دراسة مقارنة للخوارزميات المستخدمة في كشف السرقات الأدبية

د. رامز الخطيب* د. ناصر أبو صالح** م. لمى السبع***

(الإيداع: 4 حزيران 2018 ، القبول: 11 تشرين 2018)

الملخص:

العلم هو أحد أعمدة بناء الأمم وتقدمها ولا يستطيع أحد أن ينكر أنّ النمو الاجتماعي والاقتصادي في أي أمة مرتبط بالعلم والبحث العلمي الذي يعد الركيزة الأساسية في تطور أي مجتمع ولكن مع الأسف فقد طالت ظاهرة السرقات الأدبية مجال البحث العلمي حيث أصبح بعض الباحثون يلجؤون إلى نسب أفكار وكتابات شخص آخر لأنفسهم مستفيدين من التّقدم التّقني والتكنولوجي الذي سهل الوصول إلى معلومات هائلة باستخدام الانترنت.

ولم تقتصر السرقات الأدبية على البحث العلمي بل طالت جميع جوانب الحياة فأصبحت السرقة في أطروحات الدكتوراه والماجستير وفي الكتب والموسيقى والصور والمقالات الفنية والرسومات ونظراً للانعكاسات السلبية المترتبة عن هذه الظاهرة على الجامعات ومراكز البحث العلمي واعتماد أعداد كبيرة من الطلبة على الانترنت في تقديم بحوثهم الجامعية ظهرت الحاجة لتأمين أنظمة قادرة على اكتشاف هذه السرقات، وصنفت الدراسات السابقة أنواع الخوارزميات المستخدمة في هذه الأنظمة إلى نوعين أساسيين هما الخارجي حيث يتم مقارنة الوثيقة المشبوهة مع مجموعة من الوثائق الأخرى و الداخلي والتي يتم فيها دراسة الوثيقة المشبوهة ومحاولة تحديد مقاطع النص التي تختلف من حيث الصياغة والبنية عن بقية أجزاء النص، في هذا البحث سنقوم بإجراء دراسة نظرية لأهم الخوارزميات المستخدمة في كشف السرقات الأدبية واجراء مقارنة فيما بينها اعتمادا على عدة عوامل كطريقة المطابقة، العوامل التي تؤثر بنتائج كل طريقة ودقة النتائج.

الكلمات المفتاحية: السرقة الأدبية، كشف السرقات الأدبية، خوارزمية البصمة، النموذج الشعاعي، تحليل الاقتباسات، التشابه الدلالي، تطابق الوثائق.

*: عضو هيئة تدريسية في جامعة حماة – الكلية التطبيقية – قسم تقنيات الحاسوب.

**: عضو هيئة تدريسية في جامعة البعث – كلية الهندسة المعلوماتية.

***: طالبة ماجستير في جامعة البعث –كلية الهندسة المعلوماتية.

A Comparative Study of the Algorithms used to Plagiarism Detection

Dr. Ramez Alkhatib* , Dr. Nasser Abo Saleh , Lama Alsabea*****

(Received: 4 June 2018, Accepted: 11 November 2018)

Abstract:

Science is one of the pillars of nation building and progress. No one can deny that the social and economic growth of any nation linked to science and scientific research, which is the main pillar in the development of any society. Unfortunately, the phenomenon of plagiarism spread in the field of scientific research, some researchers adopted the ideas and writings of another person as their own, taking advantage of technical and technological advances that made access to vast information possible through the Internet.

Plagiarism was not confined to scientific research, but extended to all aspects of life. Plagiarism became in doctoral and master's dissertations, books, music, pictures, art articles and drawings. Due to the negative effects of this phenomenon on universities and scientific research centers and the adoption of large numbers of students on the Internet in there, university project has emerged the need to secure systems capable of detecting plagiarism. The previous studies classify the types of algorithms used in these systems into two basic types, external, where the suspicious document is compared with a set of other documents and Intrinsic, which analysis suspicious document and try to identify sections of text that differ in terms of wording and structure from the rest of the text. In this research we will conduct a theoretical study of the most important algorithms used in detecting plagiarism and comparing this algorithm, based on several factors such as the method of matching, the factors that affect the results of each method and the accuracy of the results.

Keywords: Plagiarism, Plagiarism Detection, Citation Analysis, Fingerprinting, VSM, Semantic Similarity, Document Similarity.

1- مقدمة:

إنّ التّقدم التقني والتكنولوجي في عصرنا الحديث وظهور شبكة الانترنت سهل الوصول إلى كم هائل من الكتب والمقالات والمعلومات التي أصبحت تستخدم في غايات مختلفة ليست كلها قانونية فانتشرت ظاهرة السرقات الأدبية والفكرية بشكل واسع في الأوساط العلمية والجامعات واعتبرها الكثيرون جريمة الكترونية كالقرصنة والفيروسات والاعلانات المزعجة. وحسب معجم أوكسفورد فالمقصود بالسرقة الأدبية هو نسخ أفكار، كلمات ونتائج شخص آخر ونسبها لنفسك.

فالسرقة الأدبية تعتبر وفق [14]:

1. نسب عمل شخص آخر لنفسك.
2. نسخ كلمات أو أفكار شخص آخر بدون اذن.
3. عدم استخدام علامات الاقتباس.
4. إعطاء معلومات خاطئة عن المصدر الفعلي للمعلومات.

والجدير بالذكر أن ظاهرة السرقة الأدبية ليست ظاهرة جديدة حيث قام عدة باحثون منذ عام 1920 بدراستها (Emmett D. and Brown B. ,2001) ومحاولة تحديد أسباب انتشارها والتعرف على أنواعها المختلفة(Bela G.,2014) وهي:

- ☒ نسخ/لصق: ويتم فيه نسخ أجزاء كبيرة من مصدر محدد دون ذكر المصدر.
- ☒ الاستبدال: ويتم فيه نسخ قطعة نصية بعد تغيير بعض الكلمات الرئيسية مع الحفاظ على المعلومات الأساسية للمصدر وعدم الإشارة إليه.
- ☒ الترجمة: يتم فيه ترجمة محتوى المصدر وإعادة استخدامه دون ذكر المصدر.
- ☒ السرقة الفنية: يتم فيه تمثيل محتوى المصدر بطريقة مختلفة كنص أو صورة أو فيديو دون ذكر المصدر الأساسي للمعلومة. (Ahmed O. and Naomie S. and Mohammed B., 2010).
- ☒ السرقة الفكرية: يتم فيه استخدام فكرة مطروحة من قبل شخص آخر ولكن هذه الفكرة غير شائعة.
- ☒ سرقة تجارب ونتائج.

إن أكثر المواقع التي تتم السرقة منها [15]:

- ❖ Wikipedia.com
- ❖ Slideshare
- ❖ Yahoo!Answers
- ❖ Scribd.com
- ❖ Coursehero.com

إنّ النتائج والآثار السلبية المترتبة عن انتشار هذه الظاهرة تؤثر في المستوى العلمي والثقافي لطلاب المدارس والجامعات وطلاب الدراسات العليا وطبعاً تشكك في الإنتاج العلمي الذي تقدمه هذه الجامعات.

2-الهدف من البحث:

في هذا البحث سنقوم بإجراء تحليل شامل لكل الطرق والمناهج المستخدمة سابقاً لكشف السرقة الأدبية واجراء مقارنة فيما بينها وفقاً لعدة عوامل منها المدخلات التي تطلبها كل خوارزمية، طريقة المطابقة، العوامل التي تؤثر بدقة النتائج، المساحة التخزينية وزمن التنفيذ. واعتماداً على هذه المقارنة سيتم تحديد إيجابيات وسلبيات كل طريقة لنتمكن في الأبحاث اللاحقة من تطوير خوارزمية جديدة تتجاوز السلبيات المستنتجة وتقلل منها.

3-أهمية البحث:

تعتبر السرقة الأدبية من الآفات التي ابتلت بها الجامعات والمؤسسات العلمية منذ نشوء العلم الحديث حيت تكمن خطورة هذه الآفة في أنها ترفع السارق علمياً ووظيفياً وقد ترقى به إلى أعلى المستويات السياسية والاجتماعية وهي تترك آثارا سيئة على السمعة العلمية للجامعات ومراكز البحث العلمي فهي "انحطاط ثقافي وبلطجة فكرية" . ولتحديد مدى انتشار هذه الآفة تم اجراء العديد من الدراسات أهمها دراسة قام بها Donald McCabe على 63700 طالب خلال 3 سنوات (2002-2005) فكانت النتيجة أن 38% من الطلاب الذين تحت سن التخرج و25% من طلاب التخرج اعترفوا بأنهم قاموا بعملية نسخ لجمل أو مقاطع من مصادر أخرى ضمن أبحاثهم دون الإشارة لمصدر المعلومات الحقيقي (Donald M.,2005).

إن السرقة الأدبية أو الفكرية في تزايد مستمر بسبب انتشار الانترنت حيث أصبح الطالب قادر على الوصول إلى كم هائل من المعلومات التي تستخدم في غايات مختلفة (Donald M.,2005)ومن هنا تنبع أهمية هذا البحث في اكتشاف الخوارزمية الأفضل في كشف السرقات الأدبية بعد إجراء دراسة مقارنة للخوارزميات المستخدمة في هذا المجال وبالتالي الحد من هذه الآفة التي تهدد الأمن الفكري للجامعات والمؤسسات العلمية.

1. الطرق المدروسة

استرجاع المعلومات (IR) هو علم يدرس آلية البحث عن معلومات معينة ضمن الوثائق وعن الوثائق وعن المعلومات التي توصف الوثائق بالإضافة للبحث في قواعد البيانات وشبكة الانترنت. "ويكيبيديا". ونظراً لأهمية هذا العلم في تأمين الكتب والمقالات والصور ومقاطع الفيديو المراد البحث عنها بسرعة وفعالية لذلك تم تطوير عدة طرق تختلف فيما بينها من حيث طريقة تمثيل الوثائق وآلية المطابقة واسترجاع المعلومات (Christopher M. & Prabhakar R. & Hinrich S.,2009). وبما أن الغاية من هذا البحث كشف السرقات الأدبية التي تعتبر احدى مهام استرجاع المعلومات (IR) لذلك سيتم دراسة الطرق المتبعة في كشف الاستلال من الوثائق النصية والتي تنقسم للأنواع التالية:

1.4. الأنظمة الداخلية (Intrinsic PDS)

تقوم هذه الأنظمة بتحليل الوثيقة المشبوهة لتحديد مقاطع النص التي تختلف عن بقية أجزاء النص من حيث الصياغة اللغوية وتركيب الجمل دون إجراء مقارنات مع وثائق أخرى.

حيث تقدم هذه الأنظمة تقرير عن تغير أسلوب الكتابة في الوثيقة المشبوهة كمؤشر لسرقة أدبية محتملة. (Benno S.& Nedim L. & Peter P., 2011)

وخوارزمية Stylometry تندرج تحت هذا النوع من الطرق.

Stylometry 1-1-4

تقترح هذه المقاربة طرق احصائية لتحليل أسلوب كتابة مؤلف (باحث) الوثيقة، فهي تعمل على اتخاذ قرار فيما إذا كان كاتب معين هو من كتب نص ما أم لا؟ وهذا ما يعرف بمصطلح Authorship verification . (Benno S. & Nedim L. &Peter P., 2011)

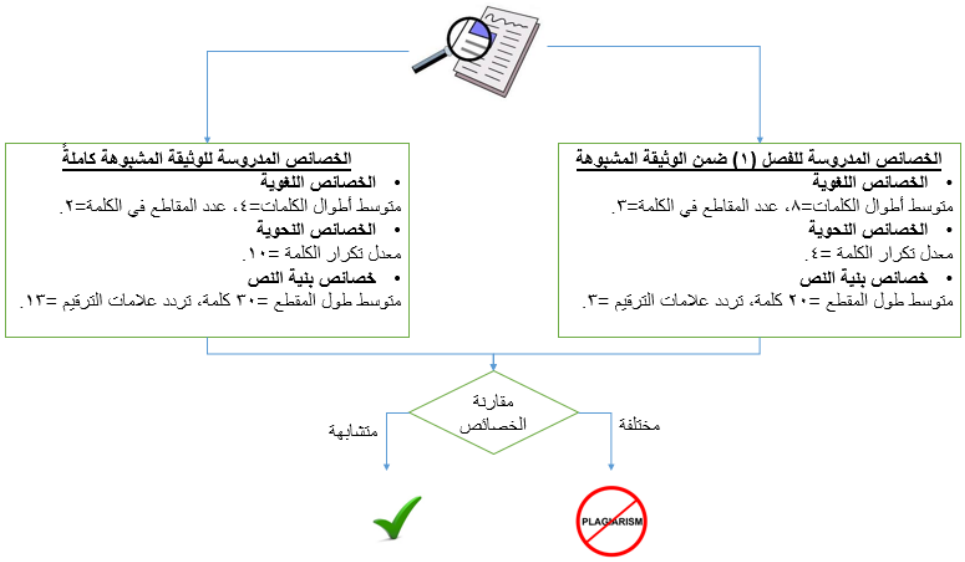
آلية العمل:

- تحديد المقاطع التي يجب مقارنتها (كلمة، جملة، فصل)
- تحديد الخصائص التي سيتم استخدامها والتي تتبع للفئات التالية:

- الخصائص اللغوية تكون هذه الخصائص اما على مستوى المحرف (تكرار الحرف) أو على مستوى الكلمة (متوسط أطوال الكلمات، عدد المقاطع في الكلمة).
- الخصائص النحوية تكرار الكلمة أو تكرار جزء من الكلام.
- الخصائص البنيوية متوسط طول المقطع أو تردد علامات الترقيم.

- مقارنة الخصائص اللغوية المحددة لمقطع النص المشبوه مع الخصائص اللغوية لكامل الوثيقة لتحديد المقاطع المختلفة.

حيث يوجد عدد من الطرق المستخدمة لإجراء عملية المقارنة باستخدام تقنيات تعلم الآلة (Ramnial H., Panchoo S., Pudaruth S. ,2016).



الشكل رقم (1): آلية عمل خوارزمية Stylometry

4-1الأنظمة الخارجية (External PDS)

تقارن هذه الأنظمة الوثيقة المشبوهة مع مجموعة الوثائق الموجودة في قائمة المراجع. وهذه المقارنة تتطلب:

- تمثيل للوثيقة بطريقة معينة تختلف وفقاً للخوارزمية المستخدمة.
- تحديد طريقة حساب نسبة التطابق بين الوثيقة المشبوهة والوثائق الموجودة في قائمة المراجع.
- إرجاع كل الوثائق التي تكون درجة تشابهها مع الوثيقة المشبوهة أكبر من عتبة معينة.

(Jan K. and Michal B.,2010)

الخوارزميات التي تنتمي لهذه الطريقة هي:

4-2-1 خوارزمية البصمة (FingerPrinting)

تعتبر من أكثر الطرق استخداماً في أنظمة كشف السرقات الأدبية.

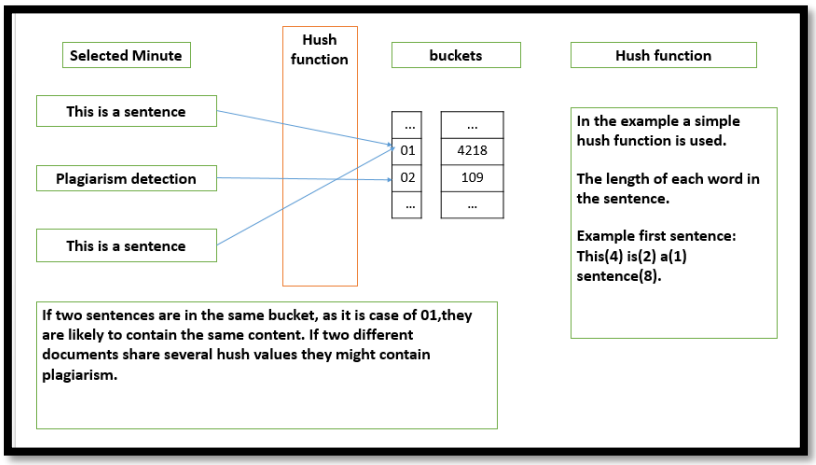
آلية العمل:

- يتم تمثيل الوثيقة المشبوهة بتقطيعها إلى سلاسل نصية.
- اختيار مجموعة من هذه السلاسل لتكون بصمة لهذه الوثيقة (fingerprint).

  كل سلسلة نصية من المجموعة السابقة تدعى minutiae (تفصيله).
- تقوم هذه الطريقة بتطبيق تابع تقطيع ما على كل تفصيله.

- ولقياس نسبة التطابق بين وثيقتين يتم المقارنة بين بصمتي الوثيقتين، أي يتم مقارنة كل تفصيله من الوثيقة المشبوهة مع تفصيلات كل الوثائق الموجودة في قائمة المراجع.
- عندما تتجاوز نسبة التطابق عتبة معينة تقترح المقاربة وجود سرقة أدبية.(Benno S. and Sven E.,2006)

الشكل التالي يوضح آلية عمل هذه الخوارزمية



الشكل رقم (2): آلية عمل خوارزمية البصمة

4-2-2تحليل الكلمات الواردة (Term Occurrence Analysis)

4-2-2-1 مطابقة النصوص (String Matching)

تقوم خوارزميات مطابقة النصوص بالتحقق من ورود سلسلة نصية معينة (Pattern) ضمن نص آخر. تستخدم هذه الخوارزميات في محركات البحث، فلترة الإعلانات المزعجة، معالجة اللغات الطبيعية، البيوانفورماتيك وفي كشف السرقات الأدبية ومن الخوارزميات المستخدمة لمطابقة النصوص (Naïve, Knuth-Morris-Pratt, Boyer-Moore, Rabin-Karp) وهي موضحة بالدراسة (Marc G.,2014)

آلية العمل:

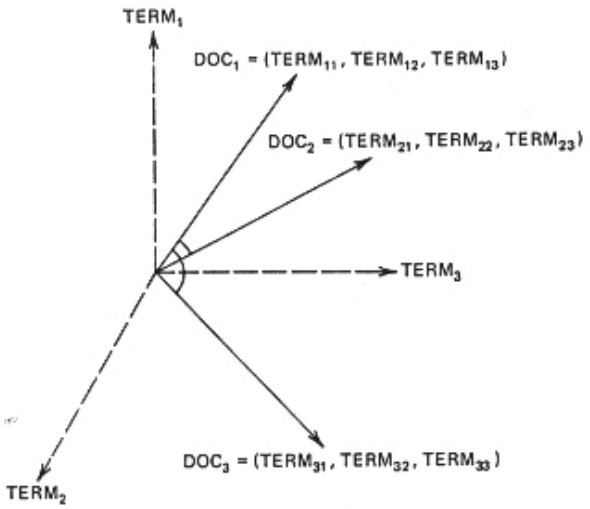
- يتم تمثيل الوثيقة المشبوهة وكل الوثائق الموجودة في قائمة المراجع باستخدام بنى المعطيات اللاحقة (N. Sandhya and Dr. A. Govardhan and Y. Sri Lalitha and Dr. K. Anuradha ,2011)
- من أجل كل سلسلة نصية في الوثيقة المشبوهة يتم البحث عنها ضمن تمثيل وثائق المراجع.
- نسبة التطابق بين وثيقتين تعتمد على عدد السلاسل المتطابقة بين الوثيقتين.

4-2-2-2 نموذج الفضاء الشعاعي (Vector Space Model)

يعتبر أحد المفاهيم الأساسية بمجال استرجاع المعلومات .(Christopher M. & Prabhakar R. & Hinrich S.,2009)

آلية العمل:

- يتم تحليل الوثيقة المشبوهة لتحديد المصطلحات ذات الأولوية الأعلى بالنسبة لموضوع الوثيقة.
- يتم تمثيل الوثائق كأشعة اعتماداً على المصطلحات التي تم تحديدها سابقاً. والشّكل (3) يوضّح مثال عن طريقة تمثيل الوثائق باستخدام الأشعة.
- لإيجاد درجة التطابق بين وثيقتين يتم مقارنة التمثيل الشعاعي لكل منها، ويوجد العديد من التوابع المستخدمة لقياس التشابه (Martin P. and Benno S., and Maik A.,2008)



الشكل رقم (3): التمثيل الشعاعي للوثائق باستخدام نموذج الفضاء الشعاعي

4-2-3 الخوارزميات التي تعتمد على الاستشهادات المرجعية

تم طرح هذه الخوارزميات في عام 2014 من قبل الباحث Bela Gipp من جامعة Konstanz حيث قدم الباحث أربع خوارزميات لكشف السرقات الأدبية معتمداً على المقاييس التي تستخدم الاستشهاد المرجعي في إيجاد الوثائق المرتبطة.

4-2-3-1 ترابط قائمة المراجع Bibliographic Coupling

تعتبر من أقدم الطرق المستخدمة في كشف السرقات الأدبية، تقوم هذه الخوارزمية بإيجاد عدد المراجع المشتركة بين الوثيقة المشبوهة والوثيقة من قائمة المراجع. فإذا تجاوز عدد المراجع المشتركة عتبة معينة يقترح النظام احتمال وجود سرقة أدبية (Bela G.,2014)

4-2-3-2 أطول سلسلة مشتركة من الاستشهادات المرجعية (Longest Common Citation Sequence)

ويرمز لها اختصاراً LCCS.تقيس هذه الخوارزمية درجة التشابه بين وثيقتين اعتماداً على أطول سلسلة جزئية مشتركة من الاستشهادات المرجعية الواردة بنفس الترتيب في الوثيقتين المدروستين. (Bela G.,2014)

الشكل (4) يوضح مثال عن تطبيق خوارزمية LCCS بين الوثيقتين A,B حيث يتم تمثيل كل وثيقة بسلسلة الاستشهادات المرجعية التي تتكون منها، ثم يتم إيجاد الاستشهادات المرجعية المتطابقة بينهما وبنفس الترتيب وهي (3,4,5)وبالتالي تكون أطول سلسلة جزئية مشتركة بينهما تتكون من 3 استشهادات مرجعية.

Doc A: 2, 3, 1, 4, 6, 8, 5, 9
Doc B: 3, 8, 9, 4, 10, 11, 5
LCCS: 3 4 5

LCCS=3

الشكل رقم (4): آلية عمل خوارزمية LCCS

4-3-2-3 Greedy Citation Tiling (GCT)

تقيس هذه الخوارزمية درجة التشـــابه بين وثيقتين اعتماداً على عدد الســـلاســـل الجزئية المشـــتركة بنفس الترتيب فقط في الوثيقتين المدروستين. (Bela G.,2014)

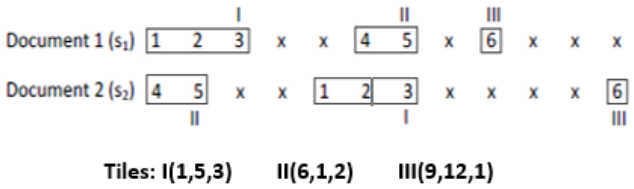
حيث تقوم هذه الخوارزمية بإيجاد كل السلاسل وتخزن كل سلسلة كtile ويتم تمثيل tile كثلاثية

$$T= (s1, s2, L)$$

حيث S1: موقع بداية tileفي السلسلة الأولى

S2: موقع بداية tileفي السلسلة الثانية

L: طول tile



الشكل رقم (5): مثال عن خوارزمية GCT

في الشكل السابق نلاحظ تشارك الوثيقتين المدروستين ب 3 سلاسل جزئية وتم تمثيل كل منها كtile.

ولقراءة أول tile في الشكل (5):

I(1,5,3) تبدأ هذه السلسلة من الدليل رقم 1 من الوثيقة الأولى ومن الدليل رقم 5 من الوثيقة الثانية وبطول 3 استشهادات مرجعية متتالية.

4-3-2-4 تقطيع الاستشهادات المرجعية Citation Chunking

قطعة الاستشـهادات المرجعية (chunkذCitation –) هو سلسـلة جزئية من سلسـلة الاسـتشـهادات المرجعية الواردة في الوثيقة مضافاً اليها قيمة عددية تمثل حجم القطعة.

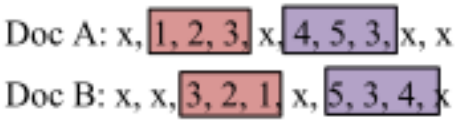
تتألف هذه الخوارزمية من 3 خطوات (Bela G.,2014):

a) تشكيل القطع

تقوم الخوارزمية بالبحث عن الاستشهادات المرجعية المشتركة بين الوثيقتين المدروستين لتشكيل القطع. اعتماداً على احدى الاستراتيجيات التالية:

- القطعة الواحدة تتكون من مجموعة من الاستشهادات المرجعية المشتركة فقط.

الشكل (6) يبين أن القطعة الأولى تتكون من الاستشهادات المرجعية (1و2و3) فقط لكونها مشتركة بين الوثيقتين. والاستشهادات المرجعية (x) غير محتواه داخل القطعة نظراً لأنها غير مشتركة.



الشكل رقم (6): مثال عن تشكيل القطع وفق الاستراتيجية الأولى

➢ المسافة للاستشهاد المرجعي المشترك السابق ≤ عدد الاستشهادات المرجعية في القطعة الحالية

يتم اضافة الاستشهاد المرجعي للقطعة الحالية إذا كان عدد الاستشهادات المرجعية غير المشتركة التي تفصله عن آخر استشهاد مشترك(N) أقل من عدد الاستشهادات المرجعية المشتركة الموجودة في القطعة الحالية(S).

$$N \leq S$$

Doc A: x, 1, 2, 3, x, x, 4, 5, x, x, x, x, x, x, 6, 7

Doc B: 3, 2, x, 1, x, x, 4, x, x, x, x, x, 5, 6, 7, x

الشكل رقم (7): مثال عن تشكيل القطع وفق الاستراتيجية الثانية

من الشكل (7) سندرس انتماء الاستشهاد رقم (5) للقطعة الأولى في كلا الوثيقتين A,B

الوثيقة A:

1. الاستشهاد المرجعي رقم (5) يلي مباشرة آخر استشهاد مشترك وهو رقم (4) إذاً عدد الاستشهادات المرجعية غير المشتركة التي تفصل بين الاستشهاد 4 والاستشهاد 5 هو N=0
2. عدد الاستشهادات المرجعية المشتركة في القطعة الحالية (الأولى) يساوي 3(1,2,3,4) أي S=4

من 1&2نجد N<S اذاً الاستشهاد رقم (5) ينتمي للقطعة الأولى في الوثيقة A.

الوثيقة B:

1. عدد الاستشهادات المرجعية غير المشتركة التي تفصل بين الاستشهاد المرجعي 4 والاستشهاد المرجعي 5 هو N=5
2. وعدد الاستشهادات المرجعية المشتركة في القطعة الحالية (الأولى) يساوي 4 أي S=4

إذاً N>S لذلك لا يمكن أن نضم الاستشهاد المرجعي رقم 5 للقطعة الأولى.

➢ المسافة للاستشهاد المرجعي المشترك السابق ≤ مجال معين

الاستشهادات المرجعية تشكل قطعة إذا كانت المسافة الفاصلة فيما بينها في النص أقل من عتبة محددة. ولتحديد المسافة العظمى لتباعد الاستشهادات المرجعية قامت احدى الدراسات باستخدام متوسط عدد الكلمات مثال:

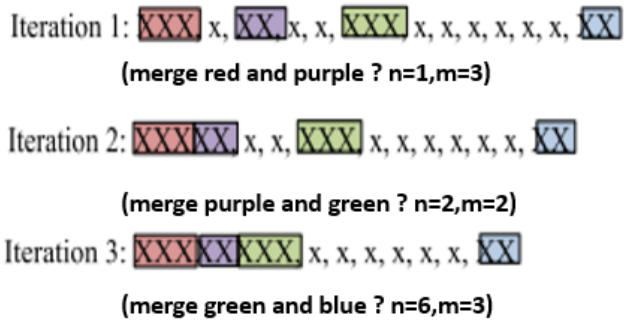
في وثيقة ما إذا كان المقطع يحوي كمتوسط 120 كلمة و720 محرف عندها إذا كان عدد الكلمات التي تفصل بين استشهاد مرجعي وآخر أقل من 120 كلمة عندها كلا الاستشهادين المرجعيين تابعين لقطعة واحدة.

b) دمج القطع (هذه الخطوة اختيارية)

يتم دمج القطع إذا كان:

عدد المراجع غير المشتركة (n) ≤ عدد المراجع المشتركة (m)

حيث m طويلة القطعة السابقة المراد الدمج بها.



الشكل رقم (8): يوضح آلية دمج القطع

من الشـكل (8) في المرحلة الأولى لندرس إمكانية دمج القطعتين الأولى والثانية نلاحظ أن عدد الاسـتشـهادات المرجعية غير المشتركة بينهما تساوي 1 (n=1) ،وطويلة القطعة الأولى يساوي 3 (m=3) ومنه نستنتج أن m>n لذلك يمكن دمج القطعتين المدروستين.

c) المقارنة بين القطع

يتم مقارنة القطع بغض النظر عن ترتيب المراجع داخل القطعتين، حيث عدد المراجع المتطابقة بين قطعتين يحدد درجة التشابه بينهما.

يوجد طريقتين لإجراء عملية المقارنة بين القطع (Bela G.,2014):

- نقارن كل قطعة من الوثيقة الأولى مع كل قطعة من الوثيقة الثانية ثم نخزن القطع التي تحقق أكبر درجة تشابه.
- نقارن قطعة من الوثيقة الأولى مع سلسلة المراجع من الوثيقة الثانية.

2. دراسة مقارنة

5-1 تعاريف

5-1-1 عامل الانتقاء:

يحدد عامل الانتقاء حجم النص المختار من الوثيقة المدروسة. ونعبر عنه بالعلاقة التالية:

$$y= \frac{xi}{n}$$

حيث n تمثل حجم الوثيقة المدروسة، وxi هي عدد السلاسل المختارة من الوثيقة المدروسة.

وبما أنّ أي سلسلة نصية xi تتكون من مجموعة من الكلمات wj لذلك يمكن تمثيل السلسلة النصية بالعلاقة التالية

$$Xi = w1w2.....wj$$

ومنه يكون عامل الانتقاء موضح بالعلاقة التالية:

$$y= \frac{\sum_{j=1} wj}{n}$$

5-1-2 عدد الاستشهادات المرجعية

يمثل عدد الاستشهادات المرجعية الواردة ضمن الوثيقة المدروسة. ونعبر عنه بالعلاقة التالية:

$$\boldsymbol{C}\mathrm{d} = \sum_{\boldsymbol{j=1}} \boldsymbol{cj}$$

حيث Cd يمثل عدد الاستشهادات المرجعية الكلية في الوثيقة d، وCj هو الاستشهاد المرجعي رقم j ضمن الوثيقة d.

5-1-3 عامل الاستشهاد المرجعي

يحدد هذا العامل حجم قطعة الاستشهادات المرجعية (Citation – chunk) التي يتم اختيارها من الوثيقة المدروسة ونعبر عنه بالعلاقة التالية:

$$\beta = \frac{\boldsymbol{C}\mathrm{h}}{\boldsymbol{C}\mathrm{d}}$$

حيث Ch تمثل عدد الاستشهادات المرجعية الواردة في القطعة، وCd يمثل عدد الاستشهادات المرجعية الكلية في الوثيقة d.

5-1-4 عامل التطابق

يحدد هذا العامل عدد المراجع المشتركة بين وثيقتين d1,d2 ونعبر عنه بالعلاقة التالية:

$$R=Rd1 \cap Rd2$$

حيث Rd1,Rd2 تمثل المراجع الموجودة في الوثيقتين d1,d2.

5-2 المقارنة

إن هدف هذا البحث دراسة الخوارزميات المستخدمة في كشف السرقات الأدبية واجراء مقارنة فيما بينها اعتماداً على عدة عوامل وهي:

5-2-1 آلية المطابقة

يوجد نوعين لآلية المطابقة بين النصوص (Bela G.,2014) وهي:

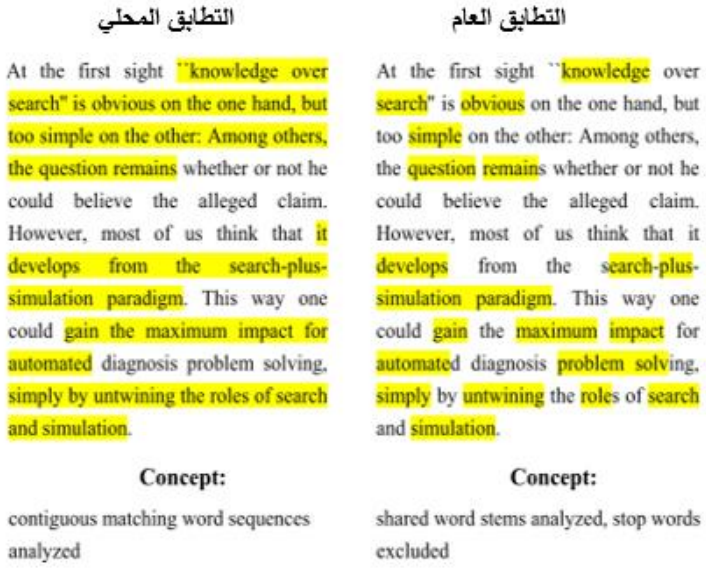
- **التطابق المحلي**

يتم إيجاد التطابق بين الوثائق على مستوى سلاسل من النص.

- **التطابق العام**

يتم إيجاد التطابق بين الوثائق على مستوى الوثيقة كاملة باستخدام الكلمات الأساسية في النص فقط.

الشكل (9) يوضح الفرق بين النوعين السابقين حيث نلاحظ في القسم اليساري يتم معالجة النص حسب التطابق المحلي فيتم إيجاد كل السلاسل النصية المتطابقة بين وثيقتين. بينما في القسم اليميني والذي يعتمد على التطابق العام يتم استخدام كلمات محددة من النص فقط.



الشكل رقم (9): مفهوم التطابق المحلي والتطابق العام

5-2-2 نوع السرقة التي تكتشفها الخوارزمية المتبعة:

ممكن أن تكون نوع أو أكثر من أنواع السرقات الأدبية (نسخ/لصق، استبدال، الترجمة...).

5-2-3 المساحة التخزينية

5-2-4 دقة النتائج التي يتم الحصول عليها عند تطبيق هذه الخوارزمية

5-2-5 المعالجة السابقة: بعض الخوارزميات تتطلب معالجة مسبقة للوثيقة المدروسة أو لكل الوثائق الموجودة في قاعدة البيانات لتمثيلها بطريقة معينة، وفي الجدول (1) تم ذكر المعالجة التي تطلبها كل خوارزمية ان وجدت.

5-2-6 زمن التنفيذ

يعتبر زمن التنفيذ عامل مهم لتقييم الخوارزميات والمقارنة فيما بينها، لذلك سنقوم بمقارنة خوارزميات كشف السرقات الأدبية اعتماداً على زمن التنفيذ.

حيث زمن تنفيذ الخوارزمية يتعلق بالزمن اللازم لإجراء المعالجة السابقة للوثائق المدروسة (tp) وبالزمن اللازم لقراءة وتحليل النص المختار (tr) ونعبر عنه بالعلاقة:

$$T=tp+(s*tr)$$

حيث:

T: الزمن الكلي لتنفيذ الخوارزمية

tp: هو الزمن الذي تطلبه المعالجة السابقة للوثائق.

S: المساحة التخزينية للنص المدروس

tr: زمن قراءة النص المدروس وهو يتناسب طرداً مع المساحة التخزينية للنص.

الخوارزمية	آلية المطابقة	نوع السرقة المكتشفة	العوامل		المساحة التخزينية	دقة النتائج	المعالجة السابقة	زمن التنفيذ
خوارزمية البصمة	تطابق محلي	نسخ/لصق	حجم القطعة "عدد الكلمات في القطعة الواحدة"		تناسب طردي مع y	تناسب طردي مع y	لا تتطلب معالجة سابقة	• يتناسب طرداً مع y • Tp=0
			عدد السلاسل التي تمثل الوثيقة		تناسب طردي مع y	تناسب طردي مع y	لا تتطلب معالجة سابقة	• يتناسب طرداً مع y • Tp=0
			آلية اختيار القطع	السلاسل الأكثر تكرارا	تناسب طردي مع y	تناسب طردي مع y	تتطلب معالجة سابقة لكافة الوثائق لاستخراج السلاسل	• تناسب طردي مع y و tp
				السلاسل المتشابهة	تناسب طردي مع y	تناسب طردي مع y	لا تتطلب معالجة مسبقة للوثائق	• الزمن يتناسب طرداً مع y • Tp=0
خوارزميات مطابقة النصوص	تطابق محلي	نسخ/لصق	حجم الوثيقة		تناسب عكسي مع y	دقة عالية لأنها تتحقق من كل سلاسل النص	تتطلب معالجة سابقة لكل الوثائق لتمثيلها باستخدام بنى المعطيات اللاحقة	• تناسب عكسي مع y • تناسب طردي مع tp
Vsm	تطابق عام	نسخ/لصق استبدال	عدد المصطلحات التي تمثل الوثيقة		تناسب طردي مع y	اعتماداً على عدد المصطلحات وأهميتها في النص	تتطلب استخراج المصطلحات ذات الأهمية الأعلى في الوثيقة	تناسب طردي مع y,tp
Stylometry	تطابق محلي	–	حجم النص		تناسب عكسي مع y	تناسب عكسي مع y	لا تتطلب معالجة سابقة	• تناسب عكسي مع y • tp=0
			عدد الخصائص اللغوية والنحوية المدروسة		تزداد بزيادة الخصائص المدروسة	تزداد بزيادة الخصائص المدروسة	تتطلب معالجة سابقة للوثيقة لاستخراج الخصائص المدروسة	تناسب طردي مع n,tp
			عدد الباحثين المشتركين		تزداد المساحة التخزينية بزيادة عدد الباحثين لأننا بحاجة لتخزين السمات اللغوية والنحوية لكل باحث	تقل الدقة كلما زاد عدد الباحثين المشتركين في البحث	يزيد التعقيد بزيادة عدد الباحثين لأننا بحاجة لمعرفة خصائص الكتابة لكل باحث	تناسب طردي مع n,tp

BC	تطابق عام	–	عدد المراجع		تناسب طردي مع عامل التطابق R	تناسب طردي مع عامل التطابق R	لا تتطلب معالجة سابقة	• تناسب طردي مع R • Tp=0
LCCS	تطابق عام	نسخ/لصق	عدد الاستشهادات المرجعية في النص		تناسب طردي مع Cd	تناسب طردي مع Cd	تتطلب معالجة سابقة لاستخراج الاستشهادات المرجعية من الوثائق المدروسة	تناسب طردي مع Cd ,tp ,n
GCT	تطابق محلي	نسخ/لصق، استبدال، الترجمة	عدد الاستشهادات المرجعية في النص		تناسب طردي مع Cd وβ	تناسب طردي مع Cd وβ	تتطلب معالجة سابقة لاستخراج الاستشهادات المرجعية من الوثائق المدروسة	تناسب طردي مع Cd وβ وn وtp
Cit-chunk	تطابق محلي	نسخ/لصق، استبدال، الترجمة	حجم القطعة "عدد الاستشهادات المرجعية في السلسلة الواحدة"		تناسب طردي مع β	تناسب طردي مع β	تتطلب معالجة سابقة لاستخراج الاستشهادات المرجعية من الوثائق المدروسة	تناسب طردي مع β وn وtp
			آلية تشكيل القطع	الاستشهادات المرجعية المشتركة	تناسب طردي مع Cd	تناسب طردي مع Cd	تتطلب معالجة سابقة لاستخراج الاستشهادات المرجعية من الوثائق المدروسة	تناسب طردي مع Cd,tp
				الاستشهادات المرجعية غير المشتركة	تناسب طردي مع Cd	تناسب طردي مع Cd	تتطلب معالجة سابقة لاستخراج الاستشهادات المرجعية من الوثائق المدروسة	تناسب طردي مع Cd,tp
				المسافة بين الاستشهادات المرجعية	تناسب طردي مع Cd بالإضافة للحاجة لتخزين المسافات بين الاستشهادات المرجعية	تزداد الدقة عند أخذ المسافة بين الاستشهادات المرجعية بعين الاعتبار لأنها تدل على ارتباط قوي للوثائق	تحتاج معالجة مسبقة للوثائق لتحديد المسافة بين الاستشهادات المرجعية	تناسب طردي مع Cd, tp

6-النتائج:

- من خلال دراستنا تبين أن خوارزمية البصمة مناسبة لكشف السرقات من نوع نسخ/لصق وهي تستخدم لإيجاد التطابق المحلي بين الوثائق حيث تقوم بتحديد السلاسل النصية المسروقة ضمن الوثيقة المشبوهة. ولقد تبين من خلال الدراسة أن زمن تنفيذ هذه الخوارزمية هو الزمن الذي تطلبه المعالجة السابقة للوثائق المدروسة لاستخراج السلاسل الأكثر تكراراً في الوثيقة أو لاستخراج السلاسل المتشابهة بين الوثائق المدروسة بالإضافة للزمن الذي تحتاجه الخوارزمية لقراءة هذه السلاسل والتحقق منها فكلما زاد عدد هذه السلاسل وحجمها تزداد المساحة التخزينية لها وبالتالي يزيد زمن تنفيذ هذه الخوارزمية ودقتها في كشف السرقة لأننا نقوم بالتحقق من أجزاء أكبر من النص.
- وتعتبر خوارزميات مطابقة النصوص من أقدم الخوارزميات المستخدمة لكشف السرقة من نوع نسخ/ لصق فهي تعمل على تحديد النص المسروق في الوثيقة المشبوهة وأداء هذه الخوارزمية يعتمد على حجم الوثيقة المشبوهة فهي تعمل على التحقق من كل سلاسل النص لذلك تتطلب مساحة تخزينية كبيرة جداً لتخزين كل السلاسل بعد تمثيلها بإحدى بنى المعطيات اللاحقة وبالتالي زمن تنفيذ هذه الخوارزمية كبير ودقتها عالية جداً لأنها تتحقق من كل السلاسل النصية في الوثيقة المشبوهة.
- بينما تقوم خوارزمية نموذج الفضاء الشعاعي بإيجاد التطابق العام بين الوثائق التي يتم تمثيلها كأشعة اعتماداً على عدد من المصطلحات المستخرجة منها وبالتالي زمن تنفيذ هذه الخوارزمية يعتمد على الزمن اللازم لتحليل كافة الوثائق لانتقاء المصطلحات ذات الأولوية الأعلى وتمثيلها كأشعة بالإضافة إلى الزمن اللازم لإيجاد التطابق بين هذه الأشعة، واعتماداً على ما سبق نلاحظ أن المساحة التخزينية لهذه الخوارزمية تتناسب طرداً مع عدد المصطلحات المستخرجة من الوثيقة ودقتها أقل من بقية الخوارزميات فهي عاجزة عن تحديد السلاسل المسروقة في الوثيقة المشبوهة.
- وإن زمن تنفيذ خوارزمية Stylometry كبير جداً فهي تقوم بتحليل النص لتحديد السمات اللغوية والنحوية لكل مقاطع النص ثم مقارنة هذه الخصائص مع أسلوب كتابة باحث الوثيقة، ومنه نلاحظ أن دقة النتائج التي يتم الحصول عليها تعتمد بشكل كبير على عدد الباحثين المشتركين في كتابة الوثيقة وتكون عرضة للفشل عند تعدد الباحثين حيث يصبح من الصعب تحديد كاتب كل مقطع في الوثيقة المشبوهة وهذا النوع من الخوارزميات يحتاج لتخزين أسلوب كتابة جميع الباحثين لنتمكن من اجراء عملية المقارنة بين الخصائص المدروسة فهي تتطلب اذاً مساحة تخزينية كبيرة ومما سبق نستنتج أن فعالية هذه الخوارزمية سيئة مقارنة مع باقي الخوارزميات.
- ثم قمنا بدراسة الخوارزميات الأربع التي تعتمد على الاستشهادات المرجعية ولاحظنا أن:

❖ زمن تنفيذ خوارزمية BC قليل نسبياً لأنها لا تحتاج لإجراء معالجة سابقة للوثائق فهي تقوم باكتشاف التطابق العام بين الوثائق عن طريق إيجاد عدد المراجع المشتركة بين الوثيقتين المدروستين وبالتالي المساحة التخزينية التي تطلبها هذه الخوارزمية تتناسب طردا مع عدد المراجع المستخرجة من الوثائق وكلما زاد عدد المراجع المشتركة بين وثيقتين (R) يعتبر ذلك مؤشر قوي على احتمالية ورود سرقة محتملة.

❖ بينما خوارزمية LCCS تحتاج لاستخراج الاستشهادات المرجعية من كل الوثائق وتحليلها للحصول على أطول سلسلة جزئية مشتركة بين الوثائق المدروسة لذلك يكون زمن تنفيذ هذه الخوارزمية يعتمد على الزمن اللازم لاستخراج الاستشهادات المرجعية من النص وعلى الزمن اللازم لإيجاد أطول سلسلة مشتركة فيما بينها وكلما زاد عدد هذه الاستشهادات المرجعية زادت المساحة التخزينية التي تطلبها الخوارزمية ،ولقد تبين لنا من خلال الدراسة أن أداء هذه الخوارزمية من حيث دقة النتائج التي نحصل عليها يكون مناسب لكشف السرقات من نوع نسخ/ لصق ولكنها تصبح عاجزة عن كشف السرقات التي يتم فيها تغيير بسيط في ترتيب ورود الاستشهادات المرجعية.

- ويعتبر أداء خوارزمية GCT أفضل فهي قادرة على كشف السرقات التي يعمد الباحث فيها على اجراء تغييرات في ترتيب الاستشهادات المسروقة لأن هذه الخوارزمية تعمل على إيجاد كل السلاسل الجزئية المشتركة من سلاسل الاستشهادات المرجعية بين الوثائق المدروسة وبالتالي زمن تنفيذ هذه الخوارزمية يعتمد على الزمن اللازم لإجراء المعالجة السابقة لاستخراج سلاسل الاستشهادات المرجعية بالإضافة للزمن اللازم لتحديد السلاسل الجزئية المتشابهة بينها ودقة النتائج التي نحصل عليها عند تطبيق هذه الخوارزمية تعتمد على عدد الاستشهادات المرجعية.
- إن أداء خوارزمية cit-chunk في إيجاد التطابق المحلي بين الوثائق يعتمد على عدد الاستشهادات المرجعية وعلى حجم قطعة الاستشهادات المرجعية فكلما زادت قيمة هذين العاملين تزداد المساحة التخزينية التي تطلبها الخوارزمية وإن الاستراتيجية المستخدمة لتشكيل القطع يؤثر على حجم القطعة فعند اختيار الاستشهادات المشتركة فقط أو الاستشهادات غير المشتركة تكون المساحة التخزينية أقل مما تطلبه الاستراتيجية الثالثة التي تعتمد على إيجاد الاستشهادات المرجعية المشتركة بالإضافة للمسافة الفاصلة بين هذه الاستشهادات وبالتالي زمن تنفيذ هذه الخوارزمية يعتمد أيضاً على الاستراتيجية المتبعة لتشكيل القطع وعلى الزمن اللازم لمطابقة هذه القطع، ومن خلال هذه الدراسة تبين لنا أن دقة النتائج التي نحصل عليها عند أخذ المسافة الفاصلة بين الاستشهادات المرجعية بعين الاعتبار تكون عالية مقارنة بالاستراتيجيات الأخرى.

6-الخاتمة والأعمال المستقبلية

من خلال هذه الدراسة تبين لنا أن أداء الخوارزميات التي تعتمد على الاستشهادات المرجعية من حيث المساحة التخزينية أفضل من بقية الخوارزميات فهي تقوم بتخزين الاستشهادات المرجعية الواردة في النص مع إمكانية تخزين المسافة الفاصلة فيما بينها، تليها خوارزمية VSM التي تعمل على تخزين مجموعة من المصطلحات المرتبطة بمحتوى الوثيقة ثم خوارزمية BC التي تخزن المراجع التي تعتمدها الوثيقة بينما تقوم خوارزمية البصمة بتخزين عدد من السلاسل النصية للوثيقة المشبوهة وتعتمد خوارزمية stylometry على تخزين الخصائص اللغوية والنحوية للباحث ولكل مقطع من مقاطع النص وتعتبر المساحة التخزينية التي تطلبها خوارزميات مطابقة النصوص هي الأكبر فهي تقوم بتخزين كل سلاسل النص للوثيقة المشبوهة وللوثائق الأخرى.

واعتماداً على آلية المطابقة نجد أن كل من خوارزمية البصمة و stylometry ومعظم الخوارزميات التي تعتمد على الاستشهادات المرجعية تقوم بإيجاد التطابق المحلي فهي قادرة على تحديد السلاسل المسروقة في الوثيقة المشبوهة بينما تعجز خوارزميات vsm,Bc,Lccs عن ذلك.

وبما أن زمن التنفيذ عامل مهم في مقارنة الخوارزميات لذلك قمنا في هذه الدراسة بتحديد زمن تنفيذ كل خوارزمية فتبين لنا أن خوارزمية Bc تعد الأفضل من حيث زمن التنفيذ تليها الخوارزميات التي تعتمد على الاستشهادات المرجعية ثم خوارزمية vsm ثم خوارزمية البصمة التي يتعلق زمن تنفيذها بالزمن اللازم لاستخراج السلاسل مضافا اليه الزمن اللازم لمطابقة هذه السلاسل ويعتبر أداء خوارزميات مطابقة النصوص الأسوأ من حيث زمن التنفيذ الذي تطلبه الخوارزمية فهي تحتاج لزمن كبير لتمثل كل الوثائق بإحدى بنى المعطيات اللاحقة ثم مقارنة هذه البنى.

بناءً على هذه الدراسة نستنتج أن أداء الخوارزميات التي تعتمد على الاستشهادات المرجعية أفضل من أداء بقية الخوارزميات المستخدمة لكشف السرقات الأدبية. ويمكن في الأبحاث القادمة العمل على تطوير أنظمة قادرة على كشف السرقات الأدبية بتطبيق الخوارزميات التي تعتمد على الاستشهادات المرجعية مع إمكانية دمجها مع واحدة أو أكثر من الخوارزميات الأخرى التي تم دراستها من خلال هذا البحث أو تطوير خوارزمية جديدة تعتمد على تقنيات الوب الدلالي إضافة لخوارزميات الاستشهادات المرجعية.

المراجع العلمية

1. Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan.(2010) Plagiarism Detection Using Graph-Based Representation. JOURNAL OF COMPUTING, VOLUME 2, ISSUE 4, APRIL 2010, ISSN 2151-9617.
2. Ramnial H., Panchoo S., Pudaruth S. (2016) Authorship Attribution Using Stylometry and Machine Learning Techniques. In: Berretti S., Thampi S., Srivastava P. (eds) Intelligent Systems Technologies and Applications. Advances in Intelligent Systems and Computing, vol 384. Springer, Cham.
3. Hoad. TC, Zobel. J, 2003, Methods for Identifying Versioned and Plagiarized Documents. Journal of the American Society for Information Science and Technology 54(3):203–215.
4. Kasprzak. J, Brandejs. M, 2010, Improving the Reliability of the Plagiarism Detection System , Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy.
5. McCabe DL, 2005, Cheating among College and University Students: A North American Perspective. International Journal for Academic Integrity,1(1),1-11.
6. Manning. CD, Raghavan. P, Schütze. H, 2009- An Introduction to Information Retrieval. Online edition edn. Cambridge University Press, Cambridge, England.
7. Potthast. M, Stein. B, Anderka. M, 2008- A Wikipedia-based Multilingual Retrieval Model. In: Proceedings of the 30th European Conference on Advances in Information Retrieval, Springer, 522–530.
8. Stein. B, Meyer. S, 2006, Near Similarity Search and Plagiarism Analysis, Springer, 430–437.
9. Stein. B, Lipka. N, Prettenhofer. P, 2011, Intrinsic Plagiarism Analysis. Language Resources and Evaluation 45(1), 63–82.
10. GIPP.B, 2014- Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis. Springer Vieweg Research, Berlin, 400.
11. http://www.nytimes.com/2012/08/31/education/harvard-says-125-students-may-have-cheated-on-exam.html
12. Marc GOU ,July 30, 2014,Algorithms for String matching
13. Emmett Dennis, Brown BS (2001) Explaining Variations in the Level of Academic Dishonesty in Studies of College Students: Some New Evidence. College Student Journal 35(4):529–538.
14. www.plagiarism.org
15. www.allenschool.edu
16. N. Sandhya, Dr. A. Govardhan, Y. Sri Lalitha, Dr. K. Anuradha (2011 An improved Approach for Document Retrieval Using Suffix Trees. international Journal of Advanced Computer Science and Applications, Vol. 2, No. 9.