TAG BASED STORAGE AND RETRIEVAL SYSTEM FOR ORGANIZATION RELATED NEWS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Kübra PARKIN

June 2019
İZMİR

We approve the thesis of **Kübra PARKIN**

Examining Committee Members:

Prof. Dr. Oğuz DİKENELLİ
Department of Computer Engineering, Ege University

Asst. Prof. Dr. Selma TEKİR
Department of Computer Engineering, İzmir Institute of Technology

Assoc. Prof. Dr. Tuğkan TUĞLULAR
Department of Computer Engineering, İzmir Institute of Technology

27 June 2019

Assoc. Prof. Dr. Tuğkan TUĞLULAR
Supervisor, Department of Computer Engineering
İzmir Institute of Technology

Assoc. Prof. Dr. Tolga AYAV
Head of the Department of
Computer Engineering

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. Tuğkan Tuğlular for sharing his information with me, directing me throughout my thesis and supporting me in this long period of time.

I thank Bimar for providing data and giving me this opportunity and being insightful to me during my thesis.

Finally, I would like to thank to my family for their unlimited love, patience and support during this thesis and all my life.

ABSTRACT

TAG BASED STORAGE AND RETRIEVAL SYSTEM FOR ORGANIZATION RELATED NEWS

For corporate organizations, it becomes more and more important to gather information about opponents or partners, or any kind of information that can be related to the organization. In a rapidly changing world, ensuring competitiveness for organizations and making consistent strategic decisions are becoming increasingly difficult. Gathering news about the business has an undeniable effect on the decisions of companies. It is essential to keep up with this race in order not to get out of the race. Therefore, what corporate companies need is to have a retrieval system that collects and evaluates information that is relevant to the organization. However, it can be difficult to make use of large amounts of information. What needed is to store that information based on a pattern and make it easy to analyses.

ÖZET

KURUMA İLİŞKİN HABERLER İÇİN ETİKET TEMELLİ SAKLAMA VE ERİŞİM SİSTEMİ

Kurumsal kuruluşlar için, rakipler, ortaklar veya kuruluşla ilgili olabilecek her türlü bilgiyi toplamak gittikçe önem kazanmaktadır. Hızla değişen dünyada, kuruluşlar için rekabet gücü sağlamak ve tutarlı stratejik kararlar almak giderek zorlaşmaktadır. İşletmeler hakkında haber toplamak, şirketlerin kararları üzerinde yadsınamaz bir etkiye sahiptir. Piyasadaki konumu korumak ve gelişebilmek için gündemi takip edip bu doğrultuda kararlar almak çok önemlidir. Öyleyse, kurumsal şirketlerin ihtiyaç duyduğu şey, organizasyonla ilgili bilgileri saklayan ve değerlendiren bir sisteme sahip olmaktır. Ancak, büyük miktarda bilgiden yararlanmak zor olabilir. Bu durumda ihtiyaç duyulan şey, bu bilgileri bir desene dayanarak depolamak ve bilgilerin analiz edilebilmesini kolaylaştırmaktır.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

CHAPTER 1

INTRODUCTION

Today, timely information is critical for companies. For a corporate company, being aware of the latest news has a significant role in determining the decisions to be taken in the future. It is important that data are kept and presented as required by the organization. Especially, the news on the Internet is very important in terms of taking quick action. However, trying to make sense of information without a proper method would take quite a long time. However, the news collected from the Internet does not have a definite scheme. The format of the news may vary depending on the source or type. Storage is also important because of the data size. As the data grows, it is important to protect data reading speed.

In addition to storing the news, it is also necessary to read and analyze the news data. The fact that the number of news being high makes it difficult for a person to read and understand one by one and decide whether it is important or not. To understand the news, the content should be read. But considering the number of news, the time of reading all the news will be very high. For this reason, it is necessary to analyze the news data set. The analysis of the news can be done via the article body or the tags of the news. It is possible to return the data as a search result by the tags defining the news. Tags can be used to analyze a lot of news. Weighting can be based on the position of that keyword in the news and the number of repetitions in the news text.

Tags are important for the news to make inferences at first glance. The methods of storing with its tags are becoming increasingly popular. Indexing the tags can result in a faster return of search results.

The tags can be used in the analysis of news, and the news articles can also be used to analyze and categorize the news. Especially, to have an idea about the news in a certain period is possible with the text mining methods. This allows for faster grouping.

In this thesis, the storage of tag-based data and the fast data reading in a specific format are compared in different databases. In addition, the news texts are analyzed by means of machine learning methods.

1.1. Motivation

In the transportation sector, it is very important to track information about customers and competitors. Rapid response to risks and opportunities is increasing the value in terms of gain or loss prevention. For this reason, the monitoring of corporate news is getting important in the Internet environment where information flow is fast. The earlier the information on the business areas of the companies is reached, the more successful decisions can be made.

One of the focal points of information technology is to provide faster information to people. Information sources on the Internet can be easily accessed. However, news on the Internet has a high volume and it can take a very long time to have all of them screened by a person. Therefore, the keywords that indicate the interests of the institution were issued and given to a company and the corporate news on the Internet was scanned. Internet news transmitted from external sources are stored and presented to users. However, the rapid growth of the amount of data slows down the processes of storing and handling data. Therefore, a more efficient data storage method is needed.

Collecting and analyzing the bulk data without looking at the individual data one by one will make the job of the users easier. Therefore, it is planned to group data by examining the texts of daily news.

1.2. Thesis Goals and Contributions

The aim of this thesis is to store and analyze the news published on the Internet about the corporate firms themselves, their customers or their competitors and to examine the methods to be used for the analysis of the news. The performance of querying the tag-based data set is compared in the relational and non-relational databases and the reading performances are compared. Then, it is aimed to group data of the news into clusters by doing text analysis to lighten people’s work who try to analyses this news.

1.3. Outline of Thesis

This thesis is organized as follows. The next chapter provides a literature overview. Chapter 3 gives a background information about Relational and NoSQL databases, and text mining algorithms concepts and their applications. In Chapter 4, the proposed approach is explained. In Chapter 5, we explain the dataset and experiments that we have conducted. Finally, Chapter 6 provides final remarks and discusses future research.

CHAPTER 2

RELATED WORK

In recent years, a news storage method was needed because of the rapid growth of data size. Relational databases can handle a large volume of data, but the management of data and retrieval efficiency become a problem. NoSQL databases are known as providing a solution for large volume and unstructured data. However, NoSQL databases can perform equal or sometimes better performance on structured data comparing to relational databases. Research shows that MongoDB has better runtime performance for inserts, updates and simple queries. SQL performs better when updating and querying non-key attributes, as well as for aggregate queries [1].

Today, relational databases are still used for information management systems. However, in most cases when a large dataset is needed to query it is not effective. Especially multi-table join queries can be very slow. But NoSQL data management systems which are relatively new can perform better. In an experiment, a NoSQL database is used to replace the relational database for a traditional information management system [2]. And as a result, the performance comparison is made for these two schemes.

Database management systems have been used for many years, they are one of the main components of software projects. In recent years NoSQL databases are preferred to store large amounts of data. There are many NoSQL database types. Five most popular NoSQL databases: Cassandra, HBase, MongoDB, OrientDB and Redis are compared in terms of query performance, based on reads and updates [3]. Cassandra has shown outstanding performance in most of its workloads, followed by comes MongoDB. Each database has advantages for specific conditions, performance will vary according to the data to be stored.

In a study in 2019, performance comparisons were made on MongoDB, Cassandra and HBase [4]. This study shows that MongoDB performed very well with low throughput, but not as well with high throughput. Cassandra and HBase performed very well under heavy loads due to their optimized designs.

Information technologies grow increasingly. This growth causes a lot of data more than people can analyze. So, it becomes important for companies to process, organize,

categorize and analyze this big data automatically. The people who placed in strategic roles in companies need to collect and analyze data to give important decisions. The VIPAR project which contains data gathering and processing has been described [5]. The Virtual Information Center (VIC) at US Pacific Command, gathers, analyzes, and summarizes information from Internet-based newspapers daily. News are gathered and categorized, and every new document is dynamically categorized.

In a database management system, one of the important subjects is scalability. A research has been conducted to examine the scalability of OLTP-style applications in SQL and NoSQL databases [6]. In the research consistency mechanisms, storage mechanisms, durability guarantees, availability, query support are contrasted. The results show that some databases sacrifice transaction consistency to gain more availability.

Clustering is a useful technique. It organizes many unlabeled text documents into a small number of meaningful clusters. Therefore, clustering provides a basis for intuitive use. For clustering, there are several similarity measures. In a research, K-Means algorithm is implemented, and five distance/similarity measures used in text clustering [7]. The results show that the dataset has a significant impact on the performance of similarity measurement.

In a study, a comparison is made to find out the advantages and disadvantages between relational databases and NoSQL databases [8]. Mainly, the disadvantages are as follows; NoSQL databases are immature, there is no standard query language and maintenance is difficult.

When text data is clustered, each word becomes a feature. In other words, there are as many features as the number of words in a document. Considering that a document contains a small portion of the features in the data set, clustering in such a data set will consume a lot of resources. For this reason, in a research, it is investigated that analyzing data with the features which are frequently used [9]. The result of this research shows that standard clustering algorithms perform better when the frequent features are used.

Document clustering algorithms play an important role in easily interpreting large data in small and meaningful clusters. Hierarchical clustering algorithms sort the data at different levels according to their similarities and help people to visualize and make sense of this information. In a study, it has been shown that partitional algorithms in large data sets always give better results than agglomerative algorithms [10]. In this research, they present a new algorithm called constrained agglomerative algorithm. In this new

algorithm, features of partitional and agglomerative algorithms are combined. Results show that it performs better than agglomerative or partitional algorithms alone.

In a survey, methods to find text similarity is discussed by dividing three categories [27]. These categories are string-based, corpus-based and knowledge-based. Existing works on text similarity can roughly group into these categories. There are also a number of studies that try to combine some of these methods and create hybrid solutions.

The common features of two objects determine their similarities. Regardless of semantic meanings, the words contained in a text are features of that text and the similarity of texts is estimated by looking at the number of common features. Analyzing the semantic similarity of the documents can increase the rate of obtaining accurate information. In a study, a method has been proposed to measure the semantic similarities of texts [28]. As a result, it is indicated that text similarity detection should be handled not only structural but also lexical, syntactic and semantic knowledge.

Text document clustering is very important since it helps to organize large sets of documents into a small number of clusters. Mostly text clustering solutions only relate documents that use identical terminology and they ignore semantic similarity of the terms. In a study, they investigate the benefits of using ontologies like WordNet in text clustering on news data [29]. They perform clustering on Reuters news dataset. As a result, it is indicated that considering different words which are the subconsepts of the same element has positive effects on clustering results.

CHAPTER 3

RESEARCH BACKGROUND

In this chapter Relational Databases, NoSQL Databases, Artificial Intelligence, Machine Learning and their methods are introduced.

3.1. Streaming API

Streaming APIs are used for real time communication between client and server. It is different from traditional APIs because it leaves the HTTP connection open as long as it is possible. To leave the connection open is called persistent connection. In traditional Rest APIs, client makes a request and according to this request, a result is returned. That means client needs to request data. But this method is not valid if a real time communication is necessary. So, to make a persistent connection and establish real time communication streaming APIs are used.

After the connection is opened, it stays open. If new data is available, server pushes data to client through HTTP connection. There is no need for the client to check server for newer data. This approach decreases network latency by assuring almost real time communication.

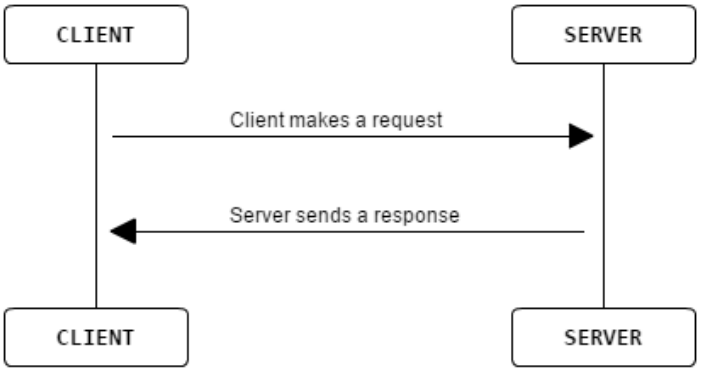

Figure 3.1. Streaming API

Streaming APIs mostly used for reading data. Streaming APIs are relatively new subjects. Social applications like Twitter, Facebook etc. are using Streaming APIs. Connecting a Streaming API requires an implementation that ensures an open connection. For some reason if the connection is down, it should be open as soon as possible. Since server does not close the socket, client can make another request on the same socket.

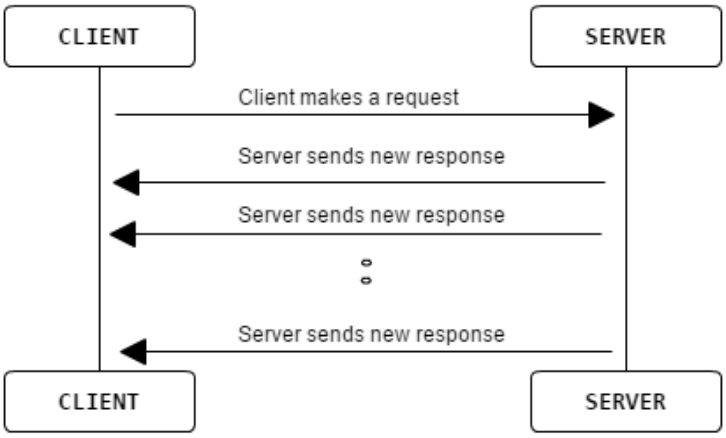

Figure 3.2. Rest API

As the new data comes, server sends it to the client. In the client, the listener application can directly transfer data to the end user (by writing data to database and displaying it on the screen) or a queue can be created to collect specific amount of data and then store into database.

3.2. Relational Databases

In 1983, Andreas Reuter and Theo Haerder first introduced ACID as a shortcut for Atomicity, Consistency, Isolation and Durability [11]. Relational databases provide these four principles of the ACID model.

- Atomicity of transactions; guaranties that if a database transaction fails then the entire transaction will be rolled back.
- Consistency; if an atomic transaction breaks the consistency of the database then the entire transaction fails.

- Isolation; if multiple transactions are occurring at the same time or near then database engine guaranties isolation.
- Durability; when a transaction is committed to the database then it is permanently preserved for the use of backup and transaction logs.

Relational databases support transactions. A database transaction is a unit of work performed in a database management system. It must provide ACID properties. Providing these properties help database to keep correct data in all partitions. In operational systems, the transaction guarantee for operations that work on records is very important.

3.3. NoSQL Databases

The formerly applied waterfall business development process has lost its influence in the current project management. Because now business development processes aim to quickly generate code in short sprints to respond to changing needs quickly. The applications that previously targeted a specific audience are now expected to be always on and must be scaled by user volume. For these reasons, companies that used to be able to meet the needs of vertical scaling servers have started to keep their resources and servers in the cloud and change their architecture. Mostly relational databases do not meet the requirements of today's scale and agility needs.

NoSQL is an approach to database design. It stands for “not only sql”. It provides alternative to traditional relational databases. In traditional databases data is placed in tables and schema is carefully designed. Recently applications have begun to produce large amounts of data. And the types of these data may change over time, for example, even if data is structured, the need to return to the unstructured may arise.

While relational databases meet the principles of ACID, NoSQL databases meet the principle of BASE [12].

- Basic availability; focuses on the availability of data even if there are failures. To achieve, it uses distributed approach to database management. Instead of a single large data store NoSQL database spread data many storage systems. When an instance fails, this does not mean that complete database failures.

- Soft state; means that the state of the system can change over time, consistency is not guaranteed. NoSQL system indicates that data consistency is developer's problem, and database does not need to handle consistency.
- Eventual consistency; data will be consistent over time, but not immediately. At some point in the future data will be in a consistent state.

NoSQL has become a general concept used for non-relational databases. Many databases with different characteristics are known as NoSQL. There are four types of NoSQL databases. These are Key-Value Stores, Wide-Column Stores, Graph Stores and Document Databases.

3.3.1. Key value stores

Key value Stores [13] are the simplest type of NoSQL. Data is stores as key and equivalent values to the keys. They are known to scale well. A key value is also referred to as a dictionary or hash. The key should be unique identifier to access the value related with the key. The key could be anything. A DBMS can bring limitations on key, while the others don't. For example, Redis uses key value storage. Its maximum key size allowed is 512 MB. Binary sequence, or a string of text, or contents of an image file can be key. But for performance reasons key should not be too long.

The value in a key value store also can be anything, text, number, markup code such as html, programming code like php or image. The value could also be a list or another key-value pair in an object. Key value stores are mostly preferred to store session information. Redis, Oracle NoSQL, Voldemorte are the examples of key value stores.

3.3.2. Wide Column Stores

In Wide Column Stores [14], in the same table columns can be different in every row. It is like a two-dimensional key value store. It uses columns to store data. Related columns are grouped and created a row as shown in the Figure 3.3.

Column family contains the line key in the first column; this row is unique for that colon family. The interior columns have a unique key that identifies that column. so that

the column values can be read by key. Figure 3.4. demonstrates storage in detail. It uses tables, rows, and columns, but unlike a relational database, the names and formats of columns can vary from row to row in the same table.

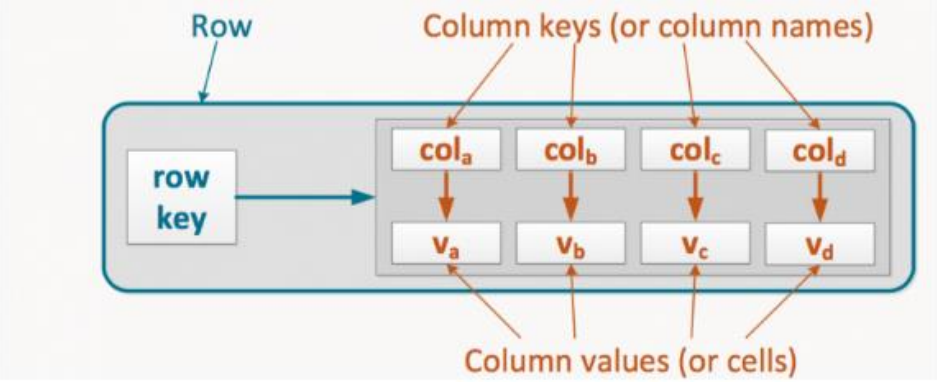

Figure 3.3. Storage of Wide Column Store Databases (Source: [24])

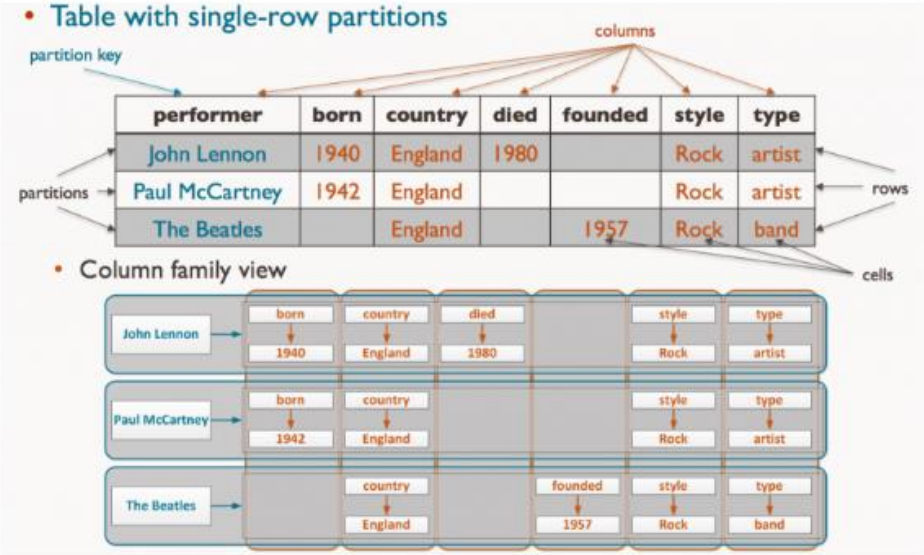

Figure 3.4. Wide Column Store Databases in detail (Source: [24])

Wide column stores provide high performance and they are highly scalable. That's why it is used for personalization and content recommendation. Cassandra and Hbase are the most known wide column stores.

3.3.3. Graph Stores

Graph stores [14] use nodes and relationships. A node is an entity (for example user or category is an entity), a relationship is the association of two nodes. Graphical databases use nodes that have relationship lists in them. The registration of relationships replaces the time-consuming search or matching process in relational databases.

For example, the process of joining three tables for search in relational database is shown in Figure 3.5.

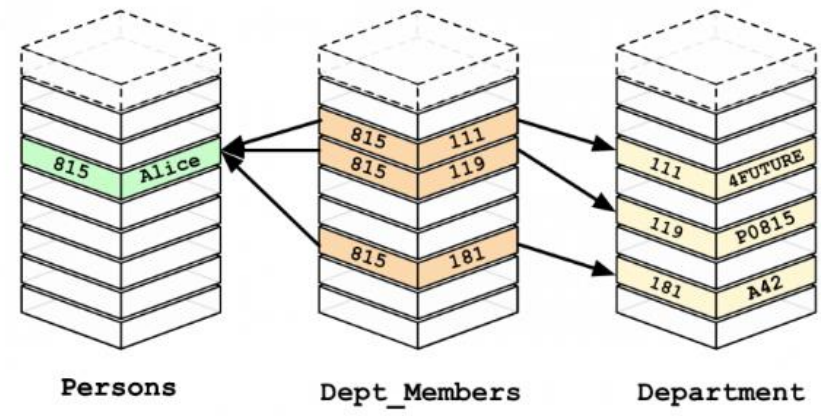

Figure 3.5. Joining table in relational databases (Source: [24])

Figure 3.6. shows that the same process is done in graph store; the connections are already defined on nodes.

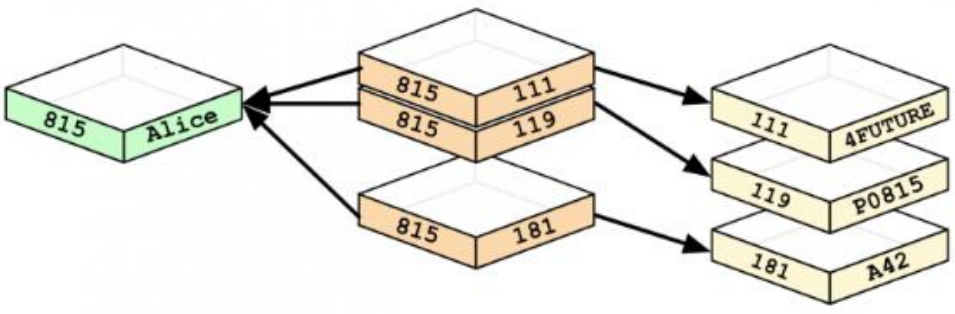

Figure 3.6. Joining tables in Graph Stores (Source: [24])

Graph stores are useful when data sets grow too large and target high query performance. Examples of graph stores are Neo4j and JanusGraph.

3.3.4. Document Databases

It is always aimed to normalize data in relational databases. For example, a user can have multiple addresses, so user and address tables are separated. However, joins are needed to pull information later. The process of aggregate is costly in runtime.

The data in the document database is kept denormalized [14]. All information is written in a single document without dividing. to retrieve single person information does not require join anymore, single read operation is enough to get data. With denormalized data, less queries and updates are required. This data model provides better read performance.

But this does not mean that document databases contain one collection and all data is denormalized and saved as a document. In theory there is no limit for fields in a document, and values that corresponds these field could be infinite. As the document size grows, the performance of reading writing data process can slow down. In that case it is practical to split document into collections.

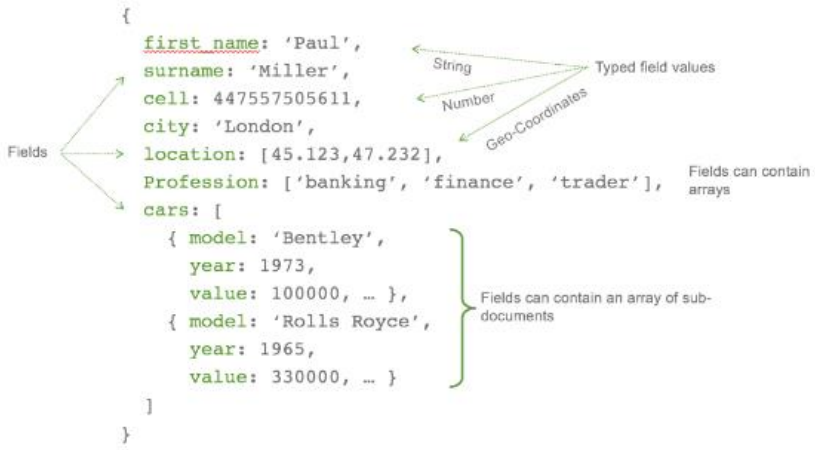

Figure 3.7. A sample document in document databases (Source: [24])

Document stores use JSON, XML or BSON documents. Each document can contain different fields but for the query performance index should be on common fields. Values can be string, integer, array or even another document. Figure 3.7 shows a sample

document. Document databases are suitable for content management applications such as blogs and video platforms. If the data model needs to be changed, only the relevant documentation should be updated. No schema update is required to make changes. Popular document databases are Amazon DynamoDB, MongoDB and Couchbase.

3.3.4.1. MongoDB[1]

MongoDB is an open-source, non-relational document database. It can store JSON like data. It supports flexible data model, so you can store data in any structure. Provides full index support, sharding and replication.

- MongoDB is a document-based database which is flexible and scalable.
- It stores data in JSON like documents, it means that the fields in set can be different in different documents or data structure can change over time.
- Since it stores data in document model, it is easy to map objects in application code to database
- It has powerful indexing and real time aggregation,
- MongoDB provides high availability and horizontal scaling means adding more machine to database resource pool to get more performance when needed.
- MongoDB is open-source and free.

NoSQL databases can be used for multimedia storage. MongoDB offers two options to store multimedia. GridFS is a specification for storing and retrieving large files in MongoDB. But if multimedia is not big (if it is smaller than 16 MB) than multimedia can be stored in a field in MongoDB document.

With MongoDB, you can store all related data in a single document. This model is generally known as denormalized model. Demoralized data models allow applications to store related pieces of information in the same database record. So that applications run fewer queries to perform operations.

References in normalized data models are used to define relationships between documents. Normalized data model is good to use when modeling large hierarchical data sets. If denormalized model causes duplication and does not provide efficiency, then

[1] https://www.mongodb.com/

normalized method can be used. Query capabilities are very strong in MongoDB. It has an aggregation pipeline which is a framework modelled for data processing pipelines. It is an alternative to map reduce. Map reduce can be used for more complex operations.

Map reduce simplifies the analysis on large data. Google engineers have found that the existing web crawling techniques are inadequate as the time passes and the data set grows. These engineers thought that distributed systems built with cheap computers would solve the problem. Map reduce is the algorithm used by google in the background of data processing and indexing. In the case of distributed architectures, collecting and analyzing data in systems with many machines. The data is collected in the map phase, and the data collected in the reduce phase is analyzed. The desired data for analysis is selected at the map stage. To summarize, instead of using highly equipped computers to work on large data sets, operations with map reduce over a set of commodity hardware are more efficient. This is the reason why Google is more successful than other search engines.

Map is equivalent to the select operation in relational database. Reduce is equivalent to aggregation operations like count, sum etc. Map reduce simplifies these operations in MongoDB. In addition to the map and reduce functions, there is also a finalize function. This function can be used optionally on the reduced data set, if there are any additional changes in result set.

In MongoDB, map reducing operations are done with JavaScript. It is used for count, sum and having operations in classical databases. Many libraries are written with different languages using the map reduce algorithm. The most well-known of these is Hadoop. Hadoop is a library written in Java that allows processing on multiple machines on large data sets.

3.4. Artificial Intelligence

For many years, people have imagined the existence of machines that behave like human beings. In order to make a machine that imitates human beings, human characteristics must be well understood. Intelligence is among the most important concepts of human. People can comment and draw conclusions by using their minds. These abilities are intended to be taught to the machine in artificial intelligence.

The goal of artificial intelligence is to create a machine that thinks and acts like human beings [15]. There are some ways to test the capability level of machine. Usually the method recommended by A. Turing is used to measure the level of the machine. Turing's 1950 seminal paper in the philosophy journal Mind is a major turning point in the history of Artificial Intelligence. The paper crystallizes ideas about the possibility of programming an electronic computer to behave intelligently, including a description of the landmark imitation game that we know as Turing's Test [16]. In this test method, human and computer answer the questions asked. If the referee does not understand which one of the answers belongs to the machine, the machine has passed the test.

Artificial intelligence can be used in many areas. Voice recognition, machine translation, driverless vehicles, cancer cell detection, fraud detection and object or person detection / tracking are some of them.

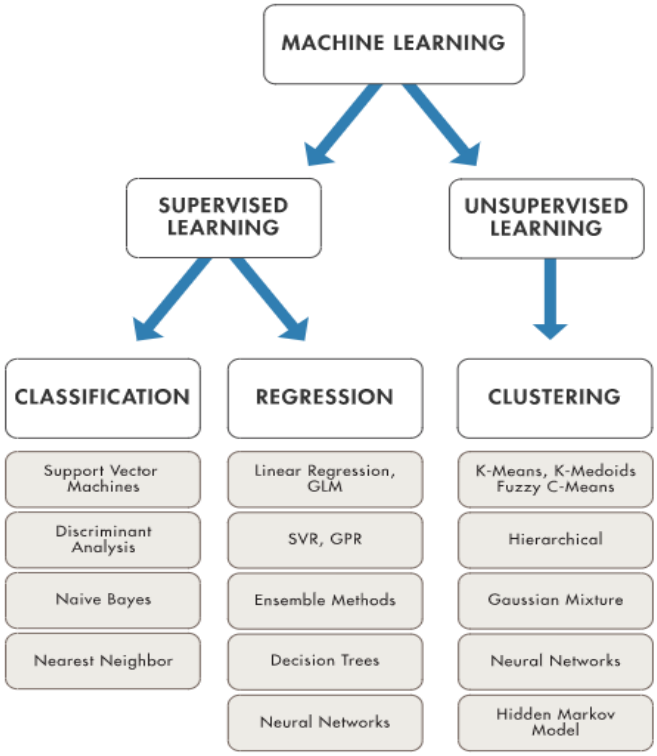

Figure 3.8. Machine Learning Types and some of the algorithms (Source: [25])

3.4.1. Machine Learning

Machine learning is a subset of Artificial Intelligence. Generally, to examine and make sense of the large data sets of computers is called machine learning [17]. Machine learning algorithms can learn and develop from data without people. The focus of machine learning is to give the ability to perceive patterns and to make rational decisions based on data. It is closely related to areas such as statistics, data mining, pattern recognition etc. There are two most common types of machine learning [17]; supervised and unsupervised. Figure 3.8 shows the machine learning types and some of the algorithms.

3.4.1.1. Supervised Learning

The inputs and outputs in the data set are clear. Inputs and outputs are given to the machine and it is expected to learn how to extract the results from this information. There are two kinds of supervised learning methods, they are regression and classification.

3.4.1.2. Unsupervised Learning

Only inputs are available in the data set. Outputs are unknown or not given to the machine. Modeling of the underlying structure of the data is desirable. Without labels, it is aimed to find the structure of data. Clustering algorithms are one of the most commonly used unsupervised learning techniques. It is used for exploratory data analysis to find groups which is hidden in the data.

CHAPTER 4

PROPOSED METHOD

In the transportation sector, it is important to track current events. The fastest follow-up of events or developments can be done on the Internet. The capture of a sectoral news is more efficient because the news will appear sites on the Internet faster than the printed publications. However, there may be a lot of news on the Internet and it takes time for a person to read and analyze them individually. The selection of important topics at first glance will shorten the time for users to review the news.

In this project, it is aimed to keep the news data in the most suitable environment and to obtain summary information for rapid analysis.

4.1. Prediction Challenges

The news can be collected from different sources. Digital newspapers and magazines or online publications like Internet news can be information source. So, these sources should be tracked and collected news should be stored. News objects are being transferred by an external source via the Streaming API. When news objects are captured, they are immediately written to the API. Therefore, new objects must always be read through an open connection.

It is aimed to analyze news by its tags, but it is not clear whether the tags give enough information about news. Tags are the keywords that has given the company which tracks the news on the Internet with respect to these keywords. Tags give general information about news, may not be enough to explain the content of the news to find similarities.

4.2. Model Architecture

The proposed approach is as follows. The news read from the data source are written to both MongoDB and Microsoft SQL Server and their performances are compared with a sample query. The results of keeping the same data set in different databases are compared and the database providing the most efficient environment is determined accordingly.

Then the analysis of news data with tags will be evaluated. In this experiment, the repeats of the news tags on the data label is examined. The frequency of labels alone, in binary groups, or in triplet groups can help us understand the data set. The process of checking the replicas of the labels is done by using the map-reduce algorithm in MongoDB. The equivalent operation is implemented with the C # code for SQL Server.

Finally, the contents of the news are analyzed with the help of clustering methods. Two clustering algorithms are implemented in Python on Visual Studio Code.

4.2.1. K-Means

The K-Means algorithm is an unsupervised learning clustering algorithm [18]. K-Means algorithm aims to find groups in data. The K value in K-Means determines the number of clusters. So, it uses predefined count of groups and this value must pe passed to the algorithm. Algorithm assigns each data point to one of the groups based on the features. After the value of K is decided, K-Means algorithm selects the random cluster center. It calculates the distance between each data and randomly selected cluster centers. Then it assigns the data to a cluster according to the nearest cluster center. Then, new cluster center is selected for each cluster and clustering is performed according to the new center. This continues until the system is stable.

4.2.1. Hierarchical Clustering

Hierarchical clustering contains creating clusters with the order based on similarities [19]. There are two types of hierarchical clustering. They are Agglomerative

and Divisive. In Agglomerative Hierarchical clustering all data is set into a cluster initially. If there are N elements, then N cluster is created, and each element assigned to each cluster. Then the clusters that are close to each other are joined together to form a new cluster. This continues until the system is stable. Divisive is the exact opposite of the Agglomerative. At first all data is created in a single cluster. This cluster is then broken smaller clusters.

There are many ways to calculate the distance in agglomerative hierarchical clustering. This distance calculation is also used to create dendrogram.

- Single Linkage calculates the closest distance between two sets.
- Complete Linkage calculates the longest distance between two clusters.
- Average Linkage calculates the average distance between two sets.
- Besides these, there are Ward, Weighted, Centroid and Median methods.

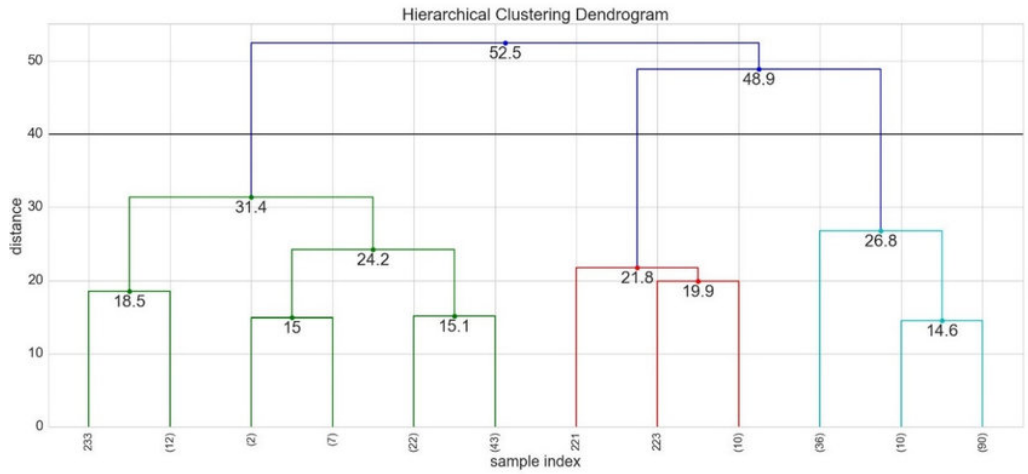

Figure 4.1. Optimal cluster calculation on dendrogram (Source: [26])

A dendrogram is a tree diagram that illustrates the relationships between hierarchical clusters or similar data sets. It shows how many clusters there will be. The horizontal line drawn from the longest leg shows the optimal cluster count. In the Figure 4.1., the number of vertical lines that cut the horizontal line gives the number of clusters.

CHAPTER 5

EXPERIMENTAL RESULTS

In this chapter, data collection and experimental study is explained. Data was collected from an external source. This data was stored on different types of databases. Some experiments were performed to see storage performances. Furthermore, text data was analyzed by machine learning methods to measure document similarity. Two different clustering methods were used and their results are compared.

5.1. Preparing Dataset

News data is provided from a third-party source. On average 3500 news are collected per day. News object contains headline, article body, publish date, source url, tags, image and other descriptive fields. A news object contains 16 tags on average. Article body of news is about 4000 characters. News objects are sent as JSON text over a Streaming API.

For data storage there are various options. The important issue is to find method that fits data. For text-based data, relational or NoSQL databases can be used. Regarding data volume, read/update frequency or schema the storage method can be decided. As the data size increases relational data storage systems can be slow on data retrieval.

Additionally, multimedia like images storage is important. Multimedia storage is available in relational databases. However, it is strongly recommended not to store images in relational database. It can drop read-write performance of rows.

Another option is to store media files in a file system. Files can be stored on disks and its reference can be kept in database. However, this method is hard to maintain as the file volume increases.

To collect news data, an API is implemented. In this API, data source is streamed. In other words, the channel is kept open to read when a new object is received. Therefore, this channel needs to be continuously listened.

The data read from the source is written to both SQL and MongoDB. Improvements have been made to ensure that the read write processes are fast for both databases, and it has been tried to be the most efficient. In SQL Server, index is added to all tables and all join operations are performed on these columns that have indexes. Figure 5.1. shows database schema in SQL Server. In all tables, key and date columns are indexed to improve query performance.

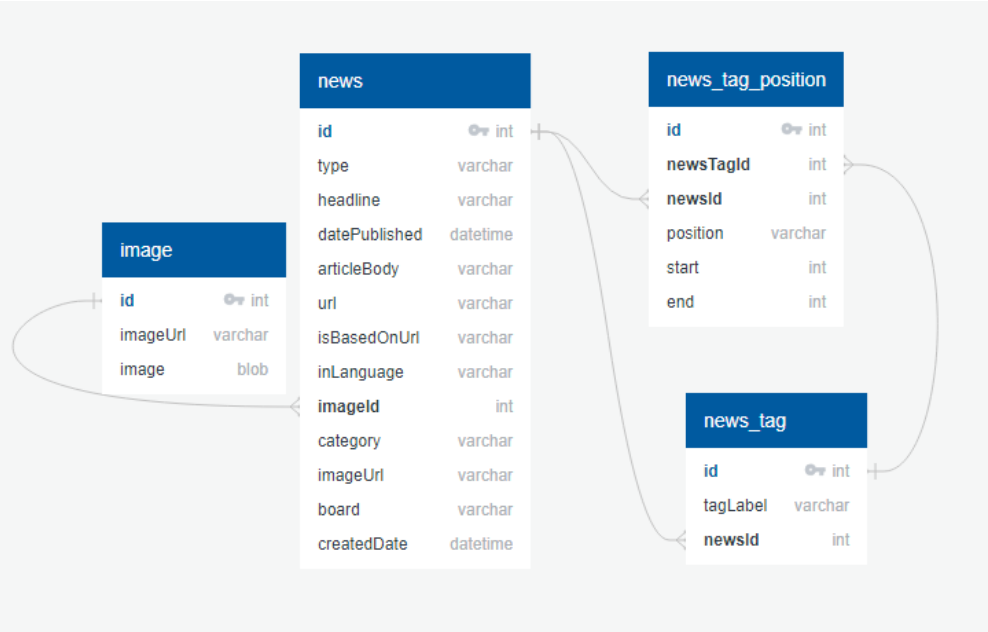

Figure 5.1. SQL Server Data Schema

Figure 5.2. show an example document structure in MongoDB. Since documents don't have a fixed schema, this structure can be different in other documents.

5.2. Tag Based Data Analysis

One of the basic needs is to accelerate the analysis of the incoming news, so it is tried to catch the agenda of the day on the news. For this, the labels of the news are used, the frequency of the tags together are checked.

For the daily analysis of the news, it is tried to find out the event of the day via news that is stored in both MongoDB and SQL. In order to analyze, the news data is kept in SQL, the Entity Framework is used to find the occurrences of the binary and triple tag

groups in the news. In MongoDB, using the map reduce structure, binary and triple tag groups are created and the frequencies of passing together on news are examined.

Key	Value	Type
(1) ObjectId("5d302336e18f2038601615aa")	{ 17 fields }	Document
_id	ObjectId("5d302336e18f2038601615aa")	ObjectId
articleBody	As Facebook Blockchain Lead David Marcus tries to simultaneously	String
url	https://www.forbes.com/sites/stevenehrlich/2019/07/17/as-facebook-	String
type	NewsArticle	String
newsId	https://www.forbes.com/sites/stevenehrlich/2019/07/17/as-facebook-	String
headline	As Facebook Struggles For Blockchain Support, A Truly Decentralize	String
category	http://cv.iptc.org/newscodes/mediatopic/20000171	String
inLanguage	en	String
isBasedOnUrl	http://www.forbes.com/	String
tagLabels	Array[32]	Array
tagDetails	Array[32]	Array
publishedDate	17.07.2019 17:00 +0000	String
imajUrl	https://thumbor.forbes.com/thumbor/600x315/https%3A%2F%2Fspe	String
createdDate	7/17/2019, 8:25:06 PM	Date
status	0	Int32
board	Şirketler	String
imaj		String

Figure 5.2. A document in MongoDB

In the first scenario, binary groups are created from news tags. The number of passes in the data set of these tag groups is calculated. Since the binary tag number in a news story can be very high, the weight of the tag in that news is calculated before this operation. If the tag is repeated less than two times in the news content, it is not included in the calculation.

- In MongoDB, this operation is done with map reduce. map reduce is written in JavaScript and is being executed using code from MongoDB C# driver.
- For SQL server; the entity framework was used to connect to the database, LINQ was used to code the map and reduce algorithms.

The following Figure 5.3. shows the time elapsed (in milliseconds) in MongoDB and SQL as the news count grows.

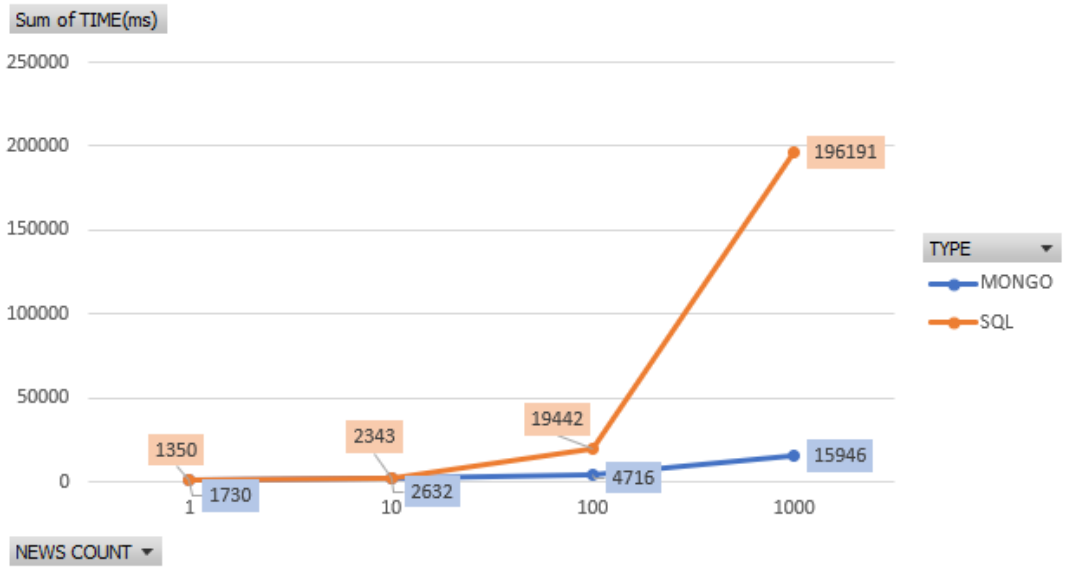

Figure 5.3. Time elapsed to find binary tag groups by queries on MongoDB and SQL Server

Figure 5.3. shows that reading data from SQL Server with Entity Framework is slowing down after the first 100 news. For this reason, instead of using Entity Framework in the process, by running the raw SQL command directly from the code, experiment was repeated. Figure 5.4. shows the time elapsed to find binary tag groups by queries on MongoDB and SQL Server without Entity Framework.

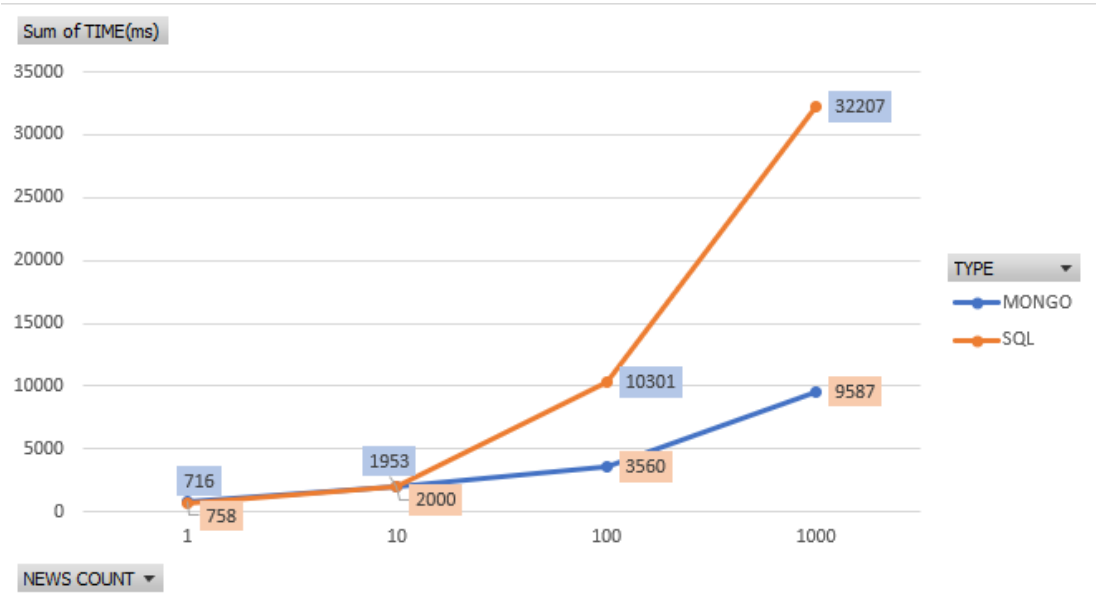

Figure 5.4. Time elapsed to find binary tag groups by queries on MongoDB and SQL Server without Entity Framework

In the second scenario, triple tag groups are created. The number of passes in the data set of these tag groups is calculated. Figure 5.5. shows the change in time as the data set grows.

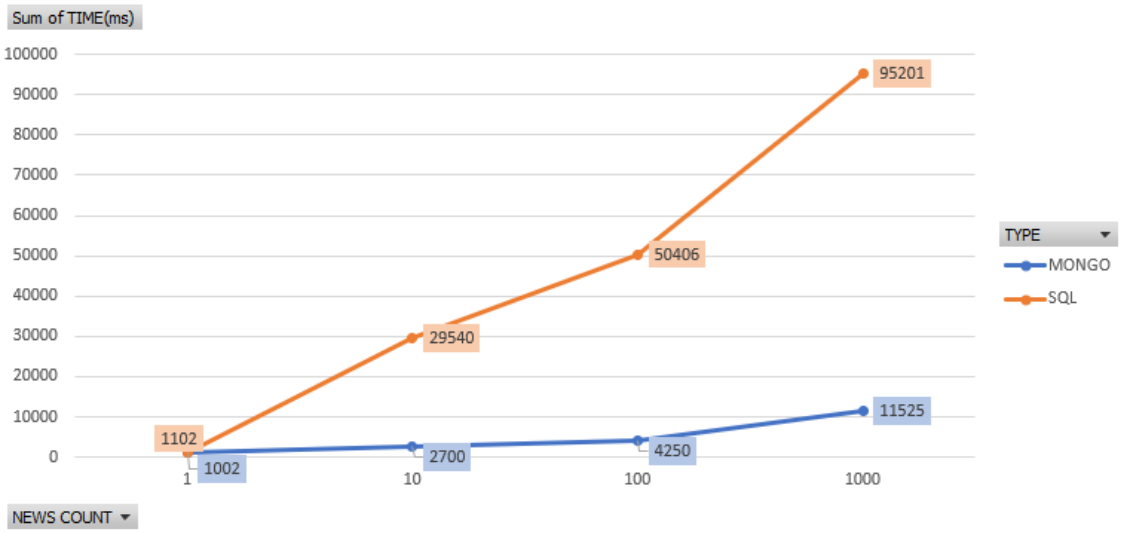

Figure 5.5. Time elapsed to find triples by queries on MongoDB and SQL Server to find triples

In both cases, with the increase in the number of news, the process of analyzing the label groups from the SQL Server starts to slow down. In relational database, when a new join is added to the query, it gets slower dramatically.

Also, when dataset contains more than 1000 news, SQL Server performance gets very slow even with one join operation. As a result, MongoDB's tag-based data retrieval processes work more efficiently as the query becomes more complex and the data set grows. For this reason, MongoDB has been selected as the data storage.

After the storage environment was decided, Machine Learning algorithms were used for detailed analysis of the data. In addition to news tags, text mining methods were used to examine in more detail.

5.3. Analyzing Data with Text Mining Methods

News objects contain article body and headline. These text data define news objects, so it is very useful for extracting relations between news. With clustering

methods, the agenda in a certain period can be perceived. Extracting agenda is important because instead of scanning all documents with eye, summary is checked.

As described in previous sections data is stored in MongoDB database in BSON format. BSON is very similar to JSON, so it can be used as the data set for clustering algorithm. Clustering algorithm is implemented in Python. Python has some libraries to connect MongoDB and pull data. In this thesis, PyMongo is used for communication with the database. PyMongo is the Python driver for MongoDB.

There are about 150,000 news documents in MongoDB collection. The following experiment was performed on data collected during a day, 2 April 2019. In this period, there are 3981 news collected. Statistics about news collected in this period are shown in the Table 5.1.

Table 5.1. Statistics about news

Total number of news	Average length of article body	Number of distinct tags	Most popular tag	Least popular tag
3981	4000	2202	Usa	Aba

Usa, China, Commerce, Turkey and Iran are the most frequent tags for this specific period. Figure 5.6. shows that this information show that Usa tag occurs most of the news. This insight also shows that these most common tags appear together.

Before start analyzing, article body texts must be preprocessed to simplify the words. So, the following operations were applied to the data set.

- Special characters are removed.
- Numbers are removed
- Texts are converted to lowercase
- Stop words are removed
- Lemmatization is applied

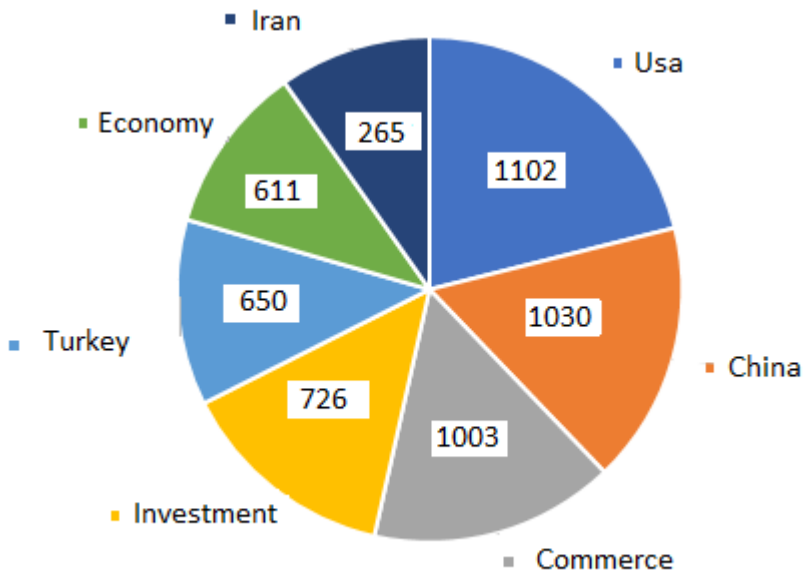

Figure 5.6. Most common tags and occurrences in news

At the end of these processes, all text documents are simplified and ready for text analysis.

Table 5.2. Example of document before and after processed

Before Being Processed	After Being Processed
(Reuters) - Numerous investigations spun out of U.S. Special Counsel Robert Mueller's probe are still alive and kicking, presenting potential ongoing legal…	reuter numerous investigation spun out of us special counsel robert mueller probe be still alive and kick present potential ongoing legal…

Next step sentences were converted into their vectoral representations. The Vector Space Model is a vector-defined state of a text object [31]. Each dimension of the vector corresponds to each term of the document. If a term is used in the text, its value in the vector cannot be zero. There are many methods of expressing the value of the term in the text. One of them is Tf-Idf method [20]. Tf-Idf weighting is calculated as the product of

Tf and Idf values.TF (Term Frequency) is the frequency of a term in a document. It specifies the number of times the word is used in the document. Idf (Inverse Document Frequency) is a measure of how much information the word contains. It specifies whether the word is common or rare in all documents. It is calculated as the logarithm of the number of documents containing the term of the total number of documents. The more documents a term refers to, the less information it contains. Commonly used terms, such as stop words, can appear in every document. If this term is used in every document, it can be said that it does not contain any information. After documents are converted to their vectoral representations Cosine Similarity method is used. The angle between document vectors is measured by this method and it is a measure of the similarity between texts [20].

	ability	able	about	above	abroad	absence	accelerator	accenture	accept	access	accommodate
0	0	0	4	0	0	0	0	0	0	1	0
1	0	0	2	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	1	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	1	0	1	0	0	0	0	0	0	10	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	3	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0

Figure 5.7. Document Term Matrix

	ability	able	about	above	abroad	absence	accelerator	accenture	accept	access	accommodate
0	0.000000	0.000000	0.005690	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.001422	0.000
1	0.000000	0.000000	0.002999	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
2	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
3	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
4	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
5	0.000000	0.000000	0.001779	0.00000	0.000000	0.000000	0.000000	0.001779	0.000000	0.000000	0.000
6	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
7	0.001063	0.000000	0.001063	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.010627	0.000
8	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
9	0.000000	0.000000	0.002825	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
10	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000

Figure 5.8. Normalized Document Term Matrix

Figure 5.7. shows the number of repetitions of words in the document. Figure 5.8. shows the normalized values of these occurrences. Resulting Tf-Idf matrix can be used to extract some information of shown in Table 5.3.

Table 5.3. Normalized Document Terms by two features for first 10 documents

	usa	china
0	0.002381	0.000000
1	0.004796	0.004796
2	0.013158	0.000000
3	0.013158	0.000000
4	0.037037	0.000000
5	0.003077	0.000000
6	0.037037	0.000000
7	0.001689	0.001689
8	0.037037	0.000000
9	0.001754	0.000000

Tf-Idf is important because of document similarity. By knowing which documents are similar it is easy to find related documents and group documents into clusters. For clustering algorithm, K-means clustering is used. It is important to choose the K, cluster count because it has effect on result. Without knowing the data in detail, it is hard to estimate the cluster count. The optimal number of clusters depends on the selected features which is chosen to measure similarity.

There are some methods to find optimal number of K. One of them is Elbow method. In Elbow Method [21], a range is selected for number of clusters. For each number K in this range K-Means algorithm runs. And for every K the Sum of Squares Error (SSE) is calculated. The aim is to find the smallest number K with low sum of squares error value. When the SSE values are displayed as a graph by cluster counts, the point where the SSE decrease starts to slow gives the optimal number of cluster count.

In text databases, a document collection defined by a document by term D matrix (of size m by n, m: number of documents, n: number of terms) number of clusters can roughly be estimated by the following formula m * n / t , where t is the number of non-zero entries in D. Note that in D each row and each column must contain at least one non-zero element [22].

In Figure 5.9. it is shown that for this dataset 8 cluster is the optimal number of clusters.

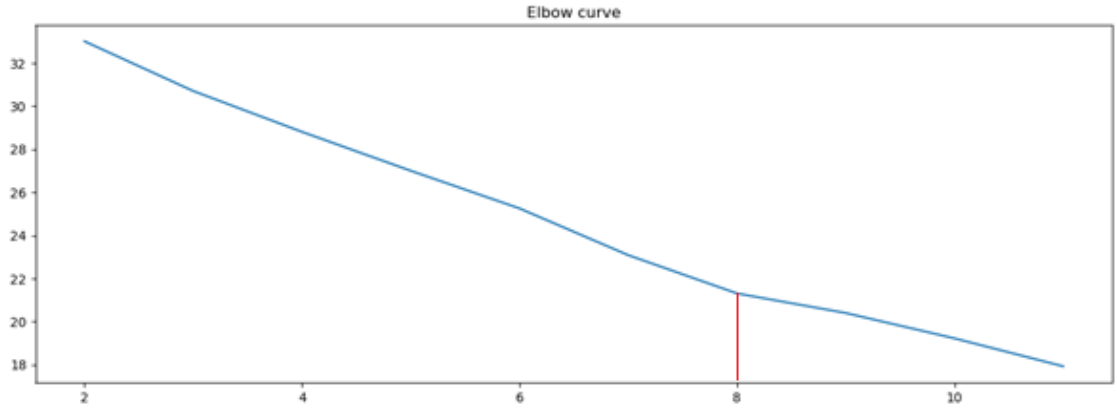

Figure 5.9. Elbow Method

Initially algorithm is run with to split data into eight clusters. Figure 5.10. shows the words that are descriptive for each cluster.

```
Top terms per cluster:
Cluster 0: lng gas project natural carrier energy vessel company china agreement
Cluster 1: company business market logistics service year billion investment one
Cluster 2: port terminal container cargo shipping news vessel service ship
Cluster 3: vessel ship shipping maritime tanker fuel fleet company sea
Cluster 4: percent stock index market usa share bank rose growth data
Cluster 5: sure review network policy information activity term support make service
Cluster 6: oil crude fuel usa price production per export company market
Cluster 7: country china trade usa state president russia government april said
```

Figure 5.10. Top terms per cluster

After running K-Means algorithm for eight clusters, Multidimensional Scaling (MDS) is used to visualize similarity in data set. MDS is a way of visualizing that the level of similarity between documents in a data set [23]. The result of this visualization is demonstrated Figure 5.11. These figure shows the distribution of eight clusters. Each color represents a cluster. There are some clear clusters like dark green, yellow and light green cluster. However, there are also some clusters that data is noisy.

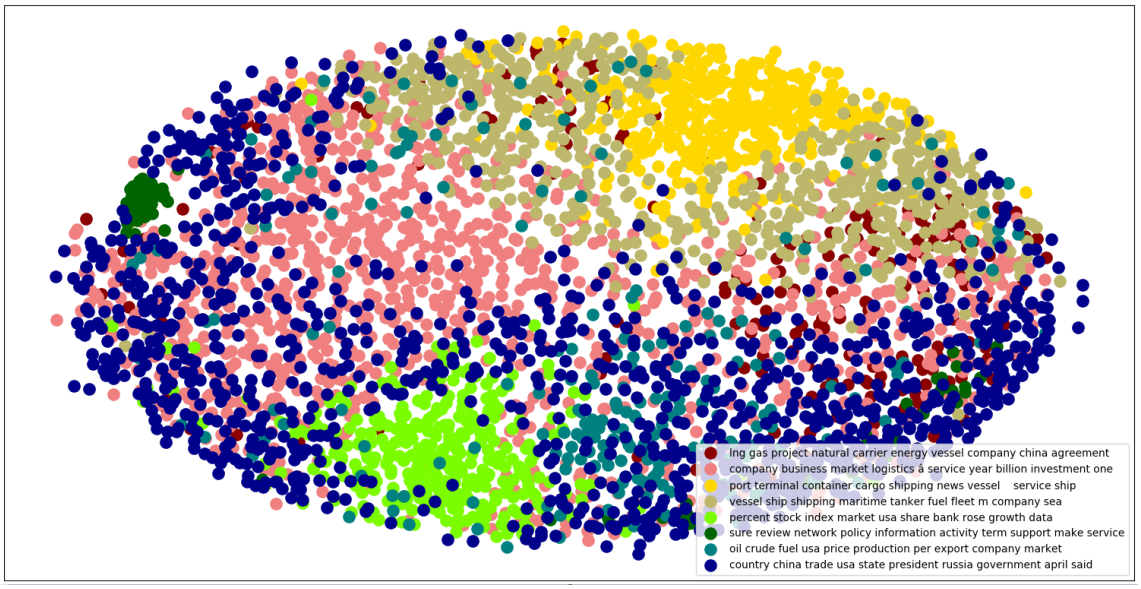

Figure 5.11. K-Means Clustering

The same data set is used in Hierarchical Clustering. To calculate the distance Ward's method is used [30]. In Hierarchical Clustering, news documents are displayed according to their similarities as shown in Figure 5.12. Initially every document is considered as a single cluster and shown on the x axis. Then two most similar documents are combined into a new cluster until there is one big cluster left.

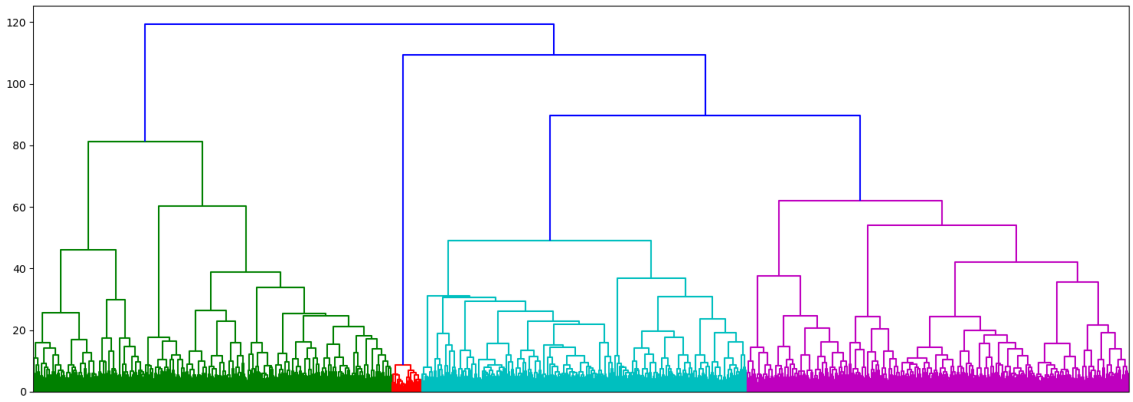

Figure 5.12. Hierarchical Clustering

In Figure 5.12., there is a small group displayed in red. When the small piece of the documents in this group are examined in detail, the following Figure is obtained.

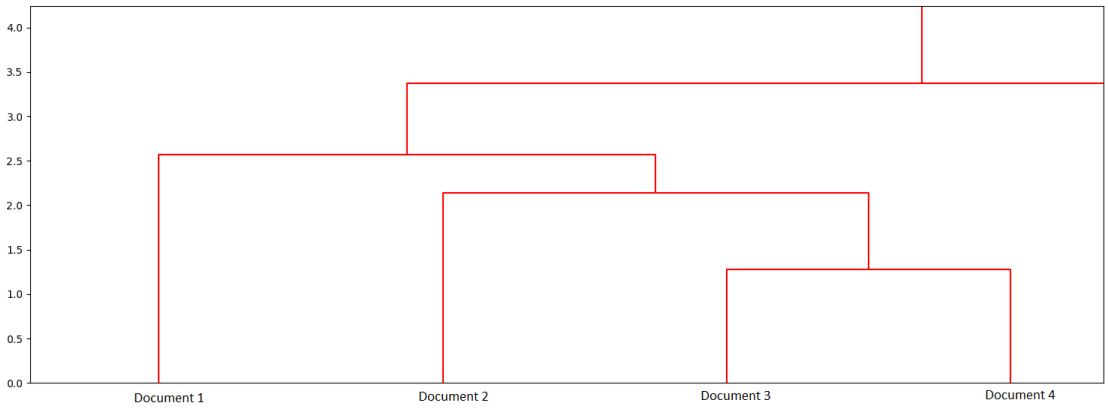

Figure 5.13. Hierarchical Clustering in detail

Figure 5.13. shows that document 3 and document 4 are the most related documents. They are similar with document 2. Document 1 is similar with them. Table 5.4. shows the headlines of these documents.

Table 5.4. Documents displayed in hierarchical clustering example

Document 1: The Asian Development Bank has cut its economic growth forecasts for India and Southeast Asia for 2019
Document 2: India power industry welcomed a top court decision to strike down central bank rules that tightened loan default guidelines
Document 3: More of India smaller banks may become acquisition targets
Document 4: Reserve Bank of India Governor Shaktikanta Das has an “out of the box” suggestion

When the result of hierarchical clustering is examined in detail, it is seen that the results are quite successful in finding similarities.

5.4. Clustered Search Results

In a screen user enters date and keywords to search news articles. The related results are returned against user query. Firstly, these keywords and date is used to query database. Then the documents in the query result are given to clustering algorithm as an

input dataset. Clustering results are displayed in the user interface. Figure 5.14. shows the screen that user enters search parameters.

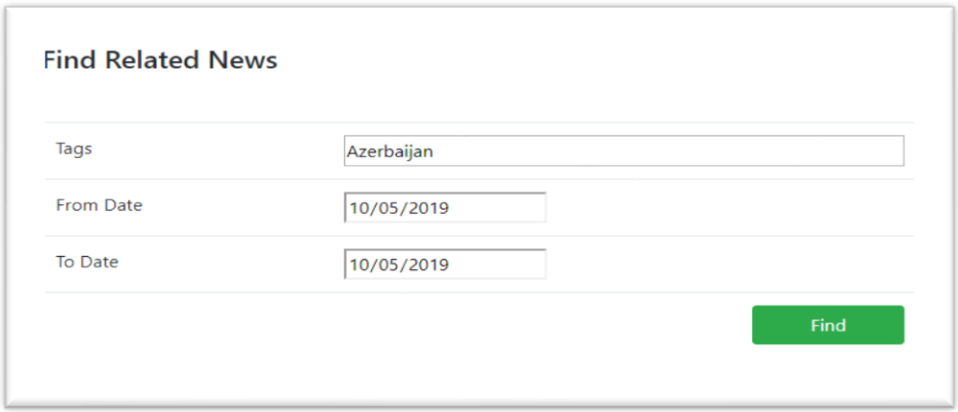

Figure 5.14. Search screen

When user clicks the find button process begins. The results of search with these parameters are returned and given to the clustering algorithms as input set. On this period there are 58 news objects containing the parameter of "Azerbaijan". This experiment and other sample queries are tried both MongoDB and SQL Server. Table 5.5. shows that basic searches perform similar in both databases.

Table 5.5. Query times in milliseconds

Query #	SQL Server	MongoDB
1	597	936
2	3173	2881
3	4205	4155
4	4532	4860
5	5690	5530

K-Means and Hierarchical Clustering algorithms are performed on the same dataset. Clustering results are shown below. Figure 5.15 shows the K-Means clustering results. Figure 5.16 shows the Hierarchical clustering results.

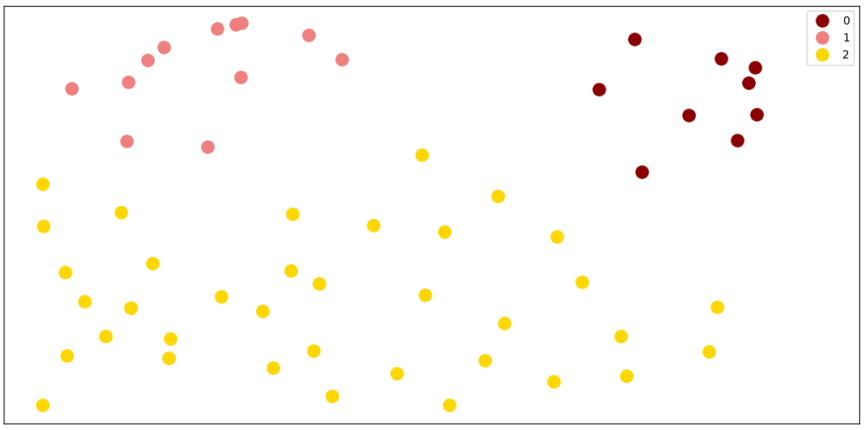

Figure 5.15. K-Means results for search

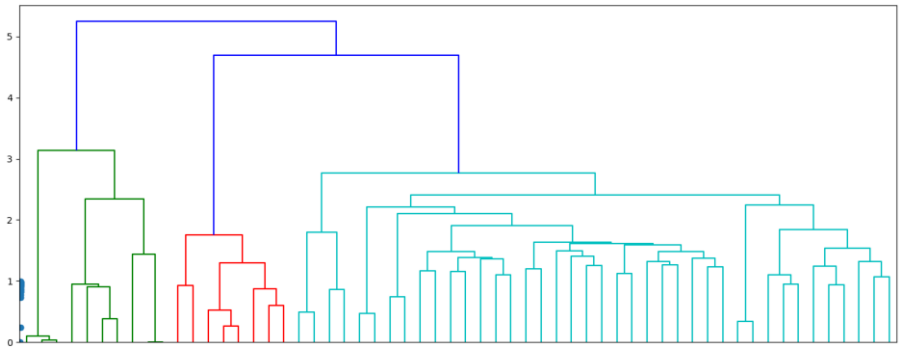

Figure 5.16. Hierarchical clustering results for search

5.5. Evaluating Clustering Results

Clustering algorithms are unsupervised learning methods and runs on dataset which is not labelled. So, it is hard to measure performance of the algorithm. To get some insights about performances of two clustering algorithms, news tags are used. Table 5.6. shows the news that are assigned to the same cluster in Hierarchical clustering. This news objects belong the red cluster in Figure 5.16. When tags are examined in detail, it can be seen that tags are similar to each other. In different clusters, tags are mostly different. K-Means algorithm assigns one more object to this cluster. In Figure 5.15. brown cluster is equivalent to the red cluster in Figure 5.16.

Table 5.6. News headlines and tags for the first cluster

headline	tags
How much foreign companies invested in Azerbaijan's oil & gas sector?	Sector,Oil,Gas,Azerbaijan,Energy,Turkey,Georgia,Sea,Albania,Bulgaria,Chevron,Greece,Italy,BP., Petronas
Envoy: It is important to complete TAP project in accordance with schedule	Sea,Leg,Italy,EU,Chain,Energy,TREND,Albania,Greece,Cross,Azerbaijan,Construction,BP.,Network ,Natural, Oil
Turkey's illegal drilling in Cyprus' EEZ revives geopolitical rivalries in Eastern	Cyprus,Turkey,Export,Russia,EU,Natural,Network,Energy,Egypt,Oil,Israel,Greece,Italy,Haifa,Chair, Alternative,Georgia,Libra,Taurus,Stretch,Water,Transfer,Azerbaijan,Flocked,Gas Pipeline,Investment,Construction,Inject
Southern Gas Corridor may be expanded	Gas,Pipeline,Natural,Turkmenistan,Construction,Turkey,TREND,Georgia,Azerbaijan,EU,Oil,Italy,E nergy,Sea,Albania,Case,Greece,Eskişehir,City,Summer,Western,Opening Ceremony,Yatak Odası Ürünleri,Maison,Desenli
Hungary interested in Southern Gas Corridor's reaching Central Europe	Hungary,PORT of CONSTANTA,Opening Ceremony,Terminal,Azerbaijan,European Union,Energy,Romania,Oil,Coast,Georgia,Sea,Black,Made,Trade,Erzurum,Sector,Austria,Bulgaria, Natural,Solution,Turkey,Eskişehir,Beauty
USA Senator: TAP, IGB are on track to become reality	Cross,Bulgaria,Greece,EU,Azerbaijan,BP.,Export,Energy,Sea,Chain,TREND,Israel,Cyprus,Made,Tra nsit,Natural,Construction,Network,Italy,Gas Pipeline,Albania
Italian PM: EastMed could join TAP	Leg,Cyprus,EU,Construction,Azerbaijan,BP.,TREND,Network,Natural,Sea,Italy,Chain,Energy,City,A lbania,Greece,Cross,Gas Pipeline,Israel,Desenli
Drilling of second wellbore at Absheron field underway	Sea,Azerbaijan,TREND,Stand,Summer,Oil,Kitchen Product,Maison

Table 5.7. shows the news that are another cluster in this dataset. K-Means algorithm assigns two more objects to this cluster. In Figure 5.15. pink cluster is equivalent to the green cluster in Figure 5.16.

Table 5.7. News headlines and tags for the second cluster

headline	tags
Weekly review of Azerbaijani currency market	Turkish Lira,Official,Exchange Rate,TREND,Set,Azerbaijan,Dollar
Currency rates for May	Dollar,Hong Kong,Azerbaijan,New Zealand,Polish,Set,Taiwan,Turkish Lira,Exchange Rate,Official,TREND
Azerbaijani Currency rates for May 10	Dollar,Azerbaijan,Turkish Lira,Exchange Rate,Official
Demand exceeds supply for Azerbaijani Central Bank's notes	Short,Azerbaijan,Cut-off,Only,TREND,Exchange,Set,Alt
Most of daily turnover at Baku Stock Exchange accounts for CBA notes	Exchange,Bonds,TREND,Azerbaijan,NOTE
Baku Stock Exchange to hold auction for placement of mortgage bonds	Bonds,Exchange,Next,TREND,Azerbaijan,Bond
Gold, silver prices down in Azerbaijan	Azerbaijan,Silver,Gold,Iridium,PALLADIUM,Platinum,TREND, Rate
Gold prices in Azerbaijan	Silver,Gold,Iridium,Platinum,TREND,Azerbaijan
Weekly turnover at Baku Stock Exchange exceeds 252M manats	Exchange,Azerbaijan,Bonds,Dollar,TREND,Oil
Gold, silver prices up in Azerbaijan	Azerbaijan,Silver,Gold,Iridium,PALLADIUM,Platinum,TREND

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1. Conclusion

Today, it is very important for corporate companies to follow the news and to be aware of the events in advance. Because managers who are in a strategic position within the organization take decisions based on these news'. For this purpose, there is a department which analyzes the collected news. However, it is quite time consuming to read and analyze a large number of news on the Internet. For this reason, the need for storage and retrieval of the news was formed for the data that captured by desired keywords.

In order to decide the storage environment of the transferred data, the same data set was kept in both the Microsoft SQL Server, which is the relational database and the MongoDB, which is the NoSQL database. Then, databases were analyzed against complex queries for the sample scenarios. These scenarios have been created according to current needs in order to make sense of the news data set without examining one by one. Experiments have shown that although the small data set has worked fast, SQL Server's performance has decreased rapidly as the data set grows beyond.

News objects also have an article body. This text can be used to investigate dataset in more detail. That's why news content was analyzed by clustering algorithms. The same dataset is analyzed by K-Means and Hierarchical Clustering. K-Means algorithm is successful in identifying news groups which are very different from others. However, news dataset isn't well separated into clusters. In other words, clusters with clear borders may not be common in the news dataset. That's why Hierarchical Clustering performs better. Instead of trying to find a cluster in the news data set, focusing on the similarities of the news gives better results.

6.2. Future Work

In addition to the tags sent from the third-party source, the labels extracted from the text with the data mining methods can be written to the database. It is possible to use indexing array in MongoDB to return faster results. In this way, users can get quick answers to search results and easily find related news.

In this thesis, it is proposed to store images with other news data in a document in MongoDB. Since MongoDB has a specification for storing files, it can be used to store images instead of storing all data in a document. This reduces the document size in MongoDB and may affect the performance of the query.

Data analysis, which is planned to be performed daily, can be examined in a certain period and it can be checked whether the results will change. It may not be needed to analyze the news daily, and this range may be extended. Algorithms can be improved, and more accurate results can be obtained.

In this thesis, it is decided to renew the Tf-Idf calculation every day. However, if the matrix does not change daily, the recalculation interval can be extended by observing the change in the Tf-Idf matrix.

One of the information that needs to be analyzed in corporate companies is to see how a specific issue develops over time. This analysis, which is currently being performed manually, can be done by machine learning algorithms. News on a topic can be found by taking advantage of topic tracking algorithms.

The frequent occurrence of a word in a sentence does not necessarily mean that it is of great importance in the sentence. For this reason, analyzing the text in semantic detail can improve the results. In this thesis, only structural text similarity is studied. In the following periods, news dataset can be analyzed by considering semantic similarity.

REFERENCES

1. Parker, Zachary, Scott Poe, and Susan V. Vrbsky. "Comparing nosql mongodb to an sql db. " Proceedings of the 51st ACM Southeast Conference. ACM, 2013.

2. Wei-Ping, Zhu, L. I. Ming-Xin, and Chen Huan. "Using MongoDB to implement textbook management system instead of MySQL." 2011 IEEE 3rd International Conference on Communication Software and Networks. IEEE, 2011.

3. Abramova, Veronika, Jorge Bernardino, and Pedro Furtado. "Which nosql database? a performance overview. " Open Journal of Databases (OJDB) 1.2 (2014): 17-24.

4. Kailas, Gaurav. (2019). "A Comparison of NoSQL Database Systems: A Study On MongoDB and Apache HBase".

5. Potok, T., et al. "VIPAR: Advanced Information Agents discovering knowledge in an open and changing environment." Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics, Orlando FL. 2003.

6. Cattell, Rick. "Scalable SQL and NoSQL data stores." Acm Sigmod Record 39.4 (2011): 12-27.

7. Huang, Anna. "Similarity measures for text document clustering." Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. Vol. 4. 2008.

8. Nayak, Ameya, Anil Poriya, and Dikshay Poojary. "Type of NOSQL databases and its comparison with relational databases." International Journal of Applied Information Systems5.4 (2013): 16-19.

9. Fung, Benjamin CM, Ke Wang, and Martin Ester. "Hierarchical document clustering using frequent itemsets." Proceedings of the 2003 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2003.

10. Zhao, Ying, and George Karypis. "Evaluation of hierarchical clustering algorithms for document datasets." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.

11. Haerder, T.; Reuter, A. (1983). "Principles of transaction-oriented database recovery". ACM Computing Surveys. 15 (4): 287.

12. Pritchett, Dan. "Base: An acid alternative." Queue 6.3 (2008): 48-55.

13. Han, Jing, et al. "Survey on NoSQL database." 2011 6th international conference on pervasive computing and applications. IEEE, 2011.

14. Moniruzzaman, A. B. M., and Syed Akhter Hossain. "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison." arXiv preprint arXiv:1307.0191 (2013).

15. Rich, Elaine, and Kevin Knight (1994). "Artificial Intelligence". McGraw-Hill

16. Buchanan, Bruce. (2005). "A (Very) Brief History of Artificial Intelligence". AI Magazine. 26. 53-60.

17. Royal Society Working Group. Machine learning: the power and promise of computers that learn by example. Technical report, 2017.

18. Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666

19. Zhao, Ying, George Karypis, and Usama Fayyad. "Hierarchical clustering algorithms for document datasets." Data mining and knowledge discovery 10.2 (2005): 141-168.

20. Chowdhury, Gobinda G. Introduction to modern information retrieval. Facet publishing, 2010.

21. Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1.6 (2013): 90-95.

22. Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". ACM Transactions on Database Systems.

23. Borg, Ingwer, and Patrick Groenen. "Modern multidimensional scaling: Theory and applications." Journal of Educational Measurement 40.3 (2003): 277-280.

24. Vargas, Kathryn. The Main NoSQL Database Types. Studio 3T, 3T Software Labs Limited, 21 Feb 2018, https://studio3t.com/whats-new/nosql-database-types/.

25. "Machine Learning in MATLAB. " MathWorks, The MathWorks, Inc. , https://in.mathworks.com/help/stats/machine-learning-in-matlab.html?w.mathworks.com

26. Kisliakovskii, Ilia, et al. "Towards a simulation-based framework for decision support in healthcare quality assessment." Procedia computer science 119 (2017): 207-214.

27. Gomaa, Wael H., and Aly A. Fahmy. "A survey of text similarity approaches." International Journal of Computer Applications68.13 (2013): 13-18.

28. Pandya, Abhinay, and Pushpak Bhattacharyya. "Text similarity measurement using concept representation of texts." International Conference on Pattern Recognition and Machine Intelligence. Springer, Berlin, Heidelberg, 2005.

29. Hotho, Andreas, Steffen Staab, and Gerd Stumme. "Ontologies improve text document clustering." Third IEEE international conference on data mining. IEEE, 2003.

30. Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function." Journal of the American statistical association 58.301 (1963): 236-244.

31. Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.