**ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ**

**ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ**

**ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ**

**Καντηλιεράκης Γεώργιος**
**Μεταπτυχιακός Φοιτητής**

**Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης**
**Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τζίτζικας**

**Τρίτη, 21/01/2020, 10:00**
**Αίθουσα B106, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης**

# " Αναζήτηση μέσω Λέξεων-Κλειδιών επί RDF Δεδομένων Χρησιμοποιώντας Εγγραφο-κεντρικά Συστήματα Ανάκτησης Πληροφοριών"

## ΠΕΡΙΛΗΨΗ

Υπάρχουν χιλιάδες σύνολα δεδομένων που δημοσιεύονται σύμφωνα με τις αρχές των Συνδεδεμένων Δεδομένων (Linked Data) και του Σημασιολογικού Ιστού (Semantic Web). Πολλά από αυτά, όντας οργανωμένα σε RDF, βρίσκονται είτε σε Γνωσιακές Βάσεις ανοικτού τομέα (π.χ. DBpedia, Wikidata) είτε σε συλλογές κλειστού τομέα (π.χ. DrugBank, MarineTLO) και η εξερεύνησή τους είναι εφικτή μόνο μέσω συστημάτων πλοήγησης και δομημένων γλωσσών όπως η SPARQL. Ωστόσο, οι τεχνικές αυτές είναι σύνθετες, στερούνται ευελιξίας και συνήθως απαιτούν από κάποιον γνώση της οντολογίας που περιγράφει τα δεδομένα. Αυτό έχει ως αποτέλεσμα να καταλήγουν να αξιοποιούνται μόνο από ειδικούς χρήστες.

Η αναζήτηση μέσω λέξεων-κλειδιών (keyword-search) είναι η πιο ευρέως χρησιμοποιούμενη μέθοδος αναζήτησης καθώς είναι φιλική προς τον χρήστη και προσφέρει άμεση πρόσβαση στο περιεχόμενο, ενώ παράλληλα διατηρεί μεγάλη εκφραστικότητα. Τα Συστήματα Ανάκτησης Πληροφοριών (Information Retrieval Systems) είναι σχεδιασμένα για την αποτελεσματική αναζήτηση λέξεων-κλειδιών πάνω από μεγάλο όγκο εγγράφων κειμενικής πληροφορίας. Για αυτόν τον σκοπό, υπάρχουν

διαθέσιμες διάφορες εξαιρετικά αποτελεσματικές και αποδοτικές μηχανές αναζήτησης. Ένα τέτοιο παράδειγμα είναι η Elasticsearch, μια κατανεμημένη μηχανή αναζήτησης κειμένου, η οποία παρέχει δυνατότητα κλιμακώσιμης αναζήτησης σε οποιοδήποτε είδος πληροφοριών κειμένου.

Σε αυτήν την εργασία αναπτύξαμε μία υλοποίηση για αναζήτηση μέσω λέξεων-κλειδιών πάνω από RDF δεδομένα, προσαρμόζοντας τις παραδοσιακές τεχνικές ανάκτησης πληροφορίας (IR) για την ευρετηρίαση και την ανάκτηση. Συγκεκριμένα, δοκιμάζουμε τρόπους με τους οποίους μια κυρίαρχη στην αγορά μηχανή αναζήτησης, όπως η ElasticSearch, μπορεί να χρησιμοποιηθεί για την ευρετηρίαση RDF δεδομένων και την παροχής αναζήτησης λέξεων-κλειδιών σε αυτά. Παρέχουμε μια ανάλυση των διαφορετικών προσεγγίσεων που ακολουθήσαμε για να αντιμετωπίσουμε τις προκλήσεις της ευρετηρίασης και της ανάκτησης δομημένης πληροφορίας και την αξιοποίηση των δυνατοτήτων που μας δίνει ο RDF γράφος. Η απάντηση του συστήματος αποτελείται από ταξινομημένες RDF τριπλέτες. Επίσης, παρέχουμε πολιτικές για την κατάταξη των διαφορετικών οντοτήτων που περιέχονται στις τριπλέτες προκειμένου να υποστηριχθεί και ο στόχος της κατάταξης οντοτήτων (entity-ranking). Τα αποτελέσματα της αξιολόγησης των διαφορετικών προσεγγίσεων μας περιλαμβάνουν (α) την αποδοτικότητα της ευρετηρίασης και της ανάκτησης και (β) την ποιότητα της ανάκτησης. Δοκιμάζουμε την αποτελεσματικότητα του συστήματός μας αξιολογώντας τη συνάφεια των οντοτήτων που κατασκευάζουμε πάνω από τη συλλογή DBpedia-Entity, σχεδιασμένη για την αναζήτηση οντοτήτων μέσω της γνωσιακής βάσης DBpedia και συγκρίνοντας τα αποτελέσματά μας με διάφορα συναφή συστήματα. Στα αποτελέσματά μας παρουσιάζουμε την αποδοτικότητα της προτεινόμενης φιλικής προς τον χρήστη προσέγγισης, η οποία εκμεταλλεύεται τα ισχυρά χαρακτηριστικά των κλιμακώσιμων μηχανών αναζήτησης, ενώ μπορεί να εφαρμοστεί πάνω από οποιοδήποτε σύνολο δεδομένων RDF χωρίς προηγούμενη γνώση του τομέα. Τα αποτελέσματα της αξιολόγησης καταδεικνύουν ότι η Elasticsearch μπορεί να υποστηρίξει αποτελεσματικά την αναζήτηση μέσω λέξεων-κλειδιών επί δεδομένων RDF, προσφέροντας αποτελεσματικότητα εφάμιλλη εκείνης των συστημάτων που έχουν δημιουργηθεί αποκλειστικά για RDF και χρησιμοποιούν οντο-κεντρικά ευρετήρια.

**Georgios Kantilierakis**

**M.Sc. Thesis**

**Computer Science Department**

**University of Crete**

**Master's Thesis Supervisor: Associate Professor, I. Tzitzikas**

**Tuesday, 21/01/2020, 10:00**

**Room B106, Computer Science Dept., University of Crete**

**“Keyword Search over RDF using Document-centric Information Retrieval Systems”**

## ABSTRACT

There are thousands of datasets published according to the principles of Linked Data and Semantic Web. Many of those datasets, organized in RDF, are maintained either in cross-domain Knowledge Bases (e.g. DBpedia, Wikidata) or domain specific repositories (e.g. DrugBank, MarineTLO), and are mainly used through navigation and structured query languages like SPARQL. However these techniques are complex, lack flexibility and possibly require a full knowledge of the underlying ontology. As a result, these datasets are exploited by expert users only.

On the other hand, keyword search is the most widely used method for searching. Keyword search is user friendly, offers instant content access, and keyword queries support a wide range of expression while being extremely flexible. Information Retrieval systems are designed for performing efficient keyword search in large data of information, usually organized as full text documents. There are various highly performant and effective state of the art search engines readily available. Such a search engine is Elasticsearch, a distributed full text search engine that provides scalable search over any kind of textual information.

In this thesis we introduce an approach for keyword-search over RDF datasets, by adapting traditional IR techniques for both indexing and retrieval. Specifically, we test how a dominant IR engine such as Elasticsearch, can be adapted for indexing RDF data and enable keyword search. We provide a systematic analysis of different approaches to cope with the challenges of indexing and retrieving structured information and exploiting the graph capabilities of RDF. The response of the system comprises ranked RDF triples. We also provide policies for ranking the different entities that are contained in these triples, in order to support the requirements of entity search.
We report evaluation results of the different approaches in terms of: (i) the efficiency of indexing and retrieval and (ii) the quality of retrieval. We test the effectiveness of our system by evaluating the relevance of the constructed entities against the DBpedia-Entity test collection, designed for entity search over the DBpedia KB and compare our results to various state of the art systems. Our results showcase the effectiveness of the proposed user friendly approach , that exploits the powerful features of scalable state of the art search engines, and can be applied in any RDF dataset, with no prior knowledge of the domain. The results show that Elasticsearch can effectively support keyword search over RDF data, offering effectiveness comparable to that of systems built from scratch for the task per se, that use entity-oriented and dataset-specic index structures.