International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Structured Document Digitalization System

Omar Isa, Subah Ahmed Al Sowaidi, Osama Pervaiz Alam

[1] *Department of Information Systems, University of Bahrain, Sakhir, Kingdom of Bahrain*

E-mail address: emg.omar.exe@gmail.com, subah1997@gmail.com, osamanote341@gmail.com

Received 11 Nov. 2019, Revised 11 April. 2020, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: While observing current work environments where documents are not handled automatically there is a higher chance of errors occurring in documents and the placement on the work materials. Companies are held back, being unable to quickly and efficiently transfer to an online well-structured system. An ultimate solution to implement a system contains several aspects. Combining these technologies together to build a structured digitalized system using artificial intelligence, cloud computing, text recognition, database and web development acquire a lot of solutions in this field. Database is created through a smooth and easy transition without much efforts, along with aspects that add to document filtering methods and advance search methods using artificial intelligence. With the purpose of improvement and change in the workplace environment through the goal of digitalizing existing physical archives to replace the manual work. A website is provided in the system for client to access and transmitting documents. This will help create a productivity boost and a much more efficient and organized document management system and it will decrease the chances of misplacement of documents and shall maintain order. Moreover, a survey was made to question the percipients on if there are similar systems and if they ever had such smart system to acquire the job automatically. Results were measured to give an inspection on the current position in Bahrain market.

Keywords: cloud computing system, artifcial intellgence, documanting, digtlizing documantes, smart archiving system , title detection

1. INTRODUCTION

To begin with consistency nowadays is essential for every company in many terms such as work efficiency, availability, redundancy, and security. Once the idea of digitalization is introduced it change the market and help many businesses in many aspects. **Digitalization** is the use of digital technologies to change a business model and provide new revenue and value-producing opportunities it is the process of moving to a digital business. Digitalization is being driven by new technologies. The arrival of cloud computing, artificial intelligence enhanced the process of digitalization as cloud computing is the most important digital technology being used nowadays along with artificial intelligence in terms of storage and processing respectively. Digitalizing favors the company by doing tasks faster and at lower expenses than rivals, this would put the company at a competitive advantage. In terms of security it is a huge profit to digitalize the work as authentication helps protect valuable information. Enabling a user to create an account and log in, is equipping them with security for and control over their information. A user can see the information they've given you in their user profile, they can understand what to use to connect with them. Security also helps the enterprise customers who want to have everything within their system secured, which means providing a sign up that works with enterprise login requirements. Availability of the system all times is crucial in accessing data from anywhere and anytime without disruption of service. Digitalization also has a trait which is maintaining and saving time, this is done threw automated services connected with each other and are triggered by a certain condition. This benefits the employees in terms of efficiency as well and creates new job opportunity. The IT staff will be involved in the business itself as transactions are automated. Thus, the days where the IT staff were completely isolated from the rest of the business is over. Digitalizing the business will require a set of highly skilled IT members. The customer communication with the employees is enhanced with the use of websites that allow customers to get answers to their questions immediately. The use software technologies to communicate with a business benefits because, better communication produce s a stronger public image.

A. Problem Statement

Many companies and government institutions in Bahrain are still running on paper-based systems. These account for the majority of archiving for all sorts of important documents. A lot of the workplace tasks and sifting is carried out within this environment of paper and hardcopies. In the age of digitalization, it is essential to have a digital archive solution. This solution will aid establishments in gaining a centralized database, thereby giving remote access, security, and advance search and filtering methods by combining different technology aspects such artificial intelligence and cloud computing. The goals is to build and design a system which brings a smooth transition from paper-based archives to digital archives, as well as provide the intended professionals tools to enhance productivity, practicality and remote access. This will entail workplace associates quicker access to a document management system, along with the relevant features integrated within the database.

2. LITERATURE REVIEW

The availability of these services has a leverage to many companies nowadays with the launch of new services from companies such as Amazon and Apple, they are now expecting every organization to deliver products and services swiftly, with a seamless user experience [1].The enterprise customers would want a quick and seamless digital experience, and they want it now this would ease the flow of access for the users as it provides Customers want to sign in to their electricity account onli ne and see their usage chart in real time. Customers expe ct their telecommunications company to activate the mobile once it's been purchased and have it enabled and installed immediately [1]. For instance, one of the advantages of digitalization is efficiency and effectiveness [1]. Also, it aids the management of a major bulk of data [2]. Digitalization further exemplifies commercial endeavors. It adds flexibility and effectiveness to the commercialization to a company [3]. With these pros in mind, the effective approach is through digitization. In terms of security majority of modern-day businesses are endangered to security risks and vandalism. The use of technology to protect financial data, executive confidential decisions an d other proprietary information leading to competitive ad vantages. Basically, technology helps to protect the businesses from their competitions for their new innovative ideas. By having computers with passwords, a company can ensure that the competition does not copy any of its upcoming projects [5]. There would be a future growth will be determined by the Internet of Things, cloud computing, AI, 3D printing, virtual/augmented reality and even by block chain, even if at a low level. The percentage of companies that use artificial intelligence has, by contrast, only grown from 6 to 7 percent. (Bitkom Research, Trendstudie "Unterwegs zu digitalen Welten" (2018), vgl. auch die Deloitte-Studie: „Zukunft der Consumer Technology" (2018))[4].

3. SYSTEM DESIGN AND COMPONENTS

The system is made up of many components which are the cloud service, artificial intelligence and website integrated with database. The cloud service is an online storage space for all sort of files. Cloud storage is a computer data processing system that stores digital data in logical pools. The use of text recognition service that is provided by Amazon web services for the extraction of text from files. Artificial intelligence sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. An online database the data is accessed online there is no need to download the online database builder. You just need to create an account to access the online database builder. You can now create your own online database builder without any trouble once you have an account. Online server creator does not need a technician because the use of the internet makes it easy to access it.

A. Cloud Computing

AWS (Amazon Web Services) is Amazon's expansive, emerging cloud computing system that provides a combination of service infrastructure (IaaS), application platform (PaaS) and service package (SaaS) offerings. AWS systems may provide tools like computing power, server processing, and content delivery services to an enterprise. Amazon S3 or Amazon Simple cloud storage service is a service that Amazon Web Services (AWS) provides to store items via a web service interface. [6] Amazon S3 uses the same flexible storage system that Amazon.com uses to operate its international e-commerce network. Amazon S3 can be used to store any form of object which enables uses such as network storage, backup and recovery, disaster recovery, information repositories, analytics data lakes, and hybrid cloud storage. [6] Amazon S3 has high reliability which promises 99.9 percent uptime in its service-level agreement, which results in less than 43 minutes of downtime per month. [6] the S3 cloud provide us a great storing environment for the system. The storing of the documents on the cloud will provide the company with remote access to the documents. The cloud storage will reduce the amount of archiving hardcopy documents inside Bahrain companies which will lead to reducing in wasting company infrastructure to store the hardcopies. Moreover, the response time for acquiring a document will be faster, as the hard copy requires a lot of effort and searching method is done in manual way. Also, the availability of accessing remotely is reliable and the company can access the document without being physically in the company. Cloud storage provided an efficient solution in terms of archiving to make handling the paperwork easily which increased the scalability.

B. Text Recognition (OCR)

Optical character reader (OCR) is an electronic conversion of typed, handwritten or printed text images into machine-encoded text, whether from a scanned document, a document photograph, a scene photograph such as a text on signs and billboards in a landscape photograph or a subtitle text superimposed on an image such as a television broadcast. In the developed system Amazon Textract was used. Amazon Textract is a system that extracts text and information from documents that are scanned automatically. [6] Amazon Textract goes beyond simple optical character recognition (OCR) to recognize field content in tables stored forms and data. Amazon Textract makes it easy to extract data from documents, forms and tables quickly and accurately. Amazon Textract recognizes the structure of a file and the key elements on the page automatically, understands the information relationships in any embedded types or tables, and extracts everything in its context. [6]

C. Artificial Intelligence (AI)

AI is a subset of machine learning which is the scientific study of algorithms and mathematical models used by computer systems to perform a particular task without using explicit instructions, relying instead on patterns and inferences. [7] It is known to be a branch of artificial intelligence. Machine learning algorithms build a sample data-based mathematical model, known as "training data," to make predictions. Machine learning algorithms are used in a wide range of applications, such as email sorting and computer vision, where designing a traditional algorithm is impossible or unfeasible to perform the task effectively. TensorFlow is an open source software library developed by google for machine learning. It is an essential part of the system in terms of coding where the library was used in JavaScript code for HTML. A model was trained inside TensorFlow using a data set of images that can be differentiated based on the requirements. by feeding the AI model we were able to do classification of the documents in the system. A major of 4 categories were created in the system which are charts, financial statements, CV's and lastly codes. The results came clearly inside the database to be viewed by the system admin with the percentage of document in template wise and text wise. The artificial intelligence aspects in the system is to use different APIs like TensorFlow and Aylien. Then, the machine learning makes our model learn about the documents we give upload in order to create the categories and sort the documents based on the results.

D. Online Database

The online database is accessed via the website with user identification. A database is an organized data collection, typically obtained from a computer system and accessed electronically. Since databases are more complex, they are often built using techniques of systematic design and modelling. Access to these data is usually provided by a "database management system" (DBMS) consisting of an integrated collection of computer software that allows users to communicate with one or more databases and provides access to all the data stored in the database. [7]

E. Laravel Web Developing Tool

Laravel is a PHP software framework development tool and it's an open source software which is used to design web applications based on the architectural pattern of the model view controller and based on Symfony (php web development framework). Some of the features of Laravel like Modular packaging mechanism with a specialized dependent manager, with various ways to access relational databases, utilities that help deploy and maintain applications and their orientation towards syntactic sugar(It's a Syntax in programming language that makes the code easier to learn and understand). Eloquent ORM (object-relational mapping) is an innovative PHP implementation of the active record model, simultaneously providing internal methods to impose constraints on the relationships between database objects. Eloquent ORM introduces the database tables as classes that are utilizing the active record model, with their object instances connected to individual table rows. To represent the work a website was created which allows the user to surf the documents easily. Furthermore, it allows acquiring and uploading with access privileges. The website is integrated to combine all the features of the system under one roof.

F. The flow chart of the system

The following steps describes the flow of the system from the input of the document going through all the processing steps where the arrangement of document and the conversion of the document if needed is done. Firstly, the user uploads the documents either from the website or the mobile into the computer. Secondly, we check the format before the document gets uploaded the code checks if it is in PDF format if so, the document gets converted into a collection of images depending on the number of pages of the document. Thirdly, When the documents are verified, they get uploaded to the online database where they are stored after text recognition (OCR). Fourthly, The AI will run the image and compare with its dataset to classify the image into an appropriate category like financial, formal/informal letters, and contracts for projects. Fifthly, Keywords are generated to categorize the images where each image can be searched using a specific

keyword based on its category. Lastly, the results of the text recognition and classification are then sent to the database.

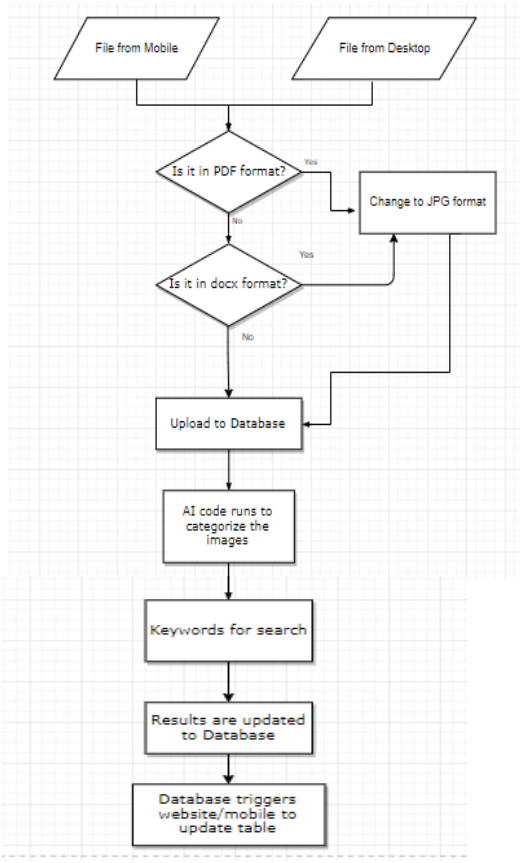

Figure 1: the flow chart of the system

G. The user flow chart

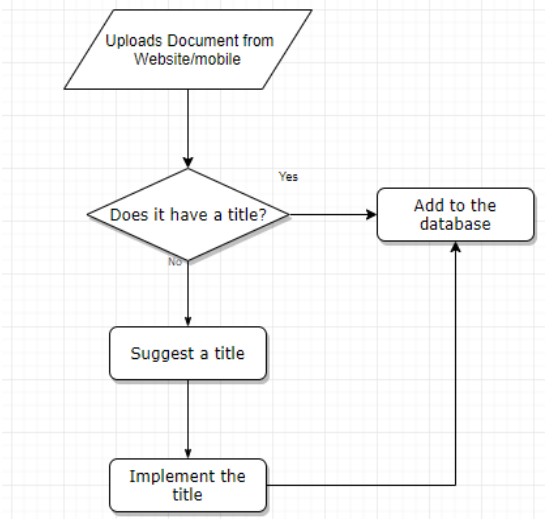

Figure 2: user flow chart

The following steps describes the flow of the system from the user side from the upload of the document going through all the processing steps where the title generation and the addition of the document to the database. The user starts by uploading the documents into the computer. After that the AI check for a title and suggests a title to user in case of there is no title based on the topic of the document and lastly, upload it to the database.

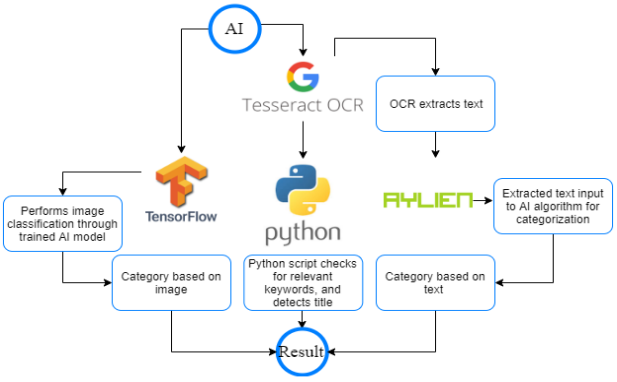

Figure 3: in depth processing

In this section the in-depth processing of the documents is done from scratch by the artificial intelligence first which uses Google's TensorFlow libraries for image classification, the image is processed and categorized. The next step is the text recognition which is performed by Google's tesseract OCR and then the text is categorized by Aylien and the keywords of the document are determined by the python code with title detection.

4. TESTING AND RESULTS

In this chapter we will discuss the results of the system and testing the main components for results. In an office environment digitalization changed the way employees work with the system significantly by providing multiple solutions to different problems that faced the employees before implementing the system. While digitization and innovation make new business ideas for organizations, they additionally change the structure of the association and its basic leadership process. Frequently, employees and their skills need to be improved in terms of technology. Digitization requires mental fortitude. Progress originates from improving and adjusting existing items to even more likely address the issues of our clients. A significant essential to digitization is the joint effort between the business department and Information Technology department. The system has positive and negative effects on the employees, but the users should fully make use of the advantages of the system. Offices with many documents and papers lying around would benefit significantly from the system, but on the other hand there are many aspects that should be considered when transforming the system.

A. Similar System

In 1960s the University of Cornell developed a similar system SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System with similar machine learning functions that classifies the text retrieval from documents and classifies them using the relevance feedback. In our system the artificial intelligence software also classifies the documents and the text within it into categories which are relevant to the user. The AI software presents the percentage of relevance of the document like financial statements, letters, charts, research papers. In terms of output form of classification of documents between our system and the similar SMART system, our system displays percentages and the SMART system gives pseudo codes that represents the relevance of the text.

B. Strengths of the System

The system has many features that makes it suitable for companies with many documents and less space in the office, the system keeps the employees in order and updated in terms of work as tasks can be stored in online cloud storage and can be accessed remotely without the need of a manager to be present to inform the employees about a specific tasks as instructional guidelines are uploaded for the employees. In addition to storage space and databases the system also includes an artificial intelligence software that is specialized in arranging and keeping track of the type of documents used by the user as it classifies the documents based on their category for example financial statements, letters, charts, research papers etc. Moreover, this trait is exact and significantly more exact in contrast with a human. The artificial intelligence software has a very low error rate and gives an accurate and precise classification of documents. The system also gives substantially more quicker computations and can-do multiple tasks which are practically difficult to be finished by a single employee manually. In terms of text extraction and title generation the software performs these tasks in under seconds with no errors in the output which saves time for the user. The frequent use of the system helps the employee in learning new topics in the technology field which also enhances the skills and knowledge. The system also makes the IT staff department of a company to be more involved in the business and creates more job opportunity for individuals.

C. Weaknesses of the System

With many benefits in hand a system would always have some negative aspects that should be considered. The system uses artificial intelligence software that needs to be updated frequently and it requires some sort of maintenance by certified and qualified programmers in that field which would cost the company for weekly scheduled maintenance. The artificial intelligence software needs to be fed new dataset for the purpose of identifying new types of documents. In addition to maintenance the security of the system should be considered as threats nowadays are increasing and the implementation of anti-viruses and anti-malware software are necessary. In the current system it is running on a single server which is a disadvantage as if the system fails this would shut down the whole system and data would be lost. In addition to this point the backups cannot be made if the system fails once in a sudden.

5. CONCLUSION

To conclude the system as a whole has many useful functionalities which ease the use of it with a friendly user interface that has proper guidelines the user can follow and use the system without the need of training or practicing. Cloud storage made the access of large stored files easy and portable that can be used anywhere and anytime. The text recognition software made the text extraction procedure fast and simple with an accurate and precise extraction. The artificial intelligence model made quick comparisons and gave precise results about the classification of documents for the user. Although the system had been very accurate and friendly there are still some limitations of the system. The cloud storage solutions depend on the speed of the internet connection for fast upload and download speed, but low latency can prevent the user from accessing the data in real time. There are many locations around the globe where internet connection cannot be reached. The costs Cloud storage platforms are viable options for an enterprise or a small business. These costs can, however, be too large for home devices to accommodate. Privacy issues are a bit of a problem concerning who owns the information after migrating the data to a third-party cloud storage provider. Is it the client, as a customer, or the provider of cloud storage? The other limited factor which is the artificial intelligence and its main limitation is that it learns from the data given which is the dataset. In contrast to human learning, there is no other way that knowledge can be integrated. This means that the results will reflect any inaccuracies in the data. The AI also requires supervision as it needs to be fed new dataset so that it continues to learn and stay updated for new tasks. The cost and maintenance of the AI software should be considered in order to adapt to the changing business environment, AI software will also require regular upgrades. before you go ahead and incorporate some AI programmer, the company needs to carefully consider the return on investment. The text recognition software also has some limitations that should be considered as well like many languages have special characters, and these characters can be lost or recognized incorrectly if the correct OCR software is not loaded properly. The handwritten text also cannot be recognized only with special configuration in the software which is also a limitation.

REFERENCES

[1]. McKinsey & Company. (2019). *Accelerating the digitization of business processes*. [online] Available at: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/accelerating-the-digitization-of-business-processes [Accessed 12 Sep. 2019].

[2]. Bradley, S. (2019). Models and Methods of University Technology Transfer.

[3] Doi.org. (2019). *Redirecting*. [online] Available at: https://doi.org/10.1016/j.tibtech.2010.12.001 [Accessed 12 Sep. 2019].

[4]. Digitalization for Companies in Germany , Conni ,[online] Available at : https://infopark.com/en/blog/digitalization-for-companies [2nd December 2019]

[5]. *Importance of technologies, (29th July 2019,)* Neil Kokemuller [online] *available at https://bizfluent.com/about-6320228-technology-important-business-.html* [Accessed 3rd December 2019].

[6]. *Amazon web services* [online] Available at: https://docs.aws.amazon.com

[7]. Beynon-Davies, Paul (2003). *Database Systems* (3rd ed.). Palgrave Macmillan.

[8]. Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval.

Omar Isa is fresh Graduate student from university of Bahrain holding BSc degree in computer engineering. Currently he is taking internship in the deanship of graduate studies and scientific research at university of Bahrain as a Manager and editor of the journals hosted by the university. Research areas include cryptography, cybersecurity, and artificial intelligence.

Subah Ahmed Al Sowaidi, Computer Engineer graduate from University of Bahrain. Earned Bsc in Computer Engineering major in 2020. Recently published an article about machine learning and cloud computing as part of senior project. His research interests include artificial intelligent, machine learning, decision support systems, big data system and information system.

Osama Pervaiz Alam is a graduate of the Computer Engineering Bachelor's program at the University of Bahrain. He has worked in multiple fields of study, some of which include artificial intelligence, web development, hardware programming, automation and networking.