Θέματα Διπλωματικών Εργασιών 2022-2023

Υπεύθυνος αν. καθ. Μακρής Χρήστος

Διαδικασία ανάθεσης

Όσοι φοιτητές ενδιαφέρονται για κάποια (ή κάποιες) από τις παρακάτω θεματικές περιοχές μπορούν να στείλουν e-mail στη διεύθυνση makri@ceid.upatras.gr (με τρεις επιλογές) μέχρι 2.10.2022. Στη συνέχεια και μέχρι 9.10.2022 θα ενημερωθείτε για την ανάθεση. Οι ενδιαφερόμενοι, καλό θα είναι να επισυνάπτουν στο email τους και ένα ηλεκτρονικό αντίγραφο της καρτέλας τους. Για όλες τις διπλωματικές είναι χρήσιμο αν και όχι υποχρεωτικό (λαμβάνεται όμως υπόψιν κατά την αξιολόγηση των αιτήσεων) οι φοιτητές να έχουν περάσει τα μαθήματα: **Δομές Δεδομένων, Εισαγωγή στους Αλγορίθμους, Ανάκτηση Πληροφορίας, Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης.**

ΕΝΗΜΕΡΩΣΗ: ΟΛΑ ΤΑ ΘΕΜΑΤΑ ΕΧΟΥΝ ΔΟΘΕΙ. ΤΟ ΑΡΧΕΙΟ ΕΙΝΑΙ ΑΝΑΡΤΗΜΕΝΟ ΓΙΑ ΕΝΗΜΕΡΩΣΗ,

1. Εφαρμογή τεχνικών μηχανικής μάθησης για βελτίωση απόδοσης σε μοντέλα ανάκτησης πληροφορίας με χρήση γραφημάτων.

Στόχος της εργασίας αυτής αποτελεί η διερεύνηση της εφαρμογής τεχνικών που χρησιμοποιούνται σε σύγχρονα μοντέλα ανάκτησης πληροφορίας σε θέματα επέκτασης ερωτημάτων - κειμένων , υπολογισμού ή και επαναυπολογισμού των διανυσμάτων που παράγονται κατά την εφαρμογή της επέκτασης του Set-Based μοντέλου με γραφήματα. Εντάσσοντας, λοιπόν, τεχνικές μηχανικής μάθησης σε μοντέλα ανάκτησης δίνεται η δυνατότητα μείωσης υπολογιστικής πολυπλοκότητας με ελάχιστο κόστος στην ποιότητα της ανάκτησης.

Ποιοτικά, θα εκτιμηθεί η μεταβολή στην ακρίβεια της ανάκτησης καθώς και η επίδραση στην χρονική πολυπλοκότητα του μοντέλου συγκριτικά με τις μέχρι τώρα υλοποιήσεις.

Αναφορές

1. Nogueira, Rodrigo et al. “Document Expansion by Query Prediction.” ArXiv abs/1904.08375 (2019): n. pag.
2. Nogueira, Rodrigo, and Kyunghyun Cho. "Passage Re-ranking with BERT." arXiv preprint arXiv:1901.04085 (2019).
3. Mallia, Antonio, et al. "Faster Learned Sparse Retrieval with Guided Traversal." arXiv preprint arXiv:2204.11314 (2022).
4. Mallia, Antonio, et al. "Learning passage impacts for inverted indexes." Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021.
5. Joel Mackenzie, Zhuyun Dai, Luke Gallagher, and Jamie Callan. 2020. Efficiency Implications of Term Weighting for Passage Retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, NY, USA, 1821–1824.
https://doi.org/10.1145/3397271.3401263
6. Kalogeropoulos, NR., Doukas, I., Makris, C., Kanavos, A. (2020). A Graph-Based Extension for the Set-Based Model Implementing Algorithms Based on Important Nodes. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 585. Springer, Cham. https://doi.org/10.1007/978-3-030-49190-1_13

2. Εφαρμογή σημασιολογικών εννοιών για δημιουργία ανεστραμμένων αρχείων

Η παρούσα διπλωματική εργασία έχει ως στόχο την ένταξη σημασιολογικών εννοιών στην διαδικασία ευρετηρίασης μοντέλων ανάκτησης πληροφορίας που εκμεταλλεύονται την ύπαρξη γραφημάτων δίνοντας έμφαση στην επέκταση του Set-Based μοντέλου με γραφήματα (G.S.B) [3]. Αρχικά,

σημασιολογική επιρροή θα εκφραστεί με την χρήση word [4] και Node [5] embeddings και θα ενταχθεί στη συνέχεια στο ευρετήριο, έχοντας ως υπόθεση ότι ένας όρος του ευρετηρίου δεν υποχρεούται να υπάρχει στο κείμενο για να μπορεί να θεωρηθεί ότι τον περιέχει. Για τον λόγο αυτό, για κάθε όρο του ευρετηρίου, η λίστα κείμενων θα παράγεται από την σχετικότητα τους στον διανυσματικό χώρο που παράχθηκε στο προηγούμενο βήμα. Η έννοια της σχετικότητας θα εκφραστεί με την μορφή της ευκλείδειας απόστασης τους στον χώρο αυτόν. Τέλος, επειδή ο αριθμός των κειμένων στη λίστα κάθε όρου θα είναι αρκετά μεγάλος θα πρέπει με την χρήση locality Sensitive Hashing (LSH) [6] να επιλεχθούν τα πρώτα k κείμενα. Ο χώρος δύνανται να περιλαμβάνει τόσο όρους όσο και τα κείμενα της συλλογής [1].

Πειραματικά, λοιπόν, θα πρέπει να εκτιμηθεί η επίδραση στην απόδοση και την ποιότητα της ανάκτησης συγκρίνοντας τα νέα μοντέλα με το απλό Set-Based μοντέλο. Η ποιότητα της ανάκτησης θα εκφραστεί ως τον αριθμό των ερωτημάτων που το εκάστοτε μοντέλο έχει μεγαλύτερη ακρίβεια από το Set–Based [2], ενώ η απόδοση από την χρονική και χωρική πολυπλοκότητα των νέων αλγορίθμων.

Αναφορές

1. Fatemeh Lashkari, Ebrahim Bagheri, Ali A. Ghorbani, Neural embedding-based indices for semantic search, Information Processing & Management, Volume 56, Issue 3, 2019, Pages 733-755,ISSN 0306-4573,https://doi.org/10.1016/j.ipm.2018.10.015.
(https://www.sciencedirect.com/science/article/pii/S0306457318302413)

2. Bruno Pôssas, Nivio Ziviani, Wagner Meira, and Berthier Ribeiro-Neto. 2002. Set-based model: a new approach for information retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02). Association for Computing Machinery, New York, NY, USA, 230–237. https://doi.org/10.1145/564376.564417

3. Kalogeropoulos, NR., Doukas, I., Makris, C., Kanavos, A. (2020). A Graph-Based Extension for the Set-Based Model Implementing Algorithms Based on Important Nodes. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 585. Springer, Cham. https://doi.org/10.1007/978-3-030-49190-1_13

4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

5. Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855-864. 2016.

6. Andoni, Alexandr, and Piotr Indyk. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions." In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pp. 459-468. IEEE, 2006.

3. Υλοποίηση επιλεγμένων τεχνικών κατηγοριοποίησης σε Apache Spark

Σκοπός της διπλωματικής είναι η υλοποίηση επιλεγμένων τεχνικών κατηγοριοποίησης (και πιο συγκεκριμένα σε μία αρχική προσέγγιση του κατηγοριοποιητή Bayesian Network [1]) σε ένα κατανεμημένο περιβάλλον ικανό να υποστηρίξει την επεξεργασία δεδομένων μεγάλου όγκου (big data). Πιο συγκεκριμένα, ως περιβάλλον υλοποίησης έχει επιλεχθεί το Apache Spark [2] και ως γλώσσα υλοποίησης δίνεται η δυνατότητα επιλογής μεταξύ Python και Scala. Μετά την περάτωση της υλοποίησης, ο αλγόριθμος θα αξιολογηθεί για την ποιότητα κατηγοριοποίησης που προσφέρει σε διάφορα σύνολα δεδομένων και σε αντιπαραβολή με άλλους αλγόριθμους που παρέχει η βιβλιοθήκη MLlib του Apache Spark. Επιπρόσθετα θα επιχειρηθεί να αξιολογηθεί η ικανότητα κλιμάκωσης που πετυχαίνει η συγκεκριμένη υλοποίηση.

Για την υλοποίηση της διπλωματικής απαιτούνται γνώσης εξόρυξης δεδομένων ενώ επιθυμητή είναι και η εμπειρία πάνω σε κατανεμημένα συστήματα.

[1] https://link.springer.com/article/10.1023/A:1007465528199
[2] https://spark.apache.org/

4. Σύστημα Συστάσεων ως Επέκταση Φυλλομετρητή

Σκοπός της διπλωματικής είναι η δημιουργία ενός πρόσθετου για φυλλομετρητές το οποίο κάθε φορά που ο χρήστης περιηγείται σε μια ιστοσελίδα θα μπορεί να ανακαλύπτει τους υπερσυνδέσμους που συμπεριλαμβάνονται σε αυτή, να τους ακολουθεί και να αξιολογεί κατά πόσο θα τον ενδιέφεραν. Για να πετύχει τον σκοπό αυτό το πρόσθετο θα αποτελείται από δύο μέρη. Το πρώτο μέρος πρόκειται για

ένα frontend το οποίο θα αναπτυχθεί σε JavaScript και θα τρέχει στον φυλλομετρητή του χρήστη με σκοπό να συγκεντρώνει τις προτιμήσεις του (θα δίνει τη δυνατότητα στον χρήστη να βαθμολογεί τις σελίδες που επισκέπτεται). Στη συνέχεια αυτά τα δεδομένα θα συγκεντρώνονται από ένα Machine Learning API υλοποιημένο σε Python το οποίο θα έχει διπλή λειτουργία:

- θα εκπαιδεύει ένα μοντέλο για τον χρήστη χρησιμοποιώντας τα χαρακτηρισμένα δεδομένα
- θα χρησιμοποιεί το μοντέλο αυτό για να μαντέψει τη βαθμολογία που θα έβαζε ο χρήστης σε μια καινούρια σελίδα

Η εκπόνηση της συγκεκριμένης διπλωματικής εργασίας απαιτούνται γνώσεις εξόρυξης δεδομένων και προγραμματισμού στον Παγκόσμιο Ιστό.

5. Σύστημα ανίχνευσης ταχύτητας μπάλας με Μηχανική Μάθηση

Σκοπός της διπλωματικής αυτής είναι να κατασκευαστεί μια εφαρμογή που θα μπορεί να λειτουργεί ως σύστημα μέτρησης ταχύτητας της μπάλας σε κάποιο άθλημα (radar gun). Η εφαρμογή θα πρέπει να τρέχει σε κάποια κινητή συσκευή και να αξιοποιεί την κάμερά που αυτή προσφέρει έτσι ώστε να καταγράφει σε video την πορεία της μπάλας, να την εντοπίζει στον χώρο και να αξιοποιεί κάποια σταθερά σημεία του γηπέδου (όπως είναι οι γραμμές του) για να μπορεί να καταλαβαίνει την απόσταση που διένυσε μέσα σε ένα συγκεκριμένο χρονικό διάστημα. Για να μπορέσει να πετύχει τα παραπάνω η εφαρμογή θα πρέπει να αξιοποιεί τεχνικές μηχανικής μάθησης οι οποίες εφαρμόζονται στο πεδίο της υπολογιστικής όρασης (π.χ. Convolutional Neural Networks).
Για την υλοποίηση της διπλωματικής απαιτούνται γνώσης εξόρυξης δεδομένων και προγραμματισμού κινητών συσκευών ενώ επιθυμητή είναι και η εμπειρία πάνω σε θέματα υπολογιστικής όρασης.

6. Κατανεμημένη Υλοποίηση Τεχνικών Νευρωνικών Δικτύων Βαθιάς Αρχιτεκτονικής (Deep Neural Networks) (DNN) με χρήση Graph Embeddings για Επισημείωση Αδόμητου Κειμένου με Οντότητες Wikipedia

Οι αλγοριθμικές τεχνικές νευρωνικών δικτύων βαθιάς αρχιτεκτονικής (Deep Learning) αποτελούν υποσύνολο αλγορίθμων μηχανικής μάθησης που επιχειρούν τη μοντελοποίηση μοτίβων και αφαιρέσεων υψηλού επιπέδου στα δεδομένα χρησιμοποιώντας βαθιά πολυεπίπεδα γραφήματα αποτελούμενα από γραμμικούς και μη γραμμικούς μετασχηματισμούς. Η επισημείωση αδόμητου κειμένου με υποκείμενη εννοιολογική πληροφορία από κάποια οντολογία, είναι σημαντικό βήμα προεπεξεργασίας σειράς εργασιών στα πεδία της ανάκτησης πληροφορίας, τεχνητής νοημοσύνης, μηχανικής μάθησης, διαχείρισης δεδομένων, κλπ. Η δημιουργία των οντοτήτων (άρθρα) και η επεξεργασία της γνώσης στη Wikipedia συγκλίνει σε κοινά αποδεκτές λεκτικές περιγραφές εννοιών που μπορεί να θεωρηθεί ότι συνθέτουν μια οντολογία. Η διαδικασία επισημείωσης κειμένου με οντότητες Wikipedia περιλαμβάνει την αναγνώριση κυρίαρχων εννοιών ενός τμήματος κειμένου, και η αντιστοίχισή τους με άρθρο αντίστοιχο του εννοιολογικού τους περιεχόμενο στο εκάστοτε πλαίσιο συμφραζομένων.
Καθώς το κλασικό bag of words μοντέλο και τα word embeddings κρίνονται ανεπαρκή στο ολοένα και πιο απαιτητικό πεδίο των εφαρμογών τεχνητής νοημοσύνης, η αξιοποίηση graph embeddings επιτρέπει τη μοντελοποίηση πληροφορίας με μεγάλη πυκνότητα, καθιστώντας ακόμα πιο αποδοτική την αξιοποίηση νευρωνικών δικτύων βαθειάς αρχιτεκτονικής για το πρόβλημα.
Η εργασία περιλαμβάνει τη μελέτη της σχετικής βιβλιογραφίας και εν συνεχεία σχεδιασμό και ανάπτυξη καινοτόμου αλγορίθμου μηχανικής μάθησης σε κατανεμημένα περιβάλλοντα μεγάλης κλίμακας (επαυξημένες δυνατότητες οριζόντιας κλιμάκωσης σε υποδομές νεφοϋπολογιστικής). Η υλοποίηση της εργασίας περιλαμβάνει πειράματα σε μεγάλου όγκου δεδομένα αδόμητου κειμένου.

Βιβλιογραφία

Palash Goyal, Emilio Ferrara, Graph embedding techniques, applications, and performance: A survey, Knowledge-Based Systems, Volume 151, 2018, Pages 78-94,
DOI: https://doi.org/10.1016/j.knosys.2018.03.022

Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural Deep Network Embedding. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1225–1234. DOI:

https://doi.org/10.1145/2939672.2939753

Επιβλέπων: Μακρής Χρήστος
Συνεπιβλέπων: Σίμος Μιχαήλ - Άγγελος (υποψήφιος Διδάκτωρ)

7. **Χρήση χρονοσειρών για την αποδοτική αναπαράσταση και διερεύνηση της συμπεριφοράς των χρηστών στα κοινωνικά δίκτυα.**

Αντικείμενο της διπλωματικής εργασίας είναι η μελέτη τεχνικών Μηχανικής Μάθησης σε χρονοσειρές από δημοσιεύσεις χρηστών σε δημοφιλή κοινωνικά δίκτυα, με σκοπό την καλύτερη κατανόηση της φύσεως των χρηστών με βάση της συμπεριφορά τους σε μοα επωνυμία. Οι τεχνικές που θα ακολουθηθούν θα είναι αλγόριθμοι ομαδοποίησης, κατηγοριοποίησης, και βαθιάς μηχανικής μάθησης συνδυαζόμενες με σημασιολογικές τεχνικές (LDA) και μετρικές καταγραφής της συμπεριφοράς χρηστών σε επωνυμίες.

Βιβλιογραφία

Eleanna Kafeza, Christos Makris, Gerasimos Rompolas, Feras N. Al-Obeidat: Behavioral and Migration Analysis of the Dynamic Customer Relationships on Twitter. Inf. Syst. Frontiers 23(5): 1303-1316 (2021)

Gerasimos Rompolas, Konstantina Karavoulia: The Use of the Twitter Graphs for Analysing User Emotion for Business. CIKM Workshops 2021

Eleanna Kafeza, Christos Makris, Gerasimos Rompolas: Exploiting Time Series Analysis in Twitter to Measure a Campaign Process Performance. SCC 2017: 68-75

8, Τίτλος: Χρήση τεχνικών ανάθεσης αναγνωριστικών σε κείμενα για αποδοτική συμπίεση ανεστραμμένων αρχείων και απάντηση ερωτήσεων τομής σε όρους

Περιγραφή

Πολλές έρευνες έχουν μελετήσει πώς να βελτιστοποιηθούν οι δομές ανεστραμμένου ευρετηρίου στις μηχανές αναζήτησης μέσω της κατάλληλης τεχνικής ανάθεσης αναγνωριστικών σε κείμενα. Αυτή η προσέγγιση προτάθηκε αρχικά για καλύτερη συμπίεση ανεστραμμένων αρχείων, αλλά είναι δυνατόν επίσης να οδηγήσει σε σημαντική επιτάχυνση συνδυαστικών ερωτημάτων
Αντικείμενο της συγκεκριμένης εργασίας είναι η μελέτη των σχετικών αλγοριθμικών σχημάτων που έχουν προταθεί στη βιβλιογραφία η πειραματική σύγκριση αυτών και η σχεδίαση νέων σχημάτων .

Βιβλιογραφία

Qi Wang, Torsten Suel: Document Reordering for Faster Intersection. Proc. VLDB Endow. 12(5): 475-487 (2019)

Josh Attenberg, Torsten Suel: Scalable techniques for document identifier assignment ininverted indexes. WWW 2010: 311-320