COMPARISON OF CLUSTERING BASED ON SEARCH ENGINE DATASET USING TANAGRA AND WEKA

PREETI BANSAL & MD. EZAZ AHMED

Department of Computer Science, ITM University, Gurgaon, Haryana, India

ABSTRACT

The paper introduced the concept of clustering on a particular dataset of some influencing factors in Search Engine Optimization using Tanagra and Weka data mining tool and find the values of factors that affect relevance. As the number of available Web pages grows, it is become more difficult for users to find documents relevant to their interest. To a search engine, relevance means more than simply finding a page with the right words. In the early days of the web, search engines didn't go much further than this simplistic step, and their results suffered as a consequence. Thus, through evolution, smart engineers at the engines devised better ways to find valuable results that searchers would appreciate and enjoy.

Today, 100s of factors influence relevance, many of which we'll discuss through this paper. Clustering is the classification of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - according to some defined distance measure. It can enable users to find the relevant documents more easily and also help users to form an understanding of the different facets of the query that have been provided for web search engine. We used clustering algorithm is K-means and EM and compare the result of clustering on Weka and Tanagra and find the values of factors like title length, number of backlinks, Domain length, keywords in title that affect search engine optimization.

KEYWORDS: Weka, Tanagra, K-Means, EM, Search Engine, Dataset