كشف الانتحال في اللغة العربية باستخدام نظرية بنية الكلام البلاغية

م. خالد عمر* **د. باسل الخطيب****

الملخص

قَدَّمَ هذا البحث دراسة مرجعية عن الخوارزميات والنظم المتوافرة لكشف الانتحال، إِذْ صُمِّمَ وبُنِيَ تطبيق لكشف الانتحال باستخدام محركات البحث المتوافرة على الشبكة العنكبوتية.

إن مسألة كشف الانتحال في الوثائق المكتوبة باللغات الطبيعية هي مسألة معقدة وتتعلق بخصائص اللغة الطبيعية المعنية نفسها.

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

تعتمد نظم كشف الانتحال على نوع محدد أو مزيج من خوارزميات كشف الانتحال؛ وذلك للحصول على نظم كشف انتحال فعالة (سريعة ودقيقة).

طُوِّرَ في هذا العمل نظام لكشف الانتحال باستخدام محرك البحثBing، وقد اِسْتُخْدِمَتْ خوارزمية تعتمد على خصائص اللغة، باستخدام نظرية بنية الكلام البلاغية(Rhetorical Structure Theory.

الكلمات المفتاحية: كشف الانتحال في اللغة العربية، نظرية بنية الكلام البلاغية، معالجة اللغات الطبيعية.

* كلية الهندسة المعلوماتية- جامعة دمشق
**كلية الهندسة المعلوماتية- جامعة دمشق

1- المقدمة:

يُعَرف الانتحال بأنه إعادة استخدام شخص لكتابات وأفكار شخص آخر أو عدة أشخاص آخرين –جهد الآخرين بشكل عام ونسبها لنفسه (دون ذكر المصدر) وقد يكون انتحال المحتوى من لغة إلى أخرى [1]، [3]، [8].

بدأ البحث في مجال كشف الانتحال في سبعينيات القرن الماضي، إذْ اِقْتُرِحَ عدد كبير من الخوارزميات والطرائق لتحديد التشابه غير العادي بين وظائف الطلاب البرمجية، وقد اتجهت جهود الباحثين والعاملين مؤخراً في مجال اللغات الطبيعية لتحديد أوجه التشابه بين نصوص اللغات الطبيعية، لكن الأمر لم يكن سهلاً؛ وذلك لعدة أسباب أهمها:

- الغموض الموجود في اللغات الطبيعية.
- عدد المفردات غير المحدود الموجود في اللغات الطبيعية.

ويوجد العديد من نظم كشف الانتحال التي تستخدم خوارزميات مختلفة ومتنوعة (هناك خوارزميات تقوم بمحاولة كشف الانتحال في عدة مستويات [10]) ،وفي هذا البحث طَوَّرْنا خوارزمية كشف انتحال تعتمد على نظرية بنية الكلام البلاغية (Rhetorical structure theory) التي تقوم بإيجاد العلاقات الموجودة بين أجزاء النص، لأنَّ هذه النظرية قابلة للتطبيق في اللغتين العربية والإنكليزية.

تحتوي هذه الورقة البحثية على دراسة مرجعية تتضمن خوارزميات كشف الانتحال بشكل عام ونقاط الضعف فيها، ومن ثم على شرح مفصل عن نظرية بنية الكلام البلاغية واستخدامها في خوارزمية كشف الانتحال المتبعة في هذا البحث، ومن ثم عرض كيفية تصميم النظام والاختبارات والنتائج، وأخيراً المراجع.

2- خوارزميات كشف الانتحال بشكل عام:

يُمكن بشكل عام أن نصنف خوارزميات كشف الانتحال ضمن صنفين رئيسين كما يأتي [3]:

- خوارزميات بصمة الملف(Fingerprinting) التي تقوم بتوليد بصمات للملفات للمقارنة فيما بينها.
- خوارزميات مقارنة المحتوى(Content comparisons) التي تعتمد على مقارنة محتويات الملفات.

2-1- خوارزميات المقارنة التي تعتمد على بصمة الملفات: تقوم بتكوين ترميز (code) للملف يسمى بصمة الملف، وتوجد عدة طرائق للحصول على بصمة الملف، إذْ من المفترض أن تُعَرف البصمة الملف بشكل وحيد، ومن أشهر خوارزميات بصمة الملف خوارزمية winnowing [11] [3] لأنَّ البصمة تتكون من رمز كودي (hashing code) ولها عدة أنواع بصمة على مستوى الأحرف وبصمة على مستوى الكلمات وبصمة على مستوى الجمل، إذْ تقوم الخوارزمية بتوليد بصمات الملفات، ومن ثم تقوم بالمقارنة بين هذه البصمات باستخدام خوارزمية السلسلة المشتركة الأطول (longest common string)، إن خوارزمية كشف الانتحال المعتمدة على بصمة الملفات هي خوارزمية ضعيفة لأنها تتأثر تأثراً كبيراً بإعادة ترتيب الكلمات في النص، كما لا يمكنها كشف الاستعاضة عن الكلمات بمرادفاتها لذا يجب الاعتماد على خوارزميات تأخذ الدلالة بالحسبان [11].

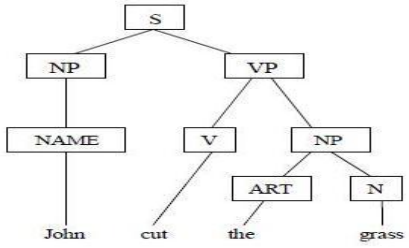
2-2-خوارزميات المقارنة التي تعتمد على مقارنة محتوى الملفات: وتتضمن خوارزميات مقارنة السلاسل النصية وخوارزميات مقارنة الأشجار الناتجة عن تحليل النصوص.

2-2-1-خوارزميات مقارنة السلاسل النصية (String matching algorithms)

تقوم هذه الخوارزميات بالمقارنة بين محتويات الملفات؛ وذلك بمقارنة السلاسل النصية الموجودة ضمن الملفات، ولكن يوجد بعض السلبيات أو نقاط الضعف عند استخدام

هذا النوع من الخوارزميات وأهمها مشكلة تطابق التقسيم (split match problem) [17]، وتحدث هذه المشكلة عند تحديد الحد الأدنى لطول السلسلة التي تقارن الخوارزمية على أساسها بين الملفات(Shortest string length to match) إذْ إنَّ الطول المناسب لهذا المعامل في الرماز البرمجي هو (10-20) كلمة، وفي حالة النصوص باللغات الطبيعية هو (4-6) كلمة [17]، إذْ إنَّ القيم الكبيرة لهذا المعامل تقود إلى أخطاء ولكنها تجعل الخوارزمية أسرع، في حين تعطي القيم الصغرى لهذا المعامل نتائجَ أدق، ولكن إذا صُغِّرَ طول السلسلة التي نقارن بها -إلى حد كبير- فإن ذلك يؤدي إلى نتائج غير صحيحة. ومن سلبيات هذه الخوارزميات أنها تتأثر بإعادة ترتيب الكلمات لأنَّها تعجز عن اكتشاف الانتحال عند إعادة ترتيب كلمات النص المسروق، كما أنها أيضاً لا تستطيع كشف الحالات التي يوجد فيها استبدال للمرادفات.

2-2-2- خوارزميات مقارنة البنى الشجرية للملفات (Tree-matching algorithms): تقوم فكرة هذا النوع من الخوارزميات على تحليل النص وبناء الأشجار المقابلة له [16] [2] [9]، إذ تحلل هذه الخوارزمية النص جملة–جملة قبل المقارنة، وذلك ببناء شجرة نحوية للجملة، أي تقوم بتوليد شجرة تمثل بنية الجملة، ويبيّن الشكل الآتي الشجرة النحوية للجملة "John cut the grass" [16]:



الشكل (1) تمثيل الجمل نحوياً

تتألف الجملة بشكل أساسي من تركيب نحوي اسمي (NP) وتركيب نحوي فعلي(VP) وتوصف الجملة باستخدام قواعد توليد الشجرة كالآتي:

1. S → NP VP
2. VP → V NP
3. NP → NAME
4. NP → ART N
5. NAME → John
6. V → cut
7. ART → the
8. N → grass

ولكن هناك مشكلة في توليد هذه القواعد إذْ من الممكن أن يكون للجملة نفسها أكثر من شجرة محتملة. لذا يجب أن يكون هناك عامل يساعدنا على تفضيل شجرة على أخرى، لذا أُدْخِلَ عامل الاحتمالات إلى الـparser الذي يولد هذه القواعد، فعندما يكون هناك أكثر من قاعدة قابلة للتطبيق تختار القاعدة الاحتمال الأعلى ذاته، بعد أن تُوَلَّدُ القواعد الخاصة بكل الجمل للملفين قيد المقارنة تقوم الخوارزمية بمقارنة هذه القواعد جملة ــ جملة حيث تُقَارَنُ القاعدة المتولدة عن الجملة الأولى في الملف الأول مع القواعد المولدة في الملف الثاني جميعها وهكذا..، ومن نقاط القوة في هذه الخوارزميات أنها تتغلب على مشكلة إعادة ترتيب الكلمات وعلى مشكلة الاستعاضة عن الكلمات بمرادفاتها، ولكن تواجهها مشكلات الغموض الموجودة في اللغات الطبيعية، التي تنتج عنها أحياناً أن يكون لجملة واحدة أكثر من شجرة تمثيل.

3-نظرية بنية الكلام البلاغية وتطبيقاتها: إن نظرية بنية الكلام البلاغية (Rhetorical Structure Theory) قابلة للتطبيق في كل من اللغتين العربية والإنكليزية، وتسمى هذه النظرية اختصاراً RST [4] [7].

وتفترض هذه النظرية ما يأتي:

- يمكن أن تُقْسَمَ النصوص إلى مجموعات من العبارات، التي تكون مترابطة فيما بينها بواسطة موصلات (Connectors)، لأنَّ هذه الموصلات تشكل صلة الوصل بين العبارات في النصوص وتولد العلاقات التي تربط أجزاء النص مع بعضها بعضاً.
- إن نظرية RST تُجَزِّئ النص إلى وحدات مترابطة فيما بينها وتحافظ على بقاء النص متجانساً، إذْ تستهدف هذه

النظرية المحافظة على المعنى الدلالي الموجود في النص؛ وذلك عند معالجة النصوص.

- إن العلاقات المتولدة من تطبيق هذه النظرية تكون بين أجزاء النص المتجاورة غير المتداخلة، وأيضاً بين أجزاء النص غير المتجاورة، إذْ يمكن أن ترتبط عبارة بأول المقطع مع عبارة بآخر المقطع [4].

تُستخدم نظرية RST في العديد من التطبيقات التي تقوم بمعالجة النصوص وأهم استخدام لها في التلخيص الآلي للنصوص إذْ يجري بالاعتماد عليها تحديد العلاقات بين عبارات النص وتُحَدَّدُ العبارات الأكثر أهميةً في النص [4].

4- استخراج العلاقات من اللغة الإنكليزية والعربية بحسب نظرية بنية الكلام البلاغية: اِستُخْدِمَتْ نظرية بنية الكلام البلاغية في العديد من التطبيقات التي تقوم بمعالجة النصوص (العربية والإنكليزية)، حيث أُجْرِيَ العديد من الدراسات والبحوث لاستخراج العلاقات الموجودة في النصوص العربية والإنكليزية[4]،[6]:

وفيما يأتي جدول يحتوي على العلاقات الموجودة في اللغة الإنكليزية والعلاقات المشتقة للغة العربية [4]:

جدول علاقات النص الإنكليزي ومقابلاتها في اللغة العربية

Rhetorical Relations	Name in English
شرط	Condition
تفسير	Interpretation
تعليل	Justification
استدراك	Recalling
توكيد	Confirmation
نتيجة	Result
تمثيل	Example
قاعدة	Base
تفصيل	Explanation
عطف	Joint
ترتيب	Sequence

مع الإشارة إلى أن الترجمة قد تختلف من مرجع إلى آخر مثل Recalling **يمكن أن يقوم بعضهم بترجمتها استرجاعاً، ولكننا في المرجع الذي أخذنا الجدول منه مترجمة على أنها استدراك وكذلك** Example **ترجمتها مثال ولكنها في المرجع تمثيل وكذلك** Sequence **يمكن أن تترجم على أنها تتالٍ بدلاً من ترتيب.**

وفيما يأتي بعض الأمثلة عن العلاقات الموجودة في النصوص العربية:

سأشتري ثلاثة كتب إذا ذهبت إلى المكتبة.

يولد الموصل "إذا " في الجملة السابقة علاقة شرطية بين العبارتين الأولى" سأشتري ثلاثة كتب" والثانية "ذهبت إلى المكتبة".

سافر محمد إلى مدينة جدة في الساعة الخامسة، أي أنه سيصل في الساعة السابعة تقريبا.

يولد الموصل" أي إنَّه " علاقة تفسير بين العبارة الأولى والعبارة الثانية.

[من المعلوم أن الفواكه مصدر أساسي للفيتامينات الضرورية للجسم.][1][ودوماً ينصح الأطباء بتناول الفواكه بشكل مستمر.][2][ومن نعم الله على عباده أن هذا المصدر متوافر طوال العام.][3][فعلى سبيل المثال نجد البرتقال والتفاح متوافرين في الأسواق بشكل مستمر، وهي من الفواكه الغنية بالفيتامينات الضرورية للجسم.][4]

نجد أن العبارة رقم 4 مرتبطة بعلاقة من نمط تمثيل (أو مثال) بالعبارات السابقة لها كلّها 1و2 و3؛ وذلك بواسطة الموصل "على سبيل المثال " إذْ تمثل هذه العبارات حقائق بالنسبة إلى العبارة الرابعة[4].

ولدينا فيما يأتي بعض الأمثلة عن العلاقات في النصوص الإنكليزية:

I am very happy because I got excellent marks in exams .

يولد الموصل "because " علاقة سببية بين العبارة الأولى "I am very happy " والعبارة الثانية " I got excellent marks in exams".

I will pass the exams if I study.

يولد الموصل "if " علاقة شرطية بين العبارة الأولى "I will pass the exams " والعبارة الثانية "I study ".

تُولِّدُ العلاقات بين العبارات في اللغتين العربية والإنكليزية بواسطة الموصلات (cue phrases)، إذْ إن كل موصل يولد نوعاً محدداً من العلاقات بين العبارات، وتشكل الموصلات الحدود المنطقية للعبارات. أي بمعنى آخر نستطيع تقسيم النص العربي والإنكليزي إلى عبارات بالاعتماد على الموصلات الموجودة في النصوص، وفيما يأتي جدول بأهم بالموصلات في النصوص العربية والإنكليزية[4]،[12]،[13]،[14]،[15]:

جدول الموصلات في اللغتين العربية والإنكليزية

Arabic cue-phrases	The closest English cue-phrases
لا ...بل ، ليس ...بل	Not…,but
لم ... إنما	But
لابد (في المرجع لابد –لكن يمكن أن تترجم إذا وفقط إذا)	If and only if
خشية أن-مخافة أن	For fear that
حتى	Even so
أو- أو لا	Or do-or do not
إما...وإما -إما...أو	This…or
إذ- وإذا –فإذا	When
إذن	The conclusion is
أو(في المرجع أو لكن يمكن أن تترجم ليس بالتأكيد)	Not sure
لكن(بمعنى الاستدراك)	But
أم	Whether
إذا...وإذا...و...-ف- سوف- فسوف	If…and if…and…,then

5- دوافع استخدام نظرية بنية الكلام البلاغية لكشف الانتحال: هناك العديد من الأسباب المقنعة التي تشجع على اعتماد نظرية بنية الكلام البلاغية للمقارنة بين النصوص وأهمها:

- إن تجزئة النصوص باستخدام نظرية RST تؤدي إلى تجزئة النص إلى مجموعة من العبارات (Clauses) دون أن تقوم بكسر أو فصل بنية الجملة الواحدة، أي إنَّه باستخدام هذه النظرية نحافظ على بنية الجملة التي هي المركب الأساسي في النص لأنَّ المعنى الدلالي يكون موجوداً في الجمل، ويؤدي تجزئة النص إلى مفردات إلى ضياع المعنى الدلالي (لأن كل مفردة لا يمكنها أن تحتوي على معنى دلالي إلا إذا وجدت ضمن جملة كاملة).
- إن تجزئة النص باستخدام نظرية RST يمكن القيام به دون الحاجة لوجود علامات الترقيم في النصوص، لأنَّ معظم النصوص لا تستخدم علامات الترقيم فيها أو تُسْتَخْدَمُ استخداماً غير كامل ودقيق.
- إن الاعتماد على بنية الجملة كمكون أساسي في عمليات المقارنة بين النصوص هو حل جيد؛ لأن الجملة تحتوي على قدر ممكن من المعلومات الدلالية أكثر من المفردات.

6- خوارزمية المقارنة المتبعة في البحث:

تتألف الخوارزمية التي سوف نتبعها لكشف الانتحال من خطوتين أساسيتين:

- مرحلة التحليل.
- مرحلة المقارنة.

سوف نشرح كل مرحلة بالتفصيل فيما يأتي:

6-1 مرحلة التحليل:

تتألف هذه المرحلة من المراحل الجزئية الآتية:

لمرحلة الأولى: تحليل النص بالاعتماد على نظرية بنية الكلام البلاغية؛ وذلك وفق مايأتي:

- أولاً تحدد الخوارزمية الموصلات في مقاطع النص كلّها.
- تقوم الخوارزمية لكل موصل بتخزين نمط العلاقة التي يولدها هذا الموصل وتُخَزِّنُ الكلمتين المحيطتين بالموصل من اليمين ومن اليسار، أي بمعنى آخر تقوم بتخزين حدود الموصل.
- تُجَزِّئُ الخوارزمية النص إلى عبارات وذلك بالاعتماد على الموصلات الموجودة في كل مقطع من مقاطع النص، ومن ثم بالنسبة إلى كلّ موصل تقوم بتخزين العبارة السابقة له والعبارة اللاحقة له.

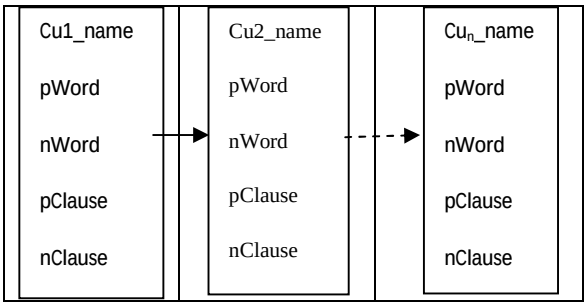
المرحلة الثانية: تحديد الأجزاء الكاملة (Complete Segments)، إذْ فضلاً عن تحديد الموصلات في النص فإن الخوارزمية تستعين أيضاً بالأجزاء الكاملة من النص، وهي:

- "*" أي جزء من النص بين علامتي تنصيص.
- (*) أي قوسين صغيرين ويوجد داخلهما نص.
- [*] أي قوسين متوسطين ويوجد داخلهما نص.
- {*} أي قوسين كبيرين ويوجد داخلهما نص.

بالنسبة إلى الأجزاء الكاملة فإنَّ الخوارزمية أيضاً تقوم بتخزين المعلومات السابقة نفسها للموصلات التي هي:

- النص الموجود داخل القوسين.
- الكلمتان المحيطتان بالجزء الكامل من اليسار ومن اليمين.
- العبارة السابقة للجزء الكامل، والعبارة اللاحقة له.

بعد أن تنتهي الخوارزمية من تحليل النص بالاعتماد على نظرية RST تُخَزَّنُ المعلومات الناتجة عن عملية التحليل هذه على الشكل الآتي:



الشكل(2) بنية المعطيات لتخزين ناتج عملية التحليل

من الشكل السابق نجد أن موصل يضم كل المعلومات المتعلقة به كلّها من اسمه والكلمة السابقة له والكلمة اللاحقة له والعبارة السابقة له العبارة اللاحقة له، إذْ إنَّ

cuName: اسم الموصل.

pWord: الكلمة السابقة للموصل.

nWord: الكلمة اللاحقة للموصل.

pClause: العبارة السابقة للموصل.

nClause: العبارة اللاحقة للموصل.

وتُخَزَّنُ أيضاً المعلومات المتعلقة بالأجزاء الكاملة ببنية تخزين المعلومات الخاصة بالموصلات نفسها.

نلاحظ أنه بعد عملية تحليل النص بالاعتماد على نظرية RST تحَول النص إلى مجموعة من الموصلات ومجموعة من العبارات الواقعة بين هذه الموصلات. وأيضاً بالنسبة إلى كلّ موصل فإنه يحتوي على الكلمات المحيطة به من اليسار ومن اليمين؛ وذلك لاستخدام هذه الكلمات المحيطة في عملية المقارنة بين الموصلات.

2-6 مرحلة المقارنة:

تقوم الخوارزمية في هذه المرحلة بالمقارنة بين معطيات ملفي الدخل اللذين حُلِّلاَ في المرحلة الأولى(مرحلة التحليل).

إذْ تقوم الخوارزمية بالمرور على الموصلات الخاصة بملفي الدخل وتبدأ بالمقارنة بين معلومات الموصلات لملفي الدخل، لكل موصل تقارن الخوارزمية ما يأتي:

- اسم الموصل (Cue Phrase name)

- حدًّا الموصل (Cue Phrase boundaries) أي الكلمتان المحيطتان بكل موصل من اليمين ومن اليسار.

وللمقارنة بين موصلين اثنين هناك عدة حالات للتطابق:

- تطابق تام: أي أسماء الموصلات متطابقة والكلمتان المحيطتان بكل منهما متطابقتان.
- تطابق من اليمين:أي أسماء الموصلات متطابقة والكلمتان السابقتان لكل منهما متطابقتان.
- تطابق من اليسار: أي أسماء الموصلات متطابقة والكلمتان اللاحقتان لكل منهما متطابقتان.

تقوم الخوارزمية أيضاً بالمقارنة بين الأجزاء الكاملة بطريقة المقارنة بين الموصلات نفسها.

توجد أيضاً طريقة أخرى للمقارنة بين الموصلات:

وهي استخدام تصنيفات الموصلات أي Cue Phrases Categories of، إذْ إنَّ هناك عدة موصلات تولد العلاقة النصَية نفسها، أي يمكن أن يكون لعدة موصلات المعنى نفسه، لإنَّ استبدال أي موصل بإحدى مرادفاته يمكن ألا يغيّر المعنى، وتصبح المقارنة باستخدام مرادفات الموصلات كالآتي:

- مقارنة الموصل بإحدى مرادفاته.
- مقارنة الكلمتين المحيطتين بالموصل.

وأخيراً لمّا كانت الموصلات تُجَزِّئُ النص إلى عبارات مترابطة منطقياً لذا أمكن المقارنة بين الموصلات دون النظر إلى الموصل بحد ذاته (أي دون النظر إلى اسم الموصل)، أي فقط نقوم بمقارنة الكلمات المحيطة به، العبارة السابقة والعبارة اللاحقة له بغض النظر عن اسم الموصل.

7- ميزات الخوارزمية المتبعة:

تتميّز الخوارزمية التي قمنا بتطويرها في هذا البحث بما يأتي:

- غير مكلفة زمنياً لأنَّ الخوارزمية تحتاج إلى N * M وحدة زمن للمقارنة بين ملفين إذْ إنَّ N هو عدد الموصلات في الملف الأول وM هو عدد الموصلات في الملف الثاني.
- 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 إذْ تقوم بتقطيعه تقطيعاً منطقياً يعتمد على وجود الموصلات في النص.
- قابلة للتطبيق على اللغة الإنكليزية واللغة العربية إذْ إنَّ خوارزمية المقارنة التي تعتمد على نظرية RST (تحديد الموصلات وتحديد العبارات اعتماداً على توضع هذه الموصلات) موجودة وقابلة للتطبيق في النصوص العربية والإنكليزية أيضاً.
- يمكن للخوارزمية مقارنة الملفات المكتوبة من دون وجود علامات الترقيم في اللغة؛ وذلك بسبب عدم اعتمادها على علامات الترقيم لتحديد العبارات.
- إن عمل الخوارزمية في تحديد بنى العبارات وحدودها لا يوجد فيه أي التباس أو غموض بالمقارنة بخوارزميات معالجة النصوص التي تعتمد على نظرية البنى الشجرية التي تقوم بتحويل النص إلى بنى شجرية مقابلة إذْ يمكن أن يكون للعبارة الواحدة أكثر من شجرة مقابلة وينتج عن ذلك غموض ونحتاج معه إلى استخدام تجريبيات لفك هذا الغموض.
- لا تحتاج الخوارزمية إلى وجود التشكيل في النص الذي تقوم بفحص الانتحال فيه (في حالة كشف الانتحال في اللغة العربية).

- لا تحتاج الخوارزمية إلى معالجة بدائية للنص قبل تحليله بخلاف أنواع خوارزميات كشف الانتحال جميعها إِذْ تحتاج كل منها إلى معالجة بدائية للنص؛ وذلك لإزالة الكلمات غير المفيدة وتقطيع الجمل الطويلة إلى جمل أصغر، كما في حالة خوارزميات مقارنة السلاسل النصية.
- تتميّز الخوارزمية المتبعة عن باقي أنواع هذه الخوارزميات بسرعتها في مرحلة المقارنة إِذْ إنَّها أسرع من أنواع الخوارزميات جميعها لأنها في مرحلة المقارنة تقارن فقط بين الموصلات وحدودها.

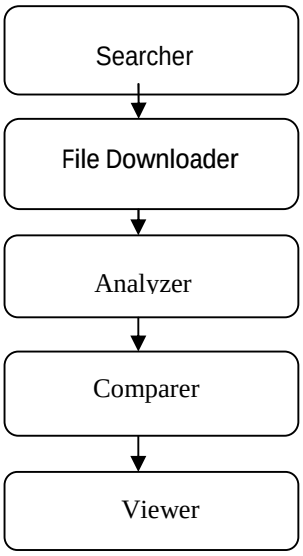
مما سبق نستنتج أن الخوارزمية المطورة في هذا البحث هي خوارزمية جيدة، وأهم ما يميّزها هو سرعتها وعدم حاجتها لمعالجة بدائية للنصوص والحفاظ على بنية الملفات التي تقوم بفحص كشف الانتحال فيما بينها.

إلا أن الخوارزمية تتأثر بإعادة ترتيب الكلمات لكن مقدار تأثرها بتغيير ترتيب الكلمات في الجملة أقل من تأثر خوارزميات المقارنة التي تعتمد على بصمات الملفات، إِذْ إن خوارزميتنا تبقى قادرة على كشف انتحال جملة ما، ما لم يجرِ تغيير الكلمات المحيطة بالموصل (أي حدود الموصل من اليمين ومن اليسار) أمَّا في خوارزميات المقارنة التي تعتمد على بصمات الملفات فإنَّ أي تغيير في أي كلمة في الجملة سوف يؤثر TD فعاليتها في كشف الانتحال.

8- تصميم النظام: يتألف النظام المنجز بشكل أساسي من الأقسام الرئيسة الآتية:

- الباحث (Searcher): يستخدم هذا القسم Bing search web service [5] ويعيد نتائج البحث الخاصة بالاستعلام الذي أدخله المستخدم، يقوم الباحث بالبحث عن الملفات من نوع doc وpdf.
- محمّل الملفات (File Downloader): يقوم بتحميل نتائج البحث من الشبكة العنكبوتية ويخزنها في قاعدة المعطيات الخاصة بالنظام.
- المحلل (Analyzer): يقوم بتحليل الملفات قبل المقارنة بينها إِذْ يمكن أن يحلل الملف تحليلاً كاملاً أي تحليل مفرداتي وتحليل معتمد على نظرية RST
- المقارن (Comparer): يقوم بالمقارنة بين مخرجات عملية تحليل الملفات باستخدام خوارزمية المقارنة التي شُرِحَتْ مسبقاً.
- عارض النتائج (Viewer): يقوم بعرض نتائج المقارنة بين الملفات وتكون الأجزاء المنتحلة ملونة بلون آخر عن باقي النص.

طُوِّرَ الباحث (Searcher) باستخدام لغة البرمجة C# أمَّا باقي أجزاء النظام فَطُوِّرَتْ باستخدام لغة البرمجة Java. صُمِّمَتْ قاعدة المعطيات الخاصة بالنظام باستخدام Sqlserver2005 وتحتوي هذه القاعدة على موصلات لكل من اللغتين العربية والإنكليزية، وأيضاً تحتوي على العلاقات الموجودة في النصوص العربية والإنكليزية، ويعرض الشكل (3) مخططاً لأقسام النظام الرئيسة.



الشكل (3) مخطط النظام

9- الاختبارات والنتائج: إن خوارزميتنا قابلة للتطبيق على اللغة العربية والإنكليزية، ولكننا قمنا باختبارها فقط لكشف الانتحال في اللغة العربية، إذْ أُجْرِيَتِ الاختبارات على عينة من الملفات(حجم العينة 100 ملف، عبارة عن بحوث علمية نُزِّلَتْ من الموقع الإلكتروني لإحدى الجامعات الحكومية السورية لأنَّها متاحة بشكل مجاني ضمن الأعداد السابقة من إصدارات المجلة، والحجم الوسطي لكل ورقة علمية هو 10 صفحات قياس A4).

وفيما يأتي نتائج اختبارات الخوارزمية: بسبب حجم الملفات التي أُجْرِيَ الاختبار عليها فإنه يعتذر علينا إظهار هذه الملفات هنا، ولكن سوف نقوم بذكر فعالية الخوارزمية المستخدمة (مع الإشارة إلى عدم المقدرة على مقارنة نظامنا المطور بالنظم الموجودة على الشبكة العنكبوتية التي أغلبها لا يدعم اللغة العربية وهي غير متاحة بشكل مجاني، ولكننا هنا قمنا بقياس مدى فعالية الخوارزمية المطورة في هذا النظام).

اِسْتُخْدِمَتْ ثلاثة أنواع لخوارزمية المقارنة المستخدمة التي تعتمد على نظرية الـ RST، وهذه الأنواع هي:

1: خوارزمية المقارنة التي تقارن معلومات الموصلات فضلاً عن الكلمات المحيطة بها.

2: خوارزمية المقارنة بين الموصلات مع الأخذ بالحسبان صفوف الموصلات Cue Phrases categories.

3: وكذلك المقارنة بين الموصلات بغض النظر عن أسماء الموصلات، أي فقط فحص الكلمات المحيطة بالموصلات واختبارها دون النظر إلى أسماء الموصلات.

إن خوارزمية المقارنة رقم 1 كانت نتائج اختبارها غير مجدية؛ وذلك لأنها تأخذ بالحسبان اسم الموصل عند المقارنة بين الموصلات، لأنَّ المنتحل عندما ينتحل أي جزء من نص ويضيفه إلى ملفه يكون احتمال تغييره للموصل سواء في بداية العبارة أو نهايتها كبيراً جداً.

وكذلك بالنسبة إلى الخوارزمية رقم 2 التي تستخدم صفوف الموصلات إذْ تنتج عدم فعاليتها من عدم وجود صفوف للموصلات بشكل واضح ودقيق.

أمَّا خوارزمية المقارنة رقم 3 التي تقوم بالمقارنة بين الملفات بالاعتماد على الكلمات المحيطة بالموصلات دون النظر إلى الموصلات ومواقعها فهي أكثر الخوارزميات المستخدمة فعاليةً.

إذْ إنَّ الخوارزمية رقم 3 أدت إلى نتائج ممتازة، وقد أدت إلى كشف معظم حالات الانتحال الحاصلة بين ملفات الدخل (أي من ملفات العينة التي اختيرت لاختبار الخوارزمية عليها) وقد بلغت نسبة كشف الانتحال 75% من الحالات.

واِسْتُخْدِمَ عامل قياس الدقة الآتي:

$$\text{Precision} = \frac{\text{Number of correct plagiarism detection results}}{\text{The total number of suspected files}}$$

الذي يمثل عدد نتائج كشف الانتحال الصحيحة مقسوماً على العدد الكلي للملفات التي نقوم باختبار كشف الانتحال فيها.

مع الإشارة إلى أن النظام المطور هنا في هذه الورقة البحثية يقوم بفحص كشف الانتحال على مستوى العبارات ضمن مقاطع النص، وخرج هذا النظام هو مجموعة الملفات التي يتضمن الملف المدخل أجزاء منتحلة منها، ويقوم النظام المطور بتلوين العبارات المنتحلة في الملفات لإرشاد المستخدم إليها، والعينة التي اِخْتُبِرَ النظام عليها المؤلفة من 100 ملف اختيرت بشكل عشوائي وغير معروف مسبقاً إذا كانت تحتوي على

أجزاء منتحلة ضمنها، وقد تم التأكد من خرج النظام على العينة المختارة (100 ملف) بشكل يدوي ولوحظ أن هناك حالات انتحال لم يقم النظام المطور بكشفها نتيجة قيام المنتحل بتغيير في بنى العبارات واستبدال المرادفات.

وتعدُّ نسبة عامل الدقة في كشف الانتحال في النظام المطور جيدة، ولاسيَّما أن هذا النظام مخصص لكشف الانتحال في اللغة العربية، ولايوجد أرقام ونسب متوافرة عن دقة نظم كشف الانتحال الأخرى المتوافرة على الشبكة العنكبوتية.

10- الخاتمة والآفاق المستقبلية: قمنا بهذا البحث بتطوير نظام لكشف الانتحال بالاستعانة بمحرك البحث bing، ويعتمد هذا النظام على خوارزمية كشف انتحال فعّالة تدعم كلاً من اللغة العربية واللغة الإنكليزية، وتتميّز هذه الخوارزمية بالسرعة والفعالية.

ويمكننا في المستقبل إدخال البعد الدلالي في خوارزمية المقارنة بين الموصلات؛ وذلك بالاستعانة بقاموس مفاهيمي عند المقارنة بين حدود الموصلات والجمل الواقعة بينها.

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.

المراجع*

1. Vinod K.R., Sandhya.S, Sathish Kumar D, Harani A, David Banji, Otilia JF Banji, Plagiarism-history detection and prevention, Journal for drugs and medicines, Vol.3, Issue:1, pp.1- 4, 2011.
2. Shizhong Wu; Yongle Hao; Xinyu Gao; Baojiang Cui; Ce Bian, Homology Detection Based on Abstract Syntax Tree Combined Simple Semantics Analysis, Web Intelligence and Intelligent Agent Technology (WI-IAT), vol.3, pp.410-414, 2010
3. Al-Khatib B., Aspel A. ,Saleh M., fares M.، Hamad M.M., plagiarism detection using the web, Damascus university,informatics engineering college, 2007.
4. Al-Sanie W., Towards an infrastructure for Arabic text Summarization using Rhetorical Structure Theory, master thesis , king Saud University, K.S.A., 2005
5. Bing , API Basics. [online] Available at: http://www.bing.com/developers/s/APIBasics.html [Accessed 15-October 2011].
6. Mathkour H., Touir A., Al-Sanea W., Parsing Arabic Texts Using Rhetorical Structure Theory, Journal of Computer Science , vol. 4, no. 9, pp. 713-720, 2008
7. Taboada M., Mann C. W.، Rhetorical Structure Theory: looking back and moving ahead, journal of Discourse Studies, vol. (8),pp. 423—459, 2006.
8. Kent K. C., Salim N., Web Based Cross Language Plagiarism Detection, Journal of computing, vol.1, issue 1, issn: 2151-9617, 2009.
9. Mozgovoy M., Tusov V., Klyuev V., The Use of Machine Semantic Analysis in Plagiarism Detection, Proceedings of the 9th International Conference on Humans and Computers, Aizu-Wakamatsu, Japan, pp. 72-77, 2006.
10. Liu Y. ، Liang L., A Dual-method Model for Copy Detection, Web Intelligence and Intelligent Agent Technology Workshops of IEEE/WIC/ACM International Conference, pp.634 – 637, 2006.
11. Schleimer S., Winnowing: Local Algorithms for Document Fingerprinting, Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp.76-85, 2003.
12. Touir A., Mathkour H., Al-Sanea W., Semantic-Based Segmentation of Arabic Texts, Information Technology Journal, Vol. 7, pp. 1009-1015, 2008.
13. Al-Sanea W., Mathkour H., Touir A, On an Arabic Text Segmentation Techniques, In proceeding of The Ninth International Conference on Information Integration and Web-based Applications Services, Jakarta, 2007.
14. Mathkour H., Member, IAENG, A Novel Rhetorical Structure Approach for Classifying Arabic Security Documents, International Journal of Computer Theory and Engineering, Vol. 1, No. 3,pp.195-200, 2009.
15. Iraky K., A Comprehensive Taxonomy of Arabic Discourse Coherence Relations, in Taibah University International Conference on Computing and Information Technology Al-Madinah Al-Munawwarah, Saudi Arabia, pp.491-496, 2012.
16. Mozgovoy M., Tuomo K., Erkki S., Using natural language parsers in plagiarism detection, in Speech and Language Technology in Education, pp.77-79, 2007.
17. Tachaphetpiboon S., Facundes N., Amornraksa T., Plagiarism Indication by Syntactic-Semantic Analysis, In proceeding of Communications, APCC 2007. Asia-Pacific Conference, 2007.

تاريخ ورود البحث إلى مجلة جامعة دمشق 2012/6/20