**Глава 1 Литературный обзор**

Данная глава последовательно даёт обзор источников информации из области современных баз данных, фреймворков (программных каркасов для построения информационных систем) и кластерного анализа данных, в качестве которых рассматриваются литература, электронные документы, компьютерные программы.

1.1 Обзор современных СУБД

Первоочередной задачей является определение наиболее подходящей системы управления базами данных (СУБД) для хранения знаний.

На данный момент выделяют следующие основные группы СУБД:

- реляционные. Наиболее популярные представители: Oracle, MySQL, MS SQL Server, PostgreSQL, DB2, SQLite;

- распределённые. Cassandra, Hbase;

- «ключ-значение». Redis, Memcached;

- документно-ориентированные. MongoDB, Amazon DynamoDB, Couchbase, CouchDB;

- графовые. Neo4j, OrientBD, ArangoDB [1].

К концу 1980-х годов реляционные СУБД стали наиболее популярным решением и сохраняют это положение на данный момент [2].

Несколько десятилетий разработчики программного обеспечения пытались приспособить связанные, полуструктурированные наборы данных к хранению в реляционных СУБД. Но хотя реляционные СУБД были изначально спроектированы для систематизации бланков и табличных структур, они плохо приспособлены для хранения ситуативных, исключительных связей, которые неожиданно возникают на практике [3].

Связи между данными являются неотъемлемой частью реляционных СУБД, однако только на уровне моделирования, как средство объединения таблиц. Зачастую необходимо снять неоднозначность семантики связей, связывающих сущности, так же как и определить их вес или силу.

Реляционные СУБД не располагают подобным функционалом. Помимо этого, по мере накопления резко разнящихся значений и всенаправленного усложнения и размытия набора данных, реляционная модель перегружается соединёнными операцией JOIN таблицами, частично заполненными записями и множеством условий на отсутствие значений.

Рост взаимосвязанности в реляционной СУБД превращается в множество JOIN-операций, которые отрицательно сказываются на производительности и усложняют дальнейшую адаптацию существующего набора данных к дальнейшим возможным изменениям в бизнес-логике [3].

СУБД только с коммерческой лицензией не соответствуют поставленным требованиям к создаваемой системе, поэтому СУБД Oracle подробно не рассматривается.

1.2 Современное состояние области баз знаний

1.3 Современное состояние области кластерного анализа данных

Рассматривая область кластерного анализа данных, нужно отметить многообразие алгоритмов и метрик. На данный момент не существует общепринятой классификации алгоритмов кластеризации, однако можно выделить несколько основных групп, имеющих схожий принцип работы.

Иерархические алгоритмы

- аггломеративные

- дивизимные

- BIRCH

- CURE

- ROCK

Chameleon

Echidna

Разделяющие

- K-means

- K-medoids

- K-modes

- PAM

- CLARANS

- CLARA

- FCM

Плотностные

Сеточные

Моделируемые