ΠΡΟΣ

1) **Όλα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**

2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**

3) **Την Επταμελή Εξεταστική Επιτροπή**

4) **Όλα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Φαφαλιού Παύλου

Την Δευτέρα, 11 Απριλίου 2016 και ώρα 15:30 στην αίθουσα Τηλεδιάσκεψης Κ206 του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψηφίου διδάκτορος του Τμήματος Επιστήμης Υπολογιστών κ. Φαφαλιού Παύλου με θέμα:

" Αξιοποίηση Διασυνδεδεμένων Δεδομένων στην Εξερευνητική Αναζήτηση"

"Exploiting Linked Data in Exploratory Search"

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια παρατηρείται μια έκρηξη στη δημοσίευση δεδομένων στον Παγκόσμιο Ιστό, κυρίως με τη μορφή Διασυνδεδεμένων Δεδομένων (Linked Data). Ένα βασικό ερώτημα όμως είναι πως αυτός ο συνεχώς αυξανόμενος πλούτος γνώσεων μπορεί να αξιοποιηθεί από απλούς χρήστες για καλύτερη αναζήτηση πληροφοριών. Αν και τα υπάρχοντα συστήματα σημασιολογικής αναζήτησης αποκρύπτουν την πολυπλοκότητα τους αξιοποιώντας φιλικά στη χρήση μέσα αλληλεπίδρασης, δεν έχουν καταφέρει ακόμα να καλύψουν κοινές – γενικού σκοπού – ανάγκες αναζήτησης και διασύνδεσης πληροφοριών. Παράλληλα, σύμφωνα με

διάφορες μελέτες, ένα μεγάλο ποσοστό των αναζητήσεων είναι εξερευνητικού χαρακτήρα. Σε τέτοιου είδους πληροφοριακές ανάγκες οι παραδοσιακές απαντήσεις που έχουν τη μορφή γραμμικής λίστας αποτελεσμάτων συνήθως δεν είναι ικανοποιητικές.

Ο σκοπός αυτής της διατριβής είναι η παροχή προηγμένων υπηρεσιών εξερευνητικής αναζήτησης οι οποίες γεφυρώνουν το χάσμα μεταξύ των κλασσικών απαντήσεων μη σημασιολογικών συστημάτων αναζήτησης (επαγγελματικής ή μη φύσεως) και σημασιολογικής πληροφορίας εκφρασμένης με τη μορφή Ανοιχτών Διασυνδεδεμένων Δεδομένων (ΑΔΔ). Προς αυτή τη κατεύθυνση, εισάγουμε και εξετάζουμε μια προσέγγιση κατά την οποία επώνυμες οντότητες (όπως για παράδειγμα πρόσωπα, περιοχές, χημικές ουσίες, κτλ.) αξιοποιούνται ως ο συνδετικός κρίκος για την αυτόματη διασύνδεση εγγράφων (αποτελεσμάτων αναζήτησης) με δεδομένα και γνώση. Μελετούμε μια προσέγγιση όπου αυτή η – βασισμένη σε οντότητες – ενοποίηση πραγματοποιείται σε πραγματικό χρόνο, κατά τη στιγμή της αναζήτησης, χωρίς εμπλοκή του χρήστη, αλλά και χωρίς την ανάγκη προκατασκευασμένων ευρετηρίων. Αυτό επιτρέπει την παροχή «φρέσκιας» πληροφορίας, την εύκολη παραμετροποίηση αυτής της λειτουργικότητας σύμφωνα με τις ανάγκες του υποκείμενου συστήματος αναζήτησης, αλλά και την εύκολη αξιοποίηση της από τα υπάρχοντα εργαλεία ανάκτησης πληροφοριών.

Η παροχή της παραπάνω λειτουργικότητας έχει διάφορες προκλήσεις. Αρχικά, τα ΑΔΔ που είναι διαθέσιμα στον Παγκόσμιο Ιστό έχουν μεγάλο μέγεθος, είναι κατανεμημένα σε πολλές Βάσεις Γνώσεων, αυξάνονται και ενημερώνονται συνεχώς, και καλύπτουν πολλές θεματικές περιοχές. Εκ τούτου, προκύπτει η ανάγκη για ένα μοντέλο διαλειτουργικότητας που να επιτρέπει τον προσδιορισμό των οντοτήτων ενδιαφέροντος και των σχετικών σημασιολογικών δεδομένων (από διάφορες βάσεις γνώσεων). Συνάμα ο αριθμός των εξορύξιμων οντοτήτων από τα αποτελέσματα αναζήτησης μπορεί να είναι αρκετά μεγάλος και το ίδιο ισχύει για το μέγεθος της σημασιολογικής πληροφορίας που μπορεί να ανακτηθεί από τα ΑΔΔ για αυτές τις οντότητες (ήτοι το πλήθος των χαρακτηριστικών τους και των συσχετίσεών τους με άλλες οντότητες). Εκ τούτου προκύπτει η ανάγκη για μεθόδους που να μπορούν να εκτιμήσουν τις πιο σημαντικές οντότητες, καθώς και τη σημαντική σχετική σημασιολογική πληροφορία, για τα αποτελέσματα της εκάστοτε αναζήτησης.

Για την επιτυχή αντιμετώπιση των παραπάνω προκλήσεων, η διατριβή προτείνει μία διαδικασία ανάλυσης αποτελεσμάτων αναζήτησης κατά την οποία τα αποτελέσματα συνδέονται με δεδομένα και γνώσεις σε πραγματικό χρόνο, χωρίς τη μεσολάβηση του χρήστη. Για την περιγραφή των οντοτήτων ενδιαφέροντος και της σχετικής σημασιολογικής πληροφορίας προτείνεται ένα γενικό μοντέλο παραμετροποίησης Συστημάτων Εξαγωγής Οντοτήτων (ΣΕΟ), ενώ για τον ακριβή προσδιορισμό της σημασιολογίας αυτού του μοντέλου, εισάγουμε ένα RDF/S λεξιλόγιο με όνομα “Open Named Entity Extraction (NEE) Configuration Model”, το οποίο επιτρέπει ένα ΣΕΟ να περιγράψει (και να δημοσιεύει ως ΑΔΔ) τις δυνατότητες του. Για να καταστήσουμε δυνατή τη συσχέτιση του αποτελέσματος της διαδικασίας εξαγωγής οντοτήτων με τις παραμέτρους που χρησιμοποιήθηκαν, προτείνουμε μια επέκταση του μοντέλου “Open Annotation” η οποία επιτρέπει και τη δημοσίευση των αποτελεσμάτων της

εξόρυξης ως ΑΔΔ. Για την διερεύνηση της επιτευξιμότητας αυτού του μοντέλου αναπτύχθηκε το σύστημα X-Link το οποίο σε αντίθεση με τα υπάρχοντα ΣΕΟ επιτρέπει με εύκολο τρόπο τον προσδιορισμό των κατηγοριών οντοτήτων και των σημασιολογικών δεδομένων που ενδιαφέρουν την υποκείμενη εφαρμογή αξιοποιώντας μια ή περισσότερες σημασιολογικές Βάσεις Γνώσεων. Για τον εντοπισμό των πιο σημαντικών σημασιολογικών πληροφοριών που σχετίζονται με τα αποτελέσματα μιας αναζήτησης, εισάγουμε και μελετάμε μια μέθοδο κατάταξης βασισμένη στο μοντέλο Τυχαίου Περιπάτου (Random Walk) που αξιοποιεί τις εξηγμένες οντότητες και τη διασύνδεσή τους. Η αξιοποίηση αυτών των σημασιολογικών πληροφοριών γίνεται είτε με την οπτικοποίηση του σχετικού γράφου ή/και στα πλαίσια ενός πολυδιάστατου μοντέλου αλληλεπίδρασης που επιτρέπει στον χρήστη να περιορίσει τον πληροφοριακό του χώρο αυξητικά. Πέραν αυτού, η διατριβή αυτή μελέτησε την αξιοποίηση τέτοιων γράφων και για την ανακατάταξη της λίστας αποτελεσμάτων με σκοπό την βελτίωση της, συγκεκριμένα για την προώθηση εγγράφων που αν και είναι συναφή με την επερώτηση δεν είναι στις πρώτες θέσεις της κατάταξης.

Η διατριβή αναφέρει εκτενή αποτελέσματα αξιολόγησης των προτεινόμενων λειτουργιών και μεθόδων. Αναφορικά με το σύστημα X-Link, η αξιολόγηση με χρήστες έδειξε την ευκολία παραμετροποίησης, ενώ μια μελέτη περίπτωσης έδειξε την απόδοση των υποστηριζόμενων λειτουργιών του. Η συγκριτική αξιολόγηση του προτεινόμενου αλγορίθμου κατάταξης των οντοτήτων και της σχετικής σημασιολογικής πληροφορίας έδειξε ότι η προτεινόμενη προσέγγιση είναι πιο αποτελεσματική σε σχέση με άλλες μεθόδους ανακατάταξης. Αναφορικά με τον τρόπο παρουσίασης των σημαντικών οντοτήτων (και των διασυνδέσεων τους) που σχετίζονται με μία απάντηση, τα αποτελέσματα μιας αξιολόγησης που έγινε με χρήστες στην περιοχή της αναζήτησης θαλάσσιων ειδών, έδειξε ότι η πλειονότητα των συμμετεχόντων (περισσότεροι από το 70%) προτιμούν μια γραφική απεικόνιση των οντοτήτων που σχετίζονται με τα αποτελέσματα αναζήτησης, ανεξαρτήτως του τύπου επερώτησης. Η αξιολόγηση του προτεινόμενου πιθανοτικού αλγορίθμου ανακατάταξης των επιστρεφόμενων αποτελεσμάτων που έγινε με συλλογές αξιολόγησης από το TREC (Text Retrieval Conference) που αφορούν τον τομέα της ιατρικής, κατέδειξε ότι η προσέγγιση αυτή μπορεί να βελτιώσει σημαντικά τη λίστα αποτελεσμάτων προωθώντας συναφή έγγραφα σε υψηλότερες θέσεις. Τέλος η υλοποίηση και τα πειραματικά αποτελέσματα της προτεινόμενης διαδικασίας αναζήτησης κατέδειξαν την επιτευξιμότητα και την απόδοσή της, και μας επέτρεψαν συνάμα να εντοπίσουμε τους περιορισμούς της.

ABSTRACT

In recent years we have witnessed an explosion in publishing data on the Web, mostly in the form of Linked Data. An important question is how typical users, who mainly use keyword search queries, can access and exploit this constantly increasing body of knowledge. Although existing interaction paradigms in Semantic Search hide their complexity behind easy-to-use interfaces, they have not managed to cover common search needs. At the same time, according to several studies, a large number of search tasks are of exploratory nature. However, in such tasks the traditional "ranked list" approach for interacting with the retrieved results is often inadequate.

The objective of this thesis is to enable effective exploratory search services which can bridge the gap between the classic responses of non-semantic search systems (e.g., Professional Search Systems, Web Search Engines) and semantic information expressed in the form of Linked Open Data (LOD). Towards this direction, we introduce an approach in which named entities (like names of persons, locations, chemical substances, etc.) are exploited as the glue for automatically connecting documents (search results) with data and knowledge. We study an approach where this entity-based integration is performed at real-time, without any human intervention and without the need of prebuilt indexes. This allows the provision of "fresh" information, the easy configuration of this functionality according to the needs of the underlying search application, as well as its easy exploitation by existing search systems.

The provision of the aforementioned functionality is challenging. At first, the LOD that are available on the Web are big, are distributed in many knowledge bases, are increased and updated continuously, and also cover many domains. Consequently, there is the need of an interoperability model that will allow the specification of the entities of interest as well as of the related and useful semantic data. In addition, the number of extractable entities from the search results can be very high and the same is true for the amount of semantic information that can be retrieved from the LOD for these entities (i.e., the number of their attributes and of their associations with other entities). Thus, there is also the need of methods that can estimate the important (for the search context) entities, attributes and associations.

To cope with above challenges, this thesis proposes a semantic analysis process in which the search results are connected with data and knowledge at real-time without any human intervention. For describing the entities of interest, as well as the related (and useful for the application context) semantic information, we propose a generic model for configuring a Named Entity Extraction (NEE) system, while for specifying the semantics of this model, we introduce an RDF/S vocabulary, called "Open NEE Configuration Model", which allows a NEE system to describe (and publish as LOD) its entity-mining capabilities. To enable associating the result of a NEE process with an applied configuration, we propose an extension of the Open Annotation Data Model which also allows publishing the annotation results as LOD. To examine the feasibility of this model, we developed the system X-Link which, contrary to existing NEE systems, allows its easy configuration by exploiting one or more semantic Knowledge Bases. To identify the important semantic information related to the search results, we introduce and study a ranking method that is based on the Random Walk model

and which exploits the extracted entities and their connectivity. The exploitation of the selected semantic information is achieved either through the visualization of the related semantic graph and/or in the context of a faceted interaction model that allows the user to gradually restrict the search space. Besides, this thesis studied the exploitation of such graphs for re-ranking the list of retrieved results aiming to promote relevant but low-ranked hits.

The dissertation reports extensive evaluation results of the proposed functionalities and methods. As regards the system X-Link, a task-based evaluation with users showed its ease of configuration, while a case study illustrated the efficiency of the supported operations. The comparative evaluation of the proposed probabilistic scheme for ranking entities and semantic data showed that the proposed approach is more effective compared to other ranking approaches (producing more than 20% better ranking). As regards the presentation of the important entities (and of their associations), the conducted survey in a marine-related search context demonstrated that the majority of participants (more than 70%) prefer to see a graph representation of entities related to the retrieved results regardless the type of the submitted query. The evaluation of the proposed probabilistic algorithm for re-ranking the retrieved search results (using TREC datasets related to the medicine domain) showed that this approach can notably improve the list of results by promoting relevant hits in higher positions. Finally, the implementation and the experimental results of the proposed search process demonstrated its feasibility and efficiency, and also enabled us to reveal its limitations.