από το παρελθόν.

Οι σημαντικότερες κατηγορίες υβριδικών συστημάτων, σύμφωνα με τη βιβλιογραφία

παρουσιάζονται στον παρακάτω πίνακα και είναι οι εξής [37]:

Πίνακας 2.1: Κατηγορίες υβριδικών συστημάτων

Σταθμισμένα	Οι βαθμολογίες αρκετών τεχνικών προτάσεων συνδυάζονται μεταξύ τους ώστε να παραχθεί μία μόνο πρόταση.
Εναλλαγής	Το σύστημα αλλάζει τεχνικές συστάσεων ανάλογα με την τρέχουσα κατάσταση.
Μικτά	Προτάσεις από πολλά διαφορετικά συστήματα σύστασης παρουσιάζονται ταυτόχρονα
Συνδυασμός χαρακτηριστικών	Χαρακτηριστικά από διαφορετικές πηγές δεδομένων σύστασης συνδυάζονται μαζί σε ένα αλγόριθμο πρότασης
Συνιστάμενα	Ένας αλγόριθμος σύστασης τελειοποιεί τις συστάσεις που δίνονται από τους άλλους.
Αύξηση χαρακτηριστικών	Η έξοδος από μία τεχνική χρησιμοποιείται ως είσοδος σε ένα άλλο σύστημα σύστασης.
Μετά- επίπεδο	Το μοντέλο που έμαθε από ένα σύστημα σύστασης χρησιμοποιείται ως είσοδος σε ένα άλλο.

Μελετώντας τη βιβλιογραφία, υπάρχουν πολλά τέτοιου είδους συστήματα σύστασης που κατασκευάστηκαν ήδη από τις αρχές του 20ου αιώνα. Συγκεκριμένα, οι Tran και Cohen [38] το 2000 ανέπτυξαν ένα σύστημα ηλεκτρονικού εμπορίου το οποίο προτείνει αγορά συναφών προϊόντων στους πελάτες. Το υβριδικό σύστημα που αναπτύχθηκε χρησιμοποιούσε τεχνικές βασισμένες στη γνώση σε συνδυασμό με τεχνικές συνεργατικού φιλτραρίσματος, ενώ η διαδραστική διεπαφή που αναπτύχθηκε μπορούσε και να συντονίσει τις λειτουργίες των δύο υποσυστημάτων για να κάνει τις καλύτερες δυνατές συστάσεις στους χρήστες [37]. Ακόμη ένα σύστημα που έχει αναπτυχθεί και παράγει υβριδικές συστάσεις, συνδυάζει την αξιολόγηση των χρηστών, τα χαρακτηριστικά των αντικειμένων, καθώς και τις δημογραφικές τους πληροφορίες [39]. Οι τρεις αυτές τεχνικές υπερτερούν των κλασσικών αλγορίθμων σύστασης, που έχουν αναπτυχθεί ως τώρα έχει καλύτερες επιδόσεις σε επίπεδα διασποράς, μετρικές όπως ROC, μέσου απόλυτου σφάλματος και προβλημάτων όπως ψυχρής εκκίνησης. Ένα εξαιρετικό παράδειγμα συνδυασμού διαφορετικών παραλλαγών αλγορίθμου προτάσεων είναι ο διαγωνισμός του βραβείου Netflix, στον οποίο συνεργάστηκαν εκατοντάδες σπουδαστές και ερευνητές βελτιώνοντας έναν αλγόριθμο πρότασης που προτείνει ταινίες υβριδοποιώντας εκατοντάδες διαφορετικές τεχνικές και προσεγγίσεις συνεργατικού φιλτραρίσματος για να βελτιώνοντας τη συνολική ακρίβεια [40]. Ακολούθως, το σύστημα σύστασης Cinematch αναλύει αυτόματα τις συνολικές βαθμολογίες ταινιών εβδομαδιαίως χρησιμοποιώντας μια

παραλλαγή της συσχέτισης του Pearson με όλες τις άλλες ταινίες για τον καθορισμό μιας λίστας "παρόμοιων" ταινιών που προβλέπει τις παρόμοιες προτιμήσεις. Όσο ο χρήστης παρέχει βαθμολογίες, το σύστημα υπολογίζει, σε πραγματικό χρόνο, μια παλινδρόμηση πολλαπλών παραλλαγών βάσει αυτών των συσχετίσεων και στην συνέχεια προσδιορίζει μια μοναδική, εξατομικευμένη πρόβλεψη για κάθε ταινία με βάση αυτές τις βαθμολογίες [40].

Τα συστήματα προτάσεων βάσει περιεχομένου μπορούν να παρέχουν προτάσεις για αντικείμενα "ψυχρής εκκίνησης" για τα οποία υπάρχουν λίγα ή καθόλου εκπαιδευτικά δεδομένα, αλλά συνήθως έχουν χαμηλότερη ακρίβεια από τα συνεργατικά συστήματα φιλτραρίσματος. Αντίθετα, οι συνεργατικές τεχνικές φιλτραρίσματος συχνά παρέχουν ακριβείς προτάσεις, αλλά αποτυγχάνουν σε αντικείμενα ψυχρής εκκίνησης. Τα υβριδικά σχήματα προσπαθούν να συνδυάσουν αυτά τα διαφορετικά είδη πληροφοριών για να δώσουν καλύτερες προτάσεις.

2.3.5 Πολυκριτηριακά Συστήματα Συστάσεων

Τα πολυκριτηριακά συστήματα σύστασης χρησιμοποιούν σαν κύριο χαρακτηριστικό τους τη βαθμολόγηση. Η βαθμολογία είναι μια πράξη με την οποία οι χρήστες δηλώνουν ρητά την άποψη και το γούστο τους σχετικά με ένα αντικείμενο. Σε ένα μονοκριτηριακό σύστημα, ο χρήστης είναι σε θέση να δώσει μόνο μία τιμή αξιολόγησης που θα ήταν η δική του βαθμολογία για το αντικείμενο αυτό, ωστόσο, σε συστήματα πολλαπλών κριτηρίων, ο χρήστης μπορεί να αξιολογήσει διαφορετικά χαρακτηριστικά ενός μοναδικού αντικειμένου [41]. Οι τιμές αξιολόγησης, εννοείται πως μπορεί να διαφέρουν ανάλογα με τη «διάσταση» του αντικειμένου που αξιολογεί ο χρήστης. Παραδοσιακά, η συντριπτική πλειοψηφία των συστημάτων σύστασης που υπάρχουν στη βιβλιογραφία έχει επικεντρωθεί στην παροχή προτάσεων μοντελοποιώντας τη χρησιμότητα ενός χρήστη (ή της προτίμησή τους) για ένα αντικείμενο ως μία αξιολόγηση προτίμησης. Ωστόσο, όπου είναι δυνατόν, η λήψη μεγαλύτερης ποικιλίας των προτιμήσεων ενός χρήστη σε διάφορες διαστάσεις — για παράδειγμα, η καταγραφή όχι μόνο της προτίμησης του χρήστη για μια δεδομένη ταινία, αλλά και των προτιμήσεών του για συγκεκριμένες πτυχές της ταινίας (όπως είδος, υπόθεση ή οπτικά εφέ) —μπορεί να προσφέρει ευκαιρίες για περαιτέρω βελτιώσεις στην ποιότητα των προτάσεων [42]. Οι πρόσθετες πληροφορίες που παρέχονται από αξιολογήσεις πολλαπλών κριτηρίων θα μπορούσαν επίσης να βοηθήσουν να βελτιωθεί την ποιότητα των συστάσεων, διότι θα μπορούσε να αντιπροσωπεύει τις πιο περίπλοκες προτιμήσεις κάθε χρήστη [42]. Λαμβάνοντας υπόψη τη βασική διαφορά μεταξύ των παραδοσιακών και πολυκριτηριακών συστημάτων σύστασης, απαιτούνται πρόσθετες τεχνικές ώστε να ενσωματωθούν τα όλα τα κριτήρια κατά την ανάπτυξη των συστημάτων [43].

Τα συστήματα πολλαπλών κριτηρίων κατηγοριοποιούνται ανάλογα με τη φύση του προβλήματος- απόφασης που υποστηρίζουν σε δύο μεγάλες κατηγορίες, εύρεσης και

κατάταξης [44]. Η συντριπτική πλειονότητα των συστημάτων πολλαπλών κριτηρίων που αναλύθηκαν στοχεύει την «εύρεση καλών αντικειμένων». Υπάρχουν, βέβαια και άλλα συστήματα πολλαπλών κριτηρίων που έχουν άλλη στοχοθεσία, αλλά σε πολύ πιο περιορισμένη κλίμακα. Η δεύτερη κατηγορία, αφορά συστήματα που στοχεύουν στην «Κατάταξη» εναλλακτικών αντικειμένων». Στη βιβλιογραφία, εντοπίζονται επίσης και συστήματα «Ταξινόμησης», «Επιλογής» και «Περιγραφής».

Πώς όμως λειτουργεί ένα σύστημα πολλαπλών κριτηρίων στην πράξη; Είναι σαφές πως ορισμένες διαστάσεις για ένα αντικείμενο (προϊόν ή υπηρεσία) θα κυριαρχήσουν και ότι αυτές θα είναι διαφορετικές για διαφορετικούς χρήστες. Ακολουθώντας αυτήν την υπόθεση, οι χρήστες συγκεντρώνονται με βάση τις προτιμήσεις κριτηρίων τους, δημιουργώντας ένα "πλέγμα προτιμήσεων". Το αποτέλεσμα της πρότασης για έναν χρήστη βασίζεται σε αξιολογήσεις από άλλους χρήστες από τα ίδια ή κοντινά πλέγματα, ενώ οι συστάσεις αποφασίζονται βάσει 3 τεχνικών: (α) χρησιμοποιώντας τη συνάρτηση συνάθροισης των κριτηρίων, (β) χρησιμοποιώντας τις συνολικές αξιολογήσεις στοιχείων και (γ) συνδυάζοντας τη συμπλέγματα με τεχνικές συνεργατικού φιλτραρίσματος [45].

Πολλοί ερευνητές, προσεγγίζουν ένα σύστημα πολλαπλών κριτηρίων, ως ένα πρόβλημα λήψης αποφάσεων πολλαπλών κριτηρίων οπότε για την επίλυση ανάλογων ζητημάτων παροχής συστάσεων, εφαρμόζουν τεχνικές που έρχονται από την επιστήμη της λήψης αποφάσεων [46].

2.3.6 Φιλτράρισμα με βάση τη γνώση

Αν μπορούσαμε να συνοψίσουμε τα συστήματα σύστασης σε δύο μεγάλες κατηγορίες αυτές είναι οι εξής: στην πρώτη κατηγορία ανήκουν αυτά που χρησιμοποιούν στατιστικές προσεγγίσεις που εξετάζουν τη συνολική συμπεριφορά προηγούμενων χρηστών ώστε να κάνουν μελλοντικές προτάσεις, ενώ τα δεύτερα εστιάζουν σε προσεγγίσεις βασισμένες στη γνώση που κατασκευάζουν ένα μοντέλο που προσπαθεί να αντικατοπτρίσει τις πιθανές επιλογές του χρήστη. Κάθε ένα από αυτά έχει μειονεκτήματα: οι στατιστικές προσεγγίσεις απαιτούν μεγάλες ποσότητες αρχικών δεδομένων και δεν μπορούν να χειριστούν τις σχέσεις που πιθανότατα θα έχουν αυτά τα δεδομένα, ενώ οι προσεγγίσεις βασισμένες στη γνώση απαιτούν γνώσεις τεχνικών βαθιάς μηχανικής [47]. Υπάρχουν περιπτώσεις σε σενάρια χρήσης και ανάπτυξης συστημάτων που δεν καλυπτόμαστε από τεχνικές συνεργατικού φιλτραρίσματος ή φιλτραρίσματος βάσει περιεχομένου. Στις περιπτώσεις αυτές χρειάζονται σαφή χαρακτηριστικά και ιδιότητες των αντικειμένων προκειμένου να χαρακτηριστεί μια σύσταση ως επιτυχημένη [48].

Τα συστήματα σύστασης βασισμένα στη γνώση (αλλιώς ως σύστασεις βασισμένες στη γνώση) είναι ένας συγκεκριμένος τύπος συστήματος συστάσεων που βασίζεται σε σαφείς γνώσεις σχετικά με αντικείμενα και χρήστες, ακολουθώντας μια προσέγγιση που δημιουργεί συστάσεις

αιτιολογώντας γιατί κάποια αντικείμενα ταιριάζουν σε συγκεκριμένους χρήστες [49].Τα συστήματα σύστασης που βασίζονται στη γνώση εκτελούν μια σημαντική λειτουργία σε ένα κόσμο με συνεχώς επεκτεινόμενους πόρους πληροφοριών. Οι πληροφορίες αυτές αποκτούν εννοιολογική σημασία όταν γίνονται γνώσεις και βάσει αυτών μπορούν να ληφθούν σωστές αποφάσεις [50].

Σε αντίθεση με άλλα συστήματα σύστασης, όπως αναφέρθηκε νωρίτερα, τα συστήματα σύστασης βασισμένα στη γνώση δε βασίζονται σε ένα μεγάλο πλήθος στατιστικών πληροφοριών με συγκεκριμένα αξιολογημένα αντικείμενα ή συγκεκριμένους χρήστες [51]. Αυτό καθιστά τα τέτοιου είδους συστήματα σύστασης ακόμη πιο πολύτιμα ιδιαίτερα σε περιπτώσεις που αυτά λειτουργούν συμπληρωματικά σε άλλα συστήματα συστάσεων (υβριδικά) [50].

Η σύσταση βασισμένη στη γνώση βασίζεται στα ακόλουθα δεδομένα: (α) ένα σύνολο κανόνων (περιορισμών) ή μετρήσεων ομοιότητας και (β) ένα σύνολο αντικειμένων [52]. Ανάλογα με τις απαιτήσεις του χρήστη τη δεδομένη στιγμή, οι κανόνες (περιορισμοί) που περιεγράφηκαν πριν, ορίζουν ουσιαστικά ποια στοιχεία πρέπει να προτείνονται [52].

2.3.7 Συστήματα βασισμένα σε περιόδους σύνδεσης

Πολλά συστήματα σύστασης που χρησιμοποιούνται στο ηλεκτρονικό εμπόριο (ειδικά αυτά των μικρών λιανοπωλητών) καθώς και από τους περισσότερους ιστοτόπους ειδήσεων δεν παρακολουθούν τα αναγνωριστικά χρήστη. Τα ψηφιακά αποτυπώματα των cookies και του προγράμματος περιήγησης μπορεί να παρέχουν κάποιο επίπεδο αναγνωρισιμότητας για τους χρήστες που επισκέπτονται αυτού του τύπου τους ιστότοπους, εντούτοις αυτές οι τεχνολογίες συχνά δεν είναι αρκετά αξιόπιστες, ενώ ταυτόχρονα εγείρουν ανησυχίες για την προστασία της ιδιωτικής ζωής. Ακόμα κι αν είναι δυνατή η παρακολούθηση, πολλοί χρήστες έχουν μόνο μία ή δύο περιόδους σύνδεσης σε έναν μικρότερο ιστότοπο ηλεκτρονικού εμπορίου και σε ορισμένους τομείς αυτών (π.χ. ταξινομημένοι ιστότοποι) [53], με αποτέλεσμα η παρακολούθησή τους με δεδομένα τέτοιου τύπου να κρίνεται μη αποτελεσματική.

Η επιλογή ενός αντικειμένου από έναν χρήστη δεν εξαρτάται μόνο από τη μακροπρόθεσμη προτίμηση του (προσκόλληση σε συγκεκριμένα αντικείμενα βάσει του χαρακτήρα του), αλλά και από τη βραχυπρόθεσμη προτίμησή του σε ένα χρονικά ευαίσθητο πλαίσιο (που περιλαμβάνει π.χ. τα στοιχεία που προβλήθηκαν ή αγοράστηκαν πρόσφατα) [54]. Ταυτόχρονα, όπως είχε αναφερθεί στην ανάλυση των δημογραφικού φιλτραρίσματος τα δεδομένα που δηλώνουν την προτίμηση ενός χρήστη, δεν είναι ένα στατικό δεδομένο αλλά δυναμικό που μεταβάλλεται με την πάροδο του χρόνου [30]. Αυτά τα δύο προβλήματα, δε μπορούσαν να λυθούν από κανένα αυτοτελές σύστημα σύστασης ωστόσο αποτέλεσαν έναυσμα για την ανάπτυξη συστημάτων σύστασης βασισμένα σε περιόδους σύνδεσης.

Τα συστήματα αυτά μαθαίνουν τις προτιμήσεις των χρηστών από τις συνδέσεις που έχει ο χρήστης και δημιουργούνται κατά τη διάρκεια της πλοήγησης του χρήστη σε μία πλατφόρμα ή σελίδα. Κάθε σύνδεση αποτελείται από πολλαπλές αλληλεπιδράσεις χρηστών-αντικειμένων που συμβαίνουν μαζί σε μια συνεχή χρονική περίοδο ενώ λαμβάνοντας κάθε σύνδεση ως δομικό στοιχείο, ένα σύστημα βασισμένο σε περίοδο σύστασης είναι σε θέση να καταγράψει τόσο τη βραχυπρόθεσμη προτίμηση ενός χρήστη όσο και τη δυναμική των προτιμήσεων του[54].

Ένα μεγάλο μέρος της έρευνας στον τομέα των συστημάτων σύστασης έχει επικεντρωθεί σε μοντέλα που λειτουργούν όταν είναι διαθέσιμο το αναγνωριστικό χρήστη (username), βοηθώντας να δημιουργηθεί ένα σαφές προφίλ χρήστη [54]. Η μέθοδος, ωστόσο πρέπει να παρέχει έναν απλό τρόπο αναπαραγωγής και συνδυασμού αναγνωριστικών στοιχείων με μεταδεδομένα. Συνήθως ένα αντικείμενο σχετίζεται με χαρακτηριστικά διαφορετικών τύπων. Για παράδειγμα, ένα προϊόν μπορεί να έχει ένα αναγνωριστικό, όνομα και περιγραφή και γενικά ανήκει σε μία ή περισσότερες κατηγορίες (μερικές φορές οργανωμένες σε ιεραρχίες κατηγοριών). Θα ήταν πιο βολικό να ορίσουμε έναν γενικό τρόπο περιγραφής ώστε να αναπαριστώνται ταυτόχρονα διαφορετικοί τύποι χαρακτηριστικών και να διαμορφώνονται από κοινού οι αλληλεπιδράσεις τους, λαμβάνοντας ωστόσο υπόψη τις σχέσεις και τις εξαρτήσεις μεταξύ τους[55].

Με βάση τις διαθέσιμες πληροφορίες από τη συγκεκριμένη συνεδρία χρήστη, το σύστημα σύστασης θα πρέπει να δημιουργήσει ένα μοντέλο για τον χρήστη και να κάνει προβλέψεις. Τα δεδομένα περιόδου σύνδεσης έχουν πολλά σημαντικά χαρακτηριστικά τα οποία είναι τα ακόλουθα [56]:

- Τα κλικ καθώς και η πλοήγηση του χρήστη κατά τη διάρκεια μιας συνεδρίας είναι διαδοχικά από τη φύση τους. Η σειρά των κλικ καθώς και η διαδρομή πλοήγησης ενδέχεται να περιέχουν πληροφορίες σχετικά με την πρόθεση του χρήστη. Για παράδειγμα, αν ο χρήστης κοιτάζει κατά την πλοήγησή του σε μια ιστοσελίδα την αγορά ενός συγκεκριμένου ρούχου (π.χ. μπουφάν) τότε υπάρχει μια μεγάλη πιθανότητα ο χρήστης να έχει όντως αυτή την ανάγκη και να προχωρήσει σε αγορά του συγκεκριμένου είδους.
- Τα αντικείμενα που προβλήθηκαν έχουν συχνά μεταδεδομένα, όπως ονόματα, κατηγορίες και περιγραφές, τα οποία παρέχουν πληροφορίες σχετικά με τις προτιμήσεις του χρήστη και τι αναζητούν. Για παράδειγμα, σε περίπτωση που ένας χρήστης επιλέξει να αγοράσει ένα αντικείμενο λ.χ. ένα αυτοκίνητο μιας συγκεκριμένης μάρκας (π.χ. γαλλικής) και στη συνέχεια κοιτάει αυτοκίνητα κατασκευαστών που προέρχονται από Γαλλία, καταλαβαίνουμε πως ο χρήστης δείχνει προσήλωση στη χώρα προέλευσης του αυτοκινήτου που τον ενδιαφέρει.

- Οι συνεδρίες είναι περιορισμένες τόσο σε χρόνο όσο και σε έκταση. Μια συνεδρία έχει έναν συγκεκριμένο στόχο και γενικά τελειώνει όταν επιτευχθεί αυτός ο στόχος: ενοικίαση ξενοδοχείου για επαγγελματικό ταξίδι, εύρεση εστιατορίου για ρομαντική ημερομηνία και ούτω καθεξής. Αυτό σημαίνει ότι η συνεδρία έχει εγγενή πληροφοριακή ισχύ που σχετίζεται με ένα συγκεκριμένο στοιχείο (όπως το ξενοδοχείο ή το εστιατόριο που έχει κλείσει τελικά).

Πολλά μοντέλα που έχουν αναπτυχθεί στην σύσταση με βάση την περίοδο σύνδεσης χρησιμοποιούν τεχνολογία αναδρομικών νευρωνικών δικτύων (ΑΝΔ). Οι τεχνολογίες αυτές περιέχουν παραλλαγές του σταθερού ΑΝΔ οι οποίες καθιστούν πιο πιθανή την παραγωγή αποτελεσματικότερων συστάσεων, όπως για παράδειγμα την συνάρτηση απώλειας κατάταξης. Τα συστήματα αυτά φαίνεται να λειτουργούν ιδιαίτερα αποτελεσματικά και σε περιπτώσεις ψυχρής εκκίνησης, προτού δηλαδή το σύστημα μάθει πολλά για τα τρέχοντα ενδιαφέροντα του χρήστη. Αυτό γίνεται χρησιμοποιώντας ένα δεύτερο RNN που χρησιμοποιείται από τις πρόσφατες συνεδρίες ώστε να προβλέψουν τις προτιμήσεις του χρήστη για την τρέχουσα περίοδο λειτουργίας [57]. Με την «προσθήκη» αυτών των πληροφοριών στο αρχικό RNN, δύναται να βελτιωθούν οι συστάσεις.

Παρόλη τη μεγάλη διάδοση των συστάσεων βασισμένων σε νευρωνικά δίκτυα, υπάρχουν έρευνες οι οποίες δεν υποστηρίζουν πως οι υβριδικές μέθοδοι αποτελούν πανάκεια των συστημάτων σύστασης. Συγκεκριμένα, υπάρχουν πειράματα τα οποία δείχνουν ότι οι απλές ευρετικές μέθοδοι που βασίζονται σε αλγορίθμους πλησιέστερων γειτόνων είναι προτιμότερες από εννοιολογικά και υπολογιστικά πιο πολύπλοκες μεθόδους που χρησιμοποιούν τα νευρωνικά δίκτυα [58]. Αυτό, ωστόσο δείχνει ότι υπάρχει ακόμη περιθώριο για ακόμη πιο αποτελεσματικές μεθόδους νευρωνικών συστάσεων στο μέλλον [59].

Ακόμη μια παραλλαγή των συστάσεων βασισμένων σε περιόδους σύνδεσης, είναι το σύστημα σύστασης περιόδου σύνδεσης βασισμένο στη ροή. Η παραλλαγή αυτή έρχεται να αντιμετωπίσει δύο προκλήσεις που θέτουν οι προκάτοχοί της και είναι (1) η αβεβαιότητα των συμπεριφορών των χρηστών και (2) η συνεχούς, μεγάλης έντασης και υψηλής ταχύτητας φύσης των δεδομένων περιόδου σύνδεσης [60]. Προκειμένου να αντιμετωπιστούν αυτά τα δύο ζητήματα, οι Guo et. All το 2019 πρότειναν το εξής μοντέλο: για να γίνει πιο εύληπτη η συμπεριφορά των χρηστών, προτάθηκε ένα μοντέλο προσοχής βασισμένο στο Matrix Factorization (MF), το οποίο βελτιώνει τον κοινόχρηστο μηχανισμό προσοχής, αξιοποιώντας τις ιστορικές αλληλεπιδράσεις του χρήστη, ενώ για να αντιμετωπιστεί η πρόκληση μεγάλου όγκου και υψηλής ταχύτητας, προτάθηκε ένα μοντέλο ροής που βασίζεται σε δεξαμενή που χρησιμοποιεί μια ενεργή στρατηγική δειγματοληψίας για τη βελτίωση της αποτελεσματικότητας της ενημέρωσης του μοντέλου [60].

Υπάρχουν ωστόσο και άλλες τεχνικές που βασίζονται στη σύσταση περιόδου βασισμένη στη ροή. Η λογική ανάπτυξης τέτοιων συστημάτων έγκειται στο γεγονός ότι τα συστήματα σύστασης πρέπει να εξετάζουν συνεχώς τα πιο πρόσφατα στοιχεία της ροής [61]. Ως ροή, μπορεί να ερμηνευθεί η λίστα όλων των δεδομένων που έρχονται στο σύστημα. Σκόπιμο είναι λοιπόν, να δοθεί βαρύτητα στα τελευταία στοιχεία που εισέρχονται σε αυτήν, καθώς τα τελευταία στοιχεία που εξετάζει ο χρήστης είναι πιο χρήσιμα στην παραγωγή σύστασης [62]. Σε αυτό το πλαίσιο, οι Sun et al πρότειναν την αρχιτεκτονική SANSR [63]. Η αρχιτεκτονική αυτή προτείνει τα εξής βήματα. Πρώτον, η διαδοχική σειρά της συνεδρίας μετατρέπεται σε μια ακολουθία σταθερού μήκους ενώ ταυτόχρονα γίνεται και η ενσωμάτωση των στοιχείων τους σε πυκνούς χώρους μικρών διαστάσεων. Το επόμενο βήμα είναι να χρησιμοποιηθούν μπλοκ αυτο-προσοχής για να καταγραφούν οι σχέσεις μεταξύ της εισόδου και της εξόδου καθώς και των αλληλεπιδράσεων μεταξύ των αντικειμένων. Στη συνέχεια, χρησιμοποιείται ένα δίκτυο πολλαπλής προσοχής για να γνωστοποιηθούν συναλλαγές των αντικειμένων υψηλού επιπέδου. Τελευταίο βήμα αυτής της αρχιτεκτονικής είναι ένα επίπεδο πρόβλεψης για να προβλεφθεί το επόμενο κλικ της περιόδου λειτουργίας, με τη χρήση μιας αντικειμενικής συνάρτησης αντιστοίχισης που χρησιμοποιεί το κριτήριο βελτιστοποίησης BPR, αλλά και μια μέθοδο περαιτέρω εκπαίδευσης του μοντέλου.

Ένα άλλο πλαίσιο που βασίζεται στην ίδια τεχνική είναι το StreamingRec. Η τεχνική αυτή εφαρμόζει ένα πρωτόκολλο αξιολόγησης βασισμένο σε επανάληψη που επιτρέπει στους αλγορίθμους να ενημερώνουν τα μοντέλα σε πραγματικό χρόνο όταν καταγράφονται νέα συμβάντα ώστε να είναι διαθέσιμα για σύσταση [61].

2.4 Περιορισμοί στα συστήματα συστάσεων

2.4.1 Το πρόβλημα της «ψυχρής εκκίνησης»

Ο όρος «ψυχρή εκκίνηση» προέρχεται από αυτοκίνητα. Όταν έχει πολύ κρύο, ο κινητήρας ενός αυτοκινήτου έχει προβλήματα με την εκκίνηση, αλλά μόλις φτάσει στη βέλτιστη θερμοκρασία λειτουργίας του, λειτουργεί ομαλά [64]. Με τα συστήματα προτάσεων, το «κρύο ξεκίνημα» σημαίνει απλώς ότι οι συνθήκες δεν είναι ακόμη οι βέλτιστες ώστε να προσφερθούν τα καλύτερα δυνατά αποτελέσματα. Το πρόβλημα της ψυχρής εκκίνησης σχετίζεται με την ανεπάρκεια των πληροφοριών (για χρήστες και αντικείμενα) που διατίθενται στον αλγόριθμο προτάσεων. Υπάρχουν τρεις τύποι προβλημάτων ψυχρής εκκίνησης και αφορούν: (α) συστάσεις για νέους χρήστες, (β) συστάσεις για νέα αντικείμενα και (γ) συστάσεις για νέα αντικείμενα σε νέους χρήστες [65]. Στη βιβλιογραφία το πρόβλημα της ψυχρής εκκίνησης εμφανίζεται και ως πρόβλημα πρώτης εκκίνησης ενώ η ετυμολογία του όρου προέρχεται από

το γεγονός πως είναι αδύνατο για το σύστημα να εκτιμήσει τους νεοεισερχόμενα αντικείμενα χωρίς μια πρώτη εκτίμηση από τους χρήστες [66].

Α) Συστάσεις για νέους χρήστες

Το πρόβλημα της εκκίνησης ψυχρής εκκίνησης για νέους χρήστες αποτελεί ένα σοβαρό πρόβλημα στα συστήματα σύστασης, καθώς μπορεί να οδηγήσει στην απώλεια νέων χρηστών που αποφασίζουν να σταματήσουν να χρησιμοποιούν το σύστημα λόγω της έλλειψης ακρίβειας στις συστάσεις που ελήφθησαν στο πρώτο στάδιο, στάδιο στο οποίο δεν έχουν ακόμη ανιχνευθεί οι προτιμήσεις τους. Αυτό φυσικά έχει ως αποτέλεσμα οι συστάσεις να είναι μη αποτελεσματικές καθώς δεν έχει ληφθεί υπ' όψιν η βαθμολογία των χρηστών [67]. Επίσης, πρόβλημα ψυχρής εκκίνησης σε επίπεδο χρηστών συναντάται σε περιπτώσεις που ένας νέος χρήστης έχει παρουσιάσει λίγες μόνο απόψεις. Σε τέτοιες περιπτώσεις, δεν υπάρχει αλληλεπίδραση μεταξύ του νέου χρήστη και των άλλων, και ως εκ τούτου δεν είναι δυνατόν να μετρηθεί η ομοιότητα μεταξύ τους [68]. Ως αποτέλεσμα, τα συστήματα σύστασης που αντιμετωπίζουν αυτό το πρόβλημα δεν είναι σε θέση να κάνουν αξιόπιστες προτάσεις.

Β) Συστάσεις για νέα αντικείμενα

Το πρόβλημα της ψυχρής εκκίνησης σε νέα αντικείμενα δημιουργείται όταν τα στοιχεία που προστίθενται στο σύστημα δεν έχουν καμία ή έχουν πολύ μικρή αλληλεπίδραση μεταξύ τους. Το γεγονός αυτό δημιουργεί πρόβλημα κυρίως στους αλγορίθμους φιλτραρίσματος διότι βασίζονται στις αλληλεπιδράσεις του αντικειμένου για να κάνουν προτάσεις. Εάν δεν υπάρχουν αλληλεπιδράσεις, τότε ένας καθαρός συνεργατικός αλγόριθμος δεν μπορεί να προτείνει το αντικείμενο. Σε περίπτωση που οι διαθέσιμες αλληλεπιδράσεις είναι λίγες, αν και ένας συνεργατικός αλγόριθμος θα είναι σε θέση να προτείνει ένα αντικείμενο, η ποιότητα αυτών των συστάσεων θα είναι κακή [67].

Γ) Συστάσεις για νέα αντικείμενα σε νέους χρήστες

Το πρόβλημα αυτό, όπως φαίνεται από τον τίτλο αφορά την εκκίνηση του συστήματος, όταν τόσο οι χρήστες όσο και τα αντικείμενα είναι νεοεισερχόμενα στο σύστημα. Αυτό δημιουργεί ελάχιστες συνθήκες σωστής σύστασης, διότι οι χρήστες δεν έχουν αξιολογήσει τα αντικείμενα.

Μία από τις κύριες λύσεις στο πρόβλημα της ψυχρής εκκίνησης στα συστήματα σύστασης, είναι η ενεργητική μάθηση, η οποία είναι ήδη γνωστή ως ένα μέρος ενός ευρύτερου ερευνητικού θέματος, της Μηχανικής Μάθησης [69]. Τι σημαίνει όμως η «ενεργητική μάθηση»; Το σύστημα πρέπει να επικεντρωθεί στη συλλογή δεδομένων μόνο υψηλής ποιότητας, το οποίο θα επιτευχθεί ελέγχοντας προσεκτικά τη διαδικασία συλλογής δεδομένων, βοηθώντας έτσι το σύστημα να ελαχιστοποιήσει το κόστος βελτιώνοντας όμως παράλληλα και την αναμενόμενη ακρίβεια από τις συστάσεις που θα παραχθούν[69]. Σε αυτό το πλαίσιο, το σύστημα ζητά ενεργά από το χρήστη να αξιολογήσει ένα σύνολο στοιχείων, τα οποία προσδιορίζονται χρησιμοποιώντας μια στρατηγική που στοχεύει στην καλύτερη «ανακάλυψη»

των ενδιαφερόντων του χρήστη και κατά συνέπεια στη βελτίωση της ποιότητας των συστάσεων [70].

Άλλες λύσεις, οι οποίες υπάρχουν στη βιβλιογραφία, εξαρτώνται κάθε φορά από το πρόβλημα το οποίο τίθεται προς επίλυση. Για παράδειγμα, ένας τρόπος αντιμετώπισης του προβλήματος χρησιμοποιεί τεχνικές συνάθροισης προτιμήσεων και κρίσεων που υπάρχουν στη θεωρία κοινωνικής επιλογής για να "αλλάξουν" τον συνεργατικό αλγόριθμο φιλτραρίσματος [71]. Ακόμη μια λύση είναι αυτή που προτάθηκε στα πλαίσια του έργου ClusterUCBscRec, το οποίο υιοθετεί μια στρατηγική «εξερεύνησης και εκμετάλλευσης» για τη συνεχή βελτίωση της προτεινόμενης απόδοσης και ως εκ τούτου περνάει γρήγορα στο στάδιο της ψυχρής εκκίνησης [72]. Σε μια άλλη λύση, σύμφωνα με την Ντούτση κ.ά , δεν αναζητούνται παρόμοιοι χρήστες σε ολόκληρη τη βάση χρηστών, αλλά γίνεται προ-διαχωρισμός τους σε ομάδες παρόμοιων και σε συνδυασμό με έναν αλγόριθμο top-k και ομαδοποιούνται οι συστάσεις, οι οποίες δεν αφορούν καθένα χρήστη χωριστά, αλλά λαμβάνεται υπ' όψιν η ομάδα στην οποία εντάσσονται [73]. Ακόμη ένας τρόπος που εμφανίζεται στη βιβλιογραφία είναι η χρήση ενός κατωφλίου προκειμένου να προσδιοριστεί η η «ψυχρότητα» ενός αντικειμένου χρησιμοποιώντας είτε το χρονικό διάστημα που υπάρχει στο σύστημα είτε τον όγκο των δεδομένων που συγκεντρώθηκαν για αυτό. Τα θερμά αντικείμενα, αποκτούν μεγαλύτερη πίστωση από τα ψυχρά σε ένα τέτοιο σύστημα σύστασης.

2.4.2 Το πρόβλημα της πλήρους εξειδίκευσης

Μερικές φορές οι χρήστες περιορίζονται στο να λαμβάνουν προτάσεις που μοιάζουν με αυτές που είναι ήδη γνωστές στα προφίλ τους και αυτό χαρακτηρίζεται ως πρόβλημα εξειδίκευσης. Ουσιαστικά ο χρήστης αποτρέπεται από την ανακάλυψη νέων αντικειμένων και άλλων διαθέσιμων επιλογών, οπότε το σύστημα συστάσεων δεν χαρακτηρίζεται από ποικιλία που είναι πλήρως επιθυμητή. Μετά την επίλυση του προβλήματος χρησιμοποιώντας γενετικούς αλγόριθμους, ο χρήστης θα διαθέτει ένα σύνολο διαφορετικών και ένα ευρύ φάσμα εναλλακτικών λύσεων [74]. Η ελευθερία στα συστήματα συστάσεων χαρακτηρίζεται από σύσταση αντικειμένων που αποτελούν όντως ενδιαφέροντα του χρήστη, ταυτόχρονα όμως συνοδεύεται και από ένα αίσθημα «έκπληξης» για χρήστες που λαμβάνουν απροσδόκητες προτάσεις [75]. Πώς όμως ορίζεται το αίσθημα του απροσδόκητου; Είναι εύλογο να διαπιστωθεί ότι όταν ένας χρήστης ακούει ένα συγκεκριμένο μουσικό κομμάτι, το πιο πιθανό είναι να του προταθεί να ακούσει κάτι από τον ίδιο καλλιτέχνη / άλμπουμ, τραγούδια τα οποία είναι ιδιαίτερα δημοφιλή, ή τραγούδια τα οποία έχουν ήδη ακουστεί από τον χρήστη. Οπότε, ένα σύστημα σύστασης το οποίο έχει διαπεράσει αυτό το πρόβλημα έχει δύο κύρια χαρακτηριστικά: παράγει συστάσεις συνειδητές σε συναφή αντικείμενα και δεύτερον, η σύσταση περιλαμβάνει πολλές φορές το συναίσθημα του απροσδόκητου.

Μια από τις πολλές τεχνικές που έχουν αναπτυχθεί για την εξέλιξη συστημάτων είναι η πρόταση Outside-The-Box (otb), η οποία εμπεριέχει κάποιο ρίσκο ώστε να βοηθήσει τους χρήστες να κάνουν νέες «ανακαλύψεις», διατηρώντας παράλληλα υψηλό δείκτη σχετικότητας σε προηγούμενες επιλογές τους [76]. Σε αυτό το επίπεδο, μπορούν να προταθούν επιπλέον λύσεις ώστε να περιλαμβάνονται ποιοτικές νέες ανακαλύψεις, το οποίο μπορεί να επιτευχθεί χρησιμοποιώντας γενετικούς αλγόριθμους που φέρνουν ποικιλομορφία στις προτάσεις που γίνονται. Το πρόβλημα θεωρείται σχετικά μικρής κλίμακας σε συστάσεις βασισμένες σε περιεχόμενο όπου μπορεί να προταθούν μη αναμενόμενα και νέα αντικείμενα.

2.4.3 Το πρόβλημα της συνωνυμίας

Το πρόβλημα της συνωνυμίας εμφανίζεται συνήθως σε περιπτώσεις όπου πολύ παρόμοια αντικείμενα δεν έχουν παρόμοια ονόματα ή καταχωρήσεις. Τα περισσότερα συστήματα σύστασης δυσκολεύονται να κάνουν διάκριση μεταξύ στενά συνδεδεμένων αντικειμένων όπως για παράδειγμα η διαφορά μεταξύ π.χ. του όρου παιδική λογοτεχνία και του παιδικού βιβλίου [77]. Σε αυτήν την περίπτωση, θα πρέπει να χρησιμοποιηθεί μια μέθοδος ώστε να διερευνήσει την ύπαρξη μιας τέτοιας λανθάνουσας συσχέτισης και να την αξιοποιήσει ώστε να δημιουργηθούν καλύτερες συστάσεις [78]. Παρόλο που το συνωνυμικό πρόβλημα υποβαθμίζει την απόδοση του συστήματος σύστασης μπορεί να επιλυθεί με διάφορες τεχνικές όπως: λανθάνουσας σημασιολογικής ευρετηρίασης, τεχνικές αποσύνθεσης μονής τιμής, οντολογίες [79], [80].

2.4.4 Το πρόβλημα της ένεσης προφίλ

Η βασική αρχή που διέπει τα συστήματα σύστασης είναι ότι ένας χρήστης που μοιράστηκε τις ίδιες προτιμήσεις με μια ομάδα άλλων χρηστών στο παρελθόν είναι πιθανό να μοιραστεί τις ίδιες προτιμήσεις μαζί τους στο μέλλον. Αυτό όμως μπορεί να χρησιμοποιηθεί κακόβουλα, από χρήστες ή εταιρείες οι οποίοι προσπαθούν να κατευθύνουν τους χρήστες σε «συγκεκριμένες» επιλογές. Τα συστήματα σύστασης εξαιρετικά ευάλωτα σε επιθέσεις που εισάγουν τέτοιου είδους προκατειλημμένες πληροφορίες σε μια προσπάθεια να οδηγηθούν οι χρήστες σε ψευδείς προτάσεις [81]. Αυτές οι επιθέσεις χαρακτηρίζονται ως «επιθέσεις ένεσης προφίλ». Οι επιθέσεις αυτές μπορούν να διαχωριστούν σε δύο επιμέρους επίπεδα: οι επιθέσεις «push», που χρησιμοποιούνται για την προώθηση ενός αντικειμένου, και οι επιθέσεις «nuke», που χρησιμοποιούνται για τον υποβιβασμό ενός αντικειμένου [81].

Οι περισσότερες επιθέσεις ένεσης ακολουθούν την εξής σειρά βημάτων: ο εισβολέας αποκτά διαφορετικές ταυτότητες μέσα στο σύστημα σύστασης και δημιουργεί ένα προφίλ χρήστη για κάθε μια από αυτές τις ταυτότητες. Αυτό αναφέρεται ως προφίλ επίθεσης. Μέσα σε κάθε

προφίλ, ο εισβολέας χειραγωγεί τη σύσταση βαθμολογώντας ή συνιστώντας ένα συγκεκριμένο στοιχείο έναντι των υπολοίπων [82].

Γενικότερα, οι επιθέσεις ένεσης προφίλ έχουν επίπτωση τόσο στην ακρίβεια και όσο και την αξιοπιστία του συστήματος σύστασης. Επομένως, ο τρόπος ανίχνευσης των επιθέσεων ένεσης είναι μια μεγάλη πρόκληση στις μελλοντικές μελέτες των συστημάτων σύστασης [83]. Προκειμένου να αντιμετωπιστούν τέτοιου είδους επιθέσεις, αναπτύσσονται ειδικοί αλγόριθμοι εντοπισμού τέτοιων προφίλ [81].

2.4.5 Το πρόβλημα των απρόβλεπτων αντικειμένων

Ακόμη ένα σημαντικό ζήτημα που προκύπτει στα συστήματα σύστασης είναι το πρόβλημα των απρόβλεπτων αντικειμένων που έχουν διαφορετική κατάταξη. Η έννοια της κατάταξης χρησιμοποιείται ευρέως στα συστήματα σύστασης και όπως αναφέρθηκε, ορίζει τη σειρά προτίμησης αντικειμένων από τον χρήστη. Με αυτόν τον τρόπο, το σύστημα σύστασης χρησιμοποιείται για να ανακαλύψει τη λίστα κατάταξης των προβλέψιμων λιστών στοιχείων καθώς και τη λίστα απρόβλεπτων αντικειμένων [84]. Η απρόβλεπτη λίστα αντικειμένων, μπορεί να περιλαμβάνει ασυνήθιστα προϊόντα που υπήρχαν πριν ή έχουν σχέση με τα ήδη υπάρχοντα, οπότε σε αυτό το σενάριο το σύστημα ενδέχεται να μην είναι σε θέση να προτείνει το προϊόν στους χρήστες. Αυτό ίσως δημιουργήσει προβλήματα σε νεοεισερχόμενα αντικείμενα, τα οποία δεν μπορούν να ενταχθούν σε κάποια κατηγορία από μόνα τους, άρα αποκλείονται από οποιαδήποτε μελλοντική σύσταση. Το πρόβλημα των «απρόβλεπτων αντικειμένων» συσχετίζεται σε μεγάλο βαθμό με το πρόβλημα της ψυχρής εκκίνησης.

Προκειμένου να επιλυθούν τέτοια ζητήματα, χρησιμοποιούνται υβριδικά μοντέλα που συνδυάζουν αλγορίθμους προσανατολισμένους στην ακρίβεια. Οι τεχνικές αυτές χρησιμοποιούν τεχνικές σημασιολογίας αντικειμένων [85]. Κατά την πρόταση νέων αλγορίθμων, οι ερευνητές τους συγκρίνουν με τις υλοποιήσεις των κορυφαίων προηγουμένων αλγορίθμων χρησιμοποιώντας προηγμένες μεθόδους αξιολόγησης [86].

2.4.6 Το πρόβλημα κινδύνου

Σε ορισμένες περιπτώσεις, μια σύσταση μπορεί να ενέχει πιθανούς κινδύνους, ανάλογα με την τελική επιλογή του χρήστη. Για παράδειγμα, όταν προτείνονται προϊόντα για αγορά, οι χρήστες μπορεί να επιθυμούν να αποφεύγουν τον κίνδυνο απότομης απώλειας αξίας του προϊόντος, προτιμώντας τα προϊόντα που έχουν χαμηλότερη αναμενόμενη αύξηση, αλλά και χαμηλότερο κίνδυνο πτώσης τιμής. Από την άλλη πλευρά, υπάρχουν χρήστες που μπορεί να αναζητούν προϊόντα που εμφανίζουν ένα κίνδυνο μείωσης της αξίας τους από την αγορά τους, έχοντας παράλληλα βλέψεις για δυνητικά υψηλό κέρδος. Σε τέτοιες περιπτώσεις μπορεί να μην επιθυμείται η αξιολόγηση μόνο της (αναμενόμενης) αξίας που γεννάται από μια πρόταση, αλλά

και της ελαχιστοποίησης του ρίσκου [87]. Ο κίνδυνος αυτός προσδιορίζεται με δύο τύπους αβεβαιότητας [88]. Ο πρώτος ονομάζεται παραμετρική αβεβαιότητα και σχετίζεται με την ελλιπή γνώση των παραμέτρων του προβλήματος. Ο δεύτερος τύπος ονομάζεται εγγενής αβεβαιότητα και σχετίζεται με τη στοχαστική φύση του συστημάτων τα οποία σε περιπτώσεις λανθασμένων επιλογών οδηγούν σε ένα καταστροφικό αποτέλεσμα. Η παροχή συστάσεων σε συστήματα που η εγγενής αβεβαιότητα έχει μεγάλη αξία είναι σημαντικά πιο δύσκολη [89]. Για παράδειγμα, δεν έχει την ίδια βαρύτητα το να προταθεί μια λάθος ταινία ή ένα λάθος βιβλίο σε ένα χρήστη, από το να του προταθεί ένα λάθος ιατρικό αντικείμενο, στο οποίο θα διαθέσει ένα μεγάλο μέρος του προϋπολογισμού του.

Ο τυπικός τρόπος αξιολόγησης τέτοιων συστημάτων είναι λαμβάνοντας υπόψη όχι μόνο το την αναμενόμενη χρησιμότητα, αλλά και τη διακύμανση της χρησιμότητας, που μπορεί να πραγματοποιηθεί εισάγοντας μια παράμετρο που σταθμίζει μια σύσταση ως επικίνδυνη ή μη [87].

2.5 Ιδιωτικότητα ενός συστήματος σύστασης

Ένα από τα επίσης πλέον σημαντικά ζητήματα ενός συστήματος σύστασης είναι η ιδιωτικότητά του [90]. Η ιδιωτικότητα σημαίνει ότι από το σύστημα σύστασης δεν πρέπει να διαρρεύσουν πληροφορίες πέρα από αυτές που εξάγονται ως συμπεράσματα ούτε σε τρίτους αλλά ούτε στον πάροχο υπηρεσιών [91]. Είναι πολύ σημαντικό ένα σύστημα σύστασης να παράγει ακριβείς προτάσεις, σεβόμενο παράλληλα και τους χρήστες του [92]. Επιπλέον, ορισμένοι ερευνητές προσθέτουν στην ιδιότητα του απορρήτου ενός συστήματος σύστασης ότι η έξοδός τους δεν πρέπει να βοηθάει τους πιθανούς «εισβολείς» να επαναπροσδιορίζουν τα αντικείμενα και τα χαρακτηριστικά τους. Σε έναν ιδεατό κόσμο, υπάρχει μια ισορροπία όπου το σύστημα είναι σε θέση να κάνει καλές προτάσεις, ενώ δεν απαιτεί από τους χρήστες να αποκαλύψουν πάρα πολλές πληροφορίες για τον εαυτό τους [93]. Οι ερευνητές έχουν εξετάσει τρόπους για τη διατήρηση του απορρήτου των χρηστών όπως π.χ. αλγόριθμους συνιστώντων αλλά και αλγόριθμους διατήρησης της ιδιωτικότητας. Μερικές εφαρμογές της ασφάλειας στα συστήματα σύστασης χρησιμοποιούν μια συμβατική προσέγγιση προσανατολισμένη σε πράκτορες [94]. Με αυτή την προσέγγιση επιλύεται το πρόβλημα της διασποράς και μπερδέματος κώδικα, αλλά ταυτόχρονα υπάρχει μικρότερη ανησυχία για την ασφάλεια του συστήματος σύστασης. Μια ακόμη τεχνική, που χρησιμοποιείται ευρέως κατά την ανάπτυξη τέτοιων αλγορίθμων χωρίζεται σε δύο βήματα [95] :

- Στο πρώτο βήμα αναγνωρίζονται όλες οι πιθανές διαδρομές επίθεσης σε ένα γράφημα, σύμφωνα με τις μεταβλητές.

- Οι πιθανές διαδρομές επίθεσης χρησιμοποιούνται σε συνδυασμό με δεδομένα ευπάθειας για τη δημιουργία ενός συστήματος σύστασης που προβλέπει μελλοντικές επιθέσεις.

Τα συστήματα σύστασης που παρουσιάζουν ιδιαίτερα προβλήματα στο απόρρητό τους είναι συνήθως δημόσια, διαθέσιμα απευθείας δηλαδή στο ευρύ κοινό. Αυτό συμβαίνει διότι συστήματα τέτοιου τύπου εμφανίζουν το πρόβλημα «επιθέσεις ένεσης προφίλ», που αναλύθηκε προηγουμένως. Οι εισβολείς μπορούν να εισαγάγουν προκατειλημμένα προφίλ για να προσπαθήσουν να προσαρμόσουν τη συμπεριφορά προτάσεων με τρόπο που να είναι επωφελής για τον εισβολέα [96]. Ωστόσο, τα συστήματα που απαιτούν ταυτοποίηση χρήστη με στοιχεία όπως (ημ. γέννησης, όνομα και επίθετο) συχνά ταυτοποιούν απευθείας ένα χρήστη. Οι προτιμήσεις του χρήστη θα μπορούσαν στη συνέχεια να χρησιμοποιηθούν για την εκ νέου αναγνώρισή του σε άλλο σύστημα. Για παράδειγμα, μια εταιρεία όπως το Netflix θα μπορούσε να αποθηκεύσει τις προτιμήσεις ορισμένων χρηστών και ξαφνικά να τις δει αποθηκευμένες σε μια άλλη ανταγωνιστική εταιρεία [97].

2.6 Νευρωνικά Δίκτυα

Πολλά συστήματα συστάσεων έχουν ως σημείο αναφοράς τα νευρωνικά δίκτυα. Για παράδειγμα, ένα συστάσεων μεγάλης κλίμακας αναπτύχθηκε στο Pinterest. Το σύστημα αυτό αποτελείται από τον αλγόριθμο Graph Convolutional Network (GCN) PinSage, ο οποίος συνδυάζει τυχαίους περιπάτους και συσπειρώσεις γράφων για τη δημιουργία ενσωματώσεων κόμβων, ουσιαστικά στοιχείων που ενσωματώνουν τόσο τη δομή του γράφου όσο και τα χαρακτηριστικά των κόμβων [98]. Μάλιστα, αυτή είναι μεγαλύτερη εφαρμογή ενσωματώσεων σε γράφους ως σήμερα και ανοίγει το δρόμο για μια νέα γενιά συστήματα συστάσεων που εφαρμόζονται σε ιστοσελίδες βασισμένες σε συνελικτικές αρχιτεκτονικές γραφημάτων.

Ένα άλλο σύστημα, που αναπτύχθηκε το 2016 στο Πανεπιστήμιο του Σιάν στην Κίνα, χρησιμοποιεί έναν νέο αλγόριθμο συστάσεων με αναγνώριση ετικετών. Στον προτεινόμενο αλγόριθμο, ένα μοντέλο νευρωνικού δικτύου, δηλαδή ένας αυτόματος κωδικοποιητής, χρησιμοποιείται για να ανακαλύψει τα σε βάθος χαρακτηριστικά του χώρου ετικετών. Αντί για τα ακατέργαστα δεδομένα χρησιμοποιούνται τα εξαγόμενα χαρακτηριστικά τα οποία αποκτούν χαρακτήρα πιο αφηρημένο και αντιπροσωπευτικό. Τα χαρακτηριστικά αυτά χρησιμοποιούνται για την παραγωγή συστάσεων, αξιοποιώντας την τεχνική φιλτραρίσματος με βάση το περιεχόμενο [99].

Την επόμενη χρονιά, η εκπαίδευση των νευρωνικών δικτύων έγινε προκειμένου να γίνει βελτίωση της ποιότητας των πολυκριτηριακών συστημάτων σύστασης. Συγκεκριμένα, το νευρωνικό δίκτυο εκπαιδεύτηκε χρησιμοποιώντας προσομοιωμένους αλγόριθμους και ολοκληρώθηκε με δύο μονοκριτηριακά συστήματα σύστασης. Το παρόν μοντέλο προτάθηκε

για να προσδιορίσει τον αντίκτυπο του εκπαιδευμένου νευρωνικού δικτύου για τη δημιουργία του προγνωστικού μοντέλου για την εκτίμηση των προτιμήσεων των χρηστών σε αντικείμενα που δεν έχουν ακόμη βαθμολογηθεί από αυτούς, προσπαθώντας να περιορίσει το πρόβλημα της ψυχρής εκκίνησης που μειώνει σημαντικά την ποιότητα των παρεχόμενων συστάσεων [100]. Το μοντέλο που προτάθηκε, ξεπέρασε κατά πολύ τις υφιστάμενες τεχνικές.

Ακολούθως, φέτος μελετήθηκε από τους Tan (et al) το πρόβλημα του βαθύ κατακερματισμού χωρίς επίβλεψη με γραφήματα νευρωνικών δικτύων, για παραγωγή συστάσεων. Προτάθηκε ένα νέο πλαίσιο HashGNN, το οποίο χρησιμοποιεί συναρτήσεις κατακερματισμού βαθιάς μάθησης και αναπαραστάσεις γραφημάτων από άκρο σε άκρο [101]. Η προτεινόμενη μέθοδος είναι ευέλικτη καθώς μπορεί να χρησιμοποιηθεί για επέκταση των ήδη υπαρχόντων νευρωνικών δικτύων. Όλη η αρχιτεκτονική είναι εκπαιδεύεται συνεχώς βελτιστοποιώντας δύο προβλήματα, δηλ. απώλεια ανακατασκευής σε ανακατασκευή των παρατηρούμενων συνδέσμων και κατάταξης διατήρησης απώλειας για να διατηρηθεί η σχετική κατάταξη ομοιότητας του κατακερματισμού.

3

Συστήματα συστάσεων ξενοδοχειακών επιχειρήσεων

3.1 Συστάσεις σε διαδικτυακές υπηρεσίες για εύρεση καταλύματος

Ο Παγκόσμιος Οργανισμός Τουρισμού (2006) προβλέπει ότι έως το 2020, οι αφίξεις τουριστών γύρω από το θα αυξηθούν πάνω από 200%. Επειδή ο τουρισμός είναι μια υπηρεσία που στηρίζεται σε μια σωρεία πληροφοριών, είναι σχεδόν απαραίτητη η εφαρμογή της τεχνολογίας πληροφοριών (IT) για την υποστήριξη του τουρισμού και των τουριστών. [102]

Σήμερα, οι ταξιδιώτες δεν βασίζονται πλέον στα ταξιδιωτικά γραφεία, αλλά αναζητούν οι ίδιοι πληροφορίες και δημιουργούν τα ταξίδια τους σύμφωνα με τις προτιμήσεις τους. Οι χρήστες πρέπει να επιλέξουν από μια πληθώρα συστημάτων σύστασης για ταξίδια και τα συστήματα σύστασης μπορεί να χρησιμοποιηθούν ως πρακτικά εργαλεία για να ξεπεραστεί η αναπόφευκτη υπερφόρτωση πληροφοριών. Η σύσταση ξενοδοχείων και άλλων σχετικών ταξιδιωτικών αντικειμένων εξακολουθεί να είναι μια δύσκολη εργασία καθώς η υπηρεσία του τουρισμού είναι ένας πολύ περίπλοκος τομέας που επηρεάζεται από πολλαπλούς παράγοντες (προτιμήσεις πελατών, προσβασιμότητα, οικονομική κατάσταση). Ο προγραμματισμός ενός ταξιδιού περιλαμβάνει συνήθως την αναζήτηση ενός συνόλου ή πακέτων προϊόντων που είναι διασυνδεδεμένα (π.χ. μέσα μεταφοράς, διαμονή, αξιοθέατα), με περιορισμένη διαθεσιμότητα ή με μερικές παραμέτρους που έχουν σημαντικό αντίκτυπο (π.χ. χρόνος, τοποθεσία, κοινωνικό περιεχόμενο). [103]

Έστω ότι ένας άνθρωπος έχει την επιθυμία να ταξιδέψει σε ένα μέρος για καλοκαιρινές διακοπές. Μπορεί να έχει αποφασίσει ότι θέλει μια συγκεκριμένη χώρα π.χ. Ελλάδα. Η Ελλάδα

όμως έχει διαφορετικά μέρη με διαφορετικές παροχές ανάλογα με την τοποθεσία τους. Έτσι ο πιθανός τουρίστας καλείται να λάβει μια απόφαση λαμβάνοντας υπόψη μια σωρεία πληροφοριών. Για παράδειγμα, μπορεί να του αρέσουν περισσότερο τα γαλάζια νερά του Ιονίου, εντούτοις στο νησί που επέλεξε να μην υπάρχουν ξενοδοχεία με τα απαιτούμενα φίλτρα επιλογής, τη συγκεκριμένη ημερομηνία και στο εύρος τιμής που μπορεί να έχει ορίσει. Οπότε γίνεται εύλογα αντιληπτό πως η διαδικασία λήψης απόφασης, γίνεται ιδιαίτερα περίπλοκη, υποκειμενική και πολυπαραγοντική. Μπορεί η επιλογή του ξενοδοχείου και μέρους να γίνει αποκλειστικά επιλέγοντας υποκειμενικούς παράγοντες (θέα ξενοδοχείου) που για έναν άλλο τουρίστα να μην έχει καμία σημασία. Είναι εύκολα αντιληπτό, το πόσες πολλές πληροφορίες καλείται να διαχειριστεί ένα σύστημα σύστασης. Όλες αυτές οι πληροφορίες, πρέπει να κατηγοριοποιούνται, ώστε να κάνουν ακόμη ευκολότερη μια πρόταση ενός συστήματος σύστασης. Μια ενδιαφέρουσα πρώιμη κατηγοριοποίηση πληροφοριών πραγματοποιείται σε τρεις διαστάσεις: το πλαίσιο της πρότασης, την ομάδα χρηστών που υποτίθεται ότι θα επωφεληθούν από ένα προτεινόμενο στοιχείο, και τις χρονικές σχέσεις μεταξύ των προτεινόμενων ειδών [104]. Στην πραγματικότητα, όλες αυτές οι πτυχές μπορούν να θεωρηθούν συμπληρωματικές και αλληλένδετες για την παραγωγή σύστασης.

Όσον αφορά την σύσταση τουριστικών προϊόντων, έχουν γίνει στο παρελθόν πολλές έρευνες για τη δημιουργία συστημάτων σύστασης. Τα τελευταία δέκα χρόνια παρατηρήθηκε μια εκρηκτική αύξηση χρήσης κινητής τεχνολογίας μεταξύ των τουριστών. Επομένως, ο τα συστήματα ηλεκτρονικού τουρισμού παρέχουν μια καλή ευκαιρία για υπηρεσίες που βοηθούν τους επισκέπτες προσφέροντας προτάσεις με βάση τις προτιμήσεις τους. Ήδη από το 2008, οι Castillo et al ανέπτυξαν ένα εργαλείο που δημιουργεί δυναμικά μοντέλα χρηστών, ενώ στη συνέχεια καθορίζει λίστες δραστηριοτήτων που μπορούν να προσφέρουν περισσότερη σχετικότητα σε έναν χρήστη, δεδομένης της προηγούμενης γνώσης του συστήματος[105]. Η ανάπτυξη αυτή βασίστηκε σε ανάπτυξη οντολογιών. Την ίδια χρονιά οι Zanker et al, ανέπτυξαν μια προσέγγιση που συνδυάζει δεδομένα περιήγησης ιστού με σχόλια χρηστών που συλλέγονται σε ένα διαδραστικό σύστημα συμβουλευτικής ταξιδίων[106].

Δύο πλατφόρμες [107] οι Tripmatcher και MetaPrint, πρότειναν συστήματα τα οποία μιμούνται τη διαδραστικότητα που παρατηρείται στα παραδοσιακά συστήματα σύστασης με ταξιδιωτικούς πράκτορες όταν οι χρήστες αναζητούσαν συμβουλές για πιθανούς προορισμούς διακοπών. Από τεχνική άποψη, χρησιμοποιούν μια προσέγγιση φιλτραρίσματος βάσει περιεχομένου, στο οποίο ο χρήστης εκφράζει ανάγκες και περιορισμούς (γνωρίσματα), εκφρασμένα σε μια συγκεκριμένη γλώσσα. Στη συνέχεια, το σύστημα ταιριάζει τις προτιμήσεις κάθε χρήστη με αντικείμενα σε έναν κατάλογο που περιλαμβάνει όλους τους πιθανούς προορισμούς (περιγράφονται με την ίδια γλώσσα που αναφέρθηκε πριν). Το VacationCoach [107] συγκεκριμένα χρησιμοποιεί μια κατηγορία για το προφίλ των χρηστών ζητώντας ρητά

από τον χρήστη να ταξινομήσει τον εαυτό του σε ένα προφίλ (για παράδειγμα, ως ένα «πλάσμα που λατρεύει την κουλτούρα») γεγονός που προκαλεί την έμμεση σκιαγράφηση του χρήστη για πληροφορίες που δεν παρέχει. Ο χρήστης μπορεί ακόμη και να εισάγει ακριβείς πληροφορίες προφίλ συμπληρώνοντας την κατάλληλη φόρμα.

3.2 Τεχνολογίες που εμφανίζονται στα υφιστάμενα συστήματα σύστασης

Τα συστήματα σύστασης που έχουν αναπτυχθεί βασίζονται κυρίως στις εξής τεχνολογίες [108]:

- Ευφυείς αυτόνομοι πράκτορες μπορούν να αναλύσουν τη συμπεριφορά ενός χρήστη, να σκιαγραφήσουν αυτόματα το προφίλ του και παρέχουν συστάσεις ανάλογα με το τρέχον περιεχόμενο που εκείνος επισκέπτεται.
- Ορισμένα συστήματα υπερβαίνουν την προσφορά μιας λίστας προτεινόμενων τουριστικών αξιοθέατων και χρησιμοποιούν αυτοματοποιημένους σχεδιαστές για να προγραμματίσουν μια ολοκληρωμένη εμπειρία περιήγησης.

Άλλες προσεγγίσεις λαμβάνουν υπόψη τις ώρες ανοίγματος και κλεισίματος των αξιοθέατων καθώς και τον χρόνο που απαιτείται ώστε ο χρήστης να διανύσει την απόσταση από ένα σημείο ενδιαφέροντος σε ένα άλλο, προσφέροντας έτσι ένα λεπτομερές χρονοδιάγραμμα επίσκεψης. Για την πραγματοποίηση μιας επίσκεψης στο μουσείο της Ακρόπολης ενδέχεται να μην ακολουθηθεί μια μέρα το κανονικό πρόγραμμα, λόγω κάποιου απρόβλεπτου παράγοντα (π.χ. κακοκαιρίας). Ωστόσο, αυτός είναι ένας πολύ περίπλοκος προγραμματισμός γεγονός που καθιστά δύσκολο να βρεθεί η βέλτιστη λύση. Ορισμένα συστήματα επιλύουν αυτήν την πολυπλοκότητα με τη χρήση τεχνικών βελτιστοποίησης τεχνητής νοημοσύνης, όπως μετα-ευρετικές επαναληπτικές μέθοδοι ή τεχνικές «αποικίας μυρμηγκιών».

Για την ταξινόμηση των τουριστών με παρόμοια γούστα ή παρόμοια χαρακτηριστικά, μπορούν να χρησιμοποιηθούν αυτοματοποιημένοι αλγόριθμοι συσταδοποίησης. Προκειμένου να επιλυθούν συχνά εμφανιζόμενα προβλήματα που σχετίζονται με την υποκειμενικότητα των προσωπικών επιλογών καθενός από τους χρήστες όπως τα Bayesian δίκτυα. Επίσης, προκειμένου να σκιαγραφηθούν αποτελεσματικότερα οι προτιμήσεις του χρήστη χρησιμοποιείται η εννοιολογική αναπαράσταση μέσω οντολογιών.

3.3 Αρχιτεκτονική

Η αρχιτεκτονική που παρουσιάζεται αποτελείται ουσιαστικά από πέντε συστατικά στοιχεία. Κάθε ένα από αυτά περιγράφονται ως εξής [109]:

- Το στοιχείο προσδιορισμού τύπου ενεργού χρήστη που εκχωρεί έναν ενεργό χρήστη σε μια ομάδα με βάση τις σιωπηρές και ρητές πληροφορίες του ιδίου και του συνόλου των τουριστικών αξιοθέατων.
- Ένα παρόμοιο στοιχείο αναγνώρισης χρήστη εξάγει ένα υποσύνολο χρηστών με βάση την ομοιότητα της κοινότητας των χρηστών ή των δημογραφικών χαρακτηριστικών τους.
- Το στοιχείο ανίχνευσης βαθμολογίας ανιχνεύει τις αξιολογήσεις για ένα αξιοθέατο που παρουσιάζονται από χρήστες με χαμηλή γνώση ή από "ψεύτικους" χρήστες.
- Το στοιχείο εξαγωγής λίστας στοιχείων υπολογίζει το ρυθμό προτίμησης για ένα αξιοθέατο με βάση την ομάδα χρηστών στην οποία ανήκει ο ενεργός χρήστης.
- Το στοιχείο παροχής λίστας προτάσεων, με βάση τις αξιολογήσεις του ενεργού χρήστη, δημιουργεί μια λίστα προορισμών.

Για κάθε ενεργό χρήστη που επιθυμεί να πραγματοποιήσει ένα ταξίδι τη δεδομένη στιγμή, το σύστημα παρέχει μια λίστα με προτεινόμενα αξιοθέατα που βασίζονται σε αξιόπιστους και παρόμοιους χρήστες. Τα περισσότερα συστήματα σύστασης που μελετήθηκαν στη βιβλιογραφία δεν παρείχαν την σύσταση τμηματοποιημένη σε μικρότερα πλαίσια σύστασης αλλά είναι κάτι που μας οδηγεί σε ασφαλέστερα συμπεράσματα ενώ ταυτόχρονα δείχνει και μεγαλύτερη δυνατότητα επαναχρησιμοποίησης ή συνεργασίας με άλλα συστήματα σύστασης.

3.3.1 Αρχιτεκτονική βασισμένη σε τεχνικές συνεργατικού φιλτραρίσματος

Το σύστημα αναλύει μηνύματα κειμένου που αποστέλλονται από χρήστες (πελάτες και ταξιδιωτικούς πράκτορες) όταν αυτοί αλληλεπιδρούν μεταξύ τους σε μια ιδιωτική συνομιλία Ιστού και με αυτόν τον τρόπο προσδιορίζονται οι περιοχές ενδιαφέροντος του πελάτη σύμφωνα με μια τουριστική οντολογία προκαθορισμένη από το σύστημα [110]. Για την αναπαράσταση των συστατικών στοιχείων του συστήματος αυτού χρησιμοποιούνται οντολογίες. Μια οντολογία είναι μια περιγραφή των «πραγμάτων» που υπάρχουν ή μπορεί να υπάρξουν σε έναν τομέα. Κάθε μήνυμα συγκρίνεται κάθε φορά την ίδια χρονική στιγμή με όλα τα θέματα που έχουν οριστεί εντός της οντολογίας. Τα θέματα που προσδιορίζονται στα μηνύματα αντιπροσωπεύουν τις προτιμήσεις του πελάτη και προωθούνται στην ενότητα σύστασης, ενώ παράλληλα χρησιμοποιούνται ως κριτήρια για την επιλογή ειδών αξιοθέατων και πόλεων από τη βάση δεδομένων. Η μέθοδος εξόρυξης κειμένου συγκρίνει το διάνυσμα που αντιπροσωπεύει το κείμενο ενός μηνύματος με πιθανά διανύσματα που αντιπροσωπεύουν τα θέματα της κάθε οντολογίας. Το συνολικό άθροισμα των στοιχείων αυτών, που περιορίζονται σε 1, είναι ο βαθμός σχέσης μεταξύ του κειμένου και του θέματος, που σημαίνει την πιθανότητα παρουσίας του σχετικού θέματος στο κείμενο ή ότι το συγκεκριμένο κείμενο εμπεριέχεται στο θέμα σε σχετικά μεγάλη και σημαντική έκταση. Η απόφαση σχετικά με το αν ένα θέμα είναι παρόν ή

όχι εξαρτάται το από το κατώφλι που χρησιμοποιείται με αποτέλεσμα να αποκλείει τις επιλογές που συγκεντρώνουν λιγότερους βαθμούς από αυτό.

3.4 Συνήθη προβλήματα που εμφανίζονται στα τουριστικά συστήματα σύστασης

Τα τουριστικά συστήματα σύστασης αποκτούν ολοένα και μεγαλύτερη αξία διότι εντάσσονται στο πλαίσιο των έξυπνων πόλεων. Με αυτόν τον τρόπο, οι λύσεις έξυπνων τουριστικών προορισμών μπορούν να βελτιωθούν μέσω της αυτοματοποίησης της διαδικασίας για την ενίσχυση των τουριστικών οργανισμών και της ανταγωνιστικότητας μεταξύ των προορισμών [111]. Τα πιο συνηθισμένα προβλήματα που εμφανίζονται στα τουριστικά συστήματα σύστασης είναι τα ακόλουθα:

Φυσικοί περιορισμοί: Η ικανότητα του ανθρώπινου εγκεφάλου να επεξεργάζεται μεγάλους όγκους πληροφοριών, κάνει τη διαδικασία λήψης αποφάσεων μια πολύ δύσκολη εργασία που μπορεί να οδηγήσει σε υπερφόρτωση πληροφοριών, μη βελτιστοποιημένη διαχείριση πόρων (π.χ. σπατάλη χρόνου, χρήματος και / ή προσπάθειας), λανθασμένες διαδικασίες λήψης αποφάσεων, σωματική εξάντληση και αγωνία ή και τον συνδυασμό τους [112].

Δυσκολία λόγω φύσης της απόφασης: ο τουρισμός είναι μια «συναισθηματική» εμπειρία, η οποία συνήθως είναι δύσκολο να περιγραφεί με λογικούς όρους · ιδιαίτερα σε πρώιμα στάδια μιας διαδικασίας απόφασης ταξιδιού, οι χρήστες δεν είναι σε θέση να εκφράσουν ρητά τις προτιμήσεις τους και συχνά δεν μπορούν να έχουν πλήρη γνώση όλων των δεδομένων, επομένως έχουν δυσκολίες στη χρήση της σωστής ορολογίας. Στην πραγματικότητα, ο χρήστης πρέπει να κάνει επιλογές πριν εξερευνήσει όλες τις δυνατότητες, διότι διαφορετικά δεν θα ήταν σε θέση να σκεφτεί την κατάλληλη φράση αναζήτησης, η οποία θα τις αντικατοπτρίζει. [113]

Έλλειψη προφίλ χρηστών: Στον τομέα του τουρισμού, ο αριθμός των αξιολογήσεων ενός χρήστη είναι συνήθως μικρότερος από ότι σε άλλους τομείς (π.χ. ταινιών ή μουσικής), γεγονός που θέτει άλλα προβλήματα κατά την εφαρμογή τεχνικών συστάσεων. Κατά συνέπεια, τα προφίλ χρηστών είναι λιγότερο ακριβή και λεπτομερή. Αυτό συμβαίνει διότι δεν είναι εύκολο ένας χρήστης να αξιολογήσει την εμπειρία του αμέσως μετά το ταξίδι του (όπως για παράδειγμα σε μια αντίστοιχη περίπτωση ταινίας), διότι εκείνη τη στιγμή πιθανότατα ταξιδεύει στον τόπο κατοικίας του ή σε κάποιο άλλο ξενοδοχείο.

Δυσκολία εφαρμογής αλγορίθμων: Μία ακόμη δυσκολία που εμφανίζουν τα συστήματα σύστασης γενικότερα και συνδυάζεται ιδιαίτερα με την προαναφερθείσα είναι η μη ύπαρξη προηγούμενης πληροφορίας. Το φιλτράρισμα με βάση το χρήστη εμφανίζει συχνά το πρόβλημα της ψυχρής εκκίνησης, πρόβλημα που όπως είχε αναλυθεί και νωρίτερα εμφανίζεται

όταν έχουμε έναν νέο χρήστη που δεν έχει δώσει βαθμολογίες στο σύστημα ή όταν έχουμε ένα νέο στοιχείο που δεν έχει βαθμολογηθεί ακόμη από τους χρήστες [cold-start].

3.5 Προκλήσεις

Στις μέρες μας, ο συντριπτικός αριθμός εικόνων που ανέβηκαν στο διαδίκτυο σε συνδυασμό με τη συνεχώς αυξανόμενη δημοτικότητα των πλατφορμών κοινής χρήσης μέσων έχει οδηγήσει σε μια αδιαμφισβήτητη ανάγκη για αποτελεσματικά συστήματα σύστασης[114]. Υπάρχουν πολλές προκλήσεις στον τομέα των συστημάτων προτάσεων του τουρισμού, οι οποίες αναφέρονται παρακάτω[115]:

1. Εξειδικευμένα συστήματα συστάσεων:
   Τα συστήματα προτάσεων δημιουργούν συστάσεις αυτόματα χωρίς να υπακούν σε κάποιο σαφές ερώτημα. Το σύστημα μπορεί να αποκτήσει μεγαλύτερο δείκτη αφοσίωσης αν γίνουν αντιληπτά και τα αιτήματα και οι απαιτήσεις του χρήστη. Επομένως, οι προβλέψεις ενός τέτοιου συστήματος γίνονται όχι μόνο με βάση τις ρητές ανάγκες του χρήστη, αλλά και με τη πρόβλεψη εισαγωγής νέων προϊόντων.
2. Αύξηση της ικανοποίησης του χρήστη:
   Ένα καλά σχεδιασμένο σύστημα συστάσεων μπορεί να βελτιώσει την εμπειρία του χρήστη όταν χρησιμοποιεί τον ιστότοπο ή την εφαρμογή του. Ο χρήστης θα βρει τις συστάσεις ενδιαφέρουσες, σχετικές και με μια σωστά σχεδιασμένη αλληλεπίδραση θα απολαύσει επίσης τη χρήση του συστήματος, αυξάνοντας την ικανοποίησή του. Ο συνδυασμός αποτελεσματικών, ακριβών συστάσεων και μιας χρήσιμης διεπαφής θα αυξήσει την υποκειμενική αξιολόγηση του συστήματος από τον χρήστη. Αυτό, με τη σειρά του, θα αυξήσει τη χρήση και τη ακρίβεια των συστάσεων αλλά και την πιθανότητα να γίνουν αποδεκτές οι προσφερόμενες προτάσεις [116].
3. Μέθοδοι αξιολόγησης:
   Οι τρέχουσες μέθοδοι αξιολόγησης για συστήματα σύστασης τις περισσότερες φορές χρησιμοποιούν τα σχόλια του κάθε χρήστη. Οι πιο δημοφιλείς τεχνικές που χρησιμοποιούνται είναι η απόκλιση ρίζας μέσου τετραγώνου (RMSE) και του μέσου τετραγωνικού σφάλματος (MAE), τεχνικές οι οποίες, όπως προαναφέρθηκε βασίζονται σε σχόλια χρηστών. Τα συστήματα προτάσεων στον τομέα του τουρισμού πρέπει να είναι σε θέση να μετρήσουν το επίπεδο ικανοποίησης των χρηστών μετρώντας τη συναισθηματική τους κατάσταση με έναν ελάχιστα-παρεμβατικό τρόπο προκειμένου να διατηρήσουν την αντικειμενικότητά τους. Μερικές από τις πιθανές μεθόδους περιλαμβάνουν ανάλυση των μέσων που ο χρήστης μοιράζεται στα κοινωνικά δίκτυα, φωτογραφίες, σχόλια και αντιδράσεις όπως αρέσει / δεν μου αρέσει [117].

4. Συστήματα συστάσεων σε μια ομάδα:

   Ιδιαίτερα σημαντικά θέματα εμφανίζονται όταν το σύνολο των μελών μιας ομάδας καθορίζουν ρητά τις προτιμήσεις τους και όπου δεν είναι σε θέση να αλληλεπιδρούν πρόσωπο με πρόσωπο. Τα ζητήματα αφορούν τόσο τον σχεδιασμό κατάλληλων μεθόδων και συγκέντρωσης προτιμήσεων και τρόπους ευαισθητοποίησης των μελών για τις προτιμήσεις και τα κίνητρα του άλλου, όπως η χρήση αντιπροσώπων για όλα τα μέλη της ομάδας[118]. Βέβαια, τα συστήματα συστάσεων έχουν επικεντρωθεί πολύ περισσότερο σε γκρουπ επισκεπτών, παρά σε μεμονωμένα μέλη [119].

5. Μεταβλητότητα δεδομένων:

   Ακόμη και οι πρώτες προσπάθειες ανάπτυξης συστημάτων σύστασης ανεξάρτητα από τον τομέα που αυτές εφαρμόστηκαν, ήρθαν αντιμέτωπες με ένα πολύ σημαντικό πρόβλημα το οποίο εντοπίζεται στην αλλαγή των προτιμήσεων των χρηστών, πρόβλημα που προτείνεται πλέον να λαμβάνεται υπ' όψιν κατά τον σχεδιασμό όλων των συστημάτων σύστασης. Οι επιλογές του χρήστη είναι υποκειμενικές και εξαρτώνται από τις μεταβλητές τους προτιμήσεις, τη συναισθηματική τους κατάσταση, τα ερεθίσματα που δέχονται από το περιβάλλον τους.

6. Τα συστήματα σύστασης που παράγουν δεδομένα σε συνεδρίες, παρέχουν χειροπιαστά αποδεικτικά στοιχεία που αποτυπώνουν τη χρονική δυναμική των προτιμήσεων του χρήστη βελτιώνοντας την απόδοση της πρότασης, συνεπώς και την αποτελεσματικότητα των συστάσεων[120]. Ένας σωστός διαχειριστής προορισμού πρέπει να γνωρίζει τις λεπτομέρειες συγκεκριμένων τοποθεσιών που επισκέπτονται οι τουρίστες, τι προσελκύει τους τουρίστες σε κάθε τοποθεσία, τις προσωπικές σκέψεις /εμπειρίες των τουριστών αλλά και τις μελλοντικές προθέσεις τους σε μεταγενέστερα ταξίδια [121]. Σε γενικές γραμμές, οι τρέχουσες προσεγγίσεις δεν είναι σε θέση να αντιμετωπίσουν αυτά τα ζητήματα με έναν ολοκληρωμένο τρόπο διότι επικεντρώνονται στην εύρεση απαντήσεων σε προκαθορισμένες ερωτήσεις. Σε αυτό προστίθεται και το γεγονός πως τα δεδομένα δεν αναλύονται με βάση τη σημασιολογική σημασία τους, με αποτέλεσμα η πολυτροπικότητά τους να αγνοείται [122]. Το πολυδιάστατο μοντέλο δεν περιορίζεται μόνο στην αλληλεπίδραση με τουριστικές υπηρεσίες, αλλά λαμβάνει επίσης πληροφορίες με βάση τα συμφραζόμενα στη σύσταση προϊόντων. Έτσι, το πολυδιάστατο μοντέλο μπορεί να ξεπεράσει τους περιορισμούς του δισδιάστατου μοντέλου[123].

4

Ορισμός προβλήματος

4.1 Εισαγωγή

Όπως αναλύθηκε στην προηγούμενη ενότητα, ο τουρισμός αποτελεί τη μεγαλύτερη πηγή εσόδων για πολλές χώρες. Ωστόσο, ένα πρόβλημα που αφορά την κατάλληλη επιλογή ενός καταλύματος από τους χρήστες δέχεται πολλαπλές παραμέτρους διαφορετικές κάθε φορά. Αυτό συμβαίνει διότι η επιλογή τουριστικού μέρους και διαμονής διαφέρει από άνθρωπο σε άνθρωπο.

Είναι εύληπτη η αντικειμενική χρησιμότητα επίλυσης ενός τέτοιου προβλήματος τόσο ως προς τους χρήστες που δεν θα καταναλώνουν πόρους σε λανθασμένες επιλογές, βρίσκοντας ευκολότερα αυτό που τους ταιριάζει, όσο και για τις επιχειρήσεις οι οποίες δύναται ευκολότερα να βελτιώσουν τις παροχές τους.

Για τη διερεύνηση ενός τέτοιου προβλήματος, είναι απαραίτητη η χρήση δεδομένων από μια αξιόπιστη πλατφόρμα την οποία επισκέπτεται ένας μεγάλος αριθμός χρηστών. Μια τέτοια πλατφόρμα που δέχεται μεγάλο αριθμό χρηστών είναι η πλατφόρμα trivago. Η trivago είναι μια μηχανή αναζήτησης τιμών και πληροφοριών για ξενοδοχεία. Η ιστοσελίδα συγκρίνει τις τιμές 700.000 ξενοδοχείων, προερχόμενων από περισσότερες από 200 ιστοσελίδες ξενοδοχειακών κρατήσεων, όπως για παράδειγμα την booking.com και η Hotels.com. Τα κεντρικά γραφεία της εταιρείας βρίσκονται στο Ντίσελντορφ της Γερμανίας και η ιστοσελίδα φιλοξενεί 18 εκατομμύρια χρήστες κάθε μήνα, στις 52 διεθνείς πλατφόρμες που παρέχει Η πλατφόρμα αναπτύχθηκε από 3 φίλους που γνωρίστηκαν στο πανεπιστήμιο: Rolf Schrömgens, Peter Vinnemeier και Stephan Stubner. Όπως οι περισσότερες νεοσύστατες επιχειρήσεις, η πρώτη διεπαφή του ιστότοπου αναπτύχθηκε σε ένα γκαράζ [124]. Η trivago συγκρίνει σε πραγματικό χρόνο τιμές ξενοδοχείων για τον επιλεγμένο. από τον χρήστη. προορισμό, αναζητώντας σε πάνω από 200 διαφορετικές ιστοσελίδες ξενοδοχειακών κρατήσεων πληροφορίες όπως κριτικές, φωτογραφίες, περιγραφές και τιμές. Ο κάθε επισκέπτης μπορεί να

βρει το ιδανικό για τις ανάγκες του ξενοδοχείο, βασισμένος στη σύγκριση τιμών αλλά και στις εντυπώσεις άλλων χρηστών που το έχουν ήδη επισκεφτεί. Η σύγκριση ξενοδοχειακών τιμών παρέχει πληροφορίες προερχόμενες από διάφορες πηγές, όπως ιστοσελίδες ξενοδοχειακών κρατήσεων, χρήστες και ξενοδόχους. Αυτές οι πληροφορίες είναι έτσι δομημένες, ώστε να παρέχουν στον χρήστη την καλύτερη δυνατή εικόνα κάθε ξενοδοχείου [125].

Ο πλουραλισμός τον δεδομένων που προέρχεται από πολλές διαφορετικές πηγές, η ευκολία επιλογής των επιθυμητών ξενοδοχείων μέσω χρήσης κατάλληλων φίλτρων, ο μεγάλος αριθμός τουριστών από διαφορετικός χώρες δίνουν στον ιστότοπο μια πληθώρα δεδομένων, που θα μπορούσαν να οργανωθούν ευκολότερα σε πληροφορίες. Οι πληροφορίες αυτές μπορούν να οργανωθούν εύκολα και να οδηγήσουν σε συμπεράσματα για το θέμα της παρούσας διπλωματικής το πώς επιλέγουν ξενοδοχεία οι χρήστες. Αναγνωρίζοντας την αξία των αυτών των πληροφοριών, κινήθηκε και ο διαγωνισμός RecSys, ο οποίος θα αναλυθεί λεπτομερέστερα στην επόμενη ενότητα.

4.2 Συνέδριο RecSys

Το ACM Conference on Recommender Systems (RecSys) είναι το κορυφαίο διεθνές συνέδριο για την παρουσίαση νέων ερευνητικών αποτελεσμάτων, συστημάτων και τεχνικών στα συστήματα συστάσεων. Το συνέδριο λαμβάνει χώρα από το 2007 και κάθε χρόνο, σε μία διαφορετική πόλη. Έχουν πραγματοποιηθεί μέχρι στιγμής 15 συνέδρια, ενώ υποβολές δέχεται σήμερα το 16ο σε σειρά. Οι πόλεις, στις οποίες έχει διεξαχθεί μέχρι στιγμής το συνέδριο με χρονολογική σειρά από το 2007-2021 είναι οι ακόλουθες: Μινεσότα, Λουϊζιάνα, Νέα Υόρκη, Βαρκελώνη, Σικάγο, Δουβλίνο, Χονγκ Κονγκ, Σίλικον Βάλεϋ, Βιέννη, Βοστώνη, Κόμο, Βανκούβερ, Κοπεγχάγη, διαδικτυακά (λόγω της πανδημίας του κορωνοϊού) και Άμστερνταμ. Την φετινή χρονιά, θα πραγματοποιηθεί από τις 18-23 Σεπτεμβρίου του 2022 στο Σιάτλ, ΗΠΑ. Η RecSys συγκεντρώνει τόσο τις μεγαλύτερες διεθνείς ερευνητικές ομάδες που εργάζονται σε συστήματα συστάσεων, όσο και πολλές από τις κορυφαίες εταιρείες στον κόσμο που δραστηριοποιούνται στο ηλεκτρονικό επιχειρείν και σε συναφείς τομείς [126].

Το συνέδριο έχει ως στόχο να φέρει σε επαφή ερευνητές και επαγγελματίες από τον ακαδημαϊκό χώρο και τη βιομηχανία για να παρουσιάσουν τα πιο αποτελέσματά τους σε μια πρόκληση, ενώ παράλληλα να εντοπιστούν νέες τάσεις και προκλήσεις στην παροχή συστατικών συστάσεων. Εκτός από το κύριο τεχνικό κομμάτι, το πρόγραμμα RecSys περιλαμβάνει ομιλίες, σεμινάρια που καλύπτουν την τελευταία λέξη της τεχνολογίας σε αυτόν τον τομέα, εργαστηριακά προγράμματα, ένα βιομηχανικό κομμάτι.

Η γενικότερη φιλοσοφία του RecSys στηρίζεται στον πλουραλισμό, στη συμπερίληψη και στην ένταξη. Η διαφορετικότητα αυτή διαφαίνεται στο επιστημονικό υπόβαθρο, στις επαγγελματικές σχέσεις (π.χ. ισχυρή εκπροσώπηση τόσο από τον ακαδημαϊκό χώρο όσο και από τη βιομηχανία) και τους τομείς εφαρμογής των λύσεων που παρουσιάζονται. Οι συμμετέχοντες στο συνέδριο λαμβάνουν μέρος σε ποικίλες επιστημονικές συζητήσεις, ενώ ταυτόχρονα αλληλεπιδρούν σε μια ζωντανή κοινότητα. Σε αυτό το πλαίσιο, έχουν αναπτυχθεί οδηγίες βέλτιστων πρακτικών, κατά την προετοιμασία των εισηγήσεών τους και την αξιολόγησή τους από ομοτίμους [127].

4.2.1 RecSys Challenge 2019

Η λήψη του συνόλου των δεδομένων έγινε στα πλαίσια του RecSys Challenge 2019. Για την πρόκληση RecSys 2019, δόθηκαν πραγματικά δεδομένα από την trivago. Για πρώτη φορά, η πρόκληση RecSys ασχολήθηκε με μένα πραγματικό σενάριο από τον τομέα του τουρισμού. Ο τουρισμός είναι ένα ιδιαίτερα ενδιαφέρον και προκλητικό πεδίο για συστήματα συστάσεων λόγω πολλαπλά ενδιαφερόμενων μερών (επιχειρήσεις, ταξιδιώτες), ποικιλίας καταλυμάτων, δυναμικών κριτηρίων αναζήτησης και υπολογιστικών απαιτήσεων σχετικά με το πόσο γρήγορα πρέπει να παραδοθούν τα αποτελέσματα για μια καλή εμπειρία χρήστη. Οι συμμετέχοντες υπέβαλαν ενεργά λύσεις σε όλη την πρόκληση σύμφωνα με τις οδηγίες υποβολής που περιορίζονται σε 1 υποβολή κάθε 12 ώρες. Κατά μέσο όρο μια ομάδα υπέβαλε 5,3 λύσεις ενώ η πιο ενεργή ομάδα υπέβαλε 119 λύσεις [128]. Η πρόκληση που καλούνταν να λύσουν οι συμμετέχοντες αναλύεται περιγραφικά στην επόμενη ενότητα.

4.3 Επιλογή δεδομένων

Τα δεδομένα που επιλέχθηκαν δόθηκαν στο πλαίσιο του RecSys Challenge του 2019. Τα δεδομένα που παρέχονται αποτελούνται από ένα σετ εκπαίδευσης και δοκιμών καθώς και ένα σετ μεταδεδομένων για τα καταλύματα (ιδιότητες). Όλα τα αρχεία είναι ανοιχτά, προσβάσιμα και διαθέσιμα στον ιστότοπο του Kaggle.com [129]. Τα αρχεία μπορεί να τα κατεβάσει οποιοσδήποτε χρήστης, σε συμπιεσμένη μορφή (.zip), συνολικού μεγέθους 263, 77,3 και 22,5 MB. Κάθε ένα από αυτά τα συμπιεσμένα αρχεία, περιλαμβάνει ένα αρχείο διαχωρισμένων τιμών (.csv) μεγέθους 1,95 GB, 510 MB, 245 MB αντίστοιχα. Συνολικά, σε αυτόν τον ιστότοπο, το σύνολο των δεδομένων έχει μέχρι στιγμής 3836 προβολές, ενώ το κατέβασαν 462 διαφορετικοί χρήστες.

Ένας από τους λόγους που επιλέχθηκε το συγκεκριμένο σύνολο δεδομένων είναι η μορφή τους (τα αρχεία .csv μπορούν να χρησιμοποιηθούν από πολλαπλά περιβάλλοντα και να μετασχηματισθούν ευκολότερα) , η ευκολία πρόσβασης, η φύση των δεδομένων (η χρήση

τουριστικών δεδομένων έχει αναμενόμενη χρησιμότητα, καθώς όπως αναλύθηκε παραπάνω αποτελεί ένα μέρος σημαντικών εσόδων μιας χώρας, αυξάνοντας των ΑΕΠ ιδιαίτερα σε τουριστικές χώρες όπως η Ελλάδα).

4.4 Περιορισμοί

Προκειμένου τα δεδομένα να έρθουν σε μία μορφή ώστε να μπορούν να αναπαρασταθούν εύκολα από έναν γράφο, επεξεργάστηκαν χρησιμοποιώντας συγκεκριμένο τεχνολογικό πλαίσιο, ανάλογα με το τι χρειαζόταν κάθε φορά, ώστε τα δεδομένα να έρθουν σε μια επιθυμητή μορφή, η οποία ήταν δυνατό να αποτυπώσει το ως προς επίλυση πρόβλημα σε γράφο. Το τεχνικό πλαίσιο στο οποίο επεξεργάστηκαν τα δεδομένα, αναλύεται στα επόμενο υποκεφάλαιο αυτής της ενότητας.

Καθώς το μέγεθος των δεδομένων εκπαίδευσης, ήταν ιδιαίτερα μεγάλο και οδηγούσε σε μεγάλες καθυστερήσεις, αποφασίστηκε από το σύνολο δεδομένων να μείνουν μόνο αυτά που αφορούν την Ελλάδα. Αυτό καθορίζεται μέσω μιας μεταβλητής που υπάρχει ήδη στο σύνολο δεδομένων και αναλύεται παρακάτω και επιτρέπει στον χρήστη να προεπιλέξει τη γεωγραφική τοποθεσία που θα επιθυμούσε να πραγματοποιήσει την αναζήτησή του.

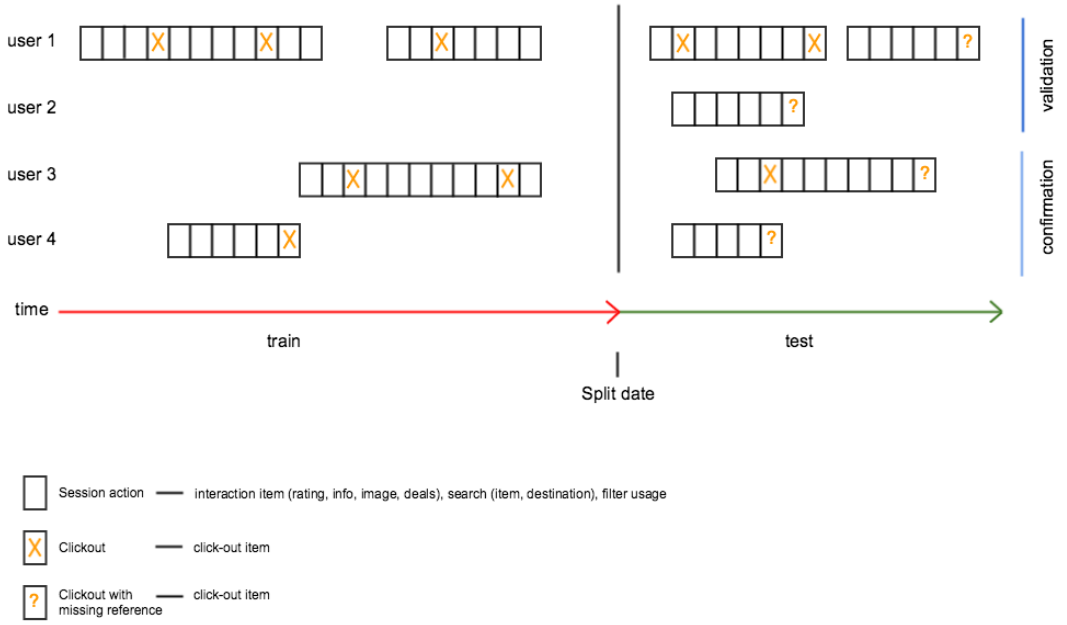
Σε ανάλογο πλαίσιο, είχε αποφασιστεί αρχικά να εντοπιστούν εκείνες οι ιδιότητες των ξενοδοχείων που επηρεάζουν την επιλογή του χρήστη περισσότερο από τις υπόλοιπες όπως για παράδειγμα: ξενοδοχείο ενός ή πολλαπλών αστέρων, δυνατότητα χρήσης θέσης στάθμευσης, ξενοδοχεία με πρόσβαση ΑμεΑ, ξενοδοχεία που δέχονται παιδιά. Μια τέτοια παραδοχή όμως είναι αυστηρώς υποκειμενική και επαφίεται στην υποκειμενική παραδοχή που θα επιθυμούνταν από εμάς. Αυτό αποτελεί και ένα από τα προβλήματα που εμφανίζεται στα συστήματα σύστασης: οι επιλογές του χρήστη στηρίζονται στην ψυχολογική του κατάσταση και οι επιθυμίες του και είναι αδύνατον να προβλεφθούν με ακριβή τρόπο.

4.5 Περιγραφή του προβλήματος

Το σετ εκπαίδευσης περιέχει τις ενέργειες του χρήστη μέχρι μία καθορισμένη χρονική στιγμή (ημερομηνία διαχωρισμού). Μπορεί να χρησιμοποιηθεί για τη δημιουργία μοντέλων αλληλεπιδράσεων χρηστών και καθορίζει τον τύπο ενέργειας που έχει πραγματοποιηθεί (χρήση φίλτρου, βελτιώσεις αναζήτησης, αλληλεπίδραση του χρήστη με ένα στοιχείο, αναζήτηση ενός στοιχείου, click σε ένα συγκεκριμένο στοιχείο), καθώς και πληροφορίες σχετικά με τη λίστα των αντικειμένων και των τιμών που παρουσιάστηκαν στον χρήστη κατά τη διάρκεια του click-out.

Οι συστάσεις πρέπει να παρέχονται για ένα σετ δοκιμών που περιέχει πληροφορίες σχετικά με τις περιόδους σύνδεσης μετά την ημερομηνία διαχωρισμού, αλλά λείπουν οι πληροφορίες σχετικά με τα καταλύματα στα οποία έγινε κλικ στο τελευταίο μέρος των περιόδων. Το απαιτούμενο αποτέλεσμα είναι μια λίστα με μέγιστο αριθμό 25 ειδών για κάθε click-out που έχει ταξινομηθεί βάσει προτιμήσεων για τον συγκεκριμένο χρήστη. Όσο υψηλότερο το πραγματικό στοιχείο εμφανίζεται στη λίστα, τόσο υψηλότερη είναι η βαθμολογία. Το ακόλουθο σχήμα απεικονίζει τη ρύθμιση του προβλήματος και το διαχωρισμό των δεδομένων σε σετ εκπαίδευσης και δοκιμών [130].

Σχήμα 4.1: Διαχωρισμός σετ δεδομένων



Η υποβολή αποτελείται από μια λίστα προτεινόμενων ξενοδοχείων για κάθε click-out που λείπει στο σετ δοκιμών. Η μορφή της υποβολής επιτρέπει τη σαφή αναγνώριση του συγκεκριμένου click-out. Επομένως, το αρχείο που θα υποβληθεί στα πλαίσια της διπλωματικής εργασίας επιστρέφει μια λίστα με τα ξενοδοχεία τα οποία πρέπει να προταθούν στον χρήστη, ανάλογα με το πρώτο του clickout.

4.6 Μεταβλητές Trivago

Το σετ δεδομένων της trivago περιλαμβάνει 3 αρχεία με επέκταση .csv. Τα επιμέρους δεδομένα που στοιχειοθετούν μια συνεδρία χρήστη (session), είναι τα ακόλουθα:

- **user_id**: αναγνωριστικό του χρήστη
- **session_id**: αναγνωριστικό κάθε περιόδου σύνδεσης

- **timestamp**: timestamp σε UNIX για την ώρα της αλληλεπίδρασης
- **step**: βήμα στην ακολουθία ενεργειών εντός της περιόδου σύνδεσης
- **action_type**: ενέργεια του χρήστη
    1. **item clickout**: ο χρήστης κάνει ένα κλικ στο στοιχείο και προωθείται σε έναν ιστότοπο συνεργάτη. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    2. **interaction item rating**: ο χρήστης αλληλεπιδρά με μια αξιολόγηση ή κριτική ενός αντικειμένου. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    3. **interaction item info:** ο χρήστης αλληλεπιδρά με πληροφορίες ενός αντικειμένου. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    4. **interaction item image:** ο χρήστης αλληλεπιδρά με μια εικόνα ενός αντικειμένου. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    5. **interaction item deals:** κλικ του χρήστη στο κουμπί προβολής περισσότερων προσφορών. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    6. **change of sort order:** ο χρήστης αλλάζει τη σειρά ταξινόμησης. Η τιμή αναφοράς για αυτήν την ενέργεια είναι η περιγραφή της σειράς ταξινόμησης.
    7. **filter selection:** ο χρήστης επιλέγει ένα φίλτρο. Η τιμή αναφοράς για αυτήν την ενέργεια είναι η περιγραφή του φίλτρου.
    8. **search for item:** ο χρήστης αναζητά ένα κατάλυμα. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το item_id.
    9. **search for destination:** ο χρήστης αναζητά έναν προορισμό. Η τιμή αναφοράς για αυτήν την ενέργεια είναι το όνομα του προορισμού.
    10. **search for poi:** ο χρήστης αναζητά ένα σημείο ενδιαφέροντος (POI). Η τιμή αναφοράς για αυτήν την ενέργεια είναι το όνομα του POI.
- **reference:** τιμή αναφοράς της ενέργειας όπως περιγράφεται για τους διαφορετικούς τύπους ενεργειών.
- **platform**: χώρα πλατφόρμας που χρησιμοποιήθηκε για την αναζήτηση, π.χ. trivago.de (DE) ή trivago.com (ΗΠΑ)
- **city:** όνομα της τρέχουσας πόλης που εισήχθη στο περιβάλλον αναζήτησης
- **device**: συσκευή που χρησιμοποιήθηκε για την αναζήτηση
- **current_filters**: λίστα φίλτρων διαχωρισμένων με | που ήταν ενεργά στη δεδομένη χρονική σήμανση (timestamp)
- **impressions:** λίστα αντικειμένων διαχωρισμένων με | που εμφανίστηκαν στον χρήστη τη στιγμή ενός κλικ (δείτε action_type = clickout_item)
- **prices**: λίστα τιμών διαχωρισμένων με | των αντικειμένων που εμφανίστηκαν στον χρήστη τη στιγμή ενός κλικ (βλ. action_type = clickout_item)

Τα δεδομένα που αφορούν τις ιδιότητες των ξενοδοχείων, περιέχονται σε ξεχωριστό αρχείο .csv, το οποίο περιλαμβάνει τις εξής δύο στήλες:

- **item_id: id** του καταλύματος όπως χρησιμοποιείται στις τιμές αναφοράς για τύπους ενεργειών που σχετίζονται με στοιχεία, π.χ. clickout_item ή λίστα impressions
- **properties**: λίστα φίλτρων που χρησιμοποιήθηκαν για το δεδομένο αντικείμενο διαχωρισμένα με | .

Κάποιες βασικές παραδοχές για το σετ δεδομένων, οι οποίες προέκυψαν και από την περιγραφή των δεδομένων αλλά και την προσωπική τους μελέτη, είναι οι ακόλουθες:

1. Τα impressions είναι τα ξενοδοχεία που εμφανίζονται στον χρήστη με το που κάνει κλικ στο ξενοδοχείο που τον ενδιαφέρει. Δεν είναι κάποιο συναίσθημα (θετικό ή αρνητικό) του χρήστη.

2. Οι συστάσεις γίνονται κυρίως με βάση την τοποθεσία που έχει εισάγει ο χρήστης μέσω της μεταβλητής **destination.**

3. Οι προτάσεις πρέπει να παρέχονται για ένα σύνολο δοκιμών αλλά λείπει το τελευταίο μέρος κλικ που έγινε. Πρακτικά το **reference_item**.

4. Με βάση τα ξενοδοχεία που προτάθηκαν και την τοποθεσία, πρέπει να συμπεράνουμε πού έχει κάνει κλικ ο χρήστης.

5. Η σειρά των κλικ έχει σημασία. Το τελικό κλικ είναι περισσότερο σχετικό με τα αποτελέσματα σε σχέση με το αρχικό.

6. Κάπως πρέπει να οριστεί πότε μια session πρέπει να περιοριστεί (αν έχουμε πολλά άκυρα κλικ) – π.χ. πολλά search για destination και POI.

7. Δεν υπάρχει κάποιος τρόπος ορισμού επιτυχίας μιας session καθώς ο χρήστης απλά ανακατευθύνεται στο site με την καλύτερη τιμή και δεν εγγυάται κάποιος ότι με αυτή την τιμή θα κλείσει δωμάτιο.

5

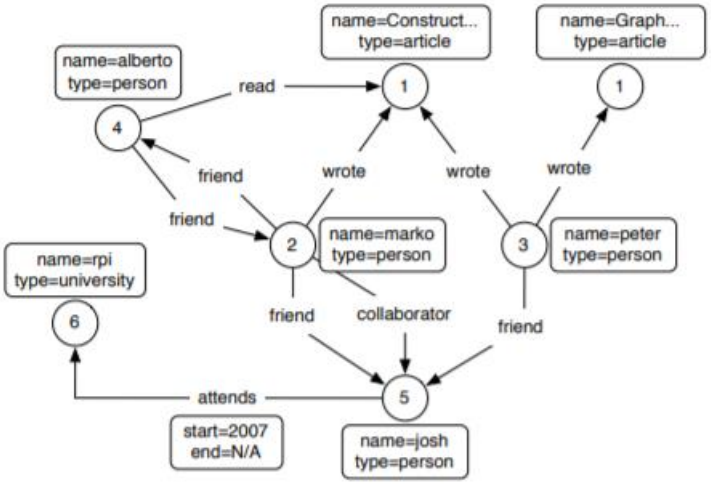
Τεχνολογικό πλαίσιο

Οι σχεσιακές βάσεις δεδομένων έχουν χρησιμοποιηθεί ευρέως για την ανάπτυξη εφαρμογών λογισμικού από τη δεκαετία του '80 έως σήμερα. Τα δεδομένα αποθηκεύονται με δομημένο τρόπο και οργανώνονται σε σχέσεις που συνίστανται μεταξύ των γραμμών και των στηλών με τη βοήθεια ενός προκαθορισμένου σχήματος. Αυτές οι σχέσεις δημιουργούνται με σύνδεση πρωτεύοντος-κλειδιού και ξένου κλειδιού. Η δημιουργία ενός προκαθορισμένου σχήματος απαιτεί περισσότερο χρόνο προκειμένου να γίνει ορθά η ανάκτηση ερωτημάτων σε βάσεις δεδομένων τεράστιες σε μέγεθος. Το μεγάλο μέγεθος συνεπώς δημιουργεί προβλήματα τόσο υπολογισμού όσο και πόρων μνήμης [131].

Οι περιορισμοί των παραδοσιακών βάσεων δεδομένων, ιδίως των προαναφερθέντων σχεσιακών βάσεων δεδομένων, για την κάλυψη των διαφορετικών και αυξανόμενων απαιτήσεων, οδήγησε στην ανάπτυξη νέων τεχνολογιών, οι οποίες ονομάζονται βάσεις δεδομένων NOSQL [132]. Με αυτόν τον τρόπο, τα δεδομένα που εισήχθησαν αποκτούν μια μεγαλύτερη ευκολία διαχείρισης, ξεφεύγοντας από την τετράγωνη λογική των σχεσιακών βάσεων δεδομένων. Τα πλεονεκτήματα των βάσεων δεδομένων που βασίζονται σε μη σχεσιακό τρόπο οργάνωσης και που οδήγησαν στην επιλογή μιας τέτοιας επιλογής ήταν η υψηλή διαθεσιμότητα και επεκτασιμότητα στο καθώς και ο γρήγορος χρόνος πρόσβασης και επεξεργασίας δεδομένων [133]. Μαζί με την επεκτασιμότητα η οποία μάλιστα όταν πρόκειται να συμβεί είναι χαμηλού κόστους, μια μη-σχεσιακή βάση δεδομένων υποστηρίζει τη μαζική αποθήκευση πολλών εγγραφών ταυτόχρονα [134]. Μάλιστα, κατά τα επόμενα ένα ή δύο χρόνια, ο O'Grady προέβλεπει ότι οι χρήστες θα υιοθετήσουν τέτοιου τύπου βάσεις δεδομένων κυρίως για εξειδικευμένα έργα, όπως αυτά, που περιλαμβάνουν μεγάλες ποσότητες δεδομένων ή δεδομένων που πρέπει να κλιμακωθούν [135].

5.1 *Γράφος ιδιοτήτων (Property Graph)*

Τι είναι όμως ένας γράφος; Ένας γράφος είναι μια συλλογή από άκρες και κορυφές [136] ή με πιο επιστημονικά λόγια ένα σύνολο από κόμβους και σχέσεις οι οποίες αναπτύσσονται μεταξύ αυτών των κόμβων. Οι γράφοι απεικονίζουν τις οντότητες ως κόμβους και τον τρόπο που αυτές οι οντότητες συνδέονται στην πραγματικότητα μεταξύ τους ως σχέσεις. Το σύστημα βάσης δεδομένων που βασίζεται σε γράφο ακολουθεί το πρότυπο CRUD (δημιουργία, ανάγνωση, ενημέρωση, διαγραφή) που στηρίζεται στις προαναφερθείσες μεθόδους, αλλά επίσης χρησιμοποιεί ευρετήριο χωρίς γειτνίαση [137]. Η προαναφερθείσα τεχνική ευρετηρίου χωρίς γειτνίαση είναι σημαντική προκειμένου να εξασφαλιστεί υψηλή απόδοση. Επίσης, ένα ακόμη πλεονέκτημα των γράφων είναι πως επιτρέπουν πολλαπλές σχέσεις μεταξύ πολλαπλών κόμβων [138]. Οι γράφοι τέτοιου τύπου επιτρέπουν την ανάθεση ετικετών (labels) και ιδιοτήτων (properties) στις κορυφές και τις άκρες του γράφου. Επιπλέον, κάθε κορυφή και άκρη έχει ένα μοναδικό αναγνωριστικό το οποίο μπορεί να χρησιμοποιηθεί ως σημείο αναφοράς για την συσχέτιση επιπλέον μετα-δεδομένων σε κάθε κορυφή ή άκρη, με την μορφή ζευγών κλειδιού-τιμής. Ο γράφος ιδιοτήτων αποτελεί μία βολική δομή επειδή περιέχει τα βασικά χαρακτηριστικά όσον αφορά την μοντελοποίηση γράφων ενώ απλές μορφοποιήσεις του έχουν ως αποτέλεσμα την δημιουργία διαφορετικών τύπων γράφων [139]. Έτσι τα συστήματα βάσεων δεδομένων που υποστηρίζουν το μοντέλο αυτό, υποστηρίζουν έμμεσα και άλλους τύπους γράφων. Παρακάτω, δίνεται μία οπτική αναπαράσταση των βασικών στοιχείων ενός γράφου ιδιοτήτων και πως αυτά συνδέονται μεταξύ τους.



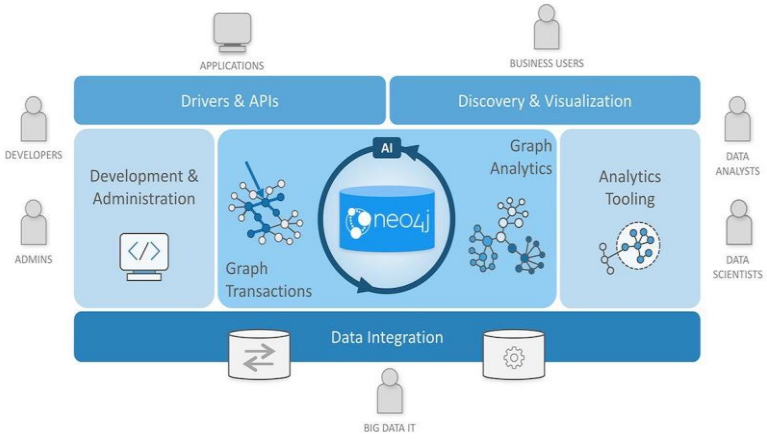
Σχήμα 5.1: Γράφος ιδιοτήτων

5.2 Neo4j

Το Neo4j είναι ένα σύστημα διαχείρισης βάσεων δεδομένων που στηρίζεται στο μοντέλο γράφου ιδιοτήτων όπως αναλύθηκε στην προηγούμενη ενότητα [140]. Όταν ασχολούμαστε με βάσεις δεδομένων τύπου γράφου, πρέπει να λάβουμε υπ' όψιν τις δύο βασικές ιδιότητές τους: τον τρόπο επεξεργασίας του γράφου καθώς και τον τρόπο αποθήκευσής του. Ανάλογα με την υλοποίηση που ακολουθεί μια βάση δεδομένων, μπορεί να αποθηκεύει τα δεδομένα είτε σε βάσεις σχεσιακού τύπου είτε σε άλλες βάσεις γενικού σκοπού. Αντιθέτως, η Neo4j υποστηρίζει την 'κανονική αποθήκευση γράφων', η οποία έχει ως αποτέλεσμα την αύξηση της αποτελεσματικότητας και της αποδοτικότητας. Τα δεδομένα αποθηκεύονται σε δύο διαφορετικά αρχεία, ανάλογα με το τμήμα της πληροφορίας που υπάρχει στον γράφο. Με αυτόν τον διαχωρισμό, οδηγούμαστε στην αποτελεσματικότερη εκτέλεση των ερωτημάτων, παρόλο που ο χρήστης δεν έχει πρόσβαση στον γράφο όπως αυτός αποθηκεύεται στη φυσική μνήμη [141]. Αναλυτικότερα, το Neo4j είναι μια βάση δεδομένων προσανατολισμένη στο δίκτυο - βασισμένη σε Java, ενώ αποθηκεύει δεδομένα δομημένα σε δίκτυα και όχι σε πίνακες. Ουσιαστικά είναι αυτό που τη μετατρέπει σε μια βάση δεδομένων προσανατολισμένη στο δίκτυο [142]. Θεωρείται η πιο δημοφιλής και χρησιμοποιημένη βάση δεδομένων γραφημάτων παγκοσμίως, έχει τη μεγαλύτερη θέση σε αναφορές και επομένως έχει μια ουσιαστική αναγνώριση στην ερευνητική και εμπορική κοινότητα [143,144]. Επιπλέον, το Neo4j διαθέτει συνεχείς πηγές ενημέρωσης, διοργανώνει σχετικές εκδηλώσεις, ενώ υπάρχει και μια ολόκληρη κοινότητα η οποία υποστηρίζει τους χρήστες με τις απορίες τους [145]. Οι ιστοσελίδες αυτές έχουν ως σκοπό την εξοικείωση των χρηστών με τις λειτουργίες της, την συνεχή ενημέρωση των χρηστών με τις νέες λειτουργίες/ δυνατότητες που προσφέρει το σύστημα, αλλά και διαρκή επίλυση απορίων ώστε οι χρήστες να νιώθουν σίγουροι προκειμένου να εντάξουν ένα τέτοιο σύστημα στη λειτουργία τους.

Τι είναι όμως αυτό που κάνει τη Neo4j να ξεχωρίζει; Για δεδομένα υψηλής σύνδεσης, είναι χιλιάδες φορές πιο γρήγορη σε σχέση με τις σχεσιακές βάσεις δεδομένων, καθιστώντας το ιδανική επιλογή για διαχείριση σύνθετων δεδομένων σε πολλούς τομείς, από τα οικονομικά έως τα κοινωνικά, τηλεπικοινωνίες ή γεωχωρικά δεδομένα [146].

Από το βασικό προϊόν (τη βάση δεδομένων γραφημάτων) έως την οπτικοποίηση για την αποτελεσματικότερη διαχείριση των δεδομένων για επαγγελματική χρήση, παρέχεται ένα ολοκληρωμένο πλαίσιο που αποτελείται από διαφορετικές συνιστώσες. Καθένα από τα στοιχεία που παρατίθενται στην εικόνα σχεδιάστηκε για να καλύψει μια επιχειρηματική ή τεχνική [147].



Σχήμα 5.2: Πλαίσιο Neo4j

- Neo4j Graph Database - η βασική βάση γραφημάτων που έχει δημιουργηθεί για αποθήκευση και ανάκτηση συνδεδεμένων δεδομένων. Το λογισμικό κυκλοφόρησε για πρώτη φορά το 2007 και χωρίζεται σε δύο μεγάλες κατηγορίες: Enterprise, Community. Η έκδοση Community είναι η δοκιμαστική, την οποία μπορεί να δοκιμάσει οποιοσδήποτε χρήστης. Η έκδοση Enterprise, είναι η πιο ολοκληρωμένη έκδοση από την έκδοση Community την οποία μπορεί να χρησιμοποιήσει για να δοκιμάσει οποιοσδήποτε χρήστης δωρεάν για 30 ημέρες. Οι κύριες διαφορές μεταξύ των δύο κύριων εκδόσεων του Neo4j (Community and Enterprise) είναι οι ακόλουθες: η ύπαρξη online backup, το υψηλό επίπεδο απόδοσης της κρυφής μνήμης, ένα αναλυτικό σύστημα παρακολούθησης, ισχυρή διαχείριση των κλειδιών στο βάση δεδομένων και μεγαλύτερη επεκτασιμότητα της βάσης δεδομένων. Όλα αυτά είναι πλεονεκτήματα της έκδοσης Enterprise τα οποία δεν εμφανίζουν οι άλλες δύο εκδόσεις [148].
- Neo4j Desktop - εφαρμογή για τη διαχείριση τοπικών παραδειγμάτων του Neo4j. Η δωρεάν λήψη περιλαμβάνει άδεια έκδοσης Enterprise, που αναλύθηκε προηγουμένως.
- Neo4j Browser - διαδικτυακή διεπαφή προγράμματος περιήγησης για αναζήτηση και προβολή των δεδομένων στη βάση δεδομένων. Βασικές δυνατότητες οπτικοποίησης με χρήση της γλώσσας ερωτημάτων Cypher.
- Neo4j Bloom - εργαλείο οπτικοποίησης για επαγγελματίες χρήστες που δεν διαθέτουν δεξιότητες κώδικα ή προγραμματισμού για την προβολή και ανάλυση δεδομένων.
- Neo4j AuraDB - προσφορά βάσης δεδομένων ως υπηρεσία που διαχειρίζεται η Neo4j για βάσεις δεδομένων γραφημάτων στο cloud. Τα δέκα βασικότερα πλεονεκτήματα της υπηρεσίας αυτής είναι τα ακόλουθα:
    - **Καινοτόμα**: Το Neo4j AuraDB Enterprise περιλαμβάνει ένα ισχυρό εργαλείο εξερεύνησης και οπτικοποίησης γραφημάτων, Neo4j Bloom. Η χρήση αυτού

του εργαλείου οδηγεί σε προβολή των δεδομένων από διαφορετική οπτική γωνία, γεγονός που προσδίδει πλουραλισμό στις σχέσεις και στην ερμηνεία τους.

- **Έμπιστη**: Το Neo4j AuraDB έχει σχεδιαστεί ειδικά για αποθήκευση, ανάλυση σχέσεων δεδομένων και ανάκτηση πληροφοριών σε πραγματικό χρόνο. Με το AuraDB, οι προγραμματιστές επικεντρώνονται σε ό,τι είναι σημαντικό: χτίσιμο και συντήρηση εφαρμογών χωρίς να υπάρχει ανησυχία για τη διαχείριση της υποδομής βάσεων δεδομένων.
- **Άκοπη**: η διαχείριση και διατήρηση της βάσης είναι πλήρως αυτοματοποιημένη για το cloud και πραγματοποιείται από καταρτισμένους ειδικούς.
- **Κλιμακούμενη**: Κλιμακούμενη ανάπτυξη και διαχείριση της εκάστοτε εφαρμογής. Προσαρμοσμένες υποδομές για την κάλυψη των ποικίλων αναγκών και απαιτήσεων κάθε περίπτωσης χωριστά.
- **Γρήγορη**: περισσότερα δεδομένα σημαίνουν μεγαλύτερη αξία. Τα γρήγορα ερωτήματα σε δισεκατομμύρια κόμβους παραδίδουν πληροφορίες σε πραγματικό χρόνο. Στις σχεσιακές βάσεις δεδομένων, είναι άκαμπτα και οποιαδήποτε αλλαγή σε αυτά ενοχλητική σε αντίθεση με την AuraDB όπου συμβαδίζουν η ταχύτητα και η ευελιξία.
- **Ασφαλής**: τα δεδομένα που αποθηκεύονται, συμπεριλαμβανομένων των εφεδρικών στιγμιότυπων κρυπτογραφούνται χρησιμοποιώντας το Προηγμένο Πρότυπο Κρυπτογράφησης (AES). Το Neo4j AuraDB Enterprise αναπτύσσει τα cluster και τα στοιχεία της υπηρεσίας σε ξεχωριστό Virtual Private Cloud (VPC), με αποκλειστική υποδομή cloud.
- **Αξιόπιστη**: το AuraDB διαθέτει υπηρεσία εγγύησης διαθεσιμότητας που ανέρχεται σε ποσοστό 99,5%. Είναι μοναδικά σχεδιασμένη ώστε να είναι πάντα ενεργοποιημένη, χωρίς προγραμματισμένη διακοπή συντήρησης. Οι συναλλαγές βάσης δεδομένων στο Neo4j AuraDB είναι πλήρως συμβατές με το πρωτόκολλο ACID.
- **Παραγωγική**: Το Neo4j AuraDB Enterprise επιτρέπει στους προγραμματιστές να δημιουργούν εφαρμογές πιο γρήγορα και πιο εύκολα χρησιμοποιώντας προεγκατεστημένους οδηγούς, εργαλεία και ενσωματώσεις για δημοφιλείς γλώσσες και τεχνολογικά πλαίσια. Αυτό σημαίνει πως είναι δυνατό να συνεργαστεί με οποιαδήποτε γλώσσα προγραμματισμού.
- **χωρίς υψηλό κόστος**: Το Neo4j AuraDB Enterprise τιμολογείται με διαφορετικό τρόπο. Η τιμολόγηση είναι απλή καθώς ο χρήστης έχει

διαφορετική τιμολόγηση βάσει χωρητικότητας, γεγονός που δίνει τη δυνατότητα στον χρήστη/εταιρεία να έχει τον έλεγχο

- **με μελλοντικές προοπτικές**: Το Neo4j παρέχει πρώτα νέες δυνατότητες στο cloud. Αυτό ελαχιστοποιεί την διάρκεια αναμονής σε σχέση με τον τυπικό κύκλο εφαρμογής της τελευταίας έκδοσης: (εγκατάσταση της νέας έκδοσης, δοκιμή, παραγωγή). Αυτή η διαδικασία διαρκεί συνήθως περίπου 18 μήνες.

- Graph Data Science - υποστηριζόμενη βιβλιοθήκη για την εκτέλεση αλγορίθμων γραφημάτων με το Neo4j. Το Graph Data Science είναι μια προσέγγιση που βασίζεται στην επιστήμη για την απόκτηση γνώσης από τις σχέσεις και τις δομές στα δεδομένα, για τη δημιουργία προβλέψεων. Είναι ένα σύνολο από τεχνικές που βοηθούν τους επιστήμονες δεδομένων χρησιμοποιώντας δεδομένα γραφημάτων να απαντήσουν σε ερωτήσεις και να εξηγήσουν τα αποτελέσματα [148]. Η βιβλιοθήκη περιέχει υλοποιήσεις κλασικών αλγορίθμων γραφημάτων όπως: εύρεσης διαδρομής, κεντρικότητας και ανίχνευσης κοινότητας. Περιλαμβάνει επίσης αλγόριθμους που είναι κατάλληλοι για τα σημαντικότερα προβλήματα επιστήμης δεδομένων, όπως η πρόβλεψη συνδέσμων και σταθμισμένης- μη σταθμισμένης ομοιότητας [149].

  - **Neo4j Graph Data Science Library (GDSL)**

  Αυτοί οι αλγόριθμοι εκτελούνται ως διαδικασίες Neo4j. Καλούνται απευθείας από το πρόγραμμα περιήγησης Neo4j, από το cypher-shell ή από τον κώδικα που ετοιμάζει ο ίδιος ο χρήστης. Για τους περισσότερους αλγόριθμους, υπάρχουν δύο διαδικασίες:

  - algo.<name> – Αυτή η διαδικασία γράφει τα αποτελέσματα πίσω στο γράφημα ως ιδιότητες κόμβου,
  - algo.<name>.stream – Αυτή η διαδικασία επιστρέφει μια ροή δεδομένων. Για παράδειγμα, nodeids και υπολογισμένες τιμές.

  Για μεγάλα γραφήματα, η διαδικασία ροής μπορεί να επιστρέψει εκατομμύρια, ή ακόμα και δισεκατομμύρια, αποτελέσματα, οπότε σε αυτήν την περίπτωση, συνίσταται να αποθηκεύονται τα δεδομένα και μπορούν να εξορυχθούν με μεταγενέστερες ερωτήσεις.

  - **NEuler: No-code graph algorithms**

  Το NEuler είναι μια διεπαφή χρήστη που βοηθά τους χρήστες να χρησιμοποιήσουν τη Neo4j Graph Data Science Library, χωρίς κώδικα. Υποστηρίζει την εκτέλεση καθενός από τους αλγόριθμους γραφημάτων καθώς και την προβολή των αποτελεσμάτων. Κατά τη δημιουργία αυτών των

ερωτημάτων μέσω της διεπαφής χρήστη, προβάλλεται ο κώδικας Cypher ώστε οι χρήστες να μπορέσουν να εξοικειωθούν ή να τον επαναχρησιμοποιήσουν.

- **Natural Language Processing (NLP)**

  Το Neo4j προσφέρει ισχυρές δυνατότητες αναζήτησης για δομημένα δεδομένα, αλλά πολλά από τα δεδομένα υπάρχουν σε έγγραφα κειμένου. Οι τεχνικές NLP μπορούν να βοηθήσουν στην εξαγωγή της λανθάνουσας δομής σε αυτά τα έγγραφα. Αυτή η δομή θα μπορούσε να είναι τόσο απλή όσο οι κόμβοι που αντιπροσωπεύουν σε μια πρόταση ή τόσο περίπλοκη όσο οι κόμβοι που αντιπροσωπεύουν οντότητες που εξάγονται χρησιμοποιώντας έναν ονομασμένο αλγόριθμο αναγνώρισης οντοτήτων.

Η σύγκριση της βάσης δεδομένων Neo4j με τις σχεσιακές βάσεις δεδομένων, διαφέρει ως προς τα εξής βασικά σημεία [150]:

- Αποθήκευση βασικών τιμών: Τα δεδομένα αποθηκεύονται σε απλούστερη μορφή, ως ζεύγος τιμών.
- Αποθήκευση προσανατολισμένη στη στήλη: Τα δεδομένα αποθηκεύονται σε στήλη τρόπο και όχι σειρές.
- Βάση δεδομένων αποθήκευσης εγγράφων: Παρέχει έναν αποτελεσματικό τρόπο διαχείρισης πληροφοριών που προσανατολίζονται σε έγγραφα σε ημιδομημένη μορφή δεδομένων. Έχει ένα στρώμα για να διατηρεί τη σχέση μεταξύ αυτών των εγγράφων.
- Βάση δεδομένων αποθήκευσης γραφημάτων: Αυτός ο τύπος αποθήκευσης δεδομένων διατηρεί τα δεδομένα σε δομή γραφήματος για την αναπαράσταση της σχέσης μεταξύ τους. Αποθηκεύει τα δεδομένα ως προς τους κόμβους, ως άκρες ή ως ιδιότητες.

Πίνακας 5.1: Σύγκριση ΣΒΔ/Neo4j

	RDBMS	NEO4J
1.	Πίνακες	Γραφήματα
2.	Σειρές	Κόμβοι
3.	Στήλες	Ιδιότητες και οι τιμές τους
4.	Ένωση	Διάσχιση

Παραδοσιακά, οι σχεσιακές βάσεις δεδομένων όπως η PostgreSQL χρησιμοποιούνταν συχνότερα, ενώ οι νεότερες βάσεις δεδομένων γραφημάτων όπως το Neo4j είχαν υποβιβαστεί στην ανάλυση των συνόλων δεδομένων κοινωνικών δικτύων και μεταφοράς [151] . Ενώ το

Neo4j είναι πιο απαιτητικό για την εφαρμογή του, τα ερωτήματά του είναι λιγότερο περίπλοκα και έχουν ταχύτερο χρόνο εκτέλεσης από συγκρίσιμα ερωτήματα που εκτελούνται στην PostgreSQL. Αυτό οδηγεί στο συμπέρασμα ότι ενώ η PostgreSQL κρίνεται επαρκής ως βάση δεδομένων σε πολλές περιπτώσεις, το Neo4j πρέπει να θεωρείται ένα καλό πλαίσιο για αποθήκευση και ανάλυση δεδομένων υγείας. Σε κάθε περίπτωση, το Neo4j χρησιμοποιείται για να αναπαραστήσει το εξεταζόμενο πρόβλημα, να διαχειριστεί τα δεδομένα του κύκλου ζωής του προγραμματισμού και να επεξεργαστεί το εφικτό προγραμματίζει με τη βοήθεια της ισχυρής του ικανότητας μοντελοποίησης και επεξεργασίας δεδομένων γραφημάτων [152].

5.3 Cypher

Η Cypher είναι η γλώσσα ερωτήσεων του Neo4j που επιτρέπει στους χρήστες να αποθηκεύουν και να ανακτούν δεδομένα από τη βάση δεδομένων των γραφημάτων. Το Neo4j θέλησε να κάνει εύκολη την εκμάθηση, κατανόηση και χρήση για όλους, αλλά και να ενσωματώσει την ισχύ και τη λειτουργικότητα άλλων τυπικών γλωσσών πρόσβασης δεδομένων.

Η σύνταξη της Cypher παρέχει έναν οπτικό και λογικό τρόπο για να ταιριάζει με τα πρότυπα κόμβων και σχέσεων στο γράφημα. Είναι μια δηλωτική γλώσσα και θεωρείται ως παράγωγο της SQL περιγράφοντας οπτικά μοτίβα σε γραφήματα χρησιμοποιώντας τη σύνταξη ASCII-Art [153]. Μας επιτρέπει να δηλώσουμε τι θέλουμε να επιλέξουμε, να εισαγάγουμε, να ενημερώσουμε ή να διαγράψουμε από τα δεδομένα των γραφημάτων μας χωρίς περιγραφή ακριβώς πώς να το κάνουμε. Μέσω της Cypher, οι χρήστες μπορούν να δημιουργήσουν εκφραστικά και αποτελεσματικά ερωτήματα για να χειριστούν τις λειτουργίες που απαιτούνται για τη δημιουργία, ανάγνωση, ενημέρωση και διαγραφή. Είναι η πιο εύκολη γλώσσα ερωτήσεων για εξόρυξη γνώσης λόγω της ομοιότητάς της με άλλες γλώσσες και της ευκολίας χρήσης της.

Η Cypher δεν είναι μόνο ο καλύτερος τρόπος αλληλεπίδρασης με τα δεδομένα και το Neo4j , είναι επίσης και ένα λογισμικό ανοιχτού κώδικα. Το έργο openCypher παρέχει μια ανοιχτή προδιαγραφή γλωσσών, ένα κιτ τεχνικής συμβατότητας και μια εφαρμογή αναφοράς για τον αναλυτή, το σχεδιασμό και το χρόνο εκτέλεσης για τη Cypher. Υποστηρίζεται από αρκετές εταιρείες στη βιομηχανία βάσεων δεδομένων και επιτρέπει στους προγραμματιστές βάσεων δεδομένων και πελατών να επωφελούνται ελεύθερα από τη χρήση και ταυτόχρονα να συμβάλλουν στην ανάπτυξη της γλώσσας openCypher.

Η κεντρική ιδέα στα ερωτήματα Cypher είναι η αντιστοίχιση μοτίβων. Τα μοτίβα στο Cypher εκφράζονται με «μορφή ASCII», δηλαδή (a)-[r]->(b). Η ρήτρα MATCH στη Cypher χρησιμοποιεί ένα τέτοιο μοτίβο και εισάγει νέες σειρές (συνώνυμες με τις εγγραφές που ταιριάζουν στο μοτίβο). Η Cypher υποστηρίζει επίσης τα μοτίβα επιστροφής και αντιστοίχισης ως μεταβλητές. Εκτός από τη ρήτρα MATCH, η Cypher περιλαμβάνει και ρήτρες ενημέρωσης/

τροποποίησης του γραφήματος. Οι βασικές ρήτρες ενημέρωσης περιλαμβάνουν τη ρήτρα CREATE για τη δημιουργία νέων κόμβων και σχέσεων, τη ρήτρα DELETE για διαγραφή οντοτήτων και SET για ενημέρωση ιδιοτήτων. Επιπροσθέτως, η Cypher παρέχει μια ρήτρα που ονομάζεται MERGE που προσπαθεί να ταιριάζει μια οντότητα με το δεδομένο μοτίβο και δημιουργεί την οντότητα εάν δεν βρέθηκε αντιστοιχία. Επιπλέον, διαθέτει ενσωματωμένη υποστήριξη για παραμέτρους ερωτήματος, καθιστώντας εύκολη την εξάλειψη των προβλημάτων που δημιουργούνται από ενέσεις ερωτημάτων.

Τα βασικά στοιχεία του Cypher είναι τα εξής:

- μοντέλο δεδομένων, που περιλαμβάνει τιμές, γραφήματα και πίνακες.
- γλώσσα ερωτήματος, που περιλαμβάνει εκφράσεις, μοτίβα, προτάσεις και ερωτήματα

5.3.1 Η Cypher ως παράδειγμα εντολών

Όλες οι βάσεις δεδομένων γραφημάτων NoSQL απαιτούν από τους προγραμματιστές και τους χρήστες να χρησιμοποιούν έννοιες γραφημάτων για αναζήτηση δεδομένων. Όπως για οποιοδήποτε άλλη βάση δεδομένων τέτοιου τύπου, στα μοτίβα ερωτημάτων, οι χρήστες είτε απαιτούν συγκεκριμένες εστιασμένες γνώσεις (π.χ. προβολή ιδιοτήτων ενός ξενοδοχείου) είτε ρωτούν για ανίχνευση τάσεων (π.χ. ανίχνευση τάσεων και συμπεριφορών όταν γίνεται μια ξενοδοχειακή κράτηση) [154].Οι βασικές ρήτρες που περιεγράφηκαν στο προηγούμενο κεφάλαιο MATCH, CREATE, SET, MERGE χρησιμοποιήθηκαν για την εισαγωγή των δεδομένων. Για να γίνει εύκολα αντιληπτή, χρησιμοποιούνται παραδείγματα που χρησιμοποιήθηκαν στο υπάρχον σύνολο δεδομένων και δομή γράφου.

```
LOAD CSV with headers FROM
"file:///datagreece_changed_onlyview.csv" AS profile
FIELDTERMINATOR';'
MERGE (e:Event {value:profile.action_type,step:profile.newstep})
MERGE (h:Hotel {value:profile.reference})
```

Το παρακάτω ερώτημα Cypher εισάγει εντός της βάσης δεδομένων από το αρχείο datagreece_changed_onlyview.csv. Οι κόμβοι οι οποίοι εισάγονται στο γράφο είναι ο κόμβος Event (Συμβάντος) καθώς και ο κόμβος Hotel (Ξενοδοχείου). Ένα συμβάν το οποίο μπορεί να έχει οποιαδήποτε τιμή όπως αναφέρεται στην ιδιότητα value και ένα ακέραιο βήμα (ανάλογα με τη σειρά του εντός της συνεδρίας) αναφέρεται σε ένα ξενοδοχείο, το οποίο έχει ως ιδιότητα

την τιμή αναφοράς του. Γίνεται εύκολα αντιληπτό, πως αν χρησιμοποιούνταν η ρήτρα CREATE αντί του MERGE σε ένα τόσο μεγάλο σύνολο δεδομένων ήταν πολύ πιθανό να δημιουργούνταν διπλοεγγραφές στις σχέσεις, κάτι το οποίο θα κόστιζε στους πόρους εκτέλεσης των ερωτημάτων στη συνέχεια. Αξίζει να αναφερθεί ότι η κατάληξη .csv οδηγεί στο να γίνει αντιληπτός ο τύπος δεδομένων που εισάγεται εντός της βάσης. Ένα αρχείο τιμών διαχωρισμένων με κόμματα (CSV) είναι ένα αρχείο απλού κειμένου που περιέχει μια λίστα δεδομένων[155]. Τέτοιου τύπου αρχεία χρησιμοποιούνται συχνά για την ανταλλαγή δεδομένων μεταξύ διαφορετικών εφαρμογών και είναι ένας συνήθης τρόπος εισαγωγής δεδομένων στη βάση δεδομένων. Εκτός όμως από την εισαγωγή ενός αρχείου CSV ο χρήστης μπορεί να επιλέξει ανάμεσα σε διαφορετικές μορφές αρχείων όπως για παράδειγμα δυναμικά αρχεία JSON, JSON Rest APIs και δεδομένα από σχεσιακού τύπου βάσεις δεδομένων (PostgreSQL).

Προκειμένου να γίνει ένας πρώτος μικρός έλεγχος για το πώς έχουν εισαχθεί τα δεδομένα μέσα στο σύστημα, έστω ότι είναι επιθυμητό από τον χρήστη να επιστραφούν μόνο οι κόμβοι που έχουν ως τιμή την πόλη Θεσσαλονίκη. Προκειμένου να επιστραφούν, χρησιμοποιείται η ρήτρα RETURN. Η ρήτρα RETURN είναι η λέξη-κλειδί που εισάγει ένα φίλτρο. Σε αυτήν την περίπτωση ένα φίλτρο που βασίζεται στη σύγκριση ισότητας μιας ιδιότητας και μιας τιμής (Θεσσαλονίκη).

```
MATCH (c:City)
WHERE c.value= ‘Thessaloniki, Greece’
RETURN c
```

Το ίδιο ερώτημα μπορούσε να γραφεί σε Cypher με τον ίδιο τρόπο:

```
MATCH (c:City {c.value= ‘Thessaloniki, Greece’})
RETURN c
```

Τα αποτελέσματα εκτέλεσης και στις δύο περιπτώσεις είναι τα ίδια, καθώς ερμηνεύονται με τον ίδιο τρόπο από τον μεταγλωττιστή. Αυτό σημαίνει ότι η επιλογή μιας σύνταξης είναι απλώς θέμα γούστου. Η πρώτη περίπτωση σύνταξης είναι πιο κοντά στην SQL. Στην πραγματικότητα,

όπως αναφέρθηκε, οι δημιουργοί της Cypher έχουν εμπνευστεί από την SQL την εισαγωγή της ρήτρας WHERE.

Σε περίπτωση που κάποιος από τους κόμβους δεν είναι επιθυμητό να βρίσκεται στο γράφημα, μπορούμε να το διαγράψουμε χρησιμοποιώντας τη ρήτρα DELETE. Χρησιμοποιώντας τη ρήτρα DETACH διαγράφουμε ταυτόχρονα και τις σχέσεις οι οποίες υπάρχουν πάνω στον κόμβο. Προκειμένου να εντοπιστεί ακριβώς ο κόμβος που είναι επιθυμητό να διαγραφεί, χρησιμοποιείται η ρήτρα ταιριάσματος MATCH. Στο παρακάτω ερώτημα, διαγράφονται τα ξενοδοχεία εκείνα που έχουν ως ιδιότητα την τιμή "Zero". Η ρήτρα MATCH έχει μια προαιρετική παραλλαγή: OPTIONAL MATCH, η οποία είναι ανάλογη με το OUTER JOIN στην SQL. Αυτή η ρήτρα παράγει σειρές για όλους τα ξενοδοχεία με τον ίδιο τρόπο που κάνει το MATCH, εξασφαλίζοντας όμως πως ότι ολόκληρο το μοτίβο βρίσκεται στο γράφημα δεδομένων [156].

```
MATCH (h:Hotel {value: 'Zero'})
DETACH DELETE h
```

Σε αυτό το ερώτημα θέλουμε να βάλουμε βάρος σε μία σχέση, τη σχέση LOCATES_IN. Εκτός από τους κόμβους, ιδιότητες μπορεί να έχουν και οι σχέσεις. Στην προκειμένη περίπτωση, δίνεται βάρος στη σχέση της τοποθεσίας ώστε να αποκτήσει μεγαλύτερη βαρύτητα κατά τον υπολογισμό του node2vec. Αυτό γίνεται με τη βοήθεια της ρήτρας SET. Σε περίπτωση που θέλουμε να προβάλλουμε μόνο τα 3 πρώτα αποτελέσματα, διότι με αυτό το ερώτημα θα μας προβληθούν όλες οι σχέσεις ανάμεσα στις οντότητες Ξενοδοχείο και Πόλη, μπορούμε να χρησιμοποιήσουμε τη ρήτρα LIMIT. Τη ρήτρα LIMIT, διαδέχεται ένας ακέραιος αριθμός, ο οποίος προσδιορίζει το σύνολο των αποτελεσμάτων που είναι επιθυμητό να προβληθούν.

```
MATCH  (h:Hotel)-[l:LOCATES_IN]->(c:City)
SET l.weight = 3
RETURN h.value
```

Έναν έτοιμο αλγόριθμο από τη βιβλιοθήκη gds, τον καλούμε χρησιμοποιώντας τη ρήτρα CALL. Η σύνταξη που ακολουθεί κάθε αλγόριθμο και οι μεταβλητές που δέχεται ως παραμέτρους, εξαρτάται κάθε φορά από το είδος του αλγορίθμου που εκτελούνται.

```
CALL gds.beta.node2vec.write ('test21', {walkLength:3,
walksPerNode:50, embeddingDimension: 10, writeProperty:
"embeddingNode2vec", relationshipTypes:['test'] })
```

Μία δυνατότητα που προσφέρει η Cypher και υπάρχει και στις σχεσιακές βάσεις δεδομένων, είναι η συγκεντρωτική προβολή/ ομαδοποίηση αποτελεσμάτων. Μερικές φορές χρειάζεται μόνο να επιστρέφεται ένας αριθμός αποτελεσμάτων που βρέθηκαν στη βάση δεδομένων, αντί να επιστραφούν τα ίδια τα αντικείμενα. Η συνάρτηση COUNT στο Cypher επιτρέπει την καταμέτρηση των οντοτήτων, σχέσεων ή αποτελεσμάτων που επιστρέφονται. Υπάρχουν δύο διαφορετικοί τρόποι με τους οποίους μπορείτε να μετρήσετε τα αποτελέσματα που επιστρέφονται από κάθε ερώτημα. Ο πρώτος τρόπος είναι χρησιμοποιώντας το COUNT(n) για να μετρηθεί ο αριθμός των εμφανίσεων του n και δεν περιλαμβάνει μηδενικές τιμές. Μπορούν να καθοριστούν κόμβους, σχέσεις ή ιδιότητες μέσα στις παρενθέσεις για να μετράει το Cypher. Ο δεύτερος τρόπος μέτρησης των αποτελεσμάτων είναι με το COUNT(*), ο οποίος μετράει τον αριθμό των σειρών αποτελεσμάτων που επιστρέφονται (συμπεριλαμβανομένων εκείνων με μηδενικές τιμές).

Πώς όμως η Cypher εκτελείται εντός της python μέσω της οποίας γίνεται η μέτρηση της ακρίβειας προβλέψεως των αποτελεσμάτων; Μέσω της εκτέλεσης του παρακάτω τμήματος κώδικα. Στην αρχή, συνήθως εισάγονται οι βιβλιοθήκες για τη neo4j ενώ στη συνέχεια δίνονται τα στοιχεία της βάσης.

```
# import the neo4j driver for Python
from neo4j import GraphDatabase
import pandas as pd
import numpy as np
from sklearn.metrics import label_ranking_average_precision_score
# Database Credentials
uri          = "bolt://localhost:11003"
userName     = "neo4j"
password     = "****"
graphDB_Driver  = GraphDatabase.driver (uri, auth=(userName,
```

Εκτελώντας αυτό το κομμάτι κώδικα στην python, μπορούμε να εκτελέσουμε από εκεί εντολές cypher οι οποίες είτε απαντούν σε ερωτήματα, είτε εκτελούν εντολές (προσθήκη, διαγραφή, ενημέρωση κόμβων).

5.4 R

Προκειμένου τα δεδομένα να έρθουν σε μία επεξεργάσιμη μορφή διαχωρίστηκαν με τη βοήθεια της γλώσσας προγραμματισμού R, έτσι ώστε να εισαχθούν σε μια επεξεργάσιμη μορφή στο Neo4j. Η R είναι ένα πρόγραμμα υπολογιστή και παράλληλα μια γλώσσα / περιβάλλον στατιστικού προγραμματισμού[157]. Αυτός ο μάλλον απίθανος γλωσσικός συνδυασμός πιθανώς δεν θα είχε προετοιμαστεί ποτέ από επιστήμονες υπολογιστών, ωστόσο, η γλώσσα έγινε εκπληκτικά δημοφιλής[158]. Εκτός του ότι είναι δωρεάν, η R είναι εκπληκτικά δημοφιλής εν μέρει επειδή παρουσιάζει διαφορετική οπτική σε διαφορετικούς χρήστες. Είναι, πρώτα απ 'όλα, μια γλώσσα προγραμματισμού - που απαιτεί εισαγωγή κώδικα μέσω μιας γραμμής εντολών, η οποία μπορεί να φαίνεται απαγορευτική για τους μη-γνώστες κώδικα[159]. Μια σημαντική ιδιότητα της R είναι ότι οι περιορισμοί αντιμετωπίζονται ομοιόμορφα με την έννοια ότι χρησιμοποιούνται για τον καθορισμό των παραμέτρων εισόδου σε ένα πρόγραμμα, είναι τα μόνα αρχικά στοιχεία που χρησιμοποιούνται στην εκτέλεση ενός προγράμματος και χρησιμοποιούνται για να περιγράψουν την έξοδο ενός προγράμματος [160].

Υπάρχουν πολλοί καλοί λόγοι για να προτιμηθεί η R έναντι άλλων γλωσσών προγραμματισμού για επιστημονικούς υπολογισμούς. Η ύπαρξη μιας συλλογής καλών στατιστικών αλγορίθμων, η πρόσβαση σε αριθμητικές ρουτίνες υψηλής ποιότητας και στα εργαλεία οπτικοποίησης ολοκληρωμένων δεδομένων είναι ίσως από τους πιο προφανής λόγους που μια τέτοια γλώσσα είναι ιδιαίτερα διαδεδομένη[161]. Η γλώσσα περιεγράφηκε αρχικά από τους Ihaka and Gentleman (1996) και ήταν το αποτέλεσμα μιας προσπάθειας συνδυασμού χρήσιμων χαρακτηριστικών δύο υφιστάμενων γλωσσών υπολογιστών, S και Scheme. Η R εφαρμόζει εμφανίζει πολλές ομοιότητες με την S γλώσσα προγραμματισμού[162]. Ενώ η σύνταξη του R είναι σχεδόν ίδια με αυτή της γλώσσας S, η σημασιολογία του R, ενώ είναι επιφανειακά παρόμοια με το της S, στην ουσία είναι αρκετά διαφορετική. Στην πραγματικότητα, το R είναι τεχνικά πολύ πιο κοντά στη γλώσσα Scheme από ό, τι στην αρχική γλώσσα S όταν πρόκειται για τον τρόπο λειτουργίας της R[163].

5.5 VBA

Για την προεπεξεργασία των δεδομένων χρησιμοποιήθηκε εκτός από την R και η Visual Basic. Συγκεκριμένα, επειδή η εισαγωγή των δεδομένων έγινε σε csv ήταν εύκολη η συγγραφή scripts

προκειμένου τα δεδομένα να εκκαθαριστούν, ώστε να δίνουν ένα ομοιόμορφο αποτέλεσμα και να απομονωθούν τα δεδομένα τα οποία θα οδηγήσουν σε ένα ασφαλές συμπέρασμα. Η Visual Basic για εφαρμογές είναι μια "φιλοξενούμενη" γλώσσα και μέρος της οικογένειας εργαλείων ανάπτυξης της Visual Basic[164]. Ο προγραμματισμός του Excel VBA δεν είναι η μόνη λύση λογισμικού που μπορεί να διαχειριστεί δεδομένα σε αυτή τη μορφή, αλλά είναι ένα από τα προγράμματα που χρησιμοποιούνται ώστε να επιλυθούν διαφορετικά προβλήματα[165]. Πολλές φορές το να γίνει κατανοητό το πλαίσιο εκτέλεσης μιας τέτοιας γλώσσας, είναι αρκετά χαοτικό. Και αυτό διότι η VBA είναι προγραμματική γλώσσα, όπως αλλά χρησιμεύει επίσης ως γλώσσα εκτέλεσης μάκρο-εντολών. Ουσιαστικά, ο δημιουργείται ένας διχασμός στον σαφή προσδιορισμό της διότι το πλαίσιο εκτέλεσης ακολουθεί την εξής διαδικασία: ένα πρόγραμμα γραμμένο σε VBA εκτελείται στο Excel επομένως είναι ένα πρόγραμμα που εκτελείται με τη βοήθεια μιας μακροεντολής[166].

Εκτός από τα βασικά συστατικά μιας λειτουργικής γλώσσας υπολογιστών, δεν υπάρχουν κάποιες σύνθετες έννοιες για να κατανοήσουν τυχόν περίεργη σύνταξη που πρέπει να τηρηθεί, κάνοντας την χρήση της ακόμη πιο ευχάριστη και κατανοητή. Σε συνδυασμό με το Excel, παρέχονται άμεσα διαθέσιμες δυνατότητες εισόδου και εξόδου κατάλληλων δεδομένων- ενώ, σε περίπτωση που είναι απαραίτητο μπορούν να αξιοποιηθούν τα ενσωματωμένα αντικείμενα του Excel, τα στοιχεία οπτικοποίησης, τις συναρτήσεις και τα φύλλα εργασίας για να χρησιμεύσει ως δομή αποθήκευσης δεδομένων[167].

5.6 Microsoft Excel

Το Microsoft Excel είναι ένα πρόγραμμα λογισμικού που δημιουργήθηκε από τη Microsoft. Χρησιμοποιεί υπολογιστικά φύλλα για την οργάνωση αριθμών και δεδομένων με τύπους και συναρτήσεις. Η ανάλυση Excel είναι πανταχού παρούσα σε όλο τον κόσμο και χρησιμοποιείται από επιχειρήσεις όλων των μεγεθών για την εκτέλεση χρηματοοικονομικών αναλύσεων. Το Microsoft Excel είναι μέρος του Microsoft Office [168]. Ένα υπολογιστικό φύλλο Excel ορίζεται ως μια συλλογή στηλών και γραμμών που σχηματίζουν έναν πίνακα. Τα αλφαβητικά γράμματα συνήθως εκχωρούνται σε στήλες και οι αριθμοί συνήθως εκχωρούνται σε σειρές. Το σημείο όπου συναντώνται μια στήλη και μια γραμμή ονομάζεται κελί. Η διεύθυνση ενός κελιού δίνεται από το γράμμα που αντιπροσωπεύει τη στήλη και τον αριθμό που αντιπροσωπεύει μια σειρά.

6

Μεθοδολογία και πειραματικά αποτελέσματα

6.1 Οντότητες και σχέσεις

Ο γράφος ιδιοτήτων (property graph) που δημιουργήθηκε προκειμένου να απεικονίσει το πρόβλημα εύρεσης και σύστασης κατάλληλων ξενοδοχείων ανά περιόδους σύστασης περιλαμβάνει τις βασικές μεταβλητές του αρχείου δεδομένων ως κόμβους αυτού του γράφου.

Οι οντότητες που παρουσιάζονται είναι οι εξής:

- **City** : πόλη στην οποία ανήκει το ξενοδοχείο, με ιδιότητα το όνομα της πόλης
- **Hotel**: ξενοδοχείο με ιδιότητα τον αριθμό του (χρήσιμο για την μοναδικότητα του κάθε ξενοδοχείου και τις επερχόμενες συστάσεις)
- **Profile**: μοναδικό αναγνωριστικό ενός προφίλ χρήστη το οποίο ξεκινά μια συνεδρία (Session). Έχει μοναδικό αριθμό που παραμένει ίδιος μέχρι ο χρήστης να κλείσει εντελώς τη συνεδρία. Ένας χρήστης, μπορεί να κάνει περισσότερες από μια συνεδρίες, εντούτοις στο σύνολο δεδομένων που χρησιμοποιούμε το προφίλ και ο αριθμός συνεδρίας μεταβάλλονται με τον ίδιο τρόπο.
- **Session**: συνεδρία χρήστη με ιδιότητα το μοναδικό της αναγνωριστικό.
- **Properties**: Ιδιότητες των υπαρχόντων ξενοδοχείων. Σε αυτές υπάγεται ένα σύνολο ιδιοτήτων που περιλαμβάνει τις πιθανές παροχές του ξενοδοχείου. Αναφέρονται ενδεικτικά μερικές όπως: δορυφορική τηλεόραση, αστέρια, ρεσεψιόν διαθέσιμη όλο το 24ωρο, καλές αξιολογήσεις, χώρος στάθμευσης, πολυτελές ξενοδοχείο, κατάλληλο για μη καπνίζοντες, προσβάσιμο σε άτομα ΑμεΑ, στεγνωτήρας μαλλιών, τηλεόραση, γήπεδο τένις ή γκολφ, κλιματισμός, κατάλληλο για οικογένειες, δωρεάν ασύρματη πρόσβαση στο ίντερνετ, ψυγείο, ραδιόφωνο, πισίνα, δυνατότητα φύλαξης παιδιών, μπαλκόνι, να επιτρέπονται τα ζώα, παιδική χαρά.

- **Event**: γεγονός που συμβαίνει κατά την περιήγηση του χρήστη εντός της ιστοσελίδας της trivago. Το γεγονός μπορεί να περιλαμβάνει τις τιμές action_type των δεδομένων, όπως εξηγείται αναλυτικά στο σετ δεδομένων της Trivago.
- **Category:** κατηγορία που μπορεί να κατατάσσεται ένα ξενοδοχείο ανάλογα με την τιμή του ανά διανυκτέρευση. Ο κόμβος αυτός μπορεί να έχει τις τιμές Cheap, Medium, Luxurious. Ο κόμβος Cheap περιλαμβάνει τα ξενοδοχεία που έχουν τιμή διανυκτέρευσης μικρότερη από 150, ο κόμβος Medium περιλαμβάνει τα ξενοδοχεία που έχουν τιμή διανυκτέρευσης μεγαλύτερη από 150 και μικρότερη από 850, ενώ ο κόμβος Luxurious περιλαμβάνει τα ξενοδοχεία που έχουν τιμή διανυκτέρευσης μεγαλύτερη από 850.

Οι σχέσεις που εμφανίζονται μεταξύ των οντοτήτων είναι οι ακόλουθες:

- **LOCATES_IN**: Η σχέση αυτή συνδέει τον κόμβο Item με τον κόμβο City. Η λογική πίσω από αυτή τη σχέση περιγράφεται ως εξής: «Κάθε ξενοδοχείο βρίσκεται σε μία πόλη».
- **HAS_PROPERTIES**: Η σχέση αυτή συνδέει τον κόμβο Item με τον κόμβο Properties. Αυτή η σχέση περιγράφεται ως εξής: «Κάθε ξενοδοχείο έχει κάποιες ιδιότητες».
- **CONTAIN**: Η σχέση αυτή συνδέει τον κόμβο Session με τον κόμβο Event. Αυτή η σχέση περιγράφεται ως εξής: «Μια συνεδρία χρήστη περιέχει γεγονότα».
- **HAS_IMPRESSION**: Η σχέση αυτή συνδέει τον κόμβο Profile και Hotel. Αυτή η σχέση περιγράφεται ως εξής: «Ένα προφίλ χρήστη έχει συγκεκριμένες προτάσεις».
- **REFERS**: Η σχέση αυτή συνδέει τον κόμβο Event και τον κόμβο Hotel. Η σχέση περιγράφεται ως εξής: «Ένα συμβάν αναφέρεται σε συγκεκριμένα ξενοδοχεία».
- **SESSION**: Η σχέση αυτή συνδέει τον κόμβο Profile με τον κόμβο Session. Η σχέση περιγράφεται ως εξής: « Ένα προφίλ χρήστη αναφέρεται σε μια συγκεκριμένη συνεδρία». Στο σετ δεδομένων που χρησιμοποιήθηκε, τόσο η οντότητα Profile όσο και ο κόμβος Session μεταβάλλονται με τρόπο ανάλογο. Δεν υπάρχει ίδιο προφίλ χρήστη σε διαφορετική συνεδρία και επίσης δεν υπάρχει διαφορετική συνεδρία σε ίδιο προφίλ χρήστη.
- **HAS_ITEMS**: Η σχέση αυτή συνδέει τον κόμβο Session με τον κόμβο Hotel. Η σχέση περιγράφεται ως εξής: «Μια συνεδρία χρήστη έχει συγκεκριμένα ξενοδοχεία».
- **HAS_CATEGORY**: Η σχέση αυτή συνδέει τον κόμβο Hotel με τον κόμβο Category. Η σχέση περιγράφεται ως εξής: « Ένα ξενοδοχείο συνδέεται με μια συγκεκριμένη κατηγορία».

Εδώ θα παρουσιάσουμε πειράματα αξιολόγησης των τεχνικών μας.

6.2 Περιορισμοί

Οι σχέσεις που αναπτύσσονται μεταξύ των οντοτήτων στη Neo4j είναι μονοσήμαντες [169]. Αυτό σημαίνει ότι οι σχέσεις έχουν μια συγκεκριμένη κατεύθυνση ακριβώς όπως περιεγράφηκαν προηγουμένως. Στη γλώσσα της λογικής, για παράδειγμα μπορούμε για τη σχέση HAS_ITEMS να συμπεράνουμε πως εφόσον «μια συνεδρία έχει συγκεκριμένα ξενοδοχεία» και ότι «κάθε ξενοδοχείο ανήκει σε μια συγκεκριμένη συνεδρία χρήστη», κάτι που όμως δεν αποτυπώνεται ως σχέση κατά την αναπαράσταση του γράφου στη Neo4j.

6.3 Φιλτράρισμα δεδομένων

Τα δεδομένα υπάρχουν σε δύο σετ δεδομένων, item_metadata.csv και test.csv. Και τα δύο αυτά σετ δεδομένων εισήχθησαν σε ένα γράφο και αυτό προκειμένου να διαχειριστούν με το βέλτιστο δυνατό τρόπο. Απομονώθηκαν τα δεδομένα που αναφέρονται μόνο στην Ελλάδα, ώστε να εξαχθούν κατά το δυνατόν ασφαλή συμπεράσματα με τη λιγότερη επεξεργαστική ισχύ. Τα δεδομένα απομονώθηκαν με τη χρήστη της vb στο περιβάλλον Excel. Όσον αφορά τα properties, ενώ υπήρχε αρχική σκέψη να απομονωθούν και να εισαχθούν μόνο τα υποκειμενικά σημαντικότερα για την επιλογή ενός ξενοδοχείου (πλήθος αστεριών, αξιολόγηση, τύπος ξενοδοχείου, χώρους για καπνίζοντες/στάθμευσης) αποτιμώντας το σύνολο το δεδομένων κρίθηκε καλύτερη η εισαγωγή όλων των ιδιοτήτων. Κάθε μία από τις ιδιότητες γράφονται στην ίδια γραμμή χρησιμοποιώντας τον διαχωριστή | προκειμένου να διακριθούν οι διαφορετικές τιμές στα δεδομένα.

6.4 Πειράματα

Οι μεταβλητές που λειτούργησαν ως παράμετροι για τις δοκιμές είναι οι ακόλουθες:

- **walkLength**: όπως αναλύθηκε και όταν περιεγράφηκε ο αλγόριθμος node2vec, είναι ο αριθμός των βημάτων σε έναν μόνο τυχαίο περίπατο. Είναι φανερό, πως για να επιτευχθεί καλύτερος βαθμός ακρίβειας, η τιμή του walklength πρέπει να είναι μικρός, ώστε να μην λαμβάνονται υπόψη ιδιαίτερα απομακρυσμένοι κόμβοι.
- **walksPerNode**: Ο αριθμός των τυχαίων περιπάτων που δημιουργούνται για κάθε κόμβο. Για να επιτευχθεί το ιδανικότερο αποτέλεσμα, είναι σημαντικό ο αλγόριθμος να δοκιμάσει πολλές διαδρομές, προκειμένου να καταλήξει στη βέλτιστη.
- **embeddingDimension**: Μέγεθος των υπολογισμένων ενσωματώσεων κόμβων. Τα ιδανικά embeddings για να εκτελεστεί ο αλγόριθμος είναι αριθμοί γύρω στα δέκα.

- **RELATIONSHIP**: Ο γράφος που απεικονίζει το πρόβλημα της σύστασης των καλύτερων ξενοδοχείων αναπαραστάθηκε με τη βοήθεια ακμών και κόμβων, οντοτήτων και σχέσεων. Οι σχέσεις αυτές μπορούν να αποκτήσουν μεγαλύτερη ή μικρότερη «σημασία» έχοντας μια ιδιότητα βάρους. Την ιδιότητα αυτή του βάρους, τη λαμβάνει ως παράμετρο ο αλγόριθμος κατά την εκτέλεσή του.
- **Weight**: βάρος της σχέσης που λαμβάνεται υπόψη κατά την εκτέλεση του αλγορίθμου.
- **No. Database Results**: Τα αποτελέσματα της βάσης δεδομένων που επιστρέφονται στο cypher query είναι μεγάλης σημασίας για την τιμή της μέσης αμοιβαίας κατάταξης. Κατά την εκτέλεση του συστήματος, υπολογίζονται αντίστοιχα τα 100, 200, 500, 1000 πιο όμοια ξενοδοχεία μεταξύ τους, μέσω της ευκλείδειας απόστασης. Στη συνέχεια, λαμβάνεται το ξενοδοχείο που έχει κάνει κλικ τελικά ο χρήστης και υπάρχει στο σύνολο των δεδομένων και συγκρίνεται με τη λίστα αποτελεσμάτων. Ανάλογα με το πόσο «κοντά» ή μακριά είναι στη λίστα, υπολογίζεται ο MRR.

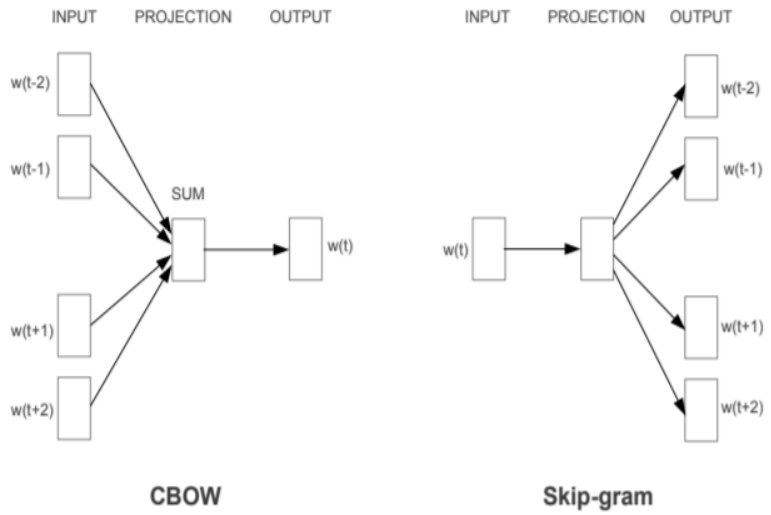
6.5 Αλγόριθμοι εύρεσης πλησιέστερων ξενοδοχείων

6.5.1 Word2vec

Το Word2vec είναι μια τεχνική για την επεξεργασία φυσικής γλώσσας που δημοσιεύτηκε για πρώτη φορά το 2013. Ο αλγόριθμος word2vec χρησιμοποιεί ένα μοντέλο νευρωνικού δικτύου για να βρει συσχετίσεις λέξεων από ένα μεγάλο σώμα κειμένου. Μόλις εκπαιδευτεί, ένα τέτοιο μοντέλο μπορεί να ανιχνεύσει συνώνυμες λέξεις ή να προτείνει πρόσθετες λέξεις για μια πρόταση. Όπως υποδηλώνει το όνομα, το word2vec αντιπροσωπεύει κάθε ξεχωριστή λέξη με μια συγκεκριμένη λίστα αριθμών που ονομάζεται διάνυσμα. Τα διανύσματα επιλέγονται προσεκτικά έτσι ώστε μια απλή μαθηματική συνάρτηση (η ομοιότητα συνημιτόνου μεταξύ των διανυσμάτων) να υποδεικνύει το επίπεδο σημασιολογικής ομοιότητας μεταξύ των λέξεων που αντιπροσωπεύονται από αυτά τα διανύσματα[170] .

Οι ενσωματώσεις μπορούν να ληφθούν χρησιμοποιώντας δύο μεθόδους (και οι δύο περιλαμβάνουν νευρωνικά δίκτυα):τη μέθοδο SkipGram και τη μέθοδο Common Bag Of Words (CBOW). Η μέθοδος CBOW παίρνει το περιεχόμενο κάθε λέξης ως είσοδο και προσπαθεί να προβλέψει τη λέξη που αντιστοιχεί στο πλαίσιο. Για παράδειγμα, έστω ότι έχουμε τη φράση «Σήμερα έχει καλό » και θέλουμε να προβλεφθεί η λέξη καιρό. Η είσοδος στο νευρωνικό δίκτυο είναι η λέξη «καλό». Πιο συγκεκριμένα, χρησιμοποιούμε την κωδικοποίηση της λέξης εισόδου και μετράμε το σφάλμα εξόδου σε σύγκριση με μία κωδικοποίηση της λέξης-στόχου. Στη διαδικασία πρόβλεψης της λέξης- στόχου (καιρό), μαθαίνουμε τη διανυσματική αναπαράσταση της λέξης στόχου. Η είσοδος στις περιπτώσεις και των δύο μοντέλων μπορεί να είναι είτε μία, είτε περισσότερες από μία λέξεις.

Η αρχιτεκτονική της μεθόδου CBOW και της μεθόδου SkipGram (σε περιπτώσεις που εισάγονται περισσότερες από μια λέξεις) φαίνεται στην παρακάτω εικόνα:



Σχήμα 6.1: Αρχιτεκτονική των μεθόδων CBOW/ SkipGram

Το παραπάνω μοντέλο παίρνει λέξεις περιεχομένου C. Το W() χρησιμοποιείται για τον υπολογισμό των εισόδων κρυφών επιπέδων, παίρνουμε έναν μέσο όρο για όλες αυτές τις εισόδους λέξης περιεχομένου C. Έτσι, είδαμε πώς δημιουργούνται αναπαραστάσεις λέξεων χρησιμοποιώντας τις λέξεις περιβάλλοντος. Για να επιτευχθεί το ίδιο αποτέλεσμα είναι δυνατό να χρησιμοποιηθεί απευθείας η λέξη-στόχος (της οποίας την αναπαράσταση θέλουμε να δημιουργήσουμε) για να προβλεφθεί το πλαίσιο και κατά τη διάρκεια, παράγονται οι αναπαραστάσεις.

Το Word2vec δεν είναι το πρώτο, το τελευταίο ή το καλύτερο που χρησιμοποιεί διανυσματικά κενά, ενσωματώσεις, αναλογίες, μετρήσεις ομοιότητας κ.λπ. Αυτό που κάνει αυτόν τον αλγόριθμο να ξεχωρίζει είναι απλός και προσβάσιμος [171, 172]. Το μοντέλο Skip-gram έχει ακριβώς την αντίστροφη αρχιτεκτονική από το μοντέλο CBOW. Και τα δύο έχουν τα δικά τους πλεονεκτήματα και μειονεκτήματα. Σύμφωνα με τον Mikolov, το Skip Gram λειτουργεί καλά με μικρό όγκο δεδομένων και αντιπροσωπεύει καλά τις σπάνιες λέξεις. Από την άλλη πλευρά, το CBOW είναι πιο γρήγορο και έχει καλύτερες αναπαραστάσεις για πιο συχνές λέξεις [173].

6.5.2 Node2vec

Από την εφεύρεση του word2vec, το μοντέλο skip-gram προώθησε σημαντικά την έρευνα ενσωμάτωσης κόμβων, όπως η πρόσφατη εμφάνιση των προσεγγίσεων DeepWalk, LINE, PTE και node2vec [174]. Προκειμένου να προτείνουμε κατάλληλα ξενοδοχεία στους χρήστες, προσπαθούμε να εντοπίσουμε κοινές διαδρομές κατά την επιλογή των κλικ ενός ξενοδοχείου. Αυτό το επιτυγχάνουμε με τη βοήθεια του Node2Vec. Το Node2Vec είναι ένας αλγόριθμος ενσωμάτωσης κόμβου που υπολογίζει μια διανυσματική αναπαράσταση ενός κόμβου βάσει

τυχαίων περιπάτων στο γράφημα. Η γειτονιά χρησιμοποιεί ως δείγμα τυχαίους περιπάτους [175]. Χρησιμοποιώντας αυτή τη σειρά από τυχαία δείγματα, ο αλγόριθμος εκπαιδεύει ένα νευρωνικό δίκτυο κρυφών επιπέδων. Το νευρωνικό δίκτυο είναι εκπαιδευμένο να προβλέπει την πιθανότητα εμφάνισης ενός κόμβου σε μια διαδρομή με βάση την εμφάνιση του σε έναν άλλον κόμβο [176]. Έστω ένα γράφημα γνώσεων K που περιλαμβάνει χρήστες U και στοιχεία I (το αντικείμενο των συστάσεων, π.χ. ξενοδοχεία στη δική μας περίπτωση) και άλλες οντότητες E (αντικείμενα που συνδέονται με στοιχεία, π.χ. πόλη ξενοδοχείου). Το node2vec δημιουργεί διανυσματικές αναπαραστάσεις των χρηστών x_u και των στοιχείων x_i (και των άλλων οντοτήτων x_e). Έτσι, χρησιμοποιείται ως συνάρτηση κατάταξης τη σχετικότητα μεταξύ του χρήστη και των διανυσμάτων στοιχείων: ρ (u, i) = d (xu, xi) όπου d είναι το συνημίτονο της ομοιότητας σε αυτό το έργο.

Στην πρόβλεψη συνδέσμων, μας δίνεται ένα δίκτυο με έναν αριθμό άκρων που έχει αφαιρεθεί, άκρα που θέλαμε να προβλέψουμε τις τιμές τους. Το σύνολο δεδομένων των ακμών δημιουργείται ως εξής: το 50% των άκρων που επιλέχθηκαν τυχαία αφαιρούνται από το δίκτυο διασφαλίζοντας παράλληλα ότι το νέο δίκτυο σχηματίστηκες αφού συνδεθούν οι αφαιρέσεις άκρων και για να δημιουργηθούν αρνητικά παραδείγματα, δειγματοληπτείται τυχαία ίσος αριθμό ζευγών κόμβων από το δίκτυο που δεν έχει άκρη που να τα συνδέει [177].

Στην περίπτωσή μας, χρησιμοποιείται η έτοιμη συνάρτηση gds.beta.node2vec.write που προσφέρεται ως plugin πακέτο της Neo4j. Συγκεκριμένα, κάθε κόμβος αποκτά δικό του embedding ως property γεγονός που μας βοηθά να εντοπίσουμε τους κόμβους με τη μεγαλύτερη ομοιότητα.

Παράμετροι της συνάρτησης:

Πίνακας 6.1: Παράμετροι συνάρτησης gds.beta.node2vec.write

Όνομα	Τύπος μεταβλητής	Προκαθορισμένο	Προαιρετικό	Περιγραφή
walkLength	Ακέραιος	80	Ναι	Ο αριθμός των βημάτων σε έναν μόνο τυχαίο περίπατο.
walksPerNode	Ακέραιος	10	Ναι	Ο αριθμός των τυχαίων περιπάτων που δημιουργούνται για κάθε κόμβο.
inOutFactor	Ακέραιος	10	Ναι	Τάση του τυχαίου περιπάτου να παραμένει κοντά στον κόμβο εκκίνησης ή να βγαίνει στο

				γράφημα. Υψηλότερη τιμή σημαίνει να παραμείνει τοπικά.
returnFactor	Ακέραιος	1.0	Ναι	Τάση του τυχαίου περιπάτου για επιστροφή στον τελευταίο κόμβο που επισκέφτηκε. Μια τιμή κάτω από 1.0 σημαίνει υψηλότερη τάση.
relationshipWeightProperty	Ακέραιος	κενό	Ναι	Όνομα της ιδιότητας σχέσης που θα χρησιμοποιηθεί ως βάρη για να επηρεάσει τις πιθανότητες των τυχαίων περιπάτων. Τα βάρη πρέπει να είναι >= 0. Εάν δεν προσδιορίζεται, ο αλγόριθμος εκτελείται χωρίς στάθμιση.
windowSize	Ακέραιος	10	Ναι	Μέγεθος του παραθύρου περιβάλλοντος κατά την εκπαίδευση του νευρωνικού δικτύου.
negativeSamplingRate	Ακέραιος	5	Ναι	Αριθμός αρνητικών δειγμάτων προς παραγωγή για κάθε θετικό δείγμα.
positiveSamplingFactor	ακέραιος	0.001	Ναι	Παράγοντας που επηρεάζει την κατανομή για θετικά δείγματα. Μια υψηλότερη τιμή αυξάνει την πιθανότητα δειγματοληψίας συχνών κόμβων.
negativeSamplingExponent	ακέραιος	0.75	Ναι	Ο εκθέτης εφαρμόζεται στη συχνότητα του κόμβου για να ληφθεί η αρνητική κατανομή δειγματοληψίας. Μια τιμή 1.0 δειγμάτων ανάλογα με τη συχνότητα. Μια τιμή 0.0 δειγματοληψία κάθε κόμβου εξίσου.
embeddingDimension	ακέραιος	128	Ναι	Μέγεθος των υπολογισμένων ενσωματώσεων κόμβων.

iterations	ακέραιος	1	Ναι	Αριθμός επαναλήψεων εκπαίδευσης.
initialLearningRate	ακέραιος	0.01	Ναι	Ρυθμός εκμάθησης που χρησιμοποιείται για την εκπαίδευση του νευρωνικού δικτύου. Ο ρυθμός εκμάθησης μειώνεται μετά από κάθε επανάληψη της εκπαίδευσης.
minLearningRate	ακέραιος	0.0001	Ναι	Χαμηλότερο όριο για το ρυθμό εκμάθησης καθώς μειώνεται κατά τη διάρκεια της εκπαίδευσης.
randomSeed	ακέραιος	τυχαίο	Ναι	Τιμή για τη γεννήτρια τυχαίων αριθμών που χρησιμοποιείται για τη δημιουργία των τυχαίων περιπάτων.
walkBufferSize	ακέραιος	1000	Ναι	Ο αριθμός των τυχαίων περιπάτων που πρέπει να ολοκληρωθούν πριν από την έναρξη της εκπαίδευσης.

Ο παραπάνω αλγόριθμος είναι μη-ντετερμινιστικός, γεγονός που σημαίνει πως για διαφορετικές τιμές της τυχαίας μεταβλητής, της παραμέτρου randomSeed στην προκειμένη περίπτωση, το αποτέλεσμα θα είναι διαφορετικό. Η τιμή randomSeed, παρότι μπορεί θεωρητικά να οριστεί σε συγκεκριμένη τιμή, στην ουσία και πάλι παράγεται τυχαία και με άγνωστο τρόπο κάθε φορά από τον αλγόριθμο.

6.5.3 Ευκλείδεια απόσταση

Η ευκλείδεια απόσταση είναι μια μετρική που χρησιμοποιείται για τον υπολογισμό της απόστασης δύο διανυσμάτων στον χώρο [178]. Αν $x = x_1, x_2, \dots, x_\sigma$ και $y = y_1, y_2, \dots, y_n$ είναι δύο σημεία του $\mathbb{R}^n$, τότε η ευκλείδεια απόσταση τους $d(x, y)$ είναι το μήκος του ευθύγραμμου τμήματος που τα ενώνει.

$d(x, y) = ||x - y|| = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2$

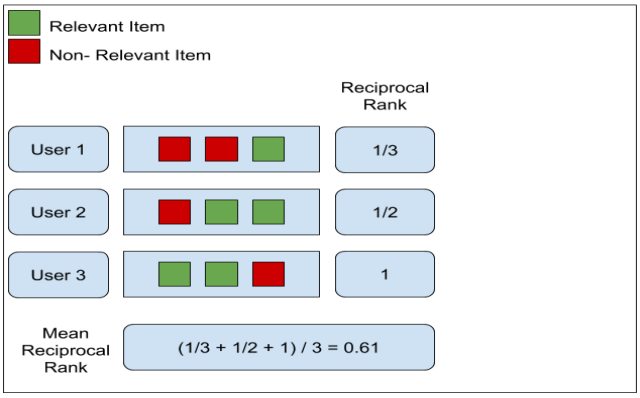
Η απόσταση μεταξύ δύο αντικειμένων που δεν είναι σημεία συνήθως ορίζεται ως η μικρότερη απόσταση μεταξύ ζευγών σημείων από τα δύο αντικείμενα. Οι τύποι είναι γνωστοί για τον υπολογισμό αποστάσεων μεταξύ διαφορετικών τύπων αντικειμένων, όπως η απόσταση από

ένα σημείο σε μια γραμμή. Στα προχωρημένα μαθηματικά, η έννοια της απόστασης έχει γενικευθεί σε αφηρημένους μετρικούς χώρους και έχουν μελετηθεί άλλες αποστάσεις εκτός από την Ευκλείδεια.

6.6 *Πρωτόκολλο αξιολόγησης*

Προκειμένου να υπολογιστεί η ακρίβεια των συστάσεων χρησιμοποιούμε τον δείκτη αξιολόγησης συστάσεων MRR (Mean Reciprocal Rank). Ο δείκτης MRR είναι ένα στατιστικό μέτρο για την αξιολόγηση οποιασδήποτε διαδικασίας που παράγει μια λίστα πιθανών απαντήσεων σε ένα δείγμα ερωτημάτων, ταξινομημένα με βάση την πιθανότητα ορθότητας. Το ερώτημα στην προκειμένη περίπτωση το οποίο καλείται να αξιολογήσει ο αλγόριθμος είναι το πόσο παραγωγικές είναι οι συστάσεις που παράγονται από τον αλγόριθμο node2vec. Με πιο απλά λόγια, το πόσο σωστά προτείνει το σύστημα σύστασης, ξενοδοχείο στον εκάστοτε χρήστη της πλατφόρμας.

Η μέση αντίστροφη κατάταξη είναι ένα στατιστικό μέτρο για την αξιολόγηση οποιασδήποτε διαδικασίας που παράγει μια λίστα πιθανών απαντήσεων σε ένα δείγμα ερωτημάτων, ταξινομημένων με βάση την πιθανότητα. Η αμοιβαία κατάταξη μιας απάντησης ερωτήματος είναι η πολλαπλασιαστική αντίστροφη της κατάταξης της πρώτης σωστής απάντησης: 1 για την πρώτη θέση, 1⁄2 για τη δεύτερη θέση, 1⁄3 για την τρίτη θέση και ούτω καθεξής. Η μέση αμοιβαία κατάταξη είναι ο μέσος όρος των αμοιβαίων βαθμολογιών των αποτελεσμάτων για ένα δείγμα ερωτημάτων Q: όπου αναφέρεται στη θέση κατάταξης του πρώτου σχετικού εγγράφου για το i-ο ερώτημα [179]. Όσο μεγαλύτερη τιμή έχει αυτός ο δείκτης, τόσο καλύτερες αξιολογούνται οι συστάσεις. Το RR είναι 1 εάν στη περίπτωση των συστάσεων ανακτήθηκε στην πρώτη θέση, αν ανακτήθηκε στην θέση 2 είναι 0.5 κ.ο.κ [180].



Σχήμα 6.2: Mean Reciprocal Rank

Η τιμή του δείκτη MRR αλλάζει κάθε φορά που μετακινείται το σχετικό αντικείμενο, αν και η αλλαγή είναι πολύ μεγαλύτερη όταν υπάρχει μετακίνηση από την 1η θέση στην 2η (η αλλαγή είναι 0,5) σε σύγκριση με τη μετάβαση από τη 100η στη 1.000η (μεταβολή 0,009) [173].

Ο αλγόριθμος node2vec όπως αναφέρθηκε είναι μη-ντετερμινιστικός, αυτό σημαίνει πως κατά τη διενέργεια των πειραμάτων η τιμή της μεταβλητής randomSeed, έδινε ένα διαφορετικό αποτέλεσμα. Γι' αυτόν τον λόγο, έγινε η διενέργεια πειραμάτων με τις ίδιες παραμέτρους 5 διαφορετικές φορές. Σε κάθε μία από τις εκτελέσεις αυτές άλλαζε μόνο η τυχαία μεταβλητή.

Η διενέργεια συγκριτικών πειραμάτων με πραγματικά δεδομένα, για διαφορετικές τιμές του αλγορίθμου είχε τα αποτελέσματα στον παρακάτω πίνακα. Οι παράμετροι της συνάρτησης που φάνηκε να είχαν μεγαλύτερη επίδραση στην εκτέλεση των αποτελεσμάτων ήταν οι ακόλουθοι: walksPerNode, embeddingDimension. Ακολούθως, η τιμή της στήλης RELATIONSHIP δείχνει την πιθανότητα ύπαρξης σχέσης με βάρος (στην προκειμένη περίπτωση τοποθεσία ξενοδοχείου), ενώ στη στήλη weight δηλώνεται το βάρος που δίνεται στη συγκεκριμένη σχέση. Στην περίπτωση '*', δηλώνεται πως όλες οι σχέσεις συμμετέχουν με το ίδιο ακριβώς βάρος στη δημιουργία των embeddings, βάσει των οποίων υπολογίζεται η ευκλείδεια απόσταση.

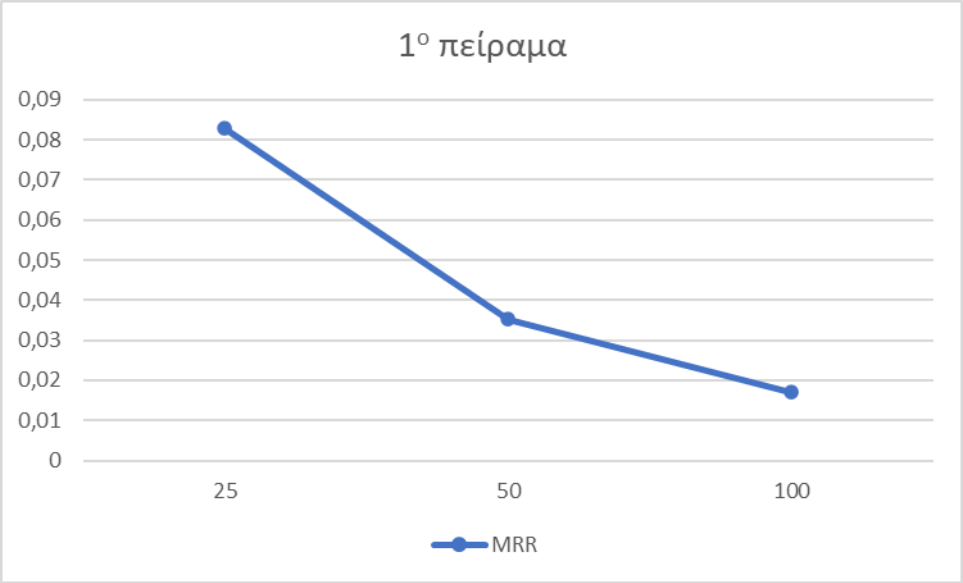
Πίνακας 6.2: Αποτελέσματα εκτέλεσης πειραμάτων

walksPerNode	embeddingDimension	RELATIONSHIP	weight?	No. Database Results	Mean(MRR)
10	10	*	-	100	0,01697173
10	10	*	-	50	0,03526266
10	10	*	-	25	0,08286616
100	10	*	-	100	0,08111297
100	10	*	-	50	0,05857971
100	10	*	-	25	0,11220729
10	10	LOCATES_IN	1.5	100	0,0582817
10	10	LOCATES_IN	1.5	50	0,03673448
10	10	LOCATES_IN	1.5	25	0,06295322
50	10	LOCATES_IN	1.5	100	0,03270654
50	10	LOCATES_IN	1.5	50	0,11034629
50	10	LOCATES_IN	1.5	25	0,05109456

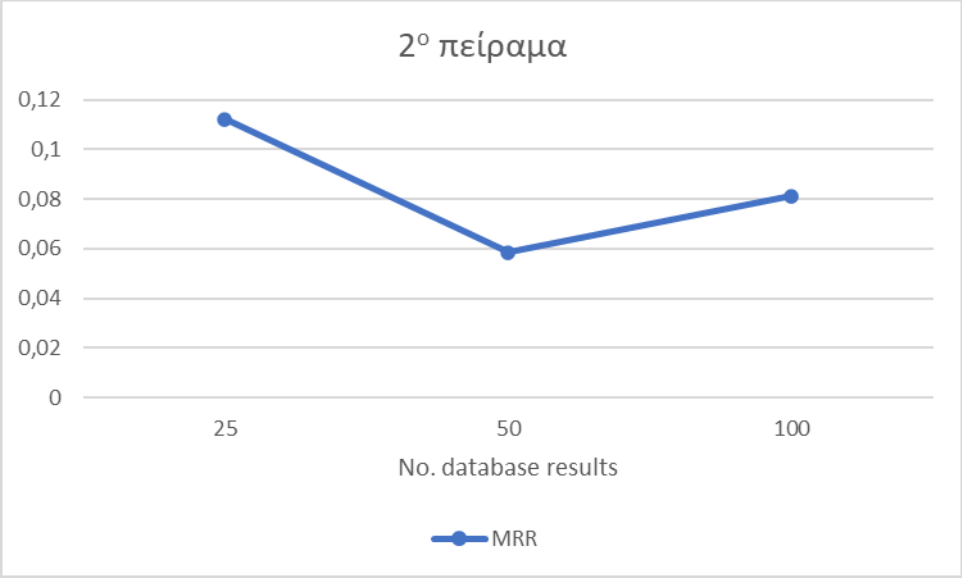
Η στήλη No. Database Results δηλώνει των αριθμό των πιο όμοιων ξενοδοχείων με αυτό που ζητά κάθε φορά ο χρήστης. Στον διαγωνισμό Recsys επιλέγονται τα 25 πιο όμοια ξενοδοχεία με αυτά που κάνει κλικ ο χρήστης, ενώ εμείς επιλέξαμε να αυξήσουμε προς τα πάνω αυτό το νούμερο, καθώς περιοριστήκαμε μόνο στον ελλαδικό χώρο.

Από το πίνακα συνάγουμε πως παρ' όλη τη μη-ντετερμινιστική συμπεριφορά που παρουσιάζει ο αλγόριθμος, έχουμε ως μέσο όρο έναν καλό δείκτη MRR>10%. Αυτό σημαίνει ότι το επόμενο ξενοδοχείο που πραγματικά επιλέγει να δει ο χρήστης είναι μέσα στα δέκα πρώτα που προβλέπει ο αλγόριθμος. Θυμίζουμε ότι το συνολικό πλήθος των ξενοδοχείων είναι πάνω από 3158. Αυτό καταδεικνύει την επιτυχία των προβλέψεων. Μάλιστα, στις περιπτώσεις όπου η τιμή της μεταβλητής randomSeed επιλεχθεί ως πιο «κατάλληλη» για τα πειράματα, έχουμε αποτέλεσμα τιμή που αγγίζει το 25%.

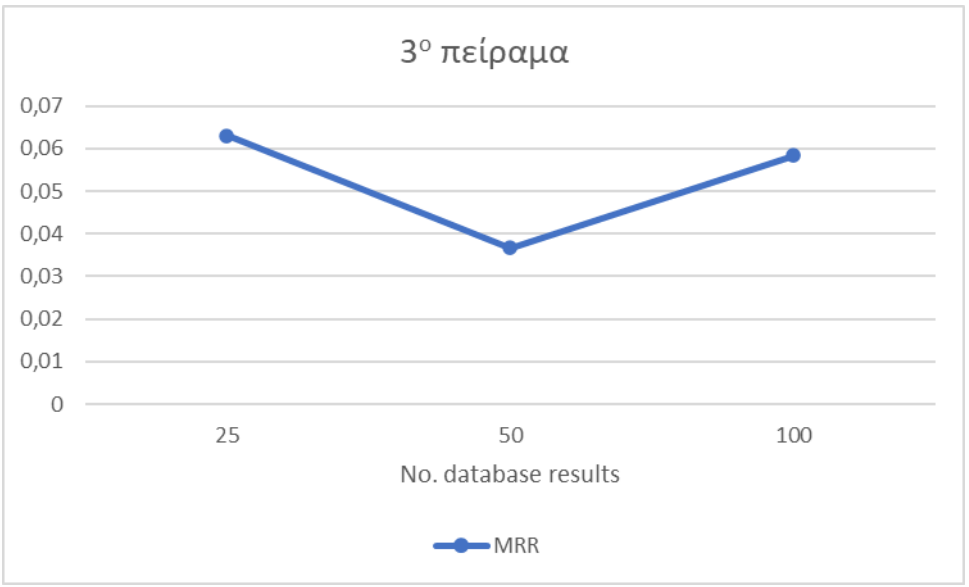
Τέλος, τα πειράματα αναπαριστώνται και γραφικά προκειμένου τα αποτελέσματα να είναι συγκρίσιμα ως προς τις δύο βασικές παραμέτρους των πειραμάτων, του αριθμού των πιο όμοιων ξενοδοχείων που μας επιστρέφει το σύστημα σύστασης και του δείκτη MRR.



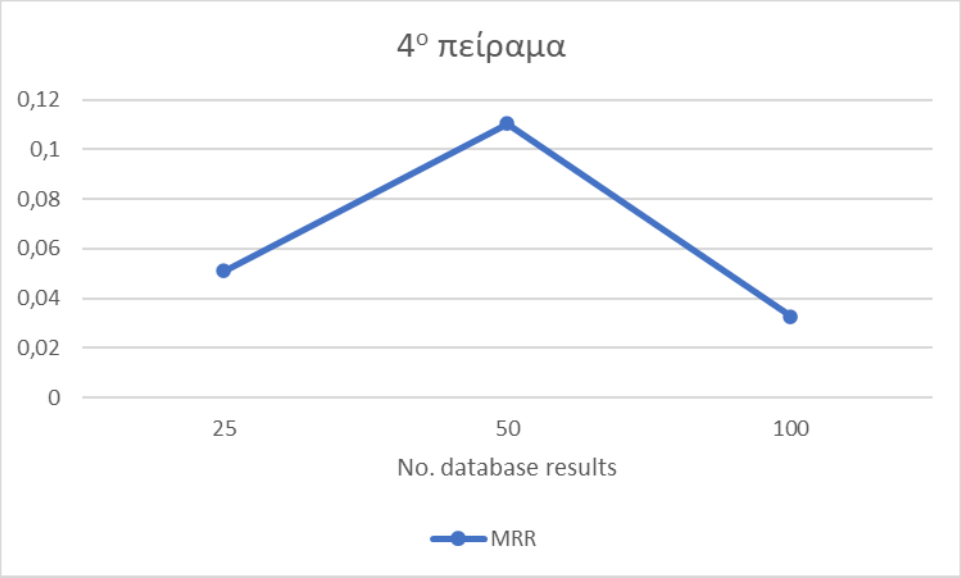
Γράφημα 6.1: Εκτέλεση 1ου πειράματος



Γράφημα 6.2: Εκτέλεση 2ου πειράματος



Γράφημα 6.3: Εκτέλεση 3ου πειράματος



Γράφημα 6.4: Εκτέλεση 4ου πειράματος

7

Επίλογος

7.1 Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική εργασία έγινε αρχικά μια θεωρητική μελέτη των πιο γνωστών συστημάτων σύστασης και των τεχνικών αξιολόγησης τους. Στην συνέχεια, ακολούθησε μια μελέτη για την πρόβλεψη των καταναλωτικών συμπεριφορών των ανθρώπων για την επιλογή του κατάλληλου τουριστικού καταλύματος.

Παρόλο που τα περισσότερα συστήματα συστάσεων εμφανίζουν μια σωρεία προβλημάτων (έλλειψη προφίλ, φυσικοί περιορισμοί, πρόβλημα της ψυχρής εκκίνησης, υποκειμενικότητα επιλογών), τα πλεονεκτήματα ανάπτυξης και χρήσης τέτοιων συστημάτων προσφέρουν μια σειρά πλεονεκτημάτων τόσο για τους καταναλωτές (προβολή κατάλληλων προϊόντων) όσο και για τις επιχειρήσεις (αύξηση κερδών, εστίαση στις ανάγκες των πελατών). Ένα σύστημα σύστασης θεωρείται πετυχημένο όχι όταν προτείνει ακριβώς αυτό που χρειάζεται ο κάθε καταναλωτής, αλλά όταν ένα μεγάλο μέρος των καταναλωτών έχουν κατατάξει στις πρώτες θέσεις προτίμησης το προϊόν που τους προτείνεται. Είναι κατανοητό πως οι προτιμήσεις των ανθρώπων μπορεί να επηρεαστούν από μια σειρά υποκειμενικών παραγόντων, όπως για παράδειγμα προσωπικότητα, ψυχολογία τη δεδομένη στιγμή, συνθήκες περιβάλλοντος σχεδόν οτιδήποτε δηλαδή επηρεάζει τις καταναλωτικές μας συνήθειες και τη λήψη αποφάσεων στην καθημερινή μας ζωή. Αυτό κάνει ακόμη πιο δύσκολη την επιτυχημένη σύσταση, θέτοντας μεγαλύτερες προσδοκίες για την περαιτέρω ανάπτυξη των συστημάτων σύστασης τα επόμενα χρόνια.

Είναι πασιφανές πως η ανάπτυξη ενός τέτοιου πολύπλευρου συστήματος απαιτεί και «ποιοτικά» δεδομένα. Δόθηκε μεγάλη βάση στην προεπεξεργασία των δεδομένων, ώστε να απομονωθούν εκείνα τα οποία δίνουν μεγαλύτερο νόημα στην απεικόνιση, μιας και το τεχνολογικό πλαίσιο που χρησιμοποιήθηκε για την αναπαράσταση των δεδομένων είναι το

Neo4j. Τα δεδομένα, αφορούν μόνο τις κρατήσεις ξενοδοχείων εντός Ελλάδος, προκειμένου τα ξενοδοχεία να απεικονισθούν και να επεξεργαστούν με τον καλύτερο δυνατό τρόπο. Στο πλαίσιο αυτό δημιουργήθηκαν και νέοι κόμβοι (π.χ. ο κόμβος κατάταξης των ξενοδοχείων σε οικονομικά, μέτρια, ακριβά). Μέσω της οπτικοποίησης, δίνεται η δυνατότητα να γίνουν αντιληπτές οι μεταβλητές καθώς και οι σχέσεις που αναπτύσσονται μεταξύ τους.

Ο αλγόριθμος node2vec ήρθε να δώσει τη δική του πινελιά στο σύστημα συστάσεως προτείνοντας τα δύο μεταξύ τους πιο όμοια αντικείμενα. Στη συνέχεια, η θέση του αντικειμένου που επιλέγει τελικά ο χρήστης συγκρίνεται με τη θέση του στον πίνακα των 100 πρώτων συστάσεων που προτείνει το σύστημα σύστασης. Το γεγονός ότι γνωρίζουμε εκ των προτέρων την τελική επιλογή του χρήστη είναι ιδιαίτερα βοηθητικό καθώς έτσι υπολογίζεται η θέση της τελικής επιλογής του χρήστη σε σχέση με τα αντικείμενα τα οποία έχουν προταθεί από τον αλγόριθμο.

Κατά τη διάρκεια ανάπτυξης του συστήματος σύστασης, αντιμετωπίστηκαν προκλήσεις που ήδη αντιμετώπιζαν τα συστήματα σύστασης με κυριότερο την ψυχρή εκκίνηση και φυσικά την έλλειψη ενός σκιαγραφημένου προφίλ. Δεν είναι καθόλου εύκολο, να γίνει πρόταση σε έναν άγνωστο χρήστη, ενώ η σύσταση ξενοδοχείων γίνεται με κριτήριο μόνο την περιοχή και την τιμή που αυτός επέλεξε στην πρώτη του αναζήτηση. Σε αυτό το σημείο μπορεί να γίνει κάποια παραδοχή όπως π.χ. να προτείνονται τα δημοφιλέστερα σε σειρά ξενοδοχεία ή τα ξενοδοχεία με τις καλύτερες κριτικές. Η πρόταση ξενοδοχείων ωστόσο με βάση έναν εξ' ολοκλήρου υποκειμενικό παράγοντα, συνήθως δεν είναι η καλύτερη δυνατή πρακτική. Προκειμένου αυτό το πρόβλημα να ξεπεραστεί, οι συστάσεις ξεκίνησαν από την πρώτη στιγμή που ο χρήστης έκανε κλικ σε ένα αντικείμενο.

Η τελική τιμή της σύστασης υπολογίζεται βάσει του δείκτη MRR είναι μεγαλύτερη από 10%, αυτό σημαίνει πως το σύστημα σύστασης που κατασκευάστηκε έχει αποτελεσματικές συστάσεις. Από υπολογιστικής άποψης, ο αλγόριθμός αποδεικνύεται πολύ αποδοτικός· το πιο βαρύ υπολογιστικό κόστος κάθε επανάληψης είναι αποθήκευση των διασυνδέσεων μεταξύ των γειτόνων κάθε κόμβου $O(a^2|V|)$ ενώ ο αριθμός επαναλήψεων που απαιτείται για την εξαγωγή αποτελεσμάτων βέλτιστης ποιότητας, στα πειράματα αξιολόγησης που πραγματοποιήσαμε, ήταν πάντα πολύ μικρός.

7.2 Μελλοντικές επεκτάσεις

Υπάρχει άραγε το τέλειο σύστημα σύστασης; Η απάντηση είναι πως όχι και αυτό διότι τόσο τα συστήματα σύστασης όσο και η ποιότητα των δεδομένων εξελίσσονται με ραγδαίο τρόπο, γεγονός που έχει ως αποτέλεσμα την κατασκευή καλύτερων συστημάτων σύστασης.

Το σύστημα σύστασης θα μπορούσε να επεκταθεί στο σύνολο των δεδομένων της trivago, ώστε να παρέχει εξειδικευμένες πληροφορίες στο σύνολο των τουριστών και όχι μόνο στους τουρίστες οι οποίοι προτίθενται να επισκεφθούν την Ελλάδα. Το σύστημα σύστασης, μέχρι στιγμής παρέχει τη δυνατότητα παροχής συστάσεων με βάση την πρώτη επιλογή του χρήστη. Αυτό σημαίνει, πως όλα τα ξενοδοχεία έχουν ένα βαθμό ομοιότητας μεταξύ τους. Ο βαθμός ομοιότητας, είναι πολυπαραγοντικός και δεν εξαρτάται καθόλου από το προφίλ του χρήστη, μιας και το παρόν σύνολο δεδομένων δε παρέχει τη δυνατότητα αυτή. Η εφαρμογή trivago παρέχει τη δυνατότητα δημιουργίας προφίλ χρήστη, η οποία μπορεί να εξατομικεύσει ακόμη περισσότερο τις συστάσεις. Με τη σύσταση προφίλ του χρήστη, μπορεί το υπάρχον σύστημα σύστασης να βελτιωθεί ακόμη περισσότερο, καθώς μπορούν να προστεθούν κάποια χαρακτηριστικά/ προτιμήσεις του χρήστη, οι οποίες επηρεάζουν σε μεγάλο αριθμό την επιλογή ξενοδοχείου (π.χ. ταξιδεύει συνήθως με κατοικίδιο ζώο, πάρκινγκ αυτοκινήτου, καπνιστής/ μη καπνιστής).

Εκτός από τις επιπρόσθετες πληροφορίες που θα μπορούσαν να σκιαγραφήσουν καλύτερα το προφίλ του χρήστη, θα μπορούσαν να αποδοθούν λεπτομερέστερα και τα ίδια ξενοδοχεία. Στο πλαίσιο αυτό, είναι επιθυμητή η ύπαρξη κριτικών ώστε τα ξενοδοχεία να κατατάσσονται βαθμολογικά από τους χρήστες. Με τη βαθμολογία των χρηστών θα μπορούσε να προταθεί ξανά ένα ξενοδοχείο που έχει ήδη βαθμολογηθεί καλά από τον χρήστη, ενώ, αντίστοιχα να μην προταθεί ένα που είχε βαθμολογηθεί άσχημα. Οι προσθήκες αυτών των δεδομένων θα εμπλουτίσουν το σύστημα σύστασης και θα αυξήσουν ακόμη περισσότερο την ακρίβειά του.

Σε συνδυασμό με τις πληροφορίες των δεδομένων, μπορούν τα ξενοδοχεία να ταξινομηθούν ως ξενοδοχεία καλοκαιρινών/ χειμερινών διακοπών. Ουσιαστικά κάθε κατάλυμα μπορεί να έχει ένα χαρακτηριστικό περιγραφής του ως κατάλυμα χειμερινών ή καλοκαιρινών διακοπών ανάλογα με την εποχή στην οποία εκτελεί ο χρήστης την αναζήτηση. Για παράδειγμα, είναι πολύ πιθανό σε μια εποχή όπως στα Χριστούγεννα, να προτείνουμε ξενοδοχεία που βρίσκονται κοντά σε μια περιοχή τουριστικού ή εποχικού ενδιαφέροντος. Η εποχικότητα ενός ξενοδοχείου, είναι καλό να εμπλουτίσει το σύνολο δεδομένων καθώς είναι μια πληροφορία η οποία δε μεταβάλλεται, ενώ λαμβάνει υπόψη μια βασική παράμετρο που περνά από το μυαλό του χρήστη όταν βλέπει ξενοδοχεία.

Μια ανάλογη υπηρεσία είναι ήδη διαθέσιμη στο site της trivago, καθώς με το που επισκέπτεται ο χρήστης (ανώνυμος ή συνδεδεμένος) την ιστοσελίδα και κάνει κλικ σε ένα ξενοδοχείο που

τον ενδιαφέρει προτείνονται τα 7 πιο παρόμοια ξενοδοχεία με αυτό που έχει επισκεφθεί. Μια ενδιαφέρουσα πρόταση είναι οι συστάσεις να επεκταθούν σε ολοκληρωμένα πακέτα ταξιδιωτικής σύστασης προορισμών. Αυτό επιτυγχάνεται με το σταδιακό χτίσιμο ενός αντιπροσωπευτικού προφίλ για τον χρήστη, όπου συστήνεται με βάση τα προηγούμενα ταξίδια του, τους τόπους που επιλέγει να ταξιδεύει (εσωτερικό / εξωτερικό), του λόγους για τους οποίους επιλέγει να ταξιδεύει (αναψυχής/ επαγγελματικούς), τις παροχές ξενοδοχείων που θεωρεί απαραίτητες κατά τη διαμονή του (δωμάτιο μη καπνιστών/ πρόσβαση σε άτομα ΑμεΑ/ επιτρέπονται τα κατοικίδια). Μελλοντικές επεκτάσεις άλλωστε και του ίδιου του αλγορίθμου node2vec θα μπορούσαν να περιλαμβάνουν δίκτυα με συγκεκριμένη δομή, όπως ετερογενή δίκτυα πληροφοριών, δίκτυα με ρητά χαρακτηριστικά για κόμβους, ακμές και κατευθυνόμενες σχέσεις.

Όσον αφορά την αξιολόγηση της ποιότητας της σύστασης, ο προτεινόμενος τρόπος αντιμετωπίζει πολλούς περιορισμούς. Οι συστάσεις αξιολογούνται με βάση τη μοναδική ληφθείσα παρατήρηση και το που έκανε κλικ τελικά. Στον πραγματικό κόσμο όμως, είναι πιθανό ο χρήστης να έκανε κλικ στη συγκεκριμένη περιοχή είτε κατά λάθος, είτε γιατί είχε ταξινομήσει στο μυαλό του τη σειρά προτίμησής του στα 5 εμφανιζόμενα της ιστοσελίδας. Η σειρά προτίμησης έρχεται να συμπληρώσει τη «σκιαγράφηση της εικόνας του χρήστη» που αναλύσαμε προηγουμένως. Πώς θα ήταν δυνατό άλλωστε να αποκρυφθεί ένα ξενοδοχείο παρόμοιο με αυτό που έχει επισκεφθεί ο χρήστης και είχε κακή εμπειρία;

Ένα άλλο χαρακτηριστικό των συστημάτων είναι η διαλειτουργικότητα, η αμφίδρομη επικοινωνία που συνήθως έχει ένα σύστημα συστάσεων για διαρκή άντληση δεδομένων. Σε αυτήν την περίπτωση, το υπάρχον σύστημα θα μπορούσε να επικοινωνεί με τα τυχόν κοινωνικά δίκτυα που έχει λογαριασμό ο χρήστης (LinkedIn, Facebook, YouTube) και να αντλεί δυναμικά δεδομένα για επίσκεψη σε κοντινή περιοχή (γεωγραφική τοποθεσία), κάποια χαρακτηριστικά τα οποία δε συμπληρώθηκαν από την εγγραφή του (προσωπικά δεδομένα που δεν έχουν συμπληρωθεί) ή και αν έχει υποστηρίξει σελίδες ξενοδοχείων στα κοινωνικά δίκτυα με την αντίστοιχη εξουσιοδότησή του.

Συνολικά, το έργο που παρουσιάζεται στην παρούσα διπλωματική εργασία υπογραμμίζει το πόσο ωφέλιμο μπορεί να γίνει ένα σύστημα σύστασης για την πρόβλεψη της καταναλωτικής συμπεριφοράς των χρηστών διαδικτυακών υπηρεσιών κράτησης ξενοδοχείου. Τονίζεται επίσης η σημασία ώστε τα δοθέντα δεδομένα από την πλατφόρμα να είναι σωστά οργανωμένα, δομημένα με έναν πιο εύληπτο τρόπο απεικόνισης. Οι γράφοι ιδιοτήτων, συνετέλεσαν στην σαφή και κατανοητή απεικόνιση των δεδομένων ενώ μέσω της ανάπτυξης κατάλληλων σχέσεων, ήταν δυνατή η ανίχνευση παρόμοιων διαδρομών μέσα στον γράφο. Παρά την δυναμική που διαθέτει η προτεινόμενη τεχνική, το πρόβλημα της σύστασης εξακολουθεί να είναι ακανθώδες τόσο στους παρόχους (προϊόντων και υπηρεσιών) όσο και τους ίδιους τους

καταναλωτές. Ο δρόμος της δυναμικής εξέλιξης των συστημάτων σύστασης φέρνει συναρπαστικές προκλήσεις σε έρευνα και ανάπτυξη νέων τεχνικών.

8

Βιβλιογραφία

[1] Ricci, F., Rokach, L., & Shapira, B. (n.d.). *Introduction to Recommender Systems Handbook*. https://doi.org/10.1007/978-0-387-85820-3_1

[2] Davidson, J., Liebald, B., Liu, J., Nandy, P., & Van Vleet, T. (2010). The YouTube video recommendation system. *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, 293–296. https://doi.org/10.1145/1864708.1864770

[3] https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48 (Τελευταία πρόσβαση 11/2020)

[4] Hsu, C. C., Chen, H. C., Huang, K. K., & Huang, Y. M. (2012). A personalized auxiliary material recommendation system based on learning style on Facebook applying an artificial bee colony algorithm. *Computers & Mathematics with Applications*, *64*(5), 1506-1513.

[5] https://www.amazon.science/the-history-of-amazons-recommendation-algorithm (Τελευταία πρόσβαση 11/2020)

[6] Huang, Yin-Fu & Wang, Pei-Lun. (2017). Picture Recommendation System Built on Instagram. 10.1145/3080845.3080868.

[7] https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada (Τελευταία πρόσβαση 11/2020)

[8] Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, *9*(2), 75–79. https://doi.org/10.1145/1345448.1345465d

[9] Sharma, R., & Singh, R. (2016). Evolution of recommender systems from ancient times to modern era: A survey. *Indian Journal of Science and Technology*, *9*(20). https://doi.org/10.17485/ijst/2016/v9i20/88005

[10] Zhao, B. Y., Huang, L., Stribling, J., Rhea, S. C., Joseph, A. D., & Kubiatowicz, J. D. (2004). Tapestry: A resilient global-scale overlay for service deployment. *IEEE Journal on selected areas in communications*, *22*(1), 41-53.

[11] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to Weave an Information tapestry. *Communications of the ACM*, *35*(12), 61–70. https://doi.org/10.1145/138859.138867

[12] Singh, R., & Rani, A. (2017). A survey on the generation of recommender systems. *International Journal of Information Engineering and Electronic Business*, *9*(3), 26.

[13] Hao, M., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011). Recommender systems with social regularization. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, 287–296. https://doi.org/10.1145/1935826.1935877

[14] Muñoz-Organero, M., Ramíez-González, G. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2010). A collaborative recommender system based on space-time similarities. *IEEE Pervasive Computing*, *9*(3), 81–87. https://doi.org/10.1109/MPRV.2010.56

[15] Levandoski, J. J., Sarwat, M., Eldawy, A., & Mokbel, M. F. (2012). LARS: A location-aware recommender system. *Proceedings - International Conference on Data Engineering*, 450–461. https://doi.org/10.1109/ICDE.2012.54

[16] Ghazanfar, M., & Prugel-Bennett, A. (2011). Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution.

[17] Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

[18] Valcarce, D., Landin, A., Parapar, J., & Barreiro, Á. (2019). Collaborative filtering embeddings for memory-based recommender systems. Engineering Applications of Artificial Intelligence, 85, 347–356. https://doi.org/10.1016/j.engappai.2019.06.020

[19] Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

[20] Symeonidis, P., Nanopoulos, A., Papadopoulos, A. N., & Manolopoulos, Y. (2008). Collaborative recommender systems: Combining effectiveness and efficiency. *Expert Systems with Applications*, *34*(4), 2995–3013. https://doi.org/10.1016/j.eswa.2007.05.013

[21] Yu, K., Schwaighofer, A., Tresp, V., Xu, X., & Kriegel, H. P. (2004). Probabilistic Memory-Based Collaborative Filtering. *IEEE Transactions on Knowledge and Data Engineering*, *16*(1), 56–69. https://doi.org/10.1109/TKDE.2004.1264822

[22] https://developers.google.com/machine-learning/recommendation/content-based/basics (Τελευταία πρόσβαση 12/2020)

[23] Recommender Systems: An Introduction - Dietmar Jannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich - Βιβλία Google. (n.d.). Retrieved December 13, 2020, from https://books.google.gr/books?hl=el&lr=&id=eygTJBd_U2cC&oi=fnd&pg=PR5&dq=content+based+filtering+jannach&ots=mWt4c6FRvH&sig=iGbQHDnoenDOpCJIo88iMzPVinU&redir_esc=y#v=onepage&q=content based filtering jannach&f=false

[24] Van Meteren, R., & Van Someren, M. (2000, May). Using content-based filtering for recommendation. In *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop* (Vol. 30, pp. 47-56).

[25] Vanetti, M., Binaghi, E., Carminati, B., Carullo, M., & Ferrari, E. (2010, September). Content-based filtering in on-line social networks. In *International Workshop on Privacy and Security Issues in Data Mining and Machine Learning* (pp. 127-140). Springer, Berlin, Heidelberg.

[26] Beel, J., Langer, S., Nürnberger, A., & Genzmehr, M. (2013). The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems. Lecture Notes in Computer Science, 396–400. doi:10.1007/978-3-642-40501-3_45

[27] https://support.google.com/analytics/answer/2799357?hl=el (Τελευταία πρόσβαση 01/2020)

[28] Beel, J., Langer, S., Nürnberger, A., & Genzmehr, M. (2013, September). The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. In *International Conference on Theory and Practice of Digital Libraries* (pp. 396-400). Springer, Berlin, Heidelberg.

[29] Zhao, X. W., Guo, Y., He, Y., Jiang, H., Wu, Y., & Li, X. (2014, August). We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1935-1944).

[30] Wang, Y., Chan, S. C. F., & Ngai, G. (2012, December). Applicability of demographic recommender system to tourist attractions: a case study on trip

advisor. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 97-101). IEEE.

[31] Εμμανουήλ Πρατικάκης, "Ανάπτυξη ενός συστήματος εξατομικευμένων συστάσεων για ηλεκτρονικές κρατήσεις ξενοδοχείων βασισμένο στη πολυκριτήρια ανάλυση αποφάσεων", Διπλωματική Εργασία, Σχολή Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης, Χανιά, Ελλάς, 2017

[32] Ghazanfar, M. A., & Prugel-Bennett, A. (2010, January). A scalable, accurate hybrid recommender system. In *2010 Third International Conference on Knowledge Discovery and Data Mining* (pp. 94-98). IEEE.

[33] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction, 12*(4), 331-370.

[34] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (n.d.). Hybrid recommendation approaches. *Recommender Systems,* 124–142. doi:10.1017/cbo9780511763113.007

[35] Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web* (pp. 377-408). Springer, Berlin, Heidelberg.

[36] Prasad, R. V. V. S. V., & Kumari, V. V. (2012). A categorical review of recommender systems. *International Journal of Distributed and Parallel Systems, 3*(5), 73.

[37] Tran, T., & Cohen, R. (2000, July). Hybrid recommender systems for electronic commerce. In Proc. Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04, AAAI Press (Vol. 40).

[38] Gunawardana, A., & Meek, C. (2009, October). A unified approach to building hybrid recommender systems. In *Proceedings of the third ACM conference on Recommender systems* (pp. 117-124).

[39] Ghazanfar, M. A., & Prugel-Bennett, A. (2010, January). A scalable, accurate hybrid recommender system. In *2010 Third International Conference on Knowledge Discovery and Data Mining* (pp. 94-98). IEEE.

[40] Bennett, J., & Lanning, S. (2007, August). The netflix prize. In *Proceedings of KDD cup and workshop* (Vol. 2007, p. 35).

[41] Farokhi, N., Vahid, M., Nilashi, M., & Ibrahim, O. (2016). A multi-criteria recommender system for tourism using fuzzy approach. *Journal of Soft Computing and Decision Support Systems, 3*(4), 19-29.

[42] Adomavicius, G., Manouselis, N., & Kwon, Y. (2011). Multi-criteria recommender systems. In *Recommender systems handbook* (pp. 769-803). Springer, Boston, MA.

[43] Hassan, M., & Hamada, M. (2017). A neural networks approach for improving the accuracy of multi-criteria recommender systems. *Applied Sciences*, *7*(9), 868.

[44] Manouselis, N., & Costopoulou, C. (2007). Analysis and classification of multi-criteria recommender systems. *World Wide Web*, *10*(4), 415-441.

[45] Hdioud, F., Frikh, B., & Ouhbi, B. (2013, December). Multi-criteria recommender systems based on multi-attribute decision making. In *Proceedings of international conference on information integration and web-based applications & services* (pp. 203-210).

[46] Καλογριδάκης, Σ. (2019). Ανάπτυξη συστήματος Συστάσεων Πολυκριτήριας Ανάλυσης για την αγορά ή ενοικίαση ακινήτων. Πολυτεχνείο Κρήτης, Χανιά.

[47] Gorakala, S. K., & Usuelli, M. (2015). *Building a recommendation system with R*. Packt Publishing Ltd.

[48] Trewin, S. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information science*, *69*(Supplement 32), 180.

[49] Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., & Tang, J. (2019). Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.

[50] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology, 58(4), 479–493. doi:10.1002/asi.20508

[51] Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, *69*(Supplement 32), 175-186.

[52] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reiterer, S., & Stettinger, M. (2013). *Basic Approaches in Recommendation Systems. Recommendation Systems in Software Engineering*, *15–37*. doi:10.1007/978-3-642-45135-5_2

[53] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

[54] Wang, S., Cao, L., & Wang, Y. (2019). A survey on session-based recommender systems. *arXiv preprint arXiv:1902.04864*.

[55] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

[56] https://livebook.manning.com/book/graph-powered-machine-learning/chapter-6/v-3/ (Τελευταία πρόσβαση: 02/2021)

[57] Ruocco, M., Skrede, O. S. L., & Langseth, H. (2017, August). Inter-session modeling for session-based recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems* (pp. 24-31).

[58] Gharahighehi, A., & Vens, C. (2021). Personalizing diversity versus accuracy in session-based recommender systems. *SN Computer Science*, *2*(1), 1-12.

[59] Ludewig, M., & Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, *28*(4-5), 331-390.

[60] Guo, L., Yin, H., Wang, Q., Chen, T., Zhou, A., & Quoc Viet Hung, N. (2019, July). Streaming session-based recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1569-1577).

[61] Jugovac, M., Jannach, D., & Karimi, M. (2018, September). Streamingrec: a framework for benchmarking stream-based news recommenders. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 269-273).

[62] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019, July). Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 346-353).

[63] Sun, S., Tang, Y., Dai, Z., & Zhou, F. (2019). Self-attention network for session-based recommendation with streaming data input. *IEEE Access*, *7*, 110499-110509.

[64] https://yuspify.com/blog/cold-start-problem-recommender-systems/ (Τελευταία Πρόσβαση 01/2020)

[65] Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065-2073.

[66] http://p-comp.di.uoa.gr/resources/Liapatas_MScThesis.pdf (Τελευταία πρόσβαση 01/2020)

[67] Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065-2073.

[68] Rohani, V. A., Kasirun, Z. M., Kumar, S., & Shamshirband, S. (2014). An effective recommender algorithm for cold-start problem in academic social networks. *Mathematical Problems in Engineering*, *2014*.

[69] Elahi, F. B. M. M. (2019). Cold Start Solutions For Recommendation Systems. IET.

[70] Elahi, M., Braunhofer, M., Ricci, F., & Tkalcic, M. (2013, December). Personality-based active learning for collaborative filtering recommender systems. In *Congress of the Italian Association for Artificial Intelligence* (pp. 360-371). Springer, Cham.

[71] Li, L., & Tang, X. (2015). A Solution to the Cold-Start Problem in Recommender Systems Based on Social Choice Theory. Intelligent and Evolutionary Systems, 267–279. doi:10.1007/978-3-319-27000-5_22

[72] Qiao, R., Yan, S., & Shen, B. (2018, December). A reinforcement learning solution to cold-start problem in software crowdsourcing recommendations. In *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)* (pp. 8-14). IEEE.

[73] Ntoutsi, E., Stefanidis, K., Nørvåg, K., & Kriegel, H.-P. (2012). *Fast Group Recommendations by Applying User Clustering. Lecture Notes in Computer Science, 126–140.* doi:10.1007/978-3-642-34002-4_10

[74] Sharma, M., & Mann, S. (2013). A survey of recommender systems: approaches and limitations. *International Journal of Innovations in Engineering and Technology*, *2*(2), 8-14.

[75] De Gemmis, M., Lops, P., Semeraro, G., & Musto, C. (2015). An investigation on the serendipity problem in recommender systems. *Information Processing & Management*, *51*(5), 695-717.

[76] Abbassi, Z., Amer-Yahia, S., Lakshmanan, L. V., Vassilvitskii, S., & Yu, C. (2009, October). Getting recommender systems to think outside the box. In *Proceedings of the third ACM conference on Recommender systems* (pp. 285-288).

[77] Khusro, S., Ali, Z., & Ullah, I. (2016). Recommender systems: issues, challenges, and research opportunities. In *Information Science and Applications (ICISA) 2016* (pp. 1179-1189). Springer, Singapore.

[78] Liphoto, M., Du, C., & Ngwira, S. (2016, November). A survey on recommender systems. In *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)* (pp. 276-280). IEEE.

[79] Vozalis, E., & Margaritis, K. G. (2003, September). Analysis of recommender systems algorithms. In *The 6th Hellenic European Conference on Computer Mathematics & its Applications* (pp. 732-745).

[80] Dhawan, S. (2019, February). Comparision of Recommendation System Approaches. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 76-78). IEEE.

[81] Lee, J. S., & Zhu, D. (2012). Shilling attack detection—a new approach for a trustworthy recommender system. *INFORMS Journal on Computing*, *24*(1), 117-131.

[82] Zhou, W., Wen, J., Xiong, Q., Gao, M., & Zeng, J. (2016). SVM-TIA a shilling attack detection method based on SVM and target item analysis in recommender systems. *Neurocomputing*, *210*, 197-205.

[83] Cao, J., Wu, Z., Mao, B., & Zhang, Y. (2013). Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system. *World Wide Web*, *16*(5-6), 729-748.

[84] Mishra, N., Chaturvedi, S., Vij, A., & Tripathi, S. (2021, January). Research Problems in Recommender systems. In *Journal of Physics: Conference Series* (Vol. 1717, No. 1, p. 012002). IOP Publishing.

[85] Zhou, T., Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R., & Zhang, Y. C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, *107*(10), 4511-4515.

[86] Ekstrand, M. D., Ludwig, M., Konstan, J. A., & Riedl, J. T. (2011, October). Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 133-140).

[87] Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257-297). Springer, Boston, MA.

[88] Bouneffouf, D., Bouzeghoub, A., & Ganarski, A. L. (2013, November). Risk-aware recommender systems. In *International Conference on Neural Information Processing* (pp. 57-65). Springer, Berlin, Heidelberg.

[89] Calero Valdez, A., Ziefle, M., & Verbert, K. (2016, September). HCI for recommender systems: the past, the present and the future. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 123-126).

[90] Calero Valdez, A., Ziefle, M., & Verbert, K. (2016, September). HCI for recommender systems: the past, the present and the future. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 123-126).

[91] Wang, J., & Tang, Q. (2015). Recommender systems and their security concerns.

[92] Knijnenburg, B. P., & Berkovsky, S. (2017, August). Privacy for recommender systems: tutorial abstract. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 394-395).

[93] Shyong, K., Frankowski, D., & Riedl, J. (2006, June). Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In *International Conference on Emerging Trends in Information and Communication Security* (pp. 14-29). Springer, Berlin, Heidelberg.

[94] Bedi, P., & Agarwal, S. K. (2011, June). Managing security in aspect oriented recommender system. In *2011 International Conference on Communication Systems and Network Technologies* (pp. 709-713). IEEE.

[95] Polatidis, N., Pimenidis, E., Pavlidis, M., & Mouratidis, H. (2017, August). Recommender systems meeting security: From product recommendation to cyber-attack prediction. In *International Conference on Engineering Applications of Neural Networks* (pp. 508-519). Springer, Cham.

[96] Felfernig, A., Friedrich, G., & Schmidt-Thieme, L. (2007). Guest editors' introduction: Recommender systems. *IEEE Intelligent systems*, *22*(3), 18-21.

[97] Almazro, D., Shahatah, G., Albdulkarim, L., Kherees, M., Martinez, R., & Nzoukou, W. (2010). A survey paper on recommender systems. *arXiv preprint arXiv:1006.5278*.

[98] Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018, July). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 974-983).

[99] Zuo, Y., Zeng, J., Gong, M., & Jiao, L. (2016). Tag-aware recommender systems based on deep neural networks. *Neurocomputing*, *204*, 51-60.

[100] Hassan, M., & Hamada, M. (2017). A neural networks approach for improving the accuracy of multi-criteria recommender systems. *Applied Sciences*, *7*(9), 868.

[101] Tan, Q., Liu, N., Zhao, X., Yang, H., Zhou, J., & Hu, X. (2020, April). Learning to hash with graph neural networks for recommender systems. In *Proceedings of The Web Conference 2020* (pp. 1988-1998).

[102] García-Crespo, Á., López-Cuadrado, J. L., Colomo-Palacios, R., González-Carrasco, I., & Ruiz-Mezcua, B. (2011). Sem-Fit: A semantic based expert system to provide recommendations in the tourism domain. *Expert systems with applications*, *38*(10), 13310-13319.

[103] Adamczak, J., Leyson, G. P., Knees, P., Deldjoo, Y., Moghaddam, F. B., Neidhardt, J. & Monreal, P. (2019). Session-based hotel recommendations: Challenges and future directions. *arXiv preprint arXiv:1908.00071*.

[104] Ricci, F. (2020). Recommender systems in tourism. *Handbook of e-Tourism*, 1-18.

[105] Castillo, L., Armengol, E., Onaindía, E., Sebastiá, L., González-Boticario, J., Rodríguez, A., ... & Borrajo, D. (2008). SAMAP: An user-oriented adaptive system for planning tourist visits. *Expert Systems with Applications*, *34*(2), 1318-1332.

[106] Zanker, M., Fuchs, M., Höpken, W., Tuta, M., & Müller, N. (2008). Evaluating recommender systems in tourism–a case study from Austria. In *Enter* (pp. 24-34).

[107] Ricci, F. (2002). Travel recommender systems. *IEEE Intelligent Systems*, *17*(6), 55-57.

[108] Borràs, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, *41*(16), 7370-7389.

[109] Esmaeili, L., Mardani, S., Golpayegani, S. A. H., & Madar, Z. Z. (2020). A novel tourism recommender system in the context of social commerce. *Expert Systems with Applications*, *149*, 113301.

[110] Loh, S., Lorenzi, F., Saldaña, R., & Licthnow, D. (2003). A tourism recommender system based on collaboration and text analysis. *Information Technology & Tourism*, *6*(3), 157-165.

[111] Figueredo, M., Ribeiro, J., Cacho, N., Thome, A., Cacho, A., Lopes, F., & Araujo, V. (2018, March). From photos to travel itinerary: A tourism recommender system for smart tourism destination. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 85-92). IEEE.

[112] Pantano, E., Priporas, C. V., Stylos, N., & Dennis, C. (2019). Facilitating tourists' decision making through open data analyses: A novel recommender system. *Tourism Management Perspectives*, *31*, 323-331.

[113] Neidhardt, J., Seyfang, L., Schuster, R., & Werthner, H. (2015). A picture-based approach to recommender systems. Information Technology & Tourism, 15(1), 49-69.

[114] Pliakos, K., & Kotropoulos, C. (2015). Building an image annotation and tourism recommender system. *International Journal on Artificial Intelligence Tools*, *24*(05), 1540021.

[115] Nemade, G., Deshmane, R., Thakare, P., Patil, M., & Thombre, V. D. (2017). Smart tourism recommender system. *International Research Journal of Engineering and Technology*, *4*(11).

[116] Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1-34). Springer, Boston, MA.

[117] Dareddy, M. R. (2016). Challenges in Recommender Systems for Tourism. In *RecTour@ RecSys* (pp. 59-61).

[118] Jameson, A. (2004, May). More than the sum of its members: challenges for group recommender systems. In *Proceedings of the working conference on Advanced visual interfaces* (pp. 48-54).

[119] Herzog, D., Dietz, L. W., & Wörndl, W. (2019). 6. Tourist trip recommendations–foundations, state of the art, and challenges. In *Personalized Human-Computer Interaction* (pp. 159-182). de Gruyter Oldenbourg.

[120] Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions.

[121] Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *36*, 54(6), 771-785.

[122] Liang, Y., & Chen, N. (2020). A Novel Tourist Attraction Recommendation System Based on Improved Visual Bayesian Personalized Ranking. *Journal homepage:* http://iieta. org/journals/isi, *25*(4), 497-503.

[123] Achmad, K. A., Nugroho, L. E., & Diunaedi, A. (2018, August). Context Based-Tourism Recommender System: Towards Tourists' Context-Sensitive Preference Conceptual Model. In *2018 4th International Conference on Science and Technology (ICST)* (pp. 1-6). IEEE.

[124] https://company.trivago.com/our-story/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[125] https://el.wikipedia.org/wiki/Trivago (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[126] https://recsys.acm.org/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[127] https://recsys.acm.org/recsys22/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[128] Knees, P., Deldjoo, Y., Moghaddam, F. B., Adamczak, J., Leyson, G. P., & Monreal, P. (2019, September). Recsys challenge 2019: Session-based hotel recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 570-571).

[129] https://www.kaggle.com/pranavmahajan725/trivagorecsyschallengedata2019 (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[130] https://recsys.trivago.cloud/challenge/dataset/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[131] She, K. S., Haw, S. C., Loh, Y. G., & Chua, F. F. (2018). AK Tourism: A Property Graph Ontology-based Tourism Recommender System. In *Knowledge Management International Conference (KMICe), July* (pp. 25-27).

[132] Angles, R. (2012, April). A comparison of current graph database models. In *2012 IEEE 28th International Conference on Data Engineering Workshops* (pp. 171-177). IEEE.

[133] Mohamed, M. A., Altrafi, O. G., & Ismail, M. O. (2014). Relational vs. nosql databases: A survey. *International Journal of Computer and Information Technology*, *3*(03), 598-601.

[134] Han, J., Haihong, E., Le, G., & Du, J. (2011, October). Survey on NoSQL database. In *2011 6th international conference on pervasive computing and applications* (pp. 363-366). IEEE.

[135] Leavitt, N. (2010). Will NoSQL databases live up to their promise?. *Computer*, *43*(2), 12-14.

[136] Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.".

[137] Kumar Kaliyar, R. (2015, May). Graph databases: A survey. In *International Conference on Computing, Communication & Automation* (pp. 785-790). IEEE.

[138] Mendelzon, A. O., & Wood, P. T. (1995). Finding regular simple paths in graph databases. *SIAM Journal on Computing*, *24*(6), 1235-1258.

[139] Rodriguez, M. A., & Neubauer, P. (2010). Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology*, *36*(6), 35-41.

[140] Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., & Partner, J. (2015). *Neo4j in action* (Vol. 22). Shelter Island: Manning.

[141] Webber, J. (2012, October). A programmatic introduction to neo4j. In Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity (pp. 217-218).

[142] Lu, H., Hong, Z., & Shi, M. (2017, May). Analysis of film data based on Neo4j. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)* (pp. 675-677). IEEE.

[143] Guia, J., Soares, V. G., & Bernardino, J. (2017, January). Graph Databases: Neo4j Analysis. In *ICEIS (1)* (pp. 351-356).

[144] Drakopoulos, G., Gourgaris, P., & Kanavos, A. (2018). Graph communities in Neo4j. *Evolving Systems*, 1-11.

[145] https://neo4j.com/events/category/meetup/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[146] Webber, J. (2012, October). A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity* (pp. 217-218).

[147] https://neo4j.com/developer/graph-platform/ (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[148] Needham, M., & Hodler, A. E. (2020). A Comprehensive Guide to Graph Algorithms in Neo4j. Neo4j (p. 83).

[149] https://neo4j.com/developer/graph-data-science/ (τελευταία πρόσβαση Ιανουάριος 2022)

[150] Sharma, M., Sharma, V. D., & Bundele, M. M. (2018, November). Performance analysis of RDBMS and no SQL databases: PostgreSQL, MongoDB and Neo4j. In *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)* (pp. 1-5). IEEE.

[151] Stothers, J. A., & Nguyen, A. (2020). Can Neo4j Replace PostgreSQL in Healthcare?. *AMIA Summits on Translational Science Proceedings*, *2020*, 646.

[152] Zhu, Z., Zhou, X., & Shao, K. (2019). A novel approach based on Neo4j for multi-constrained flexible job shop scheduling problem. *Computers & Industrial Engineering*, *130*, 671-686.

[153] Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., ... & Taylor, A. (2018, May). Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data* (pp. 1433-1445).

[154] Castelltort, A., & Laurent, A. (2014, July). Fuzzy queries over NoSQL graph databases: perspectives for extending the cypher language. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 384-395). Springer, Cham.

[155] https://www.howtogeek.com/348960/what-is-a-csv-file-and-how-do-i-open-it/ (τελευταία πρόσβαση Ιανουάριος 2022)

[156] Panzarino, O. (2014). *Learning Cypher*. Packt Publishing Ltd.

[157] Gardener, M. (2012). Beginning R: the statistical programming language. John Wiley & Sons.

[158] Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012, June). Evaluating the design of the R language. In *European Conference on Object-Oriented Programming* (pp. 104-131). Springer, Berlin, Heidelberg.

[159] Jaffar, J., Michaylov, S., Stuckey, P. J., & Yap, R. H. (1992). The CLP (ℛ) language and system. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, *14*(3), 339-395.

[160] Tippmann, S. (2015). Programming tools: Adventures with R. *Nature News*, *517*(7532), 109.

[161] Gentleman, R. (2008). *R programming for bioinformatics*. CRC Press.

[162] Grunsky, E. C. (2002). R: a data analysis and statistical programming environment–an emerging tool for the geosciences. *Computers & Geosciences*, *28*(10), 1219-1222.

[163] Peng, R. D. (2016). *R programming for data science* (pp. 86-181). Leanpub.

[164] Lomax, P. (1998). *VB & VBA in a nutshell: The language*. " O'Reilly Media, Inc.".

[165] Wong, K. W., & Barford, J. P. (2010). Teaching Excel VBA as a problem solving tool for chemical engineering core courses. *Education for Chemical Engineers*, *5*(4), e72-e77.

[166] Walkenbach, J. (2019). Excel VBA Programming. For Dummies.

[167] https://www.morsagmon.com/blog/VBA-As-a-Programming-Language-Pros-and-Cons (Τελευταία πρόσβαση: Απρίλιος 2021)

[168] https://el.wikipedia.org/wiki/Microsoft_Excel (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[169] https://graphaware.com/neo4j/2013/10/11/neo4j-bidirectional-relationships.html (Τελευταία πρόσβαση: Φεβρουάριος 2022)

[170] Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

[171] Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

[172] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[173] Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

[174] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., & Tang, J. (2018, February). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 459-467).

[175] https://neo4j.com/docs/graph-data-science/current/algorithms/node2vec/ (Τελευταία πρόσβαση: Σεπτέμβριος 2021)

[176] Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).

[177] Palumbo, E., Rizzo, G., Troncy, R., Baralis, E., Osella, M., & Ferro, E. (2018, June). Knowledge graph embeddings with node2vec for item recommendation. In *European Semantic Web Conference* (pp. 117-120). Springer, Cham.

[178] Danielsson, P. E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, *14*(3), 227-248.

[179] https://dbpedia.org/page/Mean_reciprocal_rank (Τελευταία πρόσβαση: Νοέμβριος 2021)

[180] Liu, L., & Özsu, M. T. (Eds.). (2009). *Encyclopedia of database systems* (Vol. 6). New York, NY, USA:: Springer.