Ένα εργαλείο σε Java για την ανάκτηση πληροφοριών και τον εντοπισμό ομοιοτήτων σε ακαδημαϊκές εκδόσεις με τη χρήση τεχνικών εξόρυξης δεδομένων

Εμμανουήλ Γαρουφάλλου[1], Δημήτρης Ρουσίδης[1]

και Πάνος Μπαλατσούκας[2]

[1]Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης, [2]University of Strathclyde

drousid@gmail.com , garoufallou@gmail.com , pan-bal@hotmail.com

Περίληψη

Μια πληθώρα νέων δεδομένων και πληροφοριών προσθέτονται κάθε μέρα στον ακαδημαϊκό τομέα. Ο αριθμός των ακαδημαϊκών άρθρων αυξάνεται εκθετικά κάθε χρόνο καθιστώντας αυτή την τεράστια δεξαμενή γνώσης προβληματική ως προς την εξερεύνηση και επεξεργασία της. Προτείνεται ένα αυτοματοποιημένο εργαλείο εξόρυξης δεδομένων γραμμένο σε Java το οποίο βρίσκεται στη φάση της υλοποίησης και θα εντοπίζει σημασιολογικές ομοιότητες μεταξύ ακαδημαϊκών συγγραμμάτων. Παρέχοντας στο εργαλείο έναν τίτλο άρθρου Α, θα μπορεί να αναγνωρίζει αποτελεσματικά άλλους τίτλους άρθρων τα οποία μοιράζονται κοινώς ένα ή παραπάνω κριτήρια, όπως συγγραφείς, αναφορές, λέξεις κλειδιά, θεματολογία, εκδότες, ημερομηνίες καθώς και μεθοδολογίες και θεματικές υποενότητες. Το εργαλείο θα έχει τη δυνατότητα ανακάλυψης κρυμμένων μοτίβων, κανόνων σχέσης (association rules), κατηγοριοποίησης (classification) και συσταδοποίησης-ομαδοποίησης (clustering) στις ακαδημαϊκές εκδόσεις καθώς και απεικόνιση όλων αυτών των πληροφοριών. Προκειμένου να διευκολυνθεί η ανακάλυψη των μεθοδολογιών, το προτεινόμενο εργαλείο θα είναι σε θέση να δημιουργεί μια βάση δεδομένων ορολογιών και υπο-ορολογιών κυρίως μέσω της ανάλυσης των ευρετηρίων ηλεκτρονικών βιβλίων (e-books). Επιπροσθέτως, αυτή η βάση δεδομένων θα είναι διαθέσιμη διαδικτυακά οπού ουσιαστικά θα δημιουργηθεί ένα αποθετήριο ορολογιών και υπο-ορολογιών. Ο κύριος σκοπός της δημιουργίας αυτού του εργαλείου είναι η διευκόλυνση της ακαδημαϊκής μελέτης και έρευνας και η αύξηση της ικανότητας άντλησης πληροφοριών και συσχετισμών από τα ακαδημαϊκά περιεχόμενα.

Λέξεις κλειδιά: Μεταδεδομένα, Συστήματα υποστήριξης αποφάσεων, Αναζήτηση πληροφοριών, Ανάκτηση πληροφοριών, Εξόρυξη γνώσης, Εξόρυξη πληροφοριών

A Java tool for information retrieval and similarity measurement of scholarly publications based on the use of data mining techniques

Abstract

As the number of scholarly output increases exponentially, the cognitive process of identifying and evaluating the relevance of retrieved information becomes difficult both in terms of time and cognitive effort spent by the users. The purpose of this paper is to propose an automated data mining tool written in Java. The tool, which is currently under implementation, should identify semantic similarities amongst academic publications. For example, given an article X, the tool will be able to identify effectively other titles that share commonly one or more criteria such as authors, references, keywords, subjects, publishers, dates, as well as methodologies and thematic sub-sections. In order to facilitate the discovery of similarities between articles using less conventional metadata, such as methodologies and data analysis techniques used, the proposed tool will be able to create a database of terminologies and sub-terminologies mainly through the analysis of the indexes of e-books. In addition, this database will be available online where a repository of terminologies and sub-terminologies will be created. Also, the tool will be able to show hidden patterns, through the use of association rules and clustering algorithms, and create dynamic visualisations of a given set of scholarly articles. It is anticipated that this tool will facilitate the academic study and research and enhance the ability of extracting information and correlations from academic content.

Keywords: Metadata, Decision support systems, Information search, Information retrieval, Knowledge extraction, Data mining