Implementation of Text Base Information Retrieval Technique

Syed Ali Jafar Zaidi[1], Safdar Hussain[2]
Department of Computer Science
Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

Samir Brahim Belhaouari[3]*
Division of Information and Computing Technology
College of Science and Engineering
Hamad Bin Khalifa University, Qatar

***Abstract*—Everyone is in the need of accurate and efficient information retrieval in no time. Search engines are the main source to extract the required information, when a user search a query and wants to generate the results. Different search engines provide different Application Programming Interface (API) and Libraries to the researchers and the programmers to access the data that has been stored in servers of the search engines. When a researcher or programmer search's a query by using API, it returns a Java Script Orientation Notation (JSON) file. In this JSON file, information is encapsulated where scraping techniques are used to filter out the text. The aim of this paper is to propose a different approach to effectively and efficiently filter out the queries based on text which has been searched by the search engines and return the most appropriate results to the users after matching the searched text because the previous techniques which are used are not enough efficient. We use different comparison techniques, i.e. Sequence Matcher Method and then compare the results of this technique with relevance feedback and in the end we found that our proposed technique is providing much better results.**

***Keywords*—Information retrieval; sequence matcher method; relevance feedback**

I. Introduction

Well before the invention of the internet it was so much tough to keep in touch with the world. But, with the dramatic growth of internet (see Fig. 1), it is so much easier for the people to remain in touch with each other's and to spread the information over the world using internet. Approximately, 80% data available on internet is in textual form and is highly unstructured. So, during last two decades, the websites, web blogs and other informative material contain such a massive amount of textual and unstructured data [1]. When anyone uses internet he usually deals with text because text is the main source of information and communication on the internet, majority of internet searches are text-based. This expanding availability of text has demanded lots of research in this area [13].

Internet is the connection of two or more than two computers which can communicate with each other's. Billions of people are connected on internet and are the source of generating text over the internet [1]. As the use of text is increasing day by day and further there is growth in technology is noticed as well, this was not the case in past as in previous decades the only source to use the internet facility was our desktop computers. These desktop computers used internet with wired net but now user can interact with the internet with their laptops, tablets, smart phones and even by using their smart watches with 3G, 4G and 5G technologies [8][13]. With this fast increase in usage of electronic media technology, the speed of internet has also increased such as 3G, 4G and 5G, as these are the technologies which are intended to deliver the required information in the no time [2].

Increased text over the internet is a result of rapid increase in the internet usage (see Fig. 1). So, the chance of irrelevant data extraction fist also increased, and the filtration of irrelevant data is so much necessary in this scenario [3]. Keeping in mind the defined issues, this paper addresses the issue of fetching the most appropriate and relevant text by using the text-based queries in efficient and effective manners. From our expertise in the normal world, we noticed that people are in touch with the text of different types regularly for different reasons [12].

People in this world are in touch with newspapers, televisions, radio, graphical representations and different advisory services to understand, to learn and more specifically to remain in touch with the revolution and change in the world which is occurring in daily life. The main feature for all such activities is to judge how useful and meaningful this text can be about their requirements [11]. By association, we tend to imply that, in such exercises, people don't seem to be recently upstage beneficiaries of messages, however instead dynamic searchers of writings, and therefore the active constructors of importance from these writings. They search around for writings of potential intrigue, so as to support individuals in their information [5]. In this study we have focused on the problem of relevant, real, and precise text retrieval which guarantees correct and precise results to entertain the user needs [22].

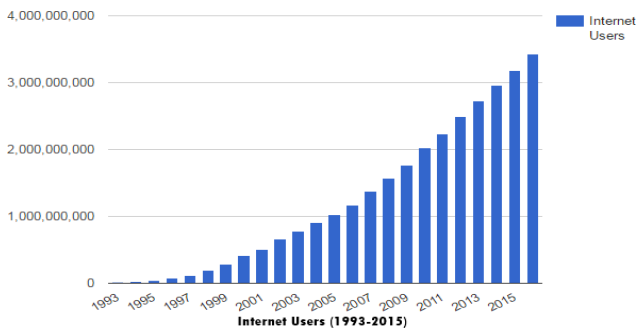

Fig. 1. The Increase of Internet users forms 1993-2015.

*Corresponding Author

II. Material and Methods

A. Approach

In this research, Bing API is used to retrieve the text by applying text-based search query. If we talk about Bing API, it is one of the very useful tools to fetch the information in the form of text or in the form of multimedia information from the server [21]. The results that are returned by this API are in the form of Java Script Orientation Notation file, we then further process the collected results to fetch the text of the information. The gathered information in maximum of the cases were not the wanted ones, thus, we have to create a system that will be helpful in retrieval of the most accurate information from the data.

B. Information Retrieval and Information Filtering

Some Modern information and text retrieval techniques are used to fetch the data in minimum possible time. The basic aim of Information retrieval model is to "discover the relevant knowledge-based information" or a document that fulfill user needs [4]. In modern days the term used is Information Retrieval (IR) rather than Information Gathering. This information can be in any form e.g. Image, Sound, Video, Text or anything which can be used in meaningful purpose [16]. When someone will search anything with the help of Search Engine it will return all the possible results which will match that query [7].

Nowadays, search engines are not the only source to retrieve information (Fig. 2). In parallel, the social media sites are also the big source of the Information Retrieval anyway [10][19]. As compared to past, the Information Retrieval is not only becoming fast but also becoming more accurate nowadays by implementing more reliable and more efficient algorithms [14]. But, in contrast, the chances of Irrelevant Information Retrieval cannot be eliminated [9]. So, to make it more efficient and more reliable, different we have implemented an information retrieval algorithm on actual data which is retrieved.

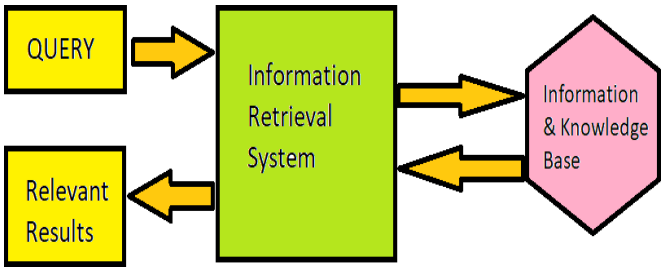

Fig. 2. Information Retrieval System.

III. Implementation

Whenever a user wants to search something on computer or any electronic device, he uses search engine(s) or social media site(s) to write a query and, as a result, receives the relevant results against the query. Relevancy of the query can be judged by different methods which are not sufficient to identify the best relevant text of the information because there are many other words and stop words that can minimize the relevancy of the relevant text. Besides this technique, as it seems not suitable or aligned sufficient, our study aims to find out the relevancy of a text, we use the Sequence Matcher Method which could be more helpful to retrieve appropriate and effective information. But to find out the efficiency of our algorithm we took the relevancy feedback from the user and then compared the precision of relevance feedback with precision Sequence Matcher Method. Where precision refers to a term that how close a result is close to the overall results which has been retrieved.

A. Relevance Feedback

The Relevance feedback refers to the feedback from the user has been taken to identify that how the returned data from the server against the query is relevant for the user [18]. The basic aim of relevance feedback is to check the relevant results of retrieval systems. The basic procedure of relevance feedback is that user will search the query, then the user will tell that which data is relevant and which one is not relevant.

B. Sequence Matcher

Sequence Matcher is basically one of the module used in python programming language, because of this module it is very easy to find out the comparison of the strings, to find out the sequence of the text and then relevancy which is computed with the help of that sequence [19][20]. Sequence Matcher Method uses equation for sequence matching which is given below.

$$D_{ro} = \frac{2*k_m}{|S_1||S_2|},$$

where, k_m denotes the number of same characters in sequence whereas |S1| & |S2| specifies the length of these both strings correspondingly.

The longest substring that is common in both S1 and S2 strings is called anchor. The right and the left part of the string must be examined once again because it will be considered as a new string and this process is repeated again and again until all the characters of S1 and S2 are examined [6][15].

C. Implementation of Sequence Matcher

To implement Sequence Matcher Method, consider the strings Waiting and Main String (Table I).

The length of the string S_2 is 7 whereas the length of string S_1 is 11. Now

$|S_1| = |11|$ *and* $|S_2| = |7|$

In S1and S2 the longest common substring between them is ING therefore ING is an anchor hence now.

$K_m = |ING| = 3$

So, in (Table II) now there is only a substring on the left side of the Km (anchor) of both strings and no substring on the right side of the anchor now the next sequence which is the longest one between them is (AI) which is the second largest sequence. So, AI is new anchor and the Km value will be

$K_m = 3 + |AI|$

$K_m = 3 + 2 = 5$

TABLE I. COMPARISON OF TWO STRINGS

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

TABLE II. LARGEST COMMON SEQUENCE IN TWO STRINGS

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

TABLE III. ALL COMMON SEQUENCES IN TWO STRINGS

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

As we can see now in (Table III) now the AI is in the both strings S1 and S2 so there is no common string on the left side of both strings. And on the right of AI there is another common substring in both strings ING so the value of Km will be 2+3 = 5. Till now we have obtained all the values which are needed to calculate the final score. Now the calculation of string MAIN STRING and WAITING is given below by using equation number.

$$D_{ro} = \frac{2 * 5}{11 + 7}$$

$$D_{ro} = \frac{10}{18} = 0.555$$

IV. RESULTS AND DISCUSSIONS

To check the efficiency and accuracy of the system, we have tested our information filtering system in contrast to the Bing API or Relevance Feedback. As a result, we got different results as compared to Bing API, this has been tested over different queries and then relevance feedback from the user has taken and then compared to Sequence Matcher Method. We have retrieved only the first twenty results against several queries using precision; although Bing API returns almost thirty-five results. Precision is the total number of relevant information over total information, the equation is given below.

$$Precision = \frac{relevant\ data}{retrieved\ data}$$

Precision is the fraction of a query which is searched by the user that is related to the particular query. Its calculation is needed to find out the relevant and non-relevant results in the evaluated documents [17]. Hence the whole results of different tests of different queries has been shown in (Table IV) which is given below.

After applying different experiments, it is clearly seen that the precision value keeps on changing for each and every result, so it shows that our system is effective and efficient. As we can see in the table that the precision of sequence matcher technique is relatively better than the precision of relevancy feedback in maximum cases.

TABLE IV. RESULTS COMPARISONS OF SEQUENCE MATCHER PRECISION WITH RELEVANCE FEEDBACK PRECISION

	Searched Query	**Relevance Feed Back**	**Relevant Sequence Order**	**Precision of Relevance Feedback**	**Precision of Relevance Sequence Order**
1	Jinnah International Airport	13	15	0.87	1.00
2	England Flag	6	12	0.40	0.80
3	Hello	10	10	0.67	0.67
4	Tom Cat Cartoon	11	11	0.73	0.73
5	Pakistan Super League	14	12	0.93	0.80
6	Fakhar Zaman Cricketer	12	12	0.80	0.80
7	Suzuki Cars	14	15	0.93	1.00
8	Indus River	13	12	0.87	0.80
9	David Miller	12	10	0.80	0.67
10	Qamar JavaidBajwa	13	14	0.87	0.93
11	Urdu Alphabets	12	13	0.80	0.87
12	Micheal Hussey	15	15	1.00	1.00
13	George Bush	14	14	0.93	0.93
14	Salman bin Abdulaziz	11	10	0.73	0.67
15	Blue Line Shark	11	13	0.73	0.87
16	National Mosque of Pakistan	12	14	0.80	0.93
17	Lawn Tennis	12	11	0.80	0.73
18	Mehdi Hassan	14	15	0.93	1.00
19	ABC Alphabets	15	15	1.00	1.00
20	Jerry Cartoon	15	15	1.00	1.00

V. CONCLUSION

In this era, data has been growing on daily basis over the internet and one of the important things is to gather data in no time. In this research, we have developed a technique of information retrieval and information filtration processes. There are several algorithms of information filtration and information retrieval that currently have been developed, but the key issue which is still to be addressed is the design of an accurate information retrieval.

In this research, a novel technique is applied which can be considered helpful and may outperform in creation of results more precisely and efficiently as compared to the original information retrieval and filtration systems. Different tests against different queries have been done and the precision of each tested query is calculated to find out the result for future use in average precision calculation of all the techniques. After comparing the results of our proposed system with relevance feedback on different queries, we have found that our proposed system has been able to improve the original Bing API collected results. Our developed algorithm provided more accurate and precise results in fetching the more relevant information as compared to the Bing API.

This study could also be helpful in the development of text-based information retrieval and filtration systems in the near future. We have used the Sequence Matcher Method individually in this paper which could be combined with relevance feedback to explore the accuracy in future work. The idea of comparing sequence matcher method with Relevance feedback is worth trying and we are sure that it would be more effective and precise.

ACKNOWLEGMENT

The authors would like to thank Qatar National Library (QNL) for supporting in publishing the paper.

REFRENCES

[1] Mok, et al: Did distance matter before the Internet? Interpersonal contact and supporting the 1970s, Social Networks, (2007).

[2] Nandha kumar Pandti1, Mohsin Nargund: Mobile Communication -Past, Present and Future: A REVIEW, (2018).

[3] Saad Farooq: Aversarial Information Retrievalon the Web (2018).

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

[5] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Margaret Mitchell. "Large Scale Retrieval and Generation of Image Descriptions." Springer Science,2015.

[6] Ilyankou, Ilya. "Comparison of Jaro-Winkler and Ratcliff/Obershelp algorithms in spell check." IB Extended Essay, 2014.

[7] Bhakar Mitra, Nick Craswell, Neural Model of Information Retrieval (2017).

[8] K.Kumaravel. "Comparative Study of 3G and 4G in Mobile Technology." IJCSI International Journal of Computer Science Issues, 2011.

[9] 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.

[10] M. Rami Ghorab, Dong Zhou, Alexander O'Connor, Vincent, "Personalised Information Retrieval: survey and classification." © Springer Science, 2012.

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

[12] Gourav Bathla, Rajni Jindal. "Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold." International Journal of Computer Applications (0975 – 8887), 2011.

[13] Diana Mok, Barry Wellman, "Did distance matter before the Internet?" Elsevier, 2007.

[14] Syed Ali Jafar Zaidi, Attaullah Buriro, Mohammad Riaz, Athar Mahboob and Mohammad Noman Riaz, "Implementation and Comparison of Text-Based Image Retrieval Schemes" International Journal of Advanced Computer Science and Applications(ijacsa), 10(1), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100177

[15] Alan F. Smeaton, Edel O'Connor, FionaRegan. "Multimedia information retrieval and environmental monitoring: Shared perspectives on data fusion." Elsevier, 2013.

[16] Christopher D. Manning, Parbhakar Raghavan, Hinrich Schutze. An introduction of Information Retrieval. Cambridge: Cambridge University Press, England, 2009.

[17] Sharan Narang, Gregory Diamos, Erich Elsen, Paulius Micikevicius, Jonah Alben, David Garcia, Boris Ginsburg, Michael Houston, Mixed Precision Training, 2017.

[18] Upendra Shardanand, Pattie Maes, Social Information Filtering Algorithms for Automating "Word of Mouth" May 1995.

[19] Nicholas J.Belkin, W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? Communications of the ACM 1992.

[20] Tanmoy Mondal, Nicolas Ragot, Jean-Yves Ramel, Umapada Pal, Flexible Sequence Matching Technique: Application to Word Spotting in Degraded Documents 2014 14th International Conference on Frontiers in Handwriting Recognition.

[21] Mike Thelwall, Pardeep Sud, "WeboMetric Research with the Bing Search API 2.0" Journal of Informetrics Pages 44-52 2012.

[22] MM Eltoukhy, I Faye, BB Samir: Breast Cancer Diagnosis Based on Texture Feature Extraction Using Curvelet Transform, International Congress on Instrumentation and Applied Sciences. Kuala Lumpur, Malaysia 2010.