مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (38) العدد (1) 2016
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (38) No. (1) 2016

كشف الانتحال في الابحاث الطبية باستخدام الأنطولوجيات الطبية

الدكتور باسل الخطيب*

الدكتورة ميسون دشاش**

خالد عمر ***

(تاريخ الإيداع 7 / 9 / 2015. قُبِل للنشر في 25/ 2 / 2016)

□ ملخّص □

يقدم هذا البحث دراسة مرجعية عن الخوارزميات والأنظمة المتوفرة لكشف الانتحال ، ويقوم بتصميم وبناء تطبيق لكشف الانتحال في الأبحاث الطبية بتوظيف الأنطولوجيات الطبية العالمية المتوفرة على الشبكة العنكبوتية .

إن مسألة كشف الانتحال في الأبحاث الطبية المكتوبة باللغات الطبيعية هي مسألة معقدة وتتعلق بالمجال الدقيق للابحاث الطبية .

يوجد العديد من الخوارزميات المستخدمة لكشف الانتحال في اللغات الطبيعية والتي تقسم بشكل عام إلى صنفين رئيسين هما خوارزميات المقارنة بين الملفات عن طريق بصمات الملفات ،وخوارزميات مقارنة محتوى الملفات والتي تتضمن خوارزميات مقارنة السلاسل النصية وخوارزميات مقارنة البنى الشجرية للملفات .

حديثا تم البحث في مجال خوارزميات كشف الانتحال ذات البعد الدلالي فتم تطوير خوارزميات كشف الانتحال الدلالية المعتمدة على تحليل نماذج الاقتباس في الأبحاث العلمية .

تمَ في هذا العمل تطوير نظام لكشف الانتحال باستخدام محرك البحث Bing ، حيث تم استخدام خوارزمية تعتمد على استخدام وتوظيف نوعين من الانطولوجيات وهي الأنطولوجيات العامة مثل وورد نت (WordNet) والأنطلوجيات الطبية العالمية أشهرها أنطولوجيا الأمراض Diseases ontology التي تحتوي على توصيف الأمراض وخصائصها وتعريفها واشتقاق الأمراض من بعضها.

الكلمات المفتاحية : الوب الدلالي، كشف الانتحال، معالجة اللغات الطبيعية، الأنطولوجيات الطبية

*أستاذ مساعد– كلية الهندسة المعلوماتية – جامعة دمشق– سورية

**مدرسة– كلية طب الأسنان –جامعة دمشق– سورية

***طالب دكتوراه– كلية الهندسة المعلوماتية – جامعة دمشق– سورية

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (38) العدد (1) 2016
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (38) No. (1) 2016

Plagiarism Detection in Medical Research Using Medical Ontology

Dr. Bassel Al-Khatib [*]
Dr. Mayssoon Dashash**
Khaled Omar***

(Received 7 / 9 / 2015. Accepted 25 / 2 / 2016)

□ ABSTRACT □

This paper presents a reference study of available algorithms for plagiarism detection and it develops semantic plagiarism detection algorithm for plagiarism detection in medical research papers by employing the Medical Ontologies available on the World Wide Web.

The issue of plagiarism detection in medical research written in natural languages is a complex issue and related exact domain of medical research.

There are many used algorithms for plagiarism detection in natural language, which are generally divided into two main categories, the first one is comparison algorithms between files by using fingerprints of files, and files content comparison algorithms, which include strings matching algorithms and text and tree matching algorithms.

Recently a lot of research in the field of semantic plagiarism detection algorithms and semantic plagiarism detection algorithms were developed basing of citation analysis models in scientific research.

In this research a system for plagiarism detection was developed using "Bing" search engine, where tow type of ontologies used in this system, public ontology as wordNet and many standard international ontologies in medical domain as Diseases ontology which contains a descriptions about diseases and definitions of it and the derivation between diseases.

Keywords: semantic web, plagiarism detection, natural language processing, medical ontologies.

[*] **Associate Professor- Information Technology Engineering College- Damascus university-Syria.**
****Assistant Professor-Faculty of Dentistry - Damascus university- Syria.**
*****Postgraduate Student -Information Technology Engineering College- Damascus university- Syria.**

مقدمة :

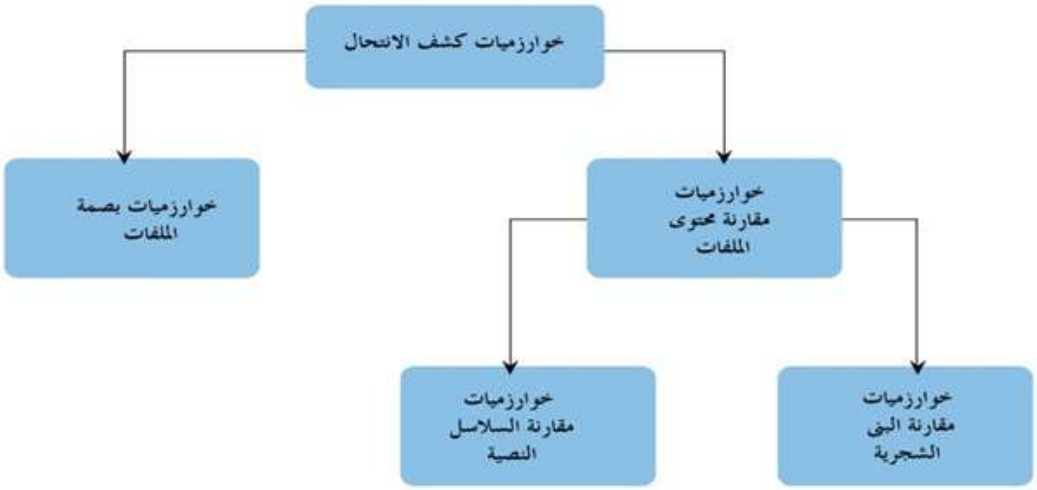
يعرف الانتحال بأنه اعادة استخدام شخص لكتابات وأفكار شخص أو عدة أشخاص آخرين أو لجهد الآخرين بشكل عام ، ونسبها لنفسه سواء بشكل مباشر أو غير مباشر (دون ذكر المصدر) [1]، وقد انتشرت ظاهرة الانتحال بشكل كبير مستفيدة من الانتشار الواسع للمعطيات على شبكة الانترنت ،

وتقوم بعض الجامعات والمؤسسات الأكاديمية بتطوير أنظمة تعتمد على التعليم الإلكتروني حيث تحتوي هذه الأنظمة على قواعد خاصة بكتابة واجبات الطلاب (assignments) وبالتالي فإن وجود مثل هذه الأنظمة يقلل من حدوث الانتحال إلى أكبر قدر ممكن .

وللحد من ظاهرة الانتحال يجب اتباع اسأليب لمنع حدوث هذه الظاهرة حيث أن الانتحال مرض اجتماعي ويجب معالجته بتوعية أفراد المجتمع حول المساوئ التي يسببها انتشار هذه الظاهرة، ووضع قوانين وأسس لضبط هذه الظاهرة، وقد تم تطوير عدة مشاريع على المستوى العالمي بهدف كشف هذه الظاهرة للحد منها، حيث أن انتشار هذه الظاهرة يؤثر بشكل سلبي على جودة البحث العلمي وفعاليته .

وتقسم خوارزميات كشف الانتحال بشكل عام إلى صنفين رئيسين [2]:

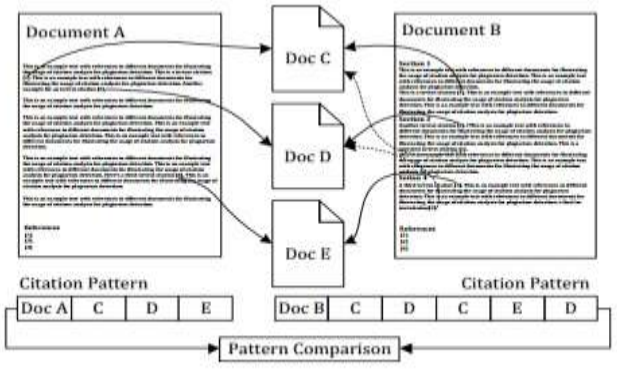
- خوارزميات بصمة الملفات(Fingerprinting) :وهي خوارزميات مقارنة تعتمد على مقارنة بصمة الملفات حيث تقوم بتوليد كود خاص (fingerprint code) لكل ملف حيث يكون هذا الكود وحيد ومميز للملف وتتم المقارنة بين أكواد الملفات الناتجة [3] .
- خوارزميات مقارنة محتوى الملفات(Content comparison) : خوارزميات مقارنة تعتمد على مقارنة محتوى الملفات وتتضمن خوارزميات مقارنة السلاسل النصية (String Matching Algorithms)[4] وخوارزميات مقارنة البنى الشجرية (Tree matching algorithms)[5] كما في الشكل التالي :



الشكل (1) :تصنيف خوارزميات كشف الانتحال [2]

تتأثر خوارزميات بصمة الملفات وخوارزميات مقارنة السلاسل النصية باعادة ترتيب الكلمات حيث أنها تعجز عن اكتشاف الانتحال عند اعادة ترتيب كلمات النص المسروق كما أن خوارزميات مقارنة السلاسل النصية تعاني من مشكلة معامل طول السلسلة الذي تقارن به بين الملفات[6] حيث أن اسناد قيم مختلفة لهذا المعامل يؤدي إلى نتائج مختلفة لذا يجب وضع قيمة هذا المعامل بدقة متناهية وذلك بالاعتماد على طبيعة الملفات التي تقارن بينها وأخذ بعين الإعتبار المجالات التي تقع ضمنها هذه الملفات ،كما انها ايضاً لا تستطيع كشف الحالات التي يتم فيها استبدال بالمرادفات ، وتتأثر خوارزميات مقارنة البنى الشجرية بالغموض الموجود في اللغات الطبيعية بشكل كبير حيث يؤدي وجود هذا الغموض إلى توليد أكثر من شجرة ممثلة لنفس النص.

مع وجود نقاط الضعف في خوارزميات كشف الانتحال التقليدية اتجهت البحوث لتطوير خوارزميات كشف انتحال ذات بعد دلالي (تعتمد بشكل أساسي على تقانات الوب الدلالي) ومن أشهر هذه الخوارزميات خوارزميات كشف الانتحال المعتمدة على تحليل نماذج الاقتباس) (Citation Patterns [7] [8] حيث تقوم هذه الخوارزميات بتحليل المعطيات الموجودة بنماذج الاقتباس في الملفات وتقوم بالمقارنة بين هذه المعطيات كما في الشكل التالي :



الشكل (2) :خوارزميات تحليل نماذج الاقتباس الدلالية [7]

يهدف هذا البحث إلى تطوير خوارزمية لكشف الانتحال ذات بعد دلالي بالاعتماد على الأنطولوجيات العالمية المتوفرة على الشبكة العنكبوتية، حيث ستقوم هذه الخوارزمية بكشف الانتحال في الأوراق البحثية المكتوبة باللغة الإنكليزية في المجال الطبي فقط، وذلك بسبب محدودية الانطولوجيات وعدم توفرها بشكل مجاني بالنسبة لكل المجالات (الهندسية، الأدبية ...)، حيث تقوم هذه الخوارزمية باستخدام نوعين من الأنطولوجيات، أنطولوجيات عامة(public Ontology).

تم في هذا البحث استخدام وورد نت WordNet كأنطولوجية عامة، واستخدام أنطولوجيات طبية تخصصية معتمدة عالمياً (medical domain ontology)في مجال البحوث الطبية.

إن هدف هذا البحث من استخدام الأنطولوجيات الدلالية semantic ontology هو التغلب على الصعوبات والتخلص من نقاط الضعف التي تعاني منها خوارزميات كشف الانتحال التقليدية، حيث أن الأنطولوجيات الدلالية تحتوي على معلومات قيمة وغنية للمفاهيم حيث تتضمن تعاريف المفاهيم والمفاهيم التي لها نفس المعنى (المفاهيم المشتقة) ومعلومات أخرى تختلف من أنطولوجية طبية إلى أخرى.

منهجية البحث:

يقوم هذا البحث باستخدام الورد نت WordNet كانطولوجيا عامة وعدد من أشهر الانطولوجيات الطبية العالمية وهي:

1- الأنطولوجيا الخاصة بتشريح جسم الانسان (Anatomy ontology)

2- الأنطولوجيا الخاصة بتوصيف الأمراض (Diseases ontology)

3-الأنطولوجيا الخاصة بعلم الجينات (Gene ontology)

4-أنطولوجيا رؤوس الموضوعات الطبية (Mesh ontology)

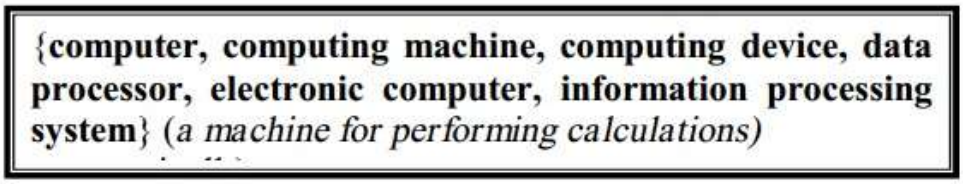
5- الأنطولوجيا الخاصة بالعلوم الطبية بشكل عام (ontology for general medical science)

6-الأنطولوجيا الطبية التي تحتوي على تكامل المعطيات والاجرائيات (EDAM ontology)

وفيمايلي شرح مبسط عن كل أنطولوجية من الأنطولوجيات السابقة

الورد نت[9] [10]

هي قاعدة بيانات معجمية كبيرة في اللغة الإنجليزية. يتم تجميع الأسماء، والأفعال، والصفات والظروف في مجموعات من المترادفات (synsets)، وتترابط هذه المفاهيم بعلاقات دلالية .وتستخدم قاعدة البيانات وورد نت بشكل كبير في اللغويات الحاسوبية ومعالجة اللغة الطبيعية، العلاقة بين الكلمات الرئيسية في وورد نت هو الترادف . ويبين الشكل التالي يبين تمثيل المجموعة ضمن وورد نت [11]



الشكل (3) : تمثيل المجموعة ضمن WordNet [11]

1- الأنطولوجيا الخاصة بتشريح جسم الانسان (Anatomy ontology)[12]

تم تطوير هذه الأنطولوجية في جامعة واشنطن لتحسين المحتوى التشريحي لنظام اللغة الطبي الموحد UMLS وذلك بالتركيز بشكل صريح على تمثيل البنية. وتستخدم هذه الأنطولوجية كأنطولوجية مرجعية حيث أنها يمكن أن تستخدم في الأنطولوجيات الأخرى التي تسخدم مفاهيم التشريح ويمكن أن تتكامل معها، تحتوي على 70.000 مفهوم ويتم تمديدها لتشمل الخلايا وأجزاء الخلايا ، وتم تنجيز هذه الانطولوجية باستخدام الأداة (Protégé4) و هي أداة لبناء وتعديل الأنطولوجيات وقد تم تطويرها في جامعة ستان فورد.

2- الأنطولوجيا الخاصة بتوصيف الأمراض (Diseases ontology)[13]

تم تطوير أنطولوجية توصيف الأمراض بغرض تزويد مجتمع الطب الحيوي بوصف ثابت ودقيق و قابل لإعادة الاستخدام للأمراض التي تصيب البشر، ولتزويد المجتمع بالخصائص الظاهرية وما يتصل بها من مفاهيم المرض وقد تم تطوير هذه الأنطولوجيا عن طريق مجموعة من الباحثين في جامعة نورث وسترن، مركز طب الجينات وكلية الطب ومعهد للعلوم الجينية في جامعة ميريلاند.

وتدمج أنطولوجيا الأمراض المرض والمفردات الطبية من خلال استخدام أنطولوجيات أخرى لربط المفاهيم الطبية مع بعضها مثل انطولوجيا التصنيف العالمي للأمراض ICD وأنطولوجيا رؤوس الموضوعات الطبية Mesh .

3-الأنطولوجيا الخاصة بعلم الجينات (Gene ontology)[14]

تم تطوير ثلاث أنطولوجيات مترابطة في هذه الأنطولوجيا وهي : أنطولوجيا وصف منتجات الجينات في سياق العمليات البيولوجيا المرتبطة بها، والمركبات الخلوية، والعمليات الجزيئية بطريقة مستقلة ودقيقة.

يهدف مشروع أنطولوجيا الجينات ليكون بمثابة منصة حيث يمكن المشرفين على المشروع على الاتفاق على كيفية و سبب استخدام مصلح معين ، وكيفية تطبيق استخدامه على الدوام، على سبيل المثال، لإقامة علاقات بين المنتجات الجينية، وتستخدم انطولوجيا الجينات في عدة مجالات منها منتجات الجينات و العمليات والوظائف أو المكونات التي هي فريدة من نوعها لطفرات أو أمراض: مثل تكون الورم ليس جين سليم وإنما سببه خلايا غير سليمة أدت لظهوره، مجالات البروتين أو الصفات الهيكلية، تفاعلات البروتين –البروتين،البيئة والتطور، الخصائص التشريحية أو النسيجية فوق مستوى المكونات الخلوية، بما في ذلك أنواع الخلايا.

4-أنطولوجيا رؤوس الموضوعات الطبية (Mesh ontology) [15]

تحتوي أنطولوجيا رؤوس الموضوعات الطبية على عناوين المواضيع الطبية وتتألف من مجموعات الواصفات في بنية هرمية الأمر الذي يسمح بالبحث في مستويات مختلفة في العناوين الطبية.

يتم ترتيب الواصفات في هذه الأنطولوجية بشكل أبجدي وهرمي، وفي المستوى العام الأعلى للبنية يوجد العناوين العريضة جدا مثل "التشريح" أو "الاضطرابات العقلية." وبالنزول إلى المستويات الضيقة(السفلى في الشكل الهرمي)يتم العثور على عناوين أكثر تحديدا ، مثل "الكاحل" و "اضطراب السلوك."

هناك 27455 واصفة في هذه الأنطولوجية في عام 2015. هناك أيضا أكثر من 220،000 مصطلح لتساعد في العثور على العناوين الأنسب للمواضيع الطبية.

وتستخدام عناوين المواضيع الطبية لفهرسة المقالات من 5400 من المجلات الطبية الرائدة في العالم لقاعدة البيانات بابميد وميدلاين (MEDLINE / PubMED)، كما أنها تستخدم في فهرسة العديد من قواعد البيانات الطبية العالمية، وهي متوفرة بشكل مقروء آليا على الشبكة العنكبوتية حيث ابتداء من يونيو عام 2015، أصبح نسخة RDF من أنطولوجيا رؤوس الموضوعات الطبية متاحة بشكل مجاني على الوب.

5- الأنطولوجيا الخاصة بالعلوم الطبية بشكل عام (ontology for general medical science) [16]

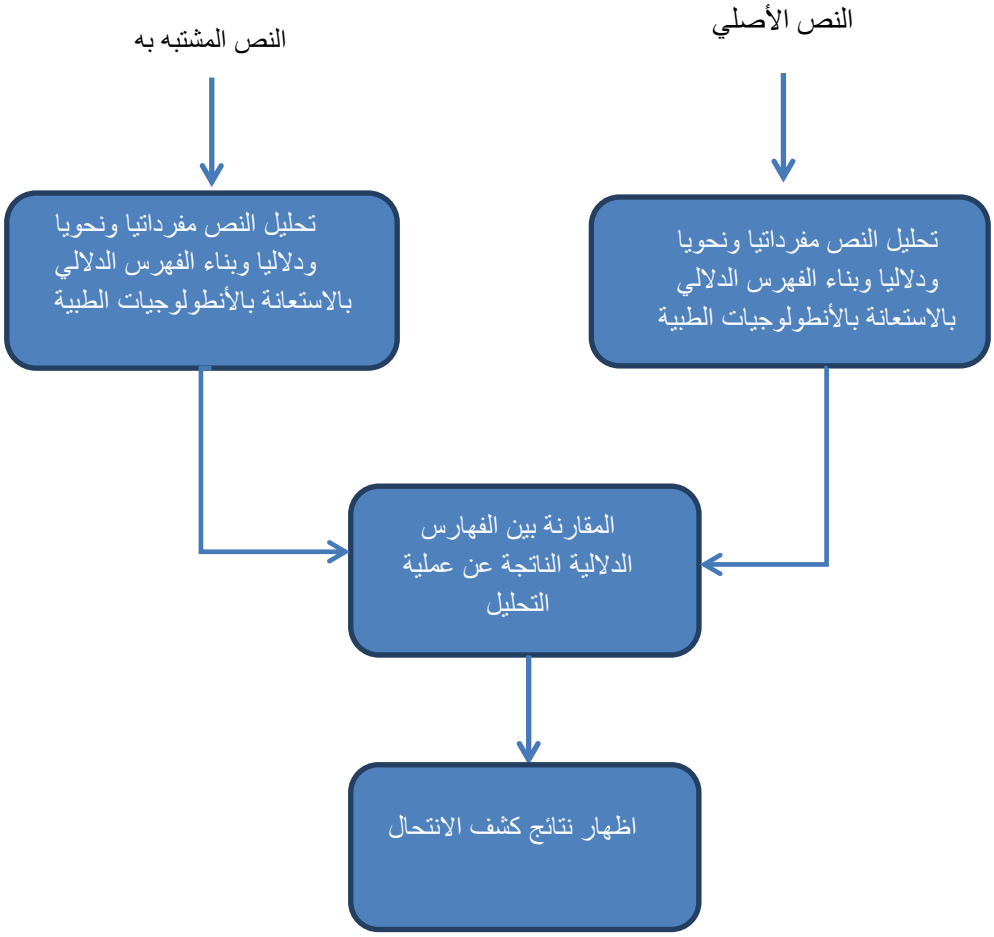
تستند الأنطولوجية الخاصة بالعلوم الطبية بشكل عام (OGMS) على الأوراق البحثية وذلك بهدف تطوير وبناء أنطولوجية للأمراض والتشخيص لمرض السرطان والأمراض الأخرى.

6-الأنطولوجيا الطبية التي تحتوي على تكامل المعطيات والاجرائيات (EDAM ontology) [17]

يأتي اسم هذه الأنطولوجية أصلا من "التشابك بين البيانات و الطرائق، تحتوي على المفاهيم المألوفة التي تقع ضمن مجال المعلوماتية الحيوية، بما في ذلك أنواع ومعرفات البيانات ، نماذج البيانات والعمليات والمواضيع – وتتألف بشكل رئيسي من مجموعة من المفاهيم مع المرادفات والتعاريف وقد بنيت بشكل هرمي سلس لتوفير سهولة في التعامل معها من قبل مطوري البرمجيات والمستخدمين النهائيين.

خطوات خوارزمية كشف الانتحال الدلالية :

الآن بعد أن تم تقديم شرح مبسط عن الانطولوجيات الطبية المستخدمة في هذا البحث سوف نقوم بشرح خوارزميتنا المقترحة لكشف الانتحال المعتمدة على الانطولوجيات الطبية كما في الشكل التالي :



الشكل (4) : خطوات خوارزمية كشف الانتحال الدلالية

وفيمايلي شرح كل خطوة بالتفصيل :

- تحليل النص تحليلا مفرداتي ونحوي ودلالي وبناء الفهرس الدلالي بالاستعانة بالأنطولوجيات الطبية (لكل من النص الأصلي والنص المشتبه به): يظهر الشكل التالي خطوات هذه المرحلة من خوارزمية كشف الانتحال الدلالية:

الشكل (5) : خطوات تحليل النص وبناء الفهارس الدلالية

في هذه المرحلة يتم تحليل النص تحليلا مفرداتيا ونحويا ودلاليا حيث تقوم الخوارزمية أولا بازالة الكلمات غير المفيدة من النص (stop – words) بالإستعانة بقائمة الكلمات غير المفيدة في اللغة الإنكليزية ، ثم تقوم الخوارزمية بتحديد الأسماء والأفعال في النص كما تقوم بتحديد علاقات الأسماء مع بعضها مثل الصفة والموصوف يتم ذلك بالإستعانة بأدوات ستانفورد لمعالجة اللغات الطبيعية (Stanford NLP tools).

بعد أن يتم الحصول على الأسماء والأفعال وتحديد العلاقات بين الأسماء (يتم استخدام أدوات ستنافورد لمعالجة اللغات الطبيعية لتحديد العلاقات بين الأسماء) حيث تقوم الخوارزمية بتحديد الأسماء المركبة التي تعبر عن مفهوم أو مفاهيم مكونة من أكثر من كلمة واحدة في النص.

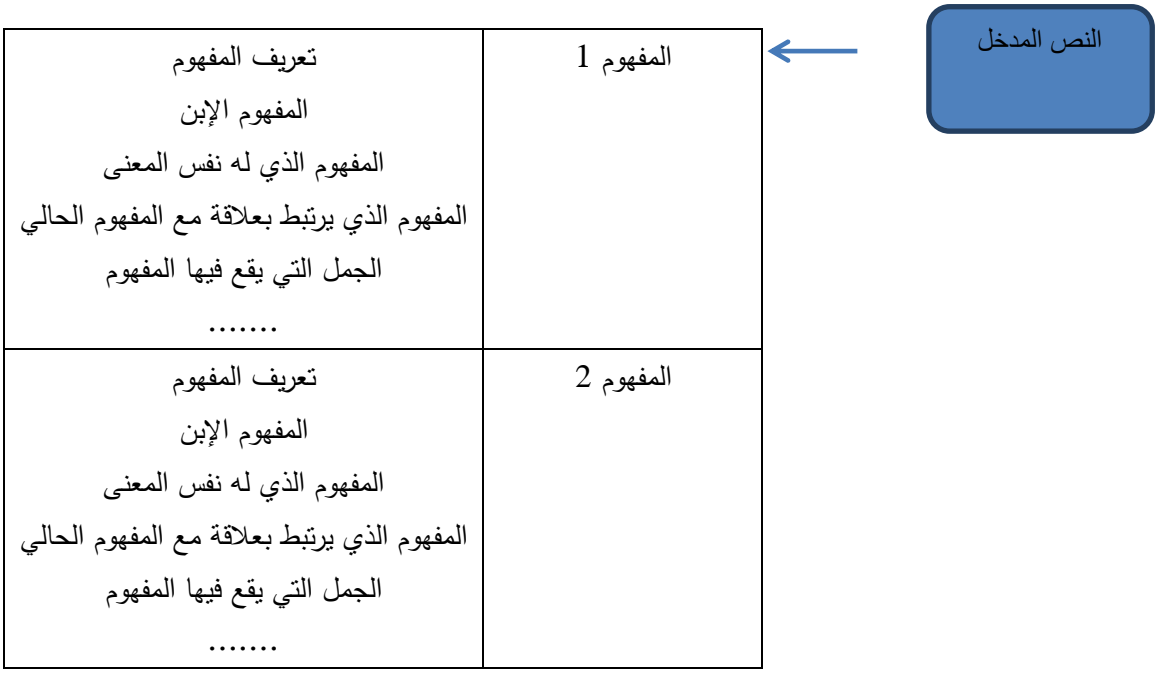
بعد الحصول على اسماء المفاهيم المكونة من كلمة واحدة أو المكونة من عدة كلمات يتم البحث عن معاني هذه المفاهيم بالأنطولوجيا العامة بشكل أولي ثم بالأنطولوجيا الطبية التخصصية (يتم البحث عن المفهوم الطبي في جميع الأنطولوجيات الطبية الآنفة الذكر).

بعد أن يتم تحديد المفاهيم المفردة والمركبة، يتم بناء فهرسين للنص :

1- فهرس خاص بالمفاهيم (concept index): حيث يحتوي هذا الفهرس على مفاتيح وقيم ، يكون مفتاح كل عنصر من هذا الفهرس هو المفهوم وقيمة هذا العنصر هي المعلومات المرتبطة بهذا المفهوم والمستخرجة من الأنطولوجيا حيث أنه تم بناء طبقة دلالية (semantic layer) للتخاطب مع الأنطولوجيات المستخدمة واستحصال المعلومات المطلوبة عن المفايهم(تحتوي الأنطولوجيات المستخدمة على مجموعة كبيرة من المعلومات والخصائص للمفاهيم وفي هذه الخوارزمية تم الإكتفاء باستحصال بعض المعلومات كتعريف المفهوم والمفهوم الإبن له والمفاهيم المشابهة بالمعنى) كما تتم إضافة معلومات أخرى تحدد مكان المفهوم في النص والجمل التي يقع فيها المفهوم وعدد ورود هذا المفهوم في النص .

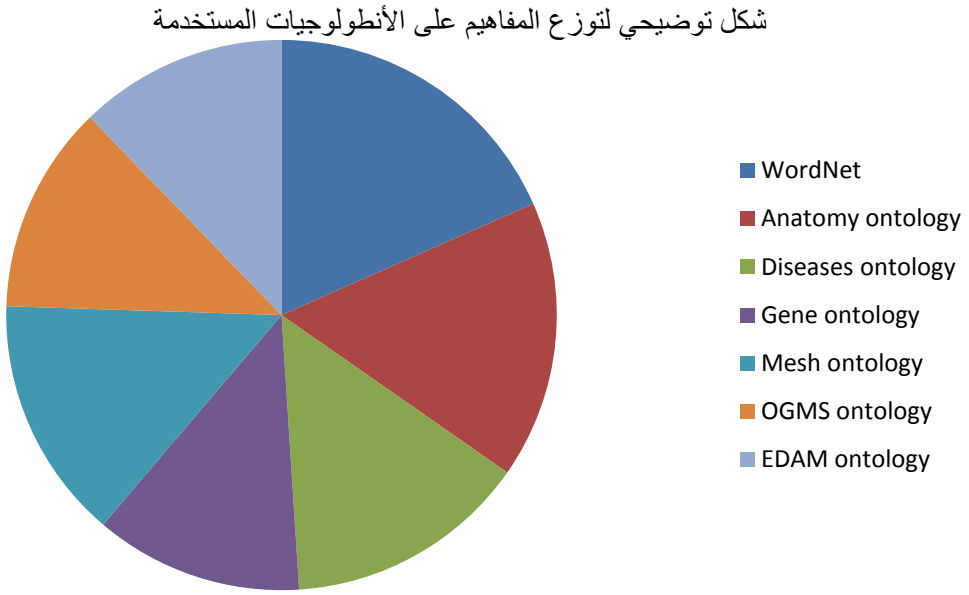
2- فهرس خاص بالأنطولوجيات (ontology index): حيث تمثل المفاتيح أنواع الأنطولوجيا في هذا الفهرس وتمثل المفاهيم التي تنتمي لهذه الأنطولوجيات قيم هذا الفهرس .

بعد الانتهاء من هذه المرحلة من الخوارزمية أصبح كل من النص الأصلي والنص المشتبه به ممثل بفهرسين ويحتوي كل من هذين الفهرسين على المفاهيم الدلالية وقيمها المتعلقة بها وبالإضافة لجمل النص الواقعة ضمنها هذه المفاهيم الدلالية، يوضح الشكل التالي تمثيل النص بعد معالجته وبناء الفهارس الخاصة به:



. . .	. . .
تعريف المفهوم المفهوم الإبن المفهوم الذي له نفس المعنى المفهوم الذي يرتبط بعلاقة مع المفهوم الحالي الجمل التي يقع فيها المفهوم	المفهوم n

الشكل (6) : تمثيل النص بفهرس المفاهيم بعد الانتهاء من مرحلة التحليل الدلالي



الشكل (7) : تمثيل النص بفهرس الأنطولوجيات بعد الانتهاء من مرحلة التحليل الدلالي

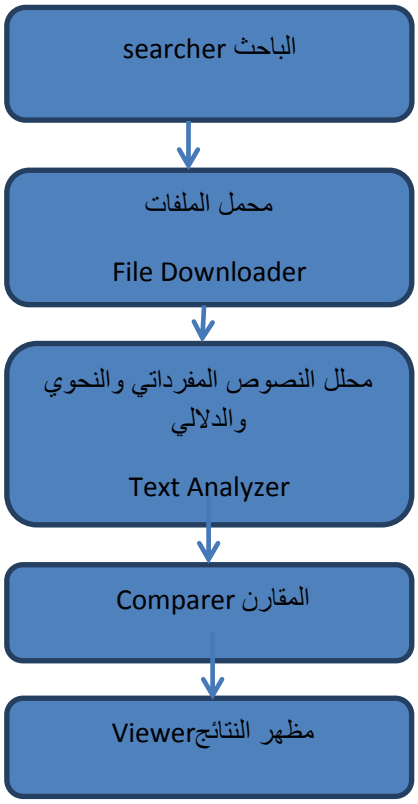
- المقارنة بين الفهارس الدلالية الناتجة عن عملية التحليل :في هذه المرحلة يتم المقارنة بين الفهارس الدلالية المتولدة عن المرحلة السابقة حيث يتم المرور على كل المفاهيم في النص الأصلي به ومقارنتها بالمفاهيم الموجودة في الفهارس الدلالية للنص المشتبه به،وتتم المقارنة على مرحلتين أو تتم المقارنة بين أعداد المفاهيم في الفهارس الخاصة بالأنطولوجيات بين الملفات حيث اذا كان عدد المفاهيم المشتركة بين فهارس الأنطلوجيات أصغر من عتبة معينة (تم اعتماد قيمة العتبة =2 مفهومين لكل أنطولوجية) يتم استبعاد الملف الحالي ولا يتم استكمال عملية المقارنة لأنه في هذه الحالة يكون الملف قيد المقارنة والذي تم الحصول عليه من نتائج محرك البحث المستخدم "Bing" (الذي سيتم الشرح عنه لاحقا في بنية النظام المطور) خارج نطاق الأنطولوجيات الطبية المستخدمة في هذا البحث، أما اذا كان عدد المفاهيم المشتركة بين الأنطولوجيات (في الملفين قيد فحص كشف الانتحال بينهما) أكبر من العتبة المحددة يتم الانتقال الى المرحلة الثانية من عملية المقارنة حيث تتم المقارنة بين فهارس المفاهيم عن طريق مقارنة المفهوم

وخصائصه المرتبطة به كالمفهوم الإبن للمفهوم وتعريفه وأيضاً المرادفات لهذا المفهوم اذا كان مفهوم عام (أي ينتمي إلى الأنطولوجية العامة المستخدمة في هذا البحث وهي وورد نت) والمفهوم الابن له اذا كان المفهوم طبي تخصصي (ينتمي لأحدى الأنطولوجيات الطبية المستخدمة).

- إظهار نتائج كشف الانتحال : يتم في هذه المرحلة اظهار الجمل التي تحتوي على مفاهيم متشابهة دلاليا وذلك حسب نتائج المقارنة بين الفهارس الدلالية في المرحلة السابقة، حيث يتم وسم الجمل التي تحتوي على مفاهيم مشتركة ومتقاربة دلاليا بلون مختلف ليتسنى للمستخدم ملاحظة كشف الانتحال بسهولة وبدون تتبع دقيق وقراءة للملفات المشتبهة بشكل كامل .

تصميم النظام :

يتألف النظام المطور من خمس مكونات برمجية حيث تم تطوير هذا النظام باستخدام لغة البرمجة جافا Java وتم استخدام NetBeans كبيئة تطوير برمجي وتم استخدام قاعدة معطيات من نوع SqlServer لحفظ البيانات الخاصة بالنظام المطور ويوضح الشكل (8) أدناه بنية النظام المطور.



الشكل (8) : بنية النظام المطور

- الباحث (Searcher) يستخدم هذا المكون البرمجي خدمة البحث التي يقدمها محرك البحث " Bing" Bing search web service [18] ويعيد نتائج البحث الخاصة بالإستعلام الذي أدخله المستخدم (الاستعلام يتم توليده بشكل اتوماتيكي من ملف الدخل الذي نريد فحص كشف الانتحال به حيث يتكون هذا الاستعلام من المفاهيم والكلمات المفتاحية الأكثر ورودا في النص)، يقوم الباحث بالبحث عن الملفات من نوع doc و pdf.

- محمل الملفات (File Downloader) : يقوم بتحميل نتائج البحث من الشبكة العنكبوتية ويخزنها في قاعدة المعطيات الخاصة بالنظام.
- محلل النص المفرداتي والنحوي والدلالي : يقوم هذا الجزء بتحليل النص المدخل تحليلا كاملا وقد تم استخدام ادوات ستانفورد لمعالجة اللغات الطبيعية (Stanford NLP Tools وStanford Stop List و Stanford Part of Speech Tagger) [19] و في التحليل الدلالي تم استخدام مكتبة التخاطب البرمجية لبيئة العمل في الوب الدلالي جينا (Jena Api Ontology library) [20] للحصول على المعلومات الدلالية الموجودة بالأنطولوجيات المستخدمة في هذا البحث حيث يحتوي هذا الجزء البرمجي من النظام المطور على طبقة دلالية تؤمن الإتصال واستحصال المعلومات من الأنطولوجيات الدلالية.
- المقارن :يقوم هذا المكون البرمجي بالمقارنة بين فهارس الملفات بعد أن تم تحليلها مفرداتيا ونحويا ودلاليا.
- مظهر النتائج : يقوم هذا المكون البرمجي باظهار نتائج كشف الانتحال حيث يقوم بتحديد الجمل التي تحتوي على مفاهيم منتحلة بلون مختلف لكي يستطيع مستخدم النظام ملاحظة الاجزاء المنتحلة من الملف المشتبه به.

النتائج والمناقشة :

تم اختبار الخوارزمية المطورة في هذا البحث على 200 ورقة علمية ضمن اختصاص الطب وبشكل خاص ضمن اختصاصات الانطولوجيات المستخدمة في هذا النظام المطور وتم استحصال الأوراق العلمية من موقع بابميد الأوروبي(Europe PubMed Central) [21] إلى استخدام محرك البحث "Bing" حيث يتيح يعد هذا الموقع غني جدا بالابحاث الطبية العلمية ويتيح مئات الألوف من الورقات العلمية الطبية بشكل مجاني.

وتم انتحال العديد من الورقات العلمية المستحصلة من الموقع المذكور بشكل يدوي ولكن بطرق مختلفة باستخدام أنواع متعددة للانتحال من تغيير الكلمات بمرادفاتها إلى انتحال أجزاء من الجمل إلى استبدال المفاهيم الطبية بمفاهيم مشابهة لها في المعنى أو مشتقة منها وتم تحديد فعالية الخوارزمية المطورة بالاعتماد على معامل الحساسية) [22] الذي يحسب من العلاقة التالية :

$$\text{معامل الحساسية} = \frac{\text{عدد حالات كشف الانتحال الصحيحة}}{\text{عدد الملفات المشبوهة الاجمالي}}$$

وفيمايلي مزايا الخوارزمية المقترحة المطورة في هذا البحث :

- مرحلة التحليل وبناء الفهارس الدلالية : تقوم الخوارزمية بتحليل وبناء الفهارس الدلالية للنص المشتبه به مرة واحدة فقط وتقوم بتحليل النصوص المعادة من نتائج محرك البحث لمقارنتها مع الملف المراد اختبار كشف الانتحال فيه.
- مرحلة المقارنة : تتميز الخوارزمية المتبعة عن باقي أنواع هذه الخوارزميات بسرعتها في مرحلة المقارنة حيث أنها أسرع من جميع أنواع الخوارزميات لأنها في مرحلة المقارنة تقارن فقط بين مفاهيم الفهارس الدلالية، حيث أن خوارزميات كشف الانتحال المعتمدة على السلاسل النصية تستغرق وقتاً M * N وحدة زمن مقارنة بين ملفين حيث أن

N هو طول سلسلة الملف الأول وM طول سلسلة الملف الثاني، وكذلك خوارزميات كشف الانتحال المعتمدة على توليد الأشجار النحوية الممثلة للنصوص تستغرق وقتاً أطول من خوارزميتنا المطورة في هذا البحث.

- غير مكلفة زمنيا حيث أن الخوارزمية تحتاج إلى N * M وحدة زمن للمقارنة بين ملفين حيث أن N هو عدد مفاتيح الفهرس في الملف الأول و M هو عدد مفاتيح الفهرس في الملف الثاني، حيث أن خوارزميات بصمات الملفات تعاني من البطء الشديد في حالة الملفات الكبيرة وأيضاً لا يمكنها كشف استبدال المرادفات ، وكذلك خوارزميات مقارنة السلاسل النصية فيوجد فيها مشكلة أساسية هي مشكلة طول السلسلة التي تقارن بها بين الملفات، وأيضاً خوارزميات مقارنة البنى الشجرية على الرغم من قدرتها على كشف استبدال المرادفات فإنها تحتاج إلى وقت كبير لتوليد الأشجار التي تمثل الجمل في الملفات.
- لا تقوم الخوارزمية بتغيير بنية الملفات التي تقارن بينها ، فهي تحافظ على بنى الجمل الموجودة ضمن هذه الملفات وهذا يساعد كثيراً في مرحلة إظهار نتائج التشابه .
- تحافظ الخوارزمية على المعنى الدلالي الموجود في النص حيث تقوم باستحصال المفاهيم الدلالية المفردة والمركبة (من أكثر من كلمة).
- قابلة للتمديد لتطبق على اللغة الإنكليزية بكل مجالاتها لأنه يوجد العديد من مصادر الأنطولوجيات المطورة باللغة الإنكليزية والمعتمدة عالميا حيث أن الخوارزمية المطورة في هذا النظام قابلة للتمديد على أنطولوجيات جديدة بمجرد إضافة هذه الأنطولوجيات وتحديد الصفات التي سيتم استخدامها من هذه الانطولوجيات.
- إن عمل الخوارزمية في تحديد بنى الجمل وحدودها لا يوجد فيه أي التباس أو غموض بالمقارنة مع خوارزميات معالجة النصوص التي تعتمد على نظرية البنى الشجرية والتي تقوم بتحويل النص إلى بنى شجرية مقابلة حيث يمكن أن يكون للجملة الواحدة أكثر من شجرة مقابلة وينتج عن ذلك غموض والحاجة إلى استخدام تجريبيات لفك هذا الغموض.
- لا تتأثر الخوارزمية بإعادة ترتيب الكلمات وباستبدال الكلمات بمرادفاتها حيث أنها تقوم بالإستعانة بالوورد نت وبالأنطولوجيات الطبية الأخرى لحل هذه المشكلة بينما خوارزميات المقارنة التي تعتمد على بصمات الملفات تتأثر تأثر كبير باستبدال الكلمات بمرادفاتها وكذلك الأمر بالنسبة لخوارزميات مقارنة السلاسل النصية .

مما سبق نستنتج أن الخوارزمية المطورة في هذا البحث هي خوارزمية جيدة وفعالة وأهم ما يميزها هو سرعتها وقدرتها على كشف الانتحال في حال استبدال الكلمات والمفاهيم بمرادفاتها وعدم تغيير بنية الملفات التي تقارن التي تقوم بفحص كشف الانتحال فيما بينها، حيث أثبتت التجارب أن نسبة معامل الحساسية للخوارزمية المطورة في كشف حالات الانتحال هو 70%، (مع الإشارة إلى أن الخوارزمية المطورة تقل فعاليتها عند وجود الأشكال والجداول في النص حيث لا تتكمن الخوارزمية المطورة من استحصال المعلومات الدلالية الموجودة ضمن هذه الصور والأشكال).

الاستنتاجات والتوصيات :

قمنا بهذا البحث بتطوير نظام دلالي لكشف الانتحال بالاستعانة بمحرك البحث "Bing" ويعتمد هذا النظام على خوارزمية كشف انتحال دلالية فعالة مختصة في كشف الانتحال بالابحاث الطبية وتتميز هذه الخوارزمية بالسرعة والفعالية.

من خلال التجارب والاختبارات التي أُجريت على خوارزمية كشف الانتحال المتبعة في هذا النظام وجدنا أن هذه الخوارزمية قللت من الاختبارات غير المجدية بين نصوص الملفات التي تقوم بفحص الانتحال فيما بينها ويمكن مستقبلا ان يتم توسيع الأنطولوجيات الطبية المستخدمة لتصبح الخوارزمية قادرة على كشف الانتحال في مجال العلوم الطبية بأكمله (يتم توسيع الأنطولوجيا إما بإضافة أنطولوجيات طبية جديدة على النظام المطور أو باستخدام خوارزميات تعلم الأنطولوجيات من المصادر المفتوحة بشكل أتوماتيكي.

المراجع :

1. Vinod K.R.،Sandhya.S،Sathish Kumar D،Harani A،David Banji and،Otilia JF Banji 2011 ، *Plagiarism history ،detection and prevention*, Hygeia, Vol.3-Issue.1-Page 1-4.
2. Maxim mozgovoy, *enhancing computer-aided plagiarism* , university of joensuu computer science and statistics dissertations 18.
3. Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). *Winnowing: Local Algorithms for Document Fingerprinting*. Proceedings of the 2003 ACM SIGMOD International Conference on on Management of Data - SIGMOD '03, 76–85.
4. Shivaji, S. K. (2015). *Plagiarism Detection by using Karp-Rabin and String Matching Algorithm Together*, International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 23, April 2015
5. Mozgovoy, M., Kakkonen, T., & Sutinen, E. (2007). *Using natural language parsers in plagiarism detection*. Proceedings of SLaTE' 07 Workshop. Farmington. Pennsylvania, 7–9. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.320&rep=rep1&type=pdf
6. Clough, P., Studies, I., & Street, P. (2003). *Old and new challenges in automatic plagiarism detection*, University of Sheffield,UK,Plagiarism Advisory Service (February) 2003.
7. Gipp, B., & Meuschke, N. (2011). *Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence*. Proceedings of the 11th ACM Symposium on Document Engineering, 249–258. http://doi.org/10.1145/2034691.2034741
8. Citation-based Plagiarism Detection. [online] Available at: http://www.sciplore.org/projects/citation-based-plagiarism-detection/ [Accessed 15-march 2015].
9. RDF/OWL Representation of WordNet . [online] Available at: http://www.w3.org/TR/wordnet-rdf/ [Accessed 1-may 2015].
10. WordNet a lexical database for English . . [online] Available at: https://wordnet.princeton.edu/ [Accessed 5-may 2015].
11. ابراهيم ، ربى طلال.ذنون ، باسل يونس. *طريقة مقترحة للكشف عن تشابه النصوص في الوثائق الإنكليزية باستخدام الشبكات الدلالية* .المجلة العراقية للعلوم الإحصائية (25) 2013، عدد خاص بوقائع المؤتمر العلمي السادس لكلية علوم الحاسوب والرياضيات ، ص ص [456 – 473].
12. Foundational model of anatomy. [online] Available at: http://sig.biostr.washington.edu/projects/fm/ [Accessed 20-may 2015].
13. Diseases Ontology . [online] Available at: http://disease-ontology.org/ [Accessed 25-may 2015].

14. Gene Ontology Consortium**.** Available at: http://geneontology.org/ [Accessed 20-may 2015].
15. U.S. National library of Medicine. [online] Available at: http://www.nlm.nih.gov/mesh/ [Accessed 25-may 2015].
16. The Open Biological and Biomedical Ontologies. [online] Available at: http://www.obofoundry.org/cgi-bin/detail.cgi?id=OGMS [Accessed 25-may 2015].
17. EDAM Ontology**.** [online] Available at: http://edamontology.org/page [Accessed 25-may 2015].
18. Bing, API Basics. [online] Available at: http://www.bing.com/developers/s/APIBasics.html [Accessed 15-October 2015].
19. The Stanford Natural Language Processing Group. [online] Available at: http://nlp.stanford.edu/software/ [Accessed 15-October 2015].
20. Apache Jena. [online] Available at: https://jena.apache.org/ [Accessed 15-October 2015].
21. Europe PubMed Center. [online] Available at: http://europepmc.org/ [Accessed 15-October 2015].
22. Alzahrani, S. M., & Salim, N. (2008). *Plagiarism Detection In Arabic Scripts Using Fuzzy Information Retrieval*, Proceedings of 2008 Student Conference on Research and Development (SCOReD 2008), 26-27 Nov. 2008, Johor, Malaysia