Hindawi
Journal of Applied Mathematics
Volume 2021, Article ID 6662984, 10 pages
https://doi.org/10.1155/2021/6662984

Research Article

Document Plagiarism Detection Using a New Concept Similarity in Formal Concept Analysis

Jirapond Muangprathub, Siriwan Kajornkasirat, and Apirat Wanichsombat

Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand

Correspondence should be addressed to Apirat Wanichsombat; apirat.w@psu.ac.th

Received 3 October 2020; Revised 24 February 2021; Accepted 13 March 2021; Published 24 March 2021

Academic Editor: Md Sazzad Hossien Chowdhury

This paper proposes an algorithm for document plagiarism detection using the provided incremental knowledge construction with formal concept analysis (FCA). The incremental knowledge construction is presented to support document matching between the source document in storage and the suspect document. Thus, a new concept similarity measure is also proposed for retrieving formal concepts in the knowledge construction. The presented concept similarity employs appearance frequencies in the obtained knowledge construction. Our approach can be applied to retrieve relevant information because the obtained structure uses FCA in concept form that is definable by a conjunction of properties. This measure is mathematically proven to be a formal similarity metric. The performance of the proposed similarity measure is demonstrated in document plagiarism detection. Moreover, this paper provides an algorithm to build the information structure for document plagiarism detection. Thai text test collections are used for performance evaluation of the implemented web application.

1. Introduction

Recently, plagiarism has increased because of easy access to data on the World Wide Web. For this reason, producing a written document can be easy and quick [1–4]. However, plagiarism or copying in a different style is a problem in education, research, publications, and other contexts. Software for detecting such problems has been mostly developed based on text string comparisons [1, 2, 5, 6]. Grouping prior documents based on their similarity has been demonstrated to reduce the search time. Formal concept analysis (FCA) is widely used to identify groups of objects sharing common attributes [7–10]. This work focuses on using FCA to group documents.

FCA is a popular approach for knowledge representation and data analysis in many applications and has become popular in information and document retrieval [11–15]. Document retrieval is used to retrieve the plagiarism candidate documents, so the suspect document is the query, while the stored source documents are retrieved [2, 16, 17]. FCA is one approach for grouping documents in a hierarchy that supports browsing. It automatically provides generalization and specialization relationships among the formal concepts for documents represented in a concept lattice [10–18]. Thus, this work applied FCA to detect document plagiarism. Moreover, this method provides related documents or groups of documents to the user. However, the application requires a similarity measure to retrieve source documents or to identify groups of similar documents in a concept hierarchy. Thus, the choice of the concept similarity measure is a challenging problem for identifying different concepts that are semantically close.

Many measures have been proposed for concept similarity based on set theory in binary weighting form (e.g., [19–25]). However, the weights determined from all content can be used to improve the precision of concept retrieval. Formica [12, 13, 26] used information in this manner. Later, concept similarity measures were developed with flexibility to adapt to user preferences (e.g., [24, 25, 27, 28]). The previous studies mostly used only intensions, instead of both

intensions and extensions of the formal concept. In addition, the appearance frequency of formal concepts could be useful for improved concept retrieval.

Thus, this paper proposes such similarity measures for formal concepts that use the above ideas. Later, mathematical proof is provided that a formal similarity metric has been defined. Concept similarity of FCA has gained importance from its application to plagiarism detection, which has to assess the similarity between formal concepts to find relevant information. We present and investigate a candidate algorithm to support plagiarism detection with the proposed concept similarity measures. Finally, plagiarism detection test cases are evaluated from collected Thai text data.

This report is organized as follows. Section 2 provides details of the formal concept analysis, Section 3 discusses related prior work, Section 4 presents the research methodology and the proposed system, Section 5 has results and discussion, and Section 6 is the conclusion.

2. Formal Concept Analysis

Formal concept analysis (FCA) is applied in many fields for data analysis and knowledge representation [9, 10, 14]. In this section, the basic definitions are presented to understand the useful notions of FCA, following the book [29]. We start with a formal context with the following definitions.

A formal context $\mathbb{K} = (G, M, I)$ consists of two sets G and M and a relation I between G and M. The elements of G are called the objects, and the elements of M are called the attributes of the context. $(g, m) \in I$ can be used to express that an object g is in a relation I with an attribute m and read as "the object g has the attribute m," also denoted by gIm. For a set $A \subseteq G$ of objects, A' is defined as follows $A' := \{m \in M \mid gI\, m \text{ for all } g \in A\}$. Correspondingly, for a set $B \subseteq M$ of attributes, B' is defined as follows $B' := \{g \in G \mid gIm \text{ for all } m \in B\}$. A formal concept of the formal context (G, M, I) is a pair (A, B) with $A \subseteq G, B \subseteq M, A' = B$, and $B' = A$. We call A the extent and B the intent of the formal concept (A, B). $\mathfrak{B}(G, M, I)$ denotes the set of all formal concepts of the formal context (G, M, I).

The above definitions show a group of documents and their shared keywords. Practically, these definitions are presented to identify groups of source documents sharing common keywords. FCA is, as such, applicable to a formal context which contain only binary values, 0 or 1. However, a typical database will hold collected data not restricted to only binary values. Since the database holds a finite set of objects and their attributes, the set of attribute values is also finite: this is called a many-valued context. A many-valued context (G, M, W, I) consists of a set of G, M, and W and a ternary relation I between G, M, and W (i.e., $I \subseteq G \times M \times W$) for which it holds that

$$(g, m, w) \in I \text{ and } (g, m, v) \in I \text{ always imply } w = v. \quad (1)$$

The elements of G are called objects, those of M (many-valued) attributes and those of W attribute values. If W has n distinct elements, it is called an n-valued context. The condition in the above definition states that there is at most one attribute value given for an object and an attribute, so we can again have an information matrix with single entries for object rows and attribute columns. We read $(g, m, w) \in I$ as "the attribute m has the value w" for the object g and can write $m(g) = w$ or $(g, m, w) \in I$. To obtain a concept lattice from a many-valued context, it has to be transformed to a formal context. The transformation can be done with *conceptual scales*. In practice, each many-valued attribute is represented by a collection of binary attributes.

A scale for the attribute m of a many-valued context is a (one-valued) context $Sm := (Gm, Mm, Im)$ with $m(G) \subseteq Gm$. The objects of a scale are called *scale values*, and the attributes are called scale attributes.

The scales of each context are joined to make a one-valued context (formal context), for which the simplest method is called plain scaling. In plain scaling, the derived formal context is obtained from a many-valued context (G, M, W, I) and the scale contexts $Sm, m \in M$ where the attribute set of Sm is replaced by $Mm := m \times Mm$. Thus, the new formal context (G, N, J) is derived from a many-valued context by plain scaling with this formal transformation

$$N := \bigcup_{m \in M} M_m, \quad (2)$$

and $gJ(m, n) :\iff m(g) = w \text{ and } wImn$.

We will later use this plain scaling approach to transform a many-valued context into a formal context so that FCA becomes applicable. Afterwards, the formal concept form is generated to obtain the knowledge structure, and to use this knowledge, the concept similarity measure is applied. Many similarity measures have been developed for use in retrieving formal concepts, surveyed in [30]. Lengnink [31] proposed similarity measures by using the averages of fractional overlaps of objects and attributes, relative to all objects (attributes) in the concepts compared as a local measure and relative to all objects (attributes) available overall as a global measure. Saquer and Deogun [22] and Dau et al. [32, 33] applied rough sets for concept approximation to improve the information retrieval with guiding query refinement. They use symmetric differences between the objects (attributes), similarly as the previous definitions using overlap. Therefore, we actually have distance measures instead of similarity measures, with zero distance given for identical concepts (instead of similarity one). Next, Formica [12, 13] improved the concept similarity by using the attribute intent with the attribute extent. The researchers used approximate extensions for information content. Moreover, they inserted weight parameters that can be adjusted by the user to tune the method. In addition, Wang and Liu [24] applied rough sets to evaluate a formal concept of the interval between upper neighbors and lower neighbors. However, only intension of formal concepts was applied on determining their similarity based on Tversky's model [27] instead of both intension and extension [34, 35]. Alqadah [19] applied set theory using intension in formal concepts to propose similarity measures. In summary, the challenge of concept similarity measures can be considered for intent and extent attributes.

However, the previous works have not considered merging intent and extent attributes. The studies have used only intent or only extent attributes to compute the similarity weights. Inspired by these reviewed similarity measures or indices, this work focuses on using intent and extent in all formal concepts.

3. Related Works

Plagiarism detection was divided into two approaches, namely, external and intrinsic plagiarism detection. External plagiarism detection involves identification of the source documents by using a database, while intrinsic plagiarism detection is not available to the text, but it is a plagiarized text converted with the use of synonyms. This paper is aimed at detecting external plagiarism by using a database. Computer-assisted plagiarism detection is an information retrieval (IR) task supported by specialized IR systems, which are referred to as plagiarism detection systems or document similarity detection systems. Thus, this section presents a plagiarism detection system based on IR and FCA for IR, as applied in this work.

Many studies have applied semantic role labeling (SRL) [1, 3, 35–37]. Abdi et al. [1] present an external plagiarism detection system that employs a combination of SRL with semantic and syntactic information. The semantic role labeling technique is here used to handle active to passive and vice versa transformations. The proposed method is able to detect different types of plagiarism, such as exact verbatim copying, paraphrasing, transformation of sentences, or changing word structure. Osman et al. [35, 36] applied SRL to extract arguments from the sentences and then compare arguments to detect the plagiarized part from the text. Paul and Jamal [37] also improved SRL for the ranking of sentences to identify direct copy-paste, active-passive transformation, and synonym conversions with faster execution times. Moreover, machine learning of both supervised and unsupervised types has been applied to detect document plagiarism [16, 38]. Vani and Gupta [3, 38, 39] studied and compared different methods of document categorization for external plagiarism detection. They applied the K-means algorithm and the general N-gram. The K-means gave promising results when dealing with highly obfuscated data. Rahman and Chow [16] proposed a new document representation to enhance the classification accuracy using a new hybrid neural network model to handle the document representation. They represent the document in a tree structure that has a superior ability to encode document characteristics.

The IR was applied to enhance the performance of plagiarism detection [4, 39–42]. Ekbal et al. [40] propose a technique based on textual similarity for external plagiarism detection by using a vector space model, which is one technique in IR to compare source and suspect documents. The results show encouraging performance with a benchmark setup, but not with language translation. Ahuja et al. [4] use the Dice measure as a similarity measure for finding the semantic resemblances between pairs of sentences. It also uses linguistic features like path similarity, a depth estimation measure, to compute the resemblance between pairs of words, and these features are combined by assigning different weights to them. It is capable of identifying cases of restructuring, paraphrasing, verbatim copying, and synonymized plagiarism. Moreover, the vector space model was applied in [39, 41, 42] to improve recall performance. These studies represent suspected and source documents as vectors using VSM and TF-ISF weighting. However, this work's conceptual IR systems are aimed at addressing the limitations of the classical keyword systems and identifying the conceptual associations and links between the documents. Thus, FCA can be used to fulfil an IR system in order to obtain the document relationship. Hierarchical order visualization of formal concepts in the concept lattice structure is an important concern for practical applications of FCA [43]. In addition, Kumar et al. [44] discussed the use of FCA for results in LSI and SVM. The authors applied FCA to discover dependencies in the data for clustering documents [45–47].

IR is concerned with selecting appropriate information from an information collection. Traditionally, the process is begun by submitting a query, matching the query with information collection, seeing the ranked information, and submitting a newly revised query, until the target information is found or the user quits [48]. FCA has been successfully applied to enhance the efficiency and effectiveness of each task in this process. Mostly, an information collection is analysed with FCA to form a hierarchy, and retrieving information from such structures requires suitable methods. Next, we briefly review interesting work on FCA for IR. A retrieval task is composed of three natural subtasks: query, matching, and ranking. The matching is based on a similarity measure of the kind that was reviewed in the previous section. Query refinement allows users to recover from situations where the returned solution set is too large or too small. By the use of related keywords (attributes) in a concept lattice, the retrieval process performed on the initial query can also retrieve further relevant keywords. For this reason, concept lattice techniques have been developed for query refinement to improve web search engines (e.g., [20, 21, 28, 43, 44, 48, 49]). Nafkha et al. [15] applied FCA to retrieve solutions by using the cooccurrence of documents inside formal concepts. Qadi et al. [11] applied FCA in both refining the query and in ranking the solutions. An ontology for image processing was used in this retrieval process. They ranked the solutions by counting the number of documents in the retrieved concepts.

4. The Proposed Document Plagiarism Detection Approach

4.1. System Overview. The document plagiarism detection using FCA is aimed at detecting good matches between the source document in storage and a suspect document. In this section, we discuss the proposed system shown in Figure 1. The source documents will be subjected to text operations such as word segmentation and stopwords to extract keywords. We applied the Thai segmentation library [50] to obtain keywords (or words in general) from the source documents. That set of extracted words is represented with the attributes of the formal context, and the source documents provide the objects of the formal context. Afterwards, the

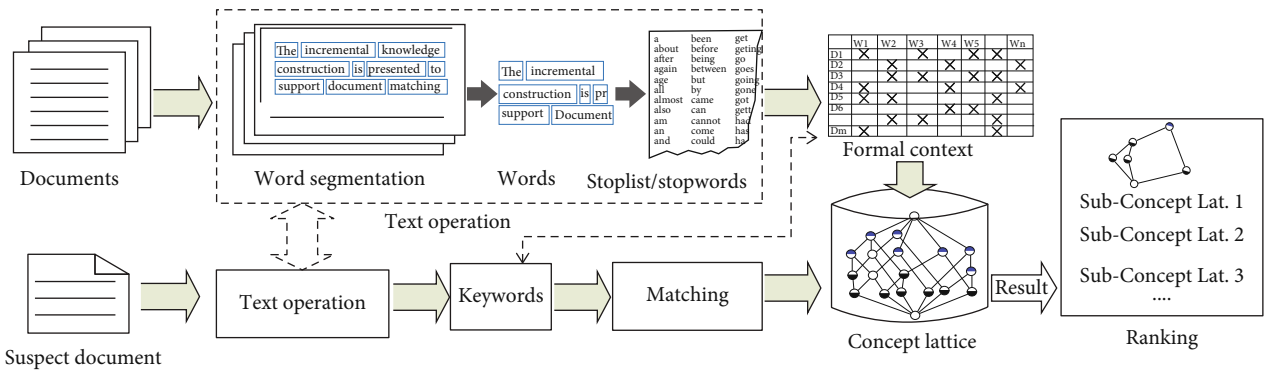

Figure 1: The proposed system overview.

formal context will be processed into a concept lattice to retrieve the relevant documents in document plagiarism detection. Likewise, the suspect document will be subjected to text operations to match and retrieve from the concept lattice.

The concepts in the lattice are used to index source documents. This structure is incrementally and automatically rebuilt when new cases are added or existing cases are updated. The new source documents are collected to prepare data with text operations. Next, the keywords are rebuilt with a new concept as a new node in the lattice structure. To initially find a new node and its position simultaneously for the updated concept lattice, we applied the algorithm in [18] to insert into concept lattice according to its position, in a scalable knowledge structure. This is used to retrieve a similarity concept from the suspected document in the subconcept lattice form by using the new concept similarity proposed in the next section.

4.2. A New Concept Similarity. We propose two concept similarity measures that not only are applicable within a concept lattice but also give similarity values between any existing formal concept and a tentative concept formed from available objects and attributes, which need not be an element of the lattice. These similarities use both object extent and attribute intent based on their appearance frequencies in the concept of the lattice. The proposed method allows ranking by the similarity values.

Both new similarity measures are introduced based on extension and intension. The first measure weighs the objects and attributes equally. The second weighs them based on existing concepts in the lattice. In this section, we define the building blocks used in our approach.

We define C_p as a formal concept represented by a pair (E_p, I_p) in the formal concepts $\mathfrak{B}(G, M, I)$, where $E_p \subseteq G$, $I_p \subseteq M$, E_p, I_p are the extent and intent of formal concept, respectively. A new formal concept is defined as $C_N = (E_N, I_N)$, where E_N is a set of the retrieved object(s) and I_N is a set of new attributes provided by the suspect document. Thus, a new concept similarity measure between formal concepts in $\mathfrak{B}(G, M, I)$ and new formal concept is defined as $\mathrm{sim}(C_N, C_p)$. The proposed concept similarity measures are based on an appearance frequency of formal concepts denoted with $f(C_N, C_p)$ according to (3) and (4). The closer $\mathrm{sim}(C_N, C_p)$ is to 1, the greater the similarity of C_N and C_p.

Given a formal concept $C_p = (E_p, I_p)$ and a new formal concept $C_N = (E_N, I_N)$ in a formal context (G, M, I), concept similarity equally weighting objects and attributes is defined as

$$\mathrm{sim}(C_N, C_p) = \frac{1}{2}\left(\frac{f(C_N, C_p)_{\mathrm{meet}}}{f(C_N, C_p)_{\mathrm{join}}} + \frac{|I_N \cap I_p|}{|I_N \cup I_p|}\right). \quad (3)$$

When the objects are used to weigh existing attributes, the concept similarity is defined as

$$\mathrm{sim}(C_N, C_p) = \frac{f(C_N, C_p)_{\mathrm{meet}}}{f(C_N, C_p)_{\mathrm{join}}} * \frac{|I_N \cap I_p|}{|I_N \cup I_p|}, \quad (4)$$

where $f(C_N, C_p)_{\mathrm{meet}}$ is the frequency of objects in a formal concept $\mathfrak{B}(G, M, I)$, $I_N \cap I_p \neq \varnothing$ and $f(C_N, C_p)_{\mathrm{join}}$ is the total number of objects in formal concept $\mathfrak{B}(G, M, I)$.

In equations (3) and (4), the frequency of objects is applied because the concept lattice is derived from the formal concept. If the number of formal concepts is high, this shows that it is general knowledge, and it shows in the upper concept lattice. Thus, $f(C_N, C_p)_{\mathrm{meet}}$ and $f(C_N, C_p)_{\mathrm{join}}$ are applied in this work. We apply these similarity measures for document plagiarism detection and provide mathematical proof of having a formal similarity metric in Theorem 1 [18].

Theorem 1. *$sim(C_N, C_p)$ is the degree of similarity between the formal concepts C_p and the formal concept C_N in concept lattice $\mathfrak{B}(G, M, I)$ if $sim(C_N, C_p)$ satisfies the following conditions [15]:*

(1) $0 \leq sim \leq 1$

(2) $sim(C_N, C_p) = 1$ if $C_N = C_p$

(3) $sim(C_N, C_p) = sim(C_p, C_N)$

(4) $sim(C_N, C_O) \leq sim(C_N, C_p)$ and $sim(C_N, C_O) \leq sim(C_p, C_O)$ if $C_N \subseteq C_p \subseteq C_O$, $C_O \in \mathfrak{B}(G, M, I)$

The proposed similarity measures are applied in the system for document plagiarism detection. Normally, the similarity measure between source document (D_i) and a suspect document (Q) is defined as $\mathrm{sim}(Q, D_i)$.

Let D_i and Q be a set of keywords (wd) where D_i is defined as $D_i = \{wd_{d1}, wd_{d2}, \cdots, wd_{dn}\}$ where n is the total number of keywords of source document i. Similarly, $Q = \{wd_1, wd_2, \cdots, wd_m\}$, where m is the total number of keywords of a suspect document. Given a formal concept $C_p = (E_p, I_p)$ and a new formal concept $C_N = (E_N, I_N)$ in a formal context $\mathbb{K} \coloneqq (G, M, I)$, for any element D_i in EP and a suspect document Q is a new formal concept where I_N represent set of keywords of a suspect document. From definitions (3) and (4), we can apply to document plagiarism detection as follows:

$$\mathrm{sim}(Q, D_i) = \frac{1}{2}\left(\frac{f(Q, D_i)_{\mathrm{meet}}}{f(Q, D_i)_{\mathrm{join}}} + \frac{|Q \cap D_i|}{|Q \cup D_i|}\right), \tag{5}$$

$$\mathrm{sim}(Q, D_i) = \frac{f(Q, D_i)_{\mathrm{meet}}}{f(Q, D_i)_{\mathrm{join}}} * \frac{|I_N \cap I_p|}{|I_N \cup I_p|}, \tag{6}$$

where for any source document D_i and any suspect document Q, we define $f(Q, D_i)_{\mathrm{meet}}$ = the frequency of source document D_i in a formal concept $\mathfrak{B}(G, M, I)$, where $Q \cap D_i \neq \varnothing$, and $f(Q, D_i)_{\mathrm{join}}$ = the total number of formal concepts which contain source document D_i = the frequency of source document D_i in a formal concept $\mathfrak{B}(G, M, I)$.

From equations (5) and (6), we get the following equation:

$$\mathrm{sim}\left(C_N, C_p\right) = \max\left\{\mathrm{sim}(Q, D_i) | D_i \in E_p\right\}. \tag{7}$$

The proposed concept similarity measure for document plagiarism detection is mathematically proved to be a formal similarity metric following Theorem 1. Namely, our concept similarity measure is the degree of similarity according to Theorem 2.

Theorem 2. *$sim(C_N, C_p) = \max\{sim(Q, D_i) | D_i \in E_p\}$ is the degree of similarity between the formal concepts C_p and the formal concept C_N.*

Proof.

(1) This work will prove that $0 \leq \mathrm{sim}(C_N, C_p) \leq 1$. To prove this, we first consider that any D_i is a source document in formal concept C_p and $Q = I_N$,

$$0 \leq \mathrm{sim}(Q, D_i) \leq 1. \tag{8}$$

From definition of $f(Q, D_i)_{\mathrm{meet}}$ and $f(Q, D_i)_{\mathrm{join}}$, it is obvious that

$$0 \leq f(\mathrm{Q}, D_i)_{\mathrm{meet}} \leq f(Q, D_i)_{\mathrm{join}}. \tag{9}$$

Then, we have $|Q \cap D_i|/|Q \cup D_i| \leq 1$, where $Q \cup D_i \neq \varnothing$. Now, we get the following result:

$$0 = \frac{1}{2}(0 + 0) \neq \mathrm{sim}(Q, D_i) = \frac{1}{2}\left(\frac{f(Q, D_i)_{\mathrm{meet}}}{f(Q, D_i)_{\mathrm{join}}} + \frac{|Q \cap D_i|}{|Q \cup D_i|}\right) \leq \frac{1}{2}(1 + 1) = 1. \tag{10}$$

From equation (14), we also get that $0 \leq \mathrm{sim}(Q, D_i) \leq 1$.

(2) Let $C_N = C_p$; now, we have $E_N = E_p$ and $I_N = I_p$. Since the suspect document Q is the I_N, $Q = I_p$.This implies that $Q \cap D_i \neq \varnothing$ for all documents D_i in formal concept C_p. Hence, we get by definitions of $f(Q, D_i)_{\mathrm{meet}}$ and $f(Q, D_i)_{\mathrm{join}}$ that $f(Q, D_i)_{\mathrm{meet}} = f(Q, D_i)_{\mathrm{join}}$. Thus, $f(Q, D_i)_{\mathrm{meet}}/f(Q, D_i)_{\mathrm{join}} = 1$. We consider only the case of $D_i = Q$. We get that $|Q \cap D_i|/|Q \cup D_i| = 1$. Thus,

$$\mathrm{sim}(Q, D_i) = \frac{1}{2}\left(\frac{f(Q, D_i)_{\mathrm{meet}}}{f(Q, D_i)_{\mathrm{join}}} + \frac{|Q \cap D_i|}{|Q \cup D_i|}\right) = \frac{1}{2}(1 + 1) = 1. \tag{11}$$

Thus, $\mathrm{sim}(C_N, C_p) = 1$.

(3) Since C_N is a new formal concept which needs to be assigned similarity with the given formal concept C_p, it is obvious that $\mathrm{sim}(C_N, C_p) = \mathrm{sim}(C_p, C_N)$

(4) Suppose $C_N \subseteq C_p \subseteq C_O$, we have $E_N \subseteq E_p \subseteq E_O$ and $I_N \subseteq I_p \subseteq I_O$. Firstly, we show that $\mathrm{sim}(C_N, C_p) = \mathrm{sim}(C_N, C_O)$. It is clear by careful inspection that for any source document D_k in $E_O \setminus E_N$, we get that $f(I_N, D_k)_{\mathrm{meet}} = 0$ and $|I_N \cap D_k| = 0$. Now, we have

$$\frac{f(I_N, D_k)_{\mathrm{meet}}}{f(I_N, D_k)_{\mathrm{join}}} + \frac{|I_N \cap D_k|}{|I_N \cup D_k|} < \frac{f(I_N, D_i)_{\mathrm{meet}}}{f(I_N, D_i)_{\mathrm{join}}} + \frac{|I_N \cap D_i|}{|I_N \cup D_i|}. \tag{12}$$

Hence, we can conclude that for any D_k in $E_O \setminus E_N$, $\mathrm{sim}(I_N, D_k)$ is not maximal in $\{\mathrm{sim}(Q, D_i) | D_i \in E_O\}$. This implies

$$\begin{aligned}\mathrm{sim}(C_N, C_O) &= \max\{\mathrm{sim}(Q, D_i) | D_i \in E_O\} \\ &= \max\{\mathrm{sim}(Q, D_i) | D_i \in E_N\} \\ &= \mathrm{sim}(C_N, C_N).\end{aligned} \tag{13}$$

Similarly, we have $\mathrm{sim}(C_N, C_p) = \mathrm{sim}(C_N, C_N)$. Hence, we get $\mathrm{sim}(C_N, C_p) = \mathrm{sim}(C_N, C_O)$. Next, we show that $\mathrm{sim}(C_N, C_O) \leq \mathrm{sim}(C_p, C_O)$. It is obvious that $f(I_N, D_i)_{\mathrm{join}} = f(I_p, D_i)_{\mathrm{join}}$ for all D_i in E_O because both of them are total

Input: Source document collection in formal context.
Output: Set of the formal concepts $\mathfrak{B}(G, M, I)$.
Method:
1. $SetExt = FindExent(M)$ //*find initial set of extent*
2. $SetInt = \varnothing$
3. *For* $i = 0$ *to* $|SetExt|$
4. { $SetInt[i] = SetExt[i]'$
5. $SetConcept[i] = (SetExt[i], SetInt[i])$
6.}
7. *Return (all formal concept SetConcept)*
Function $FindExtent(M)$
1. *For* $i = 0$ *to* $|M|$ // *find initial set of extent*
2. $SetExtInitial[i] = m[i]'$
3. *For* $j = 0$ *to* $|SetExtInitial[i]| - 1$ // *find the set of extent*
4. *For* $k = j$ *to* $|SetExtInitial[i]| - 1$
5. {
6. $IntersecExt = SetExtInitial[i] \cap SetExtinitial[j]$
7. *If* $(IntersecExt \notin SetExtInitial[i])$
8. $SetExt = \bigcup IntersecExt$
9. }
10. *Return SetExt*

Algorithm 1: Building knowledge base of document plagiarism detection.

Input: Set of formal concepts $\mathfrak{B}(G, M, I)$.
: Source document collection in formal context,
: A suspect document (Q)
Output: Set of the closest source document(s).
Method:
1. For i=0 to |G|
2. { $f(Q, D_i)_{meet}[i] = FindExtInt(SetConcept, Query)$
3. $f(Q, D_i)_{join}[i] = FindExtInt(SetConcept, g[i])$
4. $sim1(Q, D_i) = 1/2((f(Q, D_i)_{meet}/f(Q, D_i)_{join}) + (|Q \cap D_i|/|Q \cup D_i|))$,
5. $sim2(Q, D_i) = f(Q, D_i)_{meet}/f(Q, D_i)_{join} * |I_N \cap I_p|/|I_N \cup I_p|$,
6. }
7. Return set of maximum of sim1 and sim2.
Function $FindExtInt(SetConcept, QorG)$
1. For i=0 to $|SetConcept|$
2. If $QorG \cap SetInt[i] \neq 0$
3. feqDoc =feqDoc+1
4. Return feqDoc;

Algorithm 2: Retrieving the source documents with the proposed similarity.

numbers of formal concepts which contain source document D_i. Since $I_N \subseteq I_p$, then $I_N \cap D_i \subseteq I_p \cap D_i$. This implies that if $I_N \cap D_i$ is empty, then $I_p \cap D_i$ may be not empty. So, we get that $f(I_N, D_i)_{\text{meet}} \leq f(I_p, D_i)_{\text{meet}}$. This leads to

$$\frac{f(I_N, D_i)_{\text{meet}}}{f(I_N, D_i)_{\text{join}}} \leq \frac{f(I_p, D_i)_{\text{meet}}}{f(I_p, D_i)_{\text{join}}}. \tag{14}$$

It is clear by careful inspection that for any source document D_i in E_O, $I_O \subseteq D_i$. So, we get that $I_N \subseteq I_p \subseteq I_O \subseteq D_i$ for all D_i in E_O. This implies that if $I_N \subseteq I_p \subseteq I_O$, then $|I_N \cup D_i| = |I_p \cup D_i| = |D_i|$. Now, we have

$$\frac{|I_N \cap D_i|}{|I_N \cup D_i|} \leq \frac{|I_p \cap D_i|}{|I_p \cup D_i|}. \tag{15}$$

By (14) and (15), we get

$$\frac{f(I_N, D_i)_{\text{meet}}}{f(I_N, D_i)_{\text{join}}} + \frac{|I_N \cap D_i|}{|I_N \cup D_i|} \leq \frac{f(I_p, D_i)_{\text{meet}}}{f(I_p, D_i)_{\text{join}}} + \frac{|I_p \cap D_i|}{|I_p \cup D_i|}. \tag{16}$$

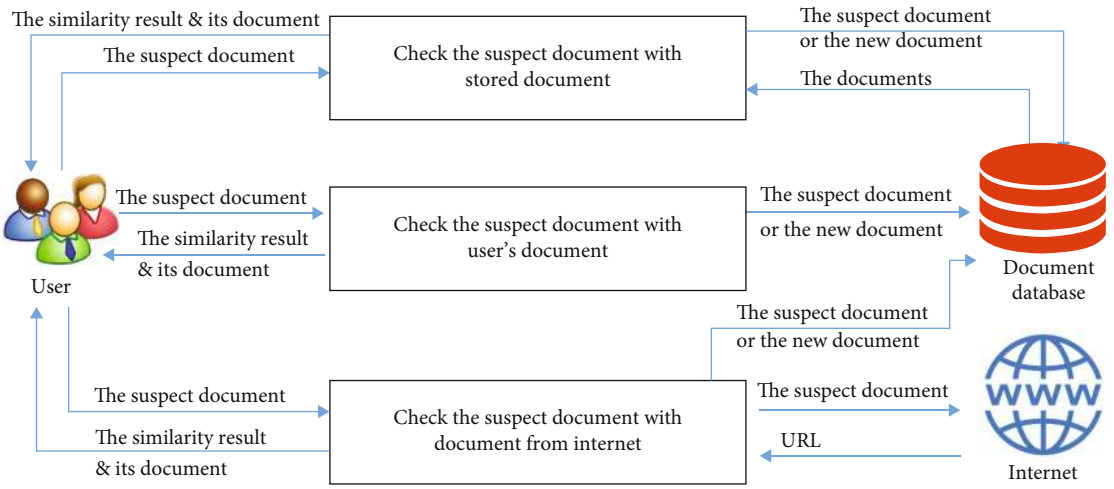

Figure 2: The workflow for the user.

Figure 3: An example of the proposed application.

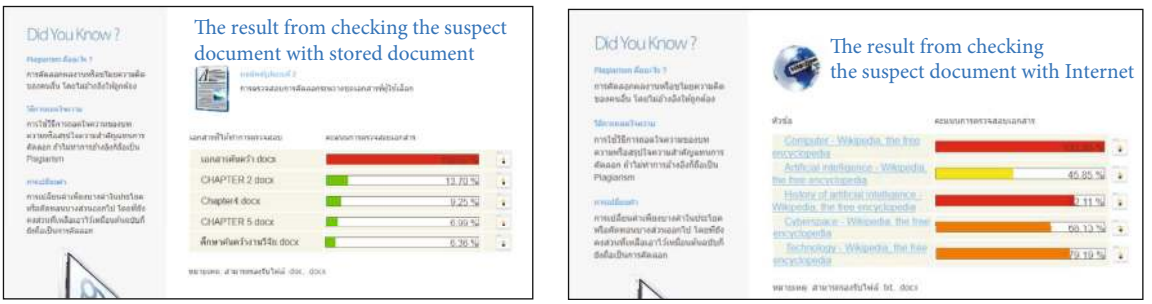

Figure 4: An example of displaying document similarities.

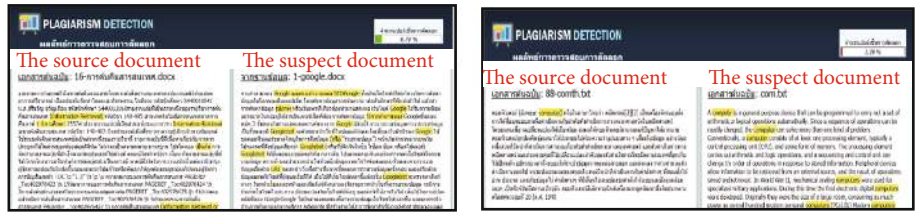

Figure 5: An example of comparing source and suspect documents.

Table 1: The experimental results of the average document plagiarism for each file.

Level of copy text	% of the result of text copy					Average	Plagiarism accuracy of the proposed system
	File1	File2	File3	File4	File5		
100%	100%	100%	100%	100%	100%	100%	100.00%
80%	78%	68%	73%	79%	75%	75%	93.73%
50%	45%	51%	47%	58%	57%	52%	96.00%
30%	28%	35%	37%	38%	36%	35%	83.33%
0%	3%	2%	5%	2%	4%	3%	97.00%
The overall plagiarism accuracy							94.01%

Consider the following result:

$$\begin{aligned}\operatorname{sim}(C_N, C_O) &= \max\{\operatorname{sim}(Q, D_i) | D_i \in E_O\} \\ &= \max\{\operatorname{sim}(I_N, D_i) | D_i \in E_O\} \\ &= \max\left\{\frac{1}{2}\left(\frac{f(I_N, D_i)_{\text{meet}}}{f(I_N, D_i)_{\text{join}}} + \frac{|I_N \cap D_i|}{|I_N \cup D_i|}\right) | D_i \in E_O\right\} \\ &= \max\left\{\frac{1}{2}\left(\frac{f(I_p, D_i)_{\text{meet}}}{f(I_p, D_i)_{\text{join}}} + \frac{|I_p \cap D_i|}{|I_p \cup D_i|}\right) | D_i \in E_O\right\} \\ &= \max\{\operatorname{sim}(I_p, D_i) \mid D_i \in E_O\} = \operatorname{sim}(C_p, C_O).\end{aligned} \tag{17}$$

Similarly, we can prove $\operatorname{sim}(Q, D_i)$ from equation (7) with the above 1-4.

In summary, the proposed concept similarity can be applied to retrieve source documents from knowledge storage in the concept lattice form, and this is demonstrated both empirically and theoretically in the next section.

4.3. Algorithm for Building Knowledge Base and Performance Evaluation. In this section, we evaluate the implemented system for document plagiarism detection using the proposed algorithm and use it to retrieve source documents. We provide an algorithm for building a knowledge base in formal concept form and retrieve source documents when the user inputs a suspect document. Algorithm 1 generates a set of formal concepts that consist of two parts, i.e., extent and intent. The result from this algorithm is used as a knowledge base for retrieving source documents when the user inputs a suspect document. Algorithm 2 next matches the suspect document within the set of all formal concepts to retrieve a group of source document(s) relevant to the suspect document, represented by the retrieved formal concept.

5. Implementation and Results

The proposed system was implemented with web applications as shown in Figure 2. This workflow demonstrates the process by the user. Firstly, the user has three ways to input the suspect document, namely, stored document, user's documents, or document from the internet. Next, the system provides a document database to support the comparison between the suspect document and prior source documents, using the FCA module mentioned in Section 4. This module is enabled in the back end of the web application. If the user would like to check with their documents, they select the provided option to compare the document similarity. Moreover, the user can check their suspect document with a document from the internet, for which they will get a URL (Uniform Resource Locator) for results on the suspect document on a website. Finally, the suspect document will be stored in the database to check in the future.

We developed our system as an online website. PHP language was used to implement the system, while a MySQL database is used for details of the documents. An example of the application is provided in Figure 3. After the user selects various options, the result will show the document similarity, for example, Figure 4. If the user would like to see the details of plagiarism, they can click the provided link in Figure 5.

In this work, we designed an experiment to evaluate provided document plagiarism. We provided documents with copied text to various extents, namely, with 100%, 80%, 50%, 30%, and 0% of copying. Each level of copying was designed with 5 general text files derived from news or academic publications, with different sizes of 200 kB, 400 kB, 800 kB, 1200 kB, or 1600 kB. These files were tested for the operation of the proposed approach, 10 times for each file. The results are shown in Table 1.

Table 1 shows the performance of the proposed system with an overall plagiarism detection accuracy of 94.01%.

Each level of copied text shows similarity between the source file and the provided suspect files (or documents). If the suspect document is completely copied, the proposed method will detect 100% of plagiarism. However, even if no copying occurred, the system still detects a few percent of plagiarism, because some frequent words have appeared.

6. Conclusions

This paper proposed an algorithm for detecting document plagiarism by using formal concept analysis (FCA) with the presented concept similarity candidate to retrieve relevant source documents. The proposed similarity measures employ concept approximation using frequency of the formal concepts and were mathematically proven to be formal similarity metrics. The source documents were processed and retrieved with the proposed algorithm to demonstrate performance of the proposed similarity measure in document plagiarism detection by implemented web applications. This work proposes 3 formats to prevent plagiarism: (1) to detect among documents inside the document collection, (2) to detect between the suspect document and source documents, and (3) to detect between the suspect document and other documents from the Internet. The proposed 3 formats in the system were implemented in PHP language with the MySQL database. Moreover, in the last format, the presented system applies services by Google. The proposed system was demonstrated to be efficient and effective with a case study of news and academic documents. The experiments were evaluated from two aspects: efficiency tests by type of document and an effectiveness test regarding correctness. The results show that (1) the proposed system can detect document types .docx, .pdf, and .txt as designed and (2) the proposed system can detect plagiarized documents with an average accuracy of 94.01%.

Data Availability

There are no data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Abdi, S. M. Shamsuddin, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, "A linguistic treatment for automatic external plagiarism detection," *Knowledge-Based Systems*, vol. 135, pp. 135–146, 2017.

[2] J. P. Bao, J. Y. Shen, X. D. Liu, and Q. B. Song, "A survey on natural language text copy detection," *Journal of software*, vol. 14, no. 10, pp. 1753–1760, 2003.

[3] K. Vani and D. Gupta, "Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges," *Information Processing & Management*, vol. 54, no. 3, pp. 408–432, 2018.

[4] L. Ahuja, V. Gupta, and R. Kumar, "A new hybrid technique for detection of plagiarism from text documents," *Arabian Journal for Science and Engineering*, vol. 45, pp. 1–14, 2020.

[5] S. Rao, P. Gupta, K. Singhal, and P. Majumder, "External & intrinsic plagiarism detection: VSM & discourse markers based approach," *Notebook for PAN at CLEF*, vol. 63, pp. 2–6, 2011.

[6] M. Sahi and V. Gupta, "A novel technique for detecting plagiarism in documents exploiting information sources," *Cognitive Computation*, vol. 9, no. 6, pp. 852–867, 2017.

[7] B. Ganter and R. Wille, "Applied lattice theory: formal concept analysis," in *In General Lattice Theory*, G. Grätzer, Ed., Birkhäuser, 1997.

[8] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer Science & Business Media, 2012.

[9] R. Wille, "Formal concept analysis as mathematical theory of concepts and concept hierarchies," in *Formal concept analysis*, pp. 1–33, Springer, Berlin, Heidelberg, 2005.

[10] U. Priss, "Formal concept analysis in information science," *Annual Review of Information Science and Technology*, vol. 40, no. 1, pp. 521–543, 2006.

[11] A. E. Qadi, D. Aboutajedine, and Y. Ennouary, "Formal concept analysis for information retrieval," *International Journal of Computer Science and Information Security*, vol. 7, no. 2, pp. 119–125, 2010.

[12] A. Formica, "Concept similarity in formal concept analysis: an information content approach," *Knowledge-Based Systems*, vol. 21, no. 1, pp. 80–87, 2008.

[13] A. Formica, "Ontology-based concept similarity in formal concept analysis," *Information Sciences*, vol. 176, no. 18, pp. 2624–2641, 2006.

[14] C. Carpineto and G. Romano, *Concept Data Analysis: Theory and Applications*, John Wiley & Sons, 2004.

[15] I. Nafkha, S. Elloumi, and A. Jaoua, "Using concept formal analysis for cooperative information retrieval," *The Leech*, vol. 1, no. 1, 2004.

[16] M. K. M. Rahman and T. W. Chow, "Content-based hierarchical document organization using multi-layer hybrid network and tree-structured features," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2874–2881, 2010.

[17] K. Baba, "Fast plagiarism detection based on simple document similarity," in *2017 twelfth international conference on digital information management (ICDIM)*, pp. 54–58, Fukuoka, Japan, 2017.

[18] J. Muangprathub, V. Boonjing, and P. Pattaraintakorn, "A new case-based classification using incremental concept lattice knowledge," *Data & Knowledge Engineering*, vol. 83, no. 1, pp. 39–53, 2013.

[19] F. Alqadah, "Similarity measures in formal concept analysis," in *Workshops of the 11th International Symposium on Artificial Intelligence and Mathematics (ISIAM2010)*, Fort Lauderdale, Florida, 2010.

[20] F. Alqadah and R. Bhatnagar, "Similarity measures in formal concept analysis," *Annals of Mathematics and Artificial Intelligence*, vol. 61, no. 3, pp. 245–256, 2011.

[21] F. Dau, J. Ducrou, and P. Eklund, "Concept similarity and related categories in searchsleuth," in *International conference on conceptual structures*, pp. 255–268, Berlin, Heidelberg, 2008.

[22] J. Saquer and J. S. Deogun, "Concept approximations based on rough sets and similarity measures," *International Journal of Applied Mathematics and Computer Science*, vol. 11, pp. 655–674, 2001.

[23] K. X. S. de Souza and J. Davis, "Aligning ontologies and evaluating concept similarities," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems*, pp. 1012–1029, Berlin, 2004.

[24] L. Wang and X. Liu, "A new model of evaluating concept similarity," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 842–846, 2008.

[25] R. Belohlavek and V. Vychodil, "Estimations of similarity in formal concept analysis of data with graded attributes," *Advances in Web Intelligence and Data Mining*, vol. 23, no. 1, pp. 243–252, 2006.

[26] A. Formica and E. Pourabbas, "Content based similarity of geographic classes organized as partition hierarchies," *Knowledge and Information Systems*, vol. 20, no. 2, pp. 221–241, 2009.

[27] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[28] C. Carpineto, G. Romano, and F. U. Bordoni, "Exploiting the potential of concept lattices for information retrieval with CREDO," *Journal of Universal Computer Science*, vol. 10, no. 8, pp. 985–1013, 2004.

[29] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundation*, Springer, Heidelberg, New York, 1999.

[30] J. Muangprathub, V. Boonjing, and P. Pattaraintakorn, "Information retrieval using a novel concept similarity in formal concept analysis," in *2014 international conference on information science, electronics and electrical engineering*, vol. 2, pp. 1249–1252, Sapporo, Japan, 2014.

[31] K. Lengnink, "Ähnlichkeit als Distanz in Begriffsverbänden," in *Begriffliche Wissensverarbeitung*, pp. 57–71, Springer, Berlin, Heidelberg, 2000.

[32] F. Dau, J. Ducrou, and P. Eklund, "Concept similarity and related categories in information retrieval using formal concept analysis," in *Annals of Mathematics and Artificial Intelligence*, Springer, Heidelberg-Berlin, 2009.

[33] Y. Zhao and W. Halang, "Rough concept lattice based ontology similarity measure," in *Proceedings of the first international conference on scalable information systems*, Hong Kong, 2006.

[34] Y. Zhao, X. Wang, and W. Halang, "Ontology mapping based on rough formal concept analysis," in *Advanced Int'l conference on telecommunications and Int'l conference on internet and web applications and services (AICT-ICIW'06)*, p. 180, Guadeloupe, French Caribbean, 2006.

[35] A. H. Osman, N. Salim, M. S. Binwahlan, S. Twaha, Y. J. Kumar, and A. Abuobieda, "Plagiarism detection scheme based on semantic role labeling," in *2012 international conference on Information Retrieval & Knowledge Management*, pp. 30–33, Kuala Lumpur, Malaysia, 2012.

[36] A. H. Osman, N. Salim, M. S. Binwahlan, R. Alteeb, and A. Abuobieda, "An improved plagiarism detection scheme based on semantic role labeling," *Applied Soft Computing*, vol. 12, no. 5, pp. 1493–1502, 2012.

[37] M. Paul and S. Jamal, "An improved SRL based plagiarism detection technique using sentence ranking," *Procedia Computer Science*, vol. 46, pp. 223–230, 2015.

[38] K. Vani and D. Gupta, "Using K-means cluster based techniques in external plagiarism detection," in *2014 international conference on contemporary computing and informatics (IC3I)*, pp. 1268–1273, Mysore, India, 2014.

[39] K. Vani and D. Gupta, "Investigating the impact of combined similarity metrics and POS tagging in extrinsic text plagiarism detection system," in *2015 international conference on advances in computing, communications and informatics (ICACCI)*, pp. 1578–1584, Kochi, India, 2015.

[40] A. Ekbal, S. Saha, and G. Choudhary, "Plagiarism detection in text using vector space model," in *2012 12th international conference on hybrid intelligent systems (HIS)*, pp. 366–371, Pune, India, 2012.

[41] S. Wang, H. Qi, L. Kong, and C. Nu, "Combination of VSM and Jaccard coefficient for external plagiarism detection," in *2013 international conference on machine learning and cybernetics*, vol. 4, pp. 1880–1885, Tianjin, China, 2013.

[42] P. Mahdavi, Z. Siadati, and F. Yaghmaee, "Automatic external Persian plagiarism detection using vector space model," in *2014 4th international conference on computer and knowledge engineering (ICCKE)*, pp. 697–702, Mashhad, Iran, 2014.

[43] P. K. Singh, C. A. Kumar, and A. Gani, "A comprehensive survey on formal concept analysis, its research trends and applications," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 2, pp. 495–516, 2016.

[44] C. A. Kumar, M. Radvansky, and J. Annapurna, "Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval," *Cybernetics and Information Technologies*, vol. 12, no. 1, pp. 34–48, 2012.

[45] C. A. Kumar and S. Srinivas, "Concept lattice reduction using fuzzy _K_-means clustering," *Expert systems with applications*, vol. 37, no. 3, pp. 2696–2704, 2010.

[46] C. A. Kumar, "Fuzzy clustering-based formal concept analysis for association rules mining," *Applied artificial intelligence*, vol. 26, no. 3, pp. 274–301, 2012.

[47] R. K. Chunduri and A. K. Cherukuri, "Scalable formal concept analysis algorithms for large datasets using spark," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 11, pp. 4283–4303, 2019.

[48] B. Koester, "Conceptual knowledge retrieval with FooCA: improving web search engine results with contexts and concept hierarchies," in *Industrial conference on data mining*, pp. 176–190, Berlin, Heidelberg, 2006.

[49] W. C. Cho and D. Richards, "Ontology construction and concept reuse with formal concept analysis for improved web document retrieval," *Web Intelligence and Agent Systems: An International Journal*, vol. 5, no. 1, pp. 109–126, 2007.

[50] Thai Word Segmentationhttp://www.arts.chula.ac.th/ling/wordseg.