Neurocomputing 177 (2016) 266–273

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A novel extreme learning fault diagnosis based supervision applied to mathematical formula contrastive analysis

Yuping Qin [a], Junnan Guo [b], Aihua Zhang [a,*]

[a] College of Engineering, Bohai University, Jinzhou, China
[b] College of Mathematics and Physics, Bohai University, Jinzhou, China

ARTICLE INFO

Article history:
Received 25 August 2015
Received in revised form
18 October 2015
Accepted 5 November 2015
Communicated by: Shen Yin
Available online 10 December 2015

Keywords:
Fault Diagnosis
Supervision
ELMS
MathML
Contrastive analysis

ABSTRACT

Focusing on the issue of few contrastive analysis technologies existing on mathematical formula in literatures, two methods of mathematical formula contrastive analysis subject to PDF and XML document formats are proposed. First, a novel extreme learning machine (ELM) fault diagnosis based supervision is presented and applied to mathematical formula contrastive analysis subject to PDF document. The strategy employs its merits such as fast learning speed, ease of implementation high efficiency and minimal human intervention, and not rely on the hidden neurons tuned. At the same time, the supervision mechanism is combined to realize the novel mathematical formula contrastive analysis based ELM supervision (ELMS). Second, a new mathematical formula contrastive analysis based MathML subject to XML document formats is presented. The idea first recognizes and extracts mathematical formulas in the detected XML document, normalizing the markup code of mathematical formulas. And then creating the tree presentation of mathematical the formula according to its presentation markup code, normalizing the tree structure by rule base, level traversing the tree to normalize the variable names and to get structure code of the tree. At last, searching the formula information table that named by the structure code. If the table exist, then searching the preorder traversal sequence of the tree in the table. If the records exist, then searching the postorder traversal sequence of the tree in the records. The searching result confirm if the mathematical formula belong to plagiarism. The experimental results reveal that the proposed algorithms can complete the contrastive analysis accurately of mathematical formula in XML documents whatever it existings PDF or XML format and has high performance on detection accuracy and speed.

1. Introduction

Fault diagnosis technology has been more mature and into various fields [1–4]. Moreover, every year, tens of thousands of papers would be published in various journals. Duplication detection or the same content analysis is always the main issue before they are published. Once the papers with same content are published in different journals, this will lead to plagiarism strife. Therefore, the plagiarism detection issues, such as duplication detection or the same content analysis, are inevitable. In this strong demand, lots of researchers put their eyes on plagiarism detection. Nowadays, contrastive analysis technology for text is relatively mature [5–9], and has been widely used in academic misconduct literature detection [10]. But it has not yet fully solved the problem of how to detect mathematical formulas in documents. However, with the rapid development of the Internet and digital library, the number of academic document, which contain mathematical formulas, is growing quickly, and the key ideas of many documents are often described by mathematical formulas. Therefore, how to effectively detect mathematical formulas in academic documents has attracted much attention, and has become a hot research field of information retrieval [11–16].

Gupta D. [17] proposed a distinguishing prerecession method based on Natural Language Processing (NLP) techniques in his recent academic research. This method primary focus is to detect intelligent plagiarism cases where semantics and linguistic variations play an important role. Shen Y. [18] proposed a fuzzy positivistic C-means clustering in his recent paper. Similar robust idea can be found in Refs. [19,20]. Some researchers discuss this method which could be put into use at contrastive analysis, however this idea has not employed deeply to this field. Anjali [21] proposed a novel dynamic analysis approach to software contrastive analysis. This method solves the issue about code

* Corresponding author.
E-mail address: jsxinxi_zah@163.com (A. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.11.027
0925-2312/

obfuscation techniques such as renaming of program entities, reordering of statements, etc. Bouarara [22] focuses on the issue of contrastive analysis in world of mail service, presents a novel automatic contrastive analysis approach. This strategy could realize the contrastive analysis with high efficient. In fact, fault diagnosis idea [23,24] is focused on in this field.

Moreover, machine learning methods are also payed attention to this field by researchers. Back propagation feedforward neural networks [25–27] are famous for its widely applied combining with amount of machine learning in the past decades. The feedforward neural networks with one hidden layer can realize the networks parameters adjusted with its own radial basis function. Although the feedforward neural networks have the ability to deal with the nonlinear mapping from the input space into the test models theoretically, it could not do it at all. In addition, the back propagation algorithm are proposed and put in solving to this issue. It can solve this problem via calculating the loss function gradient and updating the weights. However, the time consuming and overfitting issues are obvious. Support vector machine [28] is another common machine learning method and a promising way to estimate nonlinear system models accurately. Based on statistical learning theory and structural risk minimization principle, the support vector machine approach is capable of modeling nonlinear systems by transforming the regression problem into a convex quadratic programming problem and then solving it with a quadratic programming solver. Compared to conventional neural networks, support vector machine has the major advantages of global optimization and higher generalization capability [29]. Whatever the back propagation feedforward neural networks or the support vector machine methods, they all need to realize the precious machine learning results via adjusting parameters online. And all these are in theoretically. Then, extreme learning idea is proposed. Extreme learning machine[ELM] [30,31] is a whole different from above discussion for its distinguished characteristic, for example, more randomly hidden layers, fast responding ability, good generalization ability. In fact, whatever incremental ELM [32] or the other kinds of ELM has been discussed for a long time, and put in use in large fields. Here, we try to introduce ELM to a novel application field mathematical formula contrastive analysis. For the different exists in various forms of literatures, here another method based Mathematical Markup Language (MathML) is presented.

In general, three text types such as build via PDF, LATEX and eXtensible Markup Language (XML) are the main reference format. However, few contrastive analysis technologies focus on the XML document, especially on the mathematical formula contrastive analysis. XML is used to define documents with a standard format that can be read by any XML-compatible application. With the standardization of XML-based mathematical markup language, more and more XML documents that contain mathematical formulas are proliferating on the Internet. Here, we employ a novel ELMS fault diagnosis technology to realize the PDF mathematical formula contrastive analysis. Meanwhile, considering the Mathematical Markup Language (MathML) is an XML-based language for describing mathematical formula. It not only realizes the establishment and transmission of mathematical formula on the internet, but also realizes the reuse and conversion in other application programs. Another mathematical formula format, XML, based MathML is employed also. MathML provides two kinds of description markup, one is presentation markup, the other one is content markup. Both markups can completely describe any a mathematical formula. The presentation markup of a mathematical formula is encoded according to the symbol written order in the mathematical formula, the code are simple and intuitive. In addition, other markup code, such as content markup and OpenMath, can be converted into presentation markup [26]. Therefore, a mathematical formula contrastive analysis based on MathML is proposed. The methods of recognizing, extracting and matching mathematical formula is given, the effectiveness of the algorithm is verified by the experiments.

The rest of paper is organized as follows. Section 2 proposed a novel extreme learning for mathematical formula contrastive analysis based supervision strategy, Section 3 presented a MathML Mathematical Formula contrastive analysis method. Experimental results are presented in Section 4. Conclusion is outlined in Section 5.

2. A novel extreme learning method for mathematical formula contrastive analysis based supervision

2.1. Feature recognition and extraction of mathematical formula

Feature recognition and extraction is the first step of mathematical formula contrastive analysis. As we know that the mathematical formula symbol should include the normal 26 English letters (A–Z, a–z), special symbol $\left(+,-,\times,/,\sqrt{},x^n,\log_n\right)$, 10 Fig.s(0-9) and operator symbol $(\sum,\pm,\pi,\theta,\lambda\cdots)$. And here we let these to make up the feature database. If some effective features are selected during the feature extraction and have the same order list to the reference database, then the mathematical formula is plagiarized. To realize the precise feature extraction, we employ the quadratic feature extraction to complete this task. Quadratic feature extraction method includes two extraction parts: principal component analysis (PCA) and independent component analysis (ICA).

For demonstration purposes, PCA and ICA theory are introduced as follow.

2.1.1. PCA feature extraction

As the Refs. [33,34] saying that the PCA can extract the vital information from a mount of database by

$$\boldsymbol{x}_{n\_t} = \boldsymbol{Q}^T \boldsymbol{x}_t \tag{1}$$

Sometime $\boldsymbol{x}_{n\_t}$ is called principal components. In Eq. (1) $\boldsymbol{x}_t$ $(x_1, x_2, \cdots, x_n)^T$ denotes the database which the feature data should be extracted, and $\boldsymbol{Q}$ denotes the orthogonal matrix, whose ith columnq_i is the ith eigenvector of the covariance matrix defined by

$$C = \left(\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t^T\right)/n \tag{2}$$

2.1.2. ICA feature extraction

As the Ref. [35] saying that ICA is used to find independent sources, when observed data are mixtures of unknown sources and prior knowledge of the mixing mechanisms is not available, defined by

$$\boldsymbol{x}_t = \sum_{j=1}^{m} a_j s_j = \boldsymbol{A}\boldsymbol{s}_t \tag{3}$$

where $\boldsymbol{x}_t(x_1, x_2, \cdots, x_n)^T$ is the same defined as Eq.(1), $\boldsymbol{A}$ denotes a full rank matrix, and a_j denotes the jth column of $\boldsymbol{A}$, $\boldsymbol{s}_t = [s_1, s_2 \cdots, s_n]^T (m \geq n)$, named independent components(ICs), is random variables and mutually statistically independent with zero mean. In other words, the similar symbol can be directly observed via

$$\boldsymbol{s}_t = \boldsymbol{W}\boldsymbol{x}_t \tag{4}$$