ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μπαριτάκης Εμμανουήλ

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Επικ. Καθηγητής Ι. Τζίτζικας

Δευτέρα, 1/2/2016, 12:00

Αίθουσα Β108, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Χρήση Διασυνδεδεμένων Δεδομένων για Εξόρυξη και Αποσαφήνιση Οντοτήτων "

ΠΕΡΙΛΗΨΗ

Με τον όρο Εξόρυξη Οντοτήτων αναφερόμαστε στη διαδικασία εντοπισμού οντοτήτων σε κείμενα και αρκετά συχνά στην σύνδεσή τους με σχετικούς (διαδικτυακούς) πόρους. Αυτή η διαδικασία είναι χρήσιμη σε πολλές εφαρμογές, όπως στην απάντηση επερωτήσεων, στην επισημείωση κειμένων, στην επεξεργασία αποτελεσμάτων αναζήτησης, κ.α. Ωστόσο, είναι αρκετά σύνηθες ένα όνομα οντότητας να αντιστοιχεί σε παραπάνω από μια κατηγορίες, λόγου χάρη ο όρος Αργεντινή μπορεί να αφορά είτε το είδος ψαριού Αργεντινή, είτε την ομώνυμη χώρα. Αυτό το πρόβλημα είναι γνωστό στη κοινότητα ως πρόβλημα της Αποσαφήνισης Οντοτήτων. Επιπρόσθετα, τα υπάρχοντα εργαλεία εντοπισμού και αποσαφήνισης οντοτήτων στερούνται μιας εύκολης και «ανοικτής» παραμετροποίησης, η οποία είναι σημαντική για τη δημιουργία εξειδικευμένων εφαρμογών. Για παράδειγμα, η υποστήριξη μιας νέας κατηγορίας οντοτήτων ή ο προσδιορισμός του τρόπου σύνδεσης των οντοτήτων με δεδομένα στο διαδίκτυο, είναι από πολύ δύσκολο έως ακατόρθωτο. Σε αυτήν την εργασία επικεντρωνόμαστε στο πως μπορούμε να εκμεταλλευτούμε τις διαθέσιμες σημασιολογικά οργανωμένες πληροφορίες, συγκεκριμένα τα Διασυνδεδεμένα

Δεδομένα (Linked Data), για να παραμετροποιήσουμε ένα σύστημα εξόρυξης οντοτήτων καθώς και για να αποσαφηνίσουμε τις ευρεθείσες οντότητες. Προτείνουμε μια οντολογία RDF/S, που ονομάζεται Open NEE Configuration Model, η οποία επιτρέπει σε μια υπηρεσία εντοπισμού οντοτήτων να περιγράφει (και να εκφράζει ως Linked Data) τις προδιαγραφές της, καθώς και να παραμετροποιείται δυναμικά. Επίσης παρουσιάζουμε το X-Link, ένα εργαλείο εξόρυξης οντοτήτων που υιοθετεί το παραπάνω μοντέλο, που σε αντίθεση με άλλα συναφή εργαλεία, επιτρέπει στον χρήστη να προσδιορίζει, με εύκολο τρόπο, τις κατηγορίες οντοτήτων που τον ενδιαφέρουν για την εφαρμογή του (εκμεταλλευόμενος τα Linked Data). Εν συνεχεία, κινούμενοι ως προς αυτή την κατεύθυνση, εμβαθύνουμε στο πρόβλημα της αποσαφήνισης οντοτήτων, και πιο συγκεκριμένα στο πρόβλημα της επιλογής της κατάλληλης κατηγορίας για κάθε ευρεθείσα οντότητα. Για τον σκοπό αυτό προτείνουμε 3 μεθόδους, με κάθε μια να προσεγγίζει το πρόβλημα από διαφορετική σκοπιά. Η πρώτη βασίζεται εξολοκλήρου στα αποτελέσματα ενός NEE εργαλείου και θεωρεί ως πιθανότερη κατηγορία εκείνη με την μεγαλύτερη συχνότητα εμφάνισης στα αποτελέσματα. Η δεύτερη επεκτείνει την πρώτη και αξιοποιεί τις σημασιολογικές συσχετίσεις μεταξύ των οντοτήτων που έχουν εντοπιστεί, χρησιμοποιώντας τις σημασιολογικές τους ιδιότητες. Θεωρεί ως πιο πιθανή κατηγορία εκείνη που αντιστοιχεί στον σημασιολογικό πόρο που είναι πιο κοντά (στο σημασιολογικό γράφο) στους υπόλοιπους που εντοπίστηκαν. Η τελευταία μέθοδος χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την κατηγοριοποίηση του εκάστοτε κειμένου σε μια συγκεκριμένη κατηγορία, έχοντας πρώτα «εκπαιδευτεί» σε μια κατάλληλη συλλογή εγγράφων. Στη συνέχεια παρουσιάζουμε τα αποτελέσματα μιας εμπεριστατωμένης συγκριτικής αξιολόγησης που χρησιμοποιεί αποτελέσματα αναζήτησης από τη μηχανή αναζήτησης Bing. Τα αποτελέσματα της αξιολόγησης μας επιτρέπουν να εντοπίσουμε τα θετικά και τα αρνητικά κάθε μεθόδου. Πιο συγκεκριμένα, αξιολογήσαμε τις μεθόδους μας πάνω σε συλλογές εγγράφων διαφορετικού μεγέθους και υπολογίσαμε την ακρίβεια τους καθώς και τον απαιτούμενο χρόνο εκτέλεσης. Μετά το πέρας των πειραμάτων καταλήξαμε στο συμπέρασμα ότι η τρίτη μέθοδος (κατηγοριοποίηση εγγράφου) λειτουργεί καλύτερα σε όλες τις περιπτώσεις εκτός αυτής που το περιεχόμενο ενός εγγράφου είναι περιορισμένο, πχ. tweets, όπου έχει σχεδόν την ίδια ακρίβεια με την δεύτερη μέθοδο.

Baritakis Emmanouil

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Assistant Professor I. Tzitzikas

Monday, 1/2/2016, 12:00
Room B108, Computer Science dept., University of Crete

"Using Linked Data for Named Entity Extraction and Disambiguation"

ABSTRACT

Named Entity Extraction (NEE) is the process of identifying entities in texts and, very commonly, linking them to related (Web) resources. This task is useful in several applications, e.g. for question answering, annotating documents, processing of search results, etc. However, it is quite common for an entity name to correspond to more than one semantic categories, e.g. Argentina may refer either to Fish Species Argentina or to Country Argentina. This is the well-known Named Entity Disambiguation (NED) problem. In addition to, existing NEE and NED tools lack an open or easy configuration although this is very important for building domain-specific applications. For example, supporting a new category of entities, or specifying how to link the detected entities with online resources, is either impossible or very laborious. In this thesis we show how we can exploit semantic information (Linked Data) at real-time for *configuring* a NEE system and *disambiguating* the mined entities. We introduce an RDF/S vocabulary, called Open NEE Configuration Model, which allows a NEE service to describe (and publish as Linked Data) its entity mining capabilities, but also to be dynamically configured. We present X-Link a NEE framework that realizes this model, and contrary to the existing tools, it allows the user to easily define the categories of entities that are interesting for the application at hand (by exploiting Linked Data). Then we focus on the problem of NED in this context, i.e. on the problem of selecting the right category for each extracted entity. To this end we introduce three methods, each approaching the problem from a different perspective. The first method is based exclusively on NEE results and selects as more probable category the one with the highest occurrence frequency. The second method moves a step forward and exploits the semantic relations between the mined entities, using their semantic resources, and returns the semantic resource that is closer to the others in the semantic graph. The last method uses machine learning algorithms for classifying the entire document into a specific category based on a train set. Then we report the results of a thorough comparative experimental evaluation using search results from Bing search engine. We evaluate the introduced methods over collections of documents of different size and we measured the achieved precision and the required time for disambiguation. The results allowed us to identify the strong and weak aspects of each method. Overall, the third method works well in most cases apart from small snippets, e.g. tweets, where it achieves almost the same precision with the second method.