



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Προτεινόμενα Θέματα Διπλωματικών Εργασιών
2025

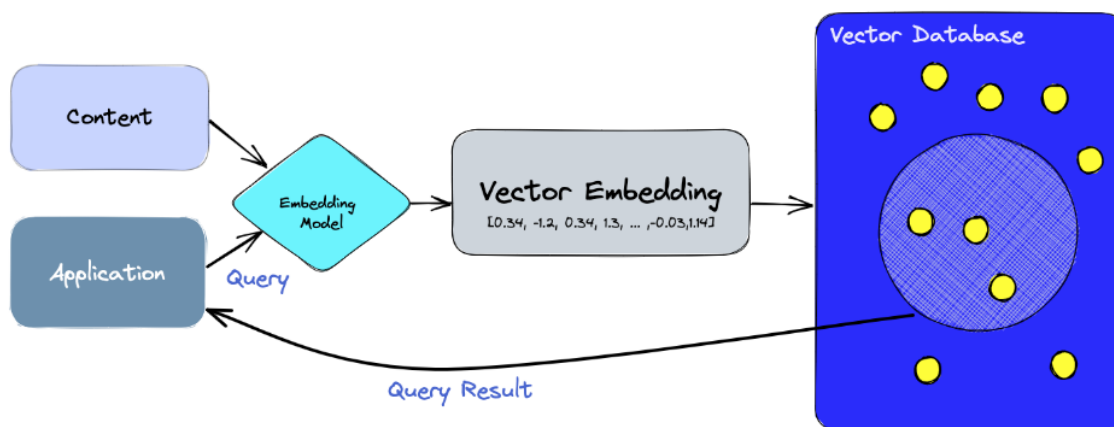
Θέμα 1

Τίτλος (Ελληνικά):	Βελτίωση Ποιότητας Αναλυτικής Κειμενικών Δεδομένων με χρήση Ενσωματώσεων
Τίτλος (Αγγλικά):	Improving Text Analytics Quality using Document Embeddings
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή(+2 μέλη):	Αθανάσιος Βουλόδημος, Γεώργιος Γκούμας
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη
Μια λέξη, χρησιμοποιώντας ένα μοντέλο μηχανικής μάθησης [1], μπορεί να μετασχηματιστεί σε ένα διάνυσμα, με τις πιο σημαντικές πληροφορίες να κωδικοποιούνται με αυτόν τον τρόπο (word embedding). Το ίδιο μπορεί να πραγματοποιηθεί δίνοντας μια ολόκληρη παράγραφο ή κείμενο το οποίο μπορεί να θα αναπαρασταθεί πλέον σε ένα διανυσματικό χώρο (document embedding, π.χ., [2, 3]). Χρησιμοποιώντας σύγκριση ανάμεσα σε διαθέσιμα δεδομένα, το σύστημα VEnOM του εργαστηρίου [4] επιτυγχάνει μοντέλα πρόβλεψης αναλυτικών τελεστών με ακρίβεια. Χρησιμοποιώντας document embeddings και αναζήτηση ομοιότητας, μπορούμε να βρούμε τα πιο κατάλληλα δεδομένα βάση του προβλήματος που θέλουμε να λύσουμε και να μοντελοποιήσουμε αναλυτικούς τελεστές για κειμενική / NLP αναλυτική. Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα: 1. Εύρεση συνόλων δεδομένων και μετατροπή τους σε document embedding διανύσματα χρησιμοποιώντας προ-εκπαιδευμένα βαθιά νευρωνικά δίκτυα αιχμής (π.χ. [3]) 2. Ανανέωση του τρόπου που λειτουργεί το VEnOM ώστε να χρησιμοποιούνται τα document embeddings. 3. Μοντελοποίηση και μέτρηση ακρίβειας τελεστών NLP	
Ενδεικτική Βιβλιογραφία: [1] Mikolov, Tomas. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013). [2] Doc2Vec, https://github.com/inejc/paragraph-vectors [3] Open-source embedding models, e.g., https://huggingface.co/Qwen/Qwen3-Embedding-0.6B , https://huggingface.co/collections/google/embeddinggemma-68b9ae3a72a82f0562a80dc4	

Θέμα 2

Τίτλος (Ελληνικά):	Υλοποίηση αλγορίθμων αναζήτησης σε Διανυσματική Βάση Δεδομένων
Τίτλος (Αγγλικά):	Implementation of state-of-the-art similarity algorithms in a Vector Database
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή(+2 μέλη):	Γ. Γκούμας, Ν. Κοζύρης
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Προγραμματισμός Συστημάτων, Προχωρημένα Θέματα ΒΔ

Σύντομη περιγραφή:



Μια Διανυσματική Βάση Δεδομένων (**vector database** [1, 2, 3]) είναι ένα σύστημα διαχείρισης βάσεων δεδομένων που έχει σχεδιαστεί ειδικά για την αποθήκευση, την ευρετηρίαση και την αποτελεσματική ανάκτηση διανυσματικών δεδομένων υψηλής διάστασης. Τα διανυσματικά δεδομένα (vectors) είναι αναπαραστάσεις αντικειμένων ή οντοτήτων ως διανύσματα σημείων σε έναν πολυδιάστατο χώρο. Αυτά μπορούν να προέρχονται από διάφορες πηγές, όπως η ενσωμάτωση κειμένου, εικόνων ή άλλων δεδομένων (embeddings). Οι διανυσματικές βάσεις δεδομένων παρέχουν: Αναζήτηση ομοιότητας για παρόμοια διανύσματα (π.χ., παρόμοια έγγραφα ή εικόνες), υποστήριξη για εφαρμογές μηχανικής μάθησης και υψηλά κλιμακούμενη απόδοση: Η αναζήτηση παρόμοιων αντικειμένων (similarity search) αποτελεί το κεντρικό σημείο της διπλωματικής. Μια αναζήτηση σε μια διανυσματική βάση μπορεί να υποστηριχτεί από πολλές μεθόδους, π.χ. Nearest Neighbor Search (NNS). Το NNS μπορεί να υποστηριχτεί από δομή πίνακα ή δένδρου/γράφου.

Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:

1. Υλοποίηση διαφορετικών αλγορίθμων αναζήτησης (π.χ., [8,9]) βασιζόμενων στο NNS χρησιμοποιώντας την βιβλιοθήκη faiss [6, 7],
2. επί διαφορετικών τύπων δεδομένων (structured, semi-structured, και unstructured), και
3. σύγκριση χρόνου εκτέλεσης τους και της ακρίβειας αναζήτησης των αντικειμένων.

Σαν παράδειγμα μπορείτε να δείτε την υλοποίηση στο [10].

Ενδεικτική Βιβλιογραφία:

- [1] <https://www.pinecone.io/learn/vector-database/>
- [2] <https://www.ibm.com/topics/vector-database>
- [3] <https://www.datacamp.com/blog/the-top-5-vector-databases>
- [4] Faiss, <https://github.com/facebookresearch/faiss/tree/main>
- [5] Douze, Matthijs, et al. "The faiss library." *arXiv preprint arXiv:2401.08281* (2024).
- [6] Jin, Yicheng, et al. "Curator: Efficient Indexing for Multi-Tenant Vector Databases." *arXiv preprint arXiv:2401.07119* (2024).
- [8] Cong Fu and Deng Cai. 2016. EFANNA : An Extremely Fast Approximate Nearest Neighbor Search Algorithm Based on kNN Graph. <https://doi.org/10.48550/arXiv.1609.07228>
- [9] Yury A. Malkov and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. <https://doi.org/10.48550/arXiv.1603.09320>
- [10] Curator, <https://github.com/hatsu3/curator/tree/main>

Θέμα 3

Τίτλος (Ελληνικά):	Αξιολόγηση μεθόδων ενσωμάτωσης για αβέβαιους Γράφους
Τίτλος (Αγγλικά):	Evaluating Embedding Methods for Uncertain Graphs
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή(+2 μέλη):	Γ. Γκούμας, Ν. Κοζύρης
Συσχετιζόμενο Μάθημα:	Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Προγραμματισμός Συστημάτων, Προχωρημένα Θέματα ΒΔ

Σύντομη περιγραφή:

Οι αβέβαιοι Γράφοι μοντελοποιούν πραγματικά δίκτυα με πιθανοτικές ακμές. Μέθοδοι όπως URGE[1], UAGE[2] και UnG-MoCha[3] ενσωματώνουν τέτοια γραφήματα για να υποστηρίξουν εργασίες όπως ταξινόμηση και ομαδοποίηση. Η παρούσα διπλωματική αξιολογεί πειραματικά αυτές τις προσεγγίσεις ως προς την ακρίβεια, την επεκτασιμότητα και την ανθεκτικότητα.

Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:

1. Εφαρμογή ή προσαρμογή μεθόδων ενσωμάτωσης αβέβαιων γραφημάτων (URGE, UAGE, UnG-MoCha)
2. Αξιολόγησή τους σε αβέβαια σύνολα δεδομένων (π.χ., πιθανοτικά κοινωνικά ή βιολογικά δίκτυα).
3. Μέτρηση της απόδοσης σε εργασίες όπως η ομαδοποίηση και η πρόβλεψη συνδέσεων.
4. Προσδιορισμός υπολογιστικών εμποδίων και ορίων ακρίβειας καθώς αυξάνεται η αβεβαιότητα.

Ενδεικτική Βιβλιογραφία:

- [1] Hu et al. (2017). URGE: UnceRtain Graph Embedding. <https://dl.acm.org/doi/10.1145/3132847.3132885>
- [2] Jiang et al. (2023). UAGE: Uncertain Attributed Graph Embedding. https://dl.acm.org/doi/abs/10.1007/978-3-031-39821-6_18
- [3] Ma et al. (2021). UnG-MoCha: Uncertain Graph Motif Counting with Neural Approximation. <https://dl.acm.org/doi/10.1145/3711896.3737170>
- [4] Chen et al. (2019). Learning Knowledge Graph Embeddings under Uncertainty (UKGE). <https://arxiv.org/abs/1811.10667>

Θέμα 4

Τίτλος (Ελληνικά):	Αλγοριθμική Αξιολόγηση Τελεστών σε Αβέβαιους Γράφους
Τίτλος (Αγγλικά):	Algorithmic Evaluation of Operators on Uncertain Graphs
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή (+2 μέλη):	Γ. Γκούμας, Δ. Φωτάκης
Συσχετιζόμενο Μάθημα:	Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Τεχνητή Νοημοσύνη, Προχωρημένα Θέματα ΒΔ
Σύντομη περιγραφή: Τα αβέβαια γραφήματα/γράφοι κωδικοποιούν τις ακμές με πιθανότητες ύπαρξης, καθιστώντας πολλούς κλασικούς αλγόριθμους γραφημάτων υπολογιστικά δύσκολους. Η παρούσα διπλωματική μελετά αλγόριθμους/τελεστές για αβέβαιους γράφους (π.χ., αξιοπιστία, προσβασιμότητα, μέτρηση μοτίβων, κ.ά.) προκειμένου να αξιολογήσει πού λειτουργούν αποτελεσματικά και πού αποτυγχάνουν. Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα: 1. Ανασκόπηση και εφαρμογή βασικών αλγορίθμων για αβέβαιες λειτουργίες γράφων. 2. Σύγκριση σε συνθετικά και πραγματικά αβέβαια δεδομένα γράφων. 3. Μέτρηση ακρίβειας, χρόνου εκτέλεσης και επεκτασιμότητας. 4. Προσδιορισμός προβληματικών σημείων και πρόταση βελτιώσεων.	
Ενδεικτική Βιβλιογραφία: [1] Danesh et al. (2023). A survey of clustering large probabilistic graphs: Techniques, evaluations, and applications. https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.13248 [2] Suman Banerjee (2021). A Survey on Mining and Analysis of Uncertain Graphs https://arxiv.org/abs/2106.07837 [3] Kassiano et al. (2016). Mining Uncertain Graphs: An Overview. https://datalab-old.csd.auth.gr/~gounaris/2016Algocloud_uncertaingraphs.pdf.	

Θέμα 5

Τίτλος (Ελληνικά):	Πρόβλεψη Ποιότητας Δυναμικών Δεδομένων με χρήση του Data Shapley
Τίτλος (Αγγλικά):	Predicting Dynamic Dataset Quality Using Data Shapley
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή(+2 μέλη):	Αθανάσιος Βουλόδημος, Γεώργιος Γκούμας
Συσχετιζόμενο Μάθημα:	Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη
Το Data Shapley [1] είναι μια έννοια από τη θεωρία συνεργατικών παιγνίων, που εφαρμόζεται στη μηχανική μάθηση για τον δίκαιο υπολογισμό της «αξίας» ή της «σημασίας» κάθε εγγραφής δεδομένων μέσα σε ένα σύνολο δεδομένων. Συνοπτικά, στο Data Shapley, κάθε δείγμα δεδομένων θεωρείται ως «παίκτης» που συμβάλλει στην ακρίβεια ή την απόδοση ενός μοντέλου. Με τον τρόπο αυτό, μπορούμε να αποφασίσουμε, για παράδειγμα, ποια δεδομένα να κρατήσουμε ή να αφαιρέσουμε για να βελτιώσουμε την απόδοση ή την αποδοτικότητα του μοντέλου. Στην εργασία [2] υλοποιήσαμε ένα γρήγορο και αποδοτικό υπολογισμό του Data Shapley για δεδομένα πινάκων. Ωστόσο, όλες οι εργασίες θεωρούν ακόμα ότι τα δεδομένα είναι στατικά. Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα: 1. Επέκταση της έννοιας Data Shapley από υπολογισμό της για ένα στατικό dataset στον αποδοτικό (επανα)υπολογισμό της για δυναμικά datasets. Συγκεκριμένα, ενδιαφέρουν (αρχικά) περιπτώσεις που στα δεδομένα προστίθενται μόνον νέες εγγραφές και πώς αυτό επηρεάζει τον αρχικό υπολογισμό. 2. Αποδοτική υλοποίηση του αλγορίθμου στο βήμα 1 για δυναμικά δεδομένα. 3. Εκτενείς μετρήσεις της αποδοτικότητας σε δεδομένα με διαφορετική συχνότητα και μέγεθος αλλαγών.	
Ενδεικτική Βιβλιογραφία: [1] Data Shapley: Equitable Valuation of Data for Machine Learning. Amirata Ghorbani, James Zou. https://proceedings.mlr.press/v97/ghorbani19c/ghorbani19c.pdf [2] http://www.cslab.ntua.gr/~dtsouma/index_files/C-DaSh_CIKM25.pdf	

Θέμα 6

Τίτλος (Ελληνικά):	Βελτιστοποίηση Ποιότητας για Σημασιολογικούς Συνδέσμους
Τίτλος (Αγγλικά):	Quality-Aware Semantic Join Optimization
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή(+2 μέλη):	N. Κοζύρης, Γ. Στάμου
Συσχετιζόμενο Μάθημα:	Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη
Οι Σημασιολογικοί Σύνδεσμοι (semantic joins) [1] επιτρέπουν τον υπολογισμό συνδέσμων, δηλαδή την αντιστοίχιση εγγραφών, που ικανοποιούν συνθήκες μέσω φυσικής γλώσσας. Τέτοιοι σύνδεσμοι μπορούν να υπολογιστούν χρησιμοποιώντας μεγάλα γλωσσικά μοντέλα (LLM) που επιλύουν νέες εργασίες χωρίς εκπαίδευση. Στην εργασία [2], υλοποιήσαμε ένα γρήγορο και αποδοτικό υπολογισμό του Data Sharpley για δεδομένα πινάκων - μια μέθοδος που υπολογίζει με αποδοτικό τρόπο μια μετρική ποιότητας σε δεδομένα. Στην παρούσα Διπλωματική καλούμαστε να προτείνουμε μια επέκταση του Semantic Join με χρήση της μεθόδου C-DaSH και ενσωματώσεων: 1. Σχεδιασμός και χρήση του C-DaSH για φιλτράρισμα ή επαναστάθμιση των εγγραφών εισόδου πριν από τις λειτουργίες σημασιολογικής ένωσης, με στόχο τη βελτίωση της ακρίβειας των αποτελεσμάτων και τη μείωση του κόστους LLM εξαιρώντας δεδομένα χαμηλής ποιότητας. 2. Συνδυασμός ενσωματώσεων σε επίπεδο συνόλου δεδομένων για αποδοτική επιλογή υποσχόμενων μπλοκ πλειάδων και, στη συνέχεια, εφαρμογή του προσαρμοσμένου αλγορίθμου block nested loop join [1] για την τελική επικύρωση από το/τα LLM.	
Ενδεικτική Βιβλιογραφία: [1] Implementing Semantic Join Operators Efficiently, Immanuel Trummer https://arxiv.org/pdf/2510.08489 [2] http://www.cslab.ntua.gr/~dtsouma/index_files/C-DaSH_CIKM25.pdf	