Reducing data volume in big data: Parallel processing based data filtering techniques

openaccessgovernment.org/article/reducing-data-volume-in-big-data-parallel-processing-based-data-filtering-techniques/172657

22 January 2024

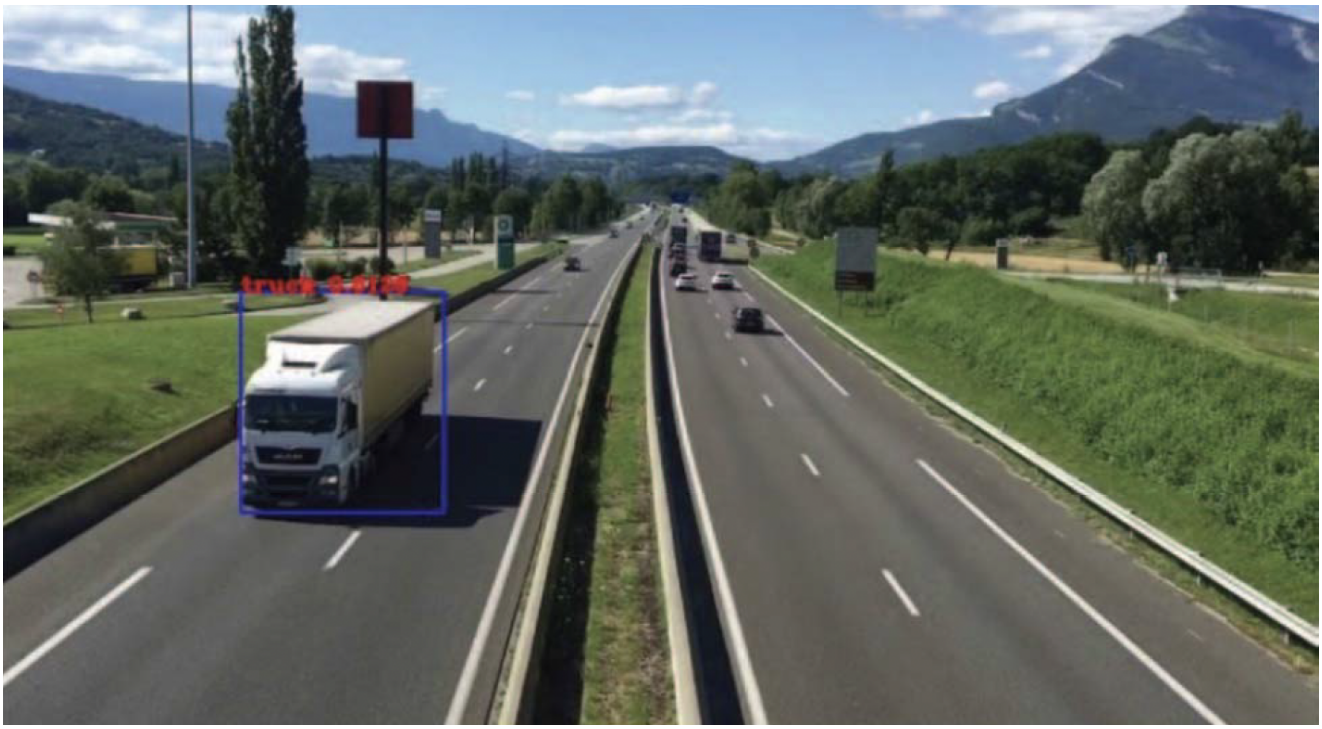


Figure 2: Example of Filtered Video Frame [from (2)]

Professor Shikharesh Majumdar from Carleton University examines reducing data volume in big data, focusing on parallel processing based data filtering techniques

Volume, velocity, and variety are three well-known characteristics of big data. The large data volumes often introduce formidable challenges to processing such data in a timely and economical manner. Research on data filtering is underway under the leadership of Shikharesh Majumdar at the Real Time and Distributed Systems Research Centre at Carleton University. Users are often interested only in a subset of the raw data.

Three different use cases

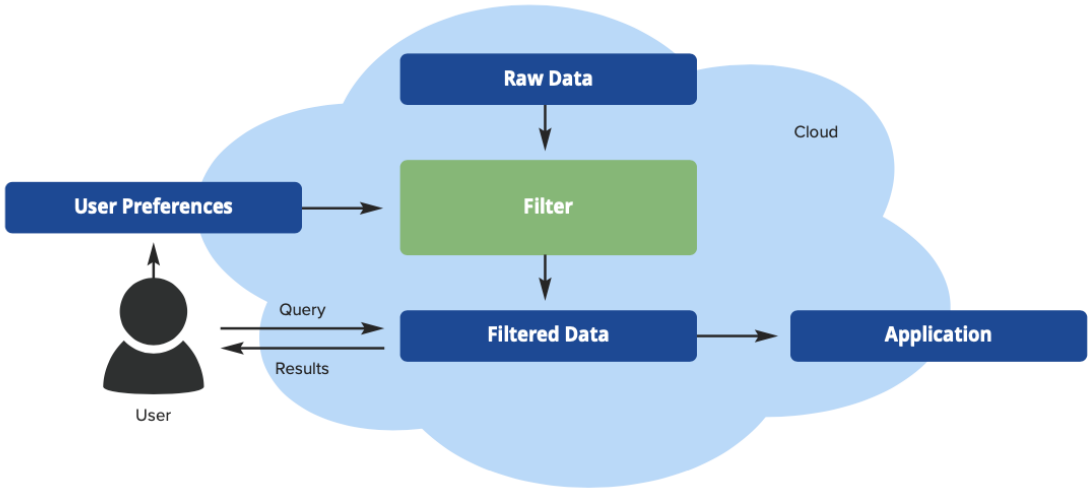
This research focuses on filtering out only this data of interest for later processing by the user, thus reducing the storage space required and the concomitant reduction in data search times. Three different use cases are considered: filtering of textual data, video data, and multimodal data, each of which is briefly discussed here.

Data filtering turns out to be computationally intensive, and parallel processing-based techniques are used in each case to complete the filtering operation in a timely manner.

Filtering of Textual Data: Users are often interested in a small number of topics of interest. For example, a specific employee may be interested in only certain issues discussed in the various company meetings. They can provide their interest in terms of “user preferences” or keywords, and the filtering algorithm can then go through all the meeting minutes and store only the contents (e.g., sentences, paragraphs) that contain these keywords in the specific user’s repository (see Figure 1). Such filtering can lead to orders of magnitude reduction in the volume of stored data. (1)

Another example is filtering the large volume of data produced by tweets to only those that contain the keywords specified by a journalist preparing a story on a specific topic.

Filtering of Video Data: Large volumes of video data are often a source of problems for data analysis. As in the case of textual data, filtering video data based on user preferences can lead to substantial savings in storage and reduced search times for the various operations performed by user applications. Users specify the class of preferred objects they are interested in, and the filtering method captures sets of consecutive frames, each containing the object of interest.



What applications need such filtered data?

Examples of applications that need such filtered data include detecting specific types of vehicles such as a truck (see Figure 2) from traffic surveillance videos or capturing frames containing closely following cars on a highway. Parallel processing-based filtering techniques that use machine learning for object detection are described in (2).

Potential other applications include tracking and counting specific objects of interest on a conveyor belt or vehicles on city roads and highways, for example.

Filtering of Multimodal Data: Documents are often characterized by multiple data types: textual, voice, and video. Consider movies, for example. A movie comprises visible scenes displayed by video data, whereas audio data and/or subtitle texts often capture the dialogues.

A user may be interested in storing only certain sections (moments) of the movie and specifying the type of objects depicted in such scenes of interest accompanied by dialogues included as audio and/or subtitle texts. Such “moment” retrievals and filtering from movies are described in [3].

Video recordings of meetings and conferences are also good candidates for moment filtering. Filtering of video and textual data is often computationally expensive. Parallel processing techniques were used in their implementation. Proof-of-concept of prototypes built using Apache Spark and deployed on an Amazon EC2 cloud demonstrated the efficacy of the techniques. [1, 2, 3]

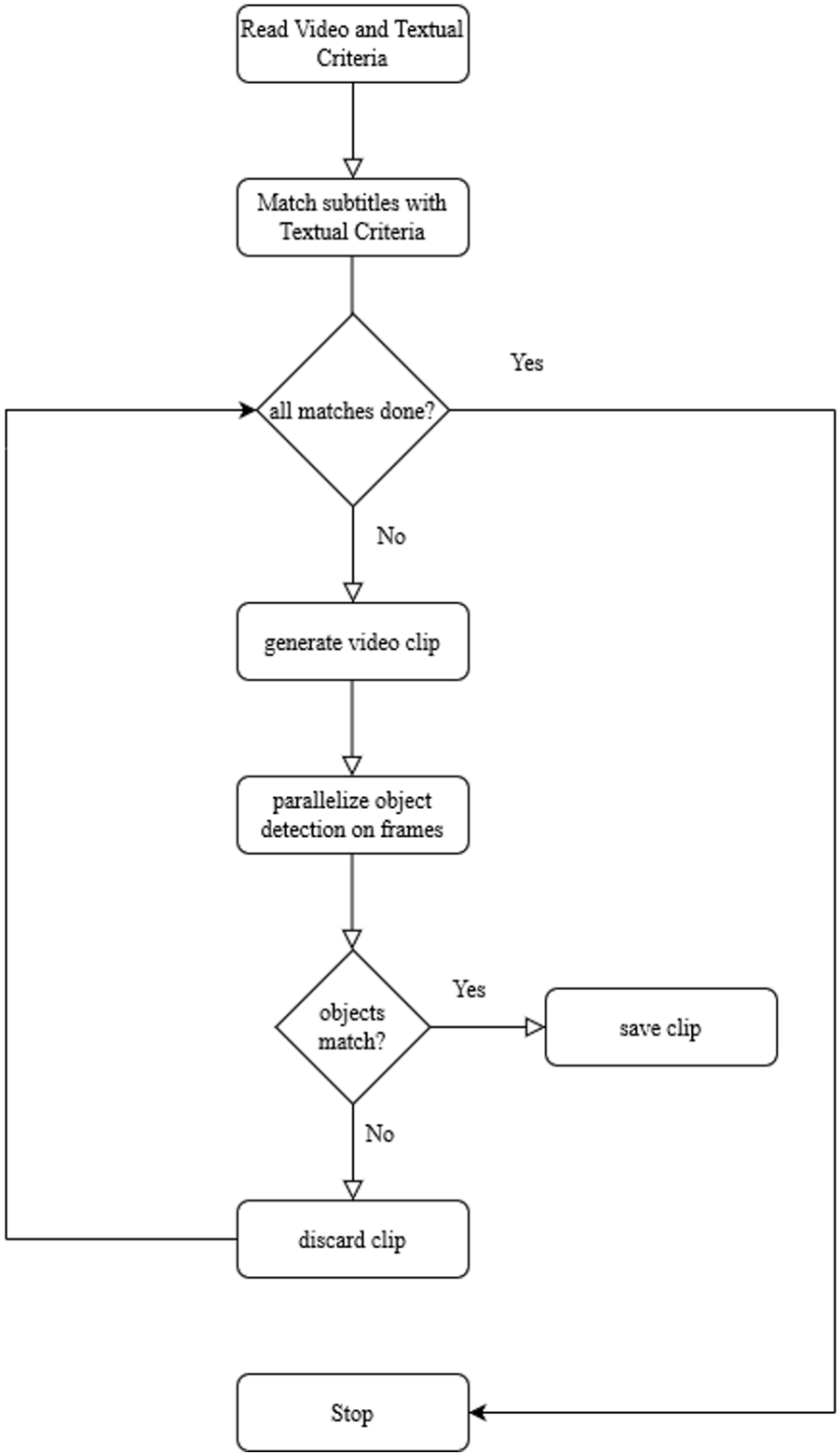


Figure 3: Steps of the Video Filtering Approach [from [3]]

References

1. Chanda, B., Majumdar, S. “A Parallel Processing Technique for Extracting and Storing User Specified Data,” Proc. IEEE 8th International Conference on Future Internet of Things and Cloud (FiCloud), Rome, August 2021.

2. Syed, A.T. and Majumdar, S., “Parallel Processing Techniques for Analyzing Large Video Files: a Deep Learning Based Approach,” Proc. 2022 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Melbourne, Australia, December 2022.
3. Syed, A.T, Majumdar, S., “Techniques for Moment Retrieval and Filtering from Large Volumes of Multimodal Data”, Proc. International Conference on Future Internet of Things and Cloud (FiCloud), Marrakech, Morocco, August 2023.

Please Note: This is a Commercial Profile