International Journal of Electronics Communications and Electrical Engineering
ISSN : 2277-7040 Volume 2 Issue 12 (December 2012)
http://www.ijecee.com/ https://sites.google.com/site/ijeceejournal/

Language detection for classification and content-based web pages filtering

Saman Bashbaghi, Abdol Hamid Pilevar
Computer Engineering Department
Bu Ali Sina University
Hamedan, Iran
{sbasbaghi, pilevar}@basu.ac.ir

Abstract. According to Daily increase of the documents on the internet, automatic language detection is getting more important. In this paper we used language detection system to classify and filtering of the immoral web pages, based on their contents. This system could detect 10 most used languages in the immoral web pages, including FARSI language. As a technique we introduce a new combined method which consists of three parts; URL Processor, page encoding processor, and text processor. In order to generate proper results this system has a voter which combines the results of these three parts. We used the immoral web pages and labeled web pages as an input data set in order to make a linguistic model for each language and system evaluation. Our experiments show 95% success in accuracy of outcome results.

Keywords: Text classification, automatic language detection, web page filtering, immoral web pages.

1. Introduction

Language recognition becomes more and more important as a result of increasing nature of web pages [9]. Automatic language detection defines as the allocation of one of the language as the label for document classification using computers. Language detection has various spectrum of application, such as search engines, optical character recognition system, information retrieval, and speech synthesis. For example search engines should detect web pages document languages, and query language, so that the documents could get retrieved in the associated language. Furthermore language detection is a preliminary state which is used in machine translation and stemmers [6]. Another example is text processing application. This application could enable the spelling error detection providing the detection of document language. And then reports all of the false words. Furthermore persons and libraries which work with multilingual languages, whilst they don't know the document language, can utilize the automatic detection systems accurately and reliably [8].

Regarding growth of immoral web pages over the internet, and its wide accessibility in the youth level of the society, webpage filtering seems to be very important. One of the methods used to detect these pages is utilization of its language. As a result of the filtering methods is automatic language detection. Another common method is using the terms in the webpage, so that we can detect the context language. Webpage address and encoding, moreover, could be used only in a way that they apply synchronously. That's because in this particular issue, it is possible that the name used in the address doesn't show the page immorality. Another reason is that, there could be many WebPages with different languages which used the same Encoding. Consequently, each method could not solve the problem by itself. It is declared in this paper that combination of these three methods has a very promising result.

The paper structure consists of: related works, problem definition, solution introduction, results interpretation, conclusion, and future works.

2. Related Works

Most of the automatic language detection systems are used on the English texts and some other in famous languages. As a result they are not applicable to Farsi. Many researchers have studied the language detection problem and they defined some detection types. One type uses the previous knowledge of the language, such as methods which are trained by use of language data. Their main deficiency occurs when the language we are using has a very low volume of data. Another type of methods performs based on linguistic information and because they are language dependant, they could not be used widely [6]. Different language detection methods usually work on the language model first. Then using similar method, form the training system linguistic model, and at the end they build a comparative algorithm which shows the input language is similar to one of the training system linguistic modes. These methods consist of unique term [2] combination that calculates frequent term in every language, and then using these frequent terms the language get detected.

Obviously, this method is so simple and for detecting languages we can use better terms combination. But the positive point of this method is the terms of language are short so we can detect the language of the text with observing a few terms.

For solving the problem of previous works, we use N-grams to consider many states of each language terms and we detect a language more precisely, obviously more time overhead in our processing. Another method uses stop words [3]. Stop words are terms that repeating frequently in the text. In this method the list of stop words are presented and because of they constitute many of texts of the language. Since frequent words such as delimiters, conjunction and prepositions are appropriate tools for language detecting, using them is proposed to language detection in this method. The main idea of this method is saving the stop words of each language in the database. Important feature of these words is they are short. Another method is using N-grams [4]. This method is used in many language detection researches and with regard to the value of N and with or without considering the range of words, has many types that each one according to the language has its advantages and disadvantages. For instance, if we don't consider the range of words, we don't need to tokenize the sentences anymore and after that the processing time is reduced but in other hand the accuracy is also reduced. In [5] the structure of web page is used to detect the language of the page. In this method according to terms and components which constitute the URL of web pages, the type of language is detected.

3. Problem Definition

In this article we implement a language detection system for content-based filtering of immoral web pages. In context of immoral websites on the internet, some research is done but they are insufficient. The statistics in [1] are presented at 2006 show in different countries many costs are spent for immoral websites. According to many surveys, 10 primary languages which are used in immoral web pages and we are going to investigate them in this paper, are: English, Chinese, Japanese, Korean, German, Russian, Spanish, Arabic, French and Farsi.

The importance of filtering these pages is: many abnormalities and problems may occur specially for children and teenager. This is why we need to implement the automatic language detection system. In this paper, Farsi is specially selected and studied.

4. Solution Introduction

We combine three subsystems consist of URL Processor, Encoding Processor and Text Processor in our method for solving the language detection problem. Each subsystem is presented with their details as follows:

4.1 Content-based language detection

We survey the different methods and algorithms which are related to this paper. In this section according to our study we are going to define a method for content detection.

Automatic language detection problem can be divided into two sub problems.

I. Creating language model of input text problem
II. Comparing created language model with language model of training set problem

Generally for solving language detection problem, first of all, a language model of input text with some criterions is created, then language model of training set with similar criterion and condition generated, and at the end a comparing algorithm that shows which language of training set is similar to the language model of input text is implemented. Since using list of stop words and immoral words that is used in the language model are not sufficient, therefore the address of pages and their encoding for language detection is used. For more accurate detection, this component and previous component are compared. The outcome of comparing the component with the voter of system results with language detection. The system inputs a web page with the XML format and assigns a number between 0 and 1 to each language which shows the probability of occurrence of that language. The highest assigned probability represents the language of the inputted web-page. At first we describe how to create the language model. In fact this step is performed as creating the language model of training set. Fig. 1 shows this step.

In this part that is accomplished locally offline, immoral web pages with HTML format which is produced by the existing crawler and their language are labeled and presented as input in the HTML processor unit. In this unit every additional information such as images and links are illuminated and its output as a XML file that consist of the page content are presented into XML Parser and separated into different parts. Then the XML file is presented as Language Model Creator component. In this component, we survey the words and texts which are used in this file to extract the language model. So our provided linguistic model consists of frequent words[1] list and also list of immoral words for each language.

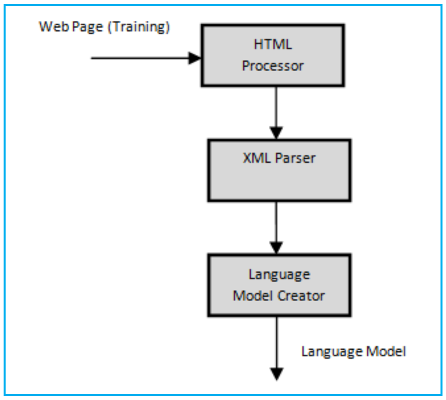

Fig. 1. Steps of creating linguistic model from training set

Now we describe our proposed method for language detecting. Fig. 2 shows components of the proposed system:

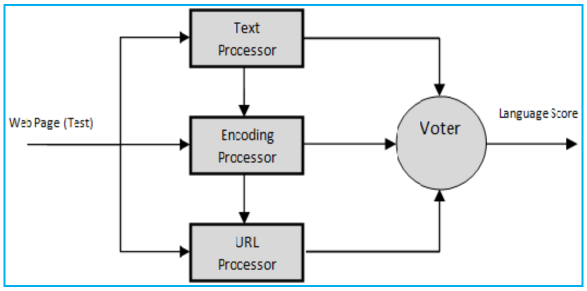

Fig. 2. Architecture of our proposed method

The input of system is web pages and we want recognize their languages. The outcome of system is the list of languages and their probability.

For detecting language in this system, we use the combination of three different methods which are explanted beside our voter system in the next sections.

4.1.1 Using page’s text (Text Processor)

Our experiments show that using the webpage text is a more accurate mechanism than language detecting methods. However this method is more time consuming. Therefore if higher speed in needed then It must be prevented.

In this section, first we build the linguistic model of input webpage as it is explained in the previous section. Then the language detection could be done, regarding comparison based on similarity and term matching of two models which consist of frequent corrupted terms.

4.1.2 Using page’s encoding (Encoding Process)

All web pages have a part, named header that consists of encoding. We can use this concept to detect language, because every language has its own encoding. So if there is any encoding, we can detect the language. In [7] all common used encoding are shown.

4.1.3 Using page’s URL (URL Process)

Every page has an address which is useful for language detection. Language detection based on URL has a lower load comparing to other methods. In this webpage language could be estimated because both URL and DOMAIN are useful for language detection. For example the domain can be used and if it is “.ru” the webpage can be selected as Russian webpage.

4.1.4 Voter

The last step of system performance is the combination of these three parts and final output generation. Using results of every part we understand that to generate the output based on voter and result combination, we assigned weight 2 to text processor, and weight 1 to two other parts. So text processor on the final stage has twice influence.

4.2 Sample Data

In this section, we explained sample data that used for training, test, and evaluation in this paper.

4.2.1 Immoral Website’s lists

A special crawler is used for downloading the WebPages , while they are selected and extracted randomly. We extracted about 100 million pages (10 blocks of 10 million pages). Regarding immoral webpage filtering in Iran, we extract filtered webpage list. We should note that this list also consists of political forbidden web pages, and also other forbidden web pages which their content is not immoral. But they are not many compare to immoral pages.

4.2.2 Using DMOZ's list

DMOZ is a classified list of webpage links building and updating manually by internet users voluntary. This website presents a tree structure in which web pages got classified based on different criteria and languages. In these web pages there is some mechanism for link evaluation which makes these web pages more valid. For example links available in the webpage gets check by the crawler in a steady manner. If the link is not valid, it gets eliminated from the list. In this list immoral web pages are extracted for each language, and utilized in linguistic model building phases.

5. Results

In this section we interpret the results of experiments. Since, acronyms of the languages are used; in the following graphs the listed acronyms of Table 1 are implemented .

Table 1. Names and acronyms of 10 implemented languages

Language Name	Acronym
English	en
French	fr
Arabic	ar
Korean	kr
Farsi	fa
Japanese	jp
Chinese	zh
Spanish	es
Russian	ru
German	de

In this section the accuracy of each part of system is evaluated. In Fig. 3, the proposed method is used for URL language detection.

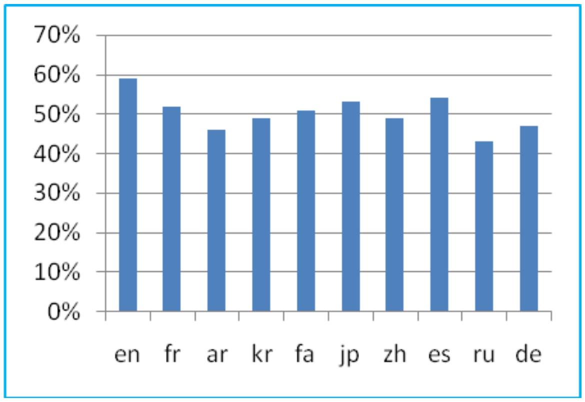

Fig. 3. Accuracy of method just for URL

Since data volume in web domain ".com" and ".net" are numerous, it is observed that the system accuracy based on URL is very low, and its accuracy is about 50.3%.

In Fig. 4 the accuracy of proposed method for encoding system is shown.

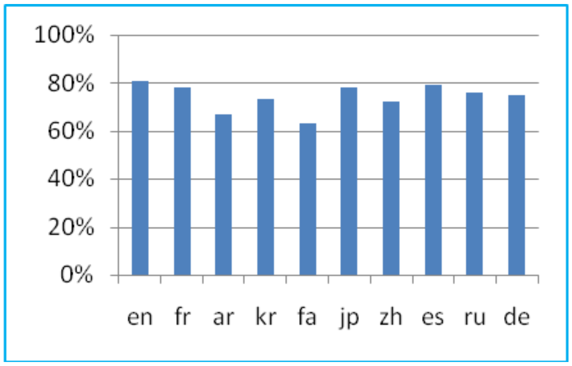

Fig. 4. Accuracy of method just for encoding

In this section the language detection method has 74.2 % accuracy. That's because: in most of the WebPages header encoding are not inserted and some of the languages using the same encoding.

Fig. 5. shows the system accuracy using page's text and its comparison with linguistic model.

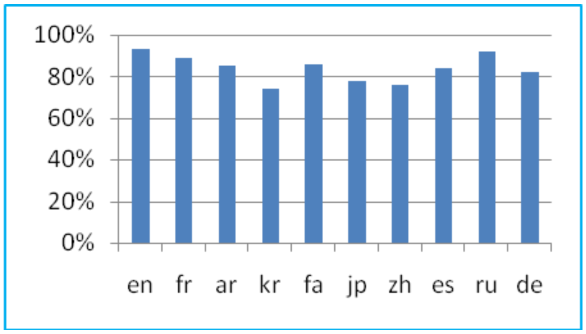

Fig. 5. Accuracy of method just for page's text

This system is 83.9 accurate, since the immoral data is insufficient. The other reason is that some words are ambiguous and has more than one meaning, and some of them are not immoral (e.g. breast, chicken breast, breast cancer).

This method is more accurate. But because of high operation volume, time consuming nature of linguistic model building, this method is very expensive.

It also has other disadvantages, such as the low probability of this kind of word in small text with less than 10 words.

This led to incorrect language classification. For languages such as Chinese that tokenization is not easy, this method is not working.

In this study, we need a good amount of knowledge from the language, when the input text is short , where most of the immoral WebPages consisting few words.

Fig. 6 shows the accuracy of method combined from these three methods.

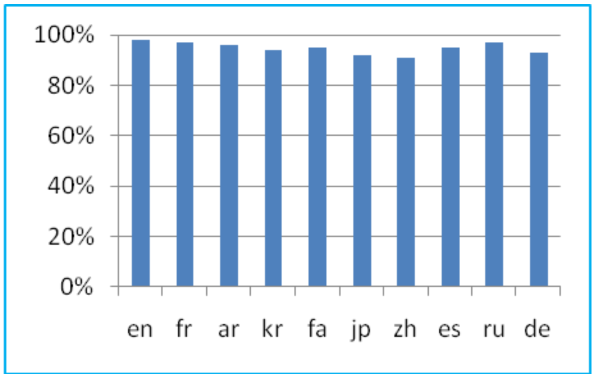

Fig. 6. Accuracy of method for combination of three methods

It is shown that the combined method has accuracy average of 94.8%.

6.Conclusion and future works

It is shown that the combination of three methods has a better performance in language detection compare to any of them separately.

For improving the system, a more precise model can be developed. For example, the common bi meaning words between medical and immoral texts could be used, and instead of distinguishing a word separately, we can translate it inside the text.

In the voter part of the proposed system, the neural network can improve the results and we are planned to do the same in our next research.

References

1. J. Ropelato, Internet Pornography Statistics, TopTenReviews, 2007.
2. G. Churcher, Distinctive character sequences, Personal communication, 1994.
3. G. Grefenstette‹"Comparing two language identification schemes" ‹In Proceedings of JADT 1995 ‹3rd International Conference on Statistical Analysis of Textual Data ‹1995.
4. W.B. Cavnar ‹J. M. Trenkle ‹"N-gram-based text categorization" ‹In Symposium on Document Analysis and Information Retrieval ‹Las Vegas ‹pp:161-175 ‹1994.

5. Eda Baykan, Monika Henzinger, Ingmar Weber, “Web page language identification based on URLs”, Proceedings of the VLDB Endowment, P.176-187, 2008.

6. Lena Grothe, Ernesto William De Luca and Andreas N¨urnberger, “A Comparative Study on Language Identification Methods”, Proceedings of 6th International Language Technologies Conference, P980-985, 2008.

7. http://www.w3.org/International/

8. Penelope Sibun, Jeffery C. Reynar, “Language Identification:Examining the Issue”, 5th Symposium on Document Analysis, 1996.

9. Lins, R. and Gonçalves, P.: Automatic Language Identification of Written Texts. Proc ACM SAC Symposium on Applied Computing, March 2004, Nicosia, Cyprus. 1128-1133.