



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

Προτεινόμενα Θέματα Διπλωματικών Εργασιών
Φεβ. 2026

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 |
| <p>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.</p> <p>Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:</p> <ol style="list-style-type: none">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 | |

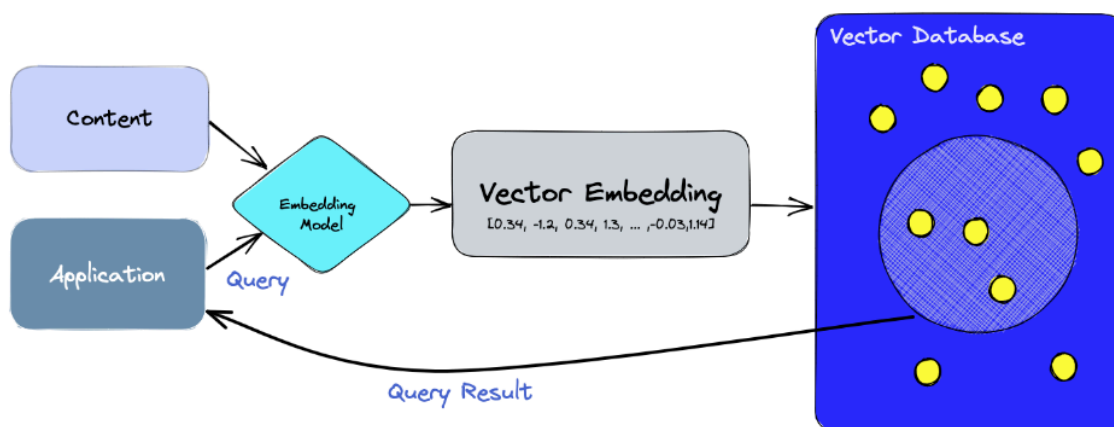
[3] Open-source embedding models, e.g., <https://huggingface.co/Qwen/Qwen3-Embedding-0.6B> ,
<https://huggingface.co/collections/google/embeddinggemma-68b9ae3a72a82f0562a80dc4>

[4] http://www.cslab.ntua.gr/~dtsouma/index_files/VEnOM_DEXA25.pdf

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 υποστηριχθεί από πολλές μεθόδους, π.χ. Nearest Neighbor Search (NNS). Το NNS μπορεί να υποστηριχθεί από δομή πίνακα ή δέντρου/γράφου.

Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:

1. Υλοποίηση ή τροποποίηση διαφορετικών αλγορίθμων αναζήτησης βασισμένων στο kNNS χρησιμοποιώντας την βιβλιοθήκη faiss [4],
2. επί διαφορετικών τύπων αλλά και πολυτροπικών δεδομένων, και
3. πιθανή ενσωμάτωση της υλοποίησης στην [4] ή σε διανυσματική ΒΔ ανοικτού κώδικα (π.χ., Milvus, Weaviate, Qdrant)

Σαν παράδειγμα μπορείτε να δείτε την υλοποίηση στο [5].

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>
- [5] Curator, <https://github.com/hatsu3/curator/tree/main>

Θέμα 3

| | |
|--|---|
| Τίτλος (Ελληνικά): | Μέθοδοι Ενσωμάτωσης για Αβέβαιους Γράφους |
| Τίτλος (Αγγλικά): | 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 |
| <p>Σύντομη περιγραφή:</p> <p>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 θα αξιολογήσει αυτές τις προσεγγίσεις ως προς την ακρίβεια, την επεκτασιμότητα και την ανθεκτικότητα.</p> <p>Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:</p> <ol style="list-style-type: none"> 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. | |
| <p>Ενδεικτική Βιβλιογραφία:</p> <p>[1] Hu et al. (2017). URGE: UnceRtain Graph Embedding. https://dl.acm.org/doi/10.1145/3132847.3132885</p> <p>[2] Jiang et al. (2023). UAGE: Uncertain Attributed Graph Embedding. https://dl.acm.org/doi/abs/10.1007/978-3-031-39821-6_18</p> <p>[3] Ma et al. (2021). UnG-MoCha: Uncertain Graph Motif Counting with Neural Approximation. https://dl.acm.org/doi/10.1145/3711896.3737170</p> <p>[4] Chen et al. (2019). Learning Knowledge Graph Embeddings under Uncertainty (UKGE). https://arxiv.org/abs/1811.10667</p> | |

Θέμα 4

| | |
|--|---|
| Τίτλος (Ελληνικά): | Αλγοριθμική Αξιολόγηση Αναλυτικών Τελεστών σε Αβέβαιους Γράφους |
| Τίτλος (Αγγλικά): | Algorithmic Evaluation of Analytics 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 |
| <p>Σύντομη περιγραφή:</p> <p>Τα αβέβαια γραφήματα/γράφοι κωδικοποιούν τις ακμές με πιθανότητες ύπαρξης, καθιστώντας πολλούς κλασικούς αλγόριθμους γραφημάτων υπολογιστικά δύσκολους. Η παρούσα διπλωματική μελετά αλγόριθμους/τελεστές αναλυτικής επεξεργασίας δεδομένων για αβέβαιους γράφους (π.χ., αξιοπιστία, προσβασιμότητα, μέτρηση μοτίβων, κ.ά.) προκειμένου να αξιολογήσει πού λειτουργούν αποτελεσματικά και πού αποτυγχάνουν.</p> <p>Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:</p> <ol style="list-style-type: none"> 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. | |
| <p>Ενδεικτική Βιβλιογραφία:</p> <p>[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</p> <p>[2] Suman Banerjee (2021). A Survey on Mining and Analysis of Uncertain Graphs https://arxiv.org/abs/2106.07837</p> <p>[3] Kassiano et al. (2016). Mining Uncertain Graphs: An Overview. https://datalab-old.csd.auth.gr/~gounaris/2016Algocloud_uncertaingraphs.pdf.</p> | |

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 |
| <p>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.</p> <p>Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:</p> <ol style="list-style-type: none"> 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. | |
| <p>Ενδεικτική Βιβλιογραφία:</p> <p>[1] Data Shapley: Equitable Valuation of Data for Machine Learning. Amirata Ghorbani, James Zou. https://proceedings.mlr.press/v97/ghorbani19c/ghorbani19c.pdf</p> <p>[2] http://www.cslab.ntua.gr/~dtsouma/index_files/C-DaSh_CIKM25.pdf</p> | |

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 |
| <p>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.</p> <p>Στην παρούσα Διπλωματική καλούμαστε να προτείνουμε μια επέκταση του Semantic Join με χρήση της μεθόδου C-DaSh και ενσωματώσεων:</p> <ol style="list-style-type: none"> 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. | |
| <p>Ενδεικτική Βιβλιογραφία:</p> <p>[1] Implementing Semantic Join Operators Efficiently, Immanuel Trummer https://arxiv.org/pdf/2510.08489</p> <p>[2] http://www.cslab.ntua.gr/~dtsouma/index_files/C-DaSh_CIKM25.pdf</p> | |

Θέμα 7

| | |
|---|--|
| Τίτλος (Ελληνικά): | Σημασιολογικοί Τελεστές για Αναλυτική Δεδομένων Υποβοηθούμενοι από Μεγάλα Γλωσσικά Μοντέλα |
| Τίτλος (Αγγλικά): | Semantic Operators for LLM-Augmented Analytical Query Processing |
| Επιβλέπων: | Δημήτριος Τσουμάκος |
| Επιτροπή(+2 μέλη): | N. Κοζύρης, Γ. Στάμου |
| Συσχετιζόμενο Μάθημα: | Προχωρημένα Θέματα ΒΔ |
| Απαραίτητες Γνώσεις: | Βάσεις Δεδομένων, Μηχανική Μάθηση |
| Επιθυμητές Γνώσεις: | Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη |
| <p>Οι σύγχρονες εφαρμογές δεδομένων υπερβαίνουν την απλή SQL και ενσωματώνουν «Σημασιολογικούς Τελεστές», δηλαδή λειτουργίες που απαιτούν μοντέλα TN για εκτέλεση (ενσωματώνοντας δυνατότητες LLM), όπως το <code>SemanticFilter</code> (π.χ., «Βρες tweets για χαρούμενα γεγονότα») ή το <code>SemanticJoin</code>. Το LOTUS [1] εισάγει σημασιολογικούς τελεστές, τελεστές φίλτρου, συνένωσης, ομαδοποίησης, top-k και συνάθροισης με υποστήριξη LLM και κατηγορήματα φυσικής γλώσσας (π.χ., <code>SEM_FILTER(papers, "related to climate change")</code>) — και παρέχει στατιστικές εγγυήσεις ακρίβειας επιτυγχάνοντας έως και 1000× μείωση κόστους. Το SwellDB [2] προχωρά παραπέρα, δημιουργώντας πίνακες κατά ζήτηση (on-the-fly) από εξωτερικές πηγές (αναζήτηση ιστού, APIs, άλλες βάσεις) με χρήση LLMs, καθιστώντας δεδομένα που προηγουμένως δεν ήταν διαθέσιμα αναζητήσιμα με τυπική SQL.</p> <p>Η παρούσα διπλωματική θα σχεδιάσει και θα υλοποιήσει ένα σύστημα που συνδυάζει σημασιολογικούς τελεστές με βελτιστοποίηση βάσει κόστους, επιλέγοντας μεταξύ ακριβούς αξιολόγησης στη ΒΔ και σημασιολογικής αξιολόγησης μέσω LLM ανάλογα με το εκτιμώμενο κόστος και τις απαιτήσεις ακρίβειας. Θα γίνει επέκταση υπάρχοντος μηχανισμού εκτέλεσης ερωτημάτων (π.χ., DuckDB ή PostgreSQL μέσω foreign data wrappers) με ένα σύνολο σημασιολογικών τελεστών:</p> <p><code>SEM_FILTER(table, predicate_nl)</code> : φιλτράρισμα με LLM βάσει κατηγορημάτων φυσικής γλώσσας.</p> <p><code>SEM_JOIN(table1, table2, condition_nl)</code> : σημασιολογική συνένωση.</p> <p>Έπειτα θα προχωρήσει στη σχεδίαση πλαισίου βελτιστοποίησης που, δεδομένου ενός ερωτήματος το οποίο αναμειγνύει τυπική SQL και σημασιολογικούς τελεστές θα εκτιμά το κόστος κάθε κλήσης σημασιολογικού τελεστή (LLM tokens, καθυστέρηση, χρηματικό κόστος) με κάποιο μοντέλο ώστε να παρέχει εγγυήσεις ακρίβειας εντός ανοχής ϵ και εμπιστοσύνης $1-\delta$ που ορίζει ο χρήστης καθώς και δημιουργία πινάκων κατά ζήτηση (εμπνευσμένη από το SwellDB) που, δοθέντος ορισμού σχήματος και προδιαγραφής πηγής δεδομένων, υλοποιεί έναν εικονικό πίνακα μέσω εξαγωγής/μετασχηματισμού με LLM από πηγές ιστού/API.</p> <p>Αξιολόγηση είναι επιθυμητή σε πραγματικά αναλυτικά καθήκοντα (π.χ., επαλήθευση γεγονότων σε ειδησεογραφικά κείμενα, ταξινόμηση βιοϊατρικής βιβλιογραφίας από το PubMed, κ.τ.λ.), με μέτρηση ακρίβειας, καθυστέρησης και κόστους σε σύγκριση με τα baselines του LOTUS και με την αφελή κλήση LLM ανά γραμμή.</p> | |
| Ενδεικτική Βιβλιογραφία: | |
| <p>[1] L. Patel et al., “Semantic Operators and Their Optimization: Enabling LLM-Based Data Processing with Accuracy Guarantees in LOTUS,” PVLDB 18(11), 2025.</p> <p>[2] V. Giannakouris & I. Trummer, “SwellDB: GenAI-Native Query Processing via On-the-Fly Table Generation,” VLDB PhD Workshop, 2025.</p> | |

[3] S. Urban & C. Binnig, "CAESURA: Language Models as Multi-Modal Query Planners," CIDR, 2024.

[4] F. Nargesian et al., "Table Union Search on Open Data," PVLDB, 2018.

Θέμα 8

| | |
|------------------------------|--|
| Τίτλος (Ελληνικά): | Ανάλυση Υπολογιστικής Προπαγάνδας: FIMIScope-Agent: Αυτόνομο Σύστημα Πρακτόρων για την Ανάλυση Υπολογιστικής Προπαγάνδας |
| Τίτλος (Αγγλικά): | Computational Propaganda Analysis: FIMIScope-Agent |
| Επιβλέπων: | Δημήτριος Τσουμάκος |
| Επιτροπή(+2 μέλη): | N. Κοζύρης, Μ. Δικαϊάκος |
| Συσχετιζόμενο Μάθημα: | Προχωρημένα Θέματα ΒΔ |
| Απαραίτητες Γνώσεις: | Βάσεις Δεδομένων, Μηχανική Μάθηση |
| Επιθυμητές Γνώσεις: | Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη |

Η διάδοση ψευδών ειδήσεων και παραπληροφόρησης στα μέσα κοινωνικής δικτύωσης έχει αναδειχθεί ως μία από τις σημαντικότερες προκλήσεις της σύγχρονης κοινωνίας. Πολυάριθμες έρευνες έχουν τεκμηριώσει τον δυνητικό αντίκτυπο των ψευδών ειδήσεων στη δημοκρατική διαδικασία και την κοινωνική συνοχή. Πρόσφατες αναλύσεις περιστατικών έχουν παράσχει αποδείξεις ότι οι ψευδείς ειδήσεις, και οι υποκείμενοι μηχανισμοί που τις καθιστούν viral στο διαδίκτυο, χρησιμοποιούνται ως όπλα από κρατικούς και μη-κρατικούς γεωπολιτικούς δρώντες που εκτοξεύουν εκστρατείες υπολογιστικής προπαγάνδας στο πλαίσιο ψυχολογικού πολέμου, επιδιώκοντας να επιτύχουν βραχυπρόθεσμους στόχους και/ή μακροπρόθεσμους στρατηγικούς σκοπούς. Σε απάντηση αυτών των ανησυχιών και πρωτοβουλιών, αρκετές ερευνητικές ομάδες, ΜΚΟ, κυβερνητικές και διακυβερνητικές οργανώσεις, συμπεριλαμβανομένης της Ευρωπαϊκής Υπηρεσίας Εξωτερικής Δράσης (EEAS), έχουν προτείνει και αναπτύξει μεθοδολογικά πλαίσια και εργαλεία για την ανάλυση και αντιμετώπιση εκστρατειών "Foreign Information Manipulation and Interference" (FIMI).

Το σύστημα FIMIScope αποτελεί ένα λογισμικό εργαλείο συνεργατικής ανάλυσης περιπτώσεων παραπληροφόρησης, συμβάντων και εκστρατειών μέσω browser-based GUI. Το λογισμικό υλοποιεί τους "FIMI canvases" ως γραφική διεπαφή χρήστη σε browser, υποστηριζόμενη από backend σύστημα που παρέχει υπηρεσίες για αποθήκευση, διαχείριση και εξαγωγή δεδομένων και μεταδεδομένων που συλλέγονται κατά τη διαδικασία ανάλυσης, υποστήριξη απομακρυσμένης συνεργασίας μεταξύ αναλυτών, και διατήρηση snapshots διαφορετικών σταδίων μιας ανάλυσης. Παρόλο που το υπάρχον FIMIScope περιλαμβάνει ένα εργαλείο Large Language Model που δημιουργεί αφηγήσεις (narratives) από τα συλλεγμένα JSON δεδομένα και επιτρέπει στους αναλυτές να αλληλεπιδρούν με τα δεδομένα μέσω διαλόγου σε φυσική γλώσσα, το σύστημα απαιτεί σημαντική χειροκίνητη προσπάθεια από τους αναλυτές για την ανακάλυψη, συλλογή και σχολιασμό FIMI περιεχομένου σε πολλαπλές πλατφόρμες κοινωνικής δικτύωσης. Η προτεινόμενη διπλωματική εργασία στοχεύει στην επέκταση του FIMIScope με δυνατότητες αυτόνομων πρακτόρων (agentic capabilities) που αυτόματα ανακαλύπτουν, συλλέγουν, παρακολουθούν και εκτελούν προκαταρκτική ανάλυση πιθανού FIMI περιεχομένου, μειώνοντας σημαντικά τον χειροκίνητο φόρτο εργασίας των ανθρώπινων αναλυτών, διατηρώντας παράλληλα την ανθρώπινη επίβλεψη για κρίσιμες αποφάσεις.

Στόχοι: Επέκταση του υπάρχοντος συστήματος FIMIScope με δυνατότητες αυτόνομων πρακτόρων (agentic capabilities) για την αυτόματη ανακάλυψη, συλλογή, παρακολούθηση και προκαταρκτική ανάλυση περιεχομένου Foreign Information Manipulation and Interference (FIMI) σε πλατφόρμες κοινωνικής δικτύωσης. Το προτεινόμενο FIMIScope-Agent θα ενσωματώσει ειδικευμένους πράκτορες όπως: (1) Discovery Agent - παρακολουθεί πλατφόρμες (Twitter/X, Facebook, Telegram, TikTok) για δείκτες FIMI, ανιχνεύει coordinated inauthentic behavior (συγχρονισμένες αναρτήσεις, bot-like συμπεριφορά), (2) Collection Agent - αυτόματο scraping και αρχειοθέτηση εντοπισμένου

περιεχομένου (posts, εικόνες, videos, metadata), (3) Analysis Agent - προκαταρκτική ανάλυση με NLP (sentiment, toxicity detection), network analysis (follower graphs, coordination metrics), visual similarity detection, temporal analysis, (4) Narrative Generation Agent - χρήση LLM για αυτόματη παραγωγή περιλήψεων συμβάντων, timelines εκστρατειών, υποθέσεων για στόχους και τακτικές, (5) Alert Agent - real-time monitoring για αναδυόμενες FIMI εκστρατείες με προτεραιοποίηση βάσει virality, reach, geopolitical context.

Ενδεικτική Βιβλιογραφία:

- [1] FIMIScope Camvases: M. Dikaiiakos <https://github.com/dikaiiakos/FIMI-Map-Canvas>
- [2] FIMIScope Software Presentation: K. Ioannidou, M. Dikaiiakos, "FIMIScope" <https://www.youtube.com/watch?v=B0D45tWfsIA>
- [3] Fake News: Zhou and Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, 2020.
- [4] Disinformation Analysis: Nimmo and Hutchins, "Phase-based Tactical Analysis of Online Operations," Carnegie Endowment for International Peace, 2023.
- [5] Disinformation Analysis: Buitrago-Lopez et al. "Frameworks, Modeling and Simulations of Misinformation and Disinformation: A Systematic Literature Review," ACM Computing Surveys, 2024.
- [6] Countermeasures: International Panel on the Information Environment, "Countermeasures for Mitigating Digital Misinformation: A Systematic Review," 2023
- [7] Foreign Information Manipulation and Interference: Wright, David (ed.), Malak Altaeb, Tim Beyer, Annye Braca Joshua Bronson, Owen Conlan, Arsenio Cuenca, Marios Dikaiiakos, Zaur Gouliev, Pablo Hernández, Richa Kumar, Evangelos Markatos, Raquel Miguel, Gary Munnelly, Alina Östling, George Pallis, Emmanouil Papadogiannakis, Demetris Paschalides, Pau Perea, Anna Pomortseva, Marianna Prysiazhniuk, Celia Ramos-Vera, Jari Räsänen, Kristian Reeson, Elodie Reuge, Mario Reyes de los Mozos, Joakim Rosell, Maria Giovanna Sessa, Brendan Spillane, Dimosthenis Stefanidis, Niklas Strand, Jaakko Tyni, Ilkka Vuorikuru, David Wright (authors), "Foreign Information Manipulation and Interference," Springer, 2025.
- [8] Fake News Detection: Paschalides, Kornilakis, Christodoulou et al. "Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web," IEEE/WIC/ACM International Conference on Web Intelligence, 298-302, 2019.

Θέμα 9

| | |
|--|--|
| Τίτλος (Ελληνικά): | Μοντελοποίηση Ενεργειακής Κατανάλωσης GPU Clusters για LLM Inference |
| Τίτλος (Αγγλικά): | Modeling of GPU clusters' energy consumption for LLM inference |
| Επιβλέπων: | Δημήτριος Τσουμάκος |
| Επιτροπή(+2 μέλη): | N. Κοζύρης, Μ. Δικαϊάκος |
| Συσχετιζόμενο Μάθημα: | Προχωρημένα Θέματα ΒΔ |
| Απαραίτητες Γνώσεις: | Βάσεις Δεδομένων, Μηχανική Μάθηση |
| Επιθυμητές Γνώσεις: | Προχωρημένα Θέματα ΒΔ, Τεχνητή Νοημοσύνη |
| <p>Ανάπτυξη μαθηματικών και machine learning μοντέλων για την πρόβλεψη της ενεργειακής κατανάλωσης GPU clusters κατά το LLM inference, λαμβάνοντας υπόψη παραμέτρους όπως: (1) Workload characteristics - incoming prompts per second, prompt length distribution (context tokens), generation length distribution (output tokens), batch size, (2) Hardware configuration - GPU τύπος (π.χ. H100, A100, B200, T4), αριθμός GPUs, μνήμη GPU, interconnect (NVLink, PCIe), (3) Model properties - μέγεθος μοντέλου (7B, 70B, 405B parameters), quantization (FP16, INT8, INT4), inference framework (vLLM, TensorRT-LLM, TGI). (4) Parallel execution of multiple prompts. Η εργασία θα διερευνήσει και συγκρίνει εναλλακτικούς τρόπους μοντελοποίησης της κατανάλωσης ενέργειας από παραμετροποιημένες συστοιχίες GPU καθώς αυτές εκτελούν εργασίες συμπερασματικών υπολογισμών.</p> <p>Ερευνητικά Ερωτήματα: (α) Ποιοι παράγοντες επηρεάζουν περισσότερο την ενεργειακή κατανάλωση; Sensitivity analysis για prompt length, batch size, model size, (β) Πώς διαφέρει η energy efficiency μεταξύ GPU γενεών; Σύγκριση H100 vs A100 vs T4 per token, (γ) Ποια είναι η σχέση μεταξύ throughput (tokens/sec) και energy consumption; Trade-offs για διαφορετικά batch sizes, (δ) Πόσο ακριβή είναι τα analytical models συγκριτικά με ML models; Validation σε πραγματικά workloads.</p> <p>Μεθοδολογία: (1) Data collection - συλλογή energy measurements από GPU clusters χρησιμοποιώντας NVIDIA SMI, NVML API, RAPL (Running Average Power Limit) για CPU, smart PDUs για rack-level power. Χρήση benchmark traces από Azure LLM Inference Dataset ή synthetic workloads από ServeGen, (2) Feature engineering - εξαγωγή χαρακτηριστικών από traces: mean/median/95th percentile prompt length, request rate variability, GPU utilization %, memory bandwidth utilization, (3) Model development - υλοποίηση analytical models (linear, piecewise linear) και training ML models με scikit-learn/PyTorch, (4) Validation - σύγκριση predictions με actual measurements σε testbed με 4-8 GPUs, υπολογισμός MAPE (Mean Absolute Percentage Error), RMSE.</p> <p>Αξιολόγηση: Μετρικές απόδοσης μοντέλων: (1) Prediction accuracy - MAPE < 10% θεωρείται αποδεκτό για production, (2) Training time - πόσο χρόνο χρειάζονται τα ML models για training, (3) Inference latency - πόσο γρήγορα μπορούν να κάνουν predictions (real-time monitoring), (4) Generalization - performance σε unseen workloads και GPU configurations. Πειραματική αξιολόγηση σε 3 scenarios: (α) Llama 3.3 70B σε 4x H100 GPUs με varying request rates (10-100 req/sec), (β) Mixtral 8x7B σε 8x A100 GPUs με different batch sizes (1, 4, 16, 32), (γ) GPT-NeoX 20B σε 2x T4 GPUs για edge deployment scenario.</p> <p>Παραδοτέα: (1) Energy profiling dataset – ανοικτά δεδομένα κατανάλωσης ενέργειας GPU inference, (2) Model library - Python package με trained models, (3) Benchmark report - σύγκριση accuracy/speed των διαφορετικών μοντέλων, (4) Integration guide - πώς να ενσωματωθούν τα models σε monitoring pipelines (Prometheus, Grafana).</p> | |

Ενδεικτική Βιβλιογραφία:

- [1] GLASSLab Paper: Tsiopani, Symeonidis, Pallis, Dikaiakos, "GLASSLab: Geo-distributed LLM Assessment for Sustainability and Scalability through Simulation" (unpublished working draft). Simulation framework για energy, carbon και water footprint σε geo-distributed LLM inference.
- [2] Energy Benchmarking: Samsi et al., "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference", IEEE HPEC 2023. Benchmarking energy consumption για διάφορα LLMs και hardware configurations.
- [3] DynamoLLM: Stojkovic et al., "DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency", IEEE HPCA 2025. Energy management για LLM clusters.
- [4] Carbon & Water Footprint: Jegham et al., "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference", arXiv:2505.09598, 2025.
- [5] Google's Environmental Impact: Elsworth et al., "Measuring the Environmental Impact of Delivering AI at Google Scale", arXiv:2508.15734, 2025. Real-world data για AI workloads σε hyperscale datacenters.

Data Sources & Tools:

- [6] Artificial Analysis: <https://artificialanalysis.ai> - Benchmarks για LLM performance, pricing, και throughput.
- [7] Azure LLM Inference Dataset: <https://github.com/Azure/AzurePublicDataset> - Real-world LLM inference traces.
- [8] NVIDIA Management Library (NVML): <https://developer.nvidia.com/nvml> - API για GPU monitoring (power, utilization, memory).
- [9] ElectricityMaps: <https://app.electricitymaps.com> - Real-time carbon intensity data ανά χώρα.