1. Εισαγωγή στα Συστήματα Συστάσεων

Σύνοψη: *Στο πρώτο κεφάλαιο θα παρουσιάσουμε λεπτομερώς τους στόχους των* ***συστημάτων συστάσεων*** *λαμβάνοντας υπόψη το μεγάλο εύρος των εμπλεκομένων μερών που ενδιαφέρονται είτε να τα προσφέρουν ως υπηρεσία (π.χ. επιχειρήσεις, οργανισμοί κ.λπ.), είτε απλώς να τα χρησιμοποιήσουν (χρήστες). Επίσης, θα προχωρήσουμε σε κατηγοριοποίηση των* ***συστημάτων συστάσεων*** *βάσει των στόχων που εξυπηρετούν (π.χ.* ***συστήματα συνεργατικού φιλτραρίσματος****,* ***διαλογικά συστήματα συστάσεων****,* ***υβριδικά*** *κ.ά.). Τέλος, θα εντάξουμε σε κατηγορίες τους αλγορίθμους συστάσεων βάσει του επιστημονικού πεδίου από το οποίο προέρχονται (π.χ.* ***εξόρυξη δεδομένων****,* ***μηχανική μάθηση****,* ***τεχνητή νοημοσύνη*** *κ.λπ.).*

Προαπαιτούμενη γνώση: *Ο αναγνώστης καλό θα ήταν να έχει στοιχειώδεις γνώσεις για αλγορίθμους* ***Εξόρυξης Δεδομένων****,* ***Ανάκτησης Πληροφορίας*** *και* ***Μηχανικής Μάθησης****.*

1.1 Στόχοι ενός Συστήματος Συστάσεων

1.1.1 Στόχοι που Αφορούν στον Χρήστη

Ο στόχος ενός *συστήματος συστάσεων* είναι να προτείνει στον χρήστη μια λίστα από στοιχεία που μπορεί να τον ενδιαφέρουν. Στη συνέχεια, παραθέτουμε τους σημαντικότερους στόχους που αφορούν στον χρήστη:

1. Η **αποτελεσματικότητα** και η **ακρίβεια** των *συστημάτων συστάσεων* είναι πολύ σημαντική, διότι του επιτρέπει να κάνει τις σωστές επιλογές και να παίρνει τις σωστές αποφάσεις.

2. Το *σύστημα συστάσεων* καλό είναι να προτείνει στον ενδιαφερόμενο, μεταξύ άλλων, και **πρωτότυπα στοιχεία**, τα οποία ως τέτοια θα πρέπει όμως να διαφοροποιούνται σημαντικά τόσο

από την πρότερη εμπειρία του χρήστη με το *σύστημα συστάσεων*, όσο και από τις προηγούμενες αλληλεπιδράσεις του μ' άλλα στοιχεία. Άλλωστε, η ύπαρξη ποικιλομορφίας στα προτεινόμενα στοιχεία κινεί το ενδιαφέρον του χρήστη με αποτέλεσμα αυτός να επανέρχεται στην υπηρεσία συστάσεων.

3. Τέλος, άλλος ένας σημαντικός στόχος των *συστημάτων συστάσεων* είναι η **επεξηγησιμότητα** (*explainability ή interpretability*). Πρόκειται δηλαδή για την ικανότητα ενός *συστήματος συστάσεων* να προτείνει ένα στοιχείο και να αιτιολογεί επαρκώς στον χρήστη τους λόγους που οδήγησαν σ' αυτήν τη σύσταση. Με αυτόν τον τρόπο το *σύστημα συστάσεων* γίνεται "διάφανο" στον χρήστη, γεγονός που τον οδηγεί στο να το εμπιστεύεται και να το χρησιμοποιεί ακόμη περισσότερο.

1.1.2 Στόχοι που Αφορούν στους Παρόχους της Υπηρεσίας Συστάσεων

Οι στόχοι των παρόχων της υπηρεσίας συστάσεων πολλές φορές έρχονται σε αντίθεση με αυτούς του χρήστη (**Promotion vs. Satisfaction**). Αυτό σημαίνει ότι συχνά οι πρώτοι ενδιαφέρονται απλώς να προωθήσουν τα προϊόντα τους χωρίς ταυτόχρονα να τους απασχολεί η ικανοποίηση των ιδιαίτερων αναγκών και προτιμήσεων του δεύτερου. Υπό αυτό το πρίσμα, παραθέτουμε παρακάτω τους σημαντικότερους στόχους των παρόχων της υπηρεσίας συστάσεων:

1. **Αύξηση πωλήσεων (sales increase)** για τα ηλεκτρονικά καταστήματα, δεδομένου ότι τα προτεινόμενα στοιχεία είναι πιθανό να ανταποκρίνονται στις ανάγκες και τις προτιμήσεις του χρήστη.

2. **Αύξηση της διασταυρούμενης πώλησης προϊόντων (product cross-selling)**, δεδομένου ότι ένας χρήστης παίρνει συστάσεις παρόμοιων προϊόντων με αυτά με τα οποία έχει ήδη αλληλεπιδράσει στο παρελθόν.

3. **Αύξηση της ικανοποίησης του χρήστη (user satisfaction)** βάσει του γεγονότος ότι ο χρήστης θα βρίσκει τα προτεινόμενα στοιχεία ενδιαφέροντα και σχετικά, και άρα θα απολαμβάνει τη χρήση της προσφερόμενης υπηρεσίας του *συστήματος συστάσεων*.

4. **Αύξηση εμπιστοσύνης και αφοσίωσης του χρήστη (trust and loyalty)**, δεδομένου ότι οι χρήστες ανανεώνουν τη συνδρομή τους (subscription) στην υπηρεσία δείχνοντας την εμπιστοσύνη και την αφοσίωσή τους.

5. **Αύξηση μέσου χρόνου χρήσης της υπηρεσίας (average time spent on service)**, δεδομένου ότι οι χρήστες απολαμβάνουν τη χρήση της υπηρεσίας και παραμένουν ολοένα και περισσότερο χρόνο σ' αυτήν.

6. **Διατήρηση των υπαρχόντων χρηστών (user retention)**, καθώς οι ικανοποιημένοι χρήστες επιστρέφουν ξανά και ξανά στην υπηρεσία.

1.2 Τύποι Συστημάτων Συστάσεων

1.2.1 Συνεργατικό Φιλτράρισμα

Το *Συνεργατικό Φιλτράρισμα* (*Collaborative Filtering*) είναι μια μέθοδος για την πρόβλεψη των ενδιαφερόντων ενός χρήστη που βασίζεται στη συνδυαστική ανταλλαγή δεδομένων χρήσης και προτίμησης πολλών χρηστών μεταξύ τους. Το *Συνεργατικό Φιλτράρισμα βάσει χρήστη* (*user-based CF*) ([Resnick et al., 1994]) σχηματίζει γειτονιές χρηστών με βάση τη μεταξύ τους ομοιότητα. Συγκεκριμένα, ο αλγόριθμος *user-based CF* υπολογίζει την ομοιότητα των χρηστών με τον *υπό εξέταση χρήστη (target user)* προκειμένου να σχηματίσει τη γειτονιά των κοντινότερων χρηστών του τελευταίου. Στη συνέχεια, ο *user-based CF* συστήνει στον υπό εξέταση χρήστη τα πιο συχνά εμφανιζόμενα στοιχεία μεταξύ των κοντινότερων γειτόνων του.

Ένας άλλος παρόμοιος αλγόριθμος [Sarwar et al., 2001] είναι γνωστός ως *Συνεργατικό Φιλτράρισμα βάσει στοιχείων* (*item-based CF*), ο οποίος και σχηματίζει γειτονιές κοντινότερων στοιχείων με βάση τις ομοιότητες μεταξύ τους. Να σημειωθεί εδώ ότι με τα συστήματα συστάσεων που βασίζονται στο συνεργατικό φιλτράρισμα θα ασχοληθούμε αναλυτικά στο 2ο κεφάλαιο.

1.2.2 Δημογραφικό Φιλτράρισμα

Η βασική ιδέα πίσω από τα *συστήματα συστάσεων δημογραφικού φιλτραρίσματος* (*demographic filtering)* είναι ότι χρήστες με παρόμοια *δημογραφικά χαρακτηριστικά* θα έχουν και παρόμοιες προτιμήσεις για προϊόντα. Ο *αλγόριθμος δημογραφικού φιλτραρίσματος*, λοιπόν, δημιουργεί για κάθε χρήστη ένα διαφορετικό *προφίλ* (*user profile*) βάσει των *δημογραφικών* δεδομένων του. Το *προφίλ χρήστη* δηλαδή είναι επί της ουσίας ένα διάνυσμα του οποίου τα στοιχεία αντιστοιχούν σε *δημογραφικά δεδομένα* του υπό εξέταση χρήστη, όπως η ηλικία, το φύλο, το εκπαιδευτικό επίπεδο, η οικογενειακή κατάσταση, το επάγγελμα, ο τόπος κατοικίας κ.ά. Στη συνέχεια, ακολουθούνται τα ίδια βήματα που περιγράφηκαν παραπάνω για τον *αλγόριθμο συνεργατικού φιλτραρίσματος*. Πιο συγκρεκριμένα, ο *αλγόριθμος δημογραφικού φιλτραρίσματος* υπολογίζει την ομοιότητα των χρηστών με τον υπό εξέταση χρήστη ως προς το *δημογραφικό προφίλ* τους προκειμένου να σχηματίσει τη *γειτονιά των πλησιεστέρων χρηστών* του. Έπειτα, ο *αλγόριθμος δημογραφικού φιλτραρίσματος* παρουσιάζει στον υπό εξέταση χρήστη τα πιο συχνά εμφανιζόμενα στοιχεία μεταξύ των πλησιεστέρων γειτόνων του.

1.2.3 Φιλτράρισμα Βάσει Περιεχομένου

Τα *Συστήματα Συστάσεων Φιλτραρίσματος βάσει Περιεχομένου* (*Content-based Filtering*) παρέχουν συστάσεις στους χρήστες αποκλειστικά με βάση τις αλληλεπιδράσεις που είχαν με τα προϊόντα στο παρελθόν, χωρίς να χρησιμοποιούνται *δεδομένα χρήσης* (*usage data*) από άλλους χρήστες [Mooney and Roy, 2000]. Ο χρησιμοποιούμενος εδώ αλγόριθμος δημιουργεί δηλαδή ένα *προφίλ χρήστη* βάσει των χαρακτηριστικών των στοιχείων με τα οποία αυτός αλληλεπίδρασε στο παρελ-

θόν. Για παράδειγμα, βάσει των ταινιών για τις οποίες ενδιαφέρθηκε ένας χρήστης σ' ένα ανάλογο σύστημα συστάσεων, τα στοιχεία στο *προφίλ χρήστη* αντιστοιχούν σε δεδομένα όπως το είδος της ταινίας, οι ηθοποιοί που συμμετείχαν, οι σκηνοθέτες κ.λπ. Στη συνέχεια, ο *αλγόριθμος φιλτραρίσματος βάσει περιεχομένου* υπολογίζει την ομοιότητα των ταινιών με τον υπό εξέταση χρήστη προκειμένου να σχηματίσει τη *γειτονιά των πλησιεστέρων ταινιών*. Τέλος, ο ίδιος αλγόριθμος συστήνει στον ενδιαφερόμενο τις πιο συχνά εμφανιζόμενες ταινίες μεταξύ των πλησιεστέρων ταινιών του.

Το πλεονέκτημα των *συστημάτων φιλτραρίσματος βάσει περιεχομένου* είναι ότι αμβλύνουν το πρόβλημα της *ψυχρής εκκίνησης* (*cold start*) που αντιμετωπίζουν τα *συστήματα συνεργατικού φιλτραρίσματος*. Αυτό αναφέρεται στα νεοεισερχόμενα στο σύστημα στοιχεία που, εξαιτίας του ότι δεν έχουν προλάβει ακόμη να αλληλεπιδράσουν οι χρήστες μαζί τους, δεν μπορούν και να είναι υποψήφια προς σύσταση. Το μειονέκτημά τους, ωστόσο, είναι ότι δεν μπορούν να παράσχουν διαφοροποιημένες σε σχέση με το *προφίλ του χρήστη* συστάσεις στοιχείων, επειδή ευνοούν μόνο τα στοιχεία που είναι παρόμοια με το ιστορικό αλληλεπιδράσεων του υπό εξέταση χρήστη. Κατατάσσουν δηλαδή σε χαμηλότερη θέση στοιχεία με διαφορετική θεματολογία από το *προφίλ του υπό εξέταση χρήστη*, γεγονός που έχει ως αποτέλεσμα να γίνονται "μονότονες" συστάσεις, χωρίς *ευρηματικότητα* και *ποικιλομορφία*. Για παράδειγμα, εάν ένας χρήστης έχει αλληλεπιδράσει μόνο με ειδησεογραφικά άρθρα που αφορούν στην πολιτική, δεν θα του προταθεί ποτέ κάποιας άλλης κατηγορίας ειδησεογραφικό άρθρο, όπως κάποιο που να αφορά στα αθλήματα ή στη μόδα. Με τα συστήματα συστάσεων που βασίζονται στο *φιλτράρισμα βάσει περιεχομένου* θα ασχοληθούμε αναλυτικά στο 3ο κεφάλαιο.

1.2.4 Συστήματα Γνώσης

Τα συστήματα συστάσεων βασισμένα στη γνώση αναπτύσσονται αποκλειστικά για έναν εξειδικευμένο τομέα υπηρεσιών (π.χ. πώληση κινητών τηλεφώνων, πώληση αυτοκινήτων κ.λπ.), και αξιοποιούν τις γνώσεις που σχετίζονται με αυτόν για τη βελτίωση των συστάσεων. Πιο συγκεκριμένα, αναπαριστούν με κατάλληλες δομές δοδομένων (π.χ. κανόνες συσχέτισης) τη γνώση που σχετίζεται με έναν εξειδικευμένο τομέα υπηρεσιών, και κατόπιν την επαυξάνουν και την επικαιροποιούν σωρευτικά στο πέρασμα του χρόνου, ώστε να μπορούν να συστήνουν τα κατάλληλα προϊόντα στον υπό εξέταση χρήστη όταν δεν υπάρχουν επαρκείς πληροφορίες αλληλεπίδρασης μεταξύ αυτού και των προϊόντων της υπηρεσίας.

Ας εξετάσουμε, λόγου χάρη, την ακόλουθη περίπτωση όπου ένας χρήστης επισκέπτεται για πρώτη φορά ένα ηλεκτρονικό κατάστημα, και άρα δεν υπάρχει το *ιστορικό* των αγορών του για να κάνουμε συστάσεις βασισμένες σε αυτό - τονίζουμε ότι το *ιστορικό* αγορών χρήστη αποτελεί απαραίτητη προϋπόθεση για τις συστάσεις που βασίζονται στο *συνεργατικό φιλτράρισμα* και στο *φιλτράρισμα βάσει περιεχομένου*. Παρόλα αυτά, όμως, το *σύστημα συστάσεων βασισμένο στη γνώση* μπορεί να αξιοποιήσει τις πληροφορίες για τα προϊόντα του ηλεκτρονικού καταστήματος που υπάρχουν διαθέσιμες, και περιλαμβάνουν τις τεχνικές προδιαγραφές και τα χαρακτηριστικά των προϊόντων.

Παράδειγμα 1.1 Έστω, λοιπόν, ότι ένα σύστημα συστάσεων για κινητά τηλέφωνα θα πρέπει να βοηθήσει έναν πελάτη να βρει ένα μοντέλο κινητού τηλεφώνου που να ταιριάζει στις ιδιαίτερες προτιμήσεις του. Με δεδομένο ότι ο μέσος πελάτης αγοράζει ένα κινητό τηλέφωνο μόνο μία φορά κάθε λίγα χρόνια, το σύστημα συστάσεων δεν μπορεί να κατασκευάσει ένα *προφίλ χρήστη* ή να προτείνει στον υπό εξέταση πελάτη κινητά τηλέφωνα που άρεσαν σε άλλους χρήστες. Στον τομέα λοιπόν των κινητών τηλεφώνων, ένα *σύστημα συστάσεων βασισμένο στη γνώση* θα μπορούσε να χρησιμοποιήσει την πρότερη γνώση που έχουν οι πωλητές για τα χαρακτηριστικά των κινητών τηλεφώνων, όπως η ανάλυση της κάμεράς τους, το βάρος τους και η τιμή τους. Επιπλέον, συγκεκριμένοι περιορισμοί θα μπορούσαν να χρησιμοποιηθούν για να οριοθετήσουν το πλαίσιο εντός του οποίου ορισμένα χαρακτηριστικά είναι σημαντικά για τον υπό εξέταση πελάτη ή όχι. Τέλος, θα μπορούσε να ζητηθεί από τον χρήστη η σχετική σημασία των χαρακτηριστικών, όπως για παράδειγμα το εάν η ανάλυση της κάμερας του κινητού τηλεφώνου είναι πιο σημαντική γι' αυτόν από ό,τι το βάρος της συσκευής.

Στο σημείο αυτό να τονιστεί ότι σε πολλά *συστήματα συστάσεων που βασίζονται στη γνώση* οι απαιτήσεις του χρήστη πρέπει να αποκαλυφθούν διαδραστικά μέσα από έναν "διάλογό" του με το σύστημα συστάσεων **(conversational recommender systems)**. Στο παράδειγμα που προηγήθηκε, θα μπορούσε να ερωτηθεί απευθείας ο χρήστης για τις απαιτήσεις του, όπως η μέγιστη τιμή του κινητού τηλεφώνου, η ελάχιστη ανάλυση της κάμερας κ.λπ. Μια τέτοια προσέγγιση, ωστόσο, απαιτεί την ύπαρξη λεπτομερών τεχνικών ανάλυσης και σύγκρισης των χαρακτηριστικών των προϊόντων. Επίσης, προϋποθέτει την ύπαρξη επιπρόσθετου γνωστικού φορτίου σε περιπτώσεις που πρόκειται να αναλυθούν/συγκριθούν προϊόντα (π.χ. αυτοκίνητα) που περιλαμβάνουν έναν πολύ μεγάλο αριθμό χαρακτηριστικών. Συνεπώς, ένα από τα μειονεκτήματα των *βασισμένων στη γνώση συστημάτων συστάσεων* είναι ότι απαιτούν την ύπαρξη *ειδικών (experts)* για τη μοντελοποίηση και επικαιροποίηση της γνώσης του τομέα στον οποίο αυτά εφαρμόζονται, γεγονός που καθιστά δύσκολη τη συντήρηση των συστημάτων αυτών. Το ζήτημα αυτό είναι γνωστό ως **"το πρόβλημα της απόκτησης γνώσης" (knowledge acquisition problem)**.

1.2.5 Υβριδικά Συστήματα

Τα *υβριδικά συστήματα* συνδυάζουν δύο ή περισσότερες μεθόδους συστάσεων για να ξεπεράσουν τα μειονεκτήματα κάθε μεθόδου σύστασης όταν αυτή εφαρμόζεται μεμονωμένα. Στο παρελθόν, αρκετά συστήματα συστάσεων είχαν προτείνει τον συνδυασμό δεδομένων περιεχομένου με δεδομένα χρήσης ([Melville et al., 2002], [Salter and Antonopoulos, 2006]). Συνδυάζοντας για παράδειγμα το *συνεργατικό* με το *φιλτράρισμα βάσει περιεχομένου*, μπορεί κανείς να μειώσει την "αραιότητα" *(data sparsity)* των δεδομένων, που οφείλεται είτε (α) σε νεοεισερχόμενους στο σύστημα συ-

στάσεων χρήστες, οι οποίοι επομένως δεν έχουν ακόμη προλάβει να αλληλεπιδράσουν με στοιχεία, και (β) είτε σε νεοεισερχόμενα στο σύστημα συστάσεων στοιχεία, με τα οποία άρα δεν έχουν προλάβει να αλληλεπιδράσουν οι χρήστες. Επισημαίνεται εδώ ότι η μείωση της *αραιότητας των δεδομένων* μπορεί να οδηγήσει σε αποτελεσματικότερες και ακριβέστερες συστάσεις. Τέλος, σ' ένα *υβριδικό σύστημα συστάσεων* μπορούν να συνδυαστούν και περισσότερες των δύο μεθόδων συστάσεων μεταξύ τους. Για παράδειγμα, το *φιλτράρισμα βάσει περιεχομένου* μπορεί να συνδυαστεί με το αντίστοιχο *δημογραφικό* αλλά και με το *σύστημα συστάσεων που βασίζεται στη γνώση*. Περισσότερες πληροφορίες για τα *υβριδικά συστήματα* δίνονται στην ενότητα 4.6 του αντίστοιχου κεφαλαίου.

1.3 Τύποι Αλγορίθμων Σύστασης

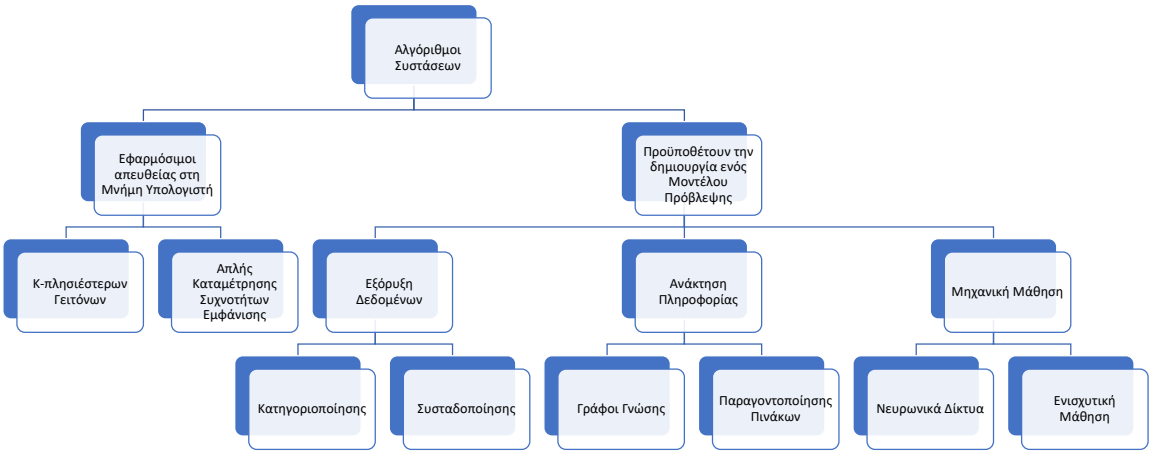
Καθώς το μέγεθος των δεδομένων που χρησιμοποιούμε στην καθημερινή μας ζωή αυξάνεται με ταχείς ρυθμούς, γίνεται όλο και πιο δύσκολο να διαχειριστούμε αυτές τις τεράστιες ποσότητες πληροφοριών. Το φαινόμενο αυτό είναι γνωστό ως **"υπερπληροφόρηση" (information overload)**. Για να αντιμετωπιστεί αυτό το ζήτημα, τα συστήματα συστάσεων λειτουργούν ως φίλτρα που προσπαθούν να μειώνουν τον "πληροφοριακό θόρυβο" και άρα να παρέχουν προτάσεις στοιχείων που πραγματικά ενδιαφέρουν τους χρήστες. Για τον σκοπό αυτό, οι *αλγόριθμοι συστάσεων* εκτελούνται σε τρεις διαδοχικές φάσεις: Πρώτα συλλέγουν τα δεδομένα αλληλεπίδρασης των χρηστών με τα στοιχεία. Στη συνέχεια, "εκπαιδεύουν" μοντέλα ικανά να προβλέπουν συσχετίσεις μεταξύ χρηστών και στοιχείων, και, τέλος, παράγουν μια λίστα σύστασης με προτεινόμενα στοιχεία.

Υπάρχουν αρκετοί αλγόριθμοι που παρέχουν συστάσεις στοιχείων σε χρήστες, όπως για παράδειγμα συστήματα που χρησιμοποιούν μεθόδους *συνεργατικού φιλτραρίσματος*, μεθόδους βασισμένες σε *δισδιάστατους πίνακες* ή *τανυστές*, μοντέλα *βασισμένα σε γράφους* κ.λπ. Όλοι αυτοί οι αλγόριθμοι μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες:

- *Αλγόριθμοι βασισμένοι στη μνήμη.* Αυτοί εκτελούν υπολογισμούς απευθείας σε ολόκληρη τη βάση γνώσης, για να προσδιορίσουν τα προτεινόμενα στοιχεία που θα συσταθούν στον υπό εξέταση χρήστη. Αξίζει να σημειωθεί ότι δεν απαιτούν την ύπαρξη ενός μοντέλου πρόβλεψης και έχει αποδειχθεί ότι επιτυγχάνουν υψηλή αποτελεσματικότητα προβλέψεων.

- *Αλγόριθμοι βασισμένοι σε μοντέλα πρόβλεψης.* Αυτοί προτείνουν στοιχεία στον υπό εξέταση χρήστη, αφού πρώτα "εκπαιδεύσουν" ένα μοντέλο πρόβλεψης. Τέτοιου είδους αλγόριθμοι όμως έχουν την επιπρόσθετη επιβάρυνση της κατασκευής και ενημέρωσης του μοντέλου.

Το Σχήμα 1.1 παρουσιάζει μια οπτική αναπαράσταση των κύριων αλγοριθμικών κατηγοριών στα συστήματα συστάσεων. Όπως φαίνεται στο Σχήμα 1.1, οι οικογένειες αλγορίθμων έχουν προέλθει από διαφορετικά ερευνητικά πεδία και κατευθύνσεις (*εξόρυξη δεδομένων, ευφυή συστήματα, σημασιολογία, νευρωνικά δίκτυα, γράφοι* κ.λπ.). Στην επόμενη ενότητα αναφερόμαστε λεπτομερώς στα πιο αντιπροσωπευτικά γνωστικά πεδία έρευνας.



Σχήμα 1.1: Οπτική απεικόνιση των κύριων αλγοριθμικών κατηγοριών.

1.3.1 Αλγόριθμοι Βασισμένοι στη Μνήμη (memory-based algorithms)

Τεχνικές που Βασίζονται στην Καταμέτρηση (counting-based algorithms)

Οι πιο απλές μέθοδοι είναι αυτές που βασίζονται στην *καταμέτρηση της συχνότητας εμφάνισης των στοιχείων* ή των *αλληλεπιδράσεών* τους με τους χρήστες. Ένα απλό *σύστημα συστάσεων σημείων ενδιαφέροντος (points of interest)*, λόγου χάρη, θα μπορούσε απλώς να μετρήσει τον αριθμό των check-in ανά σημείο, να τα κατατάξει και να συστήσει τα μέρη με τον μεγαλύτερο αριθμό check-in. Τα *συστήματα συστάσεων σημείων ενδιαφέροντος* που λαμβάνουν υπόψη και τις αξιολογήσεις των σημείων ενδιαφέροντος από τους χρήστες μπορούν να παρέχουν συστάσεις με βάση τη μέση αξιολόγηση κάθε τοποθεσίας ή συστάσεις που συνυπολογίζουν τόσο τη συχνότητα των επισκέψεων όσο και τη μέση αξιολόγηση. Τέτοια συστήματα υλοποιούνται πολύ συχνά δεδομένου ότι σε αυτά είναι εύκολο να γίνουν οι ανάλογοι υπολογισμοί και, επιπλέον, λειτουργούν σχετικά καλά σε μικρό όγκο δεδομένων.

Τεχνικές Βασισμένες στην Προβολή (projection-based)

Πολλοί αλγόριθμοι συστάσεων που έχουν σχεδιαστεί για να λειτουργούν σε *δισδιάστατους πίνακες χρηστών-βαθμολογιών* δεν μπορούν να εφαρμοστούν άμεσα σε δεδομένα περισσοτέρων διαστάσεων (βλ. π.χ. την τρισδιάστατη σχέση: χρήστης, σημείο ενδιαφέροντος, δραστηριότητα χρήστη), εκτός εάν οι τριμερείς σχέσεις αναλυθούν σε τρεις δυαδικές σχέσεις (π.χ. χρήστης-σημείο ενδιαφέροντος, χρήστης-δραστηριότητα, και σημείο ενδιαφέροντος-δραστηριότητα). Μετά την *προβολή (projection)*, οι *αλγόριθμοι συνεργατικού φιλτραρίσματος* (*user-based CF* και *item-based CF*) μπο-

ρούν εύκολα να εφαρμοστούν κυρίως στον πίνακα "χρήστης-σημείο ενδιαφέροντος", όπως φαίνεται στο Σχήμα 1.1, αλλά και επιπροσθέτως στους πίνακες "χρήστης-δραστηριότητα" και "σημείο ενδιαφέροντος-δραστηριότητα".

1.3.2 Αλγόριθμοι Βασισμένοι σε Μοντέλα Πρόβλεψης (model-based)

Οι αλγόριθμοι αυτής της κατηγορίας συνήθως αποτελούνται από δύο συνιστώσες:

- Το *μοντέλο πρόβλεψης*, και την
- *Αντικειμενική συνάρτηση*.

Το *μοντέλο πρόβλεψης* διατυπώνει τη σχέση μεταξύ των *δεδομένων εισόδου* (π.χ. χαρακτηριστικά χρηστών, χαρακτηριστικά στοιχείων, αλληλεπιδράσεις χρηστών με τα στοιχεία, άλλα χαρακτηριστικά που αφορούν στο περιβάλλον της αλληλεπίδρασης του χρήστη με τα στοιχεία κ.λπ.) και των *αποτελεσμάτων εξόδου* (προβλέψεις των βαθμολογιών των χρηστών για τα στοιχεία ή την παραγωγή μιας λίστας με προτεινόμενα στοιχεία).

Η *αντικειμενική συνάρτηση* χρησιμοποιείται για την εκπαίδευση του *μοντέλου πρόβλεψης*, ώστε να προσδιοριστούν οι παράμετροι του μοντέλου που τη βελτιστοποιούν. Με άλλα λόγια, η διαδικασία εκπαίδευσης του *μοντέλου πρόβλεψης* επιτυγχάνεται μέσω της βελτιστοποίησης της *αντικειμενικής συνάρτησης*, η οποία και οδηγεί στην εύρεση των καλύτερων παραμέτρων του *μοντέλου πρόβλεψης*.

Μοντέλο Πρόβλεψης

Ένα *μοντέλο πρόβλεψης* της βαθμολογίας ενός χρήστη για ένα στοιχείο ορίζεται ως εξής:

$$\hat{y}_{ui} = \mathcal{F}_{\boldsymbol{\theta}}(u, i)$$

όπου $\hat{y}_{ui}$ η βαθμολογία που προβλέπουμε για λογαριασμό του χρήστη u για το στοιχείο i, και $\boldsymbol{\theta}$ η μεταβλητή η οποία δηλώνει τις παραμέτρους του μοντέλου που πρέπει να μάθουμε από τα δεδομένα εκπαίδευσης. Διαφορετικές επιλογές της συνάρτησης $\mathcal{F}_{\theta}$ αντιστοιχούν σε διαφορετικές υποθέσεις σχετικά με το πώς εξαρτάται η *έξοδος* από τα δεδομένα *εισόδου*. Για παράδειγμα, το *μοντέλο παραγοντοποίησης πινάκων (Matrix Factorization Model - MFM)* μοντελοποιεί τη βαθμολογία $\hat{y}_{ui}$ του χρήστη u για το στοιχείο i ως το γινόμενο των διανυσμάτων $\boldsymbol{v}_u$ και $\boldsymbol{v}_i$, τα οποία αντιπροσωπεύουν αντίστοιχα τον χρήστη και το στοιχείο:

$$\hat{y}_{ui} = \mathcal{F}_{v}(u, i) = \boldsymbol{v}_u^{\top} \boldsymbol{v}_i$$

Το διάνυσμα $\boldsymbol{v}_u$ εκφράζει τις προτιμήσεις του χρήστη u για τα στοιχεία, ενώ το διάνυσμα $\boldsymbol{v}_i$ εκφράζει πόσο πολύ προτιμάται ένα στοιχείο i από τους χρήστες.

Αντικειμενική Συνάρτηση

Οι κύριες *αντικειμενικές συναρτήσεις* για την εκπαίδευση μοντέλων συστάσεων μπορούν να ομαδοποιηθούν σε δύο κατηγορίες: (α) *Πρόβλεψη μιας τιμής* (*point-wise prediction*), και (β) *Πρόβλεψη ζεύγους τιμών* (*pair-wise prediction*). [Η *αντικειμενική συνάρτηση του μοντέλου παραγοντοποίησης πινάκων* (*Matrix Factorization Model*) που αναφέραμε στην προηγούμενη ενότητα ανήκει στην κατηγορία *συναρτήσεων πρόβλεψης μιας τιμής*].

Οι *συναρτήσεις πρόβλεψης ζεύγους τιμών* προσπαθούν να προβλέψουν μια λίστα από διατεταγμένα στοιχεία. Ο στόχος τους είναι να ελαχιστοποιήσουν τη διαφορά μεταξύ της κατάταξης των στοιχείων στην **πραγματική** λίστα προτίμησης στοιχείων -η οποία και εξάγεται από τις αλληλεπιδράσεις του χρήστη με τα στοιχεία της υπηρεσίας στο παρελθόν- και της κατάταξης των στοιχείων στην **προβλεπόμενη** λίστα προτίμησης στοιχείων. Τονίζουμε, εδώ, ότι οι *συναρτήσεις πρόβλεψης ζεύγους τιμών* θεωρούνται συνήθως πιο κατάλληλες για την πρόβλεψη των *N-κορυφαίων προτεινόμενων στοιχείων*. Αντιθέτως, οι *συναρτήσεις πρόβλεψης μια τιμής* θεωρούνται κατάλληλες μόνο για την πρόβλεψη της βαθμολογίας που θα βάλει ο υπό εξέταση χρήστης στα στοιχεία (π.χ. σε μια κλίμακα από 1 έως 5 αστέρια, προβλέπουμε να βαθμολογεί ένα στοιχείο με 4 αστέρια). Σ' αυτό το σενάριο μας ενδιαφέρει η διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής να είναι όσο μικρότερη γίνεται.

Έστω, λοιπόν, ότι μια *αντικειμενική συνάρτηση* στη γενική της μορφή συμβολίζεται με το $f(\cdot)$ και το $\Omega(\boldsymbol{\theta})$ είναι ένας *όρος κανονικοποίησης* (*regularization term*) που ελέγχει την πολυπλοκότητα του μοντέλου και κωδικοποιεί πληροφορίες, όπως είναι η αραιότητα των δεδομένων, η ύπαρξη αρνητικών τιμών κ.λπ. Τότε, η εκπαίδευση του *μοντέλου πρόβλεψης μιας τιμής* **(point-wise loss function)** δίδεται από τον ακόλουθο μαθηματικό τύπο:

$$\sum_{(u,i)\in\mathcal{O}'} f\left(y_{ui}, \hat{y}_{ui}\right) + \lambda \cdot \Omega(\boldsymbol{\theta})$$

όπου λ ένας συντελεστής βαρύτητας της κανονικοποίησης και $\mathcal{O}'$ ένα επαυξημένο σύνολο δεδομένων που περιλαμβάνει εκτός από τα παρατηρούμενα ζεύγη αλληλεπίδρασης χρηστών-στοιχείων και ζεύγη αλληλεπίδρασης χρηστών-στοιχείων που δεν έχουν υπάρξει στην πραγματικότητα (μη παρατηρούμενα). Σημειώνουμε ότι τα μη παρατηρούμενα ζεύγη αλληλεπίδρασης χρηστών-στοιχείων χρησιμοποιούνται στην εκπαίδευση του μοντέλου έτσι ώστε να έχουμε *ισορροπία (balance)* μεταξύ θετικών (παρατηρούμενων) και αρνητικών (μη παρατηρούμενων) αλληλεπιδράσεων μεταξύ χρηστών και στοιχείων.

Όσο, δε, αφορά στη γενική μορφή της *αντικειμενικής συνάρτησης πρόβλεψης ζεύγους τιμών* **(pair-wise loss function)**, αυτή δίνεται με τον παρακάτω τύπο:

$$\sum_{(u,i,j)\in\mathcal{P}} f\Big(y_{uij}, \hat{y}_{uij}\Big) + \lambda \cdot \Omega(\boldsymbol{\theta})$$

όπου $y_{uij} = y_{ui} - y_{uj}$ η διαφορά μεταξύ της πραγματικής βαθμολογίας του χρήστη u για το στοιχείο i (με το οποίο αυτός έχει αλληλεπιδράσει) και της βαθμολογίας του για το στοιχείο j (με το οποίο αυτός δεν έχει αλληλεπιδράσει), $\hat{y}_{uij} = \hat{y}_{ui} - \hat{y}_{uj}$ η διαφορά μεταξύ της προβλεπόμενης βαθμολογίας του χρήστη u για το στοιχείο i (με το οποίο αυτός έχει αλληλεπιδράσει) και της βαθμολογίας του για το στοιχείο j (με το οποίο αυτός δεν έχει αλληλεπιδράσει), και $\mathcal{P}$ ένα σύνολο από τριπλέτες που επιλέγονται δειγματοληπτικά από το $\mathcal{O}'$, καθεμία από τις οποίες περιλαμβάνει έναν χρήστη u και ένα ζεύγος στοιχείων i και j όπου συνήθως το i είναι ένα θετικό στοιχείο (παρατηρούμε αλληλεπίδραση του χρήστη με το στοιχείο) και το j είναι ένα αρνητικό στοιχείο (δεν παρατηρούμε αλληλεπίδραση του χρήστη με το στοιχείο).

Τόσο για τις *αντικειμενικές συναρτήσεις πρόβλεψης ζεύγους τιμών* όσο και για τις *αντικειμενικές συναρτήσεις πρόβλεψης μιας τιμής*, η επιλογή του τύπου της *αντικειμενικής συνάρτησης* $f(\cdot)$ είναι σημαντική. Παρακάτω παραθέτουμε μερικές ευρέως χρησιμοποιούμενες *αντικειμενικές συναρτήσεις*:

- *Square Loss*: $f^{SL}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$,
- *Log Loss*: $f^{LL}(y, \hat{y}) = \log(1 + \exp(-y \cdot \hat{y}))$,
- *Cross-Entropy Loss*: $f^{CEL}(y, \hat{y}) = -y \log(p) - (1 - y) \log(1 - p)$, όπου $p = \sigma(\hat{y}) = 1/(1 + \exp(-\hat{y}))$.

Περισσότερες πληροφορίες για τις *αντικειμενικές συναρτήσεις πρόβλεψης μιας τιμής* και *πρόβλεψης ζεύγους τιμών* θα δοθούν στο 5ο κεφάλαιο.

1.3.3 Αλγόριθμοι Εξόρυξης Δεδομένων

Εξόρυξη Δεδομένων Παγκόσμιου Ιστού (*Web Data mining*) είναι η διαδικασία εξαγωγής γνώσης από μεγάλους όγκους δεδομένων του Διαδικτύου. Ένας κοινά αποδεκτός ορισμός για την *εξόρυξη δεδομένων Παγκόσμιου Ιστού* είναι ο ακόλουθος:

Ορισμός 1.1 *Εξόρυξη Δεδομένων Παγκόσμιου Ιστού* είναι η εξαγωγή ενδιαφέρουσας, μη προφανούς, νέας και χρήσιμης πληροφορίας από δεδομένα δομής του Παγκόσμιου Ιστού, από δεδομένα περιεχομένου ιστοσελίδων και από δεδομένα χρήσης των χρηστών.

Η εξαγόμενη πληροφορία που αναφέρεται στον Ορισμό 1.1 μπορεί να δοθεί με διάφορες μορφές. Ένας κοινά χρησιμοποιούμενος όρος για αυτήν (ανεξαρτήτως της μορφής της) είναι το *πρότυπο* (*pattern*). Πρέπει να σημειωθεί ότι στον προηγούμενο ορισμό έμφαση δίνεται στην εφαρμογή της *διαδικασίας εξόρυξης* σε μεγάλο όγκο δεδομένων. Αυτό προκύπτει, όπως προαναφέρθηκε, από τις σύγχρονες ανάγκες και αποτελεί το κίνητρο για την ανάπτυξη νέων αποτελεσματικών όσο και αποδοτικών τεχνικών που θα μπορούν να αντιμετωπίζουν αυτόν τον παράγοντα. Επομένως, η *εξόρυξη δεδομένων* έχει βοηθήσει στην ανάπτυξη αλγορίθμων που μπορούν να εφαρμοστούν σε *τεράστιους όγκους δεδομένων* (*big data*). Αυτοί οι αλγόριθμοι δεν φορτώνουν όλα τα δεδομένα μονομιάς στη μνήμη του υπολογιστή (όπως κάνουν οι απλοί *αλγόριθμοι στατιστικής*), αλλά, απεναντίας, φορτώνουν τμηματικά τα δεδομένα από τον σκληρό δίσκο στην κύρια μνήμη του υπολογιστή.

Παρότι η *εξόρυξη δεδομένων Παγκόσμιου Ιστού* χρησιμοποιεί πολλές τεχνικές δανεισμένες από την παραδοσιακή εξόρυξη δεδομένων, δεν ταυτίζεται μ' αυτήν εξαιτίας της *ετερογενούς* και *αδόμητης* μορφής των δεδομένων του Ιστού. Πρόσφατα αναπτύχθηκαν *τεχνικές εξόρυξης γνώσης* από διάφορους τύπους δεδομένων Παγκόσμιου Ιστού (π.χ. περιεχομένου, χρήσης και δομής).

Εξόρυξη γνώσης από δεδομένα δομής: Ανακαλύπτει ενδιαφέρουσα γνώση από το δίκτυο των υπερσυνδέσμων που αποτελούν τη δομή του Ιστού. Έτσι, από τους υπερσυνδέσμους μπορούμε να ανακαλύψουμε σημαντικές ιστοσελίδες. Κάτι που εφαρμόζεται ήδη από τις μηχανές αναζήτησης. Έχουμε επίσης τη δυνατότητα να εντοπίσουμε κοινότητες χρηστών που μοιράζονται κοινά ενδιαφέροντα.

Εξόρυξη γνώσης από δεδομένα χρήσης: Αναφέρεται στην ανακάλυψη προτύπων που σχετίζονται με τις προτιμήσεις του χρήστη. Τα πρότυπα αυτά εντοπίζονται από δεδομένα που έχει εισαγάγει ο χρήστης ή από την πλοήγησή του σε μια ιστοσελίδα.

Εξόρυξη γνώσης από δεδομένα περιεχομένου: Αφορά στην εξόρυξη προτύπων από τα περιεχόμενα των ιστοσελίδων, τα οποία και αντιμετωπίζονται ως κείμενο. Μπορούμε, άρα, αυτομάτως να κατηγοριοποιήσουμε και να ομαδοποιήσουμε ιστοσελίδες βάσει της θεματολογίας τους.

Τεχνικές Εξόρυξης Δεδομένων

Διαφορετικά είδη προτύπων είναι δυνατόν να εξαχθούν αναλόγως του τύπου δεδομένων και της χρησιμοποιούμενης *τεχνικής εξόρυξης* τους. Κατά συνέπεια, τα διαφορετικά είδη προτύπων μπορούν να αποκαλύψουν διαφορετικού είδους γνώση. Επομένως, είναι απαραίτητη η ύπαρξη διαφορετικών *τεχνικών εξόρυξης* που να οδηγούν σε διαφορετικά είδη προτύπων. Στη συνέχεια περιγράφονται δύο σημαντικές τεχνικές:

Κατηγοριοποίηση (classification). Με την *κατηγοριοποίηση* σχηματίζεται ένα μοντέλο κατάταξης των δεδομένων σ' ένα προκαθορισμένο σύνολο κλάσεων. Το μοντέλο σχηματίζεται μέσω των τιμών που έχουν οι ιδιότητες των αντικειμένων. Ένα μέρος των δεδομένων χρησιμοποιείται για τη φάση εκπαίδευσης (σχηματισμός του μοντέλου), και τα υπόλοιπα για την αποτίμησή του. Με τη δεύτερη πληροφορούμαστε για την ποιότητα του μοντέλου, δηλαδή την ικανότητα κατάταξης δεδομένων αγνώστου κλάσης.

Ένας από τους σημαντικούς τύπους μοντέλων είναι τα *δέντρα απόφασης* (*decision trees*), τα οποία χρησιμοποιούνται ευρύτατα σε εφαρμογές εξόρυξης λόγω της αποτελεσματικότητας (χαμηλό ποσοστό λάθους κατηγοριοποίησης) και της καλής κλιμάκωσής τους. Ένας εξίσου σημαντικός *αλγόριθμος κατηγοριοποίησης* είναι αυτός των *k κοντινότερων γειτόνων* (*k nearest neighbors - kNN*), ο οποίος βασίζεται στη χρήση μέτρων σύγκρισης της ομοιότητας μεταξύ των χρηστών ή στοιχείων. Η *τεχνική των κοντινότερων γειτόνων* υποθέτει ότι ολόκληρο το σύνολο των δεδομένων εκπαίδευσης συμπεριλαμβάνει, εκτός από τα δεδομένα του συνόλου, και την επιθυμητή κλάση για κάθε αντικείμενο. Κατά συνέπεια, το σύνολο των δεδομένων εκπαίδευσης αποτελεί και το μοντέλο κατάταξης.

Άρα, προκειμένου να γίνει η *κατηγοριοποίηση* ενός νέου αντικειμένου, πρέπει προηγουμένως να έχει μετρηθεί η απόστασή του από κάθε αντικείμενο του συνόλου δεδομένων εκπαίδευσης. Στο τέλος, το νέο αντικείμενο θα ενταχθεί σε μια ομάδα με τους k *κοντινότερους γείτονες* από το σύνολο των δεδομένων εκπαίδευσης.

Συσταδοποίηση (clustering). Εφόσον το σύνολο των κλάσεων δεν είναι προκαθορισμένο, τότε μπορούμε -σε αντίθεση με ό,τι συμβαίνει στην *κατηγοριοποίηση*- να ομαδοποιήσουμε τα δεδομένα με βάση κάποιο μέτρο ομοιότητας (ή απόστασης). Κατά αυτόν τον τρόπο, η *συσταδοποίηση* επιδιώκει την επίτευξη της κατά το δυνατόν μεγαλύτερης ομοιότητας μεταξύ των δεδομένων της ίδιας ομάδας και της κατά το δυνατόν μικρότερης ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων. Καθιερωμένοι *αλγόριθμοι συσταδοποίησης* είναι οι *διαιρετικοί* (π.χ k*-means*) και οι *ιεραρχικοί* (*hierarchical clustering*).

1.3.4 Αλγόριθμοι Ανάκτησης Πληροφορίας

Η *Ανάκτηση Πληροφορίας (Information Retrieval)* έχει να προσφέρει πολλές τεχνικές για την αναπαράσταση, αποθήκευση και οργάνωση δεδομένων του Διαδικτύου, καθώς και για την πρόσβαση σε αυτά. Βασικός στόχος της *Ανάκτησης Πληροφορίας* είναι η διευκόλυνση πρόσβασης του χρήστη σε πληροφορίες που τον ενδιαφέρουν.

Αναλυτικότερα, οι ιστοσελίδες του Παγκόσμιου Ιστού περιέχουν κυρίως κείμενο. Προκειμένου τα κείμενα να αποκτήσουν έναν επεξεργάσιμο μορφότυπο, πρέπει να υπάρξουν κάποια στάδια "κατεργασίας" τους. Αρχικά, αφαιρούνται από τα κείμενα οι τετριμμένες λέξεις (σύνδεσμοι, αντωνυμίες κ.λπ.), και αναγνωρίζονται οι ρίζες των λέξεων (*stemming*). Στη συνέχεια, αναπαριστώνται τα κείμενα σε διανύσματα, έτσι ώστε να μπορούν να εφαρμοστούν μέτρα ομοιότητας διανυσμάτων. Τέλος, εξάγωνται τα χαρακτηριστικά γνωρίσματα των κειμένων κρατώντας τελικά για κάθε κείμενο τις χαρακτηριστικότερες λέξεις-όρους. Τα τρία αυτά στάδια που αναφέραμε θα αναλυθούν λεπτομερέστερα στη συνέχεια.

Προεπεξεργασία Περιεχομένου Ιστοσελίδων

Η προεπεξεργασία του περιεχομένου των ιστοσελίδων αποτελείται από το στάδιο της αφαίρεσης των τετριμμένων λέξεων και από το στάδιο της αναγνώρισης των ριζών των λέξεων.

Όσον αφορά στο πρώτο στάδιο, εύκολα μπορεί κάποιος να παρατηρήσει το ότι υπάρχουν πολλές λέξεις σ' ένα κείμενο που εάν συμπεριληφθούν στα χαρακτηριστικά γνωρίσματά του, τότε θα λειτουργήσουν ως "θόρυβος". Τέτοιες λέξεις είναι τα άρθρα, οι σύνδεσμοι κ.λπ. Για την απομάκρυνση αυτών των λέξεων υπάρχουν αλγόριθμοι αναγνώρισης των συντακτικών μερών μιας πρότασης που μπορούν να εφαρμοστούν σ' όλο το κείμενο.

Όσον αφορά στο δεύτερο στάδιο, η ύπαρξη των ίδιων λέξεων σε διαφορετικές πτώσεις ή χρόνους μπορεί να εμποδίσει τον σωστό υπολογισμό της ομοιότητας μεταξύ δύο κειμένων. Η πιο δια-

δεδομένη μέθοδος για την αναγνώριση των ριζών των λέξεων είναι η *stemmer* που προτάθηκε από τον Martin Porter [Porter, 1980].

Αναπαράσταση Ιστοσελίδων

Η αναπαράσταση των ιστοσελίδων γίνεται με τη δημιουργία ενός διανυσματικού χώρου όπου κάθε ιστοσελίδα αποτελεί και ένα διάνυσμα. Ο βασικός τρόπος αναπαράστασης του κειμένου μιας ιστοσελίδας βασίζεται στον αριθμό εμφανίσεων μιας λέξης σε αυτό. Έτσι, κάθε λέξη αντιστοιχίζεται με μια διαφορετική διάσταση του διανυσματικού χώρου.

Σύμφωνα με τα προηγούμενα, δύο διαφορετικές λέξεις θα έχουν πάντα μηδενική ομοιότητα, παρότι πολλές φορές διαφορετικές λέξεις μπορεί να έχουν την ίδια σημασία (*συνωνυμία*). Για να αντιμετωπισθεί *το πρόβλημα της συνωνυμίας* έχει προταθεί η μέθοδος της *Λανθάνουσας Σημασιολογικής Δεικτοδότησης* (*Latent Semantic Indexing)* [Furnas et al., 1988]. Η βασική ιδέα αυτού του αλγορίθμου είναι ότι δεν είναι αναγκαίες όλες οι διαστάσεις του διανυσματικού χώρου λόγω των σημασιολογικών ομοιοτήτων που παρουσιάζουν πολλές λέξεις. Με το *LSI* όμως θα ασχοληθούμε αναλυτικά στο 5ο κεφάλαιο.

Εξαγωγή Χαρακτηριστικών Γνωρισμάτων από Ιστοσελίδες

Μετά την αναπαράσταση του κειμένου μιας ιστοσελίδας στον διανυσματικό χώρο, ακολουθεί η εξαγωγή των σημαντικότερων γνωρισμάτων της. Έτσι, για κάθε κείμενο ιστοσελίδας θα πρέπει να αναδειχθούν: (i) εκείνοι οι όροι που περιγράφουν καλύτερα το κείμενο, και, (ii) εκείνοι οι όροι που το διακρίνουν καλύτερα από τα άλλα. Το πρώτο σύνολο των όρων παρέχει ποσοτικοποίηση της ομοιότητας των όρων ενός κειμένου, ενώ το δεύτερο ποσοτικοποίηση της ανομοιότητας μεταξύ των κειμένων. Το δημοφιλέστερο μέτρο για την ποιότητα εξαγωγής χαρακτηριστικών γνωρισμάτων από κείμενα είναι η *Συχνότητα Όρου-Αντίστροφη Συχνότητα Κειμένου* (*Term Frequency-Inverse Term Frequency - TF-IDF*) [Baeza-Yates and Ribeiro-Neto, 1999]. Στο 3ο κεφάλαιο θα ασχοληθούμε αναλυτικά με το *TF-IDF*.

Παραγοντοποίηση Πινάκων

Η *Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing - LSI)* έχει χρησιμοποιηθεί εκτενώς στην *Ανάκτηση Πληροφορίας*, για να ανιχνεύσει τις λανθάνουσες σημασιολογικές σχέσεις μεταξύ εγγράφων και των λέξεων που αυτά εμπεριέχουν. Η τεχνική *LSI* δημιουργεί μια χαμηλής τάξης προσέγγιση του αρχικού *δισδιάστατου πίνακα εγγράφων-όρων (document-term)*. Συνεπώς, παράγει έναν πίνακα με λιγότερο "θόρυβο" στα δεδομένα και επομένως καλύτερο από τον πρωτότυπο ως προς την προβλεπτική του αξία. Με αυτόν τον τρόπο δημιουργούνται έννοιες ανωτέρου επιπέδου -που μπορούν δηλαδή αποτελεσματικότερα να χαρακτηρίσουν τα έγγραφα- από ό,τι οι απλές λέξεις/όροι.

Οι Furnas, Deerwester και οι λοιποί συν-συγγραφείς [Furnas et al., 1988] πρότειναν αρχικά τη μέθοδο *LSI* στο επιστημονικό πεδίο της *Ανάκτησης Πληροφορίας*. Η μέθοδος *LSI* χρησιμοποιεί την *τεχνική SVD* για να φέρει στην επιφάνεια λανθάνουσες (= κρυμμένες) σημασιολογικές συνάφειες και συσχετίσεις μεταξύ των εγγράφων και των λέξεων που αυτά εμπεριέχουν. Πιο συγκεκριμένα, η *SVD* είναι μια πολύ γνωστή τεχνική ανάλυσης πίνακα, η οποία παραγοντοποιεί έναν πίνακα σε τρεις. Ο Berry και οι λοιποί συν-συγγραφείς [Berry et al., 1994] διεξήγαγαν έρευνα και ισχυρίσθηκαν ότι η *τεχνική LSI* δημιουργεί μια προσέγγιση του αρχικού πίνακα που όμως έχει μικρότερες διαστάσεις και άρα εμπεριέχει λιγότερο "θόρυβο" στα δεδομένα από ό,τι ο πίνακας με τα αρχικά δεδομένα.

Στα συστήματα συστάσεων, ο Sarwar και οι συνεργάτες του [Sarwar et al., 2000] εφάρμοσαν πρώτοι τη μέθοδο της *απομείωσης των διαστάσεων του πίνακα βαθμολογιών χρήστη-στοιχείου*. Χρησιμοποίησαν επίσης τη *μέθοδο SVD* για τη δημιουργία προβλέψεων σχετικά με τις προτιμήσεις των χρηστών πάνω σε *στοιχεία/αντικείμενα*. Επίσης, ο Goldberg και οι συνεργάτες του [Goldberg et al., 2001] εφάρμοσαν off-line *την Ανάλυση Πρωτευουσών Συνιστωσών* (*Principal Component Analysis - PCA*) στον πίνακα χρηστών-στοιχείων με αποτέλεσμα να επιτύχουν ταχύ online υπολογισμό των προτεινόμενων στοιχείων. Περισσότερες, ωστόσο, πληροφορίες για την ανάλυση και παραγοντοποίηση πινάκων θα δώσουμε στο 5ο κεφάλαιο.

Γράφοι Γνώσης

Οι *γράφοι γνώσης* αποτελούν μια βασική δομή για την αναπαράσταση της πληροφορίας στα συστήματα γνώσης που περιγράψαμε στην Ενότητα 1.2.4. Τονίζεται ότι ένας εναλλακτικός τρόπος αναπαράστασης της γνώσης είναι οι *κανόνες συσχέτισης* (*association rules*) που επιτρέπουν την *εξαγωγή συμπερασμάτων* (*rule inference*) βάσει της *συνεπαγωγικής λογικής* (*reasoning logic*) που εφαρμόζεται σε αυτούς. Από την άλλη, οι *γράφοι γνώσης* αποθηκεύουν συνήθως διασυνδεδεμένα δεδομένα (π.χ. οντότητες από τον πραγματικό κόσμο), γεγονότα και αφηρημένες έννοιες, ενώ παράλληλα κωδικοποιούν τη σημασιολογία των σχέσεων μεταξύ των διασυνδεμένων οντοτήτων. Τελευταία, μετά την ανάπτυξη του *Σημασιολογικού Ιστού* (*Semantic Web*), οι *γράφοι γνώσης* συνδέονται συχνά και με **"διασυνδεδεμένα ανοικτά δεδομένα Διαδικτύου" (interconnected open Web data)**, εστιάζοντας στις συνδέσεις μεταξύ εννοιών και διασυνδεμένων οντοτήτων.

Ορισμός 1.2 Ένας *γράφος γνώσης*: (α) ορίζει υπερ-κλάσεις/υπο-κλάσεις εννοιών και τις μεταξύ τους σχέσεις με μια ιεραρχική δομή που ονομάζεται "οντολογία", (β) εμπεριέχει στιγμιότυπα οντοτήτων του πραγματικού κόσμου και τις μεταξύ τους σχέσεις οργανωμένες σ' έναν γράφο δεδομένων, (γ) επιτρέπει τη δυνητική διασύνδεση οντοτήτων μεταξύ τους σε διαφορετικά επίπεδα αφαίρεσης ή ανάλυσης της ορισθείσας "οντολογίας", και (δ) καλύπτει έναν πολύ συγκεκριμένο θεματικό τομέα.

Συνοπτικά, ένας *γράφος γνώσης* περιγράφει τη σημασιολογία των σχέσεων μεταξύ των οντοτήτων και επιτρέπει τη διασύνδεσή τους σε διαφορετικά επίπεδα ανάλυσης της οντολογίας. Με αυτόν τον τρόπο επιτρέπει τη *λογική εξαγωγή συμπερασμάτων* (*logic inference*) για την ανακάλυψη νέας γνώσης (π.χ. η εξαγωγή συμπεράσματος ανακαλύπτει και δημιουργεί, μέσω απλών συνεπαγωγών, νέους συνδέσμους μεταξύ των οντοτήτων του γράφου που δεν ήταν καταγγεγραμμένοι πριν και δεν ήταν αποθηκευμένοι στη βάση γνώσης). Αντιθέτως, σ' έναν *απλό γράφο δεδομένων* (που δεν είναι δηλαδή και *γράφος γνώσης*) μπορούμε μόνο να υποβάλλουμε ερωτήματα που περιορίζονται στην ανάκτηση της υπάρχουσας πληροφορίας (π.χ. αναζήτηση και ανάκτηση των συνδέσμων που συσχετίζουν τις οντότητες μεταξύ τους και που ήταν ήδη αποθηκευμένοι στη βάση γραφοδεδομένων).

Προκειμένου να αξιοποιηθούν οι *γράφοι γνώσης* σε διάφορες εφαρμογές *μηχανικής μάθησης*, εξάγονται από αυτούς κατάλληλα διαμορφωμένα διανύσματα γνωστά και ως **"ενσωματώσεις των κόμβων" (node embeddings)**. Με αυτές επιτυγχάνεται η εξαγωγή των χαρακτηριστικών και των σχέσεων των οντοτήτων σε κατάλληλα μετασχηματισμένη μορφή που επιτρέπει να εφαρμοστούν *αλγόριθμοι μηχανικής μάθησης* για τη μέτρηση της *σημασιολογικής ομοιότητας* μεταξύ των οντοτήτων/κόμβων του *γράφου γνώσης*. Μ' αυτόν τον τρόπο μπορεί κανείς να υπολογίσει τις ομοιότητες μεταξύ χρηστών ή στοιχείων προκειμένου να τις χρησιμοποιήσει για το σύστημα συστάσεων.

Στη σημερινή εποχή, *γράφοι γνώσης* χρησιμοποιούνται από μηχανές αναζήτησης όπως η *Google*, και η *Bing*, καθώς και από βοηθούς έξυπνων συσκευών όπως η *Siri* της *Apple* και η *Alexa* της *Amazon*. Περισσότερες πληροφορίες για τους *γράφους γνώσης* θα δοθούν στο 8ο κεφάλαιο.

1.3.5 Αλγόριθμοι Μηχανικής Μάθησης

Η *μηχανική μάθηση* είναι ένα γνωστικό πεδίο της *τεχνητής νοημοσύνης* που έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν και να γίνονται αποτελεσματικότερες μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Οι σχετικοί αλγόριθμοι, λοιπόν, μπορούν να μαθαίνουν (χωρίς να έχουν προγραμματιστεί μ' έναν ντετερμινιστικό τρόπο) από τα δεδομένα και να κάνουν επιτυχημένες προβλέψεις. Παραδείγματα εφαρμογών της *μηχανικής μάθησης* αποτελούν η *υπολογιστική όραση* (*computer vision*), *τα αυτόνομα κινούμενα οχήματα* (*self-driven cars*), καθώς και οι *προσωπικοί βοηθοί* (*Alexa, Siri,* κ.λπ.) των έξυπνων κινητών τηλεφώνων που προαναφέραμε και οι οποίοι στηρίζονται στην *επεξεργασία φυσικής γλώσσας* (*natural language processing*).

Νευρωνικά Δίκτυα

Ένα *νευρωνικό δίκτυο* (*neural network*) αποτελείται από κατάλληλα *διασυνδεδεμένους κόμβους (νευρώνες)* οι οποίοι σχηματίζουν μια συγκεκριμένη αρχιτεκτονική δομή που αποτελείται από τουλάχιστον τρία ή περισσότερα επίπεδα (ένα *εισόδου*, ένα *εξόδου* και ένα ή περισσότερα κρυφά επίπεδα). Εδώ να τονίσουμε ότι δεν υπάρχει σύνδεση μεταξύ των *νευρώνων* του ιδίου επιπέδου και

συνήθως ούτε και μεταξύ αυτών που ανήκουν σε μη διαδοχικά επίπεδα. Τα δεδομένα εισάγονται από τους *κόμβους εισόδου* και οδηγούνται μέσω των κόμβων των κρυφών επιπέδων στους *κόμβους εξόδου*. Επειδή ένα επίπεδο δεν μπορεί να έχει απεριόριστους *νευρώνες*, προσθέτουμε επιπλέον επίπεδα προκειμένου να μπορούμε να αυξήσουμε τον αριθμό τους, αυξάνοντας έτσι και τον αριθμό των παραμέτρων του μοντέλου πρόβλεψης, γεγονός που συνήθως ενισχύει την αποτελεσματικότητά του.

Ο τρόπος σύνδεσης μεταξύ των *νευρώνων* του κάθε επιπέδου προσδιορίζει την αρχιτεκτονική και τον τύπο του *νευρωνικού δικτύου* (π.χ. *Multi-Layer Perceptron - MLP, Convolutional Neural Network - CNN, Recurrent Neural Network - RNN* κ.λπ.). Ο στόχος ενός νευρωνικού δικτύου είναι να βρεθεί το κατάλληλο *βάρος (weight)* για κάθε *σύνδεσμο (link)* που συνδέει δυο *νευρώνες* μεταξύ τους, με απώτερο στόχο τη βελτιστοποίηση μιας *αντικειμενικής συνάρτησης* που ορίζουμε στην *έξοδό* του.

Κάθε *νευρώνας* δέχεται ένα σύνολο αριθμητικών εισόδων είτε από άλλους *νευρώνες* είτε από το περιβάλλον των χρηστών και των στοιχείων που τους προτείνονται. Με βάση αυτές τις εισόδους κάνει έναν υπολογισμό και παράγει μία έξοδο. Ειδικότερα, κάθε κόμβος ενσωματώνει εσωτερικά μια *συνάρτηση ενεργοποίησης* (*activation function*) η οποία δέχεται το άθροισμα των αριθμητικών εισόδων και παράγει μια τιμή εξόδου, συνήθως στο διάστημα [0,1].

Την τελευταία πενταετία όλες οι μεγάλες επιχειρήσεις τεχνολογίας (*Google, Amazon, Alibaba* κ.λπ.) υλοποιούν τις εφαρμογές και τις υπηρεσίες τους σχεδόν αποκλειστικά με τη χρήση *νευρωνικών δικτύων πολυάριθμων επιπέδων* λόγω της ανάπτυξης νέων δομών δεδομένων και σύγχρονων αλγορίθμων που μπορούν πια να εφαρμοστούν αποδοτικά σ' αυτές. Η τάση αυτή είναι γνωστή ως *deep learning (βαθιά μάθηση)* εξαιτίας του πλήθους των κρυφών επιπέδων του *νευρωνικού δικτύου*, αλλά και της εξαντλητικής ανάλυσης και επεξεργασίας των δεδομένων εκπαίδευσης. Περισσότερες πληροφορίες για τη *μηχανική μάθηση μέσω νευρωνικών δικτύων* θα παρουσιάσουμε στο 6ο κεφάλαιο.

Ενισχυτική Μάθηση

Η *Ενισχυτική Μάθηση* (*Reinforcement Learning*) προσπαθεί να "εκπαιδεύσει" ένα μοντέλο μάθησης μέσα από την άμεση διάδραση του συστήματος λογισμικού με το περιβάλλον των χρηστών και των στοιχείων που τους προτείνονται χωρίς όμως να καθοδηγείται το σύστημα για το ποια ενέργεια θα πρέπει να ακολουθήσει. Αντιθέτως, προσπαθεί να ανακαλύψει από μόνο του εκείνες τις ενέργειες που θα του αποφέρουν μακροπρόθεσμα το μεγαλύτερο όφελος.

Αναλυτικότερα, ο **"πράκτορας" (agent)** είναι μια οντότητα λογισμικού που συλλέγει πληροφορίες από το **"περιβάλλον" (environment)** των χρηστών και των στοιχείων με τα οποία αυτοί αλληλεπιδρούν. Αρχικά, ο *πράκτορας* χρησιμοποιεί τις πληροφορίες αυτές για να αντιληφθεί την **"τρέχουσα κατάσταση" (state)** του προφίλ του υπό εξέταση χρήστη βάσει των αλληλεπιδράσεών του με τα στοιχεία. Στη συνέχεια, ο *πράκτορας*, ανάλογα με την τρέχουσα κατάσταση του υπό εξέταση χρήστη, επιλέγει ένα, μεταξύ περισσοτέρων, υποψήφιο προς σύσταση στοιχείο. Στα συστήματα συστάσεων τα υποψήφια προς σύσταση στοιχεία είναι οι *δυνητικές ενέργειες* **(action)** μεταξύ

των οποίων μπορεί να επιλέξει ο *πράκτορας* δεδομένης της τρέχουσας κατάστασης του προφίλ του χρήστη. Το κατά πόσον ο *πράκτορας* πετυχαίνει τους στόχους του δηλώνεται από κάποια *επιβράβευση* **(reward)**. Επομένως ο *πράκτορας* δρα με τέτοιον τρόπο ώστε να μεγιστοποιήσει το άθροισμα των επιβραβεύσεων από τις συστάσεις των στοιχείων που προτείνει στον υπό εξέταση χρήστη. Στους *αλγορίθμους ενισχυτικής μάθησης* όμως θα εμβαθύνουμε στο 7ο κεφάλαιο.

1.4 Ερωτήσεις Κεφαλαίου

1. Ποιοι είναι οι αντικειμενικοί στόχοι ενός συστήματος συστάσεων από την πλευρά του χρήστη και ποιοι από την πλευρά του παρόχου της υπηρεσίας; Επιχειρηματολογήστε για το κατά πόσο οι αντικειμενικοί στόχοι των δύο πλευρών μπορεί κάποιες φορές να συγκρούονται μεταξύ τους.

2. Αναφέρετε τους τύπους συστημάτων συστάσεων και απαριθμήστε τα πλεονεκτήματα και τα μειονεκτήματά τους.

3. Ποιες είναι οι διαφορές μεταξύ των αλγορίθμων συστάσεων που βασίζονται στη *μνήμη* και εκείνων που βασίζονται σ' ένα *μοντέλο πρόβλεψης*;

4. Ποια είναι η διαφορά μεταξύ των αλγορίθμων που βασίζονται στην *απλή καταμέτρηση* και εκείνων που προϋποθέτουν προηγουμένως την εφαρμογή της πράξης της *προβολής* προκειμένου να επιλεχθούν συγκεκριμένα χαρακτηριστικά από τα δεδομένα;

5. Ποιες είναι οι κύριες διαφορές μεταξύ των *αλγορίθμων εξόρυξης δεδομένων* και των *αλγορίθμων ανάκτησης πληροφορίας*;

6. Ποιες είναι οι κύριες διαφορές μεταξύ των *αλγορίθμων μηχανικής μάθησης* και των *αλγορίθμων ανάκτησης πληροφορίας*;

7. Ποια είναι τα οφέλη που μπορεί να προκύψουν από ένα *υβριδικό σύστημα συστάσεων*;

8. Ποια είναι η σχέση μεταξύ των *συστημάτων συστάσεων βάσει γνώσης* και των *γράφων γνώσης*;

9. Ποιες κατηγορίες *αντικειμενικών συναρτήσεων* γνωρίζετε για τα συστήματα συστάσεων και ποιο πρόβλημα επιδιώκει να βελτιστοποιήσει η καθεμία;

10. Είναι η *αποτελεσματικότητα* (ως προς την *ακρίβεια*) ενός συστήματος συστάσεων αντιστρόφως ανάλογη με την *καινοτομία* των προτεινόμενων στοιχείων; Αιτιολογήστε την απάντησή σας.

Βιβλιογραφία

R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM press New York, 1999.

M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573--595, 1994.

G. W. Furnas, S. Deerwester, S. T. Durnais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings ACM SIGIR Conference*, pages 465--480, 1988.

K. Goldberg, T. Roeder, T. Gupta, and C. Perkins. Eigentaste: a constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133--151, 2001.

P. Melville, R. J. Mooney, and N. R. Content-boosted collaborative filtering for improved recommendations. In *Proceedings AAAI Conference*, pages 187--192, 2002.

R. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings ACM DL Conference*, pages 195--204, 2000.

M. F. Porter. An algorithm for suffix stripping. *Program Journal*, 1980.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering on netnews. In *Proccedings Conference Computer Supported Collaborative Work*, 1994.

J. Salter and N. Antonopoulos. Cinemascreen recommender agent: Combining collaborative and content-based filtering. *Intelligent Systems Magazine*, 21(1):35--41, 2006.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender system-a case study. In *ACM WebKDD Workshop*, 2000.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. WWW Conf.*, pages 285--295, 2001.