Abstract ID : **18**

Why does document structuring still matter and struggle in the era of GenAI and Large Language Models?

Content

Document structuring is a fundamental aspect of information management, involving the categorization and organization of documents into logical and physical structures. This presentation explores the benefits and challenges associated with document structuring, focusing on the distinctions between physical and logical structures, metadata and content, as well as addressing the implications for businesses and research centers dealing with large volumes of data encompassed in *data warehouses* and *lakes* of textual documents.

In the task of document structuring, distinctions arise between physical and logical structures. Physical structures pertain to the layout and presentation of documents, encompassing elements such as tables, figures, and images. On the other hand, logical structures refer to the organization of content within documents, including metadata that describes document attributes and content that comprises the textual information.

Implementing structured document management systems brings several benefits for business and research bodies. Firstly, structured documents target search queries more effectively, yielding more relevant search results and reducing the volume of irrelevant hits. This not only enhances search efficiency but also saves time and resources, resulting in cost savings and eco-friendly practices. Additionally, structured documents facilitate comparisons between similar structures, enabling deeper analysis and insights. Moreover, the adoption of structured documents enables the extraction of statistics and the creation of dashboards, as it allows for the identification and analysis of document elements beyond mere text.

However, document structuring still faces great challenges. Legacy documents pose a significant hurdle, particularly those with poor scans or generated through low-quality optical character recognition (*OCR*). These documents may contain noise, artifacts, or degradation, compromising the accuracy of structure recognition algorithms. Furthermore, complex layouts, heterogeneous documents, handwritten content, tables, figures, images, multilingual text, and dynamic content all contribute to the complexity of document structuring. Moreover, the scarcity of labeled data exacerbates the challenge, hindering the development of accurate and robust structuring algorithms.

While *Generative Artificial Intelligence* (*GenAI*) has demonstrated remarkable capabilities in various domains, including natural language understanding and image recognition, it still struggles with document structuring due to the complexity of document layouts, ambiguity in content, and limited contextual understanding. It faces challenges in handling diverse document formats, noisy data, and legacy documents. Additionally, *GenAI*‘s reliance on labeled data for training limits its generalization ability, hindering its performance on unseen document structures. Overcoming these challenges requires interdisciplinary collaboration and continued research to develop more robust *Artificial Intelligence* (*AI*) models capable of effectively managing the complexities of document organization and content extraction.

In conclusion, document structuring offers substantial benefits for businesses and research centers, enabling more efficient information retrieval, automated data extraction, enhanced searchability, standardization, and improved data analysis. However, overcoming these challenges requires innovative solutions and advancements in document structuring technology. By addressing these challenges, organizations can harness the full potential of structured documents to optimize workflows, facilitate knowledge management, and drive innovation.

Topic

Data innovations: Data Management/Integration/Exchange

Primary author: Dr KHEMAKHEM, Mohamed (MandaNetwork)

Presenter: Dr KHEMAKHEM, Mohamed (MandaNetwork)

Contribution Type: Long Talk

Submitted by **Dr KHEMAKHEM, Mohamed** on **Monday, 22 April 2024**