A Hybrid Neural Network Model for Unstructured Information Extraction using Page Rank

S.Nagarajan1, Dr. K.Perumal2
1Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India nagasethu2000@yahoo.com
2Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India perumalmkucs@gmail.com

Abstract

The world of technology has undergone some colossal changes in the past. The changes are evident in the scientific, engineering and societal context. People use instant messaging apps like WhatsApp and social media websites as Twitter that changed the way modern lifestyle is and their thought process. Nevertheless, a huge gap exists between computer based communication and human communication due to lack of reliable tools. A wide range of barriers still exist including language, dialect, culture, lifestyle and so on. At this point, leveraging conventional data mining through media mining and content extraction techniques becomes inevitable. In this paper, we put forth a new model titled “Unstructured Information Extraction fused Neural Network Model” (UIENNM). This innovative method will be used to make use of the principles of artificial neural network to extract data from unstructured data. The results displayed prove that the lexicon based approach has the lowest accuracy and F-measure when compared to other conventional approaches, whereas UIENNM had the highest accuracy and F-measure.

Keywords: Web Mining , Information Extraction, Page Rank, Data Mining ,Artificial Neural Network

1. Introduction

Big data in today’s business and technology environment has contributed to the complexity in accessing and retrieving relevant information for decision making [2]. The rapid growth of unstructured data is also creating a lot of problems. In 2008, Google was processing 20,000 terabytes of data (20 peta bytes) a day. For example, YouTube users uploaded 8 hours of new video every minute of the day and 1001 new websites are created every minute of the day. According to Twitter’s own research in early 2016, it sees roughly 200 million tweets every day, and has more than 525 million accounts. There are 100 terabytes of data uploaded daily to Facebook and 30 billion pieces of content shared on Facebook every month. The data production will be 44 times greater in 2020 than it was in 2016 [2]. Many information extraction algorithms have been invented in the researches of search engines [4] and most of them are focusing on a generic search in certain topic. There are also many other information retrieval tools developed for retrieving relevant information designed to search for online information that includes READWARE and ontology-based information retrieval. The aim of this paper is to propose a robust IE framework for extracting information from the unstructured data sources like web.

In order to retrieve information from the WWW[10], search engines with different capabilities and algorithms have been developed. However, the advancement of Internet made information available growth exponential through time and a robust framework for web information extraction and retrieval is critically required to process the overloaded unstructured data. Unstructured data is becoming more and more prevalent. Unstructured information is used in the representation of different media like text, video .etc. It facilitates the representation of information in the form of a tree. In this paper we propose a new model named Unstructured Information Extraction fused Neural Network Model (UIENNM). This novel method applies the principles of the artificial neural network for extracting the data from the unstructured data.

1.1 Introduction Of Neural Networks

An NN can formally be defined as: a massively parallel interconnected network of simple (usually adaptive) processing elements which is intended to interact with the objects of the real world in the same way as biological systems do. It has a device with many inputs and one output. ANN consists of a large number of simple processing elements that are interconnected with each other and layered within. The human body works with the help of neural network.

A neuron is a special biological cell that process information from one neuron to another neuron with the help of some electrical and chemical change. It is composed of a cell body or soma and two types of out reaching tree like
branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traits and plasma that holds the molecular equipments or producing material needed by the neurons.

Fig 1. Human Neurons

Fig 2. Artificial Neuron

Fig 3. Multilayered ANN

1.2. Information Retrieval (IR)

In web mining, there are four different types of tasks. We will focus only on Information Retrieval and Information Extraction. IR deals with the automatic search of relevant information contained in a set of knowledge, guaranteeing at the same time that non-relevant retrieved information is as less as possible. The aim must be to reach an improvement in retrieval results according to two key concepts in IR: recall and precision. Recall bears in mind the fact that the most relevant objects for the user must be retrieved. Precision takes into account that strange objects must be rejected.
1.3 Information Extraction

Information Extraction (IE) is a process that extracts and retrieves information that is relevant to user based on the queries posted.

A semi-structured document is a bridge between structured and unstructured data [8]. Unstructured data (also called flat data) is data that we know neither the context, nor the way information is fixed. It includes documents of mostly natural-language text, like word-processing files, e-mail, and text fields from databases or applications.

2. Literature Survey

In [8], Chen et al. have implemented several search methods in Java based on NNs and genetic search on databases, intranet, and internet. Mercure [10] is another IR system, based on multilayered networks, that allows document retrieval using a spreading activation process and query optimization using relevance back propagation. This model consists of an input layer, which represents users information needs, a term neuron layer, a document neuron layer, and an output layer representing the result of query evaluation. Lim [5] has developed a concept of visual keywords which are abstracted and extracted from visual documents using soft computing techniques. Each visual keyword is represented as an NN or a soft cluster center.

3. Proposed Unstructured Information Extraction Fused Neural Network Model (Uienmm)

The text data available on the web can be distinguished into three types: 1) unstructured data such as free text 2) semi structured data such as HTML 3) fully structured data such as in tables or databases. Mining is generally referred to as knowledge discovery in texts (KDT), is also known as text data mining or text mining. The idea behind this is to not analyze words or attribute values, but the core concepts are extracted. However, the content found in a text document presents no machine-readable semantic. To cope up with the need, some approaches suggest that restructuring the document content in a representation is a probable solution which is not recommended as they are prone to exploitation by machines. Techniques using lexicons for content interpretation are yet to reach mainstream level.

Hypertext mining is the process of mining semi-structured HTML pages which enclose hyperlinks, apart from text content. Supervised learning or classification has a crucial role in hypertext mining, as in email, newsgroup management, and maintaining web directories. Mining from services has gained significant importance due to exorbitant number of services like usenet, newsgroups, digital libraries, and mailing lists that came into being. There is a very thin line between web content mining and IR, as a group of people opine IR on the web as an instance of Webpage content mining(WCM), whereas another savant group connects WCM with intelligent IR. There are two different strategies to handle WCM: the first one directly mines the content of documents (web page content mining) whereas the other method improves the content search strategy by deploying it with other tools like search engines (search result mining). Hence, it is evident that task carried out by search engines is closely associated with WCM.
The general flow diagram for our proposed framework is shown below in Fig. 5.

![Diagram of Information Extraction in Context]

Fig 5. Distilling Structured Data from Unstructured Text

Usually the unstructured data includes textual documents, web pages[9], etc. Our proposed work starts its task extracting the relevant content from the web pages. With the large collection of retrieved documents, it is necessary to process them into a format suitable for presentation to the user. This involves sorting and ranking the documents according to their relevance to the user’s information needs. The conventional procedure is to rank the retrieved documents in descending order of relevance according to certain predetermined criteria.

The usual outcome of the ranking process applied by a search engine is a long list of document titles. The main drawback of such an approach is that the user is still required to browse through this long list to select those that are actually considered to be of interest.

Another shortcoming is that the resultant list of documents from a search engine does not make distinctions between the different concepts that may be present in the query, as the list inevitably has to be ranked sequentially. The problem lies mainly in the presentation of the list of document titles. These documents are usually listed serially irrespective of the similarity or dissimilarity in their contents—that is, it does not make distinctions between the different concepts. Thus, two documents appearing next to each other in the list may not necessarily be of a similar nature and vice versa. As the list of documents grows longer, the amount of time and effort needed to browse through the list to look for relevant documents increases.

In an attempt to avoid this problem, we proposed our novel model named Unstructured Information Extraction fused Neural Network Model (UIENNM) for extracting relevant content from the web pages with the help of page ranking algorithm. This page ranking algorithm ranks the web pages based on the rank score for the easy accessible of the data to the users. Each webpage is associated with the group of key words and phrases found in documents belonging to it. This enables the user to be selective in choosing particular group of documents to browse—thus making significant savings in browsing time and effort.

Page ranks are important since human beings find it difficult to scan through the entire list of documents returned by the search engine in response to his/her query. Therefore, it is desirable, for convenience, to get the pages ranked with respect to “relevance” to user queries so that one can get the desired documents only by scanning the first few pages. Note that it takes into consideration only the popularity of a page (reputation of incoming links) and richness of information content (number of outgoing links) and does not take care of other important factors like:

- User preference: Whether the link matches with the preferences of the user, established from his/her history?
- Validity: Whether the link is currently valid or not?
- Interestingness: Whether the page is of overall interest to the user or not?

These should also be reflected in the computation of page ranks. The learning/generalization capability of ANNs can be exploited for determining user preference and interestingness. User preference can be incorporated by training an NN-based on user history. Since ANNs can model nonlinear functions and learn from examples, they appear to be a good candidate for computing the “interestingness” of a page. Self-organizing ANNs can be used to filter out invalid pages dynamically. An ANN can compute the page rank from a combination of each of the parameters like hub, authority, reputation, validity, interestingness, and user preference with weights assigned to each which the user can modify; thereby refining the network as per his personalized interest. These factors may sometimes also be characterized by fuzzy variables. For example, variables like “close,” “far,” and “nearby” may be used to represent the distance between hits in a document for a given query.

ANNs can be used to learn the nonlinear user profiles from their previous history and reflect “relevance to user” of a document A in interest parameters. Richness and reputation is reflected in most existing page ranking systems. However, efficient computation of the page rank is an open research issue where ANNs may be used. Finally ANNs can be used to classify web pages as well as user patterns, in both supervised and unsupervised modes. Its ability in modeling complex nonlinear functions can also be exploited here. Another area where ANNs may be used is in building deductive capabilities in web mining systems. Nonlinear functions may be learned using ANNs and logical rules may be extracted from trained networks using rule extraction algorithms. The logical rules are human interpretable and help in generating deductions.
The general flow of the page rank with Neuro classifier is shown below in Fig. 6.

![Diagram](image)

**Fig 6. Page rank with neuro classifier**

As for ranking, traditional Web search engines rank Web pages based on authority and relevance score. The basic premise is that the top ranked pages (ideally the first page) contain sufficient information to satisfy the user’s information need. This paradigm is adequate for Sentiment Analysis and Opinion Mining for factual information search because one fact equals to any number of the same fact. That is, if the first page contains the required information, there is no need to see the rest of the relevant pages. The Page Rank algorithm is computed by weighting each in-link to a page proportionally to the quality of the page containing the in-link (Brin & Page, 1998). The qualities of these referring pages also are determined by PageRank.

Thus, the Page Rank of a page $p$ is calculated recursively as follows:

$$PageRank(p) = (1-d) + d \times \sum_{\text{all } q \text{ linking } p} \left( \frac{PageRank(q)}{c(q)} \right)$$

where $d$ is a damping factor between 0 and 1, $c(q)$ is the number of out-going links in a page $q$.

A Web page has a high Page Rank score if it is linked from many other pages, and the scores will be even higher if these referring pages are also good pages (pages that have high Page Rank scores). It is also interesting to note that the PageRank algorithm follows a random walk model the Page Rank of a page is proportional to the probability that a random surfer clicking on random links will arrive at that page.

### 4. Experimental Results And Discussion

In our proposed approach, four measures are compared with traditional approaches. The medical field is the domain for evaluation because many diverse users (including medical doctors, researchers, librarians and general public) seek important and high-quality information on health topics on the Web. It is also important for them to distinguish between Web pages of good and poor quality.

#### 4.1 Experimental Setup

Each of the four approaches are evaluated using cross-validation, a widely-used evaluation methodology for machine learning and text classification systems. A 50-fold cross validation was adopted, in which the 1000 documents in the data set were divided into 50 equal portions, with 20 documents each. Testing was performed for 50 iterations, in each of which 49 portions of the data (980 documents) were used for training and the remaining portion (20 documents) was used for testing.

We measured the effectiveness of each system using precision, recall, F-measure, and accuracy. Precision measures the fraction of the documents correctly classified as relevant, while recall measures the fraction of relevant documents retrieved from the data set. F-measure is a single measure that tries to combine precision and recall. Accuracy measures simply the prediction correctness of the classifiers. These measures are commonly used in text classification evaluation and have been adopted as follows:
We started with 20 documents in the first run, and increased the number of training documents by 20 in each subsequent run. There were thus 49 runs in total (from 20 to 980 training documents). In each run, a 50-fold cross validation similar to the one described above was used, and 20 documents were used for testing with rotation.

The experiment results on accuracy, precision, recall, and F-measure are displayed in Table 1 and summarized in the Fig. 7.

![Fig 7. Experimental Result Chart](image)

<table>
<thead>
<tr>
<th>Existing vs Proposed Model</th>
<th>Accuracy(%)</th>
<th>Precision(%)</th>
<th>Recall(%)</th>
<th>F- measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEXICON</td>
<td>82.83</td>
<td>65.40</td>
<td>62.36</td>
<td>0.6322</td>
</tr>
<tr>
<td>SVM- Word Extractor</td>
<td>84.57</td>
<td>84.57</td>
<td>61.39</td>
<td>0.6438</td>
</tr>
<tr>
<td>SVM- Webpage Extractor</td>
<td>87.30</td>
<td>85.35</td>
<td>61.56</td>
<td>0.5845</td>
</tr>
<tr>
<td>Proposed UIENNM (NN- Webpage)</td>
<td>92.45</td>
<td>91.78</td>
<td>73.68</td>
<td>0.7580</td>
</tr>
</tbody>
</table>

Compare to the existing models, the results of proposed approach demonstrated that it achieved the highest accuracy and F-measure.

5. Conclusion

The paper presented here describes how a Web-extraction approach can be used to Web page classification which is a combination of Web page content analysis and page ranking analysis. The new approach when compared with conventional text...
classification methods was found to be positive. We are highly confident that the method proposed can be used for wide variety of Web applications, specifically for vertical search engine development.

References