Data extraction and efficient storing of PDF in a database: A Review

AMPDS Aludeniya[1], RGC Upeksha[1], TL Weerawardane[2]

Department of Computer Science[1,], Department of Computer Engineering[2], Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

Abstract. With the broad usage of information technology in many different sectors at present, the digital world is expanding rapidly. With that, a huge amount of data is communicated around the world continuously. In there, unstructured data types are widely used than structured data types. Unstructured data doesn't have a data model or schema that has already been defined. It's the opposite of structured data, which is commonly found in relational database management systems (RDBMS). When consider about managing unstructured data in a Relational Database System, Unstructured data cannot be driven to conform to a traditional relational database's columns and rows format. Recent researchers have given their attention to finding the best method to store and manage unstructured data in a database effectively. Throughout the process I am going to do a comparative analysis based on well-defined parameters. This will be based on some existing methods such as PDF data inside relational database environment, PDF data outside relational environment. As the first step those methods will be explained separately by stating benefits as well as side effects of each. When consider about PDF data inside Relational Database, Databases are used to store information for easy lookup and better data management. The usual types of data stored are texts and numbers. Data types such as VAR or VARCHAR will let you store characters or text, while INT and FLOAT will let you store numbers. One data type called a BLOB (binary large object) will enable you to store binary files such as a DOC file, executable files and PDF files. By creating an upload form connected to your database, you can successfully store PDF files in it. The second method is an alternative to store PDF files. In that method, PDF files are storing outside the database and a file path or URL Will be used as data in the database. This work aims to review recently improved methods for storing PDF files in a database concerning the performance of the database and tools for extracting data in PDF files.

Keywords: *database, unstructured data, PDF files, performance, PDF extraction tools*