Implementation and Comparison of Text-Based Image Retrieval Schemes

Syed Ali Jafar Zaidi*, Attaullah Buriro*, Mohammad Riaz*, Athar Mahboob*, Mohammad Noman Riaz`

Department of Information Security*
Khwaja Fareed University of Engineering & IT, Rahim Yar Khan, Pakistan*
Department of Computer Science`
Virtual University of Pakistan Lahore, Pakistan`

Abstract—Search engines, i.e., Google, Yahoo pro-vide various libraries and API's to assist programmers and researchers in easier and efficient access to their collected data. When a user generates a search query, the dedicated Application Programming Interface (API) returns the JavaScript Object Notation (JSON) file which contains the desired data. Scraping techniques help image descriptors to separate the image's URL and web host's URL in different documents for easier implementation of different algorithms. The aim of this paper is to propose a novel approach to effectively filter out the desired image(s) from the retrieved data. More specifically, this work primarily focuses on applying simple yet efficient techniques to achieve accurate image retrieval. We compare two algorithms, i.e., Cosine similarity and Sequence Matcher, to obtain the accuracy with a minimum of irrelevance. Obtained results prove Cosine similarity more accurate than its counterpart in finding the maximum relevant image(s).

Keywords—Image retrieval; image filtering; cosine similarity; sequence matching

I. Introduction

Well before the advent of the internet, it was extremely difficult to remain connected with the world, no one had access to the connected world as we have in this modern computer age. Although there were railway tracks, ships, and other transportation means, however, those were slow in countryside and expensive. Most of the meetings were conducted with the neighbors and the only fastest means of communication in those days were telephone and telegraph. After 1960, new innovations in transportation and telecommunications and with the fast progress in the world cars, bullet trains, planes and phones allowed people to contact each other over the very large distances [1].

In past ages, people used to send information through letters and fax machines and the only source of information was the newspaper but with the fast progress in telecommunication field the letter and telegraph is mostly overridden by the emails and different social networking services. e.g. Facebook, Twitter and others, the source of getting information is shifted to different websites. Over the past few decades, the fast-growing use of the internet is the subject of various studies as it provides access to a large set of information. The Internet is the connections is the subject of various studies as it provides access to a large set of information The Internet is the connections of two or more computers which are connected to each other and communicate and share their resources and information with each other all over the globe [1]. Social networking service is a podium for the people to build social relations and to share their multimedia information with others who have a similar field of activities, interests, and backgrounds. The massive increase in the number of users of different social media services, people share their interest and communicate with each other through them.

Millions of people, with a continuous daily increase (as shown in Figure 1) are using social networking sites, e.g., WhatsApp & Facebook messengers, to share a lot of multimedia information, i.e., image, text, smiley etc., every second. To this end, a lot of abundant information is present on these sites. This information could be useful for the respective users, however, for irrelevant users, it is useless. Internet users may use different search engines, e.g., Google, Bing, Yahoo, etc. to acquire their required document, weblink, image or a video.

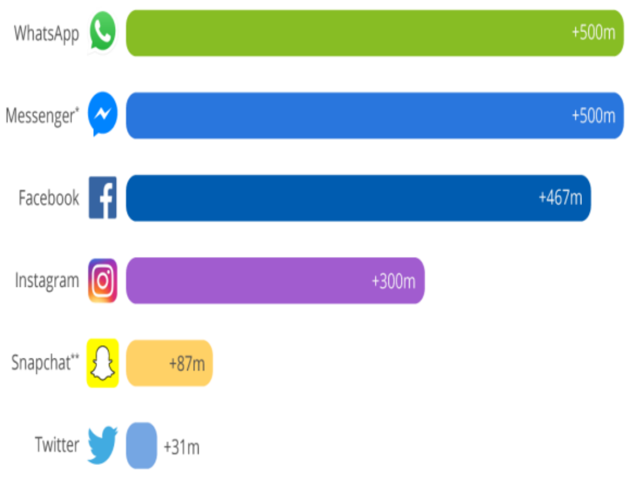

Fig. 1 Users of Different Social Media Platforms.

The technology is rapidly advancing, for example, previously smartphone users use to use Edge technology which was much slower than the present 3G and 4G technologies, which were meant to provide the desired information in the quickest possible way. In some years, 5G is also going to be deployed worldwide. Due to the differences in bandwidth [2], [3], these advanced technologies, can connect users quickly to their beloved social sites and hence to their desired multimedia data. The country-wise estimated usage of social networking websites is shown in Figure 2 which estimates that the use of these sites by the public is likely to increase during the period of five years (2017 -2022).

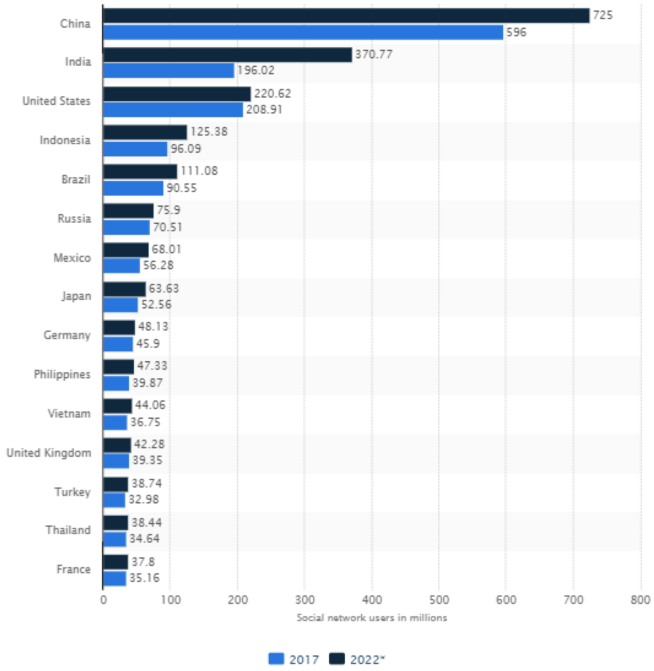

Fig. 2 Expected Increment of Internet usage in Different Coun-Tries of the World Between 2017-2022.

As shown in Figure 2, the data transaction is rapidly increasing day by day and during the next five years, the users of the internet will be increased by several million as of today. In multimedia communication, it is evident that the transaction of irrelevant data will also be increased which ultimately will overwhelm the computing devices as well the users. Hence, the filtering of relevant data in the context of the user is deemed essential. Keeping in view the latest techniques and Inspired by the new technologies and techniques of research and major issues of the knowledge domain, this paper highlights the addressed research problem and describes the research motivations and the major research objective. Additionally, in this work, we address the problem of fetching the most relevant image data using the text-based query. More technically, we address the problem of text-based image filtration problem in this work.

II. Related Works

Presently, the two methods exist in order to solve the problem of information overloading, namely: Information Retrieval Technology (IRT) and Information Filtering Technology (IFT). These information retrieval techniques were introduced by two search engine giants; Google and Yahoo, to facilitate the users in finding and fetching their desired data and information in accordance with their requirements. While searching for the desired data or information, the user has to provide a query in a detailed, precise and accurate form. If the presented query is either not accurate or precise, then this would ultimately generate unwanted results. Modern information retrieval techniques are being widely used in finding the most accurate and precise results in the minimum possible time [1]. The image retrieval community is presently focusing on two main information retrieval algorithms; namely, Collaborative Information Retrieval System (CFA) and Time Weight Algorithm (TWM). The CFA algorithm is considered as one of the most efficient and effective algorithms being used at present [19]. The CFA algorithm is based on the interests of the collaborative users which combines the results of users' interests and as a result, provides analysis at a certain point. Also, the CFA algorithm has the ability to filter out the undesired and complex impressions and ensures their subsequent settlement in real time. As far as the functionality of the TWN algorithm is concerned, it focuses on the interests of the user(s) that pertain to the finding and fetching of the desired data or information in real time. Besides ensuring this, the TWM algorithm also keeps the long-term interest of information filtering and helps in getting rid of overdue interests and, hence, saves considerable filtering time. Undoubtedly, the TWA algorithm favored over CFA or any other filtering technique or algorithm because it is considered more explicit and adaptable to user needs[16]. The most important aim of Information Retrieval (IR) model is to find out relevant knowledge-based information" or a document that fulfills the user needs. The square measure essential procedures associate IR help in demonstrating the representation of the archived information, the appreciation of the clients' information needs, and, hence, the examination of these two depictions. The client of the information recovery framework is not engaged with this procedure, and Ordering Method leads to a representation of the record [17]. Due to an explosive growth of online data repositories, individuals have gone astray within the web's info-thickets, infrequently waste abundant amount of time and money in finding out the desired and personalized data. Coping with such a pull, researchers from totally different areas have invented numerous tools. However, compared with recommended systems which automatically match the users' style supported by the historical behaviors according to the interest, these computer programmed tools are not personalized enough, and due to this inadequacy, they produce redundant results for all the users. Among several recommended systems, Collaborative Filtering (CF), is the most generally utilized in different fields, and as a result of its consecration of requiring no domain data, police investigation, Collaborative Filtering (CF), has attracted a lot of interest from each tutorial and trade field group during the last decade. Generally, there are two main kinds of CF: neighborhood and model-based approach. The essential plan of CF is that the recommendation in respect of the target user is created by predicting the preferences of uncollected items that support the neighbors. The neighbor may be a cluster of persons with similar interests. In particular, the competition of Netflix Prize (NP) has provoked totally different fields for researchers and computer scientists to propose numerous solutions to build corresponding recommended systems [18].

During the formative days of information retrieval techniques, a user- based and item-based approaches were widely applied in the domain of information retrieval, like Amazon, Flicker, Yahoo, Instagram, Google and many more were the main users. In recent years, specialists belong to both academia and trade have witnessed a terrific performance of model-based approaches, particularly the Latent Factor Model (LFM) [18]. Since the typical representative technique, Latent Factor Model (LFM), encompasses most of the methods, a Matrix factorization (MF) provides another methodology to represent the association between users and things. In LFM, users and things each depicted within the same Latent Factor Space (LFS) and, as a result, the prediction is accomplished by directly evaluating the preferences of users for uncollected things. Some Medium Frequency ways are planned in CF as a result of the high potency in handling large-scale datasets. Those approaches tend to suit the user-item rating matrix with low-rank matrix factorization and apply it to form rating predictions. Medium Frequency is economical in coaching since it assumes only a couple of factors that influence preferences in user-item ratings [18]. Due to the success of Matrix Factorization (MF) within the Netflix Prize competition, the several excellent variants area units have been projected. An MF framework with social network regularization was delineating. It provides a general methodology for increasing recommended system by incorporating social network data. These two models not only exploited the cooperative effects in the knowledge domain but also conjointly took into consideration the order, due to which things may well be viewed by the users. In addition, besides the normal Collaborative frequency (CF) ways within the recommended system, they conjointly emerged several variant ways based on applied mathematics, physics with the advent of network science. Most of these ways are supporting the divided networks. A number of these ways ensure component innovation and proof themselves effective in raising not only accuracy but also diversity and novelty conjointly. The projected recommendation rule supported the Associate Intending Integrated Diffusion on user-item-tag three-party graphs and considerably improved accuracy, diversification, and novelty of recommendations [18].

III. Problem Formulation & Solution

In this section, we define the problem and our approach we apply for solving it.

A. Problem Statement

This study focuses on the problem of accurate, effective, and relevant image(s) retrieval which ensure accurate, precise, and quick results to facilitate the user.

B. Approach

In this study, we retrieve images from the internet using Bing API by applying a text-based search query. Bing API is a very useful tool to fetch images from the server. This API also returns the results in the forms of JSON file, we further process the collected results to fetch the text of the images. The retrieved images in most of the cases were not the desired ones, thus, we need to create a system to retrieve the most accurate images.

IV. Information Retrieval and Information Filtering

Internet users use search engines to search for their required information on the internet. These so-called search engines use different web crawling algorithms to manage and maintain the information in real time. When a user searches for something in the search engine, the engine tries to answer all the underlying matches of the query, but due to the presence of a large amount of data on different blogs, web links, and on the social media, the retrieved data might not be necessarily relevant.

In this technological age, the term information gathering refers to the “Information Retrieval (IR)”. This process starts with the user’s query for any search or retrieval. More specifically, an IR process can identify numerous objects and ranks them on the basis of their similarity (however their degree of relevancy may vary) as a result of a particular search query. Needless the say, this gathered information might not necessarily contain the users required content.

A. Image Retrieval

Image retrieval process involves the searching, browsing and retrieving the image(s) from the digital image databases. Image retrieval has been the most attractive and interesting task the users would do on the internet. It has been active both in the research and commercial domains since 1990. Different IR systems have been designed and implemented for research and commercial purposes at schools, digital libraries, hospitals, and biodiversity information systems. An IR system could be used to search images by the text, examples and/or any other search methods. Currently, the two frameworks, i.e., Text-based and Content-based, are being used to retrieve the images [1], [4]. We explain below these two frameworks:

1) Text Base Image Retrieval (TBIR): Text-Based image retrieval system refers to the retrieval of images, through text as an input,e.g., keyword, etc. This text-based search may not be much use because of the chances of getting irrelevant results due to human errors, such as misspelling, unexpressed feeling, emotions, etc. As such this technique is considered as an old-fashioned technique and is not widely used anymore [5]. Each image has some text with respect to its name, caption or detail or web portal on social media as a description which is used to retrieve those image(s) [6]. The users’ search of an image is decomposed, parallelly, in the form of attributes in the metadata of the search engine and finds out the appropriate matching of the input query. Then, it finds all the images according to the attributes similarity and displays the results to the user.

2) Content Based Image Retrieval (CBIR): Content-Based Image Retrieval (CBIR) is one of the most active research topics over the past few decades. This is the system which refers to the retrieval of images on the bases of their visual context, such as color, text, shape, figure and image segmentation [7]. The static stable resemblance or remoteness roles are not often able to handle the CBIR, due to the difficulty of visual image depiction and the semantic breach challenge among the low-level visual abilities and high-level human awareness [8]. When the user gives the sample image to the IR system, the system converts that image into the feature vectors. Then the CBIR system extracts the features, e.g., colors, text, shape, etc., of the query and all the fetched similar images [9], [10]. Later the similarity is computed by comparing the extracted features of the query image to the features of all the similar images found in the dataset. The system is depicted in Figure 3.

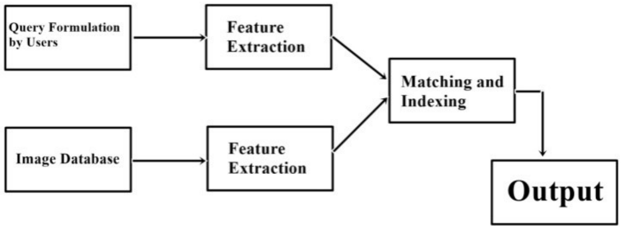

Fig. 3 Representation of Content base Image Retrieval.

B. Image Filtration

Image filtration has been seen as the method of distribution of the relevant images from search engines. Different search engines have been used to retrieve the unfiltered images, as per the users' search, and make them more accurate and suitable as per the requirements of the user. Image filtration schemes use a lot of filters to find out the appropriateness of the found results, however, due to an overabundance of the data, sometimes the irrelevant images are also retrieved [7]. Image retrieval is performed after the image filtration process.

V. Data Extraction Techniques

The demand for API is increasing rapidly as the world is getting aware of web and smartphone applications. Some web servers, i.e., Google, Bing, Yahoo, etc., are providing Open API services to the developers. We use the Python programming language as Server-end language to collect the data. Bing API, when called, returns the results closely related to the user interest but in limited numbers, meaning that Bing Image Search returns only 35 images related to the user each query. To get access to the Bing API, it is necessary to get register with the Bing. It then provides an OCP-APIM-Subscription-Key, which is unique for every user [11]. After calling this API, Bing will return all the information about the images like the name of the image, the format of the image and the web-page from where the image is retrieved and the URL of the images which is further used in the research to display the image in a web-browser.

A. Relevance Feedback

In relevance feedback, the feedback from the user is recorded to check the relevancy of the retrieved image/information. The idea of relevance feedback is introduced to improve the final results of retrieval systems. It takes the initially returned images and asks users feedback about their relevancy to the query [14].

VI. Implementation & Discussion

When a user wants to search anything using the search engine, the user is returned with the results with the relevancy to the query. Relevancy of the query can be determined by different methods. Only Term Frequency (OTF)[1] is not enough to find out the most relevant text of the picture because there are many stop words and other words that can decrease the relevancy of the most relevant text. The weight of the term using tf/idf model can be determined by the following equation:

$$wi = tfi * \log \texttt{[U+2061]} \binom{D}{dfi} \tag{1}$$

Where tf_i is the number of occurrence of the term, in i documents, whereas the term D is the total numbers of documents and dfi is the term used for the total number of the documents which contain the term i [15]. Besides this technique, as it seems not appropriate or aligned enough, the aim of our study is to find out the relevancy of a text, we use the following techniques which could be more helpful to retrieve effective data.These techniques are:

A. Cosine Similarity Technique.

B. Sequence Matcher Technique.

We will first explain, in the following sections, the working of our chosen techniques and compare their performance, afterward. In this way, we would be able to find out a better algorithm which could fetch out the more relevant images confirmed with the obtained higher accuracy.

A. Cosine Similarity Techniques

Cosine Similarity - a comparison technique based on the inner product space of two non-zero vectors measures the cosine of the angle between them. The range of cosine similarity is between -1 and 1 with 0 representing the string orthogonality (de-correlation), and intermediate values representing intermediate similarity or dissimilarity. For the text matching, the attribute vectors, A and B, are usually the term frequency vectors of the documents. In this study, as it deals with attributes of the images line by line, so we worked with attributes "not" with the documents and taking these attributes as documents. We can see the Cosine similarity as a way of normalizing the length of documentation. When we retrieve information,

[1] Term Frequency is the ratio of the occurrence of each word token to the occurrence of the all words in the document

the Cosine similarity between two documents will always be in between 0 to 1, since the term frequencies (*tf*$_i$-*df* weights) cannot be negative. The term frequencies of vectors will always be not less than 90°. In the data mining, this technique is used to find out the cohesion between the two attributes. Similarly, in this study, this technique is used to measure cohesion between the retrieved images and the queries given by the user and the attributes fetched [11]. One of the reasons to use cosine similarity technique is its effectiveness and easier implementation.

Euclidean dot product formula is used to calculate the cosine of two non-zero vectors[2].

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta \tag{2}$$

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{3}$$

where A_i and B_i are the components of of vector A and B respectively.

The range of the results will be from -1 to 1 in this perspective. If the value of the result is exactly 1, it means that the string is exactly the same and if the value is -1 then it means that the string is exactly the opposite [11]. The conclusion is that if the results are closer to 1 the string will be more closely related to them and if the value is closer to 0 means the de-correlation between the strings.

1) Implementation of Cosine Similarity: As discussed earlier, the Cosine similarity technique finds out the similarity of two non-zero vectors. As this study focuses on the text, we need to convert the text into vector first. The example of converting the text into the vector and computing the similarity between two non-zero vectors is illustrated using an example below:

For example, we need to compute the similarity of the two sentences given below:

1. Pakistan is my homeland and I love my country.

2. My country name is Pakistan this is beautiful country.

Now, to compute the similarity between these two sentences, we begin to make the list of both texts by ignoring the order. The rehashed sentences would be like:

Pakistan is my homeland and I love country name this beautiful.

Now we will find the term frequency of each word from both strings.

We are interested in two vectors rather than the words themselves. For example, there is only one instance of Pakistan in both vectors. So, we have to decide how closer these two texts are by computing one function on those two vectors.

TABLE I Words Frequency Comparison in Strings to Find Out Cosine Similarity

	Terms	**Frequency of A**	**Frequency of B**
[1]	Pakistan	1	1
[2]	Is	1	2
[3]	My	2	1
[4]	Homeland	1	0
[5]	And	1	0
[6]	I	1	0
[7]	Love	1	0
[8]	Country	1	2
[9]	This	0	1
[10]	Name	0	1
[11]	Beautiful	0	1

TABLE II Comparison between two Strings using Sequence Matcher

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

From Table I the extracted frequency of each word from both vectors are written below:

A: [1, 1, 2, 1, 1, 1, 1, 1, 0, 0, 0]

B: [1, 1, 2, 0, 0, 0, 0, 2, 1, 1, 1]

By using equations 3, the computed value of Cosine angle between the two vectors is 0.4181. So, this value indicates that that both texts are not completely related to each other, however, there exists some relevancy between them.

B. Sequence Matching Technique:

Sequence Matcher is basically a class of *difflib* module used in Python language. With the help of this sequence matcher class it is very easy to find out the comparing sequence matcher using Ratcliff/Obershelp algorithm [] to find out the sequence of the text and the relevancy, which can be computed with the help of that sequence [12]. Ratcliff/Obershelp algorithm uses the following formula for sequence matching:

$$D_{\mathrm{r}}o = (2 * k_{\mathrm{m}})/(|S_1| + |S_2|) \tag{4}$$

In this formula, $\mathrm{k_m}$ represents the number of matching characters in sequence, whereas $|S\_1|$ and $|S\_2|$ indicates the length of the corresponding strings. The longest substring that is common in $\mathrm{S_1}$ an $\mathrm{S_2}$ is called "Anchor". The left and the right part of the string must be analyzed again because it has now become a new string and this process is repeated until all the characters of $\mathrm{S_1}$ and $\mathrm{S_2}$ get analyzed [13].

[2] www.wikipedia.com

1) Implementation of Sequence Matcher:: To find out the relevancy between two strings let's consider the two strings (see Table II) *Main String* and *Matching*. The length of the string S_1 is 11 whereas the length of string S_2 is 8.

TABLE III All Common Sequences in two Strings

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

In S_1 and S_2 the longest common substring between them is *ING* (see Table III), therefore it is an anchor, hence:

$$Km = |ING| = 3$$

Now, there is only one new substring at the left side of the Km (anchor) of both strings and is no substring on the right side of the anchor. The longest possible common sequence between the two vectors now is MA (see Table IV). Hence, MA is the new Anchor. Hence, the value of km will be:

$$Km = 3 + |MA|, \implies Km = 3 + 2 = 5 \tag{5}$$

TABLE IV All Common Sequences in two Strings

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
S1	M	A	I	N		S	T	R	I	N	G
S2	M	A	T	H	I	N	G				

Now as we can see now the MA is the start of both strings S_1 and S_1 so there is no string on the left side of both strings. And on the right of MA there is no common sub string in both strings. So the value of Km will be 5 and it will not change. Now we have all the data needed to calculate the Ratcliff/Obershelp score.

$$D_{\mathrm{r}}o = (2 * 5)/(11 + 8) \implies 10/19 = 0.5261 \tag{6}$$

The resulting value shows that the two strings are not matched with each other, however, there exists a slight similarity between the two strings. If the resulting value would have been 0, means there was no similarity at all and a value of 1 would mean a perfect match.

VII. Results

A. Success Metric

We preferred reporting our obtained results in terms of *Precision*. *Precision* refers to the closeness of two measurements with each other. More technically, it is the ratio of obtained relevant data to the retrieved data. Mathematically,

$$Precision = \frac{relevant\ data}{retrieve\ data} \tag{7}$$

B. Results

We tested different queries and computed the results in terms of precision for obtained first 15 resulting values after several queries generated using Bing API and it returned 35 resulting values to the user. We explore the results obtained from Bing API and compare that using Cosine similarity and Sequence matcher techniques.

C. Discussion of Results

In this section, we discuss and explain our obtained results. We report the precision computed on the returned results of Bing API and compare these obtained results with the results of our chosen techniques, i.e., Cosine similarity and Sequence Matcher. Figure 4 summarize our results.
Obtained results, illustrated in Figure 4, are based on the returned results of a query. For example, as a result of searching any personality like Micheal Jackson, the engine also returns the results in the form of images. It is also possible that the returned images are not of our searched celebrity (Micheal Jackson). We need to check the similarity of the query to the returned images.
Another reason for getting low precision could be the fact that the user does not put the query properly, thus, the returned results would certainly be different. Additionally, the information attached to the image could also be insufficient and by applying our chosen techniques we could get more accurate results. In Figure 4, we relied on the Bing API and computed the precision on its returned results. Later, we repeat the same process and obtained the precision by applying the Cosine similarity and Sequence Matching algorithm. It is evident that Cosine similarity is more accurate than Sequence Matching because results of Cosine similarity are 4% (overall) higher than the Bing API returned results compared to 3% for Sequence Matching.

VIII. Conclusion & Future Work

In this era, the data on the internet has been increasing day by day and the need to retrieve the accurate data has been considered very important for saving the time and money. In this research, we have studied different techniques and systems of image retrieval and image filtration processes. Different information retrieval and information filtration algorithms have been designed but the issue of accurate information retrieval is not solved yet.

In this work, we have tried to solve the problem of accurate image search on the internet and have focused on text-based image retrieval system. We used Bing API to get the desired data for its manipulation using scraping from the JSON file. Then, we extract the URLs of images and content of images into a new data set and then show all that images in a web browser to the user by using HTML format in "img" tag. We have applied Relevance feedback to find out the relevancy of the images by the user to calculate how much our research is affected for image filtration. We separate the names and attributes of the images in another section of the file, i.e., text file so that using that data we would able to find the accuracy of the text and image using that data.

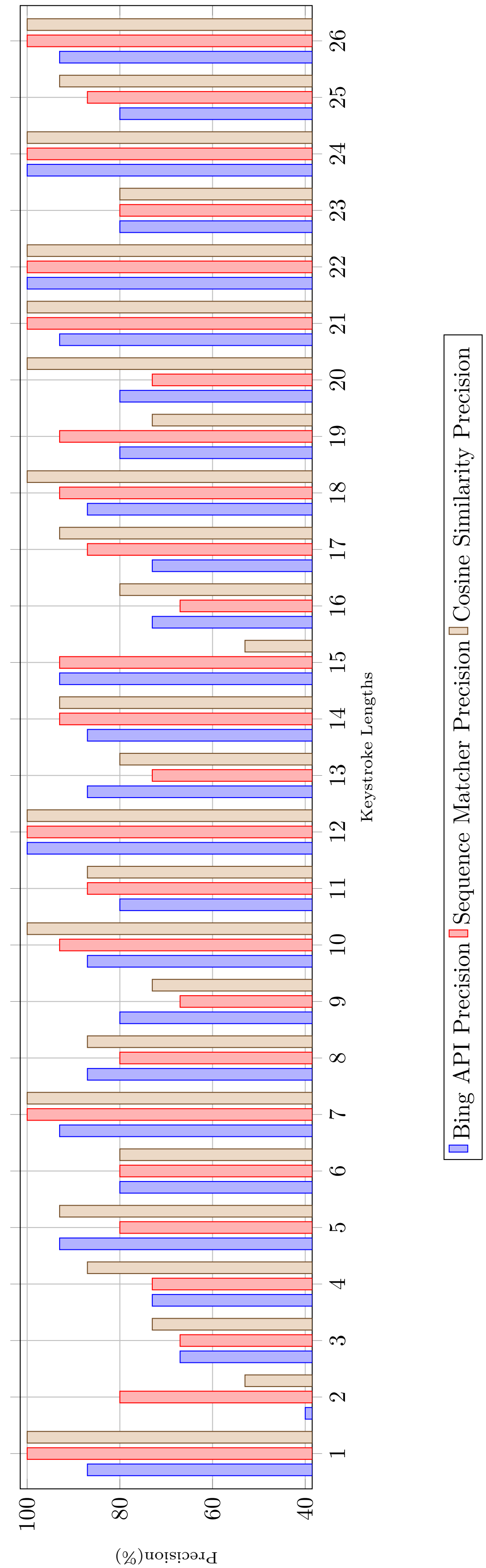

Fig. 4 Precision computed over the returned Bing API results and its comparison with the results of two techniques

In this study, we have compared the two techniques which are considered helpful in making the results more accurate and more efficient as compared to the original retrieval and filtering systems. We tested different queries and the precision of each tested query is calculated to find out the result for later use in average precision calculation of all techniques. By comparing the results of our proposed systems, i.e., using Cosine similarity and Sequence matcher techniques, we have been able to improve the original Bing API collected results. Our obtained algorithm provided more accurate results in fetching the more relevant images as compared to the Bing API.

This study could be the start towards the improvement of text-based image retrieval and filtration systems in the future. We have used the two techniques individually in this paper which could be combined to explore the accuracy in future work. The idea of combining the two scheme is worth trying and we are sure that it would be more effective and would be more efficient and accurate.

References

[1] Mok, Diana and Wellman, Barry and others, *Did distance matter before the Internet?: Interpersonal contact and support in the 1970s*, Social Networks, vol.29, no.3, pp. 430–461, 2007

[2] Kumaravel, Krishnan, *Comparative study of 3G and 4G in mobile technology*, International Journal of Computer Science Issues (IJCSI), vol.8, no.5, pp. 256, 2011

[3] Fagbohun, O, *Comparative studies on 3G, 4G and 5G wireless technology*, IOSR Journal of Electronics and Communication Engineering, vol.9, no.3, pp. 88–94, 2014

[4] Duan, Guoyong and Yang, Jing and Yang, Yilong, *Content-based image retrieval research*, Physics Procedia, vol.22, pp. 471–477, 2011

[5] Rui, Yong and Huang, Thomas S and Chang, Shih-Fu, *Image retrieval: Current techniques, promising directions, and open issues*, Journal of visual communication and image representation, vol.10, no.1, pp. 39–62, 1999

[6] Smeaton, Alan F and O'Connor, Edel and Regan, Fiona, *Multimedia information retrieval and environmental monitoring: Shared perspectives on data fusion*, Ecological informatics, vol.23, pp. 118–125, 2014

[7] Hanani, Uri and Shapira, Bracha and Shoval, Peretz, *Information filtering: Overview of issues, research and systems*, User modeling and user-adapted interaction, vol.11, no.3, pp. 203–259, 2001

[8] Wu, Pengcheng and Hoi, Steven CH and Xia, Hao and Zhao, Peilin and Wang, Dayong and Miao, Chunyan, *Online multimodal deep similarity learning with application to image retrieval*, 21st ACM international conference on Multimedia, pp. 153–162, 2013

[9] Christopher, D. Manning and Prabhakar, Raghavan and Hinrich, Schutza, *Introduction to information retrieval*, An Introduction To Information Retrieval, vol.151, no.177, 2001

[10] Rani, Deepu and Goyal, Monica, *A Research Paper on Content Based Image Retrieval System using Improved SVM Technique*, International Journal Of Engineering And Computer Science, vol.3, no.12, 2014

[11] Microsoft Azure, *https://docs.microsoft.com/en-us/azure/*, last accessed *March 2017*

[12] Python diff lib Documentation, *https://docs.python.org/2/library/difflib.html*, last accessed *May 2017*

[13] Ilyankou, Ilya, *Comparison of Jaro-Winkler and Ratcliff/Obershelp algorithms in spell check*, IB Extended Essay Computer Science, 2014

[14] Choi, Min and Jeong, Young-Sik and Park, Jong Hyuk, *Improving performance through rest open api grouping for wireless sensor network*, International Journal of Distributed Sensor Networks, vol.9, no.11, 2013

[15] Bathla, Gourav and Jindal, Rajni, *Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold*, International Journal of Computer Applications, vol.23, no.5, 2011

[16] W.F Du, G.X. Chen. ”Analysis and Research of Several Problems of Bad Short Message Filtering System.” International Conference on Computer Information Systems and Industrial Applications , 2015.

[17] Balwinder Siani, Vikram Singh and Satish Kumar. ”Information Retrieval Model and Searching.” International Journal of Advance Foundation and Research in Science Engineering (IJAFRSE) , 2014.

[18] Chu-Xu Zhang, Zi-Ke Zhang, Lu Yu, Chuang Liu, Hoa Liu, Xiao-Yong Yan. ”Information filtering via collaborative user clustering modeling.” Elsevier , 2011.

[19] Chao, C. , Qu, S. and Du, T. ”Research of Collaborative Filtering Recommendation Algorithm for Short Text.” Journal of Computer and Communications, 2, 59-66. doi: 10.4236/jcc.2014.214006.