A probabilistic-entropy approach of finding thematically similar documents with creating context-semantic graph for investigating evolution of society opinion

Moloshnikov I.A.[1], Sboev A.G.[2], Rybka R.B.[3], Gydovskikh D.V.[4]
[1]NRC "Kurchatov Institute", ivan-rus@yandex.ru
[2]NRC "Kurchatov Institute", National Research Nuclear University MEPhI, sag111@mail.ru
[3]NRC "Kurchatov Institute", rybkarb@gmail.com
[4]NRC "Kurchatov Institute", dmitrygagus@gmail.com

Abstract

An algorithm of finding documents on a given topic based on a selected reference collection of documents along with creating context-semantic graph for visualizing themes in search results is presented. The algorithm is based on integration of set of probabilistic, entropic, and semantic markers for extractions of weighted key words and combinations of words, which describe the given topic. Test results demonstrate an average precision of 99% and the recall of 84% on expert selection of documents. Also developed special approach to constructing graph on base of algorithms extract key phrases with weights. It gives the possibility to demonstrate a structure of subtopics in large collections of documents in compact graph form.