ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.8

Web Mining using Lexical Chain with Interactive Genetic Algorithm

Jyoti S. Pachare
Research Scholar
Department of Computer science
Wainganga College of Engineering and Management, Nagpur, India

Abstract:
Web is the biggest information system currently whose data has some characteristics of abundance, diverse form, many different structures and dynamic varieties, etc. A novel technique is proposed for summarizing text using a combination of Genetic Algorithms (GA) and lexical chaining. The novelty of the proposed algorithm is that fuzzy system is optimized for extractive based text summarizing and to extract images, audio, video data using interactive genetic algorithm (IGP) .The other one is using lexical chains as a representation of the lexical cohesion that exists in the text. We propose a novel approach that incorporates lexical chains into the model as a feature and learns the feature weights by genetic algorithms and for pictorial data by interactive GA. The goal is to develop an optimal intelligent system to extract important data in the texts by reducing the redundancy of data.

Keywords: Genetic Algorithm (GA), Text Summarization, Interactive Genetic Algorithm (IGA), lexical chain, Redundancy of data etc.

I. INTRODUCTION

The world-wide web is the most important source of information for most of us. Unfortunately, if there is no guarantee for the correctness of information on the web. The information which we need for us the different web sites often provide conflicting information on a subject, such as different specifications for the same product. Our analysis provides a new problem called Veracity, truth confirmation on the information, which studies how to find true facts from a large amount of conflicting information on many subjects that is provided by various web sites. A general framework for the Veracity problem by using lexical chain mechanism throught word adjacency between the words. Which utilizes the relationships between web sites and their information that a web site is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy web sites. Our study show that genetic algorithm successfully finds true facts among conflicting information, and also identifies the disadvantages on existing trustworthy web sites better than the popular search engines. With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyze their usage patterns. The world-wide web has become a necessary part of important information source for most people everyday people retrieve all kinds of information from the web for example when online shopping people find product specifications from web site like ShopZilla.com looking for interesting DVD they get information and review on web sites such as NetFlix. com or IMDB.com. Web services are the new industrial standard for distributed computing and are considered, for the first time, a real opportunity to achieve universal interoperability. Besides enabling such interoperability, web services can also be used as communication protocols for efficient and effective business application integration. At the same time, just with any new technology, these web services also bring with them some computational complexities and business challenges.

II. RELATED WORK

Vint generates rules to extract search result record srr (data records) from dynamically generated text result documents. In this approach the relevance of different features. Additionally, at least four data records have to be present in a result page to automatically build a wrapper. In the case that the data records are distributed over multiple sections (data regions), the extraction component returns only The data records found in the largest section. Viper [5] takes the approach of extracting structured data records contained in html pages. This wrapper tries to identify repetitive structures contained in html source code and finally weights and separates the patterns according to the rendering information of the browser. Moreover, this wrapper can be used efficiently to extract structured data records which acts as a disadvantage as it is unable to extract semi-structured data [10]. Vsdr [6] is a visual wrapper and is used to extract data records from semi-structured web pages. This wrapper uses the mechanism of computing the visual centre of an html page and constructs a boundary region for the data area. However the size of the boundary region may be overestimated in some case and the extracted region may include non record data. In such cases any non record data such as menu bars and navigational links will also be Identified as data records which is considered as a drawback of vsdr wrapper. Viwer is also a visual wrapper. Unlike the above wrappers it also implements dom tree properties along with visual cues during extraction of data. The web page is first converted into a dom tree using bfs extraction method. This is to identify potential data regions. This is followed by filtering to remove the noisy data using filtering stages like html tag and largest scoring function filter. However the visual cue used i.e. Bounding box of html tags only considers the centrally located data which consists of the largest data region present. Hence even if there are relevant data regions not located centrally they are not considered. Hence the wrapper discussed in this paper overcomes this disadvantage. This can be overcome by using the process of content hyperlink and keyword filtering which will below.

III.PROPOSED WORK

Step 1:

Connectivity with web and input a user query and forming a DOM tree to extract useful links. In the first modul, we will do connectivity with web. after successful connection is done a window will appear which ask for input a quarry. Now user will enter a quary and press GO button. Another window will show the links and this links will contain information. When a query is submitted to a search engine, the search engine returns dynamically generated result page containing the result records. Each Search Result Record pages (SRR) consists of a link to retrieved web page and some relevant information (snippet) of retrieved web page. The wrapper is divided into two main parts. The first part involves the parsing of the HTML web page and storing them as Document Object Model (DOM) tree. This is done in order to understand the structure of the HTML page to be processed. This method is useful in gauging the structure of unstructured or semi-structured HTML pages. In the second part, the wrapper carries out the various extraction techniques to extract the relevant link. However noisy data link is also extracted in the process. These links such as advertisements, images and menus are filtered out by the filtering stages. The DOM tree approach is the most effective way to identify the HTML tags or codes before the web data extraction. A DOM tree is created as a reference to gauge the structure of the page and is especially useful for unstructured or semi-structured source code. The approach used for DOM tree creation is explained further with an example. The user enters a tag name as input says, <a>, which forms the root node for the tree. A DOM tree must be created for each of the <a> tags present in the HTML document. Hence, first the number of <a> tags is calculated using the length() and stored in a counter. This counter is used as a terminator i.e. when the value of count is zero, it indicates the DOM trees for all <a> tags are created and the process is stopped. Now when the counter value is non-zero, the tag name is displayed and further verified to see if it is a text node r not. If it is not a text node, the descendents of the tag are found using the first Child() and displayed. They are then used for further processing as given above. This is because usually in HTML documents, text nodes are considered as leaf nodes, as a result if the tag is a text node, then backtracking is done and the sibling of its parent node is found using getSibling(). if a sibling exists, i.e. if it is not null, then it is displayed and the previous step is carried out again. If not, then it means that no further descendents of <a> exist and the DOM tree for one <a>tag has been created successfully. The counter value is decremented by one and checked to see if it is zero. The process is stopped once counter value becomes zero. In the following extraction process, a DOM tree is created first, followed by the BFS (Breadth First Search) extraction process. After carrying out this process, filtering stages are performed followed by the final extraction of potential relevant data regions separated from the irrelevant ones. After the DOM tree is created BFS is carried out. Usually potential data records in a DOM tree can be identified as those which lie in the same level in the tree and which have a common parent tag and recurring sequence of HTML tags. Hence to identify records which satisfy all of these conditions, BFS extraction is used. In this process nodes which are in the same level are checked and if they are same then they are group as a set of potential data records. However if they are not same then the search is carried out one level lower to check for similarity in the tags. Also if the similar tags are placed many nodes apart it doesn't make any difference. Hence nodes which are similar irrespective of their distance are grouped.

Step 2:

Applying various text features on text data using lexical chain.

Text Features

Initially, a parser is designed that extracts the desired features. This program parses the text into its sentences and identifies the following non-structural features for each sentence as the input of fuzzy inference system.

1. No. of title words - Feature 1 indicates the number of title words in a sentence relative to the maximum possible. This is determined by counting the number of matches between the content words in a sentence and the+- words in the title. This feature is expected to be important because the salience of a sentence according to ISO definition may be affected by the number of words in the sentence also appearing in the title.

2. length of a sentence - Feature 4, length of a sentence, is useful for filtering out short sentences such as datelines and author names commonly found in news articles. We also anticipate that short sentences are unlikely to be included in summaries.

3. Numbers of thematic words - It indicate the words with maximum possible relativity. It is determined as follows: first we remove all prepositions and reduce the remaining words to their morphological roots. The resultant content words in the document are counted for occurrence. The top 10 most frequent content words are considered as thematic words. This feature is important because terms that occur frequently in a document are probably related to its topic .

4. emphasize words - words such as very, most, etc. because the important sentences can be signified with these kinds of words.

5. sentense location - usally the intial sentense in a document are the most important ones. so the first sentense is assume to be most important.

6. sentense relative length - we assume that longer sentense contain more information. for a sentense s in a document d,the feature score is calculated as

$$F_6(d,s) = \frac{\text{sentense lenght (s)}}{\text{max sentense len ght (si)}}$$

Where ns is the number of sentense in the document.

7. Average TF - The term frequency metric is based on two assumptions

i) The importance of a term for a document is directly proportional to its number of occurences in the document .

ii) the length of the document does not affect the inportance of the terms. the tf for the term t in a document d is calculated as, where nt denotes the number of terms in a document .

$$F_6(d,t) = \frac{\text{term frequency (d,t)}}{\text{max term frequency}}$$

8. Cue words - The sentense that contain some of the words or items that contain salient information about the document. This feature count the number of cue words.

$$F_8(d,s) = \frac{\text{cue words (s)}}{\text{sentense length}}$$

9.Numerical data - Terms that are written in numerical form sometimes convey key information about a document. We test the usefullness of such terms using this feature. This feature counts the number of numerical terms in a sentence:

$$F_9(d,s) = \frac{\text{numerical terms}}{\text{sentense length}}$$

10. Sentence centrality- This feature measures the vocabulary overlap between a sentence and the other sentences in the document. This is an indication of the importance of a sentence for a document. For a sentence s in the document d, the feature score is calculated as follows: (8) where c-terms is the number of common terms that occur both in s and in a sentence d other than s, and nt is the number of terms in the document.

$$F_{10}(d,s) = \frac{c-\text{terms}}{nt}$$

where c-terms is the number of the common terms that occure in s and in sentense s other than d.

11. Synonym links - This feature is another form of sentence centrality and attempts to measure the centrality of a sentence using the number of common synonymous words in the sentences. We consider nouns only and we extract the nouns in sentences using the LingPipe part-of-speech (PoS) tagger .The synonymy relation between two nouns is determined by looking whether they have a synset in common in WordNet. The feature score is calculated as

$$F_{11}(d,s) = \frac{s-\text{links}}{ns}$$

where ns is number of the sentense in the document. and s-links is the number of synonym links between s and other sentense in document.

The selection of features plays an important role in determining the type of sentences that will be selected as a part of the summary and, therefore, would influence the performance of this fuzzy inference system.

Step 3: learning feature weight of data and take pictorial data from the extracted link.A novel aspect of the proposed approach is using the lexical chain concept as a sentence feature in the system. We first compute the lexical chains for the document, give a score to each chain, and select the strongest chains. Then, we score the important and average sentences according to their inclusion of strong chain words. The lexical relations between words are extracted using WordNet. When lexical chains are computed, each word must belong to exactly one chain. There are two challenges here. First, there may be more than one sense for a word (ambiguous word) and the correct sense must be identified. Second, a word may be related to words in different chains. The aim is to find the best way of grouping the words that will result in the longest and strongest chains. In this work, we consider only nouns as the candidate words and first determine the nouns using the LingPipe PoS tagger.

Step 4 : apply IGA i.e. interactive genetic algorithm**.** Under the proposed system experts use computer-generated images showing cells in the best condition according to their diagnosis. The system uses this information to build its knowledge base. The images that are diagnosed by experts as the best ones are generated using an Interactive Genetic Algorithm (IGA) .The IGA is the method can architect according to the sensitivity of human design and elect the information's because the human perform the portion of genetic algorithm.

Genetic Algorithm steps:

Use of genetic algorithms in optimization can be summarized as follows. For more information. A solution to a problem is defined in terms of images from different links. The process is called genetic algorithm because these images resemble DNA chains characterizing a living organism. The forming of future generation of solutions is achieved by mathematical operations that resemble crossover and mutation. The process starts by choosing *n* feasible solutions in the form of *n* image as the original population. Each solution on this population is evaluated via the objective function for the problem and assigned a fitness value. The rest of the process is summarized in the following steps. *Crossover:* Each pair of solutions is combined to generate a new pair of solutions. This combination, often referred to as crossover, is accomplished by breaking each chain at certain location and interchanging the half-chains between two image. These images are added to the population, and their fitness values are evaluated. *Selection:* A new pool of size *n* is then formed by a process referred to as selection. In this process, the new pool is formed by giving a higher probability of selection to image with higher fitness values. This population then becomes the target for a new round of crossover process. *Mutation:* Mutation causes individual image to be changed according to some probabilistic rule. Usually, only a small part of image is changed by mutation, causing the offspring t inherits most of the features of the parent. *Termination:* The process terminates when additional applications of the algorithm do not result in a significant change in the overall fitness, or the computation budget limit is reached. When this happens, the search stops, and one of the best solutions is chosen as the optimum for the problem.

IV. SIMULATION RESULT

In the first step we will make connectivity with web. After running the poject a window will appear as shown in fig 1 . this is the queary window in which user can enter any queary which he wants to search about. And then press on Go button.

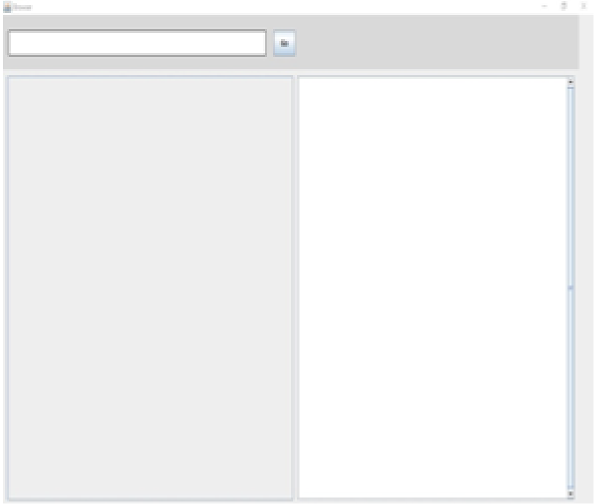
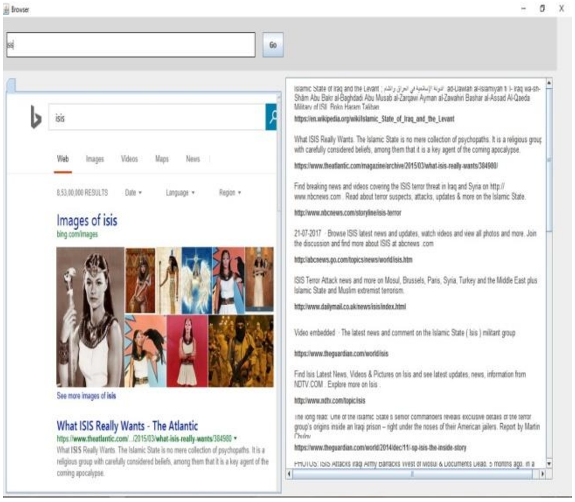


Figure.1. Window to enter a query.

After pressing a Go button a new window will appear which is shown in fig 2. This window will show a web page which consists of lots of web links and some images in left panal of the window. Now proposed algorithm will take all the links from the web page and form DOM tree. And will extract useful 10 links from web page. And after this proposed algorithm will take text data from each link and apply text feature on text data and this way will find a text summary from each link and will show in right panal of the window along with the link as shown in the fig. 2. This text summary is form by applying various text features on web data and then by finding the lexical correspondence between the text sentences.



Islamic State of Iraq and the Levant ; الدولة الإسلامية في العراق والشام ad-Dawlah al-Islāmiyah fī l-'Irāq wa-sh-Shām Abu Bakr al-Baghdadi Abu Musab al-Zarqawi Ayman al-Zawahiri Bashar al-Assad Al-Qaeda Military of ISIL Boko Haram Taliban

https://en.wikipedia.org/wiki/Islamic_State_of_Iraq_and_the_Levant

What ISIS Really Wants. The Islamic State is no mere collection of psychopaths. It is a religious group with carefully considered beliefs, among them that it is a key agent of the coming apocalypse.

https://www.theatlantic.com/magazine/archive/2015/03/what-isis-really-wants/384980/

Find breaking news and videos covering the ISIS terror threat in Iraq and Syria on http:// www.nbcnews.com . Read about terror suspects, attacks, updates & more on the Islamic State.

http://www.nbcnews.com/storyline/isis-terror

21-07-2017 · Browse ISIS latest news and updates, watch videos and view all photos and more. Join the discussion and find more about ISIS at abcnews .com

http://abcnews.go.com/topics/news/world/isis.htm

ISIS Terror Attack news and more on Mosul, Brussels, Paris, Syria, Turkey and the Middle East plus Islamic State and Muslim extremist terrorism.

http://www.dailymail.co.uk/news/isis/index.html

Video embedded · The latest news and comment on the Islamic State (Isis) militant group

https://www.theguardian.com/world/isis

Find Isis Latest News, Videos & Pictures on Isis and see latest updates, news, information from NDTV.COM . Explore more on Isis .

http://www.ndtv.com/topic/isis

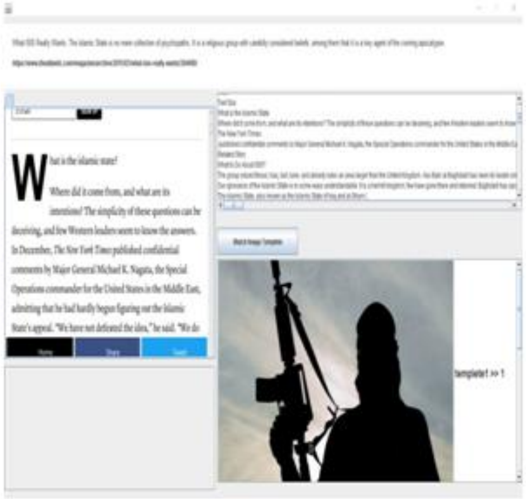
The long read: One of the Islamic State's senior commanders reveals exclusive details of the terror group's origins inside an Iraqi prison – right under the noses of their American jailers. Report by Martin Chulov

https://www.theguardian.com/world/2014/dec/11/-sp-isis-the-inside-story

PHOTOS: ISIS Attacks Iraqi Army Barracks West of Mosul & Documents Dead. 5 months ago. in a

Figure.2. window which shows all links and image data with extracted links and text summary.

Fig. 3 window shows the web page of a particular selected link from 10 extracted links by simply clicking on any link. Then a new window will appear which shows a text summary of link data and that link . a web page of selected link will shows. After that another window will shows textual data of that link. After that another window will shows the extracted data from web data by applying DOM tree technic and by applying various text features and again by applying lexical chaining mechanism. on summarised data . finally a new summary will form which is short in length ,not having redundant data, truthful data and a useful data for user.


What ISIS Really Wants. The Islamic State is no mere collection of psychopaths. It is a religious group with carefully considered beliefs, among them that it is a key agent of the coming apocalypse.

What is the islamic state?

Where did it come from, and what are its intentions? The simplicity of these questions can be deceiving, and few Western leaders seem to know the answers. In December, The New York Times published confidential comments by Major General Michael K. Nagata, the Special Operations commander for the United States in the Middle East, admitting that he had hardly begun figuring out the Islamic State's appeal. "We have not defeated the idea," he said. "We do

Figure.3. Window which shows extracted text summary and extracted image.

Another next window will shows a extracted image from a web page. Link web page has various images. So interactive genetic algorithm will find a important and unimportant images for user by applying various genetic algorithm technic like selection, crossover, mutation, termination processes and by applying DOM tree technic along with applying a classification technic on web images. So finally various extracted images will shown in another window as shown in fig. 3.

V. CONCLUSION AND DISCUSSION

In this paper, a novel technique is proposed for summarizing text using a combination of Genetic Algorithms (GA) and lexical chaining. The novelty of the proposed algorithm is that fuzzy system is optimized for extractive based text summarizing and to extract images, audio, video data using interactive genetic algorithm (IGP) .The other one is using lexical chains as a representation of the lexical cohesion that exists in the text. We propose a novel approach that incorporates lexical chains into the model as a feature and learns the feature weights by genetic algorithms and for pictorial data by interactive GA. The goal is to develop an optimal intelligent system to extract important data in the texts by reducing the redundancy of data. Another future scope is to develop algorithms for mining audio and video data

VI. REFERENCES

[1]. Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India. 1. Weiha Feng, Zhangfeng Mao. The Research of WebPages Information Extraction Based on Web, Journal ofLuoyang Technology College,3(2005)30-31 (inChinese).

[2].Zhili Zhou, Renwu Wang, A Study of Web DataAutomation Extraction and its Application, ECOMMERCE, 4(2006) 58-63 (in Chinese).

[3]. Kiani, A., Akbarzadeh, M.r 2006. Automatic Text Summarization Using Hybrid Fuzzy GA - GP. In IEEE International Conference on Fuzzy System, pp. 5465-5471.

[4]. Jarod Kelly and Panos Y. Papalambros and Colleen M. Seifert 2008 Interactive Genetic Algorithms for use as Creativity Enhancement Tools.

[5]. Luigi Troiano, Cosimo Birtolo, and Gennaro Cirillo 2009 "Interactive Genetic Algorithm for choosing suitable colors in User Interface".

[6]. Matthew Walker 2001 Introduction to Genetic Programming.

[7]. Ladda Suanmali1, Mohammed Salem Binwahlan2 and Naomie Salim 2]009 Ninth International Conference on Hybrid Intelligent Systems "Sentence Features Fusion for Text Summarization Using Fuzzy Logic".

[8]. Tao – Hsing Chang and Chia-Hoang Lee IEEE TRANSACTIONS ON FUZZY SYSTEMS, AUGUST 2007 "Subtopic segmentation for small corpus using a Novel Fuzzy Model".

[9].Xia peng and Beijing China IGARSS 2010 IEEE International "Hybrid Genetic Algorithm (GA) – based neural network for multispectral image fusion".

[10] Mine Berker, M.r. 2012. "Using Genetic Algorithm With Lexical Chains For Automatic Text Summarization".,pp. 345-348.ICAART 2012.

[11]. Santiago Segarra, Mark Eisen and Alejandro Ribeiro. Authorship Attribution Through Function Word Adjacency Networks, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 63, NO. 20, OCTOBER 15, 2015.

[12]. A.K.Tripathy1, Nilakshi Joshi2, Steffy Thomas3, Shweta Shetty4, NamithaThomas5 VEDD- A Visual Wrapper for

Extraction of Data using DOM Tree 2012 International Conference on Communication

VII. BIOGRAPHY

Jyoti Shankarrao Pachare is a Research Assistant in the Computer Science Department, Wainganga College of Engineering, Nagpur University. She studies Master of Technology (Mtech) degree now from R.T.M.N.U. Nagpur, MS, and India. Her research interests are Computer Networks, Data Mining etc.