A STUDY ON EXPERT SEARCHING FOR UNSTRUCTURED DATA

V SANGEETHA
Research Scholar
Department of Computer Science and Engineering
Shri Jagdish Prasad Jhabarmal Tibrewala University

DR. PRASADU PEDDI
CO-GUIDE
Department of Computer Science and Engineering
Shri Jagdish Prasad Jhabarmal Tibrewala University

DR. MARAM ASHOK
Department of Computer Science and Engineering
Principal
Malla Reddy Institute Of Engineering and Technology

Abstract

With the emergence of new channels and technologies such as social networking, mobile computing, and online advertising, the data generated no longer have a standard format or structure like the conventional ones and cannot be processed using relational models. They come in the form of text, XML, emails, images, weblogs, videos, and so on resulting in a surge of new data types. This formless data is either semi-structured or unstructured data and makes searching and analysis complex. Analyzing and surveying unstructured data mining and challenges in Big Data is the primary focus of this paper. Analyzing Big Data strategy issues plus implementing tools and techniques like Hadoop open source software framework would be part of future work.

***Keywords**: unstructured, Format, Structured Data*.

Introduction

Unstructured data, typically categorized as qualitative data, cannot be processed and analyzed via conventional data tools and methods. Since unstructured data does not have a predefined data model, it is best managed in non-relational (NoSQL) databases. Another way to manage unstructured data is to use data lakes to preserve it in raw form. The importance of unstructured data is rapidly increasing. Recent projections indicate that unstructured data is over 80% of all enterprise data, while 95% of businesses prioritize unstructured data management.

Data has been originally generated by organizational employees but recently it scaled-up to user generated and lately to machine generated given colossal amount of data that needs large storage and processing capability on a daily basis. The size of unstructured data generated by companies like those oil drilling, airlines, social networks, marketing, and others is tending to thousands of terabytes and the unstructured data is valuable that the companies are now devising method of extracting meaning from them. Due to this massive size of data and the variety of data types traditional data processing tools can no longer handle them. Data volume has grown exponentially because of the explosion of machine generated data and from growing human engagement within the social networks.

Literature review

Kiran Adnan (2019) Process of information extraction (IE) is used to extract useful information from unstructured or semi-structured data. Big data arise new challenges for IE techniques with the rapid growth of multifaceted also called as multidimensional unstructured data. Traditional IE systems are inefficient to deal with this huge deluge of unstructured big data. The volume and variety of big data demand to improve the computational capabilities of these IE systems. It is necessary to understand the competency and limitations of the existing IE techniques related to data pre-processing, data extraction and transformation, and representations for huge volumes of multidimensional unstructured data. Numerous studies have been conducted on IE, addressing the challenges and issues for different data types such as text, image, audio and video.

Jana Sedlakova et al (2022) Digital data play an increasingly important role in advancing medical research and care. However, most digital data in healthcare are in an unstructured and often not readily accessible format for research. Specifically, unstructured data are available in a non-standardized format and require substantial preprocessing and feature extraction to translate them to meaningful insights. This might hinder their potential to advance health research, prevention, and patient care delivery, as these processes are resource intensive and connected with unresolved challenges. These challenges might prevent enrichment of structured evidence bases with relevant unstructured data, which we refer to as digital unstructured data enrichment.

Types of data

Structured data refers to data that has definite format and length, easy to store and analyze with high degree of organization. This means that the data is organized in identifiable structure to allow it response to queries to retrieve information for organizational use. A typical example of structured data is relational database like structured query language (SQL) or Access, which contained organized numbers, dates, group of words and numbers called strings/text. Due to the database seamless structure, it is searchable with simple, straightforward search algorithms which might be by data type within the actual content. Traditional analytics focus had been on structured data in while neglecting larger amount of other types.

Semi-structured data is irregular data that may be incomplete and have a structure that changes rapidly or unpredictably but does not conform to a fixed or explicit schema. This means that it is not table oriented as in a relational database model or sorted graph as in object databases. The semi-structured data model allows information from several sources, with related but different properties, to be fit together in one whole, for example, email, XML, Doc files.

On the contrary, unstructured data has no particular structure. Unstructured data typically includes bitmap images/objects, text, email and other data types that are not part of a database. Although emails are organized in a database format like in Lotus Notes and Microsoft Exchange, the body of the message is in text format without structure in any way. In other words, unstructured data comprises documents like PowerPoint used to describe company strategy, spreadsheets of lead list, emails between coworkers, and interactions of customers on social networks. Word processing documents are another form of unstructured data, though with some formatting, the content is freeform text without any structure.

The Internet is a huge collection of data that is highly unstructured which makes it extremely difficult to search and retrieve valuable information. Due to the massive number of unstructured data, search engines that search and rank documents that contain unstructured data based on their relevance to user queries become essential for information seeking. Search engines are required to determine relevant documents within a short latency. In other words, high search efficiency is one of the key design and implementation objectives of search engines.

Unstructured data is a generic label for describing any corporate information that is not in a database. Unstructured data can be textual or non-textual. Textual unstructured data is generated in media like email messages, PowerPoint presentations, word documents, collaboration software and instant messages. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and flash video files. Unstructured information is typically text-heavy but may contain data such as dates, numbers, facts and multimedia data as well.

Nowadays, efficient indexing and searching system for unstructured data is a challenge. In this paper, we propose efficient content-based unstructured data retrieval system. The proposed system incorporates full-text search approach for text data and content-based approach for audio, video and image data. Multimedia search engines can search not only text data but also image, audio and video data that is unstructured data. They use content-based multimedia retrieval approach that combines content-based text retrieval, content-based image retrieval, content-based audio retrieval and content-based video retrieval approaches for efficient indexing and searching.

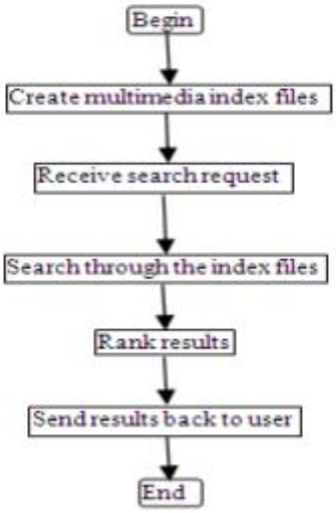
Begin
Create multimedia index files
Receive search request
Search through the index files
Rank results
Send results back to user
End

Figure 1: Functional model of multimedia search engine

Below figure describes the proposed unstructured data indexing and retrieval framework.

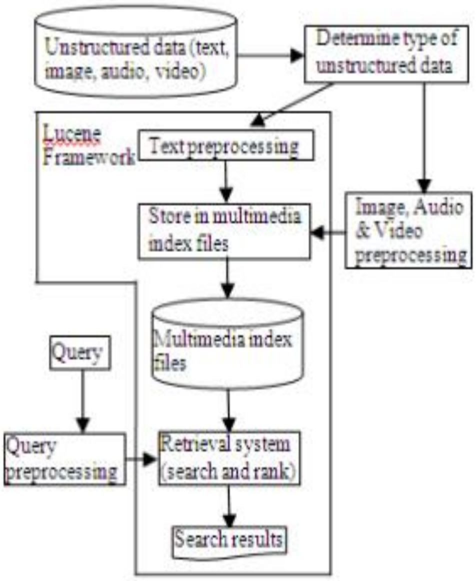
Unstructured data (text, image, audio, video)
Determine type of unstructured data
Lucene Framework
Text preprocessing
Store in multimedia index files
Image, Audio & Video preprocessing
Multimedia index files
Query
Query preprocessing
Retrieval system (search and rank)
Search results

Figure 2: Unstructured data indexing and retrieval framework

First, the system determines the type of unstructured data. If the data is text, the text preprocessing operations are performed by a text preprocessing module.

Conclusion

Gathering and analyzing unstructured data gives organizations insight into their businesses and help them to increase competitive edge, enhance productivity, and create innovations. Information derived from this analysis will help organizations to re-strategize in order to increase their market share. Data from social network site like Facebook, Instagram, and LinkedIn would have been meaningless without unstructured data. Executives can get relevant information for decision making in less time with unstructured data analysis. Unstructured data has created room for fraudulent analysis, loyalty programmes that identifies and targets the consumers and customer segmentation based on stored behavior analysis.

References

1. *Adnan, K., Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data 6, 91 (2019). https://doi.org/10.1186/s40537-019-0254-8.*
2. *Jana Sedlakova et al (2022) Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review, University of Zurich, 8006 Zurich, Switzerland.*
3. *Das, T., & Kumar, P. (2013). BIG Data Analytics: A Framework for Unstructured Data Analysis. International Journal of Engineering and Technology (IJET) , 5 (1).*
4. *Adanma Cecilia Eberendu (2016) Unstructured Data: an overview of the data of Big Data, International Journal of Computer Trends and Technology, ISSN: 2231-2803, Volume 38 Number 1.*
5. *Y.B. Ma, Z.Y. Fang, J. Liu and T.Y. Wang,"A content-based multimedia retrieval system base on MPEG-7 Metadata Schema," Proc IEEE, 1200-1201, (2008).*
6. *Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.*
7. *Lomotey RK, Deters R. Topics and terms mining in unstructured data stores. In: 2013 IEEE 16th international conference on computational science and engineering, 2013. p. 854–61.*
8. *Tapas Ranjan B., and Subhendu Kumar Pani., (2016) "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset."Procedia Computer Science 85: 862-870.*