ФОРМИРОВАНИЯ МНОЖЕСТВА ОБЪЕКТОВ ПО ЗАДАННОМУ НАБОРУ ПРИЗНАКОВ

Фищукова Н.В., Быкова А.
Национальный технический университет «Харьковский политехнический институт», г. Харьков

В области исследования рынка и рекламы часто возникает задача, когда нужно найти все сайты заданной тематики. В настоящее время в сети Интернет количество сайтов превысило миллиард. Эти сайты отличаются друг от друга по очень большому числу параметров. Они отличаются друг от друга тематикой, назначением и многими другими характеристиками. Сайты со схожими параметрами можно классифицировать в группы.

Для решения подобной задачи успешно используется кластерный анализ и алгоритмы классификации. Кластерный анализ предполагает выделение компактных, удаленных друг от друга групп объектов, отыскивает «естественное» разбиение совокупности на области скопления объектов. Он используется, когда исходные данные представлены в виде матриц близости или расстояний между объектами либо в виде точек в многомерном пространстве. Также для решения задачи формирования множества объектов применяются такие алгоритмы как например, метод к средних, генетические алгоритмы и графовые алгоритмы кластеризации. Эти алгоритмы имеют разную вычислительную сложность и некоторые недостатки. В работе решается задача классификации сайтов на основе классификации отдельных страниц сайтов и на основе данных о посещаемости веб ресурса. Данные о посещаемости получены с помощью поисковой системы google.com и Alexa.com**,** где собирается статистика о посещаемости сайтов и списки взаимосвязанных ссылок. Для определения релевантности сайта учитывалось количество релевантных страниц сайта, вес рубрики страницы. В работе разработан алгоритм классификации сайтов на основе более 2000 сайтов интернет. Для выбора наиболее подходящего проанализированы существующие алгоритмы и проведено сравнение их на основе предложенной выборки.