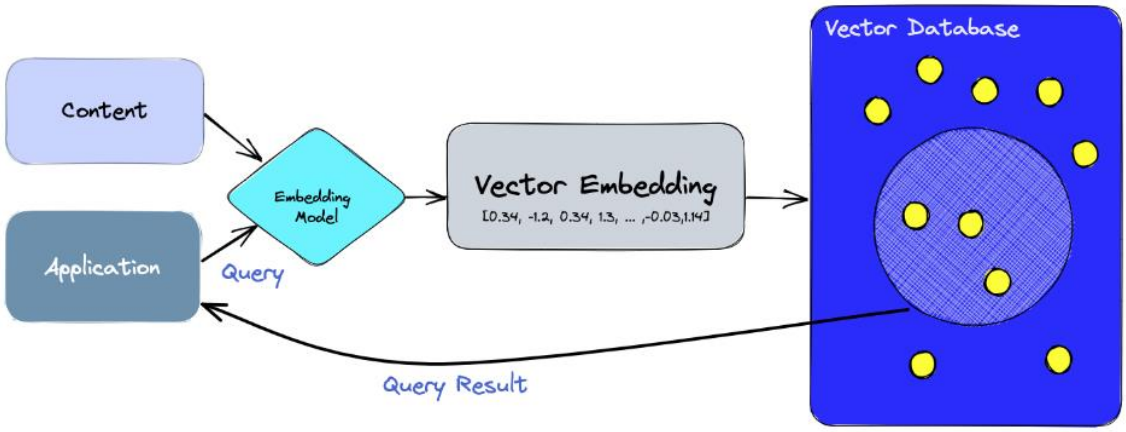
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Προτεινόμενα Θέματα Διπλωματικών Εργασιών

Θέμα 1

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

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

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

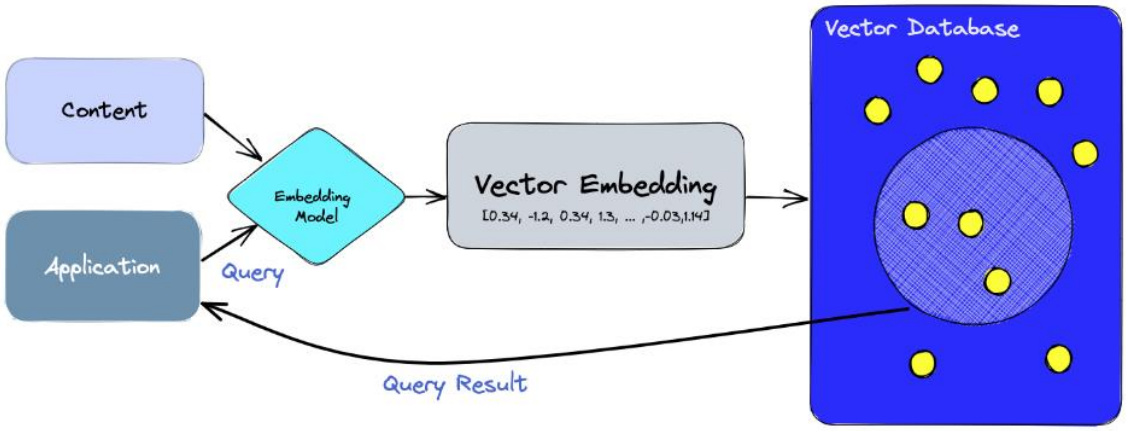
[9] Yury A. Malkov and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. https://doi.org/10.48550/ arXiv.1603.09320

[10] Curator, https://github.com/hatsu3/curator/tree/main

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 2

Τίτλος (Ελληνικά):	Ενσωμάτωση Διανυσματικής Βάσης Δεδομένων στο Σύστημα Apollo
Τίτλος (Αγγλικά):	Vector Database Integration with Apollo
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.

Το σύστημα Apollo του εργαστηρίου [4] επιτυγχάνει μοντέλα πρόβλεψης αναλυτικών τελεστών με ακρίβεια. Χρησιμοποιώντας αναζήτηση ομοιότητας ανάμεσα στα διαθέσιμα δεδομένα, μπορούμε να βρούμε τα πιο κατάλληλα δεδομένα βάσει του τελεστή και να τον μοντελοποιήσουμε με μεγάλη ακρίβεια και ταχύτητα ([5, 6]).

Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:

1. Ενοποίηση μια μοντέρνας Vector Database για αποθήκευση και επερώτηση διανυσμάτων σχετικών με τα διαθέσιμα δεδομένα (πίνακες και γράφοι).
2. Σύγκριση απόδοσης του νέου συστήματος με το υπάρχον πλαίσιο.

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] https://dtsouma.github.io/armada/

[5] T. Bakogiannis, I. Giannakopoulos, D. Tsoumakos and N. Koziris: Apollo: A Dataset Profiling and Operator Modeling System. In Proceedings of the 2019 ACM SIGMOD/PODS International Conference on Management of Data, June 30 - July 5, 2019 Amsterdam, The Netherlands.

[6] T. Bakogiannis, I. Giannakopoulos, D. Tsoumakos and N. Koziris: Predicting Graph Operator Output over Mul-

tiple Graphs. In Proceedings of the 19th International Conference on Web Engineering, June 11 - 14, 2019, Daejeon, Korea.

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 3

Τίτλος (Ελληνικά):	Βελτιστοποίηση απόδοσης ΒΔ με χρήση AutoML σε σχέση με υπάρχουσες ML4DB τεχνικές
Τίτλος (Αγγλικά):	DB performance optimization using AutoML vs. existing ML4DB techniques
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή (+2 μέλη):	Γ. Γκούμας, Ν. Κοζύρης
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Machine Learning
Επιθυμητές Γνώσεις:	Τεχνητή Νοημοσύνη, Προχωρημένα Θέματα ΒΔ

Σύντομη περιγραφή:

Η αυτοματοποιημένη μηχανική μάθηση (AutoML) είναι η διαδικασία αυτοματοποίησης των εργασιών εφαρμογής μηχανικής μάθησης σε προβλήματα του πραγματικού κόσμου. Το AutoML περιλαμβάνει δυνητικά κάθε στάδιο από την αρχή έως την κατασκευή ενός μοντέλου μηχανικής μάθησης και στοχεύει να επιτρέψει σε μη ειδικούς να κάνουν χρήση μοντέλων και τεχνικών μηχανικής εκμάθησης. Η αυτοματοποίηση της διαδικασίας προσφέρει επιπλέον τα πλεονεκτήματα της παραγωγής απλούστερων λύσεων, ταχύτερης δημιουργίας αυτών των λύσεων και μοντέλων που συχνά υπερτερούν των μοντέλων που έχουν σχεδιαστεί/επιλεγεί με μη αυτόματο τρόπο.

Το ML4DB (Machine Learning για Βάσεις Δεδομένων) είναι ένας αναδυόμενος τομέας που στοχεύει στην αξιοποίηση τεχνικών μηχανικής μάθησης για τον αυτοματισμό και τη βελτιστοποίηση διαφόρων πτυχών των συστημάτων βάσεων δεδομένων. Η βασική ιδέα πίσω από το ML4DB είναι η χρήση μοντέλων και αλγορίθμων μηχανικής μάθησης για να αντικαταστήσουν ή να ενισχύσουν τις παραδοσιακές προσεγγίσεις που βασίζονται σε κανόνες για τη διαχείριση και βελτιστοποίηση βάσεων δεδομένων. Η έννοια του ML4DB μπορεί να εφαρμοστεί σε διάφορα στάδια όπως: *Ρύθμιση Βάσεων Δεδομένων, Επιλογή Ευρετηρίων και Φυσικός Σχεδιασμός, Βελτιστοποίηση Ερωτημάτων, Πρόβλεψη Φορτίου Εργασίας και Διαχείριση Πόρων*, κλπ.

Μια πολύ αναλυτική παρουσίαση και σύγκριση λύσεων ML4DB έχει δημοσιευτεί πρόσφατα [2] με κώδικα και δεδομένα σύγκρισης ανοικτά [3]. Σε αυτή τη διπλωματική, καλείστε να χρησιμοποιήσετε μέρος της αξιολόγησης αυτής ώστε να συγκρίνετε υπάρχουσες λύσεις με τη δική σας που θα χρησιμοποιεί AutoML, σε ένα ή περισσότερα στάδια βελτιστοποίησης (π.χ., index selection, cardinality estimation, cost estimation, etc).

Ενδεικτική Βιβλιογραφία:

[1] https://www.automl.org/automl/

[2] https://dl.acm.org/doi/abs/10.14778/3636218.3636235?casa_token=56rG_HUGdHYAAAAA:M9L-JKpcLKzLz6uJCEtnXko1FvfsLpBRY1teTwl1iuiSeslBnKWLw9J4nSU6m6qF0AbuvLTX_BWX

[3] https://github.com/zhaoyue-ntu/qp_evaluation

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 4

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 Apollo 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 να μοντελοποιήσουμε αναλυτικούς τελεστές για NLP analysis.

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 Apollo 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] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.
[3] Doc2Vec, https://github.com/inejc/paragraph-vectors
[4] https://dtsouma.github.io/armada/

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 5

Τίτλος (Ελληνικά):	Πρόβλεψη Ποιότητας Συνόλων Δεδομένων με χρήση του Data Shapley
Τίτλος (Αγγλικά):	Predicting Dataset Quality Using Data Shapley
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

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, το Data Shapley λειτουργεί ως εξής:

- Βάση της τιμής Shapley: Στο Data Shapley, κάθε δείγμα δεδομένων θεωρείται ως «παίκτης» που συμβάλλει στην ακρίβεια ή την απόδοση ενός μοντέλου.
- Αναγνώριση σημαντικών δειγμάτων: Το Data Shapley βοηθά να προσδιοριστεί ποια δείγματα δεδομένων είναι πιο σημαντικά για την απόδοση του 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.
- Δίκαιη Κατανομή Αξίας: Το Data Shapley επιτρέπει τον δίκαιο υπολογισμό της αξίας κάθε δείγματος, κάτι που μπορεί να βοηθήσει στη δημιουργία στρατηγικών ανάλυσης και επιλογής δεδομένων, ειδικά σε περιπτώσεις όπου χρησιμοποιούνται δεδομένα από διαφορετικές πηγές με διαφορετική ποιότητα ή πληρότητα.

Στην παρούσα Διπλωματική καλούμαστε να μελετήσουμε τα ακόλουθα:

1. Επέκταση της έννοιας από ένα στοιχείο ενός συνόλου στην εύρεση μιας τιμής για ένα ολόκληρο dataset (δείτε π.χ., κάτι σχετικό εδώ [2])
2. Αποδοτική (κεντρικοποιημένη ή και κατανεμημένη) υλοποίηση του αλγορίθμου στο βήμα 1.
3. Επιβεβαίωση της ορθότητας και της ποιότητας της υλοποίησης μέσω σύγκρισης με χρήση του [3].

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] DU-Shapley: A Shapley Value Proxy for Efficient Dataset Valuation. Felipe Garrido-Lucero, Benjamin Heymann, Maxime Vono, Patrick Loiseau, Vianney Perchet. https://arxiv.org/pdf/2306.02071
[3] https://dtsouma.github.io/armada/

Θέμα 6

Τίτλος (Ελληνικά):	Σημασιολογική Κρύπτη Δεδομένων σε Περιβάλλοντα Cloud/Edge
Τίτλος (Αγγλικά):	Semantic Caching in Cloud/Edge environments
Επιβλέπων:	Δημήτριος Τσουμάκος & Laurent d'Orazio, Univ Rennes, CNRS, IRISA
Επιτροπή:	Δ. Πνευματικάτος, Γ. Γκούμας
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Κατανεμημένα Συστήματα

Σύντομη περιγραφή:

Digital transformation has led to unprecedented data generation creating both opportunities and challenges. In particular, the convergence of cloud computing and Big Data technologies have attracted increasing attention during the last decades. According to the market research organization MarketsandMarkets, the global big data market was estimated to be valued at $162.6 billion in revenue in 2021 and is projected to reach $273.4 billion by 2026, growing at a CAGR of 11.0% from 2021 to 2026.
Domains of application include the Internet, social networks, healthcare, smart cities and in particular vehicles/transports, smart agriculture, telecommunication or security monitoring.

Problem
This thesis aims to propose strategies for Big Data management with respect to multi-objective optimization, in particular considering quality.
Semantic caching makes it possible to add knowledge in a cache and thus increase the usability of its content. In the context of cloud computing in general and fog/edge computing in particular, semantic caching may be promising in order to balance the load over the global environment. Unfortunately, semantic caching may induce overhead and/or sometime be not necessary, in particular if part of the required data is absent from the cache. Nevertheless, existing results may be enough in some cases. This thesis will address this problem, trying to define some strategies in particular when, even if answers are not complete, the current quality is enough to not trigger additional processing on some data centres and/or edges.

Validation:
To validate the different contributions, experiments will be conducted on Grid5000 (https://www.grid5000.fr/), a large-scale and flexible testbed for experiment-driven research in all areas of computer science, with a focus on parallel and distributed computing including Cloud, HPC and Big Data and AI. Grid5000 (1) provides access to a large amount of resources: 15000 cores, 800 compute-nodes grouped in homogeneous clusters, and featuring various technologies: PMEM, GPU, SSD, NVMe, 10G and 25G Ethernet, InfiniBand, Omni- Path; (2) is highly reconfigurable and controllable: researchers can experiment with a fully customized software stack thanks to bare-metal deployment features, and can isolate their experiment at the networking layer; (3) provides advanced monitoring and measurement features for traces collection of networking and power consumption, providing a deep understanding of experiments; (4) is designed to support Open Science and reproducible research, with full traceability of infrastructure and software changes on the testbed and (5) gathers a community of 500+ users supported by a solid technical team.

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 7

Τίτλος (Ελληνικά):	Εκτέλεση Ασαφών Συνδέσμων σε Περιβάλλοντα Μεγάλης Κλίμακας
Τίτλος (Αγγλικά):	Fuzzy Joins in Large-Scale Environments
Επιβλέπων:	Δημήτριος Τσουμάκος & Laurent d'Orazio, Univ Rennes, CNRS, IRISA
Επιτροπή:	Δ. Πνευματικάτος, Γ. Γκούμας
Συσχετιζόμενο Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Μηχανική Μάθηση
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Κατανεμημένα Συστήματα

Σύντομη περιγραφή:

Digital transformation has led to unprecedented data generation creating both opportunities and challenges. In particular, the convergence of cloud computing and Big Data technologies have attracted increasing attention during the last decades. According to the market research organization MarketsandMarkets, the global big data market was estimated to be valued at $162.6 billion in revenue in 2021 and is projected to reach $273.4 billion by 2026, growing at a CAGR of 11.0% from 2021 to 2026.
Domains of application include the Internet, social networks, healthcare, smart cities and in particular vehicles/transports, smart agriculture, telecommunication or security monitoring.

Problem
This thesis aims to propose strategies for Big Data management with respect to multi-objective optimization, in particular considering quality.
Fuzzy joins are filters for similarity joins in a large-scale environment. Our experimental results have shown that there is a clear trade-off between the flexibility the user would like to allow in particular the semantic distance he would like to consider and the performance (the longer is the distance, the more expensive is the join process). In this thesis, the goal would be to include the concept quality into the similarity join process.

Validation:
To validate the different contributions, experiments will be conducted on Grid5000 (https://www.grid5000.fr/), a large-scale and flexible testbed for experiment-driven research in all areas of computer science, with a focus on parallel and distributed computing including Cloud, HPC and Big Data and AI. Grid5000 (1) provides access to a large amount of resources: 15000 cores, 800 compute-nodes grouped in homogeneous clusters, and featuring various technologies: PMEM, GPU, SSD, NVMe, 10G and 25G Ethernet, InfiniBand, Omni- Path; (2) is highly reconfigurable and controllable: researchers can experiment with a fully customized software stack thanks to bare-metal deployment features, and can isolate their experiment at the networking layer; (3) provides advanced monitoring and measurement features for traces collection of networking and power consumption, providing a deep understanding of experiments; (4) is designed to support Open Science and reproducible research, with full traceability of infrastructure and software changes on the testbed and (5) gathers a community of 500+ users supported by a solid technical team.

ΠΟΛΥΤΕΧΝΕΙΟΥΠΟΛΙΣ ΖΩΓΡΑΦΟΥ, ΗΡΩΩΝ ΠΟΛΥΤΕΧΝΕΙΟΥ 9, ΖΩΓΡΑΦΟΥ 157 80
ΤΗΛ: 210-7722515, FAX: 210-7721292
dtsouma@mail.ntua.gr
http://www.cslab.ntua.gr/~dtsouma

Θέμα 8

Τίτλος (Ελληνικά):	Επεξεργασία Ροών Δεδομένων με LSM Δέντρα Προσαρμοσμένα στο Υλικό
Τίτλος (Αγγλικά):	Hardware-Conscious LSM trees for Stream Processing
Επιβλέπων:	Δημήτριος Τσουμάκος
Επιτροπή:	Δ. Πνευματικάτος, Γ. Γκούμας
Συσχετιζόμενο Μάθημα:	Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων
Απαραίτητες Γνώσεις:	Βάσεις Δεδομένων, Λειτουργικά Συστήματα
Επιθυμητές Γνώσεις:	Προχωρημένα Θέματα ΒΔ, Κατανεμημένα Συστήματα

Σύντομη περιγραφή:

Τα LSM δέντρα είναι μια πολυ-επίπεδη δομή, η οποία επιταχύνει σημαντικά το χρόνο εισαγωγής/ενημέρωσης δεδομένων επιβάλλοντας ένα επιπλέον κόστος κατά την επερώτηση. Η μεγάλη ταχύτητα που προσφέρουν κατά την εγγραφή, τα έχει καθιερώσει ως την de facto τεχνολογία για τη διαχείριση της κατάστασης των τελεστών (operators' state) στα συστήματα επεξεργασίας ροών, όπου συνήθως παρατηρείται πολύ υψηλός ρυθμός άφιξης νέων δεδομένων.

Παραδοσιακά, η αρχιτεκτονική ενός LSM ορίζει ότι το πρώτο επίπεδο του δέντρου βρίσκεται στην κύρια μνήμη ενώ όλα τα υπόλοιπα στο δίσκο. Παρόλα αυτά, η συχνή πρόσβαση στο δίσκο εισάγει μεγάλες καθυστερήσεις κι εν τέλει μπορεί να «σκοτώσει» την επίδοση ολόκληρου του stream processing συστήματος.

Στην διπλωματική αυτή θα διερευνήσουμε πώς μπορούμε να βελτιστοποιήσουμε τα συστήματα επεξεργασίας ροών, ενσωματώνοντας διαφορετικές τεχνολογίες αποθήκευσης (πχ, DRAM, NVMe, SSD) στο ίδιο LSM.

Πιο συγκεκριμένα, ο στόχος της διπλωματικής είναι:

1. Μελέτη των χαρακτηριστικών της επίδοσης ενός LSM δέντρου για διαφορετικές τεχνολογίες αποθήκευσης
2. Σχεδίαση αλγορίθμων που κάνουν βέλτιστη χρήση των διαθέσιμων μέσων αποθήκευσης
3. Υλοποίηση και αξιολόγηση των αλγορίθμων σε ένα state-of-the-art κατανεμημένο σύστημα επεξεργασίας ροών δεδομένων (πχ, Apache Flink).

Ενδεικτική Βιβλιογραφία:

1. Luo, Chen, and Michael J. Carey. "LSM-based storage techniques: a survey." The VLDB Journal 29.1 (2020): 393-418.
2. Chen, Hao, et al. "{SpanDB}: A fast,{Cost-Effective}{LSM-tree} based {KV} store on hybrid storage." 19th USENIX Conference on File and Storage Technologies (FAST 21). 2021.
3. Athanassoulis, Manos, et al. "MaSM: efficient online updates in data warehouses." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011.
4. Lu, Lanyue, et al. "Wisckey: Separating keys from values in ssd-conscious storage." ACM Transactions On Storage (TOS) 13.1 (2017): 1-28.
5. https://flink.apache.org/