BUILDING SEMANTIC KERNEL OF WEB SITE PAGE

Dudnichenko S.M., Malugon H.V., Orekhov S.V.
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

For today site content optimization problem is one of the main part of the web-application development. At the momemtn web (site) content is being described by a semantic kernel. It is a list of favorite keywords in term of search engine and potential customers [1].

The semantic kernel of the site allows us effectively optimize the site position in search results. Also it manages the priority of sections and pages, information architecture as whole. But information architecture reflects on company's activities in the market of goods / services via Internet.

The main focus of the research is to develop the methodology of building semantic open source kernel. Our goal is to develop the software, which runs the training cycle and verifies the criteria for selecting keywords. The algorithm is based on the relevance feedback (Data mining method). As part of this method, search is considered as an iterative process, at each stage of which the user's information needs are clarified. [1].

To solve the problem, a more detailed method of tree-making was considered. It involves the graphical construction of various options for action that can be taken to solve an existing problem. As a result of the construction of the decision tree, the probability of each scenario, NPV (net present value) for each scenario is calculated, as well as a number of other fundamentally important indicators.

The TF-IDF method is used to evaluate the importance of the keyword, both in the context of the document (TF) and in the context of the document body (IDF). The metric is used with different TF and IDF calculation options.

The N-gram method calculates the probability of the last N-gram of keyword if all the previous ones are known. When using this approach for modeling the language, it is assumed that the appearance of each word depends only on the previous words. The use of N-grams is the detection of plagiarism. If you divide the text into several small fragments, represented by N-grams, they are easy to compare with each other and thus get the degree of similarity of the analyzed documents.

The clustering method is used to optimize the web application. It allows the broadest possible coverage of the semantic aspect of the search than it is possible in mechanical processing, which facilitates both the receipt of conceptually important for SEO clusters, as well as the definition of relevant to the required semantics of the pages that the user enters after performing a search query. It allows us to reduce losses in the event of a potential non-visit of the page due to an incorrectly prepared list of keywords.

References:

1. Michael Godlevsky, Sergey Orekhov, Elena Orekhova. Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning / M. Godlevsky, S. Orekhov, E. Orekhova // Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. – 2017. – Vol-1844. – P. 23-32.