Journal of Contemporary Management
Submitted on 25/10/2015
Article ID: 1929-0128-2016-01-27-09
Salha Alzahrani, Naomie Salim, and Vasile Palade

Framework for Plagiarism Detection Using Logical Tree-Structured Features and Multi-Layer Clustering

Dr. ***Salha Alzahrani*** (Corresponding Author)
College of Computers and IT, Taif University
Hawiah, 21944 Taif, Saudi Arabia
E-mail: s.zahrani@tu.edu.sa Homepage: www.c2learn.com

Prof. ***Naomie Salim***
Faculty of Computing, Universiti Teknologi Malaysia
Skudai, 81310 Johor, Malaysia
E-mail: naomie@utm.com Homepage: comp.utm.my/naomie/

Prof. ***Vasile Palade***
Faculty of Engineering and Computing, Coventry University
CV1 5FB Priory Street, Coventry, United Kingdom
E-mail: vasile.palade@coventry.ac.uk Homepage: www.cs.ox.ac.uk/vasile.palade/

***A*bstract:** Different practices of scientific misconduct have appeared recently and that impose the need for more sophisticated solutions. Logical tree-structured features describe the topology of scientific publications in terms of meaningful parts such as title, abstract, background, methods, results, and references. This paper presents the methodology proposed to uncover plagiarism in scientific publications using structural document features and multi-layer clustering. Logical tree-structured features are extracted as generic classes. Structural components such as paragraphs are organised under these generic classes. Instead of using traditional flat-based plagiarism detection methods, a layer-based clustering approach is proposed to find similar clusters and perform candidate retrieval using the top layer features. The bottom layer features are used to cluster structural components and to detect plagiarism. The suggested framework can be more efficient and reliable to detect plagiarism in scholarly articles than existing approaches.

Keywords: logical organisation, tree-structured features, clustering, plagiarism detection
JEL Classifications: C00, C82, C890

1. Introduction

The problem of plagiarism in the academic world has increased recently with the gigantic amount of digital resources and open access journals available on the Internet. Universities, publishers and individuals tend to use automatic plagiarism checkers to ensure the integrity of scholarly works. However, there are many ways to enhance the process of plagiarism detection in scientific publications in comparison with the current anti-plagiarism software. Scientific publications tend to have *consistent structure* with subsequent parts. Several studies on information extraction have addressed the *structure* of scientific publications (Burget, 2007; Hagen *et al.*, 2004; Lee *et al.*, 2003; Li and Ng, 2004; Wang *et al.*, 2005; Witt *et al.*, 2010; Zhang *et al.*, 2006). Segmentation of scholarly documents takes into consideration that the content structure is presented

by visual or physical elements, e.g. location, position, punctuations, length, font size or type, etc. They may also depend on some keywords, e.g. "chapter", "introduction", etc., to label a specific content. Several studies have defined the logical structure of the documents using different terminologies as *text-type structures* (Hagen *et al.*, 2004; Siddharthan and Teufel, 2007; Teufel and Moens, 2002), and *generic classes* in scholarly papers (Luong *et al.*, 2010). Different components extracted from the document can be generalised under these types/classes. For example, Tuefel *et al.* (2002) defined seven types of text, or argumentative zones according to a so-called rhetorical status, namely *Own, Other, Background, Textual, Aim, Basis, and contrast*.

Clustering is the process of grouping together objects or components that tend to have the same or similar features (Manning *et al.*, 2009). Each group of objects is called a cluster. Clustering differ from classification in the fact that we have no idea about the labels (i.e. name of features) in the resulting clusters. However in classification, we have a set of specific labels or categories that we want to assign each object to one of them.

Text clustering aims to discover documents, terms, passages, websites, or any textual elements which certainly share some textual similarly (Bhatia and Deogun, 1998; Manning *et al.*, 2009; Shehata *et al.*, 2010). The similarity perspective of texts can be defined in various ways. For instance, a "car" and a "horse" differ physically but similar in their functionality. Examples of text clustering include clustering of big data collections into smaller sub collections, term clustering to find shared themes or concepts in a data set, clustering of sentences from larger text objects about certain topic, clustering of websites and search results.

As an essential technique in text mining and knowledge discovery, text clustering is very useful for exploratory text analysis. Thus, it can be applied to detect plagiarism, or in other words, to get sense about highly similar textual elements and duplicates. This paper addresses the problem of plagiarism in the academic publications such as journal articles and conference papers. The contributions of this paper are twofold: (i) the use of logical tree-structured features for document segmentation and representation, and (ii) the use of clustering-based approach at different layers for plagiarism detection.

The rest of this paper is organised as follow. Section II discusses the literature review related to textual features and plagiarism detection techniques. Section III describes the logical tree-structured feature extraction method. Section IV describes the suggested algorithms for multi-layer clustering and plagiarism detection. Finally in Section V, we give concluding remarks and future works need to be done to complete the experimental works and accomplish this study.

2. Related Research

State-of-the-art research have addressed the textual data features and applied techniques for plagiarism detection (Alzahrani *et al.*, 2012b; Clough, 2000, 2003). In this section, textual data features are classified into two types: flat features and structural features. Tree-structured features of documents are described in depth. Then, we briefly summarise different plagiarism retrieval tasks and detection approaches. The relationship between the tree-structured features and clustering techniques is discussed to bridge the gap that remains highly problematic in the academic plagiarism.

2.1 Text features

Feature representation of textual documents can be classified into flat and structural features. Flat features refer to the lexical, syntax and semantic properties of the text without considering the orientation of these features throughout the document (Alzahrani *et al.*, 2012b). Examples of these features include character/word n-grams, phrases, sentences, part-of-speech (POS) tags, and others.

Structural features, on the other hand, represent the text as a *tree* with a root node and child nodes distributed in different layers (at least two layers). For example, a document (root node) can be divided into sections and sections into paragraphs (child nodes). Such representation exhibits better organisation of the scientific publications as they are highly structured. Structural features also represent better semantics of the content than the *flat* features. Structural feature extraction can be divided into *block-specific* and *content-specific* (Alzahrani *et al.*, 2012a).

Block-specific tree-structured feature representation refers to the use of specific markers such as tags or word counters to represent the tree regardless of the sections in the document. A three-level block-specific tree representation was extracted (Rahman *et al.*, 2007) as shown in Fig. 1 [a]. In (Chow and Rahman, 2009; Rahman and Chow, 2010), a hierarchical document organisation similar to (Rahman *et al.*, 2007) was used but with different dimensions for feature vectors as shown in Fig. 1 [b]. Nonetheless, block-specific features could be semantically insufficient to represent topically related content in the document. Therefore, extraction of *content-specific* tree-structured features would substantially improve the document representation. For example, scientific documents can be partitioned into *sections* and sections into *paragraphs* (Alzahrani *et al.*, 2012b). Tree representations such as *document-sections-paragraphs* or *document-concepts-chunks* would greatly characterise the semi-structured documents such as books, theses and journal articles and conference papers. However, some challenges are imposed in content-specific trees such as (i) sections have variable length in comparison to, for instance, fixed-length pages in block-specific trees, and (ii) different sections/concepts could have different degree of importance which can be exploited for different purposes such as improving the document retrieval and plagiarism detection.

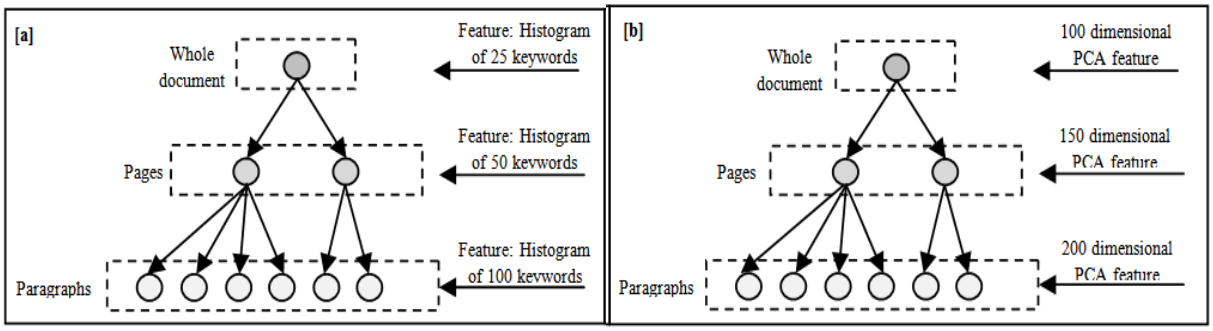

Fig. 1. Block-specific tree-structured feature representation of a document (Rahman and Chow, 2010)

2.2 Plagiarism detection

Several research works on plagiarism detection have investigated the development and evaluation of computerised techniques that address this offence. These techniques are generally working by scanning two textual documents, computing the degree of similarity, and highlighting highly similar segments as plagiarism. Most plagiarism detection techniques have utilised flat features to represent the textual data (Alzahrani *et al.*, 2012b). Few studies, on the other hand, used structural features for plagiarism detection. For example, a coarse-to-fine framework for plagiarism detection which implements *document-paragraphs-sentences* tree for a collection of web documents was proposed (Zhang and Chow, 2011). In this regard, matching sentences in the bottom layer obtained better precision in the plagiarism detection results compared with the approach in (Rahman *et al.*, 2007). Additionally, structural information has been investigated to detect significant plagiarism cases in scientific publications (Alzahrani *et al.*, 2012a).

MLSOM was used for retrieval of a set of similar documents to a suspected document and plagiarism detection (Chow and Rahman, 2009). The top layer performs document clustering and retrieval, and the bottom layer plays an important role for detecting similar, potentially plagiarised,

paragraphs. Given a query document d_q, a tree-structured document partitioning approach was firstly used to construct the tree *document-pages-paragraphs*. Secondly, feature vectors of the documents were constructed using a vocabulary table and PCA projection matrix, and used as input vector x_i. Thirdly, neurons in the upper level are matched with x_i to find the most similar neurons, i.e. documents, using Euclidean distance. A set of documents D_x is marked as having global similarity with d_q and used in the next step. Fourthly, the associated nodes of $d_x \in D_x$ in the bottom layer were compared in-depth with the third level nodes of d_q using a paragraph-to-paragraph similarity metric, and the most similar paragraph is the one with the smallest difference.

2.3 Bridging the gap

To sum up, textual features vary from simple lexical features to comprehensive structural features. Two documents having similar word-histograms at root nodes may be completely different in terms of the semantics and context. It is because of different orientation of the same set of words throughout the document, which is reflected by the discriminative lower parts of the tree data. Thus, tree structured representation can help to achieve better analysis of documents and plagiarism detection. Existing techniques applied for the problem of plagiarism detection do not consider *content-specific tree-structured features* and *multi-layer clustering*. In addition, the scope of the current methods (Chow and Rahman, 2009; Rahman *et al.*, 2007) that used block-specific features is limited to the literal plagiarism. This research work aims to bridge this gap by using *content-specific* tree-structured features representation better than the one used in (Chow and Rahman, 2009). For this aim, we propose the use of logical feature extraction from scientific documents and multi-layer clustering (i.e. the use of clustering at different layers). Clustering the root nodes will perform source document retrieval and clustering at the bottom letters will guide for in-depth analysis and plagiarism detection.

3. Logical Tree-Structured Document Model

Scientific publications have a common structure that begins with a title, authors, abstract, keywords, and the body which splits into several parts/components including headers, paragraphs, lists, tables, captions, quotes, references and so on. In contrast to the "bag-of-words"-based features used by existing methods (Barrón-Cedeño and Rosso, 2009; Grozea *et al.*, 2009; Kasprzak *et al.*, 2009; Lackes *et al.*, 2009), this work implements a feature extraction method that combines structural information and term information from scientific articles. Following sections discuss the segmentation process of scientific articles into structural components, the extraction of the logical tree-structured features, the weighting algorithm of structural components, and the construction of the vocabulary lists. A complete algorithm for the proposed tree-structured feature extraction method (TFEM) used in this study is presented in the last section to sum up the whole approach.

3.1 Component-based segmentation

One of the goals in this study is to capture the semantic organizational features of scientific publications. In this work, we proposed a tool and a method for structural components extraction based on the visual layout of the document and the raw text (Luong *et al.*, 2010). The tool works by extracting structural components using visual descriptors and keyword indicators. It can extract different constructs namely *Title, Author, Address, Affiliation, Keywords,* and *Body.* The body contains *Equations, Figures, Figure captions, Footnotes, List items, Notes, References, Section headers, Subsection headers, Sub-subsection headers, Tables, and Table captions*.

3.2 Logical tree-structured extraction

The use of the tree-structured feature representation facilitates the analysis of scientific articles in a *hierarchal,* rather than a *flat,* manner. As mentioned in Section II, *block-specific* tree-structured

features such as *document-pages-paragraphs* (Chow and Rahman, 2009) and *document-paragraphs-sentences* (Zhang and Chow, 2011) are not sufficient to represent the semantic organisation of scholarly documents. Therefore, we aim to employ *content-specific* tree-structured organisation wherein scientific articles are represented in a logical hierarchical tree namely

```
document -> generic classes -> structural components
```

By the word "generic classes", we mean a section or a group of sections that serve a unique purpose. We believe that classes convey more semantically related components than pages. To reflect the scientific topology in scholarly documents, we proposed the following generic classes:

```
G = "Title, Owner, Abstract, Introduction, Literature review,
Methodology, Evaluation, Conclusions, Acknowledgments and References"
```

3.3 Component-based weighting

A component weight $w_{C,d}$ for a structural component *C* in a document *d* can be defined as "a quantitative function which measure the weight of a structural component *C*, based on the relevance between terms in C and other structural components" (de Moura *et al.*, 2010). In this regard, $w_{C,d}$ defines a "qualitative" importance of a component *C* in scholarly documents, which can be assigned manually by an expert during the indexing phase of documents. Some methods have been developed (Bounhas and Slimani, 2010; de Moura *et al.*, 2010; Marques Pereira *et al.*, 2005; Marteau *et al.*, 2006) that use typical TF-IDF weighting but with structural components of documents taken into consideration. In this paper, we used the approach proposed in (Alzahrani *et al.*, 2012a) to compute $w_{C,d}$ automatically. Two statistical measures namely *Depth* and *Spread* (Alzahrani *et al.*, 2012a) are adapted, as below.

Spread of a term *t* in scholarly document *d* is the number of structural components in *d* that contain *t*:

$$Spread(t,d) = \sum_{C \in d} i \quad where \ i = \begin{cases} 1 & if \ t \in C \\ 0 & otherwise \end{cases} \tag{1}$$

Depth of a term *t* in a generic class *G* refers to the frequency of *t* in *G* normalized by the maximum frequency in *G* such that we do not underestimate classes with low components.

$$Depth(t,G) = \frac{tf_{t,G}}{MAX_{t',G}} \tag{2}$$

where $tf_{t,G}$ is the term frequency in generic class *G*, and $MAX_{t',G}$ is the maximum frequency gained by a term *t'* in *G*. *Spread*-based and *Depth*–based component-weight factors are defined at component level, as follows:

$$w_{C,d} = \frac{\sum_{t \in C} Spread(t,d)}{|C|} \tag{3}$$

$$w_{C,d} = \frac{\sum_{t \in C} Depth(t,G)}{|C|} \tag{4}$$

where *t* refers to index terms in a component *C*, *d* is the article that has *C*, |*C*| is the size of *C*. Finally, we combine *Depth* and *Spread* into a single factor.

$$w_{C,d} = \frac{\sum_{t \in C} Depth(t,G) \times Spread(t,d)}{|C|} \tag{5}$$

3.4 Vocabulary building

To build the vocabulary list, three steps need to be done. First is to construct the *term frequency table* which contains the terms and their occurrence information in structural components in each *d*, as follows:

$$tf'_{t,C} = tf_{t,C} \times w_{C,d} \quad (6)$$

$$tf'_{t,d} = \sum_{G \in d} \sum_{C \in G} tf'_{t,C} \quad (7)$$

where $tf_{t,C}$ is the frequency of a term t in a structural component C, $w_{C,d}$ is the combined component-weight factor given by formula (5), and $tf'_{t,C}$ *and* $tf'_{t,d}$ are the new frequency measure of terms in C and d combined with the structural information taken from the document.

We construct *term weighting table* using the frequency table in a way similar to VSM model, as follows:

$$w_{t,d} = tf'_{t,d}.log\frac{|D|}{|d \in D : t \in d|} \quad (8)$$

where $|D|$ is total number of documents in the dataset, and $|d \in D | t \in d|$ is the number of documents in the collection that contains t.

Then, the *vocabulary table T* is built which includes terms that obtain the top weights. For document features, we will consider 100 terms, while for generic classes G and structural components C, 150 and 200 top-frequency terms will be used, respectively.

3.5 Tree-structured feature extraction

The proposed algorithm for feature extraction is shown in Fig. 2. For all structural components C in each d, we will construct the feature vector f_C from term frequency computed in (6), as stated in formula (9). Then, the feature vectors for generic classes called f_G can be obtained as in equation (10).

$$f_C = [tf'_{t_1,C}, tf'_{t_2,C}, ..., tf'_{t_n,C}] \quad (9)$$

$$f_G = \sum f_C \mid C \in G \quad (10)$$

On the other hand, document feature vector f_d is constructed by using the weights computed in formula (8) as below.

$$f_d = [w_{t_1,d}, w_{t_2,d}, ..., w_{t_n,d}] \quad (11)$$

where n is the selected number of top terms to represent the feature vectors in each layer.

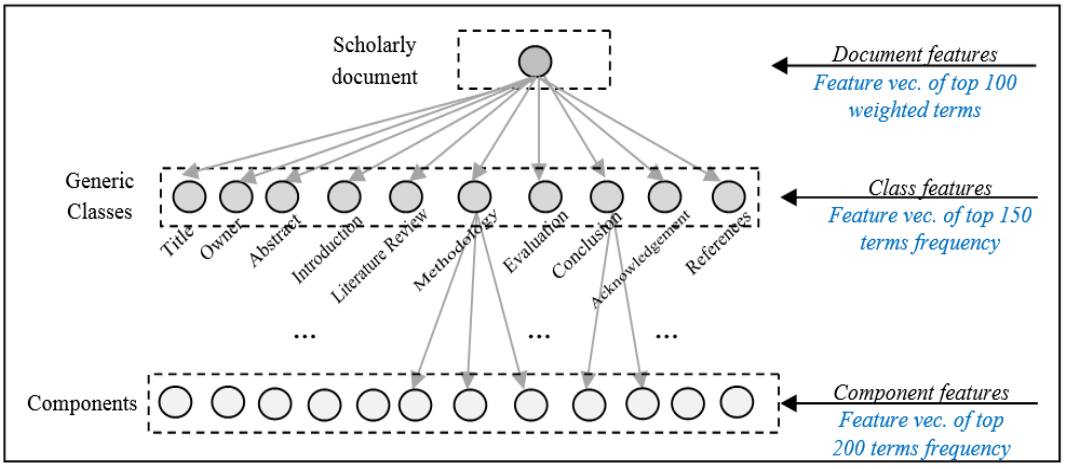

Fig. 2. Tree-structured feature extraction method (TFEM)

4. Multi-Layer Clustering and Plagiarism Detection

In plagiarism detection research, we deal with two sets of documents: source collection *D* and query documents *Q*. In this study, both sets are represented as *content-specific tree-structured* features. The proposed framework includes three main steps:

- *Step 1: Clustering at the top layer*. Clustering is performed at the top layer based on document features f_d. The aim of this step is to find a subset of the document collection $D_x c D \; \forall \; d_q \in D_q$ which is relatively smaller than D.
- *Step 2: Clustering at the middle layer*. For each query document d_q, we just used the set of relatively similar documents D_x obtained from step 1. Then, clustering on D_x is performed at the middle layer based on generic class features f_G. The aim of this step is to find similar sections or subjects between documents (i.e. generic classes), and mark them for further analysis.
- *Step 3: Clustering at the bottom layer and plagiarism detection*. This step aims to find all suspicious components C_q in $d_q | d_q \in D_q$ which are plagiarised from structural components C_x in $d_x | d_x \in D_x$ using *structural component-based comparison algorithm* explained below.

Clustering in the top and middle layers can be achieved using general text clustering techniques such as generative probabilistic models, agglomerative hierarchical clustering (Bhatia and Deogun, 1998), and *K*-means clustering algorithms (Manning *et al.*, 2009). Then, to find the cluster that is most likely to contain the set of source documents, we will use the cosine the similarity between the centre of each cluster d_j and the query d_q can be calculated as follows:

$$Sim(d_j, d_q) = \frac{d_j \cdot d_q}{\| d_j \| \cdot \| d_q \|} = \frac{\sum_{i=1}^{n} w_{t_i,d_j} w_{t_i,d_q}}{\sqrt{\sum_{i=1}^{n} w^2_{t_i,d_j}} \sqrt{\sum_{i=1}^{n} w^2_{t_i,d_q}}} \qquad (12)$$

In the last step, detailed analysis and similarity calculation are performed to find the structural components that are highly similar. Further analysis by humans may designate plagiarised components from properly cited ones. To this end, associated nodes of $d_x \in D_x$ in the bottom layer will be compared component-to-component with the feature vectors of third layer of d_q. By components we generally mean paragraphs. The similarity between the feature vectors of structural components can be calculated using vector difference. The most similar paragraph is the one with the smallest difference as stated by the equation below.

$$PD(d_q, d_x) = \forall C_q \in d_q (\min_{C_x \in d_x} | f_{C_q} - f_{C_x} |) \qquad (13)$$

where f_C are the paragraph features for documents d_q and d_x.

5. Conclusion and Future Work

Plagiarism in scientific publications is addressed in this paper. We proposed a rough-to-fine framework for feature extraction namely logical content-specific tree-structured features wherein structural components are organised under generic classes. Clustering is suggested at different layers to achieve document retrieval and plagiarism detection. The suggested methods and algorithms exhibit better understanding of the semantic content and exploratory analysis of scientific publications. Future works include the construction of a ground-truth dataset of scientific documents taking into account accurate XML tree representation. Experimental works should be performed on the dataset to evaluate the proposed framework. More in-depth analysis on structural components should be performed and information visualization methods can be used for highlighting plagiarism in a way that is different from other types of documents.

References

[1] Alzahrani, S., *et al.* (2012a). "Using structural information and citation evidence to detect significant plagiarism cases". *Journal of the American Society for Information Science and Technology (JASIST), 63*(2): 286-312.

[2] Alzahrani, S. M., Salim, N., and Abraham, A. (2012b). "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods". *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 42*(2): 133 - 149.

[3] Barrón-Cedeño, A., and Rosso, P. (2009). "On automatic plagiarism detection based on n-grams comparison". *Advances in Information Retrieval* (pp. 696-700). DOI: 10.1007/978-3-642-00958-7_69.

[4] Bhatia, S. K., and Deogun, J. S. (1998). "Conceptual clustering in information retrieval". *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 28*(3): 427-436.

[5] Bounhas, I., and Slimani, Y. (2010). "A hierarchical approach for semi-structured document indexing and terminology extraction". Paper presented at the International Conference on Information Retrieval and Knowledge Management, CAMP'10, Selangor, Malaysia.

[6] Burget, R. (2007). "Automatic Document Structure Detection for Data Integration". In: W. Abramowicz (Ed.), *Business Information Systems* (Vol. 4439, pp. 391-397): Springer Berlin / Heidelberg.

[7] Chow, T. W. S., and Rahman, M. K. M. (2009). "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection". *IEEE Transactions on Neural Networks, 20*(9): 1385-1402.

[8] Clough, P. (2000). "Plagiarism in Natural and Programming Languages: An Overview of Current Tools and Technologies", In: Department of Computer Science, University of Sheffield, UK, Technical Report CS-00-05.

[9] Clough, P. (2003). "Old and new challenges in automatic plagiarism detection". *National UK Plagiarism Advisory Service* [Online] Available at http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf.

[10] de Moura, E. S., *et al.* (2010). "Using structural information to improve search in Web collections". *Journal of the American Society for Information Science and Technology, 61*(12): 2503-2513. DOI: 10.1002/asi.21436.

[11] Grozea, C., Gehl, C., and Popescu, M. (2009). "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection". Paper presented at the 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09, Donostia, Spain.

[12] Hagen, L., Harald, L., and Petra Saskia, B. (2004). "Text type structure and logical document structure". Paper presented at the ACL Workshop on Discourse Annotation, Barcelona, Spain.

[13] Kasprzak, J., Brandejs, M., and Křipač, M. (2009). "Finding Plagiarism by Evaluating Document Similarities". Paper presented at the 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09, Donostia, Spain.

[14] Lackes, R., Bartels, J., Berndt, E., and Frank, E. (2009). "A word-frequency based method for detecting plagiarism in documents". Paper presented at the International Conference on Information Reuse and Integration, IRI'09, Las Vegas, NV.

[15] Lee, K. H., Choy, Y. C., and Cho, S. B. (2003). "Logical structure analysis and generation for structured documents: A syntactic approach". *IEEE Transactions on Knowledge and Data Engineering, 15*(5): 1277-1294.

[16] Li, Z., and Ng, W. K. (2004). "WICCAP: From semi-structured data to structured data". Paper presented at the 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, ECBS'04, Brno, Czech Republic.

[17] Luong, M.-T., Nguyen, T. D., and Kan, M.-Y. (2010). "Logical structure recovery in scholarly articles with rich document features". *International Journal of Digital Library Systems (IJDLS), 1*(4): 1-23.

[18] Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Flat Clustering Introduction to Information Retrieval* (pp. 350-374): Cambridge University Press.

[19] Marques Pereira, R. A., Molinari, A., and Pasi, G. (2005). "Contextual weighted representations and indexing models for the retrieval of HTML documents". *Soft Computing, 9*(7): 481-492.

[20] Marteau, P.-F., Ménier, G., and Popovici, E. (2006). "Weighted Naïve Bayes model for semi-structured document categorization". Paper presented at the 1st International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, Merida, Espagne.

[21] Rahman, M. K. M., and Chow, T. W. S. (2010). "Content-based hierarchical document organization using multi-layer hybrid network and tree-structured features". *Expert Systems with Applications, 37*(4): 2874-2881.

[22] Rahman, M. K. M., WangPi Yang, Tommy W.S. Chow, and Sitao Wu (2007). "A flexible multi-layer self-organizing map for generic processing of tree-structured data". *Pattern Recognition, 40*(5): 1406-1424.

[23] Shehata, S., Karray, F., and Kamel, M. (2010). "An efficient concept-based mining model for enhancing text clustering". *IEEE Transactions on Knowledge and Data Engineering, 22*(10): 1360-1371.

[24] Siddharthan, A., and Teufel, S. (2007). "Whose idea was this, and why does it matter? Attributing scientific work to citations". Paper presented at the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2007). New York, USA.

[25] Teufel, S., and Moens, M. (2002). "Summarizing scientific articles: Experiments with relevance and rhetorical status". *Computational Linguistics, 28*(4): 409-445.

[26] Wang, Z.Q., Wang,Y.C., and Gao, K. (2005). *A New Model of Document Structure Analysis, Fuzzy Systems and Knowledge Discovery* (Vol. 3614, pp. 658-666): Springer Berlin, Heidelberg.

[27] Witt, A. and Metzing, D. (2010). "Discourse Relations and Document Structure". In: N. Ide, J. Véronis, H. Baayen, K. W. Church, J. Klavans, D. T. Barnard, D. Tufis, J. Llisterri, S. Johansson & J. Mariani (Eds.), *Linguistic Modeling of Information and Markup Languages* (Vol. 40, pp. 97-123): Springer Netherlands.

[28] Zhang, H., and Chow, T. W. S. (2011). "A coarse-to-fine framework to efficiently thwart plagiarism". *Pattern Recognition, 44*(2): 471-487.

[29] Zhang, K., Wu, G., and Li, J. (2006). "Logical structure based semantic relationship extraction from semi-structured documents". Paper presented at the 15th International Conference on World Wide Web, Edinburgh, Scotland.