Аналитики компании Cisco прогнозируют, что к концу 2016 года объём переданных/принятых данных Интернета достигнет величины в один зеттабайт (1 099 511 627 776 гигабайт), причём наблюдатели отмечают нелинейный рост этого показателя, из-за чего преодоление рубежа в два зеттабайта можно ожидать в 2019 году [1].

Проблема сбора и анализа больших объёмов информации является одной из самых актуальных проблем области информационных технологий, специалисты разрабатывают различные способы её решения, совокупность которых образовала новое направление развития в отрасли под названием «большие данные».

Большие данные (англ. Big Data) — новый термин, появившийся и получивший популярность в конце 2000-х годов. Данный термин обозначает настолько объёмный и сложный массив информации, что такие традиционные инструменты разработчика, как реляционные базы данных, не способны обработать его в приемлемых затратах по времени или ресурсам [2].

В области больших данных применяется значительное число методов и техник анализа информации, одним из них является кластерный анализ данных, или обучение без учителя. Кластерный анализ данных нашёл широкое применение в индустрии информационных технологий: в качестве примеров можно упомянуть кластеризацию результатов поиска в Интернете, сегментацию изображений и интеллектуальный анализ данных.

Однако с возросшим объёмом поступающей информации задача кластерного анализа данных усложняется, в следствие чего появляется необходимость в поиске оптимального алгоритма кластеризации для работы с большими данными и его усовершенствование для повышения показателей производительности.

В то же время большой объём данных подчёркивает важность учёта семантической составляющей информации: чем больше массив данных, тем сложнее выявить в нём внутренние связи. Очевидно, в алгоритм кластеризации невозможно поместить все условия по выявлению семантики, поэтому на помощь в решении этой задачи приходят базы знаний, также являющиеся молодой, но стремительно развивающейся областью информационных технологий.

Цель работы — создание информационной системы, осуществляющей кластерный анализ больших массивов данных с учётом их семантических связей.

Для достижения поставленной цели необходимо решить следующие задачи:

- рассмотреть, изучить и проанализировать эффективность существующих свободных программных продуктов, которые могут стать компонентом разрабатываемой системы, к этой категории можно отнести базы данных, платформы и фреймворки распределённой обработки данных;

- разработать архитектуру информационной системы, включающую в себя базу данных для кластеризуемой информации, базу знаний, компонент кластерного анализа, компонент обработки данных и синхронизации;

- произвести анализ существующих алгоритмов кластеризации и выявить один или несколько алгоритмов, потенциально наиболее производительных при работе с большими данными;

- произвести анализ существующих баз знаний и разработать структуру системы хранения и обработки смысловых связей;

- усовершенствовать выбранный алгоритм кластеризации, адаптировав его к работе с большими данными и обработкой семантики;

- разработать программный каркас, позволяющий объединить все компоненты разрабатываемой системы;

- интегрировать все компоненты в единую информационную систему и произвести тестирование, оценить полученные результаты.

В диссертационной работе разработано:

...