החוג לניהול מידע וידע

The Department of Information and Knowledge Management

הפקולטה לניהול
Faculty of Management

Who are you? Identifying Unique Text Signatures of Microblog Authors

Hagit Ben Shoshan

Advisor: Dr. Osnat (Ossi) Mokryn

The use of social networking technologies enhances microblogs authors' ability to express a large amount of opinions through community-written texts. In order to generate author's unique style, we suggest representing author's corpus as weighted ranked list. Measuring that author distance from the general domain corpus should be done by calculating distance between two ranked lists. When evaluating related works in the field of distance between ranked lists, we couldn't find a suitable method that match our criteria.

The method should provide a different weight for each element in the list and the ability to compare two lists that are not identical in length (non-conjoint), hence there is partial overlap between compared elements in two lists. We like to consider weight of missing elements in the list, and give a greater importance to inconsistencies at the top of the list, than in the bottom.

Most of the popular algorithms in this domain developed to calculate distance between search engine results, or comparing weighted equal length lists.

In this work, we devised a new method for calculating distance of an author (category) from the total authors (DVR) based on words used by this author (elements) and their relative weight in its text (probabilities).

We implement this method, using Big Data technologies, while processing large amount of data. We scanned the complete data sets instead of using samples or estimations.

The generalization of our proposed method, makes it suitable for additional research areas, apart of text processing. We will explain and demonstrate the usability of this method in different domains.