ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΩΝ ΣΥΣΤΑΣΕΩΝ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ

Διπλωματική Εργασία

Αθηνά Λιακοπούλου

Επιβλέπων: Μιχαήλ Βασιλακόπουλος

Βόλος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΩΝ ΣΥΣΤΑΣΕΩΝ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΗΜΟΣΙΕΥΣΕΩΝ

Διπλωματική Εργασία

Αθηνά Λιακοπούλου

Επιβλέπων: Μιχαήλ Βασιλακόπουλος

Βόλος 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

COMPARATIVE EVALUATION OF RECOMMENDER SYSTEMS FOR SCIENTIFIC ARTICLES

Diploma Thesis

Athina Liakopoulou

Supervisor: Michail Vassilakopoulos

Volos 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Μιχαήλ Βασιλακόπουλος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Ασπασία Δασκαλοπούλου**
Επίκουρος Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Ελένη Τουσίδου**
Μέλος Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 30-6-2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κύριο Βασιλακόπουλο Μιχαήλ για την επίβλεψη και τη βοήθεια που μου προσέφερε κατά την ανάπτυξη της παρούσας διπλωματικής εργασίας, καθώς και τις κυρίες Δασκαλοπούλου Ασπασία και Τουσίδου Ελένη, που δέχτηκαν να συμμετάσχουν ως μέλη Επιτροπής.

Εν συνεχεία, ευχαριστώ πολύ τους διδακτορικούς φοιτητές κύριο Βάιο Στεργιόπουλο και κυρία Τσιανάκα Θάλεια για την πολύτιμη στήριξή του, τις συμβουλές και διορθώσεις που μου υπέδειξαν κάθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια μου για τη συνεχή στήριξη και συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Αθηνά Λιακοπούλου

30-6-2021

Περίληψη

Στη σημερινή εποχή ο όγκος των πληροφοριών που υπάρχει αυξάνεται ραγδαία. Πλέον ένα από τα μεγαλύτερα ζητήματα αποτελεί η οργάνωση και αξιοποίηση αυτών των δεδομένων με στόχο την εξαγωγή πολύτιμων συμπερασμάτων. Το θέμα της συγκεκριμένης διπλωματικής εργασίας αποτελεί η επεξεργασία δεδομένων που αφορούν διάφορα χαρακτηριστικά επιστημονικών δημοσιεύσεων, η ανάπτυξη υβριδικών συστημάτων συστάσεων και η σύγκριση της αποτελεσματικότητάς τους. Τα συστήματα συστάσεων βοηθούν τους χρήστες να ξεπεράσουν τον μεγάλο όγκο των πληροφοριών του διαδικτύου. Τα δεδομένα που χρησιμοποιούνται συλλέγονται από τον παγκόσμιο ιστό, και συγκεκριμένα από την Aminer βιβλιοθήκη, αφορούν σε δημοσιεύσεις (συγγραφείς, εκδότης, χρονολογία έκδοσης κ.λπ.) και αποθηκεύονται στη NoSQL βάση δεδομένων MongoDB. Τα δεδομένα αυτά επεξεργάζονται με τέτοιο τρόπο ώστε να είναι σε κατάλληλη μορφή να δίνονται ως είσοδος στις μεθόδους ανάπτυξης των συστημάτων συστάσεων, με τέτοιο τρόπο ώστε να μπορούν να απαντούν σε τρία διαφορετικά ερωτήματα. Τόσο η επεξεργασία των δεδομένων, όσο και η ανάπτυξη των μεθόδων έχει πραγματοποιηθεί μέσω της χρήσης της python γλώσσας προγραμματισμού.

Λέξεις Κλειδιά:
Συγκομιδή δεδομένων, Ανάλυση δεδομένων, Επιστημονικά άρθρα, Υβριδικά συστήματα συστάσεων, Aminer, Συστάσεις για δημοσιεύσεις, MongoDB, NoSQL, Βάσεις δεδομένων Εγγράφων, Python

Abstract

Nowadays, the amount of information that exists is growing rapidly. Consequently, one of the biggest issues haw become the organization and use of this data in order to draw valuable conclusions. The subject of this diplomatic work concerns the processing of data about different characteristics of scientific publications, the development of hybrid recommendations systems and the comparison of their effectiveness. In essence, recommendations systems help users to overcome the vast amount of Internet information. The data used are collected from the web, and in particular from the Aminer library, related to publications (authors, publisher, publication date, etc.) and stored in the MongoDB NoSQL database. These data shall be processed in such a way that they are in an appropriate form to serve as input to the methods of developing the recommendations systems in such a way that they can answer three different questions. Both data preprocessing and methods development have been done through the use of python programming language.

Key words:

Data collection, Data analysis, Scientific articles, Hybrid recommender systems, Aminer, Recommendations for publishing, MongoDB, NoSQL, Document-oriented database, Python

Πίνακας περιεχομένων

Ευχαριστίες **ix**

Περίληψη **xi**

Abstract **xiii**

Πίνακας περιεχομένων **xv**

Κατάλογος σχημάτων **xix**

Συντομογραφίες **xxiii**

1 Εισαγωγή **1**

1.1 Αντικείμενο της διπλωματικής 2

1.1.1 Συνεισφορά 3

1.2 Οργάνωση του τόμου 4

2 Μέθοδοι βασισμένες στα Νευρωνικά Δίκτυα **5**

2.1 Εισαγωγή 5

2.2 Προαπαιτούμενες γνώσεις βασικών μεθόδων 9

2.2.1 Η έννοια της απόστασης 10

2.2.2 Μετρική Μάθηση 11

2.2.3 Συνεργατικό Φιλτράρισμα 12

2.2.4 Νευρωνικά Δίκτυα 13

2.2.5 Γνωστές εφαρμογές των Νευρωνικών Δικτύων 16

2.3 Μέθοδοι νευρωνικών δικτύων για επιστημονικά δημοσιεύματα 20

2.3.1 Βαθιά μάθηση στην έρευνα μοντέλων σύστασης παραπομπών . . . 20

2.3.2 Συστάσεις μέσω αναδρομικών νευρωνικών δικτύων προσοχής . . . 22
2.3.3 Σύσταση παραπομπών: προσεγγίσεις και σύνολα δεδομένων 22
2.4 Μέθοδος CATA++ 24
2.5 Μέθοδος CVAE 25
2.6 Μέθοδος RVAE 27
2.7 Μέθοδος Μετρικών Συνεργασίας 29

3 Προετοιμασία και επεξεργασία βάσης δεδομένων 45
3.1 Δεδομένα 45
3.2 Σχήμα Βάσης Δεδομένων 45
3.3 Εισαγωγή στη MongoDB 52
3.4 Επεξεργασία των δεδομένων 55
3.4.1 tags.dat 57
3.4.2 My_mult_nor.mat 58
3.4.3 users.dat 60
3.4.4 items.dat 61
3.4.5 citation.dat 62
3.4.6 tag-items.dat 62
3.5 Δεύτερη Μέθοδος Επεξεργασίας των δεδομένων 64

4 Μέτρηση αποτελεσματικότητας μεθόδων 69
4.1 Εκτελέσεις των συστημάτων συστάσεων 69
4.1.1 Εκτελέσεις σε διαφορετικό λογισμικό 69
4.1.2 Εκτελέσεις με διαφορετικές παραμέτρους 71
4.2 Μετρικές αξιολόγησης των μεθόδων 73
4.2.1 Recall 73
4.2.2 DCG και nDCG 74
4.3 Υπολογισμός μετρικών 76
4.4 Recall CATA++ 76
4.4.1 Συστάσεις εκδοτικών χώρων 76
4.4.2 Συστάσεις επιστημονικών συγγραμμάτων 78
4.4.3 Συστάσεις συντακτών 80
4.5 DCG CATA++ 82

4.5.1 Συστάσεις εκδοτικών χώρων 82
4.5.2 Συστάσεις επιστημονικών συγγραμμάτων 83
4.5.3 Συστάσεις συντακτών 85
4.6 nDCG CATA++ 86
4.6.1 Συστάσεις εκδοτικών χώρων 86
4.6.2 Συστάσεις επιστημονικών συγγραμμάτων 88
4.6.3 Συστάσεις συντακτών 90
4.7 Συμπεράσματα σύγκρισης των αποτελεσμάτων CATA++ 92
4.8 Recall CVAE 92
4.8.1 Συστάσεις εκδοτικών χώρων 94
4.8.2 Συστάσεις επιστημονικών συγγραμμάτων 94
4.8.3 Συστάσεις συντακτών 96
4.9 Συμπεράσματα σύγκρισης των αποτελεσμάτων CVAE 98
4.10 Recall CML 98
4.11 Συμπεράσματα σύγκρισης των αποτελεσμάτων CML 100
4.12 Σύγκριση απόδοσης CATA++, CVAE, CML 100

5 Συμπεράσματα και Μελλοντικά Σχέδια **105**
5.1 Συμπεράσματα 105
5.2 Μελλοντικά Σχέδια 107

Βιβλιογραφία **109**

Κατάλογος σχημάτων

2.1 Πιθανά δεδομένα εισόδου [1] 6
2.2 Επίπεδα εξατομίκευσης [1] 7
2.3 Ανάπτυξη συστήματος συστάσεων [1] 9
2.4 Παράδειγμα συστήματος συστάσεων: 1.Collaborative-based και 2.Content-based [2] 10
2.5 Artificial Neural Networks 1 [3] 14
2.6 Artificial Neural Networks 2 [3] 15
2.7 Recurrent Neural Networks 1 [3] 16
2.8 Παράδειγμα Recurrent Neural Network [3] 17
2.9 Recurrent Neural Network 2 [3] 17
2.10 Convolution Neural Network παράδειγμα 1 [3] 18
2.11 Convolution Neural Network παράδειγμα 2 [3] 19
2.12 RBM [4] 20
2.13 MLP [4] 21
2.14 heterogeneous knowledge embedding based attentive recurrent neural networks [5] 32
2.15 Αρχιτεκτονική ενός πρωτότυπου συστήματος συστάσεων παραπομπών [6] 33
2.16 Μοντέλο CATA++ [7] 33
2.17 Αλγόριθμος ανάπτυξης CATA++ [7] 34
2.18 CATA++ Διάγραμμα Ροής 35
2.19 Μοντέλο CVAE [2] 36
2.20 Μοντέλο CVAE i 36
2.21 Μοντέλο CVAE ii 36
2.22 Μοντέλο CVAE iii 37
2.23 CVAE Διάγραμμα Ροής i 37

2.24 CVAE Διάγραμμα Ροής ii 38
2.25 Μοντέλο RVAE [8] 38
2.26 Μοντέλο RVAE i 39
2.27 Μοντέλο RVAE ii 39
2.28 RVAE Διάγραμμα Ροής i 40
2.29 RVAE Διάγραμμα Ροής ii 41
2.30 Μοντέλο CML [9] 42
2.31 CML Διάγραμμα Ροής 43

3.1 Σχήμα Βάσης Δεδομένων [10] 51
3.2 Περιβάλλον MongoDB Compass 55
3.3 keywords2.dat 58
3.4 tags.dat 59
3.5 users.dat 60
3.6 items.dat 61
3.7 citation.dat 62
3.8 tag-items.dat 63

4.1 Διασύνδεση με server i 71
4.2 Διασύνδεση με server ii 72
4.3 Διασύνδεση με server iii 73
4.4 Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall CATA++ 77
4.5 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall CATA++ 79
4.6 Σύστημα συστάσεων συντακτών - απόδοση Recall CATA++ 81
4.7 Σύστημα συστάσεων εκδοτικών χώρων - απόδοση DCG CATA++ 82
4.8 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση DCG CATA++ 84
4.9 Σύστημα συστάσεων συντακτών - απόδοση DCG CATA++ 85
4.10 Σύστημα συστάσεων εκδοτικών χώρων - απόδοση nDCG CATA++ 87
4.11 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση nDCG CATA++ 89
4.12 Σύστημα συστάσεων συντακτών - απόδοση nDCG CATA++ 91
4.13 Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall CVAE 93
4.14 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall CVAE 95
4.15 Σύστημα συστάσεων συντακτών - απόδοση Recall CVAE 97

4.16 Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML 101
4.17 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML 102
4.18 Σύστημα συστάσεων συντακτών - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML 103
4.19 Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML - Small1 & Small2 Dtasets 104

Συντομογραφίες

κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
RS	Recommender Systems
RDBMS	Relational Database Management System
BSON	Binary JSON
GUI	Graphical User Interface
fos	Field of Study
MF	Matrix Factorization
CF Ή CBF	Collaborative Filtering
SVM	Support Vector Machine
ML	Machine Learning
DL	Deep Learning
LMNN	Large Margin Nearest Neighbor
WRMF	Weighted Regularized Matrix Factorization
ANN	Artificial Neural Network
RNN	Recurrent Neural Network
CNN	Convolution Neural Network
RBM	Restricted Boltzmann Machine
MLP	Multi Layer Perceptron
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
BN	Batch Normalization
PMF	Probabilistic Matrix Factorization
CATA++	Collaborative Dual Attentive Autoencoder

CVAE	Collaborative Variational Autoencoder
RVAE	Relational Variational Autoencoder
CML	Collaborative Metric Learning
DCG	Discounted Cumulative Gain
nDCG	Normalized Discounted Cumulative Gain

Κεφάλαιο 1

Εισαγωγή

Τα συστήματα συστάσεων / Recommender Systems (RS) είναι εργαλεία λογισμικού που χρησιμοποιούνται για να δημιουργούν και να παρέχουν στους χρήστες προτάσεις για αντικείμενα και άλλες οντότητες, αξιοποιώντας διάφορες στρατηγικές. Τα συστήματα συστάσεων σήμερα έχουν γίνει βασικό συστατικό κάθε εμπορικού ιστότοπου. Έχουν αναπτυχθεί για να παρέχουν συστάσεις παραγόμενες από μηχανήματα και να διευκολύνουν την καθημερινή ζωή των ατόμων. Σε αντίθεση με τις συστάσεις που διατυπώνει ένας άνθρωπος, οι συστάσεις που παρέχονται από υπολογιστή μπορούν να λαμβάνουν υπόψη τεράστιες ποσότητες δεδομένων που εκφράζουν τις επιλογές και την κριτική μεγάλου αριθμού ανθρώπων, ή ακόμη και "κρυμμένη" γνώση που εξάγεται με έμμεσο τρόπο.

Σήμερα, η εκρηκτική αύξηση της ποσότητας των ψηφιακών πληροφοριών που παράγονται έχει αυξήσει την σημασία των συστάσεων που παρέχουν τα συστήματα στα άτομα. Υπάρχει αύξηση στον ετήσιο αριθμό δημοσιεύσεων, που δείχνει σαφώς το ενδιαφέρον της επιστημονικής κοινότητας για το θέμα αυτό. Με τα συστήματα συστάσεων να αυξάνονται σε δημοτικότητα, σημαντικές εταιρείες εκμεταλλεύτηκαν τη δυνατότητα να προβλέψουν τις ανάγκες των χρηστών και να τους προσφέρουν συστάσεις για τα στοιχεία (υλικά ή άυλα) που τους αρέσουν, συχνά χωρίς οι ίδιοι οι χρήστες ακόμη και να γνωρίζουν.

Κοινές εφαρμογές των συστημάτων συστάσεων έχουν σχέση με την παρακολούθηση ταινιών, την ακρόαση μουσικής, την επίσκεψη σε εστιατόρια, τα ταξίδια, την εύρεση κατάλληλων φίλων στα κοινωνικά δίκτυα, και άλλα. Χρησιμοποιούνται επίσης ευρέως στην πολιτική, τον αθλητισμό και την επιστήμη.

Ο τεράστιος όγκος των ταχέως αναπτυσσόμενων επιστημονικών δημοσιεύσεων σε όλους τους επιστημονικούς κλάδους κατακλύζει ερευνητές και μελετητές με μεγάλο αριθμό επι-

στημονικών δημοσιεύσεων για να διαβάσουν. Ταυτόχρονα, τους δίνεται μεγάλη ποικιλία επιλογών περιοδικών ή συνεδρίων για δημοσίευση των άρθρων τους, ή να εκδώσουν τα συγγράμματά τους.Τέλος, είναι δύσκολο να βρουν άλλους κατάλληλους ερευνητές - συντάκτες επιστημονικών άρθρων με τους οποίους έχουν κοινά ενδιαφέροντα και μπορούν να συνεργαστούν. Βασικό σκοπό αυτής της διπλωματικής εργασίας αποτελεί η δημιουργία συστημάτων συστάσεων που θα βοηθήσουν τους συγγραφείς ώστε να επιτύχουν όλους τους στόχους τους.

1.1 Αντικείμενο της διπλωματικής

Βασικός σκοπός της παρούσας διπλωματικής εργασίας αποτελεί η σύγκριση τεσσάρων υβριδικών συστημάτων συστάσεων επιστημονικών δημοσιεύσεων βασισμένων σε νευρωνικά δίκτυα. Αρχικά γίνεται η χρήση ενός μεγάλου αριθμού δεδομένων από τη διαδικτυακή βάση Aminer, η οργάνωση τους στη NoSQL βάση δεδομένων MongoDB και στη συνέχεια η επεξεργασία όλων αυτών των δεδομένων με στόχο την απάντηση συγκεκριμένων ερωτημάτων. Πιο συγκεκριμένα, πρώτο βήμα αποτελεί η επεξεργασία των δεδομένων ώστε να είναι στη κατάλληλη μορφή που χρειάζεται και να δοθούν ως είσοδος στις μεθόδους συστημάτων συστάσεων που θα αναπτυχθούν. Σε αυτή τη διπλωματική εργασία θέλουμε να κάνουμε τρεις βασικές συστάσεις μέσω των ίδιων δεδομένων:

1. Πρόταση σχετικά με εκδοτικούς χώρους στους συντάκτες επιστημονικών εγγράφων για την υποβολή συγκεκριμένων εγγράφων που έχουν ήδη γράψει ή που θα συντάξουν στο μέλλον, σύμφωνα με τα ερευνητικά τους ενδιαφέροντα.

2. Πρόταση επιστημονικών άρθρων για ανάγνωση και μελέτη στους συντάκτες επιστημονικών περιοδικών και βιβλίων.

3. Πρόταση ερευνητών-συντακτών επιστημονικών άρθρων σε άλλους ερευνητές-συντάκτες επιστημονικών άρθρων για πιθανή συνεργασία.

Δεύτερο βήμα αποτελεί η ανάπτυξη των συστημάτων συστάσεων που θα χρησιμοποιηθούν για την παραγωγή των παραπάνω προτάσεων. Υπάρχουν αρκετές δημοφιλείς μέθοδοι παραγωγής συστημάτων συστάσεων που έχουν χρησιμοποιηθεί ευρέως μέχρι σήμερα. Ωστόσο, οι μέθοδοι αυτές αντιμετωπίζουν αρκετά προβλήματα , και κατά συνέπεια, προτάθηκαν πρόσφατα πολλά υβριδικά μοντέλα για τη βελτιστοποίηση της απόδοσης των μέχρι

στιγμής γνωστών μεθόδων, ενσωματώνοντας πρόσθετες συναφείς πληροφορίες στη διαδικασία εκμάθησής τους. Τα υβριδικά συστήματα συστάσεων συνδυάζουν δύο ή περισσότερες στρατηγικές σύστασης με διαφορετικούς τρόπους ώστε να επωφελούνται από τα συμπληρωματικά τους πλεονεκτήματα. Στη διπλωματική αυτή αναπτύσσονται τέσσερις υβριδικές μέθοδοι που βασίζονται στα νευρωνικά δίκτυα για την εκπαίδευση των συστημάτων συστάσεων.

Τελευταίο βήμα αποτελεί η σύγκριση απόδοσης αυτών των μεθόδων. Ανάλογα με το λογισμικό στο οποίο τρέχουν τα συστήματα και ανάλογα με το μέγεθος και τη ποιότητα των δεδομένων που δίνονται ως είσοδος στις μεθόδους συγκρίνουμε την αποτελεσματικότητα των συστάσεων που προκύπτουν. Για να επιτευχθεί αυτό, χωρίζονται τα δεδομένα σε σετ εκπαίδευσης και δοκιμών / training και testing sets και στη συνέχεια υπολογίζονται οι τιμές recall, dcg και ndcg που αποτελούν μετρήσεις της απόδοσής τους.

1.1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκαν τέσσερις βασικές εργασίες ανάπτυξης υβριδικών μεθόδων συστημάτων συστάσεων που σχετίζονται με επιστημονικά δημοσιεύματα.

2. Μελετήθηκε η Aminer ως πηγή παροχής μεγάλου όγκου πληροφοριών για επιστημονικά δημοσιεύματα.

3. Μελετήθηκε η MongoDB ως βάση δεδομένων και η χρησιμότητα και διευκόλυνση που προσφέρει στην αποθήκευση, οργάνωση, διαχείριση και αξιοποίηση των δεδομένων που εξάγουμε από την Aminer.

4. Έγινε επεξεργασία των δεδομένων με διάφορες μεθόδους ώστε να λάβουν κατάλληλη μορφή και να δοθούν ως είσοδος στα συστήματα συστάσεων.

5. Έγινε εγκατάσταση τεσσάρων μεθόδων ανάπτυξης συστημάτων συστάσεων τα οποία προσαρμόστηκαν ώστε να εξυπηρετούν το σκοπό της διπλωματικής εργασίας.

6. Τροποποιήθηκαν τα συστήματα συστάσεων ώστε να υπολογίζουν τις ίδιες μετρικές αξιολόγησης, με σκοπό να μπορούν να γίνουν συγκρίσιμα και να υπάρχει ένα μέτρο αξιολόγησης ως προς την αποδοτικότητά τους.

7. Εκτελέστηκαν οι μέθοδοι με διάφορους συνδυασμούς ώστε να βρεθούν οι βέλτιστες παράμετροι εκτέλεσης για τον καθένα.

8. Δημιουργήθηκαν διαγράμματα με τα παραπάνω αποτελέσματα με σκοπό την οπτικοποίηση και βέλτιστη σύγκριση των μεθόδων.

9. Βγήκαν συμπεράσματα από τη μελέτη που έγινε ως προς την αποδοτικότητα της κάθε μεθόδου.

1.2 Οργάνωση του τόμου

Στο κεφάλαιο 2 αναλύονται όλες οι προ απαιτούμενες γνώσεις που χρειάζονται για την κατανόηση λειτουργίας των υβριδικών συστημάτων, και στη συνέχεια επεξηγούνται οι τέσσερις υβριδικές μέθοδοι. Στο κεφάλαιο 3 αναλύεται η MongoDB ως βάση δεδομένων και ο λόγος για τον οποίο επιλέχθηκε για την ανάπτυξη της διπλωματικής εργασίας, και αναλύεται ο τρόπος επεξεργασίας και διαμόρφωσης των δεδομένων ώστε να είναι κατάλληλα να δοθούν ως είσοδος στις παραπάνω μεθόδους. Στο κεφάλαιο 4 παρουσιάζονται όλες οι εναλλακτικές με τις οποίες έτρεξαν οι μέθοδοι και αξιολογείται η αποτελεσματικότητά τους μέσω των τιμών recall, dcg και ndcg. Οι τιμές οπτικοποιούνται σε μορφή διαγραμμάτων ώστε να παραχθούν συμπεράσματα ως προς την αποτελεσματικότητα των μεθόδων, να βρεθεί υπό ποιες προϋποθέσεις λειτουργεί καλύτερα κάθε μέθοδος, και εν τέλη ποια είναι η βέλτιστη. Τέλος, στο κεφάλαιο 5 παρατίθενται τα συμπεράσματα και ιδέες μελλοντικής βελτιστοποίησης και επέκτασης της εργασίας.

Κεφάλαιο 2

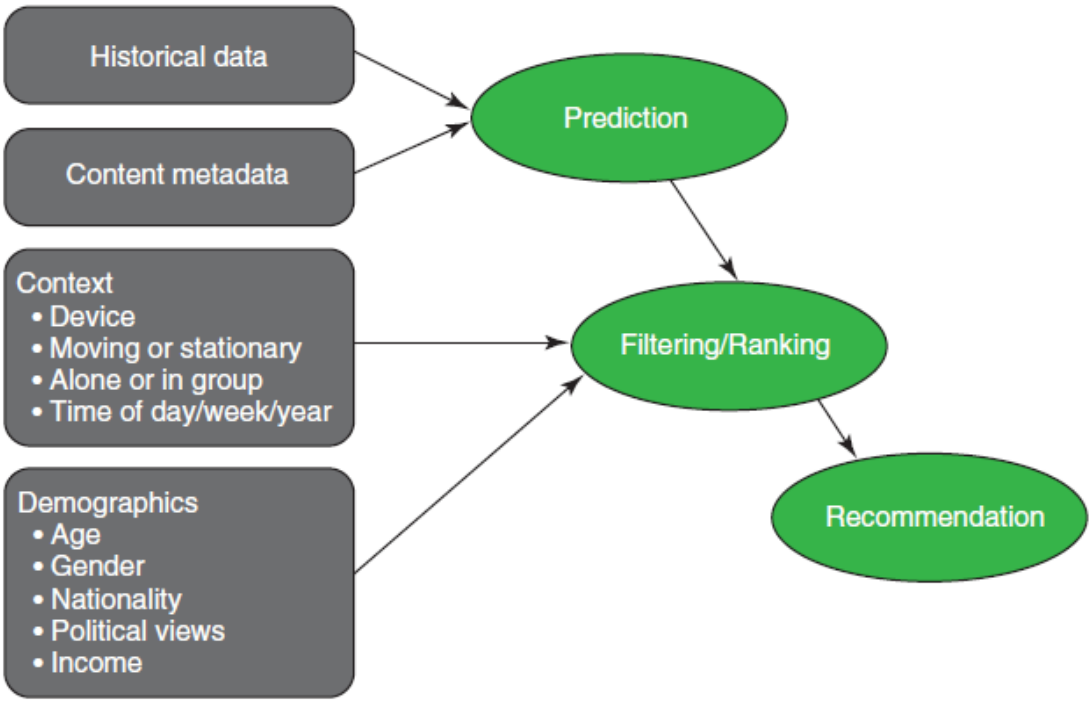
Μέθοδοι βασισμένες στα Νευρωνικά Δίκτυα

2.1 Εισαγωγή

Ορισμός 2.1. *Ένα* **σύστημα συστάσεων** *υπολογίζει και παρέχει σχετικό περιεχόμενο στο χρήστη με βάση τις γνώσεις του για το χρήστη, του περιεχομένου των στοιχείων που προτείνει, και των αλληλεπιδράσεων μεταξύ του χρήστη και του στοιχείου.*

Με την ταχεία ανάπτυξη των υπηρεσιών και εφαρμογών του διαδικτύου, οι άνθρωποι έχουν πρόσβαση σε μεγάλους όγκους περιεχομένου ηλεκτρονικών πολυμέσων, όπως ταινίες, μουσική, ειδήσεις και άρθρα. Ενώ η αύξηση αυτή επέτρεψε στους χρήστες να καταναλώσουν έναν τεράστιο αριθμό πόρων με ένα μόνο κλικ, κατέστησε επίσης δυσκολότερο για τους χρήστες να βρουν πληροφορίες σχετικές με τα ενδιαφέροντά τους. Για παράδειγμα, οι χρήστες μπορεί να μην γνωρίζουν την ύπαρξη ενδιαφερουσών ταινιών που θα ήθελαν και οι ερευνητές μπορεί να δυσκολεύονται να αναζητήσουν σημαντικά επιστημονικά άρθρα σχετικά με τον τομέα της έρευνάς τους. Συνεπώς, τα συστήματα συστάσεων αποκτούν ολοένα και μεγαλύτερη σημασία για την προσέλκυση χρηστών και την αποτελεσματική χρήση των διαθέσιμων πληροφοριών. Αυτό παρακινεί και προσελκύει τους ερευνητές να αξιοποιήσουν τα μαζικά δεδομένα για να αναπτύξουν πιο πρακτικές και ακριβείς λύσεις στους περισσότερους τομείς της επιστήμης των υπολογιστών [7].

Συνεπώς, μια σύσταση υπολογίζεται με βάση το τι αρέσει στο χρήστη, τι άρεσε σε άλλους στο παρελθόν, και τι συχνά ζητείται από το χρήστη. Οι περισσότερες συστάσεις συμβαίνουν στο χώρο του διαδικτύου, επειδή εκεί όχι μόνο απευθύνονται σε μεμονωμένους χρή-



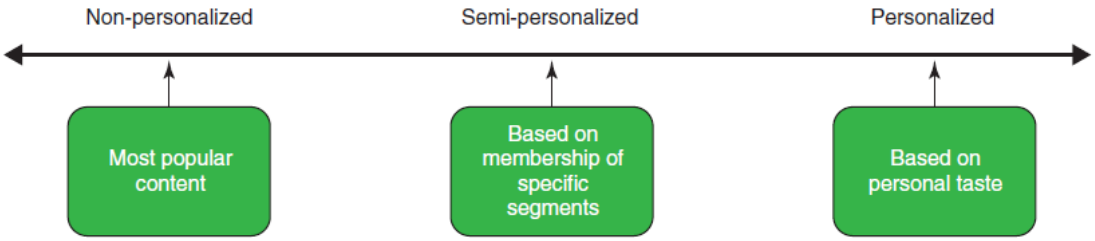
Σχήμα 2.1: Πιθανά δεδομένα εισόδου [1]

στες αλλά ταυτόχρονα είναι δυνατή η συλλογή δεδομένων συμπεριφοράς. Μια ιστοσελίδα που παρουσιάζει τις 10 κορυφαίες λίστες των πιο πωληθέντων μηχανημάτων παρασκευής ψωμιού παρέχει μη εξατομικευμένες συστάσεις. Εάν μια ιστοσελίδα για πωλήσεις εισιτηρίων συναυλιών παρουσιάζει συστάσεις με βάση τα δημογραφικά στοιχεία ή την τρέχουσα τοποθεσία, οι συστάσεις είναι ήμι - εξατομικευμένες. Εξατομικευμένες συστάσεις μπορούν να βρεθούν στην "Amazon", όπου οι πελάτες βλέπουν "Συστάσεις για εσάς". Η ιδέα της εξατομικευμένης σύστασης προκύπτει επίσης από την ιδέα ότι οι άνθρωποι δεν ενδιαφέρονται μόνο για τα δημοφιλή αντικείμενα, αλλά και για αντικείμενα που δεν είναι δημοφιλή.

Βάση των παραπάνω συνεπάγεται ότι τα περισσότερα συστήματα συστάσεων προσπαθούν να χρησιμοποιήσουν με διάφορους τρόπους τα δεδομένα που αναφέρονται στο σχήμα 2.1 με σκοπό τη βέλτιστη σύσταση συστάσεων.

Τα βασικά στοιχεία που απαρτίζουν ένα σύστημα συστάσεων είναι τα εξής:

1. Τομέας = Ο τομέας είναι ο τύπος περιεχομένου που προτείνεται. Για παράδειγμα, στο Netflix ο τομέας είναι ταινίες και τηλεοπτικές σειρές. Γενικότερα, μπορεί να είναι οτιδήποτε, όπως: αλληλουχίες περιεχομένου όπως λίστες αναπαραγωγής, οι καλύτεροι τρόποι για να παρακολουθήσετε μαθήματα ηλεκτρονικής μάθησης για να επιτύ-



Σχήμα 2.2: Επίπεδα εξατομίκευσης [1]

χετε ένα στόχο, θέσεις εργασίας, βιβλία, αυτοκίνητα, ψώνια, διακοπές, προορισμούς ή ακόμα και ανθρώπους. Ο τομέας είναι σημαντικός, επειδή καθορίζει τις συστάσεις που προκύπτουν.

2. Σκοπός = Βασικός σκοπός ενός συστήματος συστάσεων θα μπορούσε να είναι η παροχή πληροφοριών ή η παροχή βοήθειας ή η εκπαίδευση του χρήστη. Στις περισσότερες περιπτώσεις, ωστόσο, ο σκοπός είναι πιθανώς η αύξηση των πωλήσεων. Ταυτόχρονα, αυτό καθορίζει το είδος πελατών που εξυπηρετούνται: είτε επικεντρώνονται σε καταναλωτές που φθάνουν μία φορά και αναμένουν καλές συστάσεις, είτε σε πιστούς καταναλωτές που δημιουργούν προφίλ και επιστρέφουν σε τακτική βάση.

3. Πλαίσιο = Το πλαίσιο είναι το περιβάλλον στο οποίο ο καταναλωτής λαμβάνει μια σύσταση. Ένα παράδειγμα στο οποίο το περιβάλλον έχει βασικό ρόλο αποτελεί το εξής: Έστω ότι πραγματοποιείται μια αναζήτηση για καφετέρια στο Google Maps. Ο χρήστης κάθεται σε έναν υπολογιστή γραφείου και ψάχνει για μια καλή καφετέρια ή στέκεται στο δρόμο καθώς αρχίζει να βρέχει; Αναλόγως του πλαισίου η σύσταση πρέπει να είναι διαφορετική.

4. Επίπεδο εξατομίκευσης = Οι συστάσεις μπορούν να προκύψουν σε πολλά επίπεδα εξατομίκευσης, από τη χρήση βασικών στατιστικών έως την εξέταση των δεδομένων μεμονωμένων χρηστών. Το σχήμα 2.2 απεικονίζει αυτά τα επίπεδα.

5. Συστάσεις ειδικών = Οι συστάσεις των ειδικών είναι προαπαιτούμενες σε συστήματα όπως συστάσεις κρασιών, βιβλίων ή κάτι παρόμοιο. Τα συστήματα αυτά χρησιμοποιούνται σε τομείς όπου είναι γενικά αποδεκτό ότι χρειάζεται κάποιος ειδικός για να ορίσει το τι είναι καλό. Ωστόσο, οι ημέρες των ιστοσελίδων εμπειρογνωμόνων έχουν πλέον μειωθεί, καθώς σχεδόν όλες οι ιστοσελίδες χρησιμοποιούν τις απόψεις της μάζας.

6. Προστασία της ιδιωτικής ζωής και αξιοπιστία = Πόσο καλά προστατεύει το απόρρητο των χρηστών το σύστημα; Πώς χρησιμοποιούνται οι συλλεγμένες πληροφορίες; Πολλοί θεωρούν τις συστάσεις ως μορφή χειραγώγησης, επειδή παρουσιάζουν επιλογές που οι πελάτες είναι πιο πιθανό να επιλέξουν από ό,τι αν τους προσφερόταν τυχαία. Συνεπώς, η αξιοπιστία δείχνει πόσο εμπιστεύεται ο καταναλωτής τις συστάσεις αντί να τις θεωρεί διαφημίσεις ή απόπειρες χειραγώγησης.

7. Διασύνδεση = Η διασύνδεση ενός συστήματος συστάσεων απεικονίζει το είδος της εισόδου και της εξόδου που παράγει (User Interface).

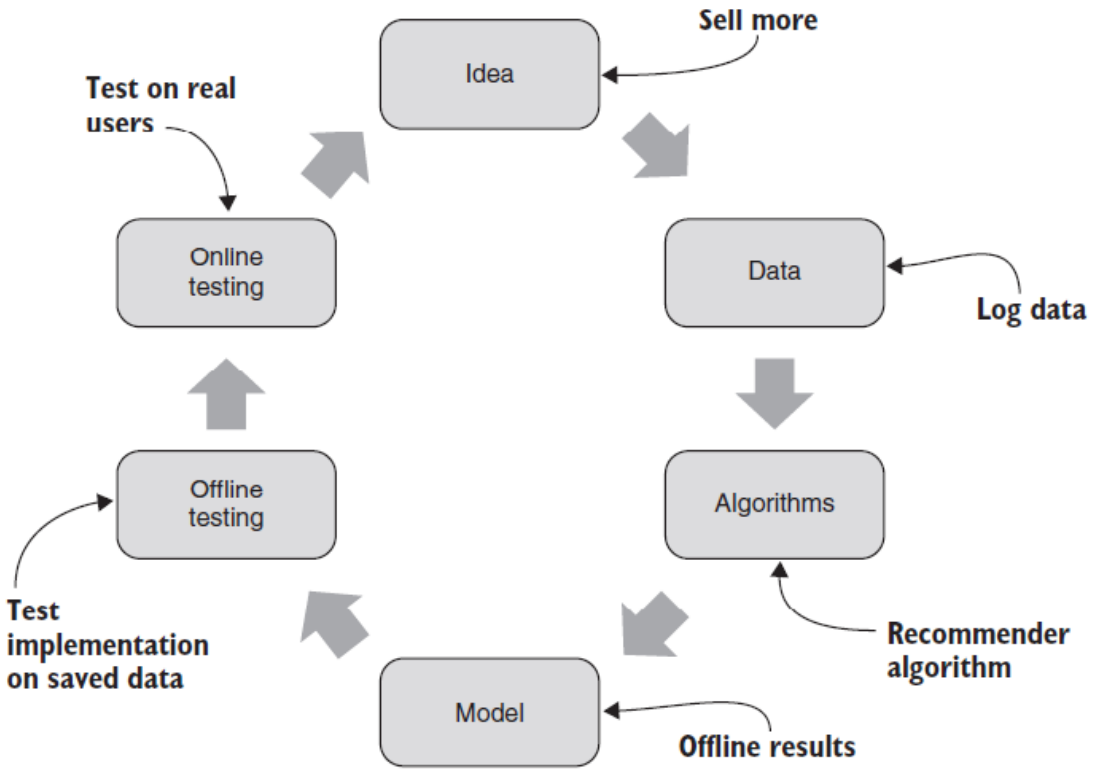
8. Αλγόριθμοι ανάπτυξης = Οι αλγόριθμοι χωρίζονται σε δύο ομάδες και εξαρτώνται από τον τύπο των δεδομένων που χρησιμοποιούνται για να προκύψουν οι προτάσεις.

   Οι αλγόριθμοι που βασίζονται στις αξιολογήσεις και προτιμήσεις του χρήστη ονομάζονται συνεργατικό φιλτράρισμα / collaborative filtering. Οι αλγόριθμοι που χρησιμοποιούν μεταδεδομένα περιεχομένου και προφίλ χρηστών για τον υπολογισμό των συστάσεων ονομάζονται φιλτράρισμα βάσει περιεχομένου / content based. Ένας συνδυασμός των δύο τύπων ονομάζεται υβριδικές συστάσεις.

Συνεπώς, για να αναπτυχθεί ένα σύστημα συστάσεων πρέπει κανείς να λάβει υπόψη όλα τα παραπάνω. Μια βασική προσέγγιση ανάπτυξης ενός συστήματος συστάσεων παρουσιάζεται στην εικόνα 2.3.

Όπως προαναφέρθηκε, στα συστήματα συστάσεων υπάρχουν δύο είδη διαθέσιμων πληροφοριών: η αξιολόγηση και το περιεχόμενο του αντικειμένου, π.χ. οι κριτικές των ταινιών και η περιγραφή τους. Οι υπάρχουσες μέθοδοι για συστήματα συστάσεων μπορούν να ταξινομηθούν χονδρικά σε τρεις κατηγορίες: μέθοδοι που βασίζονται στο περιεχόμενο / content based, μέθοδοι που βασίζονται σε συνεργασία / collaborative based, και υβριδικές μέθοδοι / hybrid methods.

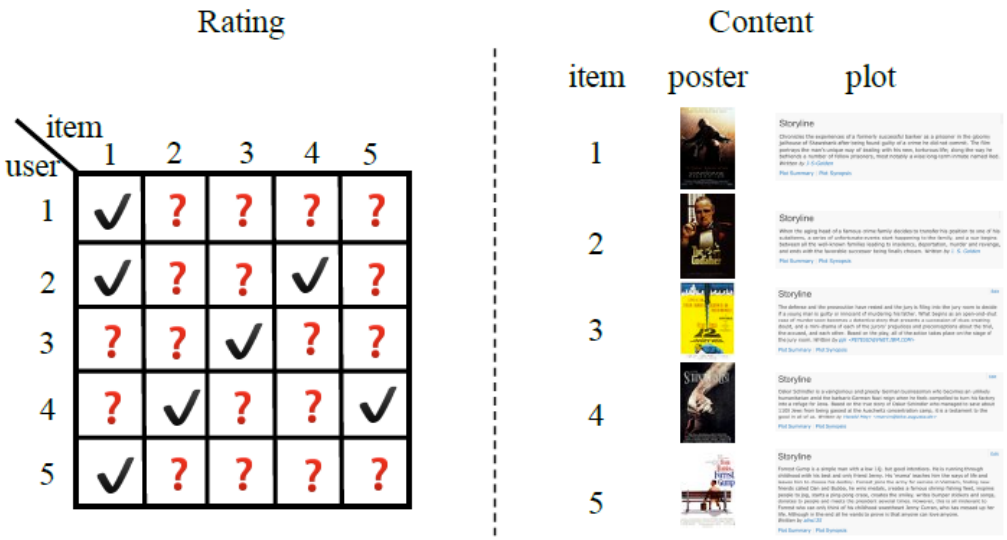
Παρακάτω αναλύονται οι πιο βασικοί και δημοφιλείς “content based” ή “collaborative based” μέθοδοι που χρησιμοποιούνται μέχρι σήμερα αλλά και μέθοδοι που βασίζονται στα νευρωνικά δίκτυα, και στη συνέχεια χρησιμοποιούνται ως βάση για την ανάπτυξη των υβριδικών μεθόδων που εξετάζει η παρούσα διπλωματική εργασία.



Σχήμα 2.3: Ανάπτυξη συστήματος συστάσεων [1]

2.2 Προαπαιτούμενες γνώσεις βασικών μεθόδων

Όπως προαναφέρθηκε, σε εφαρμογές συστημάτων συστάσεων υπάρχουν δύο τύποι διαθέσιμων πληροφοριών: η βαθμολογία / κριτική και το περιεχόμενο / η περιγραφή του αντικειμένου, όπως φαίνεται στο σχήμα 2.4. Οι μέθοδοι βάσει περιεχομένου / content based κάνουν χρήση των περιγραφών του χρήστη ή των περιγραφών των στοιχείων και προτείνονται στον κάθε χρήστη αντικείμενα παρόμοια με αυτά που του άρεσαν στο παρελθόν. Οι μέθοδοι συνεργασίας / collaborative-based κάνουν χρήση δεδομένων όπως αξιολογήσεις χρηστών σε αντικείμενα, χωρίς τη χρήση πληροφοριών περιεχομένου των αντικειμένων, και στο χρήστη συνιστώνται στοιχεία που άρεσαν στο παρελθόν σε άτομα με παρόμοιες προτιμήσεις. Μια από τις πιο δημοφιλείς content-based μεθόδους αποτελεί ο Matrix Factorization (MF), λόγω της απλότητας και της αποτελεσματικότητάς του. Οι μέθοδοι που βασίζονται στη συνεργασία επιτυγχάνουν γενικά καλύτερες συστάσεις από τις μεθόδους βάσει περιεχομένου. Ωστόσο, οι μέθοδοι που βασίζονται σε συνεργασίες έχουν αρκετούς περιορισμούς. Η απόδοση της πρότασης μειώνεται σημαντικά όταν οι βαθμολογίες είναι πολύ αραιές. Επιπλέον, δεν μπορούν να χρησιμοποιηθούν για τη σύσταση νέων στοιχείων που δεν έχουν λάβει αξιολογήσεις από



Σχήμα 2.4: Παράδειγμα συστήματος συστάσεων: 1.Collaborative-based και 2.Content-based [2]

τους χρήστες, το οποίο είναι το λεγόμενο πρόβλημα εκκίνησης / cold start. Κατά συνέπεια, οι υβριδικές μέθοδοι επιδιώκουν να αξιοποιήσουν καλύτερα και τις δύο προσεγγίσεις.

2.2.1 Η έννοια της απόστασης

Η έννοια της απόστασης, όπως εξηγείται και στη δημοσίευση[9], βρίσκεται στο επίκεντρο πολλών θεμελιωδών μεθόδων μηχανικής μάθησης, συμπεριλαμβανομένων των K-πλησιέστερων γειτόνων, K-mean και SVMs. Οι μέθοδοι μετρικής μάθησης παράγουν μια μέτρηση απόστασης που καταγράφει τις σημαντικές σχέσεις μεταξύ των δεδομένων.

Δεδομένου ενός συνόλου αντικειμένων στα οποία είναι γνωστό ότι ορισμένα ζεύγη αντικειμένων είναι όμοια ή ανόμοια, στόχο της μετρικής μάθησης αποτελεί η εκμάθηση μια μετρική απόσταση που σέβεται αυτές τις σχέσεις. Συγκεκριμένα, η εκμάθηση μιας μέτρησης που εκχωρεί μικρότερες αποστάσεις μεταξύ παρόμοιων ζευγών και μεγαλύτερες αποστάσεις μεταξύ ανόμοιων ζευγών. Μαθηματικά, μια μέτρηση πρέπει να πληροί αρκετές προϋποθέσεις, μεταξύ των οποίων η ανισότητα του τριγώνου είναι η πιο κρίσιμη για τη γενίκευση μιας μαθημένης μέτρησης. Η ανισότητα του τριγώνου δηλώνει ότι για οποιαδήποτε τρία αντικείμενα, το άθροισμα οποιωνδήποτε δύο αποστάσεων κατά ζεύγη θα πρέπει να είναι μεγαλύτερο ή ίσο με την υπόλοιπη απόσταση κατά ζεύγη. Αυτό συνεπάγεται ότι, λαμβανομένων υπόψη των πληροφοριών: το x είναι παρόμοιο και με τα y και z, μια μελετημένη μέτρηση όχι μόνο

θα τραβήξει τα δηλωμένα δύο ζεύγη πιο κοντά, αλλά και θα τραβήξει το υπόλοιπο ζεύγος (y, z) σχετικά κοντά το ένα στο άλλο. Συνεπώς μπορεί να θεωρηθεί ως μια διαδικασία διάδοσης ομοιότητας, στην οποία η μελετημένη μέτρηση διαδίδει τις γνωστές πληροφορίες ομοιότητας στα ζεύγη των οποίων οι σχέσεις είναι άγνωστες.

2.2.2 Μετρική Μάθηση

Στη δημοσίευση [9] αναλύεται επίσης η μέθοδος της μετρικής μάθησης:

Έστω $x = x_1, x_2, ..., x_n$ μια συλλογή δεδομένων στον χώρου εισόδου $\mathbb{R}^m$. Οι πληροφορίες στη μετρική εκμάθηση / Metric Learning καθορίζονται με τη μορφή περιορισμών κατά ζεύγη, συμπεριλαμβανομένου του συνόλου γνωστών παρόμοιων ζευγών:

S = $\{(x_i, x_j)|x_i x_j$ θεωρούνται παρόμοια $\}$,

D = $\{(x_i, x_j)|x_i x_j$ θεωρούνται ανόμοια $\}$.

Έτσι βασικό στόχο αποτελεί η εκμάθηση μιας μέτρησης απόστασης που συγκεντρώνει όλα τα παρόμοια ζεύγη μεταξύ τους και ξεχωρίζει τα διαφορετικά ζεύγη. Αυτός ο στόχος, ωστόσο, δεν είναι πάντα εφικτός.

Από την άλλη πλευρά, αν η μέθοδος metric learning χρησιμοποιηθεί για την ταξινόμηση k-πλησιέστερου γείτονα, αρκεί να είναι γνωστή μόνο μια μέτρηση που κάνει τους πλησιέστερους γείτονες κάθε αντικειμένου να είναι τα αντικείμενα που μοιράζονται την ίδια ετικέτα κλάσης με αυτό. Συγκεκριμένα, δεδομένου ενός στοιχείου εισόδου / input x, τα δεδομένα που είναι πιο κοντά στο x αναφέρονται ως γειτονικοί στόχοι / target neighbors. Οι γείτονες-στόχοι του x δημιουργούν μια νοητή περίμετρο που δεν πρέπει να εισβάλλουν στοιχεία με διαφορετικές ετικέτες. Τα στοιχεία με διαφορετικές ετικέτες που εισβάλλουν στην περίμετρο αναφέρονται ως απατεώνες / impostors. Ο στόχος της μάθησης, γενικά, είναι η εκμάθηση μιας μέτρησης που ελαχιστοποιεί τον αριθμό των απατεώνων.

Μια τεχνική για να επιτευχθούν τα παραπάνω αποτελεί η large margin nearest neighbor μέθοδος, η οποία χρησιμοποιεί δυο βασικές συναρτήσεις:

Pull loss = τραβάει τους γείτονες στόχους / target neighbors μιας εισόδου x πιο κοντά:

$$\mathrm{Lpull(d)} = \sum_{j \sim i} (d(xi, xj))^2$$

όπου $j \sim i$συμβολίζει ότι η είσοδος j είναι γειτονικός στόχος του i.

Push loss = απομακρύνει τους απατεώνες από τη γειτονιά και διατηρεί ένα περιθώριο

ασφάλειας γύρω από τα όρια απόφασης του kNN:

$$\text{Lpush (d)} = \sum_{i,j\sim \text{i}} \sum_{k} (1 - yik) \left[1 + d(xi, xj)^2 - d(xi, xj)^2\right] +$$

Όπου $y_{ik} = 1$ αν τα στοιχεία εισόδου i και k στοιχεία είναι με ίδιες ετικέτες,

αλλιώς $y_{ik} = 0$ και [z]+ = max(z, 0) αποτελεί το standard hinge loss.

Συνεπώς η LMNN μέθοδος έχει ως στόχο τον σταθμισμένο συνδυασμός Lpull(d) και Lpush(d).

2.2.3 Συνεργατικό Φιλτράρισμα

Τα μοντέλα φιλτραρίσματος /Collaborative Filtering (CBF) βασίζονται κυρίως στις αξιολογήσεις και προτιμήσεις του χρήστη, και λειτουργούν αποτελεσματικά για την αντιμετώπιση των προβλημάτων εκκίνησης και αραιότητας δεδομένων. Κάποιες πρώιμες εργασίες σχετικά με αυτό το θέμα εφαρμόζουν την μέθοδο naive Bayesian για προτάσεις βιβλίων από ιστοσελίδες του Amazon. Επίσης, μια άλλη εργασία αναπτύσσει ένα σύστημα συστάσεων ταινιών βασισμένο μόνο σε συνόψεις ταινιών χρησιμοποιώντας τεχνικές κατηγοριοποίησης κειμένου. Αργότερα, οι ερευνητές ενσωματώνουν τόσο το περιεχόμενο του κάθε αντικειμένου, όσο και τις αξιολογήσεις χρηστών για αυτά, με σκοπό τη βελτίωση των προτάσεων.

Κατά την τελευταία δεκαετία, το matrix factorization (MF) έχει γίνει η πιο δημοφιλής προσέγγιση CF λόγω της ανώτερης απόδοσής του. Τα αρχικά μοντέλα MF σχεδιάστηκαν για να μοντελοποιήσουν τα ρητά σχόλια των χρηστών, αντιστοιχίζοντας τους χρήστες και τα αντικείμενα έτσι ώστε οι σχέσεις χρηστών-στοιχείων (δηλαδή βαθμολογίες) να μπορούν να συλληφθούν μέσω της χρήσης παραγόντων / factors' dot product. Συγκεκριμένα, αν το r_{ij} υποδηλώσει την βαθμολογία του χρήστη i στο στοιχείο j, μαθαίνουμε το διάνυσμα του χρήστη $u_i \epsilon \mathbb{R}^r$ και το διανυσματικό στοιχείο $v_j \epsilon \mathbb{R}^r$, έτσι ώστε το τελικό προϊόν τους $u_i \mathrm{T} v_j$ να προσεγγίζει το r_{ij}. Αυτή η διατύπωση οδηγεί στο πρόβλημα βελτιστοποίησης που ελαχιστοποιεί το μέσο σφάλμα / mean squared error:

$$\min_{\mathbf{u}_*, \mathbf{v}_*} \sum_{r_{ij} \in \mathcal{K}} \left(r_{ij} - \mathbf{u}_i^T \mathbf{v}_j\right)^2 + \lambda_u \left\|\mathbf{u}_i\right\|^2 + \lambda_v \left\|\mathbf{v}_i\right\|^2$$

Ωστόσο, είναι προβληματικό να εφαρμοστεί η μέθοδος matrix factorization σε "άρρητα" / implicit δεδομένα, καθώς παρατηρούνται μόνο θετικά σχόλια. Δεν είναι ορθό να αγνοηθούν τα δεδομένα που δεν υπάρχει κριτική από τον χρήστη, αλλά επίσης, δεν μπορεί να γίνει η

υπόθεση ότι η έλλειψη κριτικής ισοδυναμεί με αρνητική κριτική, καθώς δεν είναι γνωστό αν αυτές οι αλληλεπιδράσεις δεν συνέβησαν επειδή ο χρήστης δεν του άρεσε το στοιχείο ή ο χρήστης δεν το γνώριζε. Για την αντιμετώπιση αυτών των ζητημάτων, οι Hu et al. και Pan et al. πρότειναν την μέθοδο σταθμισμένης ομαλοποίησης / weighted regularized matrix factorization (WRMF) που περιλαμβάνει όλες τις μη παρατηρούμενες αλληλεπιδράσεις χρήστη-στοιχείου ως αρνητικά δείγματα και χρησιμοποιεί το βάρος της υπόθεσης c_i, j για να μειώσει την επίδραση αυτών των αβέβαιων δειγμάτων. Έτσι:

$$\min_{\mathbf{u}_*,\mathbf{v}_*} \sum_{r_{ij}\in\mathcal{K}} c_{ij}\left(r_{ij} - \mathbf{u}_i^T\mathbf{v}_j\right)^2 + \lambda_u \left\|\mathbf{u}_i\right\|^2 + \lambda_v \left\|\mathbf{v}_i\right\|^2$$

Όπου το c_{ij} είναι μεγαλύτερο για θετική κριτική και μικρότερο για μη παρατηρούμενες αλληλεπιδράσεις.

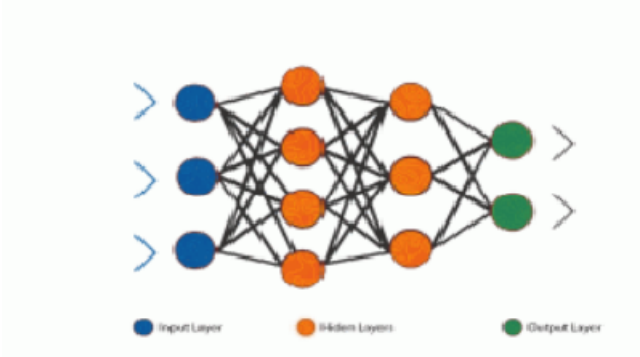
2.2.4 Νευρωνικά Δίκτυα

Πρόσφατα, τα νευρωνικά δίκτυα / neural networks πρωτοστατούν σε αυτόν τον τομέα. Η βαθιά μάθηση – Deep Learning (DL) έχει αποκτήσει αυξανόμενη προσοχή τα τελευταία χρόνια λόγω της βελτίωσης του τρόπου επεξεργασίας μεγάλων δεδομένων και της ικανότητάς της να μοντελοποιεί πολύπλοκα δεδομένα, όπως κείμενα και εικόνες. Η βαθιά μάθηση συμμετέχει στην έρευνα των συστημάτων συστάσεων και ξεπερνά τις παραδοσιακές μεθόδους. Οι βασικές υπό-κατηγορίες της βαθιάς μάθησης είναι τα: ANN, RNN και CNN τα οποία χρησιμοποιούνται για να μάθουν μια σύνθετη χαρτογράφηση του δικτύου.

Τεχνητό Νευρωνικό Δίκτυο

Ένας μόνο τεχνητός νευρώνας μπορεί να θεωρηθεί ισοδύναμος με την ανάπτυξη ενός μοντέλου λογιστικής παλινδρόμησης / logistic regression. Το Τεχνητό Νευρωνικό Δίκτυο / Artificial Neural Network (ANN) είναι μια ομάδα πολλαπλών αισθητήρων/ νευρώνων σε κάθε επίπεδο. Το ANN είναι επίσης γνωστό ως νευρωνικό δίκτυο τροφοδοσίας-προώθησης επειδή η επεξεργασία των εισόδων γίνεται μόνο προς τα εμπρός όπως φαίνεται στο σχήμα 2.5.

Ο ANN αποτελείται από 3 επίπεδα - Είσοδος, Απόκρυψη και Έξοδος. Το επίπεδο εισόδου / input layer δέχεται τις εισόδους, το κρυφό επίπεδο / hidden layer επεξεργάζεται τις εισόδους, και το επίπεδο εξόδου / output layer παράγει το αποτέλεσμα. Ουσιαστικά, κάθε



Σχήμα 2.5: Artificial Neural Networks 1 [3]

επίπεδο προσπαθεί να μάθει ορισμένα βάρη. Στο ANN κάθε επίπεδο διαθέτει ένα σύνολο νευρώνων.

Το Τεχνητό Νευρωνικό Δίκτυο είναι ικανό να μαθαίνει οποιαδήποτε μη γραμμική λειτουργία, και έχει την ικανότητα να μάθει βάρη που αντιστοιχίζουν οποιαδήποτε είσοδο στην έξοδο.

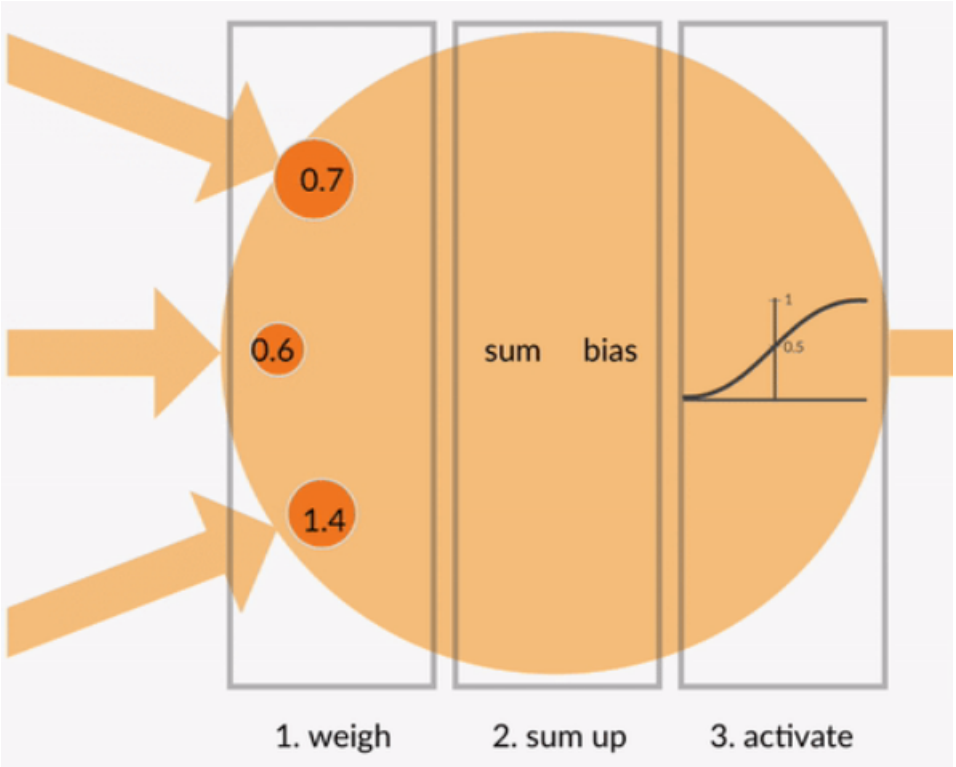
Ένας βασικό στοιχείο αποτελεί η συνάρτηση ενεργοποίησης. Οι συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικές ιδιότητες στο δίκτυο. Αυτό βοηθά το δίκτυο να μαθαίνει οποιαδήποτε σύνθετη σχέση μεταξύ εισόδου και εξόδου. Όπως φαίνεται στο σχήμα 2.6, η έξοδος σε κάθε νευρώνα είναι ένα σταθμισμένο άθροισμα των εισόδων.

Αναδρομικό Νευρωνικό Δίκτυο

Ένας περιορισμός επανάληψης στο κρυφό επίπεδο του ANN συνεπάγεται το Αναδρομικό Νευρωνικό Δίκτυο / Recurrent Neural Networks (RNN). Όπως φαίνεται στο σχήμα 2.7, το RNN έχει μια επαναλαμβανόμενη σύνδεση στο κρυφό επίπεδο. Αυτός ο περιορισμός βρόχου εξασφαλίζει ότι οι διαδοχικές πληροφορίες καταγράφονται στα δεδομένα εισόδου.

Το RNN καταγράφει τις διαδοχικές πληροφορίες που υπάρχουν στα δεδομένα εισόδου, δηλαδή την εξάρτηση μεταξύ των λέξεων στο κείμενο, ενώ κάνει προβλέψεις. Ένα παράδειγμα αποτελεί αυτό που απεικονίζεται στο σχήμα 2.8, όπου φαίνεται ότι η έξοδος (o1, o2, o3, o4) σε κάθε βήμα εξαρτάται όχι μόνο από την τρέχουσα λέξη αλλά και από τις προηγούμενες λέξεις.

Τα RNN μοιράζονται τις παραμέτρους σε διάφορα βήματα. Αυτό είναι ευρέως γνωστό ως κοινή χρήση παραμέτρων / Parameter Sharing - σχήμα 2.9, το οποίο έχει ως αποτέλεσμα τη μείωση των παραμέτρων και τη μείωση του υπολογιστικού κόστους.



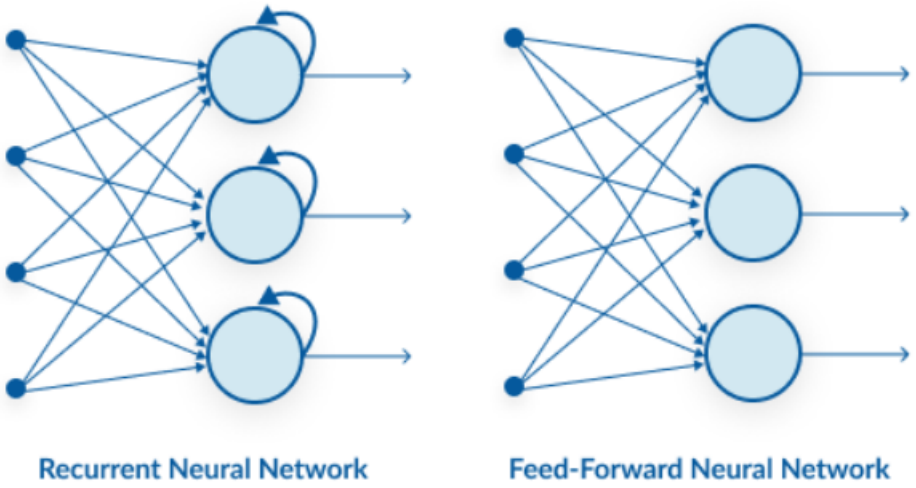
Σχήμα 2.6: Artificial Neural Networks 2 [3]

Νευρωνικό Δίκτυο συνελικτικής σύνδεσης

Τα δομικά στοιχεία των Νευρωνικών Δικτύων συνελικτικής σύνδεσης / Convolution Neural Network (CNN) είναι φίλτρα / kernels. Τα φίλτρα χρησιμοποιούνται για την εξαγωγή των σχετικών χαρακτηριστικών από την είσοδο με τη χρήση της λειτουργίας συγκέντρωσης.

Το CNN μαθαίνει τα φίλτρα αυτόματα χωρίς να τα αναφέρει ρητά. Αυτά τα φίλτρα βοηθούν στην εξαγωγή των κατάλληλων και σχετικών χαρακτηριστικών από τα δεδομένα εισόδου. Ένα παράδειγμα του CNN αποτελεί το σχήμα 2.10, όπου φαίνεται ότι το CNN αποτυπώνει τα χωρικά χαρακτηριστικά μιας εικόνας. Τα χωρικά χαρακτηριστικά αναφέρονται στη διάταξη των pixel και στη σχέση μεταξύ τους σε μια εικόνα. Μας βοηθούν στον ακριβή προσδιορισμό του αντικειμένου, της θέσης ενός αντικειμένου, καθώς και της σχέσης του με άλλα αντικείμενα σε μια εικόνα.

Το CNN ακολουθεί επίσης την έννοια της κοινής χρήσης παραμέτρων / parameter sharing. Εφαρμόζεται ένα φίλτρο σε διαφορετικά μέρη μιας εισόδου για τη δημιουργία ενός χάρτη δυνατοτήτων όπως φαίνεται και στο παράδειγμα του σχήματος 2.11, όπου ο χάρτης χαρακτηριστικών 2*2 δημιουργείται με την ολίσθηση του ίδιου φίλτρου 3*3 σε διαφορετικά μέρη μιας εικόνας.



Σχήμα 2.7: Recurrent Neural Networks 1 [3]

2.2.5 Γνωστές εφαρμογές των Νευρωνικών Δικτύων

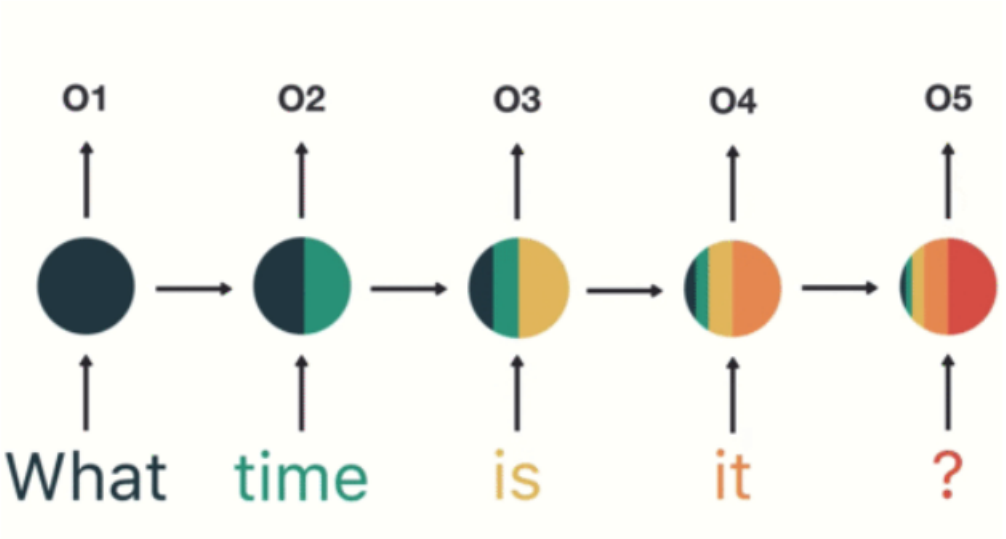
Μέθοδος Boltzmann με περιορισμούς

Η μέθοδος Boltzmann με περιορισμούς / Restricted Boltzmann Machine (RBM) είναι ένα Νερωνικό Δίκτυο / ANN δύο στρώσεων, το οποίο περιέχει μόνο ένα επίπεδο εισόδου και ένα κρυφό επίπεδο. Η απόδοση μάθησής του είναι πολύ καλή λόγω των περιορισμών που υπάρχουν στη συνδεσιμότητα των νευρώνων μεταξύ των επιπέδων. Στο κρυμμένο επίπεδο Hi υπολογίζεται η βαθμολογία για κάθε κόμβο πολλαπλασιάζοντας τέσσερις εισόδους από το ορατό επίπεδο / visible layer με τα βάρη τους. Στη συνέχεια, το άθροισμα αυτών των προϊόντων προστίθεται. Τέλος, μια συνάρτηση ενεργοποίησης μεταβιβάζει τα αποτελέσματα στην έξοδο [4].

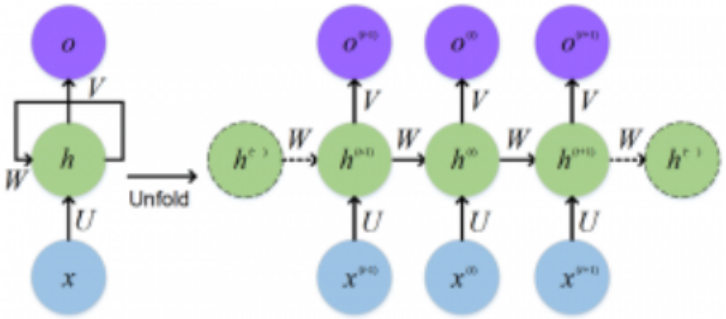
Η λογική αυτή οπτικοποιείται στο σχήμα 2.12.

Νευρωνικά Δίκτυα πολλαπλών επιπέδων

Τα νευρωνικά δίκτυα με πολλαπλά επίπεδα / Multi-Layer Perceptron (MLP) θεωρούνται ο απλούστερος τύπος ANN. Αποτελούνται από ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου. Αυτά τα επίπεδα περιέχουν συναρτήσεις ενεργοποίησης / activation functions ρυθμίζοντας τα βάρη κατά την εκπαίδευση. Όταν τα κρυφά επίπεδα λαμβάνουν την αναπαράσταση εισόδου, χαρτογραφούν την είσοδο και την έξοδο εφαρμόζοντας μη γραμμικότητα, όπως φαίνεται στο Σχήμα 2.13.



Σχήμα 2.8: Παράδειγμα Recurrent Neural Network [3]



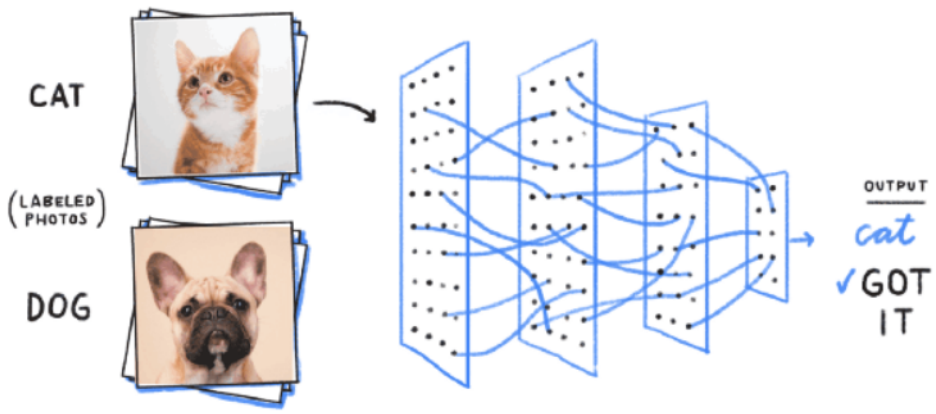
Σχήμα 2.9: Recurrent Neural Network 2 [3]

Η έξοδος κάθε νευρώνα υπολογίζεται με

$$h = g\left(W * x_i + b\right) \tag{2.1}$$

Όπου το W αντιπροσωπεύει το βάρος μεταξύ δύο νευρώνων των αντίστοιχων επιπέδων εισόδου και των κρυφών επιπέδων. Ενώ h και b υποδηλώνουν το κρυμμένο νευρώνα και το διάνυσμα πόλωσης, αντίστοιχα, το g χρησιμοποιείται ως συνάρτηση ενεργοποίησης. Μεταξύ των πιο γνωστών είναι relu, tanh, sigmoid, κλπ. Τέλος, η έξοδος του δικτύου προβλέπεται από τη διάδοση των σταθμισμένων αθροισμάτων των κρυμμένων επιπέδων στο επίπεδο εξόδου, το οποίο εφαρμόζει τη μη γραμμικότητα στην μεταβιβαζόμενη τιμή για να παράγει την τελική πρόβλεψη ως εξής:

$$\widehat{y} = g(W * h + b) \tag{2.2}$$

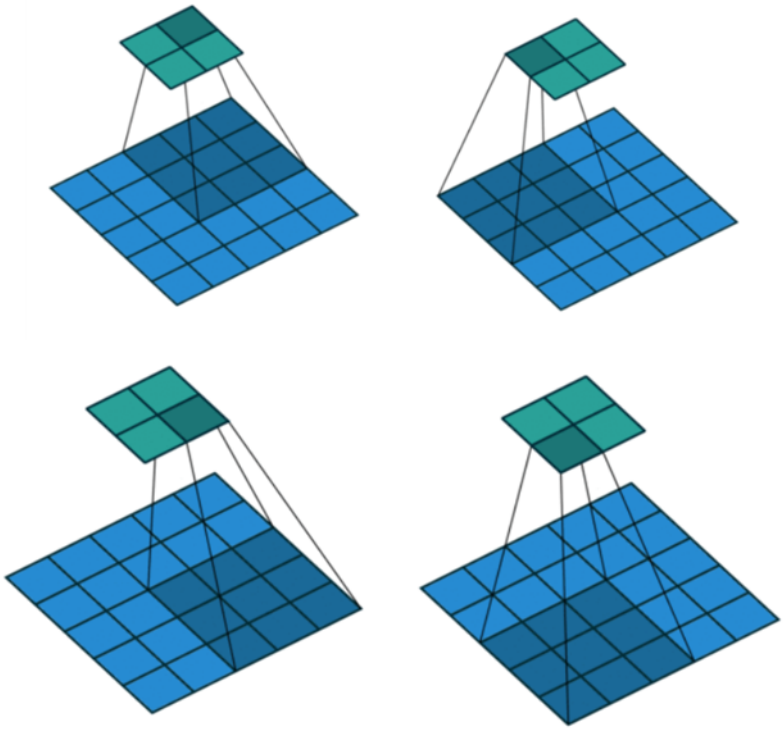


Σχήμα 2.10: Convolution Neural Network παράδειγμα 1 [3]

όπου το $\widehat{y}$ υποδηλώνει την έξοδο και το g αντιπροσωπεύει τη μη γραμμική συνάρτηση ενεργοποίησης. Για τη μετάδοση στο επόμενο επίπεδο χρησιμοποιούνται μη γραμμικές λειτουργίες ενεργοποίησης.

Αυτόματοι Κωδικοποιητές

Ένας αυτόματος κωδικοποιητής / autoencoder είναι ένα νευρωνικό δίκτυο που εκπαιδεύεται με μη εποπτευόμενο τρόπο. Οι αυτόματοι κωδικοποιητές είναι δημοφιλείς για τη μείωση της διάστασης, έτσι ώστε η είσοδος τους να συμπιέζεται σε μια αναπαράσταση χαμηλών διαστάσεων, ενώ διατηρείται η έννοια των χαρακτηριστικών της. Συγκεκριμένα, οι αυτόματοι κωδικοποιητές είναι οι βέλτιστοι στην αναπαράσταση κειμένου. Το δίκτυο του αυτόματου κωδικοποιητή αποτελείται από δύο βασικά μέρη: τον κωδικοποιητή και τον αποκωδικοποιητή. Στην αρχή, ο κωδικοποιητής λαμβάνει μια είσοδο και τη συμπιέζει σε ένα λανθάνον διάνυσμα, και από την άλλη πλευρά, ο αποκωδικοποιητής χρησιμοποιείται στη συνέχεια για να δημιουργήσει ξανά την είσοδο. Και τα δύο μέρη αποτελούνται συνήθως από πολλά κρυφά επίπεδα. Ωστόσο, όλα τα τρέχοντα υπάρχοντα μοντέλα που βασίζονται στον αυτόματο κωδικοποιητή / autoencoder-based models δεν έχουν τη δυνατότητα διαφορετικής αντιμετώπισης των διαφόρων χαρακτηριστικών, κάτι που μπορεί να οδηγήσει σε μια μη βέλτιστη απόδοση. Στην πραγματικότητα, η σημασία των διαφορετικών χαρακτηριστικών δεν είναι ομοιόμορφη και ορισμένα χαρακτηριστικά είναι πιθανό να έχουν πιο σημαντική συμβολή από άλλα, κάτι το οποίο τα μοντέλα αυτά δεν μπορούν να το αξιοποιήσουν.

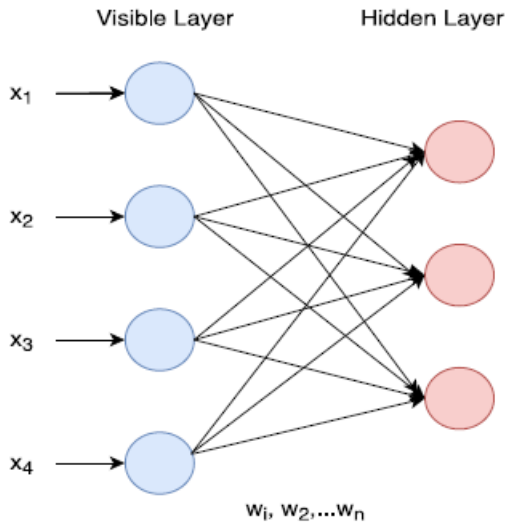


Σχήμα 2.11: Convolution Neural Network παράδειγμα 2 [3]

Μηχανισμός Προσοχής

Η ιδέα του μηχανισμού προσοχής / attention mechanism παρακινείται από την ανθρώπινη όραση, και πώς τα μάτια μας δίνουν προσοχή και επικεντρώνονται σε ένα συγκεκριμένο τμήμα μιας εικόνας, ή συγκεκριμένες λέξεις σε μια πρόταση. Με τον ίδιο τρόπο, ο μηχανισμός προσοχής στον τομέα της βαθιάς μάθησης μπορεί να εξηγηθεί βασικά ως ένας φορέας διαφορετικών βαρών που εκφράζει τον τρόπο με τον οποίο τα στοιχεία εισόδου είναι σημαντικά με διαφορετικούς τρόπους. Επομένως, η έννοια του μηχανισμού προσοχής είναι ότι δεν είναι εξίσου σημαντικά όλα τα τμήματα εισόδου, δηλαδή μόνο λίγα τμήματα είναι σημαντικά για τις προβλέψεις. Ο μηχανισμός αυτός είχε αρχικά σχεδιαστεί για ένα νευρωνικό σύστημα μετάφρασης, και στη συνέχεια εφαρμόστηκε με επιτυχία σε άλλα περιβάλλοντα όπως η ταξινόμηση εικόνας και η ταξινόμηση εγγράφων.

Πρόσφατα, ο μηχανισμός προσοχής έχει υιοθετηθεί συχνά σε εφαρμογές επεξεργασίας φυσικής γλώσσας / natural language processing (NLP), όπου μπορεί να δώσει προσοχή σε διαφορετικές λέξεις του κειμένου. Συγκεκριμένα, η προσοχή μπορεί να προσδιορίσει τη συ-



Σχήμα 2.12: RBM [4]

νεισφορά κάθε λέξης του κειμένου υπολογίζοντας το σταθμισμένο άθροισμα όλων των λέξεων και, στη συνέχεια, εκχωρώντας διαφορετικές βαθμολογίες σε κάθε λέξη μέσω ευθυγραμμισμένης βαθμολόγησης / alignment score function.

2.3 Μέθοδοι νευρωνικών δικτύων για επιστημονικά δημοσιεύματα

2.3.1 Βαθιά μάθηση στην έρευνα μοντέλων σύστασης παραπομπών

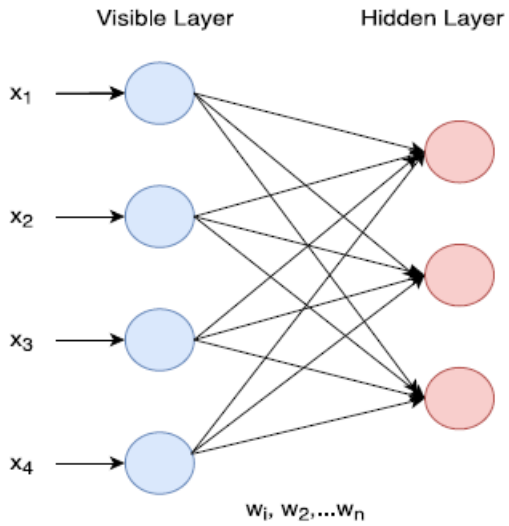
Η δημοσίευση [4] αποτελεί μια διεξοδική μελέτη που ταξινομεί τα συστήματα συστάσεων και εξετάζει τα

(α) δυνατά και αδύνατα σημεία τους,

(β) τον τρόπο αξιολόγησής τους,

(γ) τα δημοφιλή σύνολα δεδομένων και τις προκλήσεις που αντιμετωπίζουν.

Ως εκ τούτου, αυτή η έρευνα, παρουσιάζει μια νέα προσέγγιση για μοντέλα βαθιάς μάθησης που παρέχουν συστάσεις. Η προσέγγισή της μελετάει 35 μοντέλα Βαθιάς μάθησης / DL-based, και χρησιμοποιεί τα ακόλουθα έξι κριτήρια: παράγοντες δεδομένων, μέθοδοι αναπαράστασης δεδομένων, μεθοδολογίες, διαφορετικοί τύποι συστάσεων, προβλήματα που αντιμετωπίζονται και εξατομίκευση. Επιπλέον, παρουσιάζει μια συγκριτική ανάλυση των μοντέλων που χρησιμοποιούν το ίδιο σύνολο μετρήσεων αξιολόγησης και το ίδιο συνόλων



Σχήμα 2.13: MLP [4]

δεδομένων ως είσοδο. Στη συνέχεια παρέχεται μια επισκόπηση των συνόλων δεδομένων και των μετρήσεων που έχουν υιοθετήσει τα διευρυνθέντα μοντέλα. Επίσης, η έρευνα εξετάζει και αναλύει τις μετρήσεις αξιολόγησης και τα σύνολα δεδομένων που υιοθετούνται από τα μοντέλα. Στην ανάλυση αυτή, λαμβάνονται υπόψη μόνο τα μοντέλα που χρησιμοποιούν το ίδιο σύνολο μετρήσεων και σύνολο δεδομένων αξιολόγησης. Από την έρευνα προέκυψε ότι η συγχώνευση / embedding είναι η πιο δημοφιλής μέθοδος που υιοθετείται σε 15 από τα 35 μοντέλα που παρουσιάζονται. Αυτή η μέθοδος επιφέρει αξιοσημείωτη βελτίωση στην ακρίβεια αφού εκμεταλλεύεται όλες τις πληροφορίες, τη σημασιολογία, και τις βοηθητικές πληροφορίες που αντιστοιχούν σε επιστημονικά δημοσιεύματα και χρήστες. Οι παραλλαγές των RNN, όπως η LSTM, και η BiLSTM, είναι δεύτερες, δεδομένου ότι αποτυπώνουν μακροπρόθεσμες εξαρτήσεις.

Τα συστήματα συστάσεων επιστημονικών δημοσιεύσεων χρησιμοποιούν διαφορετικούς παράγοντες και χαρακτηριστικά, όπως: περιεχόμενο δημοσίευσης, ετικέτες/λέξεις-κλειδιά, προφίλ χρήστη, πληροφορίες χώρου δημοσίευσης, δίκτυο παραπομπών, πληροφορίες συντάκτη και αξιολογήσεις, ενώ συνιστώνται σχετικές αναφορές στους χρήστες. Η χρήση τέτοιων πληροφοριών μπορεί να εκπροσωπήσει καλύτερα τα συμφέροντα των ερευνητών και να βοηθήσει στην κατανόηση των αναγκών τους.

Τέλος, η έρευνα ολοκληρώνεται με τάσεις και μελλοντικές κατευθύνσεις για την περαιτέρω ενίσχυση της έρευνας σε αυτόν τον τομέα.

Για περισσότερες πληροφορίες: [4]

2.3.2 Συστάσεις μέσω αναδρομικών νευρωνικών δικτύων προσοχής

Στο δημοσίευμα [5], έχει αναπτυχθεί μια νέα μέθοδος ανάπτυξης ενός συστήματος συστάσεων : heterogeneous knowledge embedding based attentive recurrent neural networks, ώστε να συστήσει επιστημονικά δημοσιεύματα και παραπομπές.

Η μέθοδος αυτή φαίνεται σχηματική στην εικόνα 2.14.

Πρώτα, δημιουργείται ένα βιβλιογραφικό δίκτυο που περιλαμβάνει εφημερίδες, συγγραφείς, συνδέσμους, χώρους δημοσίευσης, έτος έκδοσης, παραπομπές και ημερομηνία έκδοσης. Με τη χρήση του TransD, όλες οι γραφικές οντότητες και σχέσεις μπορούν να διανυσματοποιηθούν για περαιτέρω υπολογισμούς. Επιπλέον, δημιουργήθηκαν αριθμητικά χαρακτηριστικά για τον τίτλο κάθε χαρτιού. Στη συνέχεια, δημιουργείται ένα "προσεκτικό αμφίδρομο RNN" / attentive bidirectional RNN για να προτείνει έγγραφα και αναφορές με βάση την ταυτότητα ενός χρήστη με ένα ερώτημα περιορισμένου μήκους. Τα αποτελέσματα του πειράματος δείχνουν τη σκοπιμότητα της μεθόδου αυτής, τόσο στον αριθμό όσο και στην ποιότητα των κορυφαίων-N συνιστώμενων αναφορών. Σε σύγκριση με τα υπάρχοντα μοντέλα, το μοντέλο έχει βελτιώσει τις μετρικές MRR και NDCG κατά περίπου 4,8% και 2,4%, αντίστοιχα.

Για περισσότερες πληροφορίες: [5]

2.3.3 Σύσταση παραπομπών: προσεγγίσεις και σύνολα δεδομένων

Στο άρθρο [6] πραγματοποιείται μια ενδελεχής εισαγωγή στην έρευνα για τα συστήματα συστάσεων παραπομπών. Η σύσταση παραπομπής αναφέρεται στη σύσταση κατάλληλων παραπομπών για ένα απόσπασμα κειμένου μέσα σε ένα έγγραφο. Βοηθά τον χρήστη να τεκμηριώσει ένα συγκεκριμένο κείμενο (π.χ. μια επιστημονική ιδέα) σε ένα έγγραφο εισόδου συνιστώντας δημοσιεύσεις που μπορούν να χρησιμοποιηθούν ως παραπομπές. Παρουσιάζεται μια επισκόπηση των προσεγγίσεων και των συνόλων δεδομένων για τη σύσταση παραπομπής και εντοπίζονται διαφορές και κοινές πτυχές χρησιμοποιώντας ποικίλες διαστάσεις. Το βασικό πρότυπο ανάπτυξης ενός συστήματος συστάσεων παραπομπών φαίνεται στο σχήμα 2.15

Η διαδικασία της επεξεργασίας / processing κατά το "Offline step" (= φάση εκπαίδευσης στο σύστημα μηχανικής μάθησης), στο οποίο εκπαιδεύεται ένα μοντέλο σύστασης με βάση μια συλλογή εγγράφων, αποτελείται από τα εξής βήματα:

- Εξαγωγή αναφοράς
- Εξαγωγή και αναπαράσταση περιβάλλοντος παραπομπής
- Εκμάθηση μοντέλων

Ενώ αντίστοιχα για το “Online step” όπου το σύστημα εφαρμόζεται σε ένα νέο εισερχόμενο έγγραφο κειμένου :

- Εξαγωγή αναφοράς (προαιρετική)
- Εξαγωγή και αναπαράσταση περιβάλλοντος παραπομπής
- Υπόδειγμα εφαρμογής
- Εμπλουτισμός κειμένου

Η έρευνα δείχνει ότι η ανάπτυξη του παραπάνω συστήματος συστάσεων μπορεί να γίνει με διάφορες μεθόδους, όπως: μοντέλα βασισμένα σε χαρακτηριστικά / hand-crafted feature-based models, μοντέλα με θέμα / topic models, μοντέλα αυτόματης μετάφρασης / machine translation models, και νευρωνικά μοντέλα / neural network models.

Στη συνέχεια, αναλύονται οι μέθοδοι αξιολόγησης και σκιαγραφούνται οι γενικές προκλήσεις της αξιολόγησης και ο τρόπος αντιμετώπισής τους. Εξετάζονται επίσης τα σύνολα δεδομένων που μπορούν να χρησιμοποιηθούν για την ανάπτυξη και την αξιολόγηση των προτάσεων παραπομπής. Ωστόσο, τα σύνολα δεδομένων διαφέρουν σημαντικά ως προς το μέγεθος και την ποιότητά τους. Η αξιολόγηση των συστάσεων παραπομπής μπορεί επίσης να εξαρτάται από την επιστημονική πειθαρχία και τη συγκεκριμένη περίπτωση χρήσης. Όσον αφορά την αξιολόγηση, οι προσεγγίσεις αξιολογούνται με βάση πολύ διαφορετικές μετρήσεις και διαφορετικά σύνολα δεδομένων, καθιστώντας δύσκολη την αξιολόγηση της εγκυρότητας και της προόδου των μεμονωμένων προσεγγίσεων. Επιπλέον, οι προσεγγίσεις συχνά συγκρίνονται σε περιορισμένο βαθμό με τις υπάρχουσες προσεγγίσεις. Συνεπώς, οι προσεγγίσεις έχουν αξιολογηθεί μάλλον μονομερώς και όχι σε όλους τους κλάδους.

Τέλος, αναφέρεται ότι η συγκεκριμένη έρευνα αφορά σε συστάσεις παραπομπής για επιστημονικές δημοσιεύσεις, καθώς αυτός ο τύπος εγγράφου έχει μελετηθεί περισσότερο σε αυτόν τον τομέα. Ωστόσο, πολλές από τις παρατηρήσεις και τις συζητήσεις που περιλαμβάνονται στην παρούσα έρευνα ισχύουν και για άλλους τύπους κειμένου, όπως άρθρα ειδήσεων και εγκυκλοπαίδεια.

Για περισσότερες πληροφορίες: [6]

2.4 Μέθοδος CATA++

Τα προαναφερθέντα υπάρχοντα είδη μοντέλων συστημάτων συστάσεων έχουν δύο σημαντικούς περιορισμούς. Πρώτον, υποθέτουν ότι όλα τα χαρακτηριστικά των στοιχείων συμβάλλουν εξίσου στην τελική πρόβλεψη. Δεύτερον, εστιάζουν μόνο σε ορισμένα μέρη των δεδομένων και παραμελούν άλλα μέρη που θα μπορούσαν επίσης να χρησιμοποιηθούν για τη βελτίωση της αποτελεσματικότητας των συστάσεων. Ως εκ τούτου, ο πρώτος περιορισμός αντιμετωπίζεται χρησιμοποιώντας τον "μηχανισμό προσοχής" / attention mechanism, ενώ ο δεύτερος περιορισμός χρησιμοποιώντας παράλληλα δύο "προσεκτικούς αυτόματους κωδικοποιητές" / attentive autoencoders, οι οποίοι εκπαιδεύονται χωριστά για να εντοπίζουν τα χαρακτηριστικά του κάθε αντικειμένου με μεγαλύτερη ακρίβεια. Πιο συγκεκριμένα, παρακάτω αναλύεται η μέθοδος Collaborative Dual Attentive Autoencoder (CATA++) που χρησιμοποιείται για να προτείνει επιστημονικές εργασίες. Η "τεχνική προσοχής" ενσωματώνεται κατά τη διαδικασία βαθιάς εκμάθησης των χαρακτηριστικών με στόχο να βελτιωθεί η ποιότητα των προτάσεων αξιοποιώντας τα διάφορα δεδομένα του άρθρου (π.χ. τίτλος, περίληψη, ετικέτες και παραπομπές μεταξύ των συγγραμμάτων). Το μοντέλο αυτό είναι πιθανόν το πρώτο που χρησιμοποιεί όλα τα χαρακτηριστικά του άρθρου, συμπεριλαμβανομένων τίτλου, περίληψης, ετικετών και παραπομπών, μαζί σε ένα μοντέλο βαθιάς εκμάθησης, συνδέοντας παράλληλα δύο attentive autoencoders. Τα χαρακτηριστικά που προκύπτουν από τον κάθε "Autoencoder" στη συνέχεια χρησιμοποιούνται στη μέθοδο matrix factorization (MF) για την ανάπτυξη συστάσεων. Τέλος, το μοντέλο αξιολογείται χρησιμοποιώντας σύνολα δεδομένων από το Aminer και συγκεκριμένα το dblp dataset.

Το πρώτο βήμα της μεθόδου CATA++ αποτελεί η ανάπτυξη του αυτόματου κωδικοποιητή / attentive autoencoder. Για την ανάπτυξη του κωδικοποιητή αρχικά ο κώδικας χρησιμοποιούσε ως συνάρτηση ενεργοποίησης / activation function την ReLu, αλλά κατά την ανάπτυξη της διπλωματικής εργασίας δοκιμάστηκε η χρήση διάφορων συναρτήσεων και συμπεραίνετε πως η μέγιστη αποτελεσματικότητα επιτυγχάνεται μέσω της SineRelu συνάρτησης, και χρησιμοποιώντας ως αρχικοποιητή βαρών / weight initializer τον he_normal.

Εφαρμόζεται η τεχνική ομαλοποίησης / batch normalization (BN) σε κάθε ένα από τα στρώματα του αυτόματου κωδικοποιητή για να επιτευχθεί σωστή κατανομή του αποτελέσματος. Η ενσωμάτωση της BN στο μοντέλο το καθιστά πιο αποτελεσματικό, καθώς παρέχει κανονικοποίηση στην εκπαίδευση του νευρικού δικτύου.

Επιπλέον, τοποθετείται ένα επίπεδο προσοχής / attention layer στη μέση του αυτόμα-

του κωδικοποιητή, έτσι ώστε μόνο τα σημαντικά στοιχεία της εξόδου του κωδικοποιητή να επιλέγονται για την ανασυγκρότηση της αρχικής εισόδου. Για να γίνει αυτό, χρησιμοποιείται η συνάρτηση softmax. Μετά από αυτό, εφαρμόζεται στην έξοδο της προηγούμενης συνάρτησης και στην έξοδο του κωδικοποιητή η συνάρτηση πολλαπλού πολλαπλασιασμού / element-wise multiplication function. Τέλος, εφαρμόζεται η "binary cross-entropy" συνάρτηση για καθένα από τους αυτόματους κωδικοποιητές.

Το Probabilistic Matrix Factorization (PMF) είναι ένα πιθανοτικό γραμμικό μοντέλο όπου οι προηγούμενες κατανομές των προτιμήσεων των χρηστών και των περιεχομένων των στοιχείων προέρχονται από την κατανομή Gauss. Το μοντέλο CATA++ εκμεταλλεύεται το περιεχόμενο των στοιχείων και τα εκπαιδεύει μέσω δύο ξεχωριστών, παράλληλων κωδικοποιητών. Η έξοδος αυτών των δύο διαχωρισμένων κωδικοποιητών χρησιμοποιούνται μαζί ως οι προηγούμενες πληροφορίες των παραγόντων των στοιχείων του PMF.

Τέλος, μόλις ολοκληρωθεί η εκπαίδευση, οι βαθμολογίες πρόβλεψης του μοντέλου υπολογίζονται ως το γινόμενο των διανυσμάτων των χρηστών επί των άρθρων.

Το μοντέλο του CATA++ αλγορίθμου μπορεί να απεικονιστεί γραφικά όπως φαίνεται στο Σχήμα 2.16.

Αντίστοιχα, ο αλγόριθμος ανάπτυξής του σε ψευδοκώδικα αναπαριστάται στο Σχήμα 2.17, όπου οι εξισώσεις 9 και 10 αποτελούν τις εξισώσεις που χρησιμοποιούνται κατά την ανάπτυξη του PMF.

Το διάγραμμα ροής της CATA++ μεθόδου απεικονίζεται στο Σχήμα 2.18.

Για περισσότερες λεπτομέρειες: [7]

2.5 Μέθοδος CVAE

Τα σύγχρονα συστήματα συστάσεων συνήθως χρησιμοποιούν την μέθοδο Collaborative Filtering που αναλύθηκε και παραπάνω. Ωστόσο, λόγω των μειονεκτημάτων, όπως η έλλειψη αρκετών πληροφοριών / sparsity, η δυσκολία κατά την εκκίνηση του αλγορίθμου / cold start, κ.λπ., έχει δοθεί περισσότερη προσοχή στις υβριδικές μεθόδους που λαμβάνουν υπόψη τόσο την αξιολόγηση όσο και τις πληροφορίες περιεχομένου. Οι περισσότερες μέθοδοι που έχουν ήδη αναπτυχθεί δεν μπορούν μέσω της εξέτασης της μορφής ενός κειμένου ή μέσω του περιεχομένου του να κάνουν σωστές συστάσεις, το οποίο τις καθιστά πολύ περιορισμένες. Αυτή η μέθοδος προτείνει ένα μοντέλο Bayesian που ονομάζεται collaborative variational autoen-

coder (CVAE) που εξετάζει τόσο την αξιολόγηση όσο και το περιεχόμενο για την παραγωγή ενός συστήματος συστάσεων.

Το μοντέλο μαθαίνει τόσο για τις σχέσεις των περιεχομένων μεταξύ τους χωρίς επίβλεψη, όσο επίσης και τις έμμεσες σχέσεις μεταξύ αντικειμένων και χρηστών και από το περιεχόμενο αλλά και από την αξιολόγηση.

Τα μοντέλα βαθιάς μάθησης έχουν δείξει πρόσφατα μεγάλες δυνατότητες εκμάθησης αποτελεσματικών αναπαραστάσεων και έχουν επιτύχει προηγμένες επιδόσεις. Αν και η μέθοδος αυτή είναι ελκυστική, λίγες προσπάθειες έχουν γίνει για την ανάπτυξη μοντέλων βαθιάς μάθησης για συστήματα συστάσεων. Είναι δύσκολο να αναπτυχθούν μοντέλα βαθιάς μάθησης για να συλλάβουν και να μάθουν την "άρρητη" / implicit σχέση μεταξύ αντικειμένων (και χρηστών), η οποία, αντίθετα, είναι η βέλτιστη κατά την χρήση των πιθανοτικών γραφικών μοντέλων. Αυτό απαιτεί την συγχώνευση γραφικών μοντέλων Bayesian και μοντέλων βαθιάς μάθησης για να επωφεληθούν από τα καλύτερα και των δύο κόσμων. Συνεπώς, προτείνεται ένα μοντέλο που ονομάζεται συνεργατικός διαφορικός αυτόματος κωδικοποιητής / Bayesian deep generative model called collaborative variational autoencoder (CVAE) για να μοντελοποιήσουμε από κοινού το περιεχόμενο και τις πληροφορίες αξιολόγησης σε ένα συνεργατικό φιλτράρισμα. Το μοντέλο μαθαίνει βαθιές αναπαραστάσεις από τα δεδομένα περιεχομένου χωρίς επίβλεψη και επίσης μαθαίνει έμμεσες σχέσεις μεταξύ αντικειμένων και χρηστών τόσο από το περιεχόμενο όσο και από τη βαθμολογία.

Το CVAE είναι ένα γενεσιουργό λανθάνον μοντέλο μεταβλητής / generative latent variable model, όπου τα περιεχόμενα των στοιχείων δημιουργούνται από τις λανθάνουσες μεταβλητές περιεχομένου τους και οι αξιολογήσεις των στοιχείων από τους χρήστες δημιουργούνται μέσω μεταβλητών λανθάνοντος στοιχείου. Οι μεταβλητές λανθάνοντος στοιχείου ενσωματώνουν τόσο τις πληροφορίες περιεχομένου μέσω μεταβλητών λανθάνοντος περιεχομένου όσο και τις συνεργατικές πληροφορίες μέσω λανθάνοντων μεταβλητών συνεργασίας, και συνδυάζουν τις υβριδικές πληροφορίες μαζί με βαθιά αρχιτεκτονική.

Ο παραδοσιακός PMF λαμβάνει υπόψη μόνο τις συνεργατικές πληροφορίες για την πρόβλεψη της αξιολόγησης και δεν χρησιμοποιεί το περιεχόμενο των στοιχείων. Το προτεινόμενο μοντέλο CVAE κατασκευάζει ένα μοντέλο μεταβλητής για το περιεχόμενο και εκχωρεί μια μεταβλητή λανθάνοντος περιεχομένου Z_j σε κάθε στοιχείο j.

Το δίκτυο συμπερασμάτων / inference network συνάγει το z από 2 διαδρομές:

1. δημιουργία περιεχομένου x

2. δημιουργία της βαθμολογίας R

Και απεικονίζεται στο παρακάτω Σχήμα 2.19.

Συνοπτική περιγραφή του τρόπου ανάπτυξης του CVAE μοντέλου:

1. Το πρώτο βήμα απεικονίζεται στο Σχήμα 2.20.

Ως x_j συμβολίζεται το περιεχόμενο ενός στοιχείου j και παράγεται από την λανθάνουσα μεταβλητή z_j μέσω χρήσης ενός νευρωνικού δικτύου.

Δεδομένης της λανθάνουσας μεταβλητής z, το x δημιουργείται μέσω ενός δικτύου αντίληψης πολλαπλών επιπέδων / multi-layer perception network (MLP).

2. Το δεύτερο βήμα απεικονίζεται αντίστοιχα στο Σχήμα 2.21.

Οι μεταβλητές λανθάνοντων στοιχείων συντίθενται με τη μεταβλητή λανθάνοντος περιεχομένου συνεργασίας και τη μεταβλητή λανθάνοντος περιεχομένου:

$$v_j = v_j^{\dagger} + z_j$$

3. Τέλος, το τρίτο βήμα απεικονίζεται αντίστοιχα στο Σχήμα 2.22.

Η βαθμολογία R προκύπτει από την κανονική κατανομή / Normal distribution που επικεντρώνεται στο γινόμενο μεταξύ των λανθανόντων μεταβλητών.

Το διάγραμμα ροής της CVAE μεθόδου απεικονίζεται στα Σχήματα 2.23 και 2.24.

Για περισσότερες λεπτομέρειες: [2]

2.6 Μέθοδος RVAE

Έχουν προταθεί πολλές μέθοδοι πρόβλεψης δεσμών / link prediction methods. Οι μέθοδοι που βασίζονται σε συνδέσεις αναζητούν κρυφά μοτίβα μεταξύ των δεδομένων. Κατά μήκος αυτής της γραμμής, πρόσφατα έργα για την πρόβλεψη συνδέσμων επικεντρώνονται σε λανθάνοντα μοντέλα μεταβλητών / latent variable models, συμπεριλαμβανομένων τόσο των παραμετρικών όσο και των μη παραμετρικών Bayesian μεθόδων. Τα λανθάνοντα μοντέλα μεταβλητών επιδιώκουν να μάθουν την λανθάνουσα αναπαράσταση από τα δεδομένα μέσω της μεταξύ τους σύνδεσης, και να βελτιώσουν την ευκολία εφαρμογής των μεθόδων μηχανικής μάθησης. Αν και ισχυρά, αυτά τα λανθάνοντα μοντέλα μεταβλητών αντιπροσωπεύουν μόνο τις δομές συνδέσεων των δεδομένων και αγνοούν τα χαρακτηριστικά τους.

Από την άλλη πλευρά, οι μέθοδοι που βασίζονται σε χαρακτηριστικά, κάνουν συστάσεις που βασίζονται στα χαρακτηριστικά των δεδομένων και μετατρέπουν το πρόβλημα σε

πρόβλημα ταξινόμησης.

Πρόσφατα, έχει δοθεί προσοχή σε υβριδικές μεθόδους που μοντελοποιούν από κοινού τα χαρακτηριστικά των δεδομένων και τις δομές σύνδεσης για να αξιοποιήσουν τα καλύτερα χαρακτηριστικά και των δύο μοντέλων και να επιτύχουν κορυφαία απόδοση.

Διαμορφώνοντας το πρόβλημα πρόβλεψης συνδέσμων σε ένα πιθανοτικό μοντέλο παραγωγής με νευρωνικά δίκτυα, το προτεινόμενο μοντέλο RVAE μπορεί ταυτόχρονα να μάθει μια αποτελεσματική λανθάνουσα αναπαράσταση από το περιεχόμενο και τις δομές συνδέσμων μεταξύ κόμβων για συστάσεις.

Για να αντιμετωπιστούν οι παραπάνω προκλήσεις, προτείνεται ένα Bayesian μοντέλο βαθιάς μάθησης που ονομάζεται αυτοκωδικοποιητής σχεσιακής παραλλαγής / relational variational autoencoder (RVAE) για την από κοινού διαμόρφωση των χαρακτηριστικών του περιεχομένου και των δεσμών μεταξύ των δεδομένων. Το μοντέλο μαθαίνει τις έντονα λανθάνουσες απεικονίσεις από το περιεχόμενο των δεδομένων με μη εποπτευόμενο τρόπο, ενώ μαθαίνει επίσης τις δομές σύνδεσης μεταξύ των δεδομένων τόσο από το περιεχόμενο όσο και από τις συνδέσεις.

Στο RVAE μοντέλο τα περιεχόμενα των κόμβων δημιουργούνται από τις λανθάνουσες μεταβλητές τους και οι σύνδεσμοι μεταξύ των κόμβων δημιουργούνται μέσω της αλληλεπίδρασης των λανθανόντων μεταβλητών. Το RVAE, λαμβάνοντας υπόψη τόσο το περιεχόμενο όσο και τις πληροφορίες δομής συνδέσεων, μπορεί να μάθει μια καλή αναπαράσταση ειδικά για την εργασία πρόβλεψης συνδέσεων και να επιτύχει καλή απόδοση πρόβλεψης συνδέσεων.

Το μοντέλο χωρίζεται σε 2 βασικά μέρη:

1. Δημιουργία λανθάνουσας μεταβλητής για τα στοιχεία x_i, x_j

2. Εύρεση της σύνδεσης μεταξύ τους

Και απεικονίζεται στο Σχήμα 2.25.

Συνοπτική περιγραφή του τρόπου ανάπτυξης του RVAE μοντέλου:

1. Το πρώτο βήμα απεικονίζεται στο Σχήμα 2.26.

Ως x_j συμβολίζεται το περιεχόμενο ενός στοιχείου j και παράγεται από την λανθάνουσα μεταβλητή z_j μέσω χρήσης ενός νευρωνικού δικτύου

Δεδομένης της λανθάνουσας μεταβλητής z, το x δημιουργείται μέσω ενός δικτύου αντίληψης πολλαπλών επιπέδων / multi-layer perception network (MLP).

Η ίδια διαδικασία γίνεται για δυο αντίστοιχα δεδομένα, έστω i, j όπως φαίνεται και τη φωτογραφία.

2. Το δεύτερο βήμα απεικονίζεται αντίστοιχα στο Σχήμα 2.27.

Με τις λανθάνουσες αναπαραστάσεις των δεδομένων, ο δυαδικός δείκτης σύνδεσης / binary link indicator μεταξύ των i και j μπορεί να σχεδιαστεί από τη συνάρτηση πιθανότητας σύνδεσης που ορίζεται από τα z_j και z_i :

$$r_{ij} \sim \psi\left(\cdot \mid z_i, z_j, \eta\right)$$

Όπου n είναι η παράμετρος που ορίζει πιθανότητας σύνδεσης των δεδομένων.

Το διάγραμμα ροής της RVAE μεθόδου απεικονίζεται στα Σχήματα 2.28 και 2.29.

Για περισσότερες λεπτομέρειες: [8]

2.7 Μέθοδος Μετρικών Συνεργασίας

Η μέθοδος μετρικών συνεργασίας / Collaborative Metric Learning (CML) αποτελεί μια σύνδεση μεταξύ της μετρικής μάθησης / metric learning και της συλλογικής επιλογής / collaborative filtering. Η μέθοδος Collaborative Metric Learning μαθαίνει έναν κοινό μετρικό χώρο για να κωδικοποιεί όχι μόνο τις προτιμήσεις των χρηστών αλλά και την ομοιότητα χρήστη - χρήστη και αντικειμένου - αντικειμένου.

Όπως προαναφέρθηκε, μεταβαίνοντας από "ρητά" σχόλια σε "άρρητα", το επίκεντρο του collaborative filtering δεν είναι πλέον η εκτίμηση ενός συγκεκριμένου πίνακα αξιολόγησης, αλλά η καταγραφή των σχετικών προτιμήσεων των χρηστών για διαφορετικά στοιχεία.

Σε αυτήν την ενότητα περιγράφεται ο CML ως ένας πιο φυσικός τρόπος για να συλληφθούν τέτοιες σχετικές σχέσεις. Η βασική ιδέα του CML έχει ως εξής: μοντελοποιεί την "ρητή" / implicit ανατροφοδότηση ως ένα σύνολο ζευγών στοιχείων - χρήστη S που θεωρείται ότι έχουν θετικές σχέσεις και εκπαιδεύει μια μέθοδο μέτρησης της σχέσης στοιχείου-χρήστη για την κωδικοποίηση αυτών των σχέσεων. Συγκεκριμένα, η λογική είναι να τραβά τα κοινά ζεύγη στο S πιο κοντά και να ωθεί τα άλλα ζεύγη σχετικά πιο μακριά. Αυτή η διαδικασία, λόγω της ανισότητας του τριγώνου, θα συσσωρεύσει επίσης

1. τους χρήστες που τους αρέσουν τα ίδια αντικείμενα μαζί, και

2. τα αντικείμενα που αρέσουν στους ίδιους χρήστες.

Τελικά, τα πλησιέστερα γειτονικά αντικείμενα για κάθε χρήστη θα γίνουν:

- τα στοιχεία που άρεσαν προηγουμένως σε αυτόν τον χρήστη και
- τα αντικείμενα που άρεσαν στο παρελθόν σε άλλους χρήστες που έχουν παρόμοια γούστο με αυτόν τον χρήστη

Με άλλα λόγια, μαθαίνοντας μια μέτρηση που υπακούει στις γνωστές "θετικές" σχέσεις, διαδίδονται αυτές τις σχέσεις όχι μόνο σε άλλα ζεύγη χρηστών-στοιχείων, αλλά και σε εκείνα τα ζεύγη χρηστών-χρηστών και αντικειμένων για τα οποία δεν παρατηρούνται άμεσα τέτοιες σχέσεις.

Κάθε χρήστης και κάθε στοιχείο αντιπροσωπεύονται με ένα διάνυσμα χρήστη $u_i \epsilon \mathbb{R}^r$ και ένα διάνυσμα αντικειμένου $v_j \epsilon \mathbb{R}^r$. Τα διανύσματα εκπαιδεύονται με τρόπο που η ευκλείδεια απόσταση τους : d(i,j) = || u_i - v_j || θα ακολουθήσει τις σχετικές προτιμήσεις του χρήστη i για διαφορετικά αντικείμενα, δηλαδή ένα αντικείμενο που άρεσε στο χρήστη θα είναι πιο κοντά σε αυτόν τον χρήστη από άλλα αντικείμενα που δεν του άρεσαν. Για να το επιτευχθεί αυτό χρησιμοποιείται η παρακάτω συνάρτηση :

$$\mathcal{L}_m(d) = \sum_{(i,j)\in\mathcal{S}} \sum_{(i,k)\notin\mathcal{S}} w_{ij} \left[m + d(i,j)^2 - d(i,k)^2\right]_+$$

Όπου j ένα στοιχείο που άρεσε στον χρήστη i, και k ένα στοιχείο που δεν του άρεσε, [z]+ = max(z,0) = standard hinge loss, w_{ij} = ranking loss weight. Έτσι τα αντικείμενα που αρέσουν στο χρήστη έχουν κλίση προς τα μέσα για να δημιουργήσουν μικρότερη ακτίνα.

Για αντικείμενα "απατεώνων", που είναι τα αντικείμενα που δεν άρεσαν στο χρήστη αλλά που εισβάλλουν στην περίμετρο, οι κλίσεις τους κινούνται προς τα έξω έως ότου ωθούνται έξω από την περίμετρο ασφαλούς περιθωρίου / margin, όπως φαίνεται και στο Σχήμα 2.30.

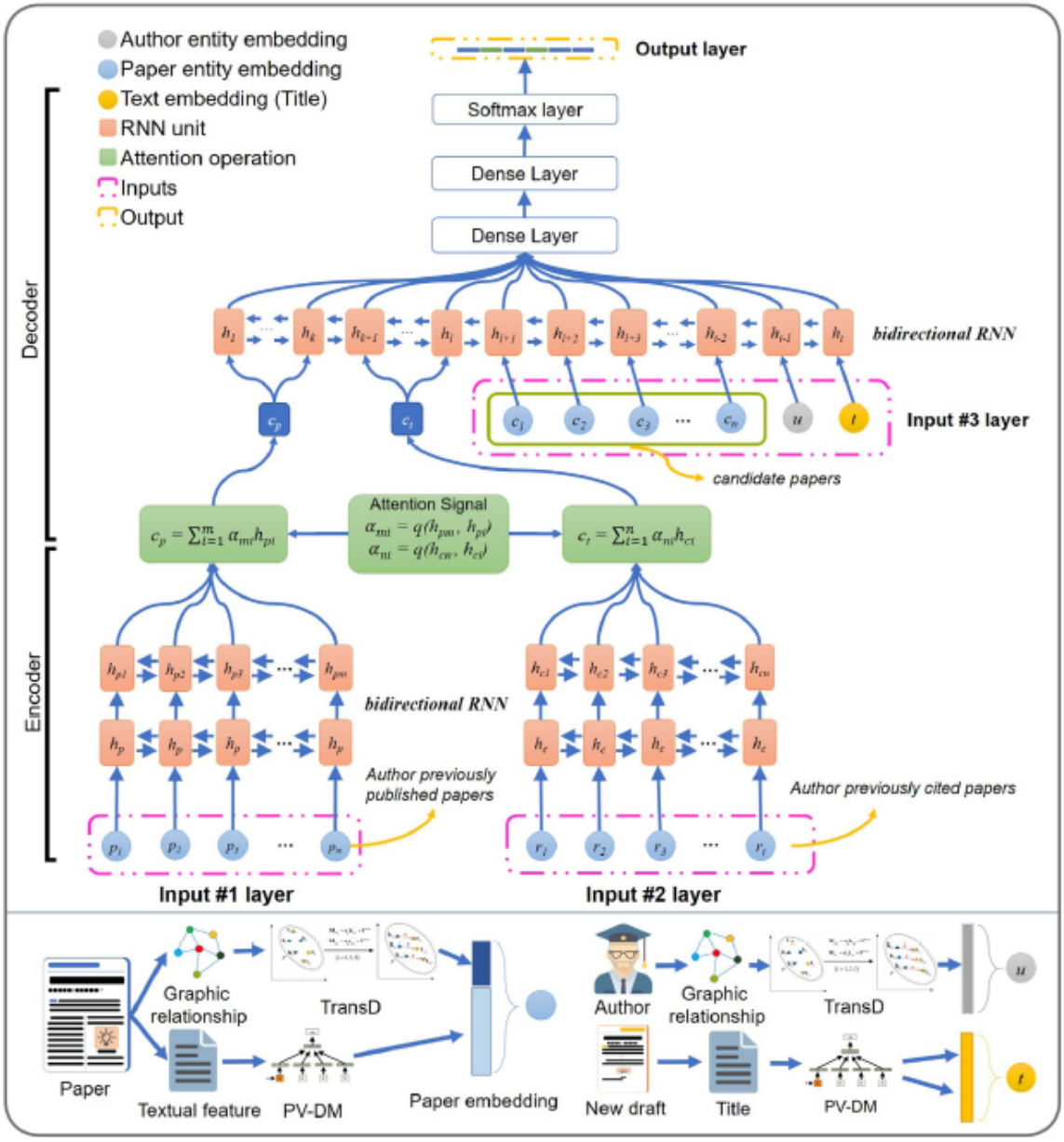
Αυτή η λειτουργία απώλειας / loss function είναι παρόμοια με εκείνη του LMNN που προαναφέρθηκε αλλά με τρεις σημαντικές διαφορές:

1. Οι γείτονες-στόχοι κάθε χρήστη είναι όλα τα στοιχεία που του άρεσαν και δεν υπάρχει γείτονας-στόχος για τα στοιχεία.
2. Δεν υπάρχει ο όρος Lpull, επειδή ένα στοιχείο μπορεί να αρέσει σε πολλούς χρήστες και δεν είναι εφικτό να το προσεγγίσουμε περισσότερο σε όλους. Ωστόσο, η απώλεια ώθησης φέρνει τα θετικά στοιχεία πιο κοντά στο χρήστη όταν υπάρχουν απατεώνες.

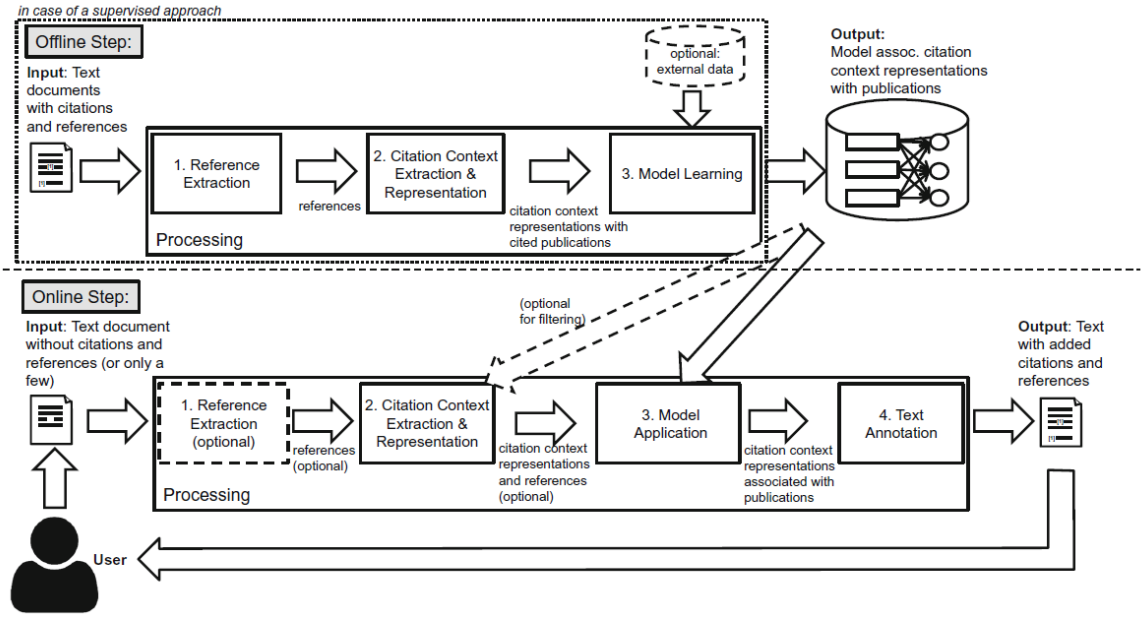
3. Χρησιμοποιείται μια σταθμισμένη απώλεια κατάταξης / weighted ranking loss για να βελτιωθούν οι Top-K συστάσεις που προκύπτουν.

Το διάγραμμα ροής της CML μεθόδου απεικονίζεται στο Σχήμα 2.31.

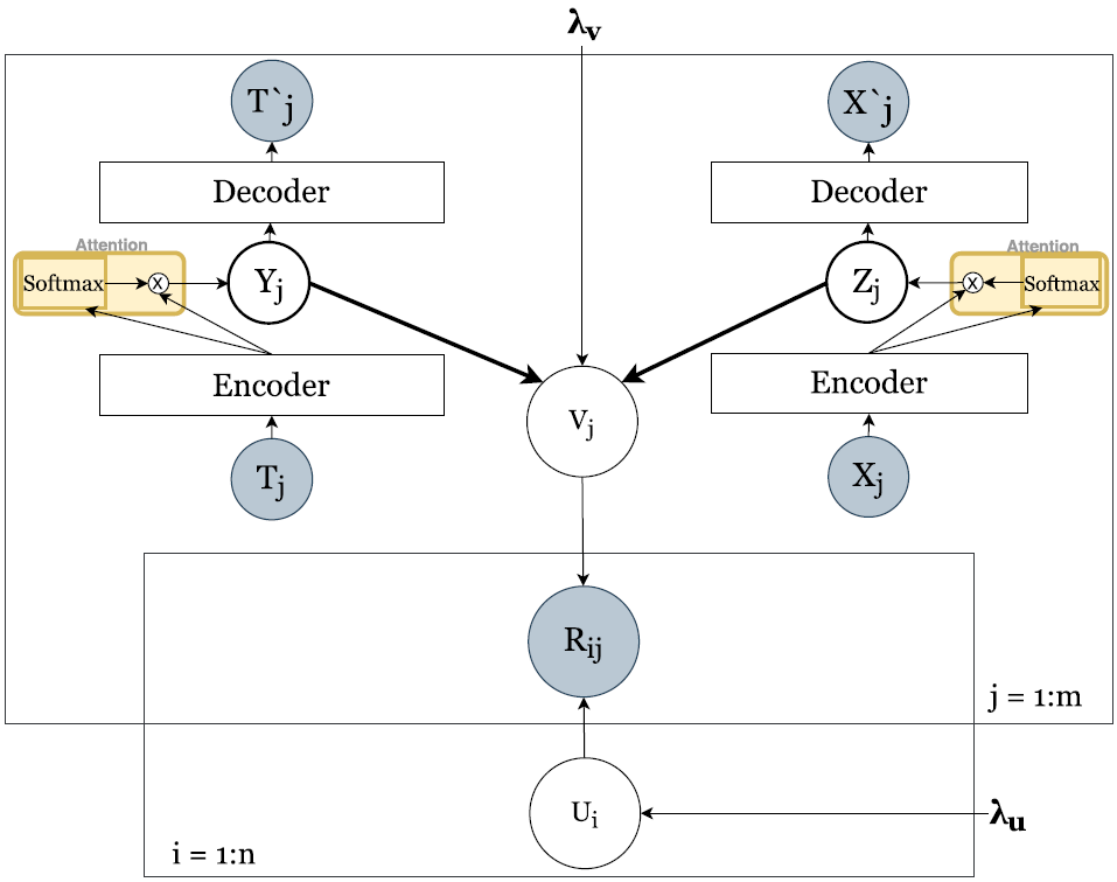
Για περισσότερες λεπτομέρειες: [9]



Σχήμα 2.14: heterogeneous knowledge embedding based attentive recurrent neural networks [5]



Σχήμα 2.15: Αρχιτεκτονική ενός πρωτότυπου συστήματος συστάσεων παραπομπών [6]

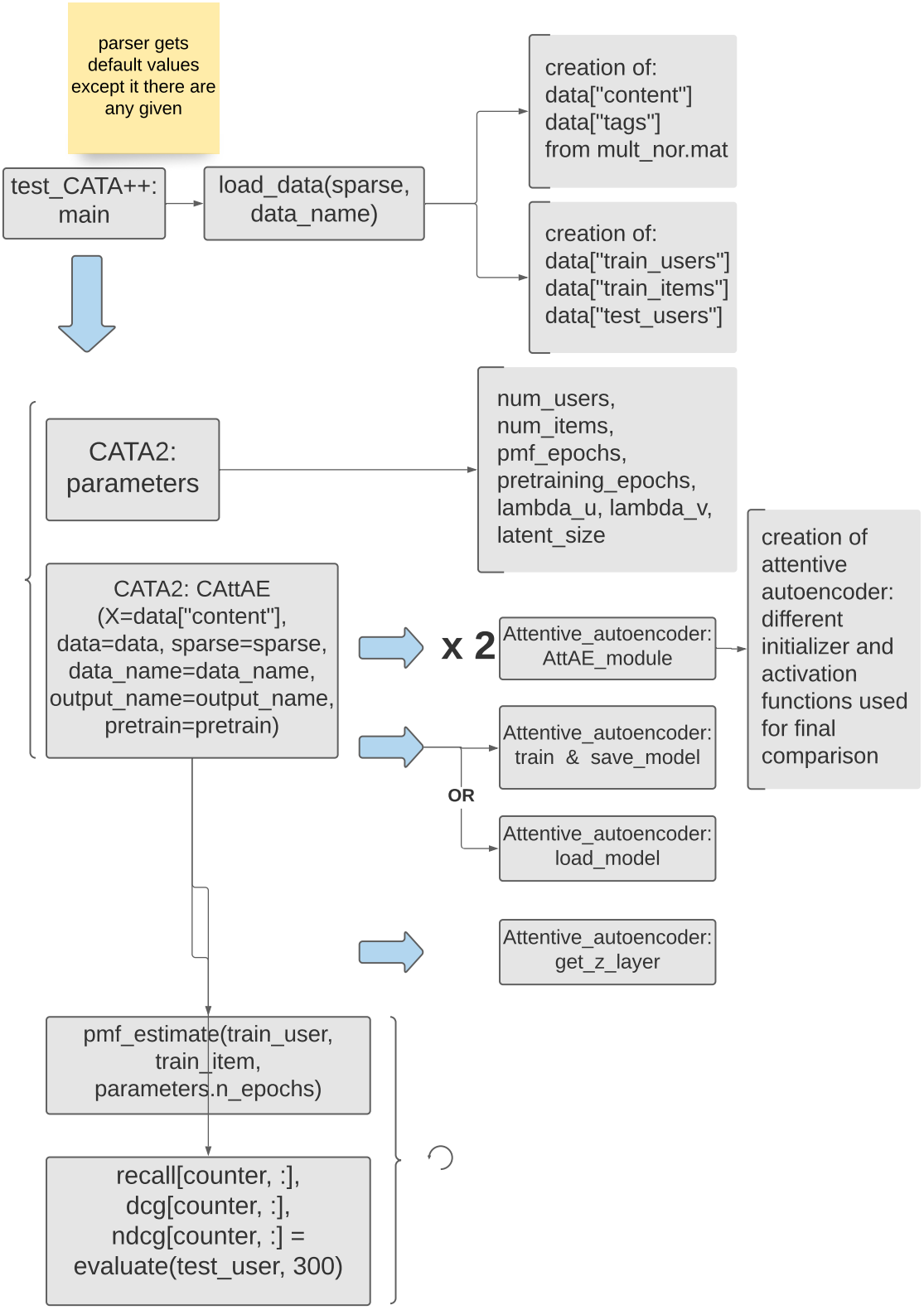


Σχήμα 2.16: Μοντέλο CATA++ [7]

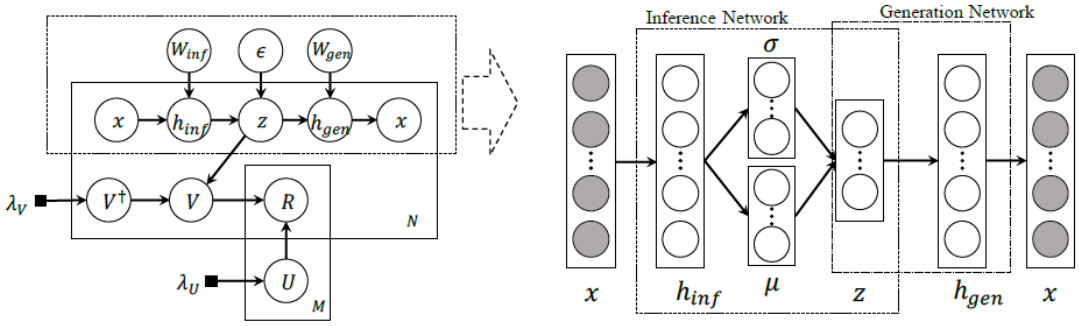
Algorithm 1 CATA++ Algorithm

pretrain 1st autoencoder using X;
pretrain 2nd autoencoder using T;
Z $\leftarrow \theta(X)$;
Y $\leftarrow \gamma(T)$;
Initialize U and V randomly;
while <model NOT converge> **do**
  for <each $user_i$> **do**
    $u_i \leftarrow$ optimize via Equation 9;
  end for
  for <each $item_j$> **do**
    $v_i \leftarrow$ optimize via Equation 10;
  end for
end while
for <each $user_i$> **do**
  $scores_i \leftarrow u_i V^T$;
  sort($scores_i$) in descending order;
end for
Recommend top-K articles;

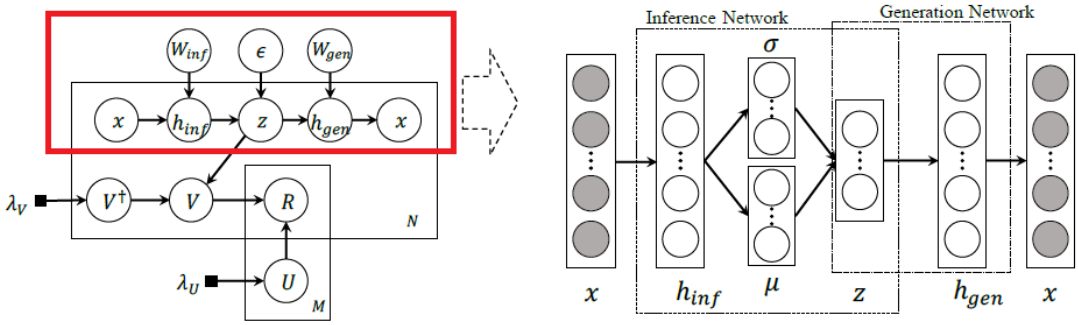
Σχήμα 2.17: Αλγόριθμος ανάπτυξης CATA++ [7]



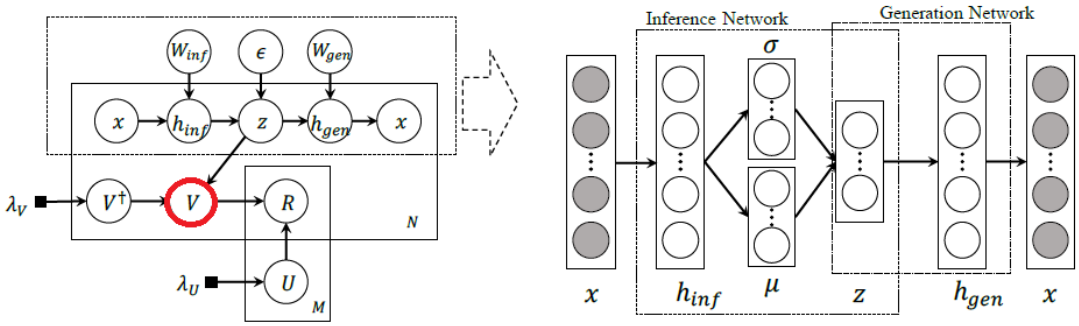
Σχήμα 2.18: CATA++ Διάγραμμα Ροής



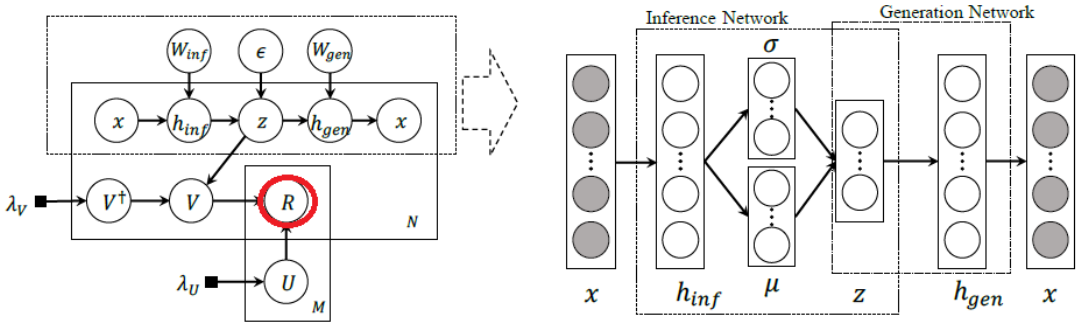
Σχήμα 2.19: Μοντέλο CVAE [2]



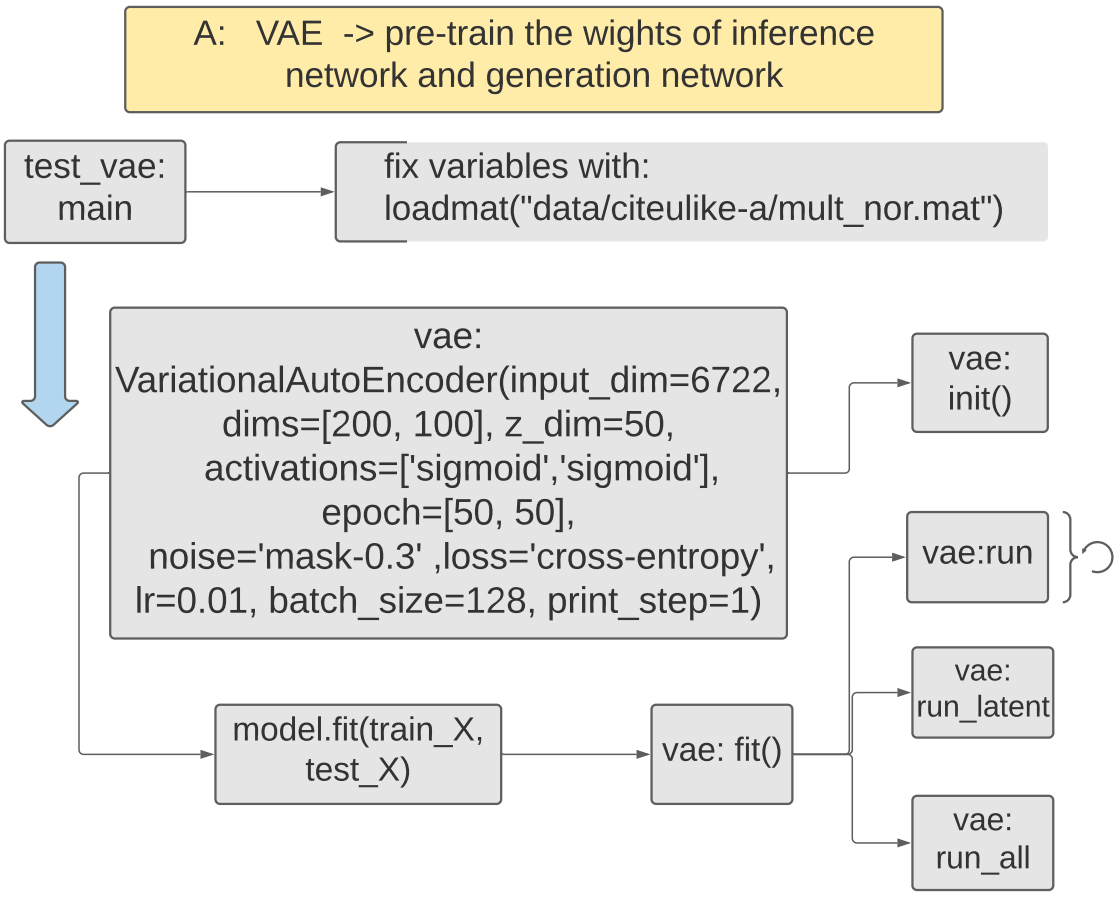
Σχήμα 2.20: Μοντέλο CVAE i



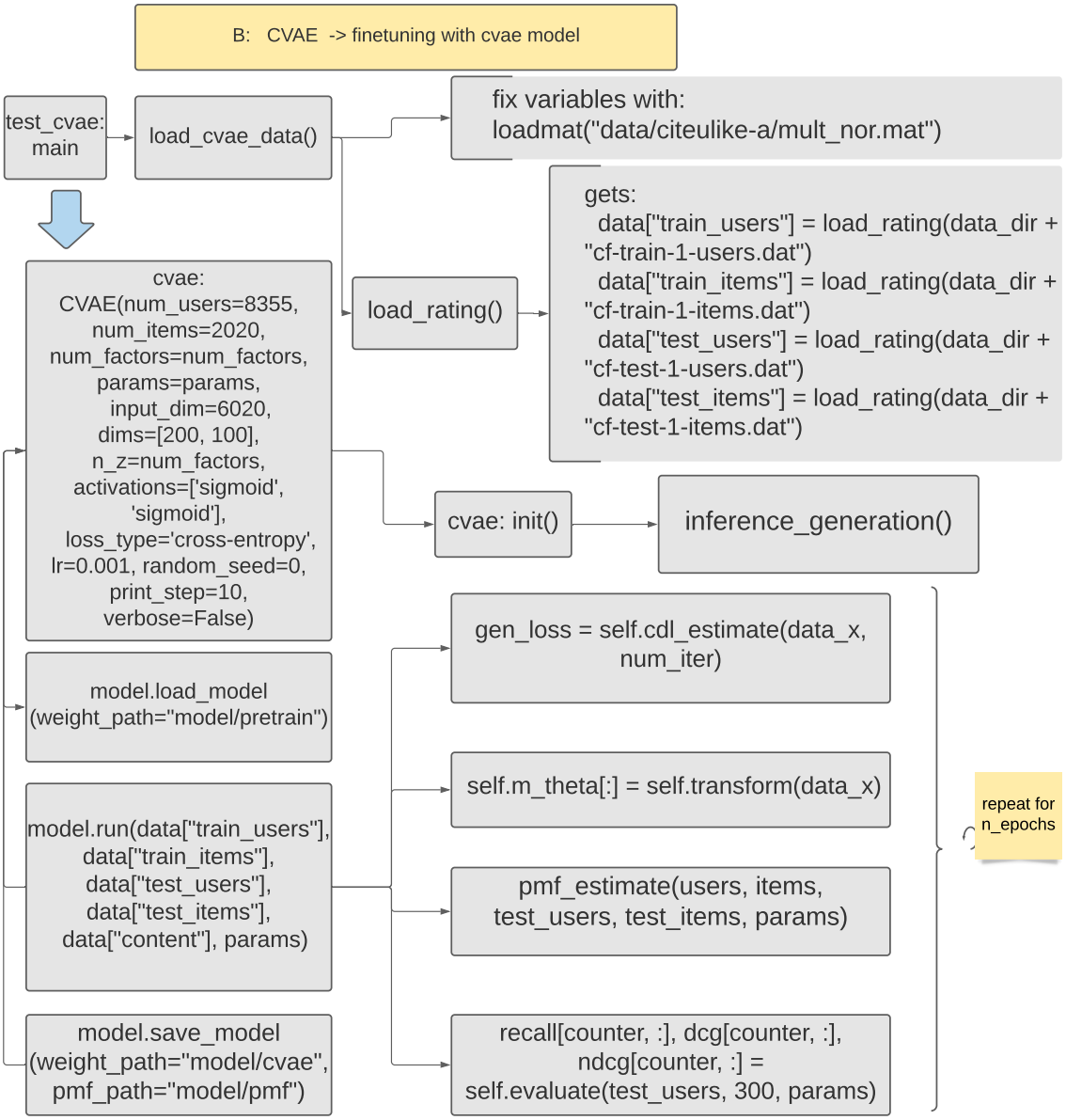
Σχήμα 2.21: Μοντέλο CVAE ii



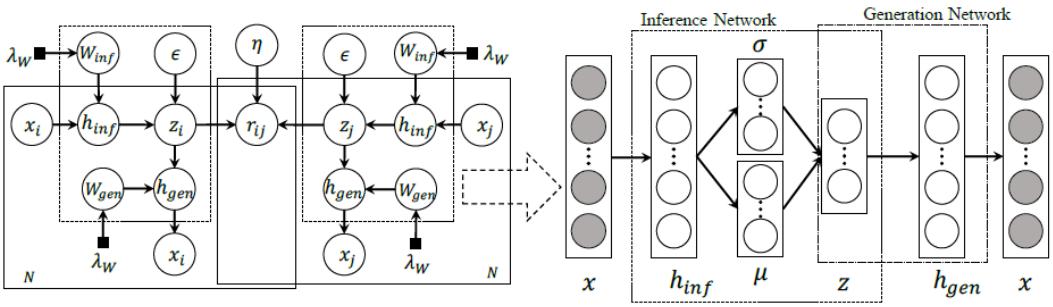
Σχήμα 2.22: Μοντέλο CVAE iii



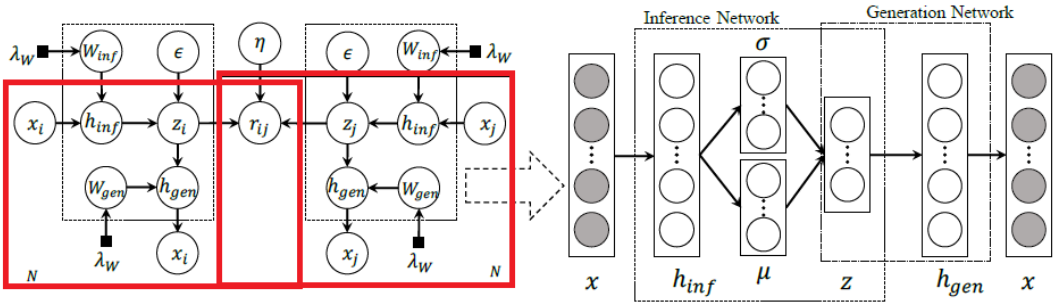
Σχήμα 2.23: CVAE Διάγραμμα Ροής i



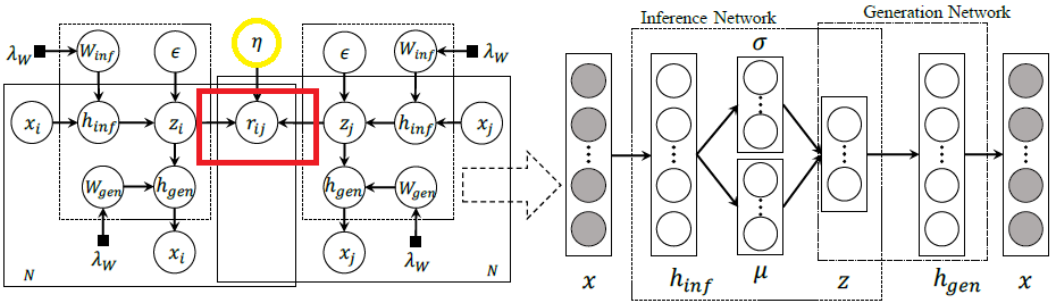
Σχήμα 2.24: CVAE Διάγραμμα Ροής ii



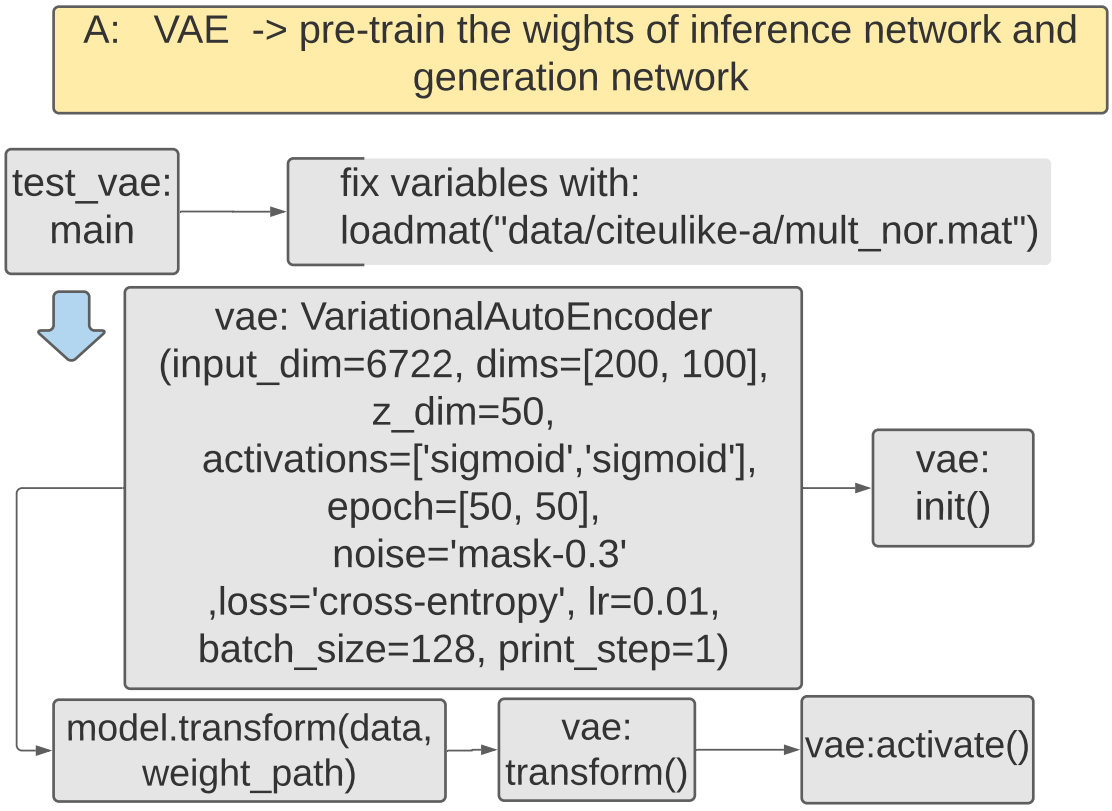
Σχήμα 2.25: Μοντέλο RVAE [8]



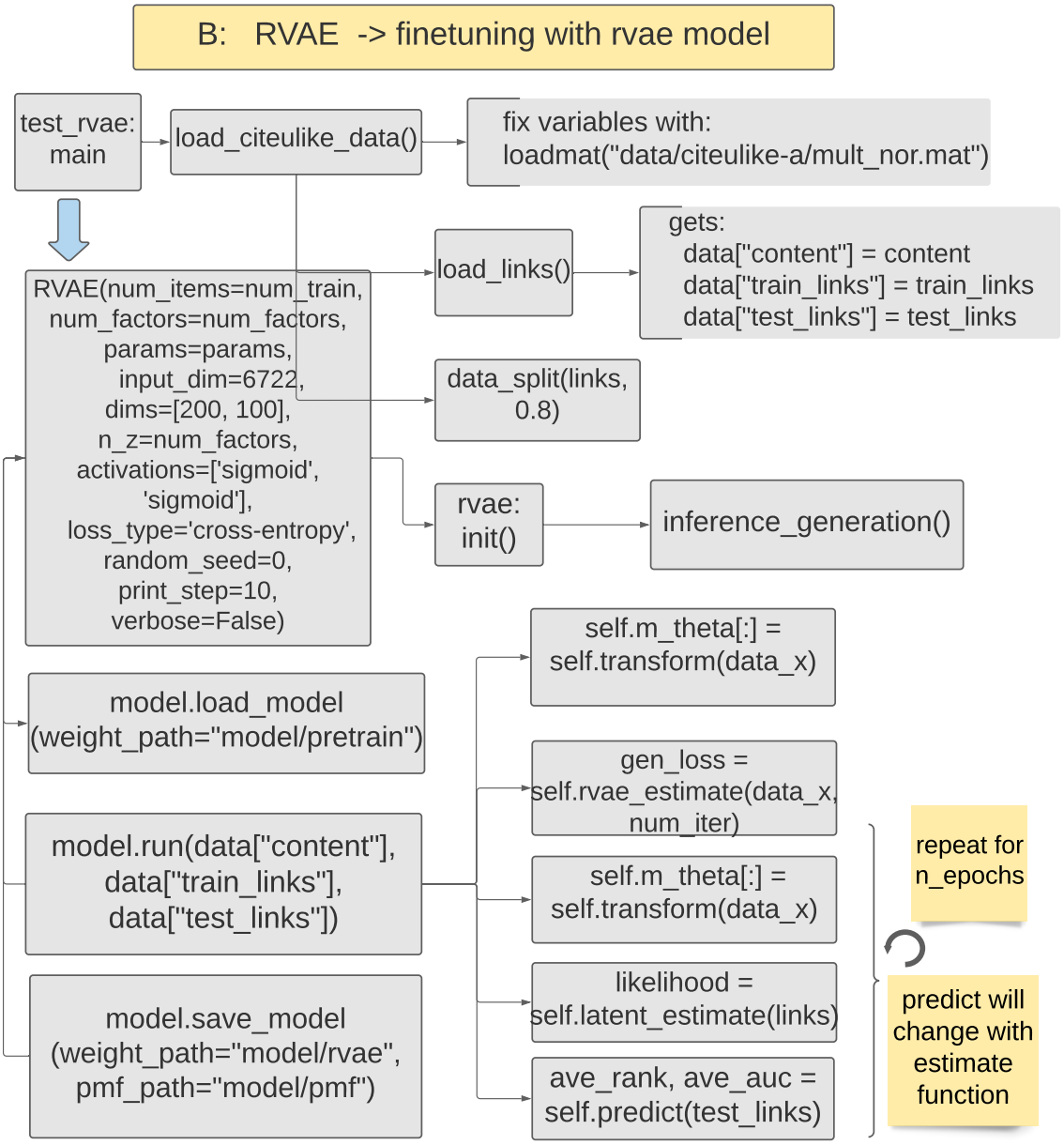
Σχήμα 2.26: Μοντέλο RVAE i



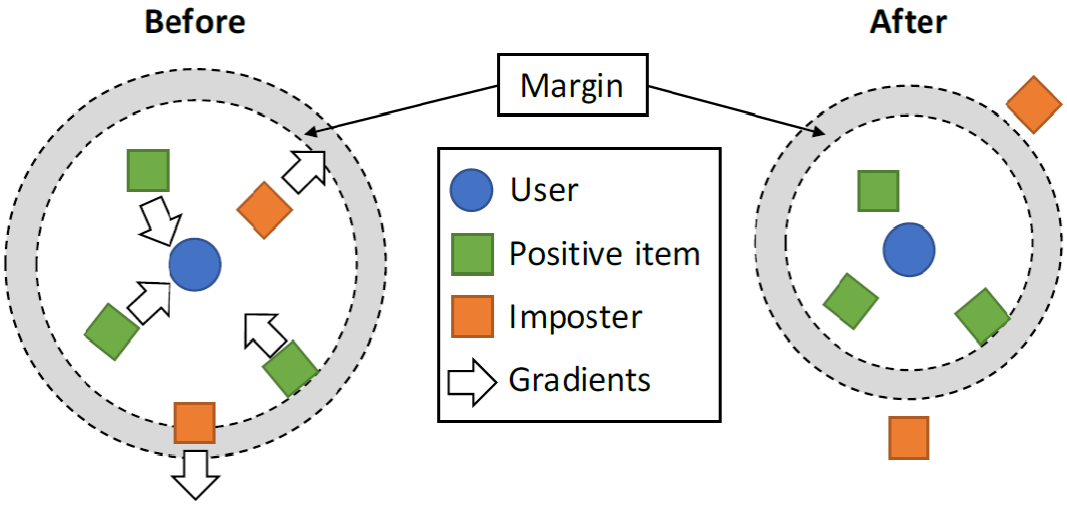
Σχήμα 2.27: Μοντέλο RVAE ii



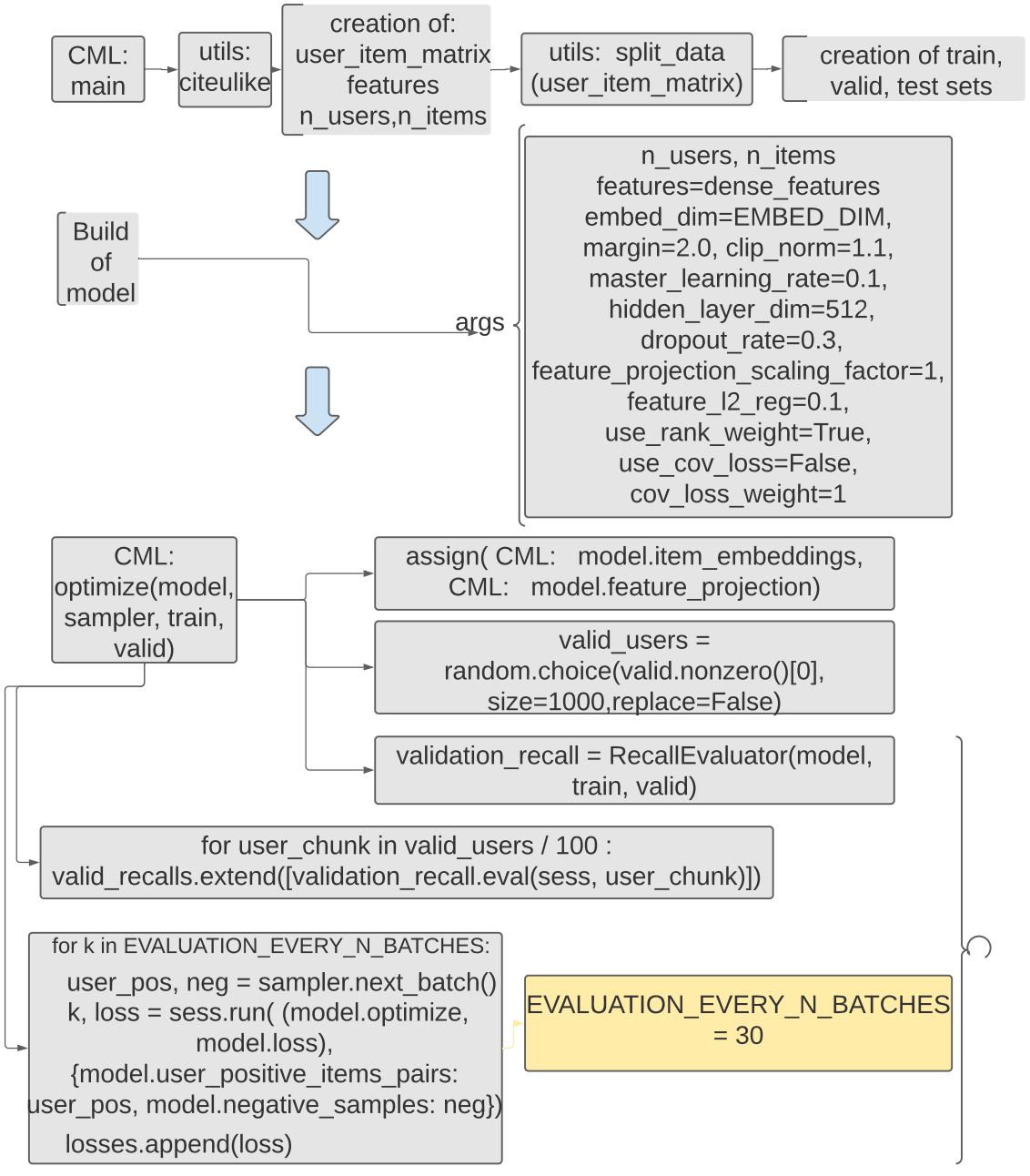
Σχήμα 2.28: RVAE Διάγραμμα Ροής i



Σχήμα 2.29: RVAE Διάγραμμα Ροής ii



Σχήμα 2.30: Μοντέλο CML [9]



Σχήμα 2.31: CML Διάγραμμα Ροής

Κεφάλαιο 3

Προετοιμασία και επεξεργασία βάσης δεδομένων

3.1 Δεδομένα

Πρώτο βήμα για την ανάπτυξη και σύγκριση των συστημάτων συστάσεων αποτελεί η εύρεση, ανάλυση και επεξεργασία των δεδομένων που θα χρησιμοποιηθούν ως αρχεία εισόδου / dataset. Για την ανάπτυξη αυτής της διπλωματικής εργασίας έχω χρησιμοποιήσει το αρχεία από την Aminer βιβλιοθήκη ([10]).

Το σύνολο δεδομένων έχει σχεδιαστεί μόνο για ερευνητικούς σκοπούς. Τα δεδομένα εξάγονται από τις DBLP, ACM, MAG (Microsoft Academic Graph) και άλλες πηγές. Για την αποθήκευσή τους χρησιμοποιήθηκε η MongoDB, και η επεξεργασία τους έγινε με δυο διαφορετικές μεθόδους οι οποίες αναλύονται παρακάτω.

3.2 Σχήμα Βάσης Δεδομένων

Τα δεδομένα που χρησιμοποιούνται κατά την εφαρμογή των συστημάτων συστάσεων μπορούν να απεικονιστούν ως ένας πίνακας με γραμμές και στήλες, όπου η κάθε γραμμή του πίνακα αναπαριστά ένα σύγγραμμα, και κάθε στήλη τα χαρακτηριστικά του. Τα δεδομένα είναι αρχικά σε json μορφή, και στη συνέχεια εισάγονται στη MongoDB ώστε να διευκολύνεται η ανάλυση και επεξεργασία τους. Για το κάθε έγγραφο έχουν αξιοποιηθεί κυρίως τα ακόλουθα χαρακτηριστικά:

- ένα id που το καθιστά μοναδικό

- το title, δηλαδή τον τίτλο του συγγράμματος
- έναν πίνακα από authors, όπου ο κάθε author έχει 3 πεδία: name, origin και id
- το year που δείχνει το έτος συγγραφής του εγγράφου
- τα keywords, δηλαδή τις κύριες λέξεις που καθορίζουν το θέμα του εγγράφου
- τα references, δηλαδή σε ποια άλλα άρθρα έχει γίνει αναφορά κατά τη συγγραφή του συγκεκριμένου εγγράφου
- το fos = field of study / πεδίο σπουδών (αποτελεί πίνακα όπου για κάθε στοιχείο υπάρχουν τα αντίστοιχα πεδία fos.name και fos.w δηλαδή το όνομα και το βάρος w - weight)
- το indexed_abstract, το οποίο αποτελεί ένα dictionary όπου περιέχονται όλες οι λέξεις από το abstract string (η συμβολοσειρά / string αποτελεί τη περιληπτική περιγραφή του θέματος του κάθε εγγράφου) και τον αριθμό των φορών που εμφανίζεται η κάθε μια.

Εκτός αυτών των βασικών χαρακτηριστικών υπάρχουν επίσης τα κάτωθι: n_citation, page_start, page_end, doc_type, publisher, volume, issue, doi και venue (αποτελεί πίνακα όπου για κάθε στοιχείο υπάρχουν τα αντίστοιχα πεδία venue.raw, venue.id και venue.type).

Ένα παράδειγμα ενός συγγράμματος σε json μορφή αποτελεί το παρακάτω:


```
{
    ''_id'': {
        ''$oid'': ''5f8f0d16ec43c662c80bf1ff''
    },
    ''id'': 1091,
    ''authors'': [{
        ''name'': ''Makoto Satoh'',
        ''org'': ''Shinshu University'',
        ''id'': {
            ''$numberLong'': ''2312688602''
        }
```

```
}, {
    ''name'': ''Ryo Muramatsu'',
    ''org'': ''Shinshu University'',
    ''id'': {
        ''$numberLong'': ''2482909946''
    }
}, {
    ''name'': ''Mizue Kayama'',
    ''org'': ''Shinshu University'',
    ''id'': 2128134587
}, {
    ''name'': ''Kazunori Itoh'',
    ''org'': ''Shinshu University'',
    ''id'': 2101782692
}, {
    ''name'': ''Masami Hashimoto'',
    ''org'': ''Shinshu University'',
    ''id'': 2114054191
}, {
    ''name'': ''Makoto Otani'',
    ''org'': ''Shinshu University'',
    ''id'': 1989208940
}, {
    ''name'': ''Michio Shimizu'',
    ''org'': ''Nagano Prefectural College'',
    ''id'': 2134989941
}, {
    ''name'': ''Masahiko Sugimoto'',
    ''org'': ''Takushoku University, Hokkaido Junior
      College'',
    ''id'': {
        ''$numberLong'': ''2307479915''
```

```
        }
    }],
    ''title'': ''Preliminary Design of a Network Protocol
      Learning Tool Based on the Comprehension of High
      School Students: Design by an Empirical Study Using a
      Simple Mind Map'',
    ''year'': 2013,
    ''n_citation'': 1,
    ''page_start'': ''89'',
    ''page_end'': ''93'',
    ''doc_type'': ''Conference'',
    ''publisher'': ''Springer, Berlin, Heidelberg'',
    ''volume'': '''',
    ''issue'': '''',
    ''doi'': ''10.1007/978-3-642-39476-8_19'',
    ''references'': [2005687710, 2018037215],
    ''indexed_abstract'': {
        ''IndexLength'': 58,
        ''InvertedIndex'': {
            ''tool.'': [42],
            ''study'': [4],
            ''aim'': [37],
            ''purpose'': [1],
            ''scientific'': [17],
            ''for'': [11],
            ''aspects'': [18],
            ''students'': [14, 46],
            ''focus'': [27],
            ''hands-on'': [47],
            ''learning'': [9, 41],
            ''experience'': [48],
            ''our'': [40],
```

```
            ‘‘we’’: [26],
            ‘‘network’’: [33, 56],
            ‘‘The’’: [0],
            ‘‘More’’: [24],
            ‘‘high’’: [12],
            ‘‘protocols.’’: [57],
            ‘‘school’’: [13],
            ‘‘and’’: [21],
            ‘‘of’’: [2, 19, 32, 55],
            ‘‘communication’’: [22],
            ‘‘protocols’’: [34],
            ‘‘gives’’: [45],
            ‘‘on’’: [28],
            ‘‘a’’: [8],
            ‘‘studying’’: [15],
            ‘‘specifically,’’: [25],
            ‘‘this’’: [3],
            ‘‘understand’’: [51],
            ‘‘is’’: [5],
            ‘‘develop’’: [7, 39],
            ‘‘Our’’: [43],
            ‘‘tool’’: [10, 44],
            ‘‘the’’: [16, 29, 36, 52],
            ‘‘help’’: [50],
            ‘‘as’’: [35],
            ‘‘principles’’: [31, 54],
            ‘‘information’’: [20],
            ‘‘networks.’’: [23],
            ‘‘to’’: [6, 38, 49],
            ‘‘basic’’: [30, 53]
        }
    },
```

```
    ''fos'': [{
        ''name'': ''Telecommunications network'',
        ''w'': 0.45139
    }, {
        ''name'': ''Computer science'',
        ''w'': 0.45245
    }, {
        ''name'': ''Mind map'',
        ''w'': 0.5347
    }, {
        ''name'': ''–Humancomputer interaction'',
        ''w'': 0.47011
    }, {
        ''name'': ''Multimedia'',
        ''w'': 0.46629
    }, {
        ''name'': ''Empirical research'',
        ''w'': 0.49737
    }, {
        ''name'': ''Comprehension'',
        ''w'': 0.47042
    }, {
        ''name'': ''Communications protocol'',
        ''w'': 0.51907
    }],
    ''venue'': {
        ''raw'': ''International Conference on
          Human–Computer Interaction'',
        ''id'': 1127419992,
        ''type'': ''C''
    }
}
```

Field Name	Field Type	Description	Example
id	string	paper ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	Department of Computer Science, University of Illinois at Urbana-Champaign
author.id	string	author ID	53f42f36dabfaedce54dcd0c
venue.id	string	paper venue ID	53e17f5b20f7dfbc07e8ac6e
venue.raw	string	paper venue name	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos.name	string	paper fields of study	Web mining
fos.w	float	fields of study weight	0.659690857
references	list of strings	paper references	["4909282", "16018031", "16159250", "19838944", ...]
n_citation	int	citation number	40829
page_start	string	page start	11
page_end	string	page end	18
doc_type	string	paper type: Journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55860-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	//static.aminer.org/upload/pdf/1254/370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.lsi.uned.es/revista/index.php/ia/ article/view/479"]
abstract	string	abstract	Our ability to generate...
indexed_abstract	dict	indexed abstract	{"IndexLength": 164, "InvertedIndex": {"Our": [0], "ability": [1], "to": [2, 7, ...]}}

Σχήμα 3.1: Σχήμα Βάσης Δεδομένων [10]

Το σχήμα της βάσης απεικονίζεται στο Σχήμα 3.1. Το συγκεκριμένο dataset έχει μέγεθος 12 GB οπότε για τη σωστή και αποτελεσματική αποθήκευση και εκμετάλλευσή του χρησιμοποιήθηκε η MongoDB βάση δεδομένων.

3.3 Εισαγωγή στη MongoDB

Το όνομα Mongo είναι ένα τμήμα της λέξης "humongous". Οι ιδρυτές της, Dwight Merriman, Eliot Horowitz και Kevin Ryan, οι οποίοι συν-ίδρυσαν τη MongoDB στη Νέα Υόρκη το 2007, ήθελαν να δημιουργήσουν μια βάση δεδομένων που θα ήθελαν οι προγραμματιστές, μια βάση δεδομένων που θα διέσχιζε τα εμπόδια στα Συστήματα Διαχείρισης Βάσεων Δεδομένων (RDBMS) που χρησιμοποιούν τη γλώσσα ερωτημάτων SQL.

Η MongoDB είναι μια βάση δεδομένων εγγράφων ανοιχτού κώδικα βασισμένη σε "scale-out" αρχιτεκτονική.

Έχει παγκόσμια παρακολούθηση στην κοινότητα προγραμματιστών όλων των ειδών που δημιουργούν επεκτάσιμες εφαρμογές χρησιμοποιώντας ευέλικτες μεθοδολογίες.

Εταιρείες και ομάδες ανάπτυξης όλων των μεγεθών χρησιμοποιούν το MongoDB επειδή:

- Το μοντέλο της βάσης δεδομένων αποτελεί έναν αποτελεσματικό τρόπο αποθήκευσης και ανάκτησης δεδομένων που επιτρέπει στους προγραμματιστές να κινούνται γρήγορα.
- Η "scale-out" αρχιτεκτονική μπορεί να υποστηρίξει τεράστιους όγκους δεδομένων και επισκεψιμότητας.
- Επιτρέπει σε προγραμματιστές να αρχίσουν να γράφουν κώδικα αμέσως.
- Μπορεί να χρησιμοποιηθεί παντού από οποιονδήποτε:
  - Δωρεάν μέσω της έκδοσης κοινότητας ανοιχτού κώδικα
  - Στα μεγαλύτερα κέντρα δεδομένων μέσω της εταιρικής έκδοσης
  - Σε οποιοδήποτε από τα μεγάλα "clouds" μέσω του MongoDB Atlas
- Η MongoDB αποτελεί μια μεγάλη πλατφόρμα, που σημαίνει:
  - Έχει μια παγκόσμια κοινότητα προγραμματιστών και συμβούλων, επομένως είναι εύκολο να ληφθεί βοήθεια

- Λειτουργεί σε όλους τους τύπους υπολογιστικών πλατφορμών, τόσο εντός εγκατάστασης όσο και στο cloud (τόσο ιδιωτικό όσο και δημόσιο cloud όπως AWS, Azure και Google Cloud)
- Μπορεί να χρησιμοποιηθεί από όλες τις μεγάλες γλώσσες
- Υπάρχει πρόσβαση στο MongoDB από όλα τα μεγάλα συστήματα διαχείρισης δεδομένων και ETL
- Έχει υποστήριξη εταιρικού επιπέδου

Αντί να αποθηκεύονται δεδομένα σε πίνακες γραμμών ή στηλών όπως η SQL βάση δεδομένων, κάθε σειρά σε μια βάση δεδομένων MongoDB είναι ένα έγγραφο σε JSON μορφή.

Οι βάσεις δεδομένων εγγράφων / Document Databases είναι εξαιρετικά ευέλικτες, επιτρέποντας παραλλαγές στη δομή των εγγράφων και επιτρέποντας την αποθήκευση εγγράφων που είναι εν μέρει πλήρεις. Ένα έγγραφο μπορεί να έχει άλλα ενσωματωμένα σε αυτό.

Τα πεδία σε ένα έγγραφο παίζουν το ρόλο των στηλών σε μια βάση δεδομένων SQL, και, όπως οι στήλες, μπορούν να ευρετηριαστούν για να αυξήσουν την απόδοση αναζήτησης.

Τα “documents” στη MongoDB είναι αρχεία JSON και BSON.

Το JSON είναι ισχυρό για πολλούς λόγους:

- Είναι μια φυσική μορφή αποθήκευσης δεδομένων
- Είναι αναγνώσιμο από τον άνθρωπο
- Δομημένες και μη δομημένες πληροφορίες μπορούν να αποθηκευτούν στο ίδιο έγγραφο
- Μπορεί να αποθηκεύσει σύνθετα δεδομένα
- Έχει ένα ευέλικτο και δυναμικό σχήμα, οπότε η προσθήκη πεδίων ή η διαγραφή ενός πεδίου δεν αποτελεί πρόβλημα

Οι προγραμματιστές προσαρμόζουν και αναδιαμορφώνουν τη βάση δεδομένων καθώς η εφαρμογή εξελίσσεται χωρίς τη βοήθεια διαχειριστή της βάσης δεδομένων. Όταν απαιτείται, η MongoDB μπορεί να συντονίσει και να ελέγξει τις αλλαγές στη δομή των εγγράφων χρησιμοποιώντας την επικύρωση σχήματος / “ schema validation ”.

Η MongoDB δημιούργησε τη μορφή Binary JSON (BSON) για να αυξήσει την αποδοτικότητα και να υποστηρίξει περισσότερους τύπους δεδομένων. Τα δεδομένα που αποθηκεύονται σε BSON μορφή μπορούν να αναζητηθούν και να ευρετηριαστούν, αυξάνοντας

σημαντικά την απόδοση. Η MongoDB υποστηρίζει μια μεγάλη ποικιλία μεθόδων ευρετηρίασης όπως κείμενο, δεκαδική, γεωχωρική και μερική.

Από την ίδρυσή της, η MongoDB χτίστηκε με "scale-out" αρχιτεκτονική, που επιτρέπει σε πολλές μικρές μηχανές να συνεργάζονται για τη δημιουργία συστημάτων που είναι γρήγορα και διαχειρίζονται τεράστιες ποσότητες δεδομένων.

Στους προγραμματιστές αρέσει επίσης το γεγονός ότι η MongoDB έχει διασφαλίσει ότι η βάση δεδομένων μπορεί να χρησιμοποιηθεί από μια μεγάλη ποικιλία γλωσσών προγραμματισμού, όπως: C, C# και .NET, C ++, Erlang, Haskell, Java, JavaScript, Perl, PHP, Python, Ruby και Scala (μέσω Casbah). Κατά την ανάπτυξη της συγκεκριμένης εργασίας χρησιμοποιήθηκε η σύνδεση της MongoDB με την python γλώσσα προγραμματισμού μέσω της PyMongo βιβλιοθήκης.

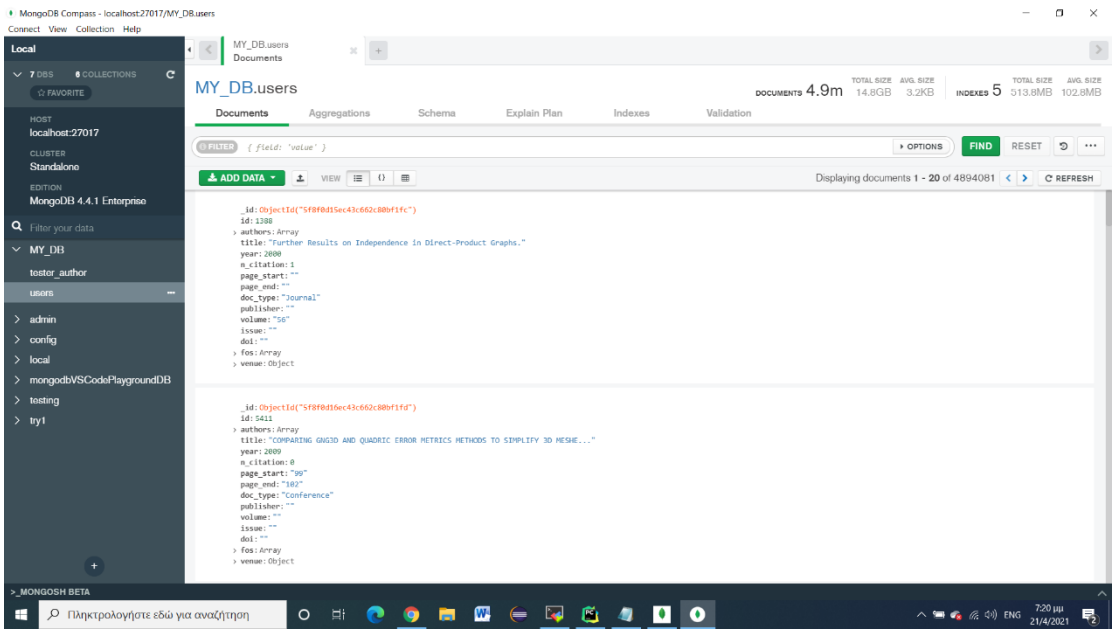
Η αρχιτεκτονική επέκτασης της MongoDB διανέμει εργασίες σε πολλούς μικρότερους (και φτηνότερους) υπολογιστές.

Οι τεχνολογικές καινοτομίες της MongoDB υποστηρίζουν τεράστιους αριθμούς ανάγνωσης και γραφής. Στο επίκεντρο αυτών των καινοτομιών βρίσκεται η προσέγγιση της MongoDB για "sharding", η οποία επιτρέπει την αποθήκευση ομάδων πληροφοριών καθώς οι πληροφορίες διαδίδονται σε όλο το σύμπλεγμα υπολογιστών. Αντιθέτως, οι περισσότερες βάσεις δεδομένων SQL χρησιμοποιούν μια αρχιτεκτονική αναβάθμισης που είναι περιορισμένη επειδή βασίζεται στη δημιουργία ταχύτερων και ισχυρότερων υπολογιστών ([11]).

Στα πλαίσια ανάπτυξης της διπλωματικής εργασίας χρησιμοποιήθηκε και η MongoDB Compass, η οποία αποτελεί το γραφικό περιβάλλον εργασίας του χρήστη / GUI για τη MongoDB. Επιτρέπει την οπτικοποίηση των δεδομένων, την εκτέλεση ερωτημάτων σε δευτερόλεπτα, την αλληλεπίδραση με τα δεδομένα με πλήρη λειτουργικότητα CRUD και την προβολή και βελτιστοποίηση της απόδοσης των ερωτημάτων. Διατίθεται σε Linux, Mac ή Windows. Η MongoDB Compass δίνει τη δυνατότητα για λήψη πιο έξυπνων αποφάσεων σχετικά με την ευρετηρίαση, την επικύρωση εγγράφων και πολλά άλλα. Στο σχήμα 3.2 διατίθεται ένα στιγμιότυπο από το περιβάλλον της MongoDB Compass.

Για να εισαχθούν τα δεδομένα στη MongoDB ακολουθήθηκαν τα παρακάτω βήματα: Αρχικά, έγινε λήψη της MongoDB από το επίσημο site της ([12]). Εκεί δίνεται η επιλογή στο χρήστη αν θέλει να κατεβάσει τη MongoDB στον υπολογιστή του ή αν θέλει να την έχει στο cloud μέσω του MongoDB Atlas, και επιλέχθηκε η εγκατάσταση στον υπολογιστή.

Στη συνέχεια, έγινε λήψη του json αρχείου με τα δεδομένα από την Aminer βιβλιοθήκη



Σχήμα 3.2: Περιβάλλον MongoDB Compass

που πρέπει να εισαχθεί στη βάση ([10]).

Έπειτα, έγινε λήψη της MongoDB Compass και πάλι από την αντίστοιχη ιστοσελίδα ([13]).

Τέλος, για την αυτόματη εισαγωγή των json αρχείων στη βάση δεδομένων υπάρχει η εντολή mongoimport. Αυτή η εντολή δεν είναι προεγκατεστημένη στο περιβάλλον της MongoDB και γι'αυτό το λόγο απαιτείται η λήψη και των `mongodb-database-tools-windows-x86_64-100.2.0`, τα οποία πρέπει να γίνουν export στο φάκελο `MongoDB\Server\4.4\bin`. Έτσι ανοίγοντας ένα command window στο σημείο αυτό (`cd C:\ProgramFiles\MongoDB\Server\4.4\bin\mongodb-database-tools-windows-x86_64-100.2.0\bin`) και εκτελώντας την εντολή `mongoimport--dbMY_DB--collectionusers--drop--jsonArray--batchSize1--file./C:\Users\hp\Desktop\ΣΧΟΛΗ\ειδικό\dblp.v12.json` δημιουργείται αυτόματα μια νέα βάση στη MongoDB με όνομα MY_DB, και ένα νέο collection με όνομα users όπου εκεί περνιούνται αυτόματα όλα τα δεδομένα από το dblp.v12.json αρχείο.

3.4 Επεξεργασία των δεδομένων

Λαμβάνοντας υπόψιν τους προαναφερθέντες λόγους, αποφασίστηκε η χρήση της MongoDB για την αποθήκευση των δεδομένων που χρησιμοποιούνται. Στη συνέχεια με τη χρήση

της python, και συγκεκριμένα της Pymongo βιβλιοθήκης, εξάγεται συγκεκριμένο "υπό-dataset" βάση συγκεκριμένων κριτηρίων και παράγονται τα απαραίτητα αρχεία εισόδου για την ανάπτυξη των αλγορίθμων συστημάτων συστάσεων.

Αρχικά, τίθεται ως κριτήριο να επιστραφούν τα πρώτα 750 έγγραφα τα οποία έχουν ημερομηνία συγγραφής μεγαλύτερη ή ίση του 2019, για να αξιοποιηθούν τα πιο πρόσφατα έγγραφα. Για το κάθε έγγραφο εξάγονται οι αναφορές / references που χρησιμοποιεί και στο τέλος αποθηκεύονται σε μια λίστα το σύνολο των ids όλων των εγγράφων που χρησιμοποιούνται ως αναφορά στα πρώτα 750 έγγραφα.

Σε αυτό το σημείο δίνεται η επιλογή της μεταβλητής depth που ορίζει πόσες φορές θα εκτελεστεί ο παρακάτω κώδικας και ανάλογα βγαίνουν διαφορετικά μεγέθη αρχείων. Αν το depth ισούται με 0 τότε προσπερνάτε το παρακάτω κομμάτι κώδικα και πηγαίνει απευθείας στο επόμενο.

Αν το depth είναι μεγαλύτερο από 0 : Από τη λίστα των συνολικών αναφορών που έχει προκύψει επιλέγεται το 1/3 των εγγράφων και προστίθενται στον συνολικό αριθμό από έγγραφα που αποσπούνται από τη βάση δεδομένων. Τίθεται και πάλι περιορισμός ώστε και τα επόμενα έγγραφα που θα επιλεχθούν να έχουν και πάλι έτος συγγραφής μεγαλύτερο ή ίσο του 2019. Στη συνέχεια επαναλαμβάνεται η παραπάνω διαδικασία ώστε να ξανά δημιουργηθεί η λίστα με τα επόμενα ids των εγγράφων που θα εξαχθούν. Η διαδικασία αυτή επαναλαμβάνεται τόσες φορές όσες και η τιμή που έχει λάβει η μεταβλητή depth.

Στο τέλος, διαπερνάται το σύνολο από έγγραφα τα οποία έχουν εξαχθεί και ελέγχεται αν στις αναφορές / references υπάρχουν τιμές από ids εγγράφων τα οποία δεν χρειάζονται. Αν ναι, τότε διαγράφονται, καθώς δεν θα έχει λογική να υπάρχει στο τελικό dataset που θα χρησιμοποιηθεί για την ανάπτυξη των αλγορίθμων κάποια αναφορά σε ένα έγγραφο το οποίο όμως δεν θα είναι δυνατό να ελεγχθεί. Έτσι σχηματίζεται ένα dictionary του οποίου κάθε γραμμή αποτελεί και ένα έγγραφο, και κάθε έγγραφο αντίστοιχα έχει ως key-value ζεύγη τις τιμές του εκάστοτε εγγράφου όπως ήταν αποθηκευμένο στη MongoDB.

Ενδεικτικά για να είναι γνωστοί οι αριθμοί των συγγραμμάτων και συγγραφέων που επεξεργάζονται:

DEPTH=0	DEPTH=1	DEPTH=2
USERS: 2111 ITEMS: 750	USERS: 10701 ITEMS: 4304	USERS: 35272 ITEMS: 18157

Στη συνέχεια αρχίζει η διαδικασία της επεξεργασίας αυτών των δεδομένων με στόχο

την παραγωγή των απαραίτητων εγγράφων που χρησιμοποιούνται ως είσοδος / input κατά την ανάπτυξη των αλγορίθμων συστημάτων συστάσεων. Αρχικά δημιουργούνται δύο πίνακες που αντιστοιχίζουν τα ids των εγγράφων αλλά και των συγγραφέων σε αύξουσες τιμές που να αρχίζουν από 1,2,...,συνολικός αριθμός εγγράφων / number of docs και 1,2,..., συνολικός αριθμός συγγραφέων / number of authors, καθώς στη βάση δεδομένων τα ids έχουν τυχαίες τιμές. Έτσι είναι πιο εύκολη η λήψη του συνολικού αριθμού από έγγραφα και συγγραφείς που δίνονται ως είσοδος στον κάθε αλγόριθμο. Οι πίνακες αυτοί ονομάζονται dict_ids_of_docs_to_use και dict_total_authors αντίστοιχα.

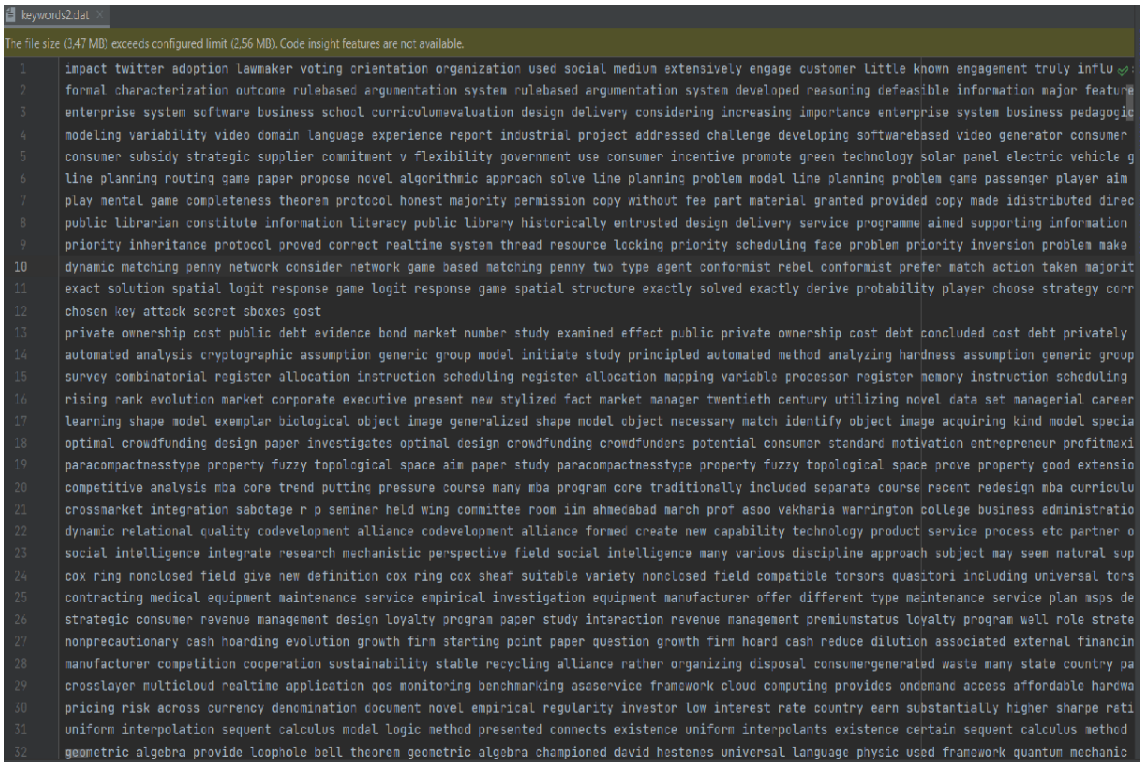
3.4.1 tags.dat

Το πρώτο έγγραφο που δημιουργείται είναι το keywords2.dat και περιέχει το σύνολο από λέξεις-κλειδιά που χρησιμοποιεί το κάθε έγγραφο. Πιο αναλυτικά:

Δημιουργείται μια ενιαία συμβολοσειρά / string από τις λέξεις που περιέχονται τόσο στον τίτλο όσο και στην περιγραφή /abstract του κάθε εγγράφου. Στη βάση δεδομένων το abstract δεν δίνεται ως ένα ενιαίο string αλλά ως ένα dictionary όπου key η λέξη και value το πόσες φορές εμφανίζεται. Συνεπώς, διατρέχοντας αυτό το dictionary παράγεται ένα string με όλες τις λέξεις, όσες φορές αντιστοιχεί στη κάθε μια. Το string αυτό όταν προκύπτει δεν είναι αναγνώσιμο αλλά αυτό δεν επηρεάζει τη μετέπειτα επεξεργασία που πραγματοποιείται. Στο string που προκύπτει γίνεται μια σειρά από επεξεργασίες με στόχο την εξόρυξη των λέξεων κλειδιών που χρησιμοποιούνται. Συγκεκριμένα, χρησιμοποιούνται Regular Expressions από τη βιβλιοθήκη Python re για να εκτελεσθούν διαφορετικές εργασίες προ-επεξεργασίας. Αρχικά, μετατρέπονται τα δεδομένα σε πεζά ώστε οι λέξεις που είναι στην πραγματικότητα οι ίδιες αλλά είναι σε διαφορετική μορφή να μπορούν να αντιμετωπίζονται ισότιμα. Στη συνέχεια αφαιρούνται όλοι οι μη λεκτικοί χαρακτήρες, όπως ειδικοί χαρακτήρες, αριθμοί κ.λπ. και καταργούνται όλοι οι μεμονωμένοι χαρακτήρες.

Στο επόμενο βήμα αφαιρούνται όλες οι “stopwords” μέσω της χρήσης της ntlk βιβλιοθήκης. Οι “stopwords” είναι λέξεις που χρησιμοποιούνται συνήθως (όπως “το”, “από”, “ένα”, “σε”) που μια μηχανή αναζήτησης έχει προγραμματιστεί να αγνοήσει, τόσο κατά την ευρετηρίαση καταχωρίσεων για αναζήτηση όσο και κατά την ανάκτησή τους ως αποτέλεσμα ενός ερωτήματος αναζήτησης.

Το τελευταίο βήμα προ-επεξεργασίας είναι η λεμετοποίηση / lemmetization. Στη λεμετοποίηση, μειώνεται κάθε λέξη σε μορφή ρίζας λεξικού. Για παράδειγμα, οι “γάτες” μετατρέ-



Σχήμα 3.3: keywords2.dat

πονται σε "γάτα". Η διαδικασία αυτή γίνεται προκειμένου να αποφευχθεί η εξόρυξη λέξεων που είναι σημασιολογικά παρόμοιες αλλά συντακτικά διαφορετικές. Για παράδειγμα, δεν πρέπει να θεωρούνται ως δύο διαφορετικές λέξεις – κλειδιά οι λέξεις "γάτες" και "γάτα", τα οποία είναι σημασιολογικά παρόμοια, επομένως εκτελείται λεμετοποίηση.

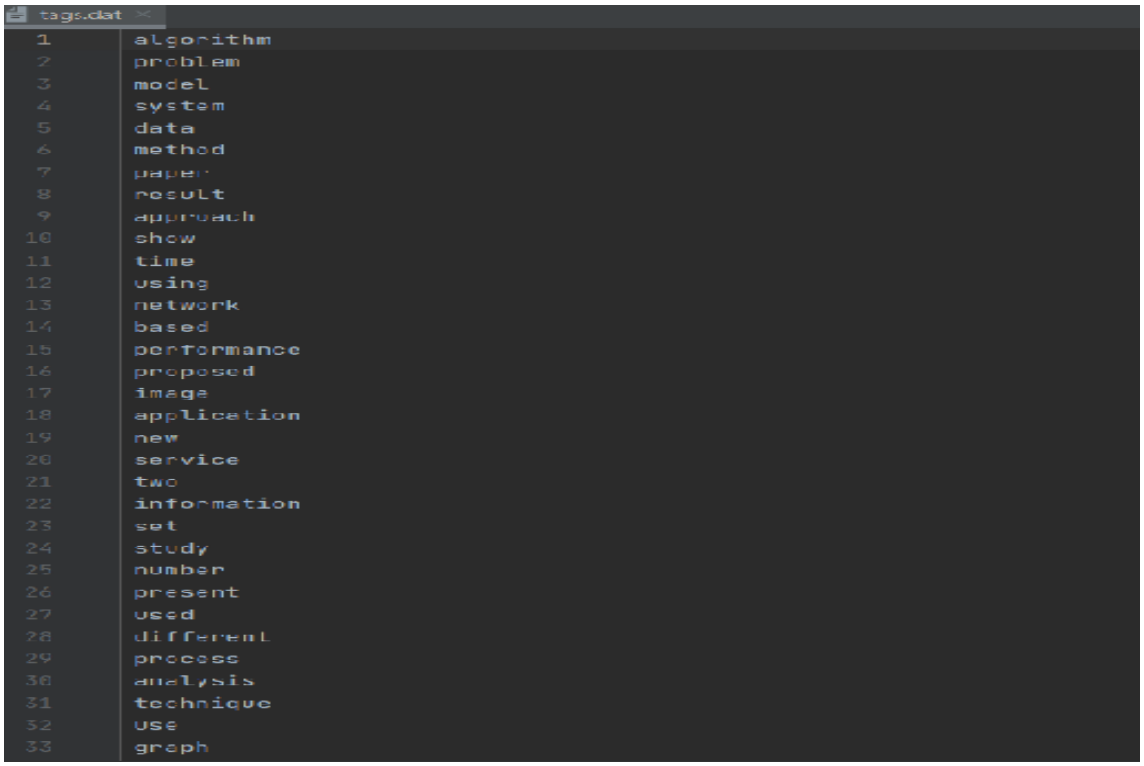
Η μορφή του keywords2.dat αρχείου απεικονίζεται στο Σχήμα 3.3. Τέλος, αφού έχει δημιουργηθεί το έγγραφο keywords2.dat, χρησιμοποιείται ώστε να εξαχθούν οι πιο επαναλαμβανόμενες λέξεις. Για να επιτευχθεί αυτό, χρησιμοποιείται το CountVectorizer από τη sklearn βιβλιοθήκη (sklearn.feature_extraction.text), όπου ορίζεται να εξαχθούν οι 1500 πιο σημαντικές λέξεις και έτσι προκύπτει το tags.dat έγγραφο.

Η μορφή του tags.dat αρχείου απεικονίζεται στο Σχήμα 3.4.

3.4.2 My_mult_nor.mat

Στο επόμενο βήμα εκτελείται TF-IDF με τη χρήση του TfidfVectorizer από τη sklearn βιβλιοθήκη (sklearn.feature-extraction.text).

Το TF-IDF σημαίνει Term Frequency - Inverse Document Frequency και είναι μια στατιστική που στοχεύει στον καλύτερο προσδιορισμό του πόσο σημαντική είναι μια λέξη για ένα



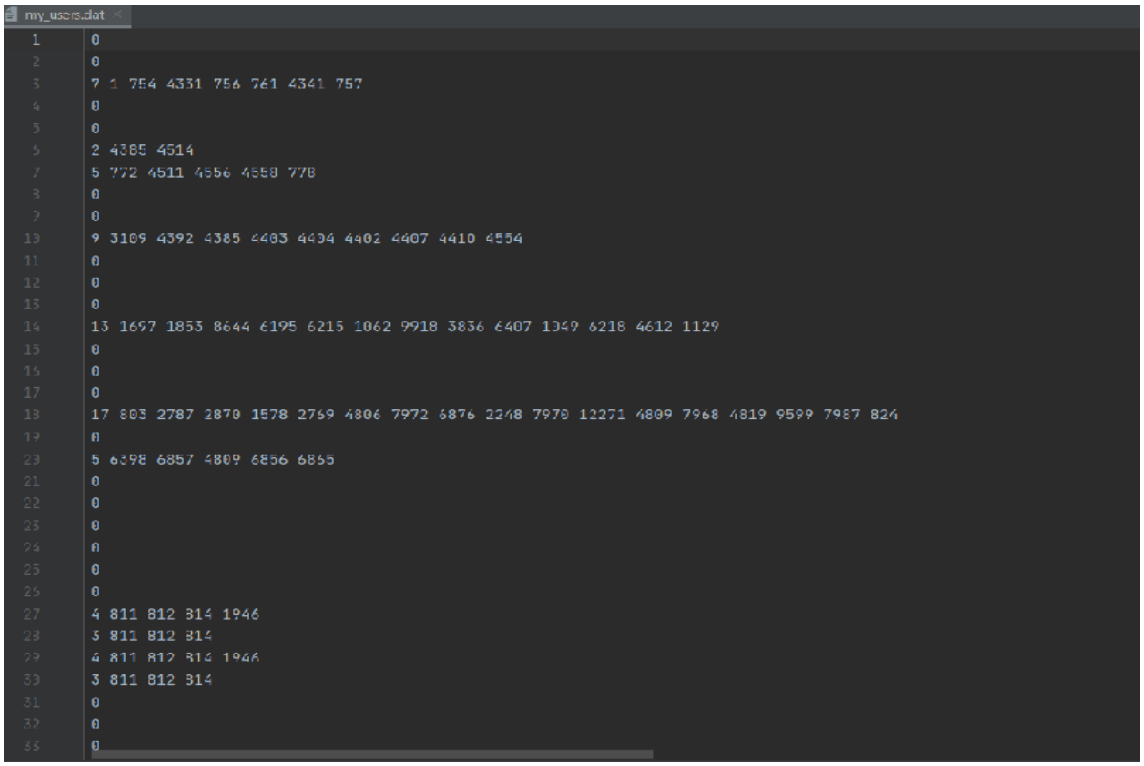
Σχήμα 3.4: tags.dat

έγγραφο, λαμβάνοντας παράλληλα υπόψη τη σχέση με άλλα έγγραφα του ίδιου σώματος.

Αυτό πραγματοποιείται εξετάζοντας πόσες φορές μια λέξη εμφανίζεται σε ένα έγγραφο, ενώ παράλληλα προσέχει πόσες φορές η ίδια λέξη εμφανίζεται σε άλλα έγγραφα στο σώμα.

Το σκεπτικό πίσω από αυτό είναι το εξής:

Μια λέξη που εμφανίζεται συχνά σε ένα έγγραφο έχει μεγαλύτερη συνάφεια με αυτό το έγγραφο, πράγμα που σημαίνει ότι υπάρχει μεγαλύτερη πιθανότητα το έγγραφο να αφορά κάτι σχετικό με τη συγκεκριμένη λέξη. Μια λέξη που εμφανίζεται συχνά σε περισσότερα έγγραφα μπορεί να εμποδίσει να βρεθεί το σωστό έγγραφο σε μια συλλογή. Η λέξη είναι σχετική είτε για όλα τα έγγραφα είτε για κανένα. Είτε έτσι είτε αλλιώς, δεν θα βοηθήσει να φιλτραριστεί ένα μόνο έγγραφο ή ένα μικρό υποσύνολο εγγράφων από ολόκληρο το σετ. Έτσι, το TF-IDF είναι μια βαθμολογία που εφαρμόζεται σε κάθε λέξη σε κάθε έγγραφο στο σύνολο δεδομένων μας. Και για κάθε λέξη, η τιμή TF-IDF αυξάνεται με κάθε εμφάνιση της λέξης σε ένα έγγραφο, αλλά μειώνεται σταδιακά με κάθε εμφάνιση σε άλλα έγγραφα ([14]). Αφού έχει εκτελεστεί το TF-IDF, χρησιμοποιείται το savemat από τη scipy.io.savemat βιβλιοθήκη για να παραχθεί το τελικό My_mult_nor.mat αρχείο.

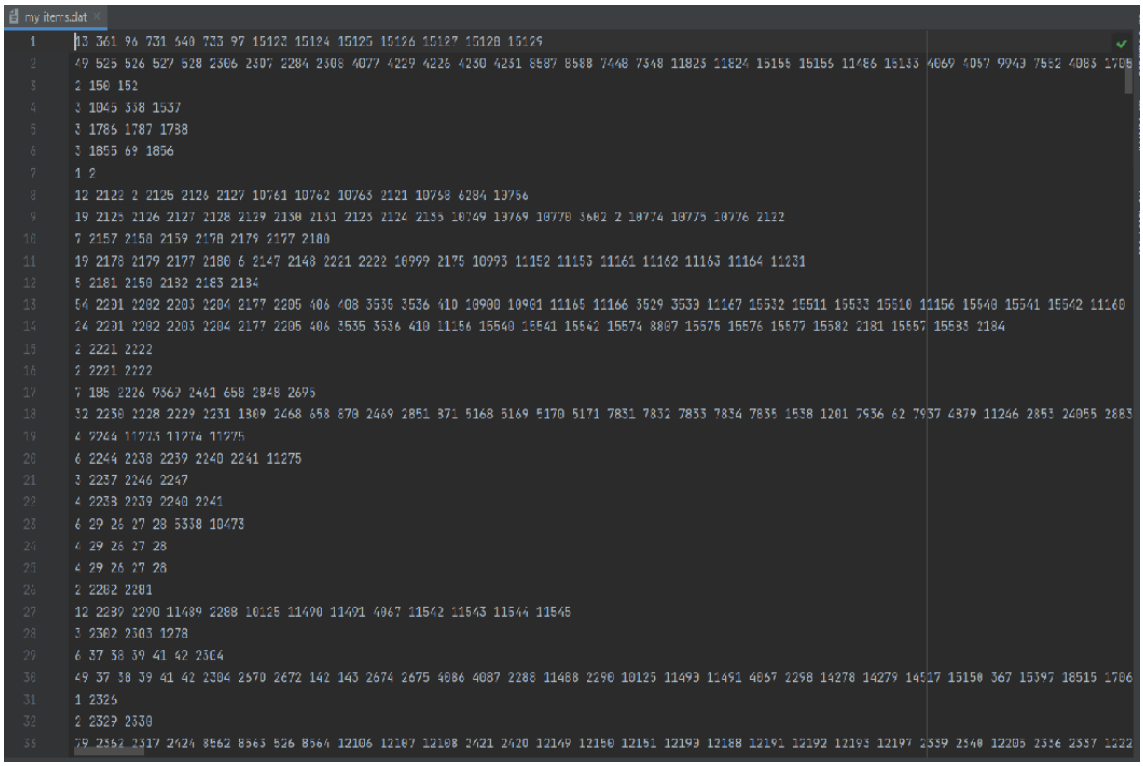


Σχήμα 3.5: users.dat

3.4.3 users.dat

Σε αυτό το αρχείο κάθε γραμμή αναπαριστά ένα συγγραφέα / author και συνεπώς ο αριθμός γραμμών του αρχείο είναι τόσος όσος και ο αριθμός των συγγραφέων / users που έχουν βρεθεί από την αρχική διαδικασία. Ο σκοπός αυτού του αρχείου είναι να δείξει πόσες αναφορές / references έχει χρησιμοποιήσει συνολικά ένας συγγραφέας στο σύνολο των συγγραμμάτων που έχει γράψει / έχει συμμετάσχει. Συνεπώς, διατρέχονται όλα τα συγγράμματα που έχουν εξαχθεί από τη βάση, και ταυτόχρονα δημιουργείται ένα dictionary όπου αποθηκεύονται όλες οι αναφορές που συναντούνται σε κάθε έγγραφο και αντίστοιχα και όλους τους συγγραφείς που το έχουν χρησιμοποιήσει ως αναφορά. Τέλος, διατρέχεται το τελικό dictionary και καταγράφεται στο αρχείο για κάθε συγγραφέα πόσες αναφορές έχει χρησιμοποιήσει και ποιες. Τα ids των άρθρων που χρησιμοποιούνται ως αναφορές αντιστοιχίζονται με τις ανάλογες τιμές που έχουν αποθηκευτεί στο dict_ids_of_docs_to_use.

Η μορφή του users.dat αρχείου απεικονίζεται στο Σχήμα 3.5.

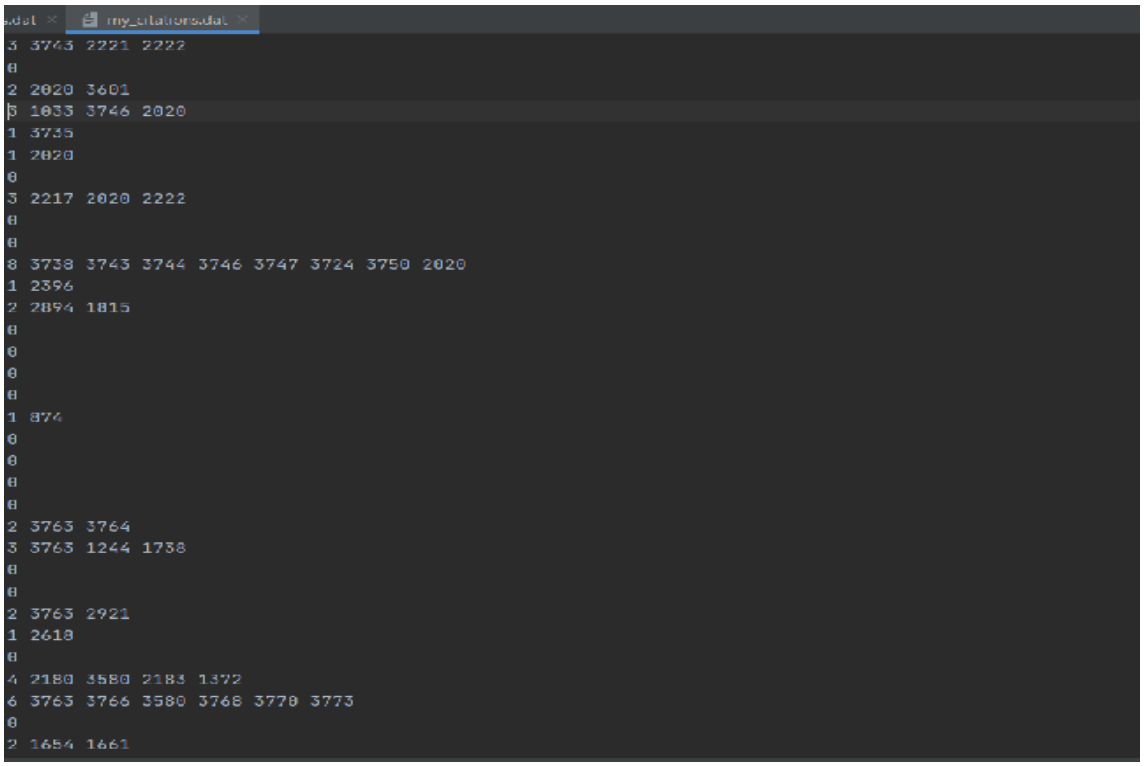


Σχήμα 3.6: items.dat

3.4.4 items.dat

Σε αυτό το αρχείο κάθε γραμμή αναπαριστά ένα έγγραφο / document και συνεπώς ο αριθμός γραμμών του αρχείο είναι τόσος όσος και ο αριθμός των εγγράφων / documents που έχουν βρεθεί από την αρχική διαδικασία. Ο σκοπός αυτού του αρχείου είναι να δείξει πόσοι συγγραφείς έχουν χρησιμοποιήσει το εκάστοτε έγγραφο ως αναφορά και ποιοι. Συνεπώς, διατρέχονται όλα τα συγγράμματα που έχουν εξαχθεί από τη βάση, και ταυτόχρονα δημιουργείται ένα dictionary όπου αποθηκεύονται όλα τα άρθρα που χρησιμοποιούνται ως αναφορά που συναντώ σε κάθε έγγραφο και αντίστοιχα και όλους τους συγγραφείς που το έχουν χρησιμοποιήσει ως αναφορά. Τέλος, διατρέχεται το τελικό dictionary και καταγράφεται στο αρχείο για κάθε έγγραφο πόσοι το έχουν χρησιμοποιήσει ως αναφορά και ποιοι. Τα ids των συγγραφέων που έχουν χρησιμοποιήσει κάποια αναφορά αντιστοιχίζονται με τις ανάλογες τιμές που έχουν αποθηκευτεί στο dict_total_authors.

Η μορφή του items.dat αρχείου απεικονίζεται στο Σχήμα 3.6.



Σχήμα 3.7: citation.dat

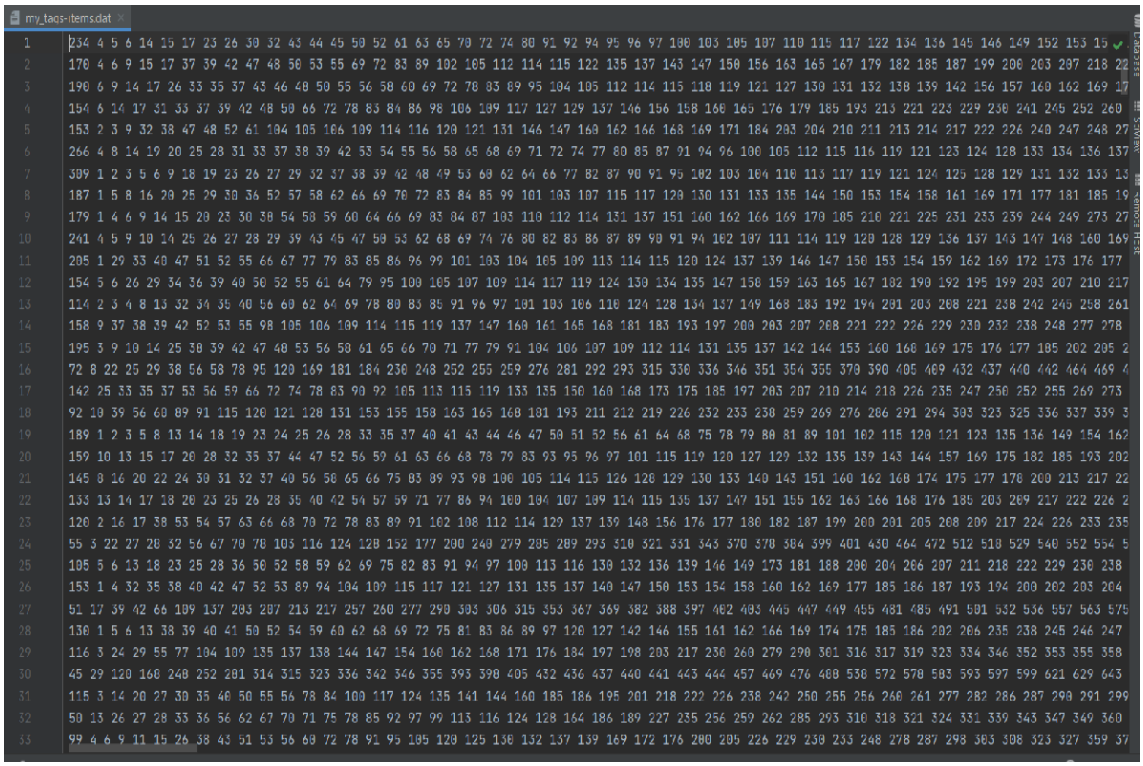
3.4.5 citation.dat

Σε αυτό το αρχείο κάθε γραμμή αναπαριστά και πάλι ένα έγγραφο / document και συνεπώς ο αριθμός γραμμών του αρχείο είναι τόσος όσος και ο αριθμός των εγγράφων / documents που έχουν βρεθεί από την αρχική διαδικασία. Ο σκοπός αυτού του αρχείου είναι να δείξει πόσες αναφορές / references έχει χρησιμοποιήσει το εκάστοτε έγγραφο και ποιες. Συνεπώς, εδώ είναι πιο εύκολο καθώς διατρέχονται όλα τα συγγράμματα που έχουν εξαχθεί από τη βάση, και ταυτόχρονα δημιουργείται ένα dictionary όπου αποθηκεύονται όλα τα άρθρα που χρησιμοποιεί το κάθε άρθρο ως αναφορά διαδοχικά. Τέλος, διατρέχεται το τελικό dictionary και καταγράφονται στο αρχείο για κάθε έγγραφο πόσες αναφορές έχει χρησιμοποιήσει και ποιες. Τα ids των εγγράφων που χρησιμοποιούνται ως αναφορά τα αντιστοιχίζονται με τις ανάλογες τιμές που έχουν αποθηκευτεί στο dict_ids_of_docs_to_use.

Η μορφή του citation.dat αρχείου απεικονίζεται στο Σχήμα 3.7.

3.4.6 tag-items.dat

Σε αυτό το αρχείο κάθε γραμμή αναπαριστά ένα tag, δηλαδή μια από τις 1500 λέξεις που έχουν εξαχθεί ως οι πιο αντιπροσωπευτικές, και συνεπώς ο αριθμός γραμμών του αρχείο εί-



Σχήμα 3.8: tag-items.dat

ναι 1500. Ο σκοπός αυτού του αρχείου είναι να δείξει πόσα έγγραφα εμπεριέχουν το κάθε tag είτε στον τίτλο τους, είτε στο abstract, και ποια. Συνεπώς, διατρέχονται για κάθε λέξη όλα τα συγγράμματα που έχουν εξαχθεί από τη βάση, αλλά εδώ αξιοποιείται το keywords2.dat όπου κάθε γραμμή αναπαριστά τις λέξεις που έχει κάθε έγγραφο τόσο στον τίτλο όσο και στο abstract. Έτσι αν υπάρχει το tag που εξετάζεται στην αντίστοιχη σειρά του keywords2.dat, τότε δημιουργείται ένα dictionary όπου αποθηκεύεται αντίστοιχα το id του άρθρου που χρησιμοποιεί το tag. Τέλος, διατρέχεται το τελικό dictionary και καταγράφεται στο αρχείο για κάθε tag πόσες φορές έχει χρησιμοποιηθεί από κάποιο άρθρο και από ποια. Τα ids των εγγράφων που έχουν χρησιμοποιήσει τα tags είναι ήδη αντιστοιχισμένα αφού αποθηκεύεται ως τιμή id του άρθρου η γραμμή του εγγράφου line που βρισκόταν στο keywords2.dat.

Η μορφή του tag-items.dat αρχείου απεικονίζεται στο Σχήμα 3.8.

Έτσι προκύπτουν όλα τα αρχεία εισόδου που δίνονται στους αλγορίθμους. Τα αρχεία αυτά ανάλογα με την τιμή της αρχικής μεταβλητής depth έχουν και ανάλογο μέγεθος και συνεπώς μπορεί να εξεταστεί ο κάθε αλγόριθμος μέχρι πόσου μεγέθους δεδομένα μπορεί να αξιοποιήσει και να βγάλει ορθά αποτελέσματα.

3.5 Δεύτερη Μέθοδος Επεξεργασίας των δεδομένων

Αντίστοιχη διαδικασία έχει ακολουθήσει και ο κύριος Στεργιόπουλος Βάιος και έχουν παραχθεί ανάλογα δεδομένα χωρίς όμως να εκμεταλλεύεται τη MongoDB βάση δεδομένων, αλλά επεξεργάζοντας τα δεδομένα απευθείας από την αρχική json μορφή τους.

Πιο συγκεκριμένα:

Αρχικά από το json αρχείο εξάγεται μια πιο μικρή βάση δεδομένων η οποία αποτελείται από τα 750 πρώτα συγγράμματα έχουν εκδοθεί από το 2020 και μετά. Στη συνέχεια, όπως και πιο πριν, εξάγονται τα references των άρθρων αυτών, αποκλείονται όσα έχουν συγγραφεί πριν το 2020, και από αυτά που απομένουν εξάγεται ένα ποσοστό ($\simeq 30\%$). Τέλος, επαναλαμβάνεται αυτή η διαδικασία άλλη μια φορά και έτσι προκύπτει στο τελικό dataset το οποίο αποθηκεύεται σε .dat μορφή αρχείου.

Στη συνέχεια, έχοντας δημιουργήσει την υπό-βάση που θα αξιοποιηθεί, ακολουθεί η επεξεργασία των δεδομένων. Για κάθε σύγγραμμα χρησιμοποιούνται τα: τίτλος, περίληψη και πεδίο σπουδών, τα οποία επεξεργάζονται όπως και προηγουμένως και προκύπτουν οι τελικές λέξεις-κλειδιά / keywords.

Πιο συγκεκριμένα, η διαδικασία επεξεργασίας των δεδομένων είναι η εξής:

1. Δημιουργία των mult_nor.mat και tag-item.dat μέσω της TagFiles() συνάρτησης. Πρώτα καλείται η tf_idf_function() η οποία εκτελεί tf_idf vectorization στο αρχείο από λέξεις κλειδιά που δημιουργήθηκε προηγουμένως. Το αποτέλεσμα αυτό αποθηκεύεται σε Matlab μορφή. Έπειτα καλείται η createTagItem(), η οποία φτιάχνει ένα dictionary που αντιστοιχίζει τις λέξεις κλειδιά με τα άρθρα στα οποία περιέχονται αυτές. Έτσι, δημιουργούνται δύο αρχεία: στο ένα υπάρχει η αντιστοιχία αριθμού γραμμής με την λέξη κλειδί, και στο άλλο σε κάθε γραμμή αναγράφεται ποια άρθρα περιέχουν την αντίστοιχη λέξη.

2. Δημιουργία του citation.dat αρχείου μέσω της CitationFile() συνάρτησης. Πρώτα καλείται η retrieve_ids_citation() συνάρτηση που δημιουργεί το citations.dat αρχείο με όλα τα ids που χρησιμοποιούνται ως αναφορά / reference στο κάθε άρθρο που εξετάζεται. Έπειτα καλείται η break_ids_citation() συνάρτηση που δημιουργεί δύο αρχεία. Το ένα αρχείο αποθηκεύει την αντιστοιχία αναγνωριστικών, έτσι ώστε ο αριθμός γραμμής να αντιπροσωπεύει το αναγνωριστικό / id του άρθρου, και το άλλο αρχείο την αντιστοιχία αριθμού γραμμής με το αναγνωριστικό / id της παραπομπής. Τέλος, καλείται

η replaceIds() συνάρτηση η οποία αλλάζει το αρχικό citations.dat αρχείο ώστε κάθε id να αντικατασταθεί από τον αντίστοιχο αριθμό γραμμής που προκύπτει από την προηγούμενη συνάρτηση. Έτσι όλα τα άρθρα έχουν αναγνωριστικά που αρχίζουν από τον αριθμό ένα και αυξάνονται κατά ένα.

3. Δημιουργία των cf-train και cf-test αρχείων μέσω της TrainingFiles() συνάρτησης. Πρώτα καλείται η retrieveUsersItems() συνάρτηση. Η συνάρτηση ανακτά όλους τους χρήστες – συγγραφείς / users και τα αντικείμενα – άρθρα / items που του αρέσουν – έχει χρησιμοποιήσει ως αναφορά. Για κάθε γραμμή της βάσης δεδομένων που αντιστοιχίζεται με ένα σύγγραμμα συλλέγονται όλα τα αναγνωριστικά των συγγραφέων, και των συγγραμμάτων που χρησιμοποιούνται ως αναφορές. Διατρέχοντας όλη τη βάση δημιουργείται ένα dictionary που αντιστοιχίζει το κάθε author id στο σύνολο των αναφορών που έχει χρησιμοποιήσει σε όλα τα συγγράμματα που έχει συμμετάσχει. Στη συνέχεια, αντικαθίσταται το αναγνωριστικό των συγγραμμάτων με τον αντίστοιχο αριθμό γραμμής του αρχείου που δημιουργήθηκε προηγουμένως. Τέλος, επιλέγονται μόνο οι συγγραφείς που έχουν χρησιμοποιήσει πάνω από πέντε συγγράμματα ως αναφορές και γράφονται στο τελικό αρχείο. Έπειτα, το αρχείο αυτό δίνεται ως όρισμα στη συνάρτηση createUsersFiles() η οποία δημιουργεί τέσσερα αρχεία:

   cf-train-10-users.dat, cf-test-10-users.dat -> χωρίζονται τα δεδομένα σε 20% training και 80% testing

   cf-train-1-users.dat, cf-test-1-users.dat -> χωρίζονται τα δεδομένα σε 10% training και 90% testing.

   Για την παραγωγή των παραπάνω αρχείων, αγνοούνται όσοι συγγραφείς έχουν λιγότερο από δύο αναφορές.

   Στη συνέχεια, καλείται η συνάρτηση createBaseItemFile (). Κατά την εκτέλεσή της, δημιουργείται ένα dictionary που έχει ως πεδίο κλειδί το αναγνωριστικό ενός άρθρου, και ως τιμή τη λίστα των συγγραφέων που το έχουν χρησιμοποιήσει ως αναφορά. Οι τιμές των αναγνωριστικών αντιστοιχίζονται και πάλι με αυτές των αριθμών γραμμής μέσω των αρχείων που έχουν ήδη δημιουργηθεί. Τέλος δημιουργείται ένα αρχείο που κάθε του γραμμή αντιστοιχίζεται με το id ενός άρθρου, και για το κάθε άρθρο αναφέρεται πόσοι και ποίοι συγγραφείς το έχουν χρησιμοποιήσει ως αναφορά. Τέλος, καλείται η createItemsFiles() συνάρτηση η οποία δημιουργεί τα “-cf-train-10-items.dat” , “-cf-

train-1-items.dat", "-cf-test-10-items.dat", "-cf-test-1-items.dat" αρχεία. Δίνοντας ως είσοδο το αρχείο που προκύπτει από την παραπάνω συνάρτηση χωρίζει τα δεδομένα βάση του αριθμού των συγγραφέων που έχουν χρησιμοποιήσει το κάθε άρθρο ως αναφορά και έτσι προκύπτει ο διαχωρισμός τους στα τέσσερα αυτά αρχεία.

Κατά τη διαδικασία ανάλυσης και σύγκρισης των αλγορίθμων χρησιμοποιούνται τα αρχεία που έχουν παραχθεί από τον κύριο Στεργιόπουλο με στόχο τη μέγιστη απόδοση των αλγορίθμων και την καλύτερη σύγκρισή τους. Τα δεδομένα αυτής της μορφής έχουν δημιουργηθεί έτσι ώστε να απαντούν σε τρία διαφορετικά ερωτήματα, αλλά και για κάθε ερώτημα έχουν αναπτυχθεί τρία διαφορετικά κριτήρια παραγωγής της υπό-βάσης δεδομένων.

Αναλυτικότερα, όπως προαναφέρθηκε χρησιμοποιούνται τα πεδία id, title, authors, year, keywords, references, fos, indexed_abstract και venue, τα οποία επεξεργάζονται με τον παραπάνω τρόπο με στόχο να μπορούν να απαντηθούν τα ακόλουθα ερωτήματα:

1. Χρήση των δεδομένων με στόχο τη σύσταση ενός χώρου δημοσιεύσεων συγγραμμάτων / publication venue recommendation στους συντάκτες επιστημονικών εγγράφων σύμφωνα με τα ερευνητικά τους ενδιαφέροντα

2. Χρήση των δεδομένων με στόχο τη σύσταση επιστημονικών άρθρων για ανάγνωση και μελέτη στους συντάκτες επιστημονικών περιοδικών και βιβλίων / article recommendation

3. Χρήση των δεδομένων με στόχο τη σύσταση ερευνητών-συντακτών επιστημονικών άρθρων σε άλλους ερευνητές-συντάκτες επιστημονικών άρθρων για πιθανή συνεργασία / user recommendation

Δημιουργούνται τρία διαφορετικά datasets για τη επίλυση του κάθε ερωτήματος:

1. Στο πρώτο λαμβάνονται υπόψιν το πεδίο field of study

2. Στο δεύτερο και πάλι λαμβάνεται υπόψιν το πεδίο field of study αλλά πλέον ζητούνται περισσότερα δεδομένα ώστε στο τέλος να προκύψει ένα μεγαλύτερο dataset και να συγκριθεί η αποτελεσματικότητα της μεθόδου δίνοντας ως είσοδο μεγαλύτερη βάση δεδομένων

3. Στο τρίτο dataset δεν λαμβάνεται υπόψιν το πεδίο field of study και ο αριθμός των δεδομένων που παράγονται είναι παρόμοιος με αυτόν που είχαν τα αρχικά δεδομένα

που αξιοποιούσε η κάθε μέθοδος απαντώντας στα ερωτήματα από την citeulike-a βάση δεδομένων, έτσι ώστε να μπορεί να γίνει βέλτιστη σύγκριση των μεθόδων χωρίς να διαφέρει το μέγεθος των δεδομένων που μπαίνουν ως είσοδος.

Ενδεικτικά ο αριθμός των συγγραμμάτων και συγγραφέων που προκύπτουν:

	Q1	Q2	Q3
FOS_DATASET_V2	USERS: 11227 ITEMS: 2075	USERS: 16898 ITEMS: 23719	USERS: 25564 ITEMS: 25564
FOS_DATASET_V3	USERS: 14687 ITEMS: 2389	USERS: 26059 ITEMS: 31487	USERS: 33132 ITEMS: 33132
RANDOM_DATASET	USERS: 8355 ITEMS: 2020	USERS: 10496 ITEMS: 17487	USERS: 18945 ITEMS: 18945

Κεφάλαιο 4

Μέτρηση αποτελεσματικότητας μεθόδων

4.1 Εκτελέσεις των συστημάτων συστάσεων

Έχοντας ετοιμάσει τις μεθόδους ανάπτυξης των συστημάτων συστάσεων, και έχοντας ήδη επεξεργαστεί και αναλύσει τα δεδομένα που δίνονται ως είσοδοι κατά την εκτέλεση των μεθόδων, τα συστήματα συστάσεων είναι έτοιμα. Παρακάτω αναλύεται ο τρόπος εκτέλεσης και αξιολόγησης των μεθόδων, και συγκρίνονται μεταξύ τους ώστε να καθοριστεί ποιος αλγόριθμος και με ποιες παραμέτρους είναι ο πιο αποδοτικός.

Τα συστήματα συστάσεων έχουν εκτελεστεί με τους εξής πιθανούς συνδυασμούς:

4.1.1 Εκτελέσεις σε διαφορετικό λογισμικό

- Σε laptop με τα εξής χαρακτηριστικά:
  - Επεξεργαστής Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz 2.00 GHz
  - Εγκατεστημένη RAM 6,00 GB
  - Τύπος συστήματος Λειτουργικό σύστημα 64 bit, επεξεργαστής τεχνολογίας x64
  - Χωρίς αξιοποίηση κάρτας γραφικών
- Σε server με τα εξής χαρακτηριστικά:
  - ο Επεξεργαστής Intel(R) Xeon(R) W-2123 CPU @ 3.60GHz
  - Εγκατεστημένη RAM 16,00 GB

- Τύπος συστήματος Λειτουργικό σύστημα 64 bit, επεξεργαστής τεχνολογίας x86_64

- NVIDIA-SMI 460.32.03 Driver Version: 460.32.03 CUDA Version: 11.2

Συγκεκριμένα, το περιβάλλον του pycharm που χρησιμοποιήθηκε για την ανάπτυξη και εκτέλεση του κώδικα δίνει τη δυνατότητα επιλογής "διερμηνέα" / interpreter. Συνεπώς, διαλέγοντας ως interpreter τον python3.6 ο κώδικας εκτελείται στο προσωπικό laptop με τα αντίστοιχα χαρακτηριστικά. Επιλέγοντας όμως ως python interpreter: remote python 3.6, ο κώδικας μπορεί να εκτελεστεί εξ' αποστάσεως / remotely σε κάποιον server. Ο UTH server διαθέτει 16GB RAM και NVIDIA GPU η οποία είναι εκμεταλλεύσιμη από το tensorflow για να εκπαιδεύει πιο γρήγορα τα Νευρωνικά Δίκτυα. Για να γίνει σύνδεση με τον server, πρέπει:

Ctr+alt+s → επιλογή του σωστού python interpreter → add new → ssh interpreter → χρειάζεται να δοθούν τα : host και username, password. Στη συνέχεια δίνεται το path που θα ανέβουν τα αρχεία στον server.

Tools → deployment→ configuration:

- CONNECTION:

  - Type = sftp

  - Δίνεται το σωστό ssh configuration για να γίνει η διασύνδεση με τον server, με test connection μπορεί να ελεγχθεί αν έχουν δοθεί τα σωστά στοιχεία.

  - Web server URL = url σύνδεσης του server

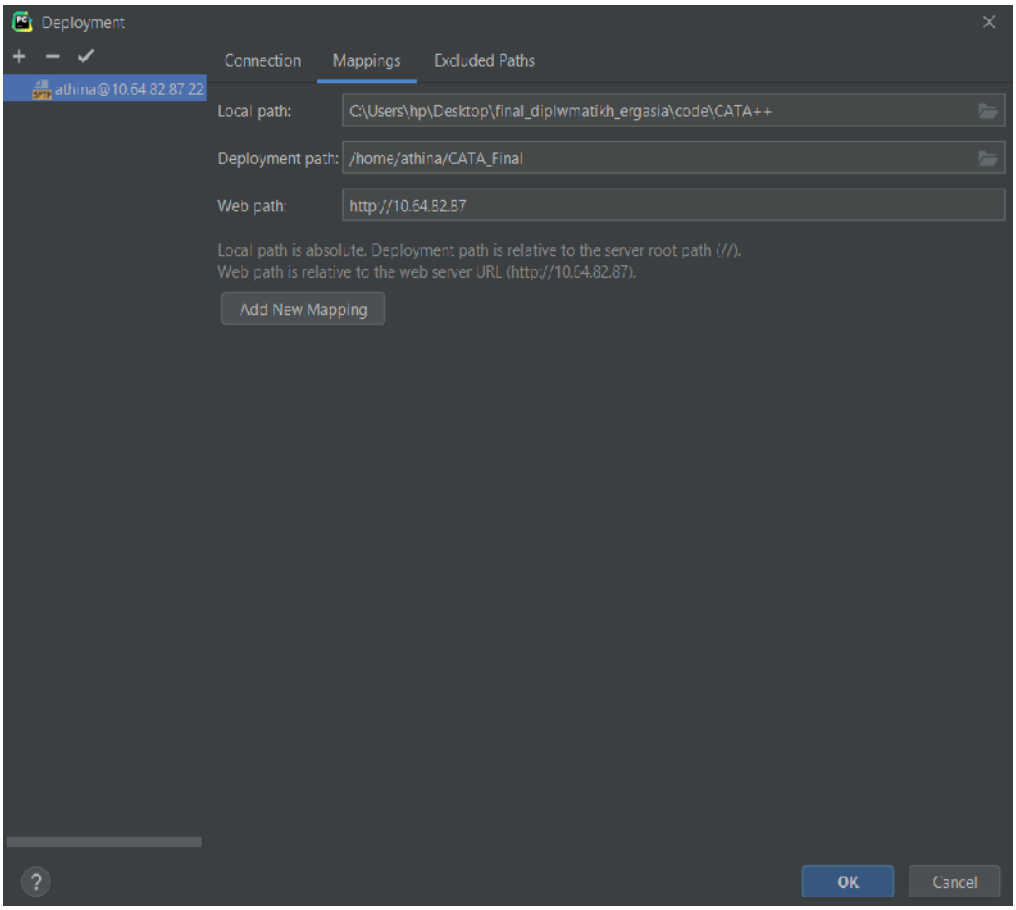
- Mappings:

  - Local path: το path του laptop στο οποίο βρίσκεται αποθηκευμένος ο κώδικας

  - Deployment path: το path του server στο οποίο γίνεται upload ο κώδικας στον server

  - Web path: είναι και πάλι η url σύνδεσης του server

Παρακάτω στις εικόνες 4.1, 4.2 και 4.3 φαίνονται οι αντίστοιχες ρυθμίσεις που έχω κάνω στο προσωπικό μου laptop ώστε να κάνω τη διασύνδεση με τον server της σχολής του Πανεπιστημίου Θεσσαλίας. Σχήμα 3.3.



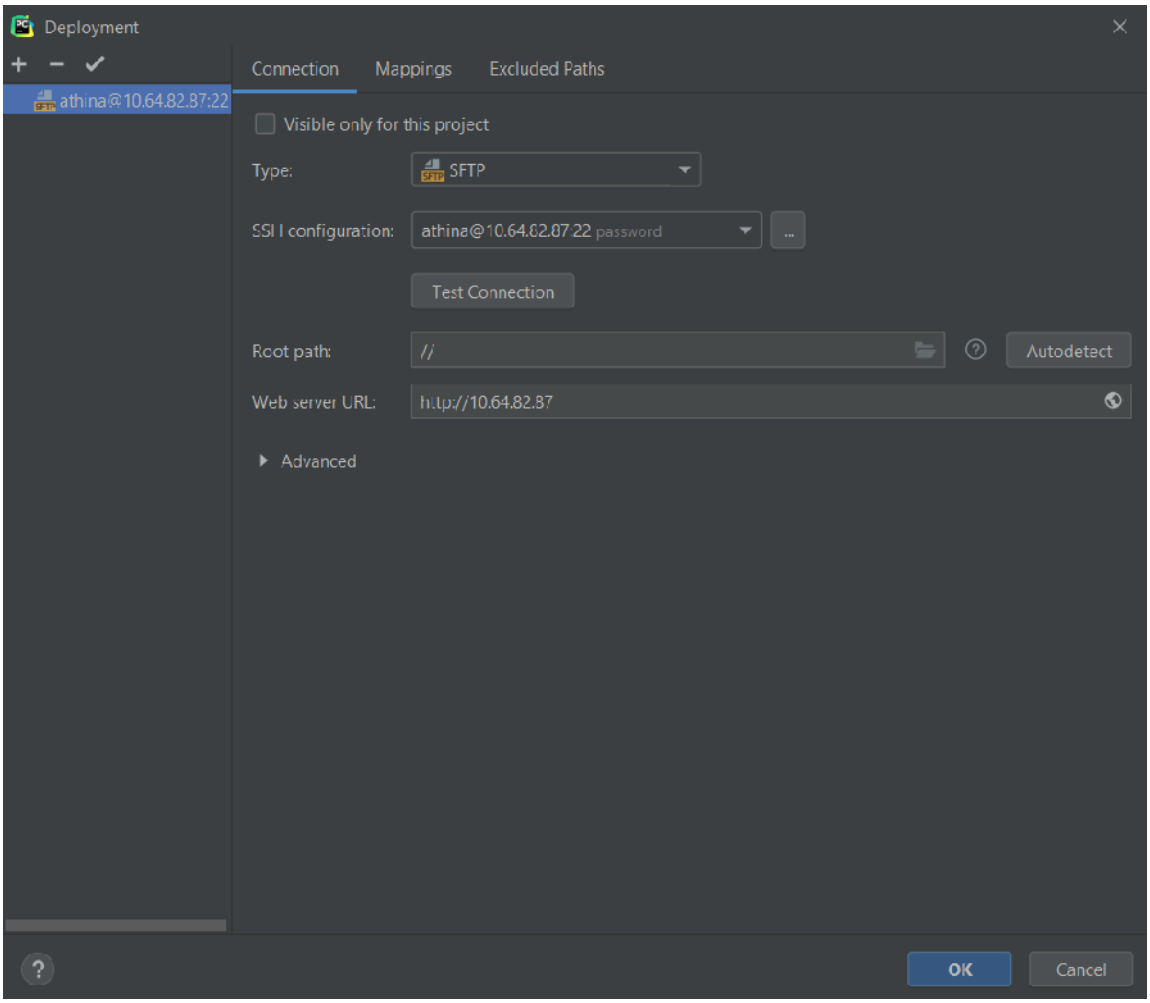
Σχήμα 4.1: Διασύνδεση με server i

Τέλος, πατώντας με δεξί κλικ πάνω στο αρχείο / φάκελο που πρέπει να ανέβει στον server: Deployment → Upload to → όνομα server, τα αρχεία ανεβαίνουν στον server και πλέον η κάθε εκτέλεση γίνεται στο περιβάλλον του server και όχι του laptop.

Η ρύθμιση του python interpreter είναι ξεχωριστή για κάθε project. Συνεπώς κάθε project μπορεί να τρέχει με διαφορετική έκδοση της python ή σε διαφορετικό περιβάλλον χωρίς να επηρεάζονται μεταξύ τους, κάτι το οποίο καθιστά το pycharm ένα πού εύχρηστο εργαλείο για την συγγραφή και εκτέλεση κώδικα.

4.1.2 Εκτελέσεις με διαφορετικές παραμέτρους

Ελέγχεται επίσης η αποτελεσματικότητα των μεθόδων βάση των αρχείων που δίνονται ως είσοδος αλλά και βάση των συναρτήσεων της keras βιβλιοθήκης που χρησιμοποιούνται ως activation functions και weight initialization. Συγκεκριμένα, δίνοντας ως είσοδο καθένα από τα 3 datasets που προαναφέρθηκαν (random_dataset, v2_fos_dataset, v3_fos_dataset) αλλά και τα αρχικά δεδομένα από το citeulike-a, μπορεί να ελεγχθεί η απόδοση των συστημάτων

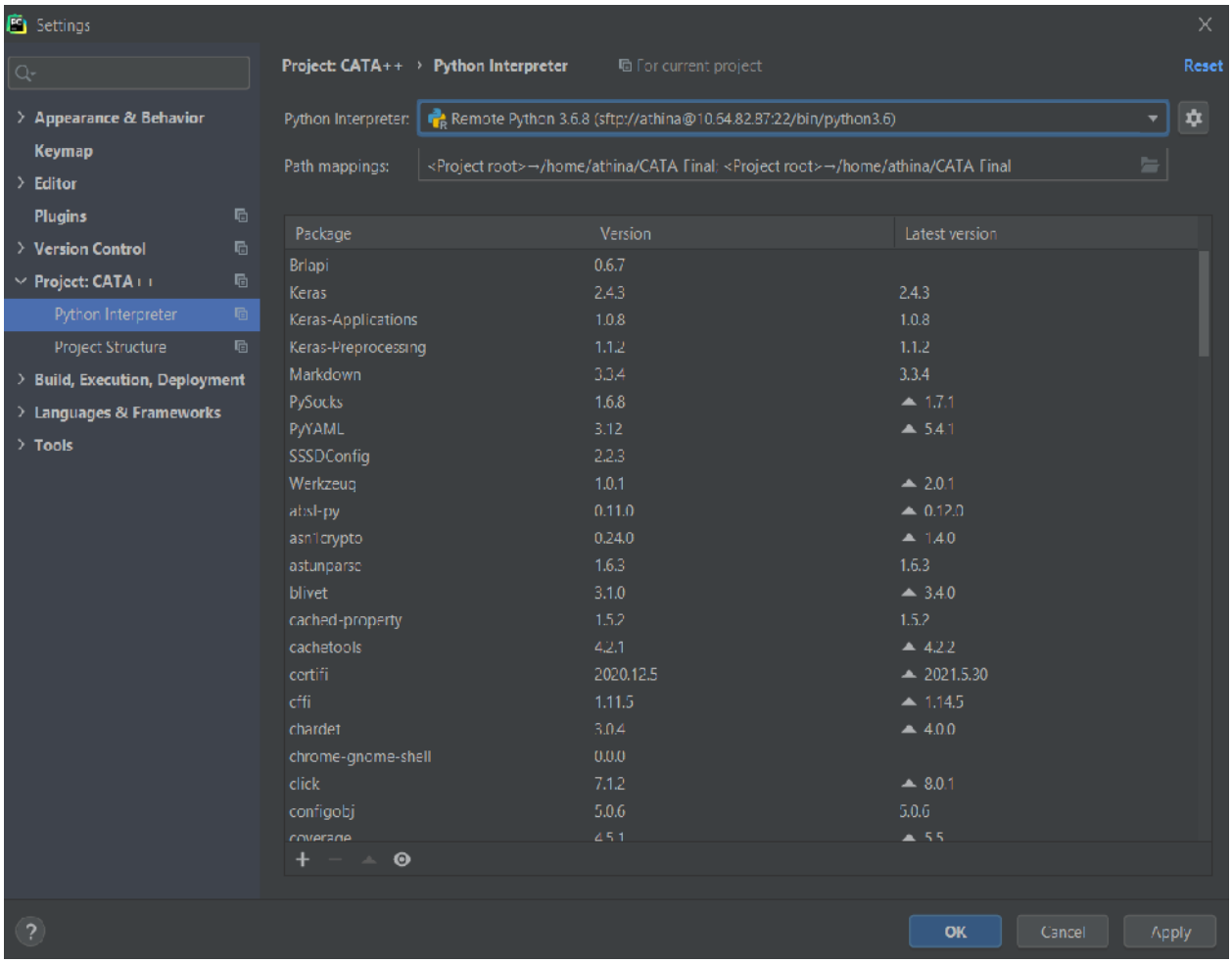


Σχήμα 4.2: Διασύνδεση με server ii

συστάσεων βάση της ποιότητας και του μεγέθους των αρχείων που δίνονται ως είσοδο για να τα επεξεργαστούν.

Επίσης, δοκιμάζοντας διάφορα activation functions και weight initiallizations συμπεραίνετε ότι χρησιμοποιώντας το SineRelu ως activation functions[[15]] και το he_normal [[16])] ως weight initialization βελτιώνεται η απόδοσή τους.

Συνεπώς, χρησιμοποιώντας τις τρεις μεθόδους αξιολόγησης των συστημάτων συστάσεων που αναφέρονται παρακάτω, και με όλους τους παραπάνω πιθανούς τρόπους εκτέλεσης, βασικό στόχο αποτελεί να βρεθεί ο βέλτιστος συνδυασμός για την ανάπτυξη του συστήματος συστάσεων.



Σχήμα 4.3: Διασύνδεση με server iii

4.2 Μετρικές αξιολόγησης των μεθόδων

Οι μετρικές αξιολόγησης των μεθόδων που έχουν χρησιμοποιηθεί είναι οι recall, dcg και ndcg.

4.2.1 Recall

$$\text{recall}\,\alpha K = \frac{|\ \text{Test Items} \cap \text{K Recommended Items}\ |}{|\ \text{Test Items}\ |} \tag{4.1}$$

Η recall υπολογίζεται ως ο λόγος των ορθώς προβλεπόμενων θετικών παρατηρήσεων προς όλες τις παρατηρήσεις. Έτσι, αναλόγως της ερώτησης η recall μας δείχνει :

Για τη πρώτη ερώτηση: τους τοπ K εκδοτικούς χώρους από τα test_publication_venues στους συντάκτες επιστημονικών εγγράφων για την υποβολή εγγράφων / συνολικό αριθμό test_publication_venues

Για την δεύτερη ερώτηση : τα τοπ K επιστημονικά συγγράμματα από τα test_articles που προτείνονται σε συντάκτες / συνολικό αριθμό test_articles

Για την τρίτη ερώτηση: οι τοπ K συντάκτες από τους test_users που προτείνονται σε άλλους ερευνητές με σκοπό μια πιθανή συνεργασία / συνολικό αριθμό test_users

4.2.2 DCG και nDCG

Η recall δεν κάνει κάποιο διαχωρισμό μεταξύ των τοπ K στοιχείων, αντιθέτως τα θεωρεί ισάξια. Γι' αυτό χρησιμοποιείται και η DCG η οποία αξιολογεί τη χρησιμότητα μιας πρότασης ανάλογα με τη θέση της στο αποτέλεσμα των τοπ K στοιχείων. Η χρησιμότητα ενός εγγράφου θεωρείται υψηλότερη στην κορυφή της λίστας των αποτελεσμάτων, και μειώνεται στις χαμηλότερες θέσεις.

Τέλος, η nDCG ταξινομεί τα αποτελέσματα που προκύπτουν βάση της σχετικής συνάφειάς τους, παράγοντας τη μέγιστη δυνατή DCG, που ονομάζεται IDCG.

Τύποι υπολογισμών των DCG και nDCG:

$$nDCG \propto K = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{DCG@K}{IDCG@K} \tag{4.2}$$

Όπου:

$$DCG@K = \sum_{i=1}^{K} \frac{\alpha(i)}{\log_2(i+1)} \tag{4.3}$$

και

$$IDCG \propto K = \sum_{i=1}^{\min(R,K)} \frac{1}{\log_2(i+1)} \tag{4.4}$$

Όπου για τα τρία ερωτήματα αναλόγως:

Για την πρώτη ερώτηση: ως |U| συμβολίζεται ο αριθμός χρηστών, i η θέση κατάταξης του μέρους δημοσίευσης / publication venue, R είναι ο αριθμός των συγγενών / σχετικών μερών δημοσίευσης, και α(i) είναι μια μεταβλητή που ισούται με 1 αν πάλι είναι σχετικό, και 0 αλλιώς.

Για τη δεύτερη ερώτηση : ως |U| συμβολίζεται πάλι ο αριθμός χρηστών, i η θέση κατάταξης του συγγράμματος, R είναι ο αριθμός των συγγενών / σχετικών συγγραμμάτων, και α(i) είναι μια μεταβλητή που ισούται με 1 αν το σύγγραμμα είναι σχετικό, και 0 αλλιώς.

Για την τρίτη ερώτηση: ως |U| συμβολίζεται πάλι ο αριθμός χρηστών, i η θέση κατάταξης ενός συντάκτη, R είναι ο αριθμός των ερευνητών, και α(i) είναι μια μεταβλητή που ισούται με 1 αν είναι παρόμοιοι, και 0 αλλιώς.

Παράδειγμα για κατανόηση των DCG και nDCG

Παράδειγμα: Έστω ότι υπάρχουν 5 έγγραφα [Doc_1, Doc_2, Doc_3, Doc_4, Doc_5]

Doc_1 είναι 100% σχετικό

Doc_2 είναι 70% σχετικό

Doc_3 είναι 95% σχετικό

Doc_4 είναι 20% σχετικό

Doc_5 είναι 100% σχετικό

Το DCG υπολογίζεται:

DCG = SUM(relivencyAt(index) / log2(index + 1)), όπου index 1 -> 5

Doc_1 είναι 100 / log2(2) = 100.00

Doc_2 είναι 70 / log2(3) = 044.17

Doc_3 είναι 95 / log2(4) = 047.50

Doc_4 είναι 20 / log2(5) = 008.61

Doc_5 είναι 100 / log2(6) = 038.69

DCG = 100 + 44.17 + 47.5 + 8.61 + 38.69

DCG = 238.97

και Ideal DCG υπολογίζεται:

IDCG = Doc_1 , Doc_5, Doc_3, Doc_2, Doc_4

Doc_1 είναι 100 / log2(2) = 100.00

Doc_5 είναι 100 / log2(3) = 063.09

Doc_3 είναι 95 / log2(4) = 047.50

Doc_2 είναι 75 / log2(5) = 032.30

Doc_4 είναι 20 / log2(6) = 007.74

IDCG = 100 + 63.09 + 47.5 + 32.30 + 7.74

IDCG = 250.63

Οπότε:

nDCG(5) = DCG / IDCG

= 238.97 / 250.63

= 0.95

4.3 Υπολογισμός μετρικών

Πρώτο βήμα για την εύρεση των παραπάνω τιμών αποτελεί η κλήση της pmf_estimate συνάρτησης όπου υπολογίζονται οι τιμές των m_U και m_V. Τα m_U και m_V αποτελούν τα λανθάνοντα διανύσματα χρηστών και αντικειμένων αντίστοιχα / “latent factors” . Για τον υπολογισμό τους απαραίτητες είναι οι ακόλουθες τιμές: latent_size, lamda_u, lamda_v, pmf_epochs, m_theta και m_gamma.

Το latent_size δείχνει το μέγεθος του διανύσματος.

Τα lamda_u, lamda_v και pmf_epochs είναι παράμετροι που μπορεί να πάρουν διάφορες τιμές και βασικό σκοπό αποτελεί να βρεθεί για ποιες τιμές το σύστημα παράγει τα βέλτιστα αποτελέσματα.

Για τις παραπάνω τιμές αξιοποιήθηκε προηγούμενη έρευνα με διάφορους πειραματισμούς που έχει πραγματοποιηθεί κατά την ανάπτυξη της CATA++ μεθόδου ([7]) που καταλήγει ότι οι βέλτιστες τιμές είναι οι : latent_size =50 , lamda_u = 10, lamda_v=0.1 και pmf_epochs=100. Τα m_theta και m_gamma αποτελούν τα αποτελέσματα της get_z_layer() η οποία ορίζεται στο Attentive_autoencoder.py αρχείο και επιστρέφει τα αποτελέσματα της model.predict().

Έχοντας υπολογίσει πλέον τα λανθάνοντα διανύσματα τα m_U και m_V καλείται η evaluate() συνάρτηση η οποία κάνει τη σύγκριση με τις πραγματικές τιμές των testing δεδομένων και κάνοντας τους υπολογισμούς που γίνονται στους παραπάνω τύπους προκύπτουν οι τελικές τιμές των recall, dcg και ndcg για τα τοπ K στοιχεία, όπου το K παίρνει τις τιμές [10, 50, 100, 150, 200, 250, 300].

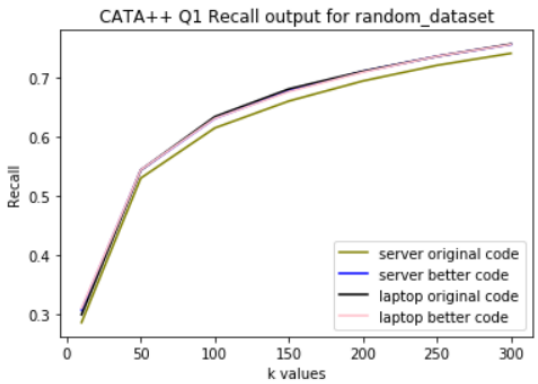
4.4 Recall CATA++

Παρακάτω παρουσιάζονται οι τιμές απόδοσης Recall της CATA++ μεθόδου που έχει εκτελεστεί με το συνδυασμό εναλλακτικών που προαναφέρθηκαν.

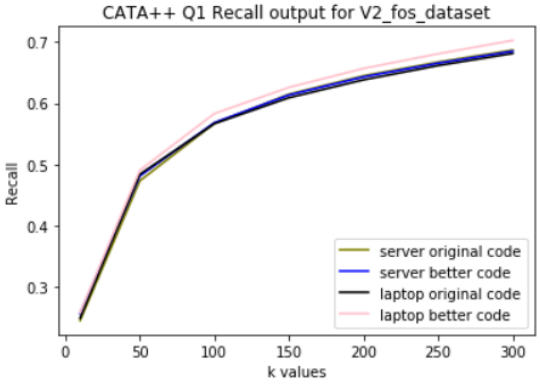
4.4.1 Συστάσεις εκδοτικών χώρων

Στο σχήμα 4.4 απεικονίζεται η απόδοση recall της Cata++ μεθόδου κατά την εκτέλεση του πρώτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων εκδοτικών χώρων.

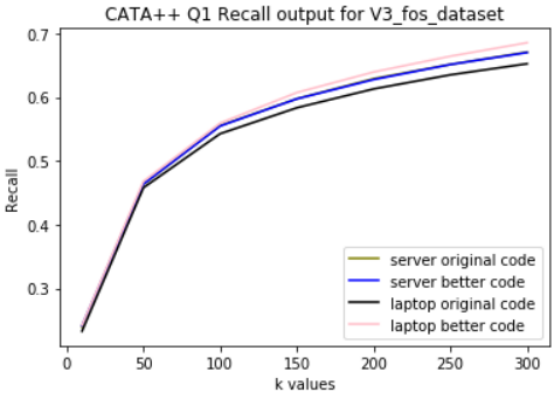
Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και



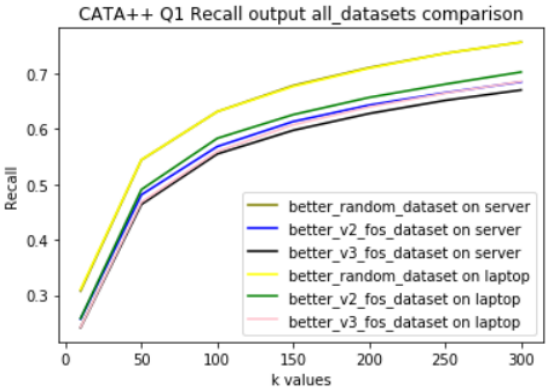
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.4: Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall CATA++

γίνεται σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι με τη χρήση των νέων συναρτήσεων επιτυγχάνεται μια ελαφρώς καλύτερη απόδοση, ενώ το λογισμικό εκτέλεσης δεν κάνει κάποια μεγάλη διαφορά.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Εδώ παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset όπου, όπως και πριν, παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το πρώτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

4.4.2 Συστάσεις επιστημονικών συγγραμμάτων

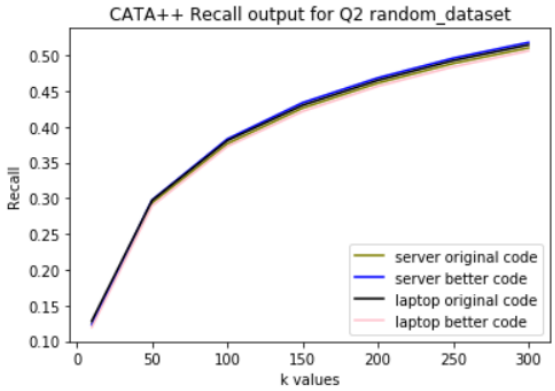
Στο σχήμα 4.5 απεικονίζεται η απόδοση recall της Cata++ μεθόδου κατά την εκτέλεση του δεύτερου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι σχεδόν όλες οι εκτελέσεις βγάζουν παρεμφερή αποτελέσματα και με ελάχιστη διαφορά καλύτερη είναι η εκτέλεση με τη χρήση των νέων συναρτήσεων και στο λογισμικό του server.

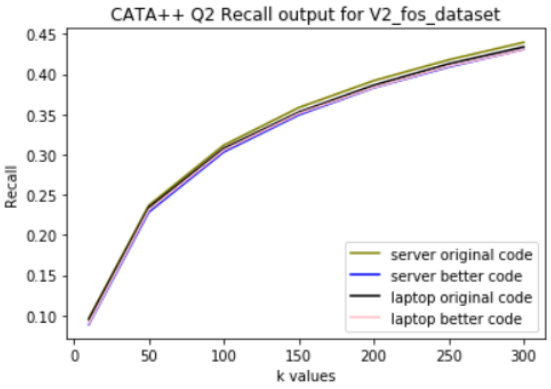
Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις αρχικές συναρτήσεις εκτελεσμένη στον server.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

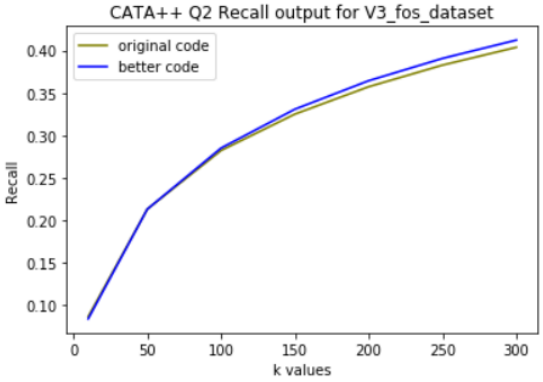
Στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων



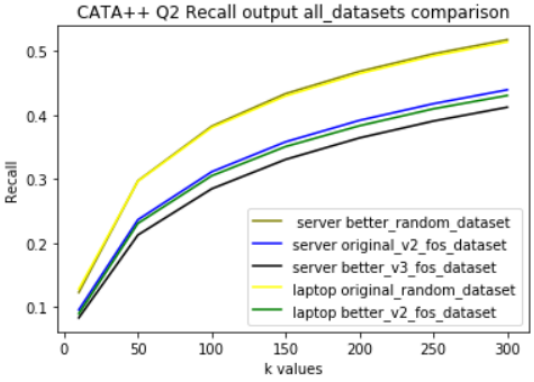
(a) Random dataset.



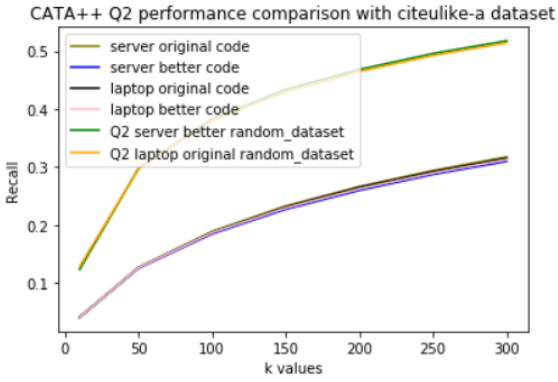
(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets



(e) Comparison between my dataset and citeulike-a

Σχήμα 4.5: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall CATA++

των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το δεύτερο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

Τέλος, στο πέμπτο διάγραμμα παρουσιάζεται η σύγκριση απόδοσης της μεθόδου βάση των δεδομένων που δίνονται ως είσοδος. Πιο συγκεκριμένα, συγκρίνεται η citeulike-a πηγή δεδομένων με το random dataset από την Aminer dblp πηγή δεδομένων καθώς αυτά τα δύο datasets έχουν παρόμοιο μέγεθος. Όπως φαίνεται και στο διάγραμμα, το λογισμικό που εκτελείται ο κώδικας και η επιλογή των συναρτήσεων δεν προκαλούν μεγάλη διαφορά στην απόδοση της μεθόδου για το citeulike-a dataset, αλλά είναι εμφανές ότι η dblp dataset είναι πολύ πιο αποδοτική, κάτι το οποίο σημαίνει ότι τα δεδομένα είναι πιο πλήρη και πιο σωστά επεξεργασμένα.

4.4.3 Συστάσεις συντακτών

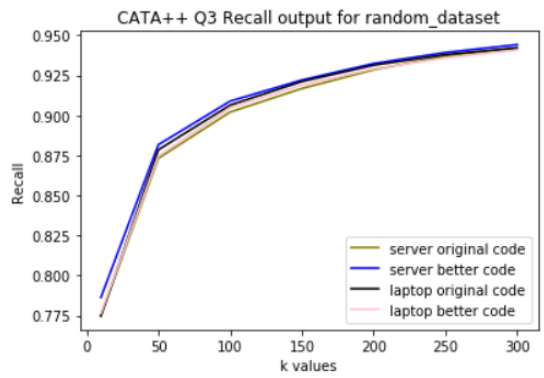
Στο σχήμα 4.6 απεικονίζεται η απόδοση recall της Cata++ μεθόδου κατά την εκτέλεση του τρίτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων συντακτών επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι σχεδόν όλες οι εκτελέσεις βγάζουν παρεμφερή αποτελέσματα και με ελάχιστη διαφορά καλύτερη είναι η εκτέλεση με τη χρήση των νέων συναρτήσεων και έχοντας τρέξει στο λογισμικό του server.

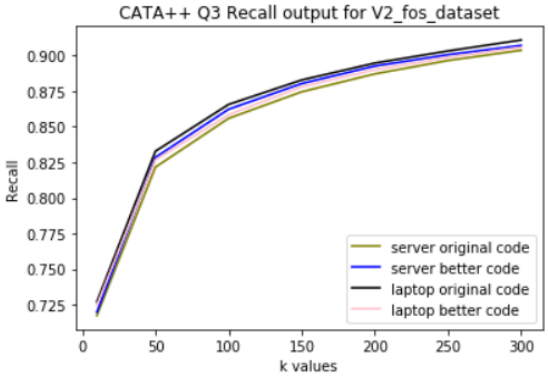
Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις αρχικές συναρτήσεις εκτελεσμένη στο laptop.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται πάλι ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

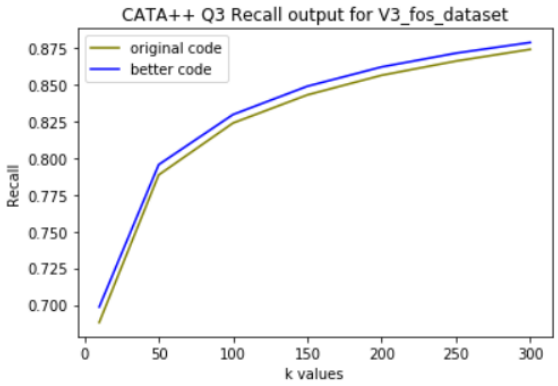
Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το τρίτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του server.



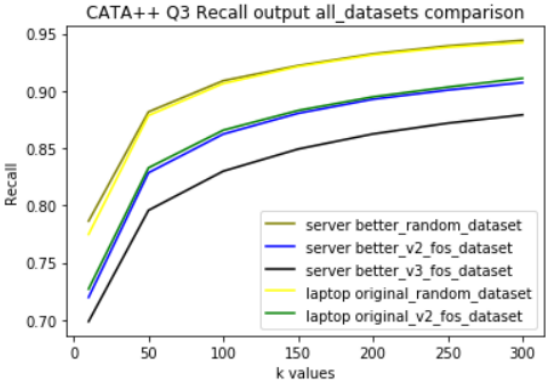
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



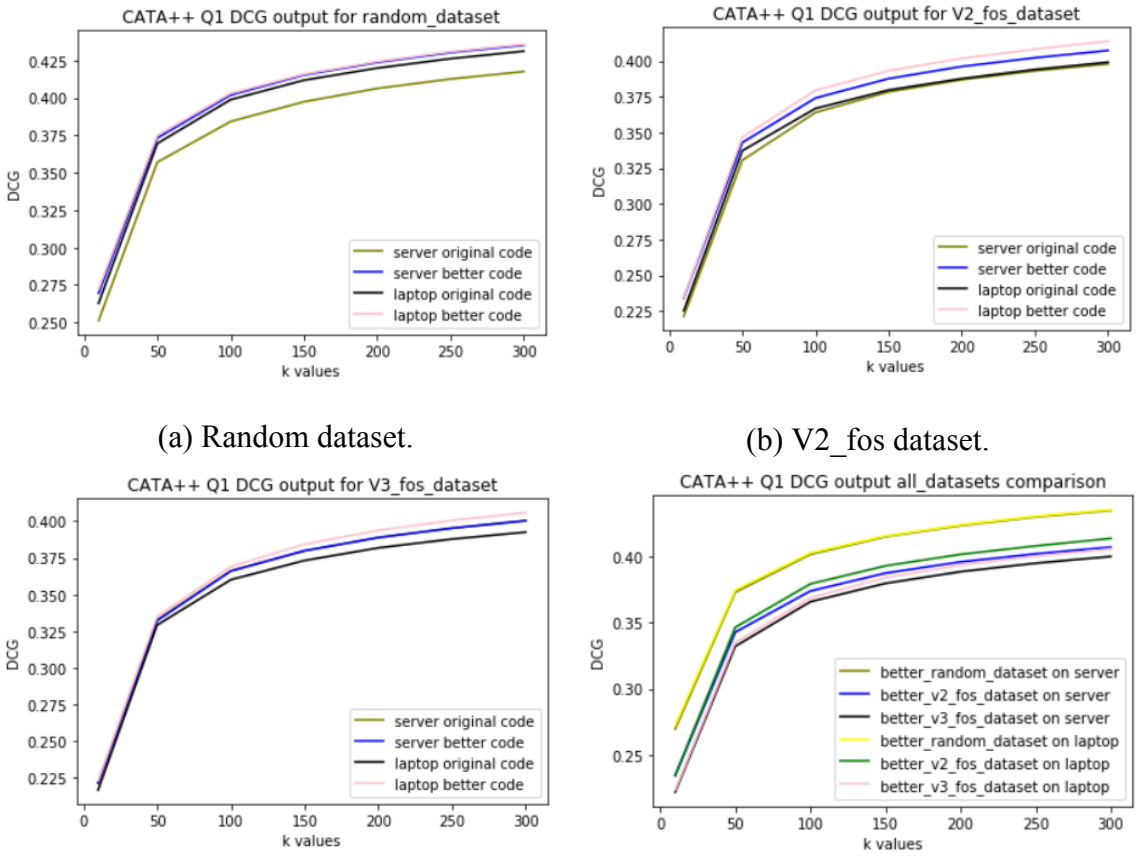
(d) Comparison between all datasets

Σχήμα 4.6: Σύστημα συστάσεων συντακτών - απόδοση Recall CATA++

4.5 DCG CATA++

Παρακάτω παρουσιάζονται οι τιμές απόδοσης DCG της CATA++ μεθόδου που έχει εκτελεστεί με το συνδυασμό εναλλακτικών που προαναφέρθηκαν.

4.5.1 Συστάσεις εκδοτικών χώρων



(a) Random dataset.

(b) V2_fos dataset.

(c) V3_fos dataset.

(d) Comparison between all datasets

Σχήμα 4.7: Σύστημα συστάσεων εκδοτικών χώρων - απόδοση DCG CATA++

Στο σχήμα 4.7 απεικονίζεται η απόδοση dcg της Cata++ μεθόδου κατά την εκτέλεση του πρώτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων εκδοτικών χώρων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι με τη χρήση των νέων συναρτήσεων επιτυγχάνεται μια ελαφρώς βέλτιστη απόδοση, ενώ το λογισμικό εκτέλεσης δεν κάνει κάποια μεγάλη διαφορά.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Εδώ παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset όπου, όπως και πριν, παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των βέλτιστων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το πρώτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

Συνεπώς, παρατηρείται ότι προς το παρόν τόσο με τη χρήση της recall όσο και της dcg μεθόδου προκύπτουν ακριβώς τα ίδια πορίσματα.

4.5.2 Συστάσεις επιστημονικών συγγραμμάτων

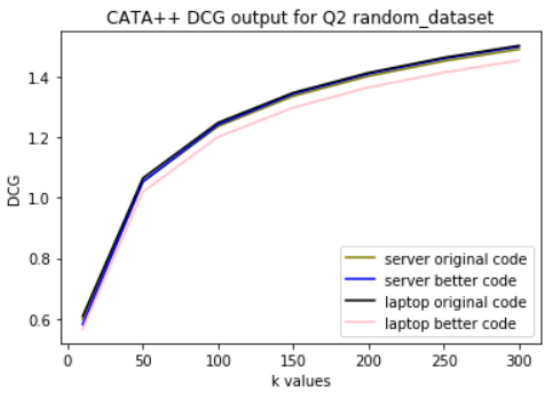
Στο σχήμα 4.8 απεικονίζεται η απόδοση dcg της Cata++ μεθόδου κατά την εκτέλεση του δεύτερου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι σχεδόν όλες οι εκτελέσεις βγάζουν παρεμφερή αποτελέσματα και με ελάχιστη διαφορά καλύτερη είναι η εκτέλεση με τη χρήση των αρχικών συναρτήσεων και έχοντας στο λογισμικό του laptop.

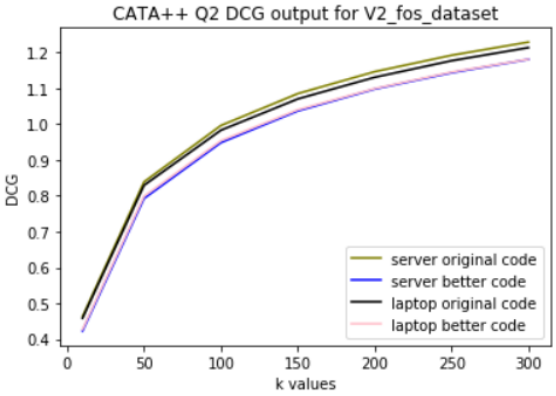
Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις αρχικές συναρτήσεις εκτελεσμένη στον server.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται και πάλι ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

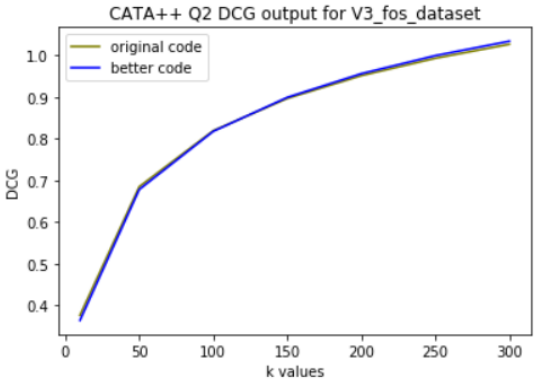
Στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το δεύτερο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές



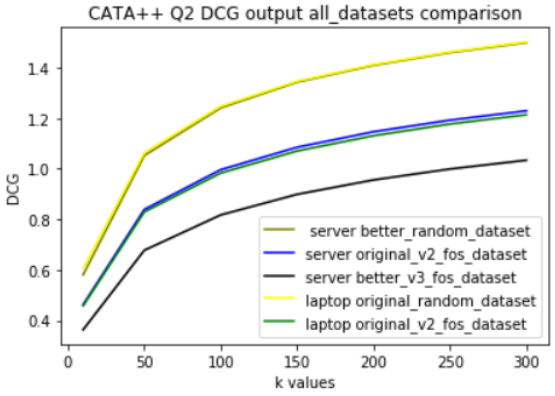
(a) Random dataset.



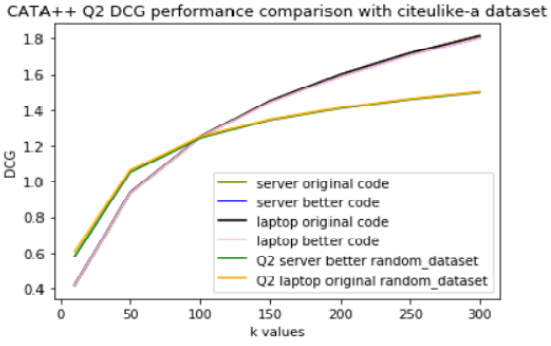
(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets



(e) Comparison between my dataset and citeulike-a

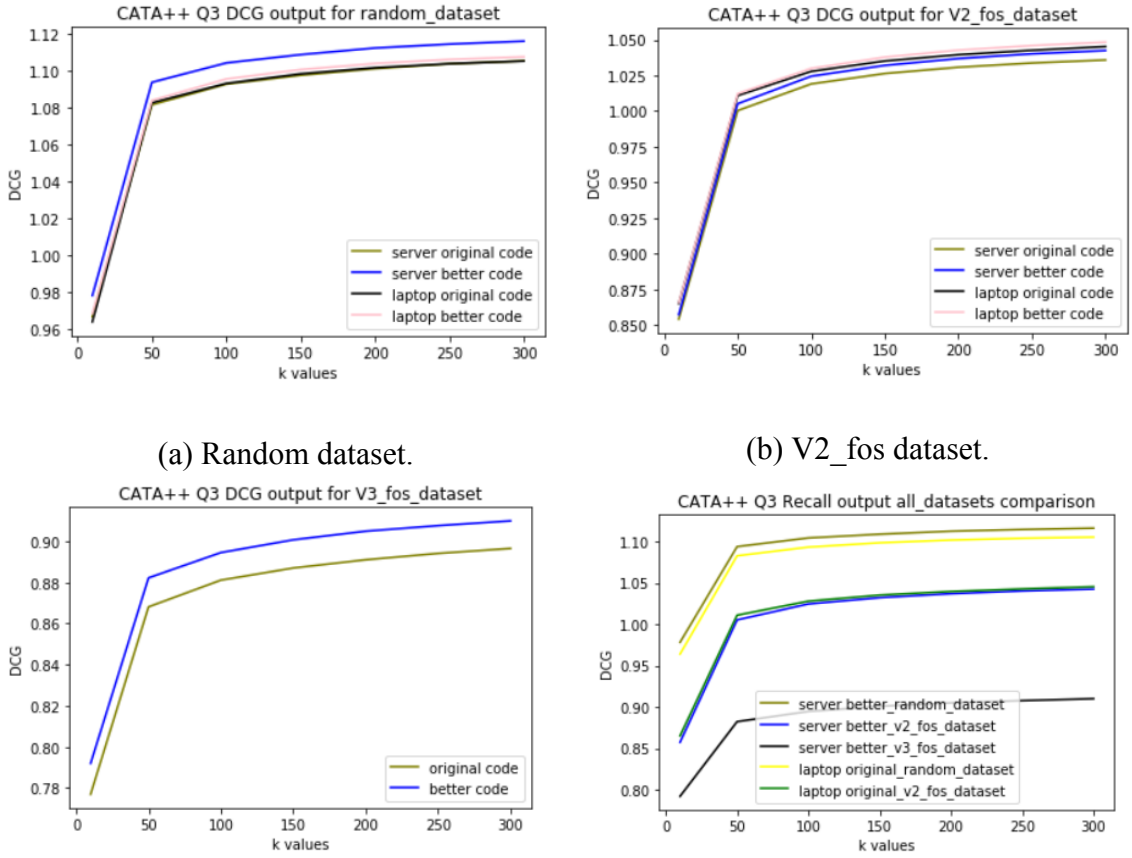
Σχήμα 4.8: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση DCG CATA++

συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

Τέλος, στο πέμπτο διάγραμμα παρουσιάζεται η σύγκριση απόδοσης της μεθόδου βάση των δεδομένων που δίνονται ως είσοδος. Πιο συγκεκριμένα, συγκρίνεται η citeulike-a πηγή δεδομένων με το random dataset από την Aminer dblp πηγή δεδομένων καθώς αυτά τα δύο datasets έχουν παρόμοιο μέγεθος. Όπως φαίνεται και στο διάγραμμα, το λογισμικό που εκτελείται ο κώδικας και η επιλογή των συναρτήσεων δεν προκαλούν μεγάλη διαφορά στην απόδοση της μεθόδου για το citeulike-a dataset, αλλά σε αντίθεση με τα αποτελέσματα που προέκυψαν από τη recall, εδώ φαίνεται ότι το citeulike-a dataset είναι πιο αποδοτικό.

Συνεπώς, παρατηρείται ότι σε αυτή τη περίπτωση η χρήση της recall διαφέρει ελαφρώς με τη χρήση της dcg μεθόδου και κυρίως όσον αφορά τα πορίσματα σύγκρισης των datasets.

4.5.3 Συστάσεις συντακτών



(a) Random dataset.

(b) V2_fos dataset.

(c) V3_fos dataset.

(d) Comparison between all datasets

Σχήμα 4.9: Σύστημα συστάσεων συντακτών - απόδοση DCG CATA++

Στο σχήμα 4.9 απεικονίζεται η απόδοση dcg της Cata++ μεθόδου κατά την εκτέλεση του τρίτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων συντακτών επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι σχεδόν όλες οι εκτελέσεις βγάζουν σχετικά παρεμφερή αποτελέσματα και με μια μικρή διαφορά καλύτερη είναι η εκτέλεση με τη χρήση των νέων συναρτήσεων και έχοντας τρέξει στο λογισμικό του server.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις νέες συναρτήσεις, εκτελεσμένη στο laptop.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται πάλι ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το τρίτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του server.

Συνεπώς, παρατηρείται ότι και εδώ, όπως και στο πρώτο ερώτημα, τόσο με τη χρήση της recall όσο και με της dcg μεθόδου προκύπτουν ακριβώς τα ίδια πορίσματα.

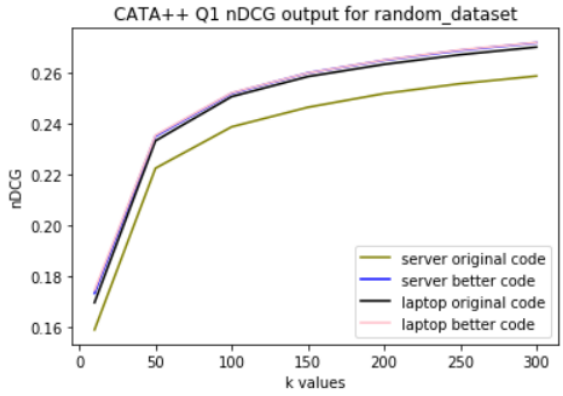
4.6 nDCG CATA++

Παρακάτω παρουσιάζονται οι τιμές απόδοσης nDCG της CATA++ μεθόδου που έχει εκτελεστεί με το συνδυασμό εναλλακτικών που προαναφέρθηκαν.

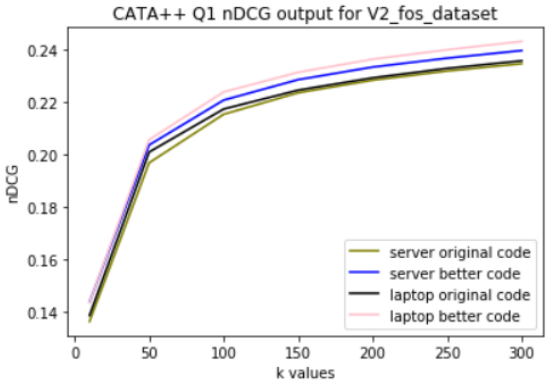
4.6.1 Συστάσεις εκδοτικών χώρων

Στο σχήμα 4.10 απεικονίζεται η απόδοση nDCG της Cata++ μεθόδου κατά την εκτέλεση του πρώτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων εκδοτικών χώρων.

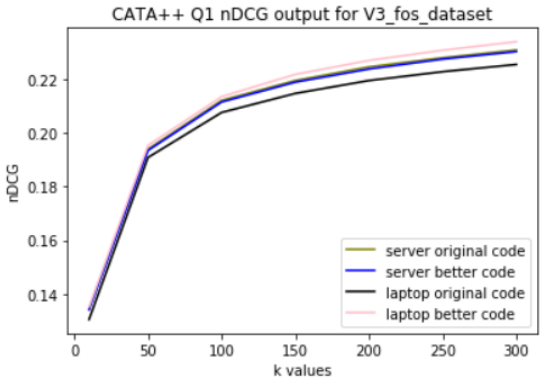
Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων



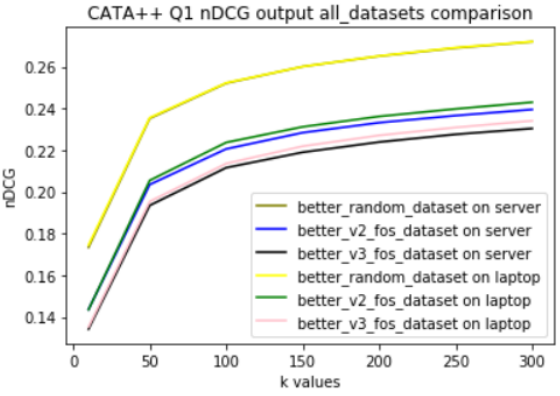
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.10: Σύστημα συστάσεων εκδοτικών χώρων - απόδοση nDCG CATA++

activation function και weight initialization. Αποδεικνύεται ότι με τη χρήση των νέων συναρτήσεων στο λογισμικό του laptop επιτυγχάνεται μια ελαφρώς βέλτιστη απόδοση.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Εδώ παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset όπου, όπως και πριν, παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται στο λογισμικό του laptop και σε συνδυασμό με τη χρήση των νέων συναρτήσεων.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το πρώτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

Συνεπώς, παρατηρείται ότι προς το παρόν τόσο με τη χρήση της recall όσο και με της nDCG μεθόδου προκύπτουν τα ίδια πορίσματα.

4.6.2 Συστάσεις επιστημονικών συγγραμμάτων

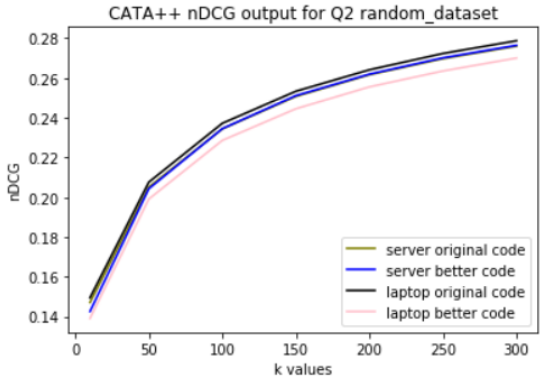
Στο σχήμα 4.11 απεικονίζεται η απόδοση nDCG της Cata++ μεθόδου κατά την εκτέλεση του δεύτερου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι σχεδόν όλες οι εκτελέσεις βγάζουν παρεμφερή αποτελέσματα και με ελάχιστη διαφορά καλύτερη είναι η εκτέλεση με τη χρήση των αρχικών συναρτήσεων και έχοντας στο λογισμικό του laptop.

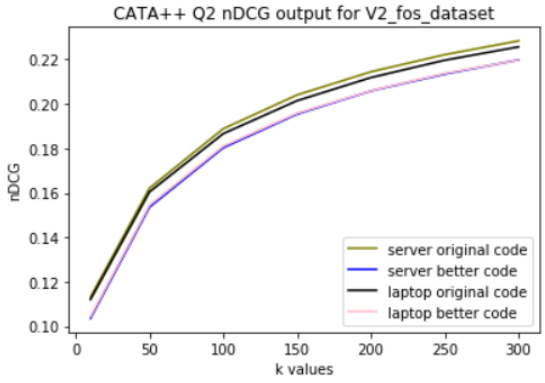
Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις αρχικές συναρτήσεις, εκτελεσμένη στον server.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

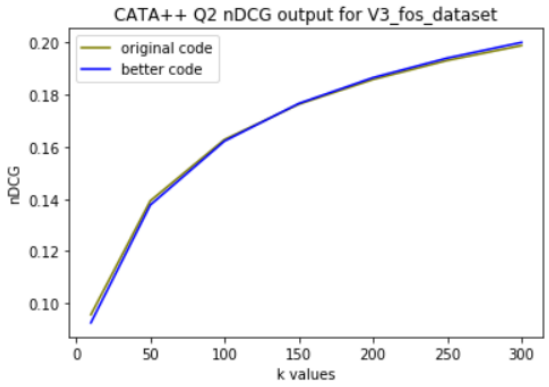
Στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων



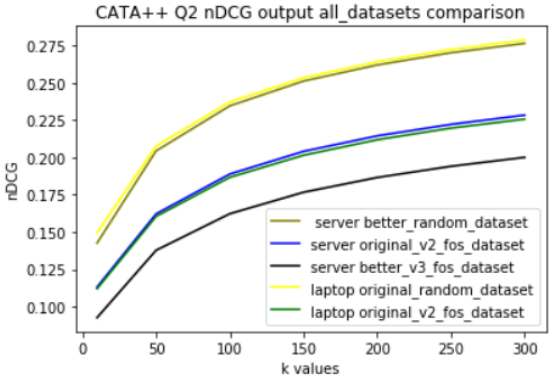
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

(e) Comparison between my dataset and citeulike-a

Σχήμα 4.11: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση nDCG CATA++

των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το δεύτερο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του laptop.

Τέλος, στο πέμπτο διάγραμμα παρουσιάζεται η σύγκριση απόδοσης της μεθόδου βάση των δεδομένων που δίνονται ως είσοδος. Πιο συγκεκριμένα, συγκρίνεται η citeulike-a πηγή δεδομένων με το random dataset από την Aminer dblp πηγή δεδομένων καθώς αυτά τα δύο datasets έχουν παρόμοιο μέγεθος. Όπως φαίνεται και στο διάγραμμα, το λογισμικό που εκτελείται ο κώδικας και η επιλογή των συναρτήσεων δεν προκαλούν μεγάλη διαφορά στην απόδοση της μεθόδου για το citeulike-a dataset, αλλά είναι εμφανές ότι η dblp dataset είναι πολύ πιο αποδοτική, κάτι το οποίο σημαίνει ότι τα δεδομένα είναι πιο πλήρη και πιο σωστά επεξεργασμένα.

Συνεπώς, παρατηρείται ότι τα αποτελέσματα της nDCG ταυτίζονται με αυτά της recall, και όχι με αυτά της dcg.

4.6.3 Συστάσεις συντακτών

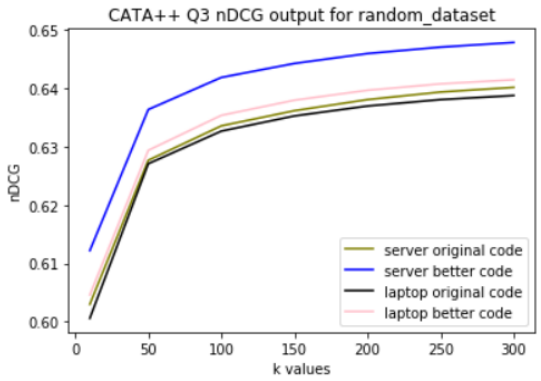
Στο σχήμα 4.12 απεικονίζεται η απόδοση recall της Cata++ μεθόδου κατά την εκτέλεση του τρίτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων συντακτών επιστημονικών συγγραμμάτων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι βέλτιστη απόδοση προκύπτει από την εκτέλεση στον server με τη χρήση των νέων συναρτήσεων.

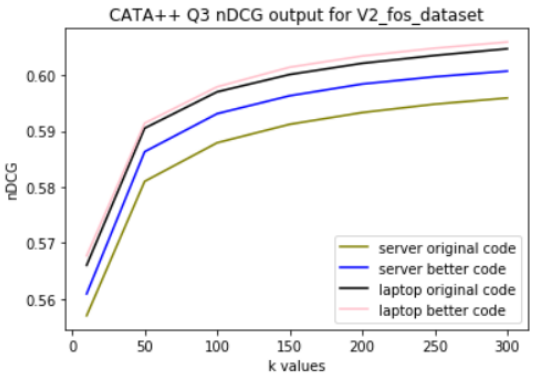
Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι όλες οι εκτελέσεις έχουν σχεδόν ίδια απόδοση, και με πολύ μικρή διαφορά προηγείται αυτή με τις νέες συναρτήσεις εκτελεσμένη στο laptop.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ παρατηρείται πάλι ότι η μέθοδος δεν κατάφερε να εκτελεστεί στο laptop καθώς η μνήμη RAM δεν ήταν επαρκής. Συνεπώς γίνεται σύγκριση των εκτελέσεων μόνο στον server, όπου και φαίνεται ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των νέων συναρτήσεων.

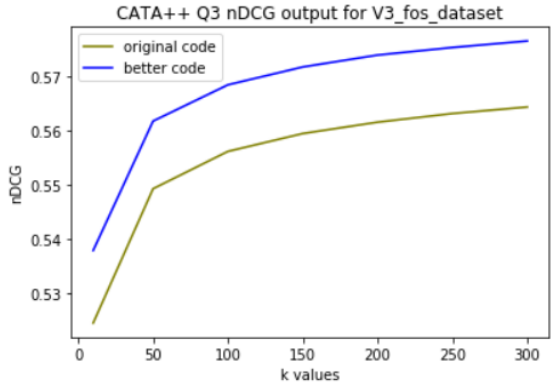
Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το τρίτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις νέες



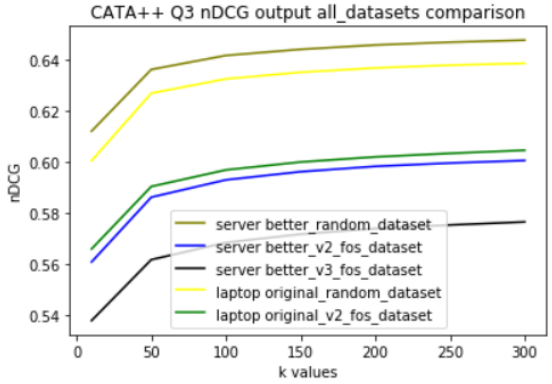
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.12: Σύστημα συστάσεων συντακτών - απόδοση nDCG CATA++

συναρτήσεις, και εκτελώντας τον κώδικα στο λογισμικό του server.

Συνεπώς, παρατηρείται ότι προς τα αποτελέσματα χρήσης της recall είναι αρκετά παρόμοια με αυτά της nDCG μεθόδου.

4.7 Συμπεράσματα σύγκρισης των αποτελεσμάτων CATA++

Βάση των παραπάνω αποτελεσμάτων προκύπτουν τα εξής συμπεράσματα:

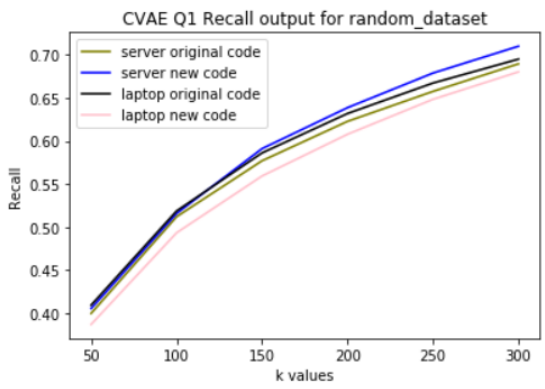
Αρχικά, γίνεται κατανοητό ότι το λογισμικό εκτέλεσης της μεθόδου δεν προκαλεί κάποια αλλαγή στην απόδοση των αποτελεσμάτων, αλλά στη παρούσα διπλωματική δοκιμάζονται δύο διαφορετικά λογισμικά με σκοπό να φανεί ότι για την εκτέλεση του παραπάνω κώδικα με αρχεία μεγέθους όσα αυτά που σχηματίστηκαν χρειάζεται ένας υπολογιστής με σχετικά μεγάλη RAM ώστε να τρέξει με διάφορους συνδυασμούς. Για αυτό το λόγο παρατηρήθηκε ότι ορισμένοι συνδυασμοί δεν ήταν δυνατό να εκτελεστούν στο λογισμικό του laptop που συγκριτικά με το server είχε μικρότερη RAM.

Επιπρόσθετα, παρατηρείται ότι η χρήση των νέων συναρτήσεων activation function και weight initialization επιφέρουν καλύτερα αποτελέσματα στην πλειονότητα των συνδυαστικών εκτελέσεων. Συνεπώς, μπορεί να θεωρηθεί ότι η επιλογή αυτών όντως βελτιώνει την απόδοση της CATA++ μεθόδου.

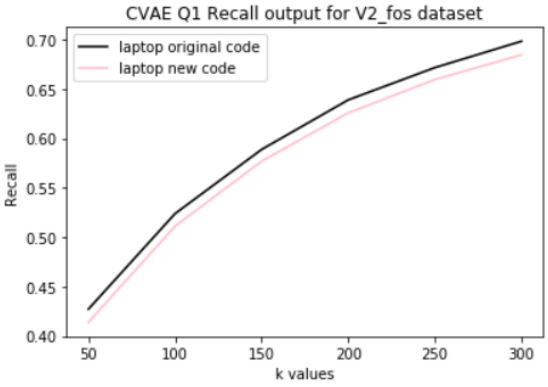
Τέλος, όσον αφορά στην επιλογή της μεθόδου αξιολόγησης των μεθόδων ανάπτυξης των συστημάτων συστάσεων παρατηρείται ότι οι μέθοδοι recall και nDCG έχουν παρόμοια αποτελέσματα, ενώ η μέθοδος DCG διαφέρει. Γι' αυτό το λόγο, για τη σύγκριση της CATA++ μεθόδου με τις υπόλοιπες αποφασίστηκε να χρησιμοποιηθεί η recall μέθοδος αξιολόγησης των μεθόδων ως η πιο απλή και αντιπροσωπευτική.

4.8 Recall CVAE

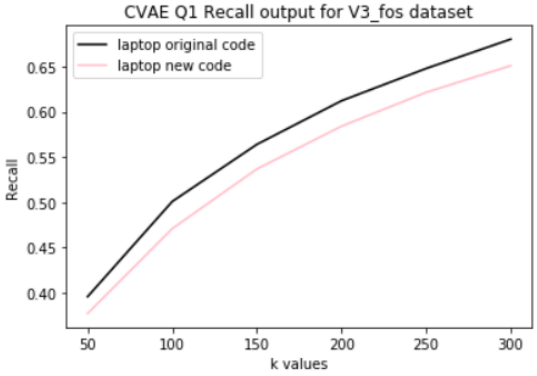
Παρακάτω παρουσιάζονται οι τιμές απόδοσης Recall της CVAE μεθόδου που έχει εκτελεστεί με το συνδυασμό εναλλακτικών που προαναφέρθηκαν.



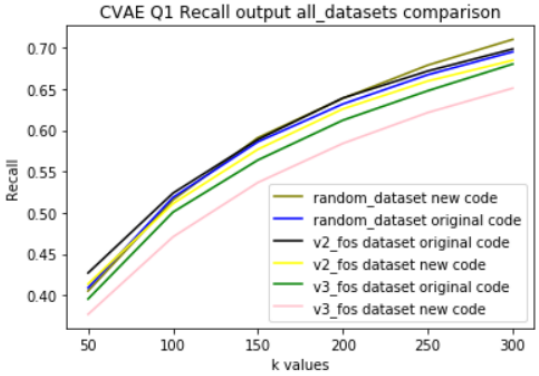
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.13: Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall CVAE

4.8.1 Συστάσεις εκδοτικών χώρων

Στο σχήμα 4.13 απεικονίζεται η απόδοση recall της CVAE μεθόδου κατά την εκτέλεση του πρώτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων εκδοτικών χώρων.

Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι με τη χρήση των νέων συναρτήσεων δεν επιτυγχάνεται καθόλου βελτίωση των αποτελεσμάτων, ενώ το λογισμικό εκτέλεσης δεν κάνει κάποια μεγάλη διαφορά. Βασίζοντας και στις προηγούμενες εκτελέσεις, συνεπάγεται πλέον ότι το λογισμικό στο οποίο εκτελείται ο κώδικας δεν επηρεάζει την απόδοση που επιτυγχάνεται, παρά μόνο το χρόνο εκτέλεσής του. Γι' αυτό το λόγο πλέον οι εκτελέσεις πραγματοποιούνται σε ένα από τα δύο λογισμικά.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Εδώ παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται με τη χρήση των αρχικών συναρτήσεων.

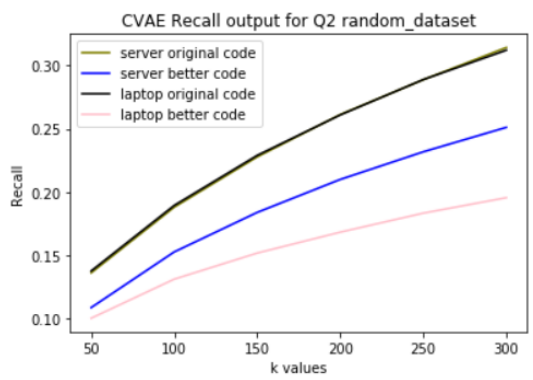
Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset όπου, όπως και πριν, παρατηρείται ότι η βέλτιστη εκτέλεση πραγματοποιείται με τη χρήση των αρχικών συναρτήσεων.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το πρώτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές συναρτήσεις, και παρατηρείται ότι οι νέες συναρτήσεις δεν βελτιστοποιούν ποτέ την απόδοση. Ταυτόχρονα, παρατηρείται ότι η CVAE μέθοδος αποδίδει καλύτερα όταν δίνεται ως είσοδος μικρότερο μέγεθος αρχείων.

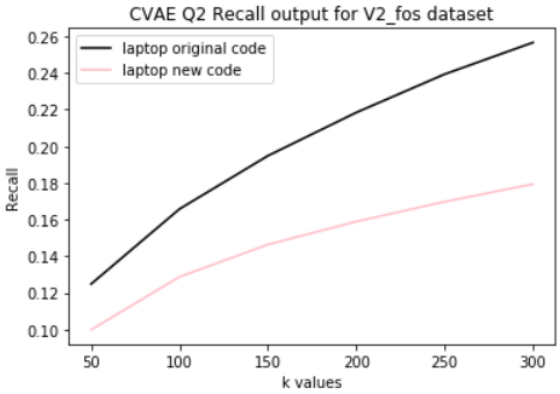
4.8.2 Συστάσεις επιστημονικών συγγραμμάτων

Στο σχήμα 4.14 απεικονίζεται η απόδοση recall της CVAE μεθόδου κατά την εκτέλεση του δεύτερου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων επιστημονικών συγγραμμάτων.

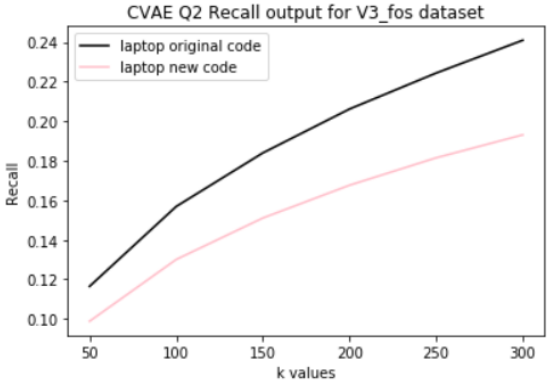
Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση του λογισμικού που χρησιμοποιήθηκε και βάση των συναρτήσεων activation function και weight initialization. Αποδεικνύεται ότι καλύτερη είναι η εκτέλεση με τη χρήση των αρχικών συναρτήσεων και στα δυο λογισμικά, ενώ η χρήση των νέων



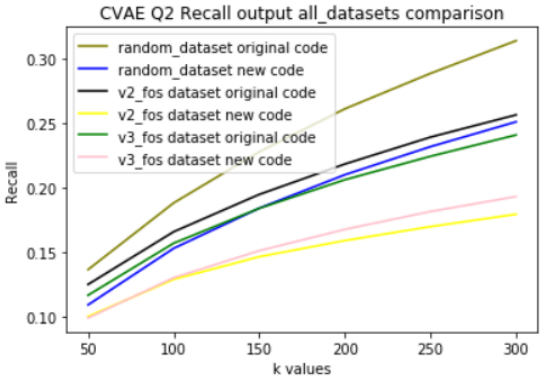
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.14: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall CVAE

συναρτήσεων ρίχνει πολύ την απόδοση. Συνεπώς, όπως και πριν συνεχίζονται οι εκτελέσεις μόνο σε ένα λογισμικό.

Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset. Και εδώ παρατηρείται ότι υπάρχει μεγάλη διαφορά της απόδοσης, καθώς οι νέες συναρτήσεις χειροτερεύουν πολύ τα αποτελέσματα.

Στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Εδώ φαίνεται και πάλι ότι η βέλτιστη απόδοση επιτυγχάνεται με τη χρήση των αρχικών συναρτήσεων.

Στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται ρητά ότι η βέλτιστη απόδοση για το δεύτερο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές συναρτήσεις, ενώ ταυτόχρονα αποδεικνύεται και πάλι ότι οι νέες συναρτήσεις δεν παράγουν καλύτερες προτάσεις και ότι η μέθοδος είναι πιο αποτελεσματική όταν δίνονται ως είσοδος πιο μικρά δεδομένα αρχείων.

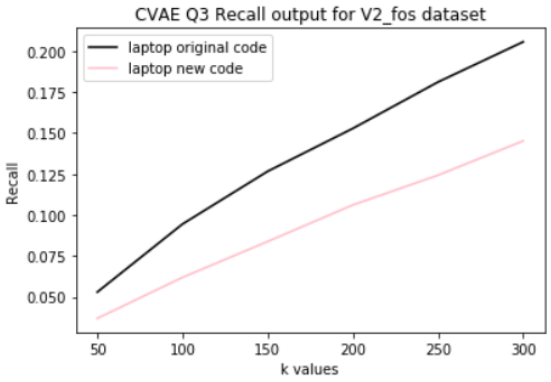
4.8.3 Συστάσεις συντακτών

Στο σχήμα 4.15 απεικονίζεται η απόδοση recall της CVAE μεθόδου κατά την εκτέλεση του τρίτου ερωτήματος, δηλαδή την ανάπτυξη συστήματος συστάσεων συντακτών επιστημονικών συγγραμμάτων.

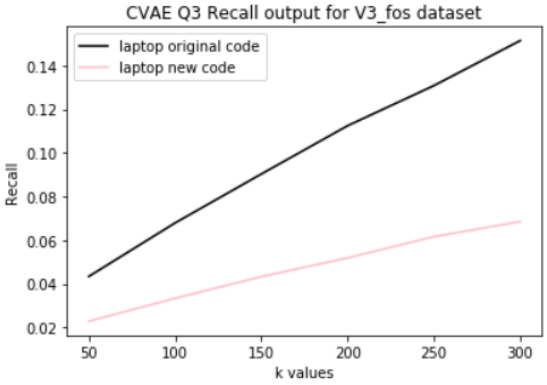
Στο πρώτο γράφημα φαίνεται η απόδοση όταν δίνεται ως είσοδος το random dataset και γίνεται πάλι σύγκριση βάση των συναρτήσεων activation function και weight initialization. Αντίστοιχα στο δεύτερο διάγραμμα φαίνεται η εκτέλεση δίνοντας ως είσοδο το V2_fos dataset, και στο τρίτο διάγραμμα παρουσιάζεται η εκτέλεση με είσοδο το V3_fos dataset. Και στα τρία διαγράμματα παρατηρείται ότι η απόδοση της CVAE μεθόδου είναι σαφώς καλύτερη όταν χρησιμοποιούνται οι αρχικές συναρτήσεις, και ότι το λογισμικό δεν επηρεάζει την απόδοση του συστήματος.

Τέλος, στο τέταρτο διάγραμμα φαίνεται η σύγκριση μεταξύ των καλύτερων συνδυασμών όλων των παραπάνω. Εδώ πλέον φαίνεται πάλι ότι η βέλτιστη απόδοση για το τρίτο ερώτημα επιτυγχάνεται με το συνδυασμό του random dataset ως είσοδο, χρησιμοποιώντας τις αρχικές συναρτήσεις, και ότι η βέλτιστη απόδοση επιτυγχάνεται για μικρότερα αρχεία εισόδου.

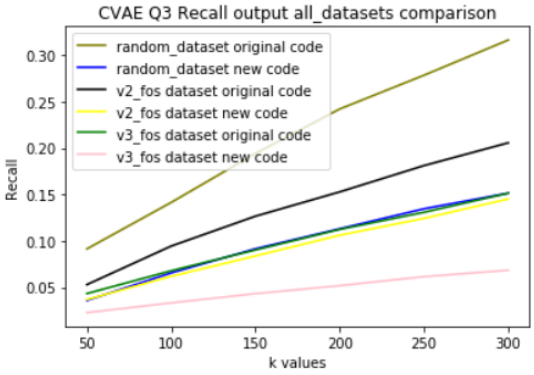
(a) Random dataset.



(b) V2_fos dataset.



(c) V3_fos dataset.



(d) Comparison between all datasets

Σχήμα 4.15: Σύστημα συστάσεων συντακτών - απόδοση Recall CVAE

4.9 Συμπεράσματα σύγκρισης των αποτελεσμάτων CVAE

Βάση των παραπάνω αποτελεσμάτων προκύπτουν τα εξής συμπεράσματα:

Αρχικά, γίνεται κατανοητό ότι, όπως και πριν, το λογισμικό εκτέλεσης της μεθόδου δεν προκαλεί κάποια αλλαγή στην απόδοση των αποτελεσμάτων, παρά μόνο στο χρόνο εκτέλεσης των προγραμμάτων, όπου ο server είναι ταχύτερος, και ως προς το μέγεθος των αρχείων που μπορεί να επεξεργαστεί ένας αλγόριθμος, καθώς ο server έχει μεγαλύτερη RAM.

Επιπρόσθετα, παρατηρείται ότι η χρήση των νέων συναρτήσεων activation function και weight initialization δεν βελτιστοποιούν τις εκτελέσεις της CVAE μεθόδου, σε αντίθεση με την CATA++ μέθοδο. Συνεπώς, μπορεί να θεωρηθεί ότι η επιλογή των αρχικών συναρτήσεων αποτελεί τη βέλτιστη επιλογή για τη CVAE μέθοδο.

Τέλος, παρατηρείται ότι η CVAE μέθοδος παράγει βέλτιστα αποτελέσματα για μικρότερου μεγέθους αρχεία, και όσο αυξάνεται το μέγεθός τους, τόσο χειροτερεύουν και οι συστάσεις. Συνεπώς, συμπεραίνεται ότι σίγουρα η CATA++ μέθοδος είναι πιο αποδοτική σε αρχεία μεγαλύτερου μεγέθους και μένει να συγκριθεί η απόδοση των δυο μεθόδων για τα πιο μικρά αρχεία εισόδου.

4.10 Recall CML

Παρακάτω παρουσιάζονται οι τιμές απόδοσης Recall της CML μεθόδου. Η CML μέθοδος ανάπτυξης συστημάτων συστάσεων αποδεικνύεται ότι απαιτεί υπερβολικά πολύ ισχυρή RAM υπολογιστή και πολύ χρόνο για να εκτελέσει τον κώδικα. Αυτό συνεπάγεται ότι η μέθοδος αυτή δεν συμφέρει για ένα ρεαλιστικό σενάριο ανάπτυξης ενός συστήματος συστάσεων. Συγκεκριμένα, δεν επιτεύχθηκε η εκτέλεση του κώδικα της CML ούτε στο προσωπικό μου laptop αλλά ούτε και στον server για τα V2_fos και V3_fos datasets τα οποία είναι πιο μεγάλα. Ο UTH server διαθέτει 16GB RAM και NVIDIA GPU η οποία είναι εκμεταλλεύσιμη από το tensorflow για να εκπαιδεύει πιο γρήγορα τα Νευρωνικά Δίκτυα. Κατά την εκτέλεση της CML μεθόδου, κατανάλωνε αμέσως τη διαθέσιμη μνήμη της κάρτας GPU (2GB) και παρόλο που υπήρχε διαθέσιμη RAM το πρόγραμμα τερμάτιζε με OOM error (Out Of Memory). Παρόλα αυτά, για να μπορέσει να υπάρξει μια μέτρηση της αποτελεσματικότητας της CML, εκτελέστηκε το random dataset το οποίο είναι το μικρότερο εκ των τριών.

Η εκτέλεση της CATA++ μεθόδου πραγματοποιήθηκε στο προσωπικό μου laptop, ενώ η εκτέλεση της CML μεθόδου πραγματοποιήθηκε σε laptop με τα εξής χαρακτηριστικά:

- Επεξεργαστής AMD Ryzen 3 CPU @ 3.60GHz
- Εγκατεστημένη RAM 16,00 GB
- Τύπος συστήματος Λειτουργικό σύστημα 64 bit, επεξεργαστής τεχνολογίας x64-based processor
- Χωρίς αξιοποίηση κάρτας γραφικών

Πιο συγκεκριμένα, βασισμένοι στα προηγούμενα συμπεράσματα, η σύγκριση γίνεται με τη χρήση της recall ως μέθοδος αξιολόγησης, και έχοντας ορίσει τη μεταβλητή K=50.

Τα αποτελέσματα των εκτελέσεων με τα αρχεία εισόδου του Random Dataset φαίνονται στον παρακάτω πίνακα:

Σύσταση εκδοτικών χώρων: 0,66263
Σύσταση επιστημονικών συγγραμμάτων: 0,6871
Σύσταση συντακτών: 0,81264

Επίσης, παρατηρώντας ότι για σύσταση επιστημονικών δημοσιευμάτων επιτυγχάνεται βέλτιστη απόδοση της CML μεθόδου, δημιουργήθηκαν δύο νέα μικρά datasets με τα εξής χαρακτηριστικά:

USERS: 2918	ITEMS: 1641	TAGS: 4788
USERS: 3458	ITEMS: 1759	TAGS: 5236

Αξιοποιώντας τα δυο νέα datasets συγκρίνεται η απόδοση της CML μεθόδου για το ίδιο ερώτημα αλλά δίνοντας ως είσοδο τα δυο νέα datasets και προκύπτουν τα παρακάτω αποτελέσματα:

Σύσταση επιστημονικών συγγραμμάτων - Small dataset: 0,68102
Σύσταση επιστημονικών συγγραμμάτων - Small2 dataset: 0,64509
Σύσταση επιστημονικών συγγραμμάτων - Random Dataset: 0,6871

4.11 Συμπεράσματα σύγκρισης των αποτελεσμάτων CML

Αρχικά,είναι προφανές ότι τα συμπεράσματα δεν είναι το ίδιο αξιόπιστα με αυτά που προηγήθηκαν καθώς δεν είναι δυνατό να εκτελεστούν όλοι οι προηγούμενοι συνδυασμοί.

Παίρνοντας ως δεδομένα τα παραπάνω αποτελέσματα προκύπτουν τα εξής συμπεράσματα:

Αρχικά, η μέθοδος αυτή είναι πολύ απαιτητική ως προς τη RAM που απαιτείται για να εκτελεστεί ο κώδικας, κάτι το οποίο δεν είναι αποδοτικό. Ταυτόχρονα, ο απαραίτητος χρόνος εκτέλεσης του κώδικα είναι πολύς, με αποτέλεσμα να μην υπερτερεί ούτε ως προς τη χρονική απόδοσή του.

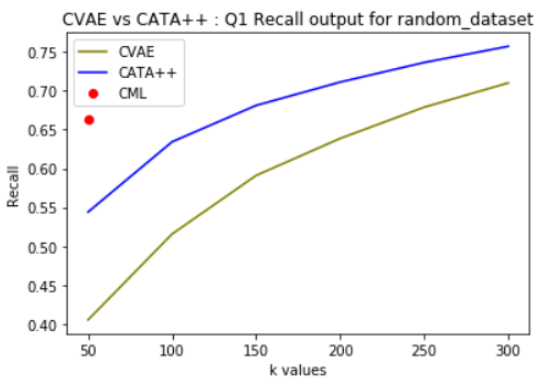
Τέλος, παρατηρείται ότι η CML μέθοδος παράγει αποτελέσματα υψηλής απόδοσης για μικρά αρχεία εισόδου, τα οποία ανεξαρτήτως μεγέθους έχουν παρόμοια αποτελέσματα εφόσον καταφέρουν να εκτελεστούν, κάτι το οποίο μπορεί να συνεπάγεται ότι αν βελτιωθεί ο τρόπος εκτέλεσής της, ίσως να είναι πιο αποδοτική για μεγαλύτερα δεδομένα εισόδου.

4.12 Σύγκριση απόδοσης CATA++, CVAE, CML

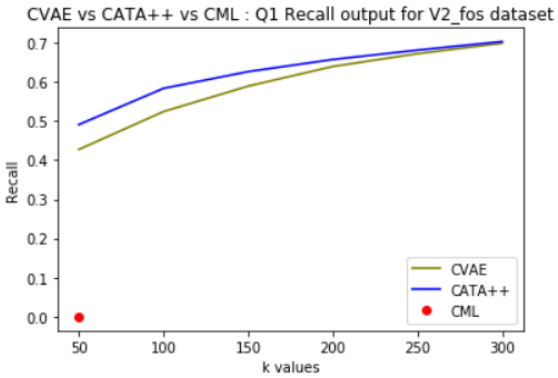
Στα παρακάτω σχήματα φαίνεται η σύγκριση των τιμών απόδοσης μεταξύ των μεθόδων CATA++, CVAE και CML μέσω της recall μεθόδου αξιολόγησης. Σε κάθε περίπτωση, η σύγκριση έχει γίνει με τη βέλτιστη τιμή που έχει επιτύχει με οποιονδήποτε συνδυασμό η κάθε μέθοδος.

Αναλυτικότερα:

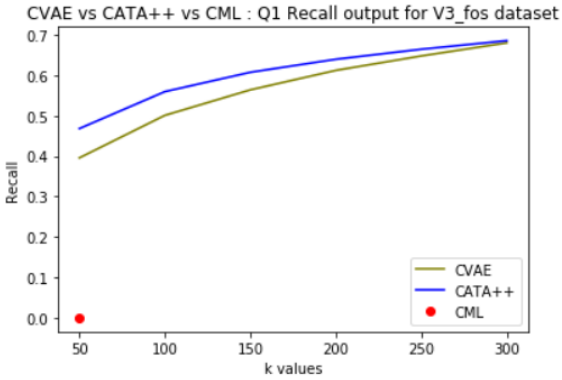
Στο σχήμα 4.16 απεικονίζεται η απόδοση όλων των μεθόδων για το πρώτο ερώτημα, δηλαδή για συστάσεις εκδοτικών χώρων. Συμπεραίνουμε ότι για το πρώτο ερώτημα, για δεδομένα μικρού μεγέθους και με K=50 φαίνεται ότι η CML μέθοδος είναι η πιο αποδοτική, αλλά γενικά η πιο αποδοτική μέθοδος είναι η CATA++ και στη συνέχεια η CVAE. Στο σχήμα 4.17 απεικονίζεται η απόδοση όλων των μεθόδων για το δεύτερο ερώτημα, δηλαδή για συστάσεις επιστημονικών συγγραμμάτων. Συμπεραίνουμε ότι, όπως και στο πρώτο, έτσι και στο δεύτερο ερώτημα, για δεδομένα μικρού μεγέθους και με K=50 φαίνεται ότι η CML μέθοδος είναι η πιο αποδοτική, αλλά και πάλι, η γενικά πιο αποδοτική μέθοδος είναι η CATA++ και στη συνέχεια η CVAE.



(a) Random dataset.

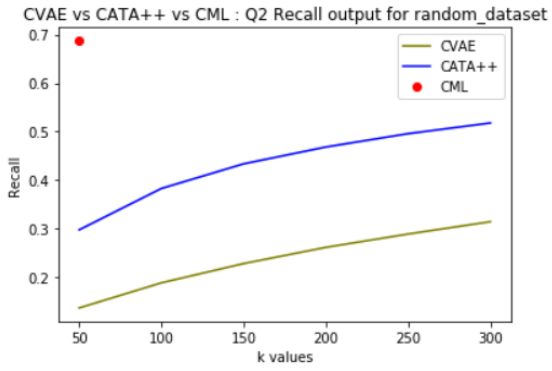


(a) V2_fos dataset.

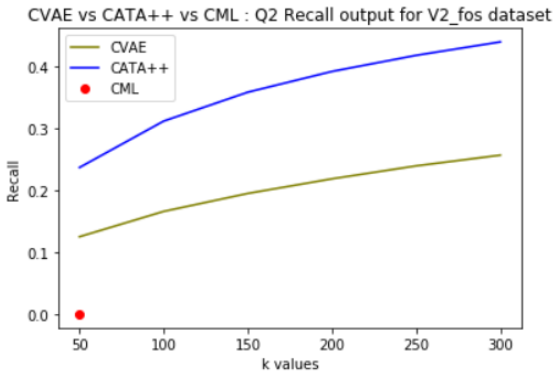


(a) V3_fos dataset.

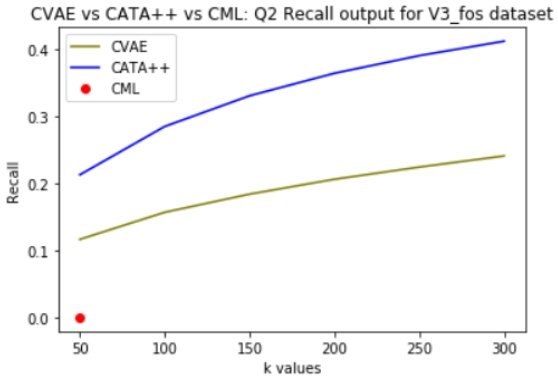
Σχήμα 4.16: Σύστημα συστάσεων εκδοτικών χώρων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML



(a) Random dataset.



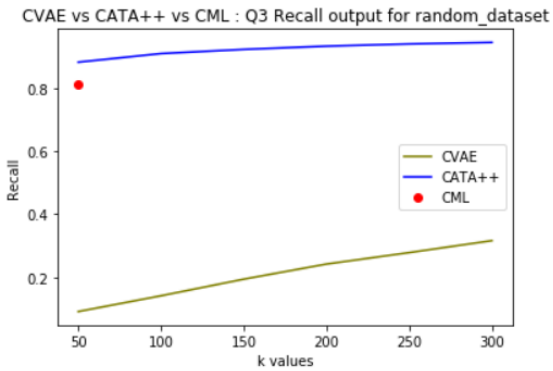
(a) V2_fos Dataset.



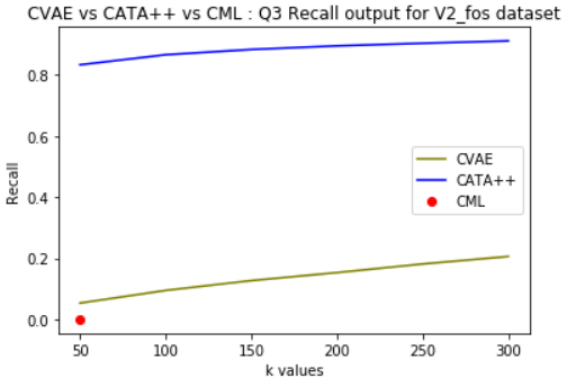
(a) V3_fos Datase.

Σχήμα 4.17: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML

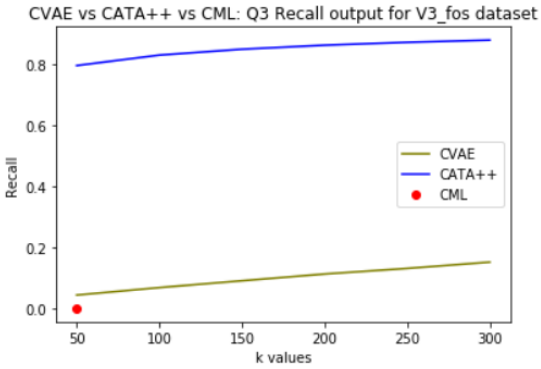
Στο σχήμα 4.18 απεικονίζεται η απόδοση όλων των μεθόδων για το τρίτο ερώτημα, δηλαδή για συστάσεις συντακτών. Συμπεραίνουμε ότι στο τρίτο ερώτημα η CATA++ μέθοδος είναι βέλτιστη σε κάθε περίπτωση, κάτι το οποίο μπορεί να προκαλείται από το γεγονός ότι κατά την εκτέλεση του τρίτου ερωτήματος ο αριθμός των δεδομένων που δίνονται ως είσοδος είναι μεγαλύτερος σε σύγκριση με τα δύο προηγούμενα αποτελέσματα. Συνεπώς, επιβεβαιώνεται ότι η CML μέθοδος είναι καλύτερη μόνο όταν δίνεται ως είσοδος αρχεία πολύ μικρού μεγέθους. Αλλιώς, η βέλτιστη μέθοδος είναι η CATA++, και στη συνέχεια ακολουθεί η CVAE.



(a) Random dataset.



(a) V2_fos Dataset.

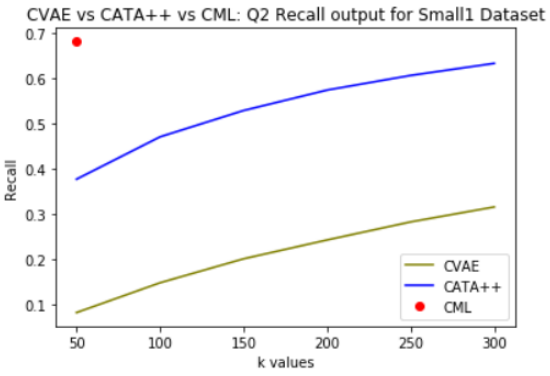


(a) V3_fos Datase.

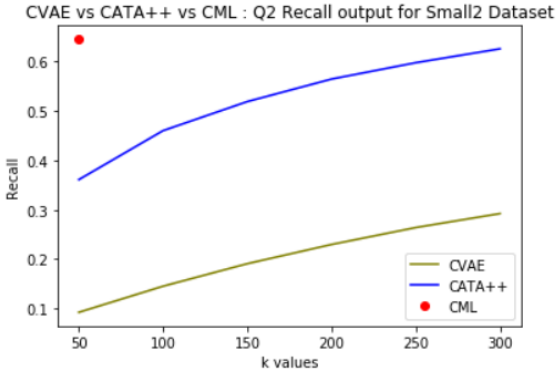
Σχήμα 4.18: Σύστημα συστάσεων συντακτών - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML

Στο σχήμα 4.19 απεικονίζεται η απόδοση όλων των μεθόδων για το δεύτερο ερώτημα, δηλαδή για συστάσεις επιστημονικών συγγραμμάτων χρησιμοποιώντας τα δυο νέα δεδομένα εισόδου, τα Small1 και Small2 dataset. Συμπεραίνουμε ότι όπως ήταν αναμενόμενο, η CML μέθοδος είναι πιο αποτελεσματική από τις άλλες δύο όταν δίνεται ως είσοδος πολύ μικρό

μέγεθος αρχείων εισόδου, ενώ στη συνέχεια ακολουθεί η CATA++ μέθοδος, και τέλος η CVAE,



(a) Small1 dataset.



(a) Small2 dataset.

Σχήμα 4.19: Σύστημα συστάσεων επιστημονικών συγγραμμάτων - απόδοση Recall - Σύγκριση απόδοσης CATA++, CVAE, CML - Small1 & Small2 Dtasets

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικά Σχέδια

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία αναπτύχθηκαν τέσσερα υβριδικά συστήματα συστάσεων που βασίζονται στα νευρωνικά δίκτυα με στόχο τη σύσταση συστάσεων επιστημονικών συγγραμμάτων. Ως δεδομένα χρησιμοποιήθηκαν αυτά από τη Aminer - dblp βάση δεδομένων από τον παγκόσμιο ιστό και επεξεργάστηκαν με τέτοιο τρόπο ώστε να λαμβάνονται μόνο τα πιο πρόσφατα συγγράμματα και να απαντούν στις ακόλουθες τρεις ερωτήσεις:

1. Πρόταση σχετικά με εκδοτικούς χώρους στους συντάκτες επιστημονικών εγγράφων για την υποβολή συγκεκριμένων εγγράφων που έχουν ήδη γράψει ή που θα συντάξουν στο μέλλον, σύμφωνα με τα ερευνητικά τους ενδιαφέροντα.

2. Πρόταση επιστημονικών άρθρων για ανάγνωση και μελέτη στους συντάκτες επιστημονικών περιοδικών και βιβλίων.

3. Πρόταση ερευνητών - συντακτών επιστημονικών άρθρων σε άλλους ερευνητές - συντάκτες επιστημονικών άρθρων για πιθανή συνεργασία.

Έχοντας υλοποιήσει τους αλγορίθμους ανάπτυξης των μεθόδων και έχοντας προετοιμάσει τα δεδομένα, ακολουθεί η εκτέλεση του κώδικα και η αξιολόγηση της αποδοτικότητάς του. Αρχικά, αξιολογείται η CATA++ μέθοδος μέσω των recall, DCG και nDCG μεθόδων, και εδώ συμπεραίνονται τα εξής:

Αρχικά, οι μέθοδοι recall και nDCG είναι σχεδόν ισάξιες όσον αφορά την αξιολόγηση της αποδοτικότητας της μεθόδου, ενώ η DCG μέθοδος δεν είναι το ίδιο αποτελεσματική.

Επίσης, παρατηρείται ότι το λογισμικό εκτέλεσης του κώδικα δεν επηρεάζει το αποτέλεσμα, αλλά η χρήση ενός λογισμικού με μεγαλύτερη RAM προσφέρει μεγαλύτερη ταχύτητα εκτέλεσης του κώδικα, αλλά και τη δυνατότητα εκτέλεσης του δίνοντας ως είσοδο αρχεία μεγαλύτερου μεγέθους.

Ταυτόχρονα, παρατηρείται η χρήση των νέων συναρτήσεων activation function και weight initialization επιφέρουν καλύτερα αποτελέσματα στην πλειονότητα των συνδυαστικών εκτελέσεων, και μπορεί να θεωρηθεί ότι η επιλογή αυτών βελτιώνει την απόδοση της CATA++ μεθόδου.

Τέλος, παρατηρείται ότι η μέθοδος CATA++ παράγει βέλτιστες συστάσεις όταν τα δεδομένα που δίνονται ως είσοδος δεν είναι πολύ μεγάλα. Συγκεκριμένα, συμπεραίνετε ότι το βέλτιστο μέγεθος αρχείων εισόδου είναι περίπου τόσο, όσο και τα δεδομένα της citeulike-a βάσης δεδομένων.

Στη συνέχεια, εξετάζεται η αποτελεσματικότητα της μεθόδου CVAE, η οποία εκτελείται και πάλι για τα 3 διαφορετικά ερωτήματα, εξετάζοντας την αποδοτικότητα των νέων συναρτήσεων σε σύγκριση με τις παλιές, και ελέγχεται και πάλι η διαφορά της απόδοσης των δυο λογισμικών. Για τη μέτρηση της αποδοτικότητάς της πλέον επιλέγεται ο υπολογισμός μόνο της μεθόδου Recall, καθώς έχει ήδη αποδειχθεί ότι είναι αρκετά αντιπροσωπευτική.

Παρατηρείται ότι η χρήση των νέων συναρτήσεων activation function και weight initialization δεν βελτιστοποιούν τις εκτελέσεις της CVAE μεθόδου, σε αντίθεση με την CATA++ μέθοδο. Συνεπώς, για την παραγωγή των βέλτιστων συστάσεων με τη χρήση της CVAE μεθόδου προτείνεται να χρησιμοποιηθούν οι αρχικές συναρτήσεις. Ταυτόχρονα, η CVAE μέθοδος παράγει βέλτιστα αποτελέσματα για μικρότερου μεγέθους αρχεία, και όσο αυξάνεται το μέγεθός τους, τόσο χειροτερεύουν και οι συστάσεις.

Έπειτα, εκτελείται η μέθοδος CML η οποία είναι η λιγότερο αποτελεσματική. Πιο συγκεκριμένα, είναι αρκετά απαιτητική σε μνήμη και πιο χρονοβόρα συγκριτικά με τα άλλα. Αυτό είχε ως αποτέλεσμα να μη μπορεί να εκτελεστεί για όλα τα αρχεία δεδομένων που εκτελέστηκαν οι υπόλοιπες μέθοδοι. Παρατηρείται ότι η CML μέθοδος παράγει αποτελέσματα παρόμοιας απόδοσης για όλα τα αρχεία, κάτι το οποίο μπορεί να συνεπάγεται ότι αν βελτιωθεί ο τρόπος εκτέλεσής της, ίσως να είναι πιο αποδοτική για μεγαλύτερα δεδομένα εισόδου.

Στη συνέχεια, ακολουθεί η σύγκριση των μεθόδων. Οι μέθοδοι CATA++ και CVAE συγκρίνονται έχοντας τιμές της ίδιας μεθόδου αξιολόγησης για τα δυο συστήματα συστάσεων

σε ίδιες συνθήκες εκτέλεσης. Για να υπάρξει μια σύγκριση μεταξύ αυτών των μεθόδων και της CML μεθόδου, αρχικά εκτελέστηκαν όλες οι μέθοδοι με τη χρήση του Random Dataset. Μέσω των παραπάνω συγκρίσεων παρατηρήθηκε ότι για μικρό μέγεθος αρχείων, ορισμένες φορές η CML μέθοδος παρήγαγε τα βέλτιστα αποτελέσματα. Γενικά όμως, πιο αποδοτική μέθοδος θεωρείται η CATA++ και στη συνέχεια η CVAE καθώς η CML είναι πολύ περιοριστική. Για να επιβεβαιωθεί όμως ότι η CML μέθοδος είναι η βέλτιστη για πολύ μικρό μέγεθος αρχείων, δημιουργήθηκαν δύο νέα σύνολα δεδομένων που δόθηκαν ως είσοδος στις μεθόδους, και εκτελέστηκε ο κώδικας κάθε μεθόδου για τα το δεύτερο βασικό ερωτήματα. Η εκτέλεση έγινε σε διαφορετικό υπολογιστή με 16 GB RAM για να αποφευχθούν μηνύματα λάθους λόγω ανεπαρκούς μνήμης. Τέλος, υπολογίστηκε η recall μέθοδος αξιολόγησης της μεθόδου συγκεκριμένα για κ=50 και στις τρεις μεθόδους, και βγήκε το εξής συμπέρασμα: Για πολύ μικρά αρχεία τις βέλτιστες συστάσεις τις πραγματοποιεί η CML μέθοδος. Όμως, ακόμα και σε αυτή τη περίπτωση ο χρόνος που απαιτείται για την ολοκλήρωση της εκτέλεσης, και η απαραίτητη RAM την καθιστούν μη αποδοτική. Συνεπώς, γενικά βέλτιστη μέθοδος ανάπτυξης συστήματος συστάσεων για επιστημονικά δημοσιεύματα θεωρείται η CATA++, ακολουθεί η CVAE, και τέλος η CML.

5.2 Μελλοντικά Σχέδια

Τη βασική έλλειψη της παρούσας διπλωματικής εργασίας αποτελεί η μη αξιολόγηση της RVAE μεθόδου με τη Recall μέθοδο αξιολόγησης. Πιο συγκεκριμένα, η RVAE μέθοδος ανάπτυξης συστήματος συστάσεων αξιολογείται μέσω της μεθόδου AUC = Area Under Curve. Η μέθοδος AUC εμπεριέχει τον υπολογισμό της recall, καθώς η τιμή της αναπαριστάται στον x'x άξονα [17]. Συνεπώς, για να επιτευχθεί η σύγκριση της RVAE μεθόδου με τις υπόλοιπες θα πρέπει να υπολογιστεί η αποδοτικότητα της μεθόδου μέσω της recall, κάτι το οποίο δεν υπήρχε επαρκής χρόνος για να πραγματοποιηθεί στα πλαίσια ανάπτυξης αυτής της διπλωματικής εργασίας.

Επιπρόσθετα, μια ακόμη αλλαγή που θα μπορούσε να εκτελεστεί είναι να χρησιμοποιηθούν δεδομένα και από άλλες πηγές όπως citeulike-t ή το sciencedirect ώστε να συγκριθούν και να αποδειχθεί ποια βάση έχει τα πιο πλήρη και σωστά δεδομένα ώστε να προκύπτουν πιο σωστές συστάσεις.

Σημαντική βελτίωση θα αποτελούσε επίσης η κατάλληλη μετατροπή του κώδικα ώστε

να μπορεί να επεξεργάζεται δεδομένα μεγάλου όγκου / big data, κάτι το οποίο θα επιφέρει πολύ πιο σωστές και ακριβείς συστάσεις, και ταυτόχρονα αποτελεί έναν ρεαλιστικό στόχο για τη χρήση του συστήματος συστάσεων σε πραγματικές συνθήκες.

Στην ίδια λογική, είναι σημαντικό να διορθωθεί ο κώδικας της CML μεθόδου ώστε να μπορεί να επεξεργάζεται τα δεδομένα χωρίς την υπερβολική απαίτηση χρήσης μνήμης και χρόνου, καθώς δεν είναι αποδοτικό και εύχρηστο.

Επίσης, η βάση δεδομένων που χρησιμοποιείται έχει πάρα πολλά πεδία, και ορισμένα από αυτά δεν εκμεταλλεύονται πλήρως. Συνεπώς, θα μπορούσε να απαντηθούν ακόμα περισσότερα ερωτήματα με τη χρήση αυτών των δεδομένων και μεθόδων, όπως για παράδειγμα θα μπορούσε να γίνει μελέτη στην έρευνα μοντέλων συστάσεων παραπομπών με τη χρήση βαθιάς μάθησης.

Επίσης, πρέπει να γίνει σύγκριση και με άλλες μεθόδους ανάπτυξης συστημάτων συστάσεων είτε βασίζονται και πάλι με μεθόδους βαθιάς μάθησης όπως οι : CDL, CVAE++, POP, κ.α. , είτε όχι και να γίνει σύγκριση απόδοσης με τις στοιχειώδης μεθόδους ανάπτυξης συστημάτων συστάσεων, όπως SVM, CF, Association rule learning, κ.α. [18].

Τέλος, για διευκόλυνση κατά την εκτέλεση του κώδικα της κάθε μεθόδου ανάπτυξης συστήματος συστάσεων, προτείνεται να δημιουργηθεί ένα script το οποίο θα τρέχει αυτόματα όλους τους πιθανούς συνδυασμούς που δίνονται σε κάθε εκτέλεση του κώδικα. Έτσι, θα εξοικονομηθεί πολύς χρόνος, κυρίως αν πρέπει να τρέξουν ακόμα περισσότεροι συνδυασμοί από αυτούς που παρουσιάστηκαν σε αυτή τη διπλωματική εργασία. Έτσι, όλοι οι κώδικες θα τρέχουν στο παρασκήνιο, και όταν θα έχουν ολοκληρωθεί όλες οι εκτελέσεις θα είναι πιο εύκολο να παραχθούν τα αντίστοιχα διαγράμματα που οπτικοποιούν τα παραπάνω αποτελέσματα.

Βιβλιογραφία

[1] K. Falk. *Practical Recommender Systems*. Manning Publications, 2019.

[2] Xiaopeng Li and James She. Collaborative variational autoencoder for recommender systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 305–314. ACM, 2017.

[3] Cnn vs. rnn vs. ann – analyzing 3 types of neural networks in deep learning. `https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/`. Ημερομηνία πρόσβασης: 9-6-2021.

[4] Zafar Ali, Pavlos Kefalas, Khan Muhammad, Bahadar Ali, and Muhammad Imran. Deep learning in citation recommendation models survey. *Expert Syst. Appl.*, 162:113790, 2020.

[5] Yifan Zhu, Qika Lin, Hao Lu, Kaize Shi, Ping Qiu, and Zhendong Niu. Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks. *Knowl. Based Syst.*, 215:106744, 2021.

[6] Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. *Int. J. Digit. Libr.*, 21(4):375–405, 2020.

[7] Meshal Alfarhood and Jianlin Cheng. CATA++: A collaborative dual attentive autoencoder method for recommending scientific articles. *IEEE Access*, 8:183633–183648, 2020.

[8] Xiaopeng Li and James She. Relational variational autoencoder for link prediction with multimedia data. In Wanmin Wu, Jianchao Yang, Qi Tian, and Roger Zimmermann,

editors, *Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, October 23 - 27, 2017*, pages 93–100. ACM, 2017.

[9] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge J. Belongie, and Deborah Estrin. Collaborative metric learning. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 193–201. ACM, 2017.

[10] Citation network dataset. `https://www.aminer.org/citation`. Ημερομηνία πρόσβασης: 25-5-2021.

[11] Mongodb. `https://www.mongodb.com/`. Ημερομηνία πρόσβασης: 25-5-2021.

[12] Mongodb download. `https://www.mongodb.com/try/download/community`. Ημερομηνία πρόσβασης: 25-5-2021.

[13] Mongodb compass. `https://www.mongodb.com/try/download/compass`. Ημερομηνία πρόσβασης: 25-5-2021.

[14] Tf-idf explained and python sklearn implementation. `https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275`. Ημερομηνία πρόσβασης: 25-5-2021.

[15] Keras layer activation functions. `https://keras.io/api/layers/activations/`. Ημερομηνία πρόσβασης: 7-6-2021.

[16] Keras layer weight initializers. `https://keras.io/api/layers/initializers/#layer-weight-initializers`. Ημερομηνία πρόσβασης: 7-6-2021.

[17] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 233–240. ACM, 2006.

[18] Pradeep Singh, Pijush Dutta Pramanik, Avick Dey, and Prasenjit Choudhury. Recommender systems: An overview, research trends, and future directions. *International Journal of Business and Systems Research*, 15:14–52, 01 2021.