Next-generation Serverless System for Contextual Search Based on Rich Media Content

Vladyslav Holubiev[1], Bohdan Ihnatiuk[2], Iryna Voytyuk[3]

Faculty of Computer Information Technologies, Ternopil National Economic University, UKRAINE, Ternopil, 8 Chekhova str., email: vladyslav.holubiev@gmail.com[1], bohdan.ihnatiuk@eleks.com[2], i.voytyuk@tneu.edu.ua[3]

Abstract: **The purpose of this research is building a next-generation knowledge management system based on the most recent developments in the industry of serverless computing. Such approach reimagines a conventional way people build software by abstracting away the complexity of maintaining highly available, highly scalable systems. The software combines many ways to parse rich media and text content such as image, audio, video to build a contextual search index.**

Keywords: **serverless system, contextual search, rich media content.**

I. Introduction

One of the problems which pop up on a day to day job of office workers - is facing a huge amount of information. Even more adds up to the problem when these files are scattered across different cloud systems, storage providers and there is no single place to store all. The next problem is that these cloud-based systems have a very inconsistent user interface and the average office employee should spend hours and hours learning a new UI, which drastically reduces performance and prolongs onboarding time. After some quick calculations, it turns out this process takes up to 500 hours, which causes new problems and diseases that emerged only in the 21st century - information overload and fatigue from decision making. As a result, an office worker cannot make critical decisions and loses productivity.

The next problem of the existing cloud storage providers - they have limited search interface. No single solution on a market can do powerful analysis and extract all the information from files and parse the content. Each of them focuses on one aspect of the user's field of knowledge, while today the organization can operate with an extremely diverse range of file formats. Even the largest search engines are focused on a thorough analysis of web pages while closing eyes on office workers in enterprise companies.

Talking about enterprises with thousands of employees, the amount of wasted money measured in millions. So it is critically important to stick to the discipline and the common idea of knowledge sharing in the organization in order not to be distracted by the search for information. Following some simple practices as avoiding people known as knowledge holders, making knowledge easy to find and share will bring you a long way ahead.

In addition, it is necessary to take into account achievements of the 21st century in the field of machine learning. This allows you to use new areas for analyzing files for a self-regulating organization. With a large number of files, human assistance is no longer effective and error-prone.

After combining all of the above-mentioned with the latest technologies of parallel computing, the system was born, which solves the problem of overloading machine resources. New developments in cloud tech and so-called "serverless computing" make it possible to use resources extremely cost-effectively, paying only for the time spent by your code per second of time and gigabyte of memory [1]. A new unit of measurement was born - GB/s.

Per-second payment cycle saves the organization's budget since computer hardware does not work at nights and other hours that are not considered to be working in the office world. Instead, the company rents a CPU power of another company which owns loads of efficient hardware. Thanks to global coverage, the tenant company provides 24-hour service in all time zones, thereby minimizing costs by reducing latency. Even by putting 100 data centers around the world in the largest cities, gives a coverage which ensures <50 ms latency regardless of your geolocation.

After analyzing the problem, assessing available solutions and researching innovative technology - a new knowledge management system was born which steps up the game.

II. Technologies Behind Content Categorization and Search Index

The process of synchronization described in this paper is aimed at reducing the negative effects outlined above. The main point is assuming file synchronization happens in terms of some folders structure. Given a normal distribution of files per folder, we can apply parallelism at the hierarchy level. With a folder structure, you can divide it into identical shards in order to squeeze the most performance of available computing resources.

In this case, several components that put a risk on files are fixed at once. Firstly, there is no queue now. Due to an efficient allocation of a tree structure of a folder hierarchy, it became possible to split all files into the so-called "branches" and process them separately and independently. Because each branch has the same small size, it takes only a few seconds to import, so the chance that there will be some kind of error at the moment is diminishingly small. This solves the problem of importing through a large number of files. The second improvement is that instead of using one server, there are now a large number of servers that perform the same job. At first glance, it may seem that it will be less effective since organizations will have to pay for all these servers and maintain them in proper condition, which multiplies the resources spent on one server. Instead, the developed software system is based on the technology of serverless computing, which can be used to leverage computing power from other companies. This is a new form of business that has gained insane popularity.

Over the last 2017, serverless computing has gained a significant growth. Thus, organizations of different sizes around the world appreciate benefits of hassle-free computing power without wasting their resources to support many own servers, which stand idle at night. Just for granted, at your will, you get essentially "unlimited" number of parallel computations, which do not exceed five minutes each. Let’s consider a way of processing text information.

So from the previous step, the system received input files in different custom formats. To search for these files you need to convert them to text format. Typically, several modern file processing techniques are used. Let’s consider in detail the technique of serverless computing. At the first step, let's collect a sample of all the different formats that can be encountered when an import is finished - a representative sample of typical files that the office worker uses most often. Next, consider the methods of processing each of these file types. Often, text files carry a lot of unimportant information to look for. Therefore, the system will process them for cleaning out irrelevant characters. This process is called "tokenization". It also includes the process of bringing words to its original form. After this stage of processing, we have text in the output that is suitable for transferring to the search index. The same principles then apply to all other types of files after they pass the step of extracting information.

After analyzing file types that are used by office workers, the most common types of photo materials were found. In 70% of cases, these are scanned documents. The rest of the files are general-purpose photographs, where no text is depicted, but objects are present (more on this later).

Let's choose two tactics of image processing (Fig. 1).

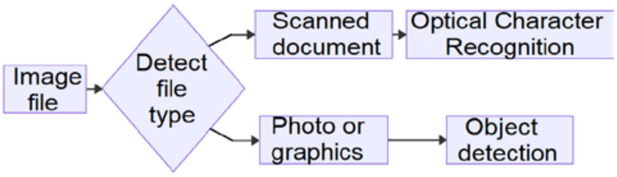


Fig. 1. Two tactics of image processing

The first tactic involves the use of machine learning algorithms that were trained on millions of characters with different projections and distortions. The software for this task is open source and is not copyrighted. The knowledge management software system has a built-in neural network model for recognition documents in three most popular languages such as English, French, and German [2].

For example, let’s take a photo of a Ukrainian bachelor's diploma. As shown in Fig. 2, a trained model recognized the text in the image with close to 100% accuracy. An office worker can find this document by quoting a word or a phrase from an image. Such rich media integration greatly expands capabilities of conventional search engines.

In case the model was unable to extract enough text from the image file after the first step, it means that we are dealing with the type of photo file without text. In such situation, another model of the neural network, which is trained on millions of images, will be used. Usually, such networks are created by training them by hand using a supervised training method. Unlike to unsupervised training, a human being required to teach the model range of objects. During this process, the model makes links between elementary image properties and the vector of available objects for recognition.

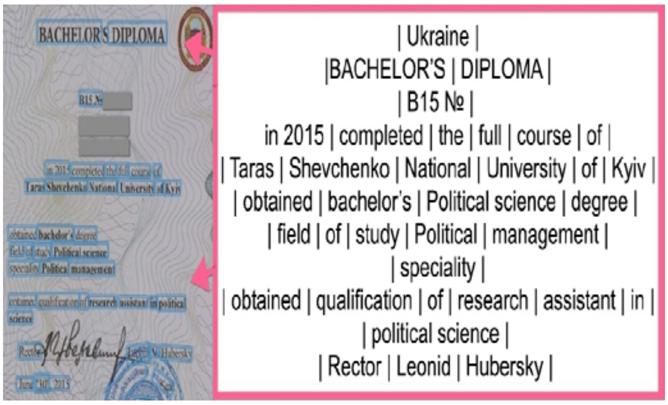
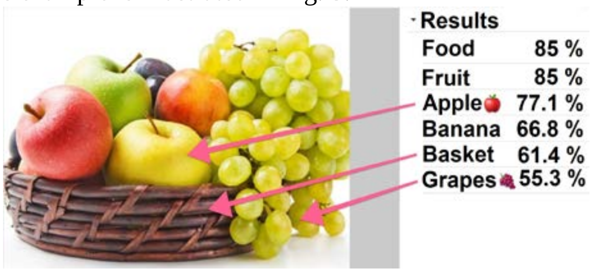

| Ukraine |
|BACHELOR'S | DIPLOMA |
| B15 № |
in 2015 | completed | the | full | course | of |
Taras	Shevchenko	National	University	of	Kyiv
obtained	bachelor's	Political science	degree
field	of	study	Political	management
speciality
obtained	qualification	of	research	assistant	in
political science
Rector	Leonid	Hubersky

Fig. 2. Example of Optical Character Recognition

However, it's worthwhile to leverage readily available developments of Google company, such as TensorFlow (open source license). This software processes an input image and outputs a scalar vector of key phrases depicted in the photo. This example is illustrated in Fig. 3.


Results

Food	85 %
Fruit	85 %
Apple	77.1 %
Banana	66.8 %
Basket	61.4 %
Grapes	55.3 %

Fig. 3. Example of recognition the scene and objects in the photo

Given an image of basket with fruit, TensorFlow model outputs a list of objects.

Given an image of a basket, Tensor Flow model outputs a list of fruit, such as apple, grapes, and the basket itself. Imagine a company file storage, filled with thousands of photos which are searchable only by file name. This kind of rich media experience opens new ways for content findability. Furthermore, detectable objects are limited to a list of known vocabulary words, so it makes easy applying search faceting to the list of objects. In a couple of clicks, an employee can drill down thousands of search results in orders of magnitude.

It is worth mentioning the fact that each recognized item is followed by a percentage of confidence, which will then allow you to weed out variants that are not relevant.

A video of the 136th episode of TNEU "Kaleidoscope of Events" was taken to demonstrate possibilities of video analysis for object detection in video media. The architecture of the video recognition process is shown in Fig. 4. The following recognition categories are available: people, faces, logos, celebrities.

The theory behind video processing is pretty much the same as images, but the scale of processing power required grows exponentially [3]. The system needs plenty of CPU resources to analyze each frame with the same grade of detail as each photo. That's where serverless computing shines in its glory. For a case when we use our own trained model for detection we just rent out a couple of seconds of runtime in cloud providers and run the analysis there. As another option, cloud providers do offer a dedicated service for video recognition. From a viewpoint of knowledge management system, such setup is considered serverless as well, as it does not provision any servers for video analysis at all.

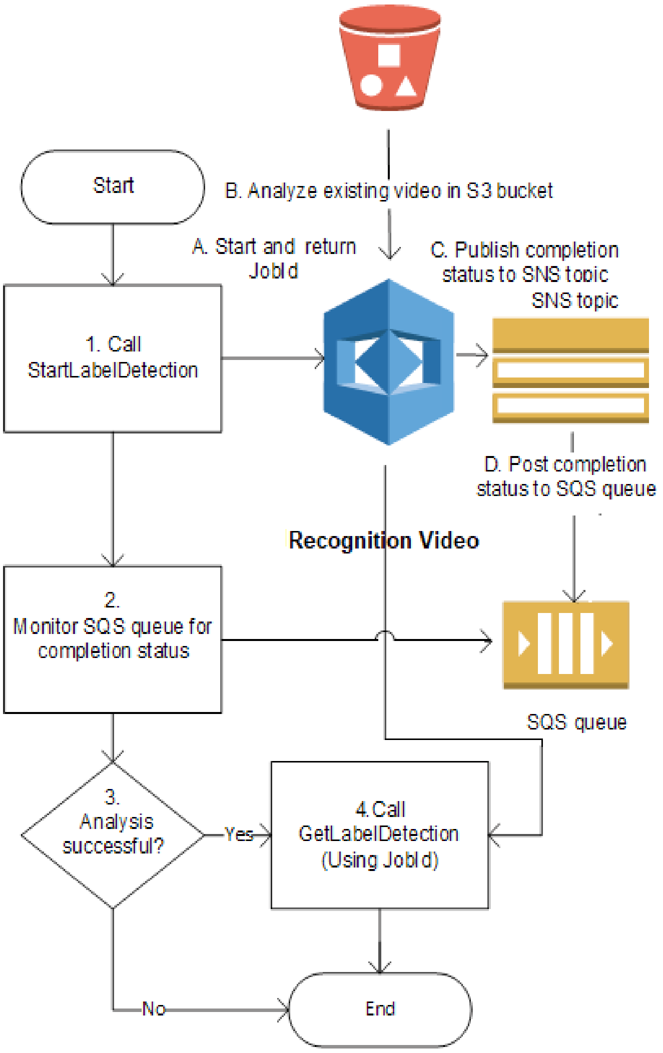
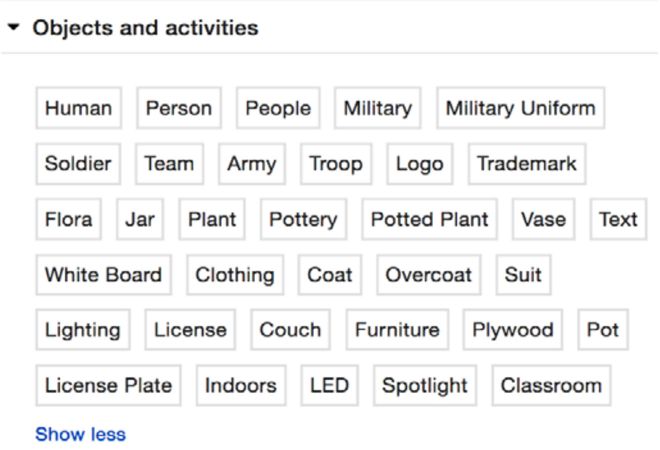

Start
B. Analyze existing video in S3 bucket
A. Start and return JobId
C. Publish completion status to SNS topic
SNS topic
1. Call StartLabelDetection
D. Post completion status to SQS queue
Recognition Video
2. Monitor SQS queue for completion status
SQS queue
3. Analysis successful?
Yes
4.Call GetLabelDetection (Using JobId)
No
End

Fig. 4. The architecture of the video file analysis process

Video processing is a very complicated process and requires enormous computing resources. Therefore, in this case, the software system solves the problem by using the AWS Rekognition software solution for analyzing video files. It's enough to send the link to the downloadable file from the cloud storage and check the status of the recognition process. As a result of analysis of the latest episode of the TNEU "Kaleidoscope of Events" it discovered plenty of entities which appeared in a frame, as shown in Fig. 5.

After manual checking the outcome of analysis on my own and watching the video, it is safe to come to the conclusion that the confidence score of the keywords is acceptable and it's good to go in the knowledge management system. Similar to images, this opens a whole new set of ways for employees to find company's content scattered across many videos. This means less time for people to put text descriptions to videos just to make them searchable.



▾ Objects and activities

Human · Person · People · Military · Military Uniform
Soldier · Team · Army · Troop · Logo · Trademark
Flora · Jar · Plant · Pottery · Potted Plant · Vase · Text
White Board · Clothing · Coat · Overcoat · Suit
Lighting · License · Couch · Furniture · Plywood · Pot
License Plate · Indoors · LED · Spotlight · Classroom

Show less

Fig. 5. Entities that are recognized in the latest video episode TNEU "Kaleidoscope of Events"

The next category of files that are usually synced from cloud storage systems is audio files. They can be divided into two categories: text from a voice recorder and phone conversations. In order to cover both categories, the software system uses the AWS Transcribe software, which provides an easy-to-use interface for interacting with audio files. At the entrance to the API, the file is sent in formats - mp3, ogg or wav, and the output will be text and timing.

An important advantage of using this software interface is the support of several speakers. An office worker will have the opportunity to select the text of the person whose record he wants to find. This functionality is provided by "uncontrolled" (unsupervised) machine learning. By entering several different voice records of different people, the algorithm of the neural network forms its guesses by breaking the audio tracks on the common features, thereby dividing the phrases that are similar to sound and assigning them to the people speaking them. Worth to not, that a few years ago, such technology was available only with a supporting video by separating audio using lips reading.

For other types of files that could not be divided into the media categories mentioned above, the Apache Tika software will be used to analyze even binary content (in total more than 1500 different MIME types of files) and extract media and information from them [4]. Examples of such information include dates, file creation, and people that created these files.

Apache Tika provides three ways to use: graphical user interface, command line software interface, and HTTP web server with a REST API.

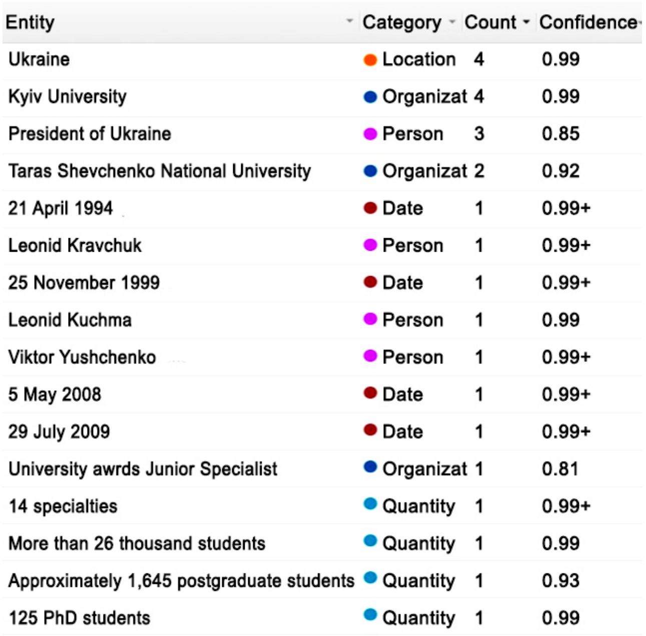
The principle of building serverless systems involves freezing OS processes. But even here it's not an issue, as you can run Apache Tika server that receives the first request and then stops the process. The process state is saved and the new file process is restored, thereby eliminating the need to keep the running server running on the server permanently up to date. Even conventional bulky software from the 90s can be adapted to run in serverless environments to cut down the cost to the bare minimum.

Data transfer happens within the system in a private availability zone, no user information leaves the closed network. Servers which run computations, do not have a network interface and even have no assigned IP address for access from the outside world. Latest OS patches are installed as soon as possible by cloud providers and require zero human intervention. As an example of security measures taken, let's remember a recent CPU vulnerability CVE-2017-5754 (Meltdown attach) which stressed out the whole world. Serverless environments were upgraded to a patched OS version within a more matter of hours after announcement without owners worrying patching a system manually.

In a knowledge management system, open source software plays a big role to identify keywords and key phrases. A group of programs for recognition and analyzing human speech comes from a range of products from the Stanford University, such as OpenNLP. It stands for Open Natural Language Processing. This framework provides an easy way to recognize keywords and entities from a text. Results are used in a search index to build filters and categories. For example, given an archive of library documents and analyzed contents, it is possible to share group thousands of materials into faceted search filters by the author, place of the event, time, country, and so on. A huge win for office workers when they search for a specific place or event in enormous blobs of text. And the second suggested use case of these rich text entities is building categories. Without training a model a system can group content related to a single date in history in one place. So it becomes easier to organize previously scattered content.

Consider an example of text analysis. The encyclopedia Wikipedia contains a fragment of the text from the article about the Taras Shevchenko National University of Kyiv for demonstration purposes.

The results of fragment recognition are shown in Fig. 6.



Entity	Category	Count	Confidence
Ukraine	● Location	4	0.99
Kyiv University	● Organizat	4	0.99
President of Ukraine	● Person	3	0.85
Taras Shevchenko National University	● Organizat	2	0.92
21 April 1994	● Date	1	0.99+
Leonid Kravchuk	● Person	1	0.99+
25 November 1999	● Date	1	0.99+
Leonid Kuchma	● Person	1	0.99
Viktor Yushchenko	● Person	1	0.99+
5 May 2008	● Date	1	0.99+
29 July 2009	● Date	1	0.99+
University awrds Junior Specialist	● Organizat	1	0.81
14 specialties	● Quantity	1	0.99+
More than 26 thousand students	● Quantity	1	0.99
Approximately 1,645 postgraduate students	● Quantity	1	0.93
125 PhD students	● Quantity	1	0.99

Fig. 6. Key entities recognition

In a matter of minutes, gigabytes of information is transferred into a search index, where a user can filter out any document by several key criteria or by a person mentioned in the document.

III. Implementation of the Knowledge Management System

This knowledge management software is a web-based portal with a graphical user interface to provide an easy access to the search and the ability to drill down numerous search results through faceted filters, categories and suggestions.

A software implementation is based on JavaScript programming language and the Node.js software platform [5]. Many additional libraries, frameworks, and servers were included to parse and analyze files, such as Apache Tika, OpenNLP, Elasticsearch, Tesseract, GraphicsMagick, AWS Rekognition Video, and AWS Transcribe. Each file has a visual display for a quick and intuitive impression of what's inside a photo or a document.

IV. Conclusion

The software system for synchronization of files from cloud storage and their analysis for the organization and contextual search on the basis of serverless computations is proposed and implemented. The developed software system is designed to handle the flow of files from cloud storage retrieved as a result of moving the directory structure tree hierarchy. The software system provides the ability to quickly search for tens of thousands of files. This solves a common problem among office workers who spend up to 500 hours a year searching for files. Such a large amount of information leads to information overload and fatigue from decision making.

The user feedbacks were analyzed on a sample of volunteers who used the system for 30 days. More than 90% of the results had positive feedback, so it is considered that the system has found its application and is meant to be useful.

The last but not the least, one important aspect of the developed system is its cost. Principles of maximum simplicity of software architecture design and low cost of maintenance were set in stone from the very beginning of development. As a result, the latest achievements in the field of computing were utilized, especially the principle of "serverless" computation. This approach hides a huge complexity from the knowledge system owner in terms of maintenance of own servers and data. Even database storage and search index implemented in serverless way ensure data persistence, availability, eventual consistency and idempotency by utilizing sharding and replication.

References

[1] M. Stigler. *Beginning Serverless Computing*. Richmond, Virginia, USA: Apress, 2018, pp. 190-199.

[2] A. Chaudhuri, Kr. Mandaviya, P. Badelia, S. K. Ghosh. *Optical Character Recognition Systems for Different Languages with Soft Computing*, 1st ed., Springer, 2016, pp. 2010-248.

[3] C. Gurturk. *Building Serverless Architectures*. Istanbul: Packt Publishing, 2017, pp. 131-219.

[4] Ch. Mattmann, Ju. Zitting. *Tika in Action*, 1st ed., Manning Publications, 2011, pp. 200-256.

[5] A. R. Young, M. Harter. *Node.js in Practice*, 1st ed., Manning Publications, 2014, pp. 400-424.