Key Phrase to Text Similarity, Clustering, and Interpretation in Hierarchical Ontologies

Boris Mirkin

Applied Mathematics and Informatics, National Research University Higher School of Economics Moscow
Computer Science and Information Systems, Birkbeck University of London
BMirkin@hse.ru

Abstract. Scoring similarity between key phrases and unstructured texts is an issue which is important in both information retrieval and text analysis. Researchers from the two fields use different scoring functions, although clear delineation between the two still is lacking. We use suffix tree based score expressing the average conditional probability of a symbol in a common substring. Usually, a domain taxonomy serves as the source of key-phrases. Given a set of entities, such as texts or projects or working groups, one can derive clusters of key-phrases using key-phrase-to-entity scores. The clusters represent common themes in the meaning of texts or in activities of working groups. To interpret them, the domain ontology should be used. If the ontology is a rooted tree, a lifting method is proposed to find the most parsimonious interpreting head subject(s), up to a few gaps and offshoots. Some applications and application issues are considered. The work is being conducted jointly with T. Fenner (London), S. Nascimento (Lisbon) and E. Chernyak (Moscow).