A Comparative Study of Data Mining and Information Retrieval

Akhetuamen Sylvester[1], Mark Uvietesiri[2], Obi, Morgan Onwemadu[3], Ubani, Chinyere[4]

[1]Department of Computer Science, Auchi Polytechnic, Edo State, Nigeria.
[2]Ministry of Science and Technology Project e-Delta, Asaba, Delta State, Nigeria.
[3]Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria.
[4]Department of Computer Science, Captain Elechi Amadi Polytechnic, Rumuola, Port Harcourt, Rivers State, Nigeria.

Abstract - *This write up describes the concept of data mining and information retrieval, focusing basically on the type of tasks that can be solved with data mining, the technologies used to perform these tasks, the advantages and challenges of data mining. It also examines the concept of information retrieval looking at the challenges in retrieving information, the measures for information retrieval and the methods of information retrieval.*

Keywords *– automated, data mining, decision making, information retrieval, precision,*

1.0 Introduction

Today we hear about the term information overload. This is a concept coined in the year 1964 by Professor Bertram Gross in his work "The Managing of Organizations". However, it was popularized by Alvin Toffler, the American writer and futurist, in his book "Future Shock" in 1970. The issue of information overload drastically affects decision making, innovation, and productivity (Hemp, 2009); this is because we are being exposed to a large surge of information from large databases and web contents. As the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly. Lying hidden in all this data is information, potentially useful information that is rarely made explicit or taken advantage of. This problem brings about the need for data mining and information retrieval. Data Mining and Information Retrieval is the blending of scientific finding and practice, whose subject matter is to collect, manage, process, analyze, and visualize the vast amount of structured or unstructured data. It has grown dramatically and became more institutionalized in the 21st Century (Liu, et al., 2019).

2.0 The Concept of Data Mining

This is the extraction of hidden predictive information from large databases. It provide a means of predicting future events from an existing store of data using some advanced algorithms, powerful multiprocessor computers, massive databases, etc (Leon & Leon, 1999).

Data mining uses the discovery model to create information about data. According to Leon et al. (1999), data mining uses methodologies that can sift through the data in search of frequently occurring patterns, can detect trends, produce generalizations about data, etc. These tools can discover these types of information with very little or no guidance from the user as illustrated in the data mining process in figure 1. The discovery of facts in data mining is not a result of a haphazard. A well designed data mining tool is one that is designed and built so that the exploration of the data is done in such a way as to yield as large a number of useful facts about the data as possible in the shortest amount of time.

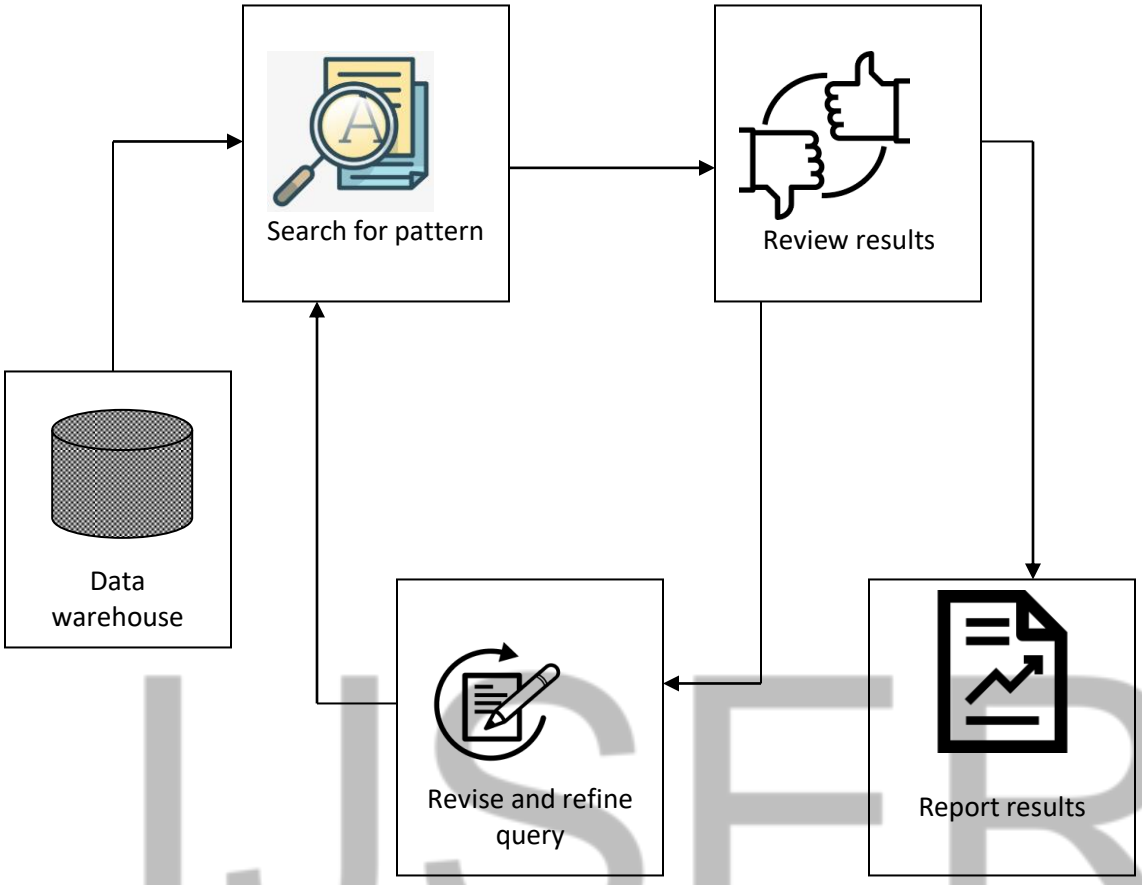

Figure 1: Data mining process (Leon & Leon, 1999)

3.1 Tasks Solved by Data Mining

According to Leon et al., (1999), the main tasks that are solved by a data mining system include:

Predicting: this involve learning a pattern from examples and using the developed model to predict future values of the target variable.

Classification: involves finding a function that maps an example into one of several discrete classes.

Detection of relations: involves searching for the most influential independent variables for a selected target variable

Explicit modeling: this task involves finding explicit formulae describing dependencies between various variables.

Clustering: this task involves identifying a finite set of categories or clusters that describe data.

Deviation detection: involves determining the most significant changes in some key measures of data from previous or expected values.

3.2 Advantages of Data Mining

Automated Decision-Making

Data Mining allows organizations to continually analyze data and automate both routine and critical decisions without the delay of human judgment. Banks can instantly detect fraudulent transactions, request verification, and even secure personal information to protect customers against identity theft. Deployed within a firm's operational algorithms, these models can collect, analyze, and act on data independently to streamline decision making and enhance the daily processes of an organization.

Accurate Prediction and Forecasting

Planning is a critical process within every organization. Data mining facilitates planning and provides managers with reliable forecasts based on past trends and current conditions.

Cost Reduction

Data mining allows for more efficient use and allocation of resources. Organizations can plan and make automated decisions with accurate forecasts that will result in maximum cost reduction.

Customer Insights
Firms deploy data mining models from customer data to uncover key characteristics and differences among their customers. Data mining can be used to create personas and personalize each touch point to improve overall customer experience.

3.3 Technologies used in Data Mining

The most commonly used techniques in data mining include:

i. **Neural networks:** non-linear predictive models that learn through training and resemble biological neural networks in structure.
ii. **Rule reduction:** the extraction of useful if-then rules from data based on statistical significance.
iii. **Evolutionary programming:** according to Leon et al, (1999) evolutionary programming is the youngest and evidently the most promising branch of data mining. The underlying idea of the method is that the system automatically formulates hypotheses about the dependence of the target variable on other variables in the form of programs expressed in an internal programming language.
iv. **Case base reasoning (CBR):** the main idea underlying this method is very simple. To forecast a future situation, or to make a correct decision, such systems find the closet past analogs of the present situation and choose the same solution, which was the right one in those past situations. This method is often called nearest neighbour method.
v. **Decision trees:** tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a data set.
vi. **Genetic algorithms:** optimization techniques that use processes such as genetic combination, mutation, and neural selection in a design based on the concepts of evolution.

3.4 Challenges of Data Mining

While a powerful process, data mining is hindered by the increasing quantity and complexity of big data. Where Exabyte of data are collected by firms every day, decision-makers need ways to extract, analyze, and gain insight from their abundant repository of data.
The challenges of big data are prolific and penetrate every field that collects, stores, and analyzes data. Big data is characterized by four major challenges: volume, variety, veracity, and velocity. The goal of data mining is to mediate these challenges and unlock the data's value.
Volume describes the challenge of storing and processing the enormous quantity of data collected by organizations. This enormous amount of data presents two major challenges: first, it is more difficult to find the correct data, and second, it slows down the processing speed of data mining tools.
Variety encompasses the many different types of data collected and stored. Data mining tools must be equipped to simultaneously process a wide array of data formats. Failing to focus an analysis on both structured and unstructured data inhibits the value added by data mining.
Velocity details the increasing speed at which new data is created, collected, and stored. While volume refers to increasing storage requirement and variety refers to the increasing types of data, velocity is the challenge associated with the rapidly increasing rate of data generation.
Finally, **veracity** acknowledges that not all data is equally accurate. Data can be messy, incomplete, improperly collected, and even biased. With anything, the quicker data is collected, the more errors will manifest within the data. The challenge of veracity is to balance the quantity of data with its quality.
These challenges of big data presents the following challenges to data mining

i. Over-Fitting Models

Over-fitting occurs when a model explains the natural errors within the sample instead of the underlying trends of the population. Over-fitted models are often overly complex and utilize an excess of independent variables to generate a prediction. Therefore, the risk of over-fitting is heighted by the increase in volume and variety of data. Too few variables make the model irrelevant, where as too many variables restrict the model to the known sample data. The challenge is to moderate the number of variables used in data mining models and balance its predictive power with accuracy.

ii. Cost of Scale

As data velocity continues to increase data's volume and variety, firms must scale these models and apply them across the entire organization. Unlocking the full benefits of data mining with these models requires significant investment in computing infrastructure and processing power. To reach scale, organizations must purchase and maintain powerful computers, servers, and software designed to handle the firm's large quantity and variety of data.

iii. Privacy and Security

The increased storage requirement of data has forced many firms to turn toward cloud computing and storage. While the cloud has empowered many modern advances in data mining, the nature of the service creates significant privacy and security threats. Organizations must protect their data from malicious figures to maintain the trust of their partners and customers.
With data privacy comes the need for organizations to develop internal rules and constraints on the use and implementation of a customer's data. Data mining is a powerful tool that provides businesses with compelling insights into their consumers. However, at what point do these insights infringe on an individual's privacy? Organizations must weigh this relationship with their customers, develop policies to benefit consumers, and

communicate these policies to the consumers to maintain a trustworthy relationship.

4.0 Information Retrieval

Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents (Han & Kamber, 2006).

According to Babu et al., (2014), the aim of efficient information retrieval should be to retrieve that information, and only that information which is deemed relevant to a given query. They further stated that a typical information retrieval system selects documents from a collection in response to a user's query and ranks these documents according to their relevance to the query (Babu, Ali, & Rao, 2014). This is primarily accomplished by matching a text representation with a representation of the query.

Information retrieval has found many applications in areas such as such as on-line library catalog systems, on-line document management systems, and more recently Web search engines.

4.1 Problems of Information Retrieval

Search based on keywords: A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some ad hoc (i.e., short-term) information need, such as finding information to buy a used car. When a user has a long-term information need (e.g., a researcher's interests), a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process is called information filtering, and the corresponding systems are often called filtering systems or recommender systems. From a technical viewpoint, however, search and filtering share many common techniques.

Uncertainty: The consent of the user is another major problem. Going through user's consent leads us to the study of his search behavior. And this, in turn, leads us to the uncertainty and probability of one's decision-making. Stating user's behavior and understanding his information needs highly affects the organization and operation of the information retrieval system (Ricard & Berthier, 1999) and may help us in predicting his decision-making.

Decision-making is connected to many factors in practice. Decision science as defined by Brugha (2001) is mostly related to philosophy, information system, psychology, culture, and management. And this is the reason that prediction of what the user wants or decides becomes very difficult. Brugha shows that the user's need and what he is looking for relates to his type of thinking. The subject approach has become a major way of finding information in the electronic age. Search engines have tried to fill the void on the Internet, yet users became more frustrated with the thousands of so called "hits" beyond their desired results.

4.2 Measures for Text Retrieval: Precision and Recall

Although there are various methods to evaluate the retrieval process as well as classification activity, recall and precision are highly recommended by authors (Araghi, 2005).

Precision: is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

$$precision = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{retrieved\}|}$$

Recall: is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$precision = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{relevant\}|}$$

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision given as:

$$f - score = \frac{recall * precision}{(recall + precision)/2}$$

Information retrieval often finds application in
Web search
Desktop search
Library search etc.

4.3 Information Retrieval Methods

Broadly speaking, retrieval methods fall into two categories: They generally either view the retrieval problem as a document selection problem or as a document ranking problem.

Document selection methods: in this method, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the Boolean retrieval model, in which a document is represented by a set of keywords and a user provides a Boolean expression of keywords, such as "car and repair shops," "tea or coffee," or "database systems but not Oracle." The retrieval system would take such a Boolean query and return documents that satisfy the Boolean expression. Because of the difficulty in prescribing a user's information need exactly with a Boolean query, the Boolean retrieval method generally only works well when the user knows a lot about the document collection and can formulate a good query in this way.

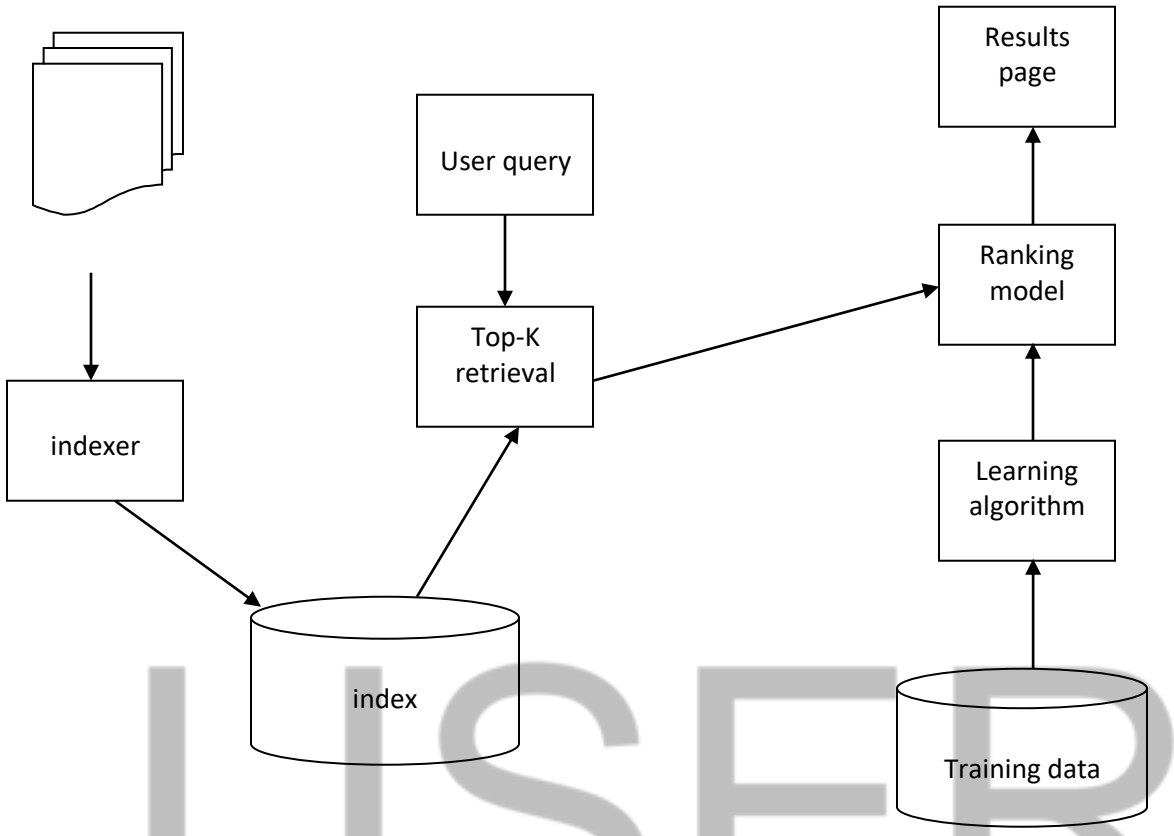

Figure 2: Architecture of Information Retrieval (MLMinds, 2018)

Document ranking methods: this method use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods. Most modern information retrieval systems present a ranked list of documents in response to a user's keyword query. There are many different ranking methods based on a large spectrum of mathematical foundations, including algebra, logic, probability, and statistics. The common intuition behind all of these methods is that we may match the keywords in a query with those in the documents and score each document based on how well it matches the query. The goal is to approximate the degree of relevance of a document with a score computed based on information such as the frequency of words in the document and the whole collection. Notice that it is inherently difficult to provide a precise measure of the degree of relevance between a set of keywords. For example, it is difficult to quantify the distance between data mining and data analysis. Comprehensive empirical evaluation is thus essential for validating any retrieval method.

Conclusion

Data mining and information retrieval are emerging trends to beat the challenge of information overload. These concepts major deals with fast and easy retrieval of customized information for quick decision making. In general, information retrieval from this research can be seen as a subset of data mining. While data mining deals on a large scale with discovering patterns from large stores of databases, information retrieval deals with retrieving customize information for a user from large collection of documents which could also include databases.

References

[1] Araghi, G. F. (2005). Major Problems in Retrieval Systems. Cataloging & Classification Quarterly,, (pp. 43-53).

[2] Babu, S., Ali, A., & Rao, S. (2014). A Study on Information Retrieval Methods in Text Mining. International Journal of Engineering Research & Technology , 184-190.

[3] Brugha, C. (2001). Implication from Decision Science for the Systems Development Life Cycle in Information Systems. Information System Frontier.

[4] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann.

[5] Hemp, P. (2009, September). Harvard Business Review. Retrieved June 2020, from Death by Information Overload: https://hbr.org/2009/09/death-by-information-overload

[6] Leon, A., & Leon, M. (1999). Database Management Systems. New Delhi: Vikas Publishing House.

[7] Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., et al. (2019). Data Mining and Information Retrieval in the 21st century: A bibliographic review. Elsevier , 1-13.

[8] MLMinds (Director). (2018). Intoduction to Information Retrieval [Motion Picture].

[9] Ricard, o. B., & Berthier, R.-N. (1999). Modern Information Retrieval. New York: ACM Press, Addison-Wesley.

[10] Whitten, T. H., & Frank, E. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.

1.