استخراج القواعد الدلالية من قاعدة بيانات نصية باستخدام التنقيب في المعطيات

أحمد بدر الدين الخضر[1] و محمد دباس الحميد[1] و لينا محمد فاروق بعاج[2]

(1) قسم هندسة البرمجيات، كلية الهندسة المعلوماتية، جامعة حلب، حلب، سورية

(2) قسم هندسة الحواسيب، كلية الهندسة الكهربائية والإلكترونية، جامعة حلب، حلب، سورية

الملخص

يعرض هذا البحث نموذجاً لتوليد القواعد الدلالية Semantic Rules من قاعدة بيانات تتضمن الكلمات الأكثر تكراراً في اللغة الانكليزية لاستخدامه في التحقق من صحة جمل اللغة الانكليزية من ناحية المعنى، ويبين الأسلوب المتبع في بناء هذا النموذج بالاعتماد على إحدى خوارزميات التنقيب في المعطيات، هذه الخوارزمية هي خوارزمية FP Growth التي تم اعتمادها لتوليد قواعد الترابط Association Rules بين الكلمات المخزنة في قاعدة البيانات، القواعد الناتجة عن الخوارزمية تتجاهل تسلسل الكلمات ضمن القواعد، لكن النظام المقترح في هذا البحث يركز على تسلسل ورود الكلمات ضمن قواعد الترابط ويعتبره مهماً لإنجاز التحليل الدلالي، لذلك تم تعديل خوارزمية FP Growth للحصول على قواعد ترابط تعطي أهمية لتسلسل ورود الكلمات ضمن هذه القواعد. وبما أن معالجة اللغات الطبيعة هي عملية متطورة بشكل مستمر فإن هذا البحث يفتح المجال نحو إنجاز التحليل الدلالي للجمل باستخدام تقنيات التنقيب في المعطيات، كما أنه يعد خطوة هامة تُظهر أهمية الأبحاث المعلوماتية في إنجاز أنظمة تهتم بمعالجة اللغات الطبيعية وتحاكي الإنسان الخبير باللغة الإنكليزية.

الكلمات المفتاحية: التحليل الدلالي، التنقيب في المعطيات، القواعد الدلالية، خوارزمية FP Growth، قاعدة بيانات، قواعد الترابط.

المقدمة

التنقيب في المعطيات Data Mining هو عملية اكتشاف المعلومات الموجودة في مجموعة ضخمة من البيانات (Han and Kamber, 2006)، والهدف الأساسي من هذه

العملية هو اكتشاف النماذج Patterns والاتجاهات الموجودة في هذه البيانات، وتعرّف أيضاً بأنها عملية آلية أو شبه آلية تهدف إلى تحليل كميات كبيرة من البيانات لاستخراج نماذج وقواعد مهمة منها ولإيجاد نماذج لم تكن معروفة مسبقاً (Witten *et al.*, 2011)، حيث أن هذه النماذج لا يمكن اكتشافها بالطرق التقليدية مثل الاستعلام من قاعدة البيانات بواسطة تعليمة الاختيار select لأسباب منها أن العلاقات بين البيانات قد تكون معقدة جداً أو بسبب ضخامة حجم البيانات. يستلزم تطبيق تقنيات التنقيب في المعطيات ما يلي:

1. توفر قاعدة بيانات ضخمة تتضمن بيانات عن المسألة المراد حلها.
2. اختيار وتطبيق خوارزمية تناسب المسألة المطروحة.

من جهة أخرى يعتبر التحليل الدلالي مهماً في الحكم على صحة معاني الجمل، لكنه يتطلب توفر قواعد دلالية. تستلزم عملية استخلاص هذه القواعد وجود إنسان خبير يقوم ببناء هذه القواعد، أو يمكن توليد هذه القواعد آلياً باستخدام الحاسب. سنعتمد على تقنيات التنقيب في المعطيات لاستخراج القواعد الدلالية من قاعدة بيانات، وذلك من خلال تطبيق إحدى خوارزميات التنقيب في المعطيات (خوارزمية FP Growth (Han and Kamber, 2006; Witten *el al.*, 2011)) على قاعدة بيانات ضخمة تتضمن بيانات عن الكلمات المترافقة الأكثر تكراراً في اللغة الانكليزية. وبعد تطبيق الخوارزمية السابقة على قاعدة البيانات نحصل على المعلومات المخبأة فيها المتمثلة بالقواعد الدلالية.

يتكون البحث من عدة مقاطع حيث نستعرض في المقطع الثاني أهمية البحث وأهدافه المتمثلة في استخدام إحدى تقنيات التنقيب في المعطيات لاستخراج القواعد الدلالية من قاعدة بيانات تخزن بيانات عن الكلمات المترافقة الأكثر تكراراً في اللغة الانكليزية. نقوم في المقطع الثالث بشرح خوارزمية FP Growth ومبدأ عملها. نقدم بعدئذ في المقطع الرابع طريقة ومنهجية البحث المتبعة التي تتضمن تجهيز قاعدة البيانات التي سيتم تطبيق خوارزمية FP Growth عليها، وبرمجة خوارزمية FP Growth بلغة

البرمجة C# ودراسة المشاكل الناتجة عند تطبيقها في مجال التحليل الدلالي لجمل اللغة الانكليزية، إضافة إلى إنجاز تعديلات على الخوارزمية تضمن حل هذه المشاكل. نعرض في المقطع الخامس دراسة وتحليلاً للنتائج التي تم التوصل إليها. وأخيراً سنقوم في المقطع السادس بعرض ملخص ما توصل إليه البحث.

أهمية البحث وأهدافه

لا زالت عملية فهم نص من قبل الحاسب ومعرفة المعنى الدلالي لهذا النص والحكم عليه فيما إذا كان صحيحاً أما لا من المشاكل الكبرى التي تواجه التطبيقات المعلوماتية في مجال معالجة اللغات الطبيعية، لذلك تعد عملية التحقق من معاني ودلالة جمل اللغة الانكليزية من الأهداف الأساسية لعلماء اللغة، تندرج هذه العملية ضمن مجال معالجة اللغات الطبيعية NLP Natural Language Processing (Manning and Schutze, 1999) التي تعد مجالاً هاماً يربط بين علم الحاسب Computer Science وعلم اللغة Linguistics وتركز على التفاعل بين الحاسب واللغات الطبيعية (Charniak, 1985). يتمحور هذا البحث حول إيجاد حل لمشكلة فهم المعنى الدلالي لنص (Yamuna and Shree 2013)، وتعود أهميته إلى أنه يقدم إنجازاً جديداً في مجال فهم ومعالجة اللغات الطبيعية من خلال إنجاز التحليل الدلالي لجمل اللغة الانكليزية.

العديد من الأبحاث خاضت تجربة استخراج القواعد الدلالية ونذكر منها (Kane and Milgram 1993) حيث تم استخدام المنطق الترجيحي propositional logic وجدول الحقيقة truth table وذلك من أجل استخراج الترابط المنطقي logical dependencies من الشبكة العصبونية المدربة trained neural network وقد قام الباحثون بتطبيق نتائجهم في مجال تحليل البيانات data analysis. باحثون آخرون قاموا باستخدام منطق الغموض fuzzy logic كم في (Masuika *et al* 1990) و (Hayashi 1990) لكن النتائج تبقى ليس عامة وهي تستند إلى نتائج تجريبية فقط empirical results.

نقوم في هذا البحث بتطبيق إحدى تقنيات التنقيب في المعطيات (خوارزمية FP Growth) مع القيام بتعديلها لاستخلاص القواعد الدلالية (قواعد المعاني) المخبأة في قاعدة بيانات تتضمن الكلمات الأكثر تكراراً في اللغة الانكليزية، لاستثمارها في التحقق من صحة الجمل من ناحية المعنى. إضافة إلى ذلك تأتي أهمية هذا البحث من كونه يفتح المجال أمام الحاسب ليكون أداة تحاكي الإنسان الخبير باللغة الإنكليزية.

طريقة البحث

تمثل الخطوات التالية منهجية العمل الذي قمنا به:

- تحديد المسألة المراد حلها.
- تحديد الأدوات اللازمة لإنجاز الحل.
- توضيح السبب الأساسي لتعديل خوارزمية FP Growth.
- توضيح آلية استثمار القواعد الدلالية الناتجة عن خوارزمية FP Growth المعدلة.
- إعداد قاعدة البيانات التي سيتم تطبيق خوارزمية FP Growth عليها.
- برمجة خوارزمية FP Growth وتطبيقها على قاعدة البيانات التي تم تجهيزها في الخطوة السابقة.
- إعادة برمجة خوارزمية FP Growth بحيث تولد قواعد ترابط تعطي أهمية لتسلسل ورود الكلمات.
- دراسة النتائج الجديدة التي تم الحصول عليها بعد التعديل.

تحديد المسألة المراد حلها

لدينا جملة صحيحة قواعدياً في اللغة الانكليزية، ونريد إيجاد القواعد الدلالية (قواعد المعاني) لاستخدامها في التأكد من أن الجملة صحيحة من الناحية الدلالية. أي أننا نريد الحصول على محرك دلالي يحاكي الإنسان الخبير بمعاني اللغة ولديه القدرة على الحكم على صحة الجمل دلالياً.

تحديد الأدوات اللازمة لإنجاز الحل

نحتاج لحل المسألة السابقة إلى ما يلي:

1. توفر قاعدة بيانات ضخمة تتضمن بيانات عن كلمات اللغة الانكليزية.
2. اختيار وتطبيق خوارزمية تنقيب في المعطيات تناسب المسألة المطروحة.

تم الحصول على قاعدة بيانات ضخمة من الانترنت سنتحدث عنها بشكل تفصيلي في الفقرة (4 - 5). أما بالنسبة لاختيار خوارزمية التنقيب المناسبة فقد اتجهنا نحو الخوارزميات التي تختص بإيجاد الترابط بين البيانات من خلال إيجاد النماذج المتكررة وقواعد الترابط Frequent Patterns and Association Rules. درسنا عدة خوارزميات مثل Apriori وFP Growth وVertical data format.

تستخدم خوارزمية Apriori طريقة التوليد والاختبار generate and test أي أنها تولد العناصر المرشحة ثم تختبر فيما إذا كانت تمثل تكراراً. هذه الخوارزمية مكلفة (في الزمن والمساحة التخزينية) خاصة إذا كانت قاعدة البيانات ضخمة وكان عدد العناصر المرشحة المختبرة كبيراً (Han and Pei, 2000; Motoda and Ohara, 2009)، لذلك تم استبعاد هذه الخوارزمية.

يقوم المبدأ الأساسي لخوارزمية Vertical data format على تحويل الأسطر في الجدول إلى الأعمدة، أي أنه لكل عنصر في قاعدة البيانات يتم تخزين قائمة بأرقام السجلات التي ينتمي إليها العنصر، وبذلك يتم تمثيل البيانات بشكل عمودي. وبعد ذلك يتم حساب مجموعات العناصر المتكررة. تعتبر خوارزمية سريعة لكن القوائم الوسيطية التي تضم أرقام سجلات العناصر ممكن أن تكون ضخمة جداً وخاصة إذا كانت قاعدة البيانات ضخمة (Han and Kamber, 2006)، لذلك تم استبعاد هذه الخوارزمية.

وجدنا أن خوارزمية FP Growth هي الأفضل لأنها تتميز بما يلي (Han and Kamber, 2006; Motoda and Ohara, 2009; Witten *et al*., 2011):

- تمر فقط مرتين على قاعدة البيانات.
- تحول قاعدة البيانات إلى بنية شجرية FP tree حجمها دائماً أصغر من حجم قاعدة البيانات الأصلية.
- لا تولد عناصر مرشحة، لذلك فهي أسرع من خوارزمية Apriori.

مبدأ عمل خوارزمية FP Growth

قبل الشروع بتوضيح السبب الأساسي لتعديل خوارزمية FP Growth سنقوم بسرد هذه الخوارزمية للمقارنة ومعرفة آلية عملها حتى يتثنى لنا تحديد التعديل فيها.

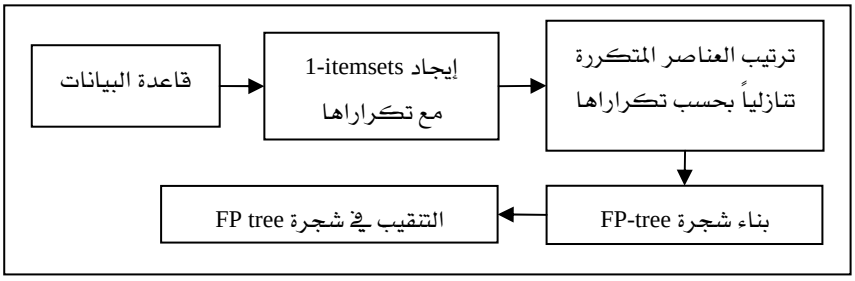
تقوم آلية عمل خوارزمية FP Growth على التنقيب في قواعد الترابط الموجودة في شجرة النموذج المتكرر FP Tree، وتعمل وفق الخطوات التالية (Han and Kamber, 2006):

1. المرور الأول على قاعدة البيانات لإيجاد 1-itemsets (مجموعة تتضمن كل عنصر بشكل مفرد) مع تكراراتها، وتخزينها في اللائحة L.
2. الحصول على العناصر المتكررة Frequent items (عناصر تكرارها >= min_sup).
3. ترتيب العناصر المتكررة ترتيباً تنازلياً بحسب قيمة تكرارها.
4. المرور الثاني على قاعدة البيانات لبناء شجرة النموذج المتكرر FP tree، حيث يتم بناء الشجرة كما يلي:

- إنشاء جذر الشجرة وتسميته null.
- لكل سجل من سجلات قاعدة البيانات، يتم ترتيب العناصر المتكررة فيه بحسب الترتيب في اللائحة L.
- يتم إنشاء فرع لكل سجل من سجلات قاعدة البيانات مع مراعاة ما يلي: إذا كانت الشجرة تتضمن عقدة لها الاسم نفسه للعنصر الحالي نزيد العداد الموجود فيها بمقدار واحد، وإلا يتم إنشاء عقدة جديدة يكون التكرار فيها مساوياً الواحد مع ربطها مع العقدة الأب لها.

5. التنقيب في شجرة FP tree: لكل نموذج بطول 1 (نموذج لاحق ابتدائي) يتم بناء قاعدة النموذج الشرطي Conditional Pattern Base (قاعدة بيانات جزئية تشكل مسارات مرتبطة في شجرة FP tree مرتبطة مع النموذج اللاحق)، بعدئذ تُبنى شجرة FP الشرطية Conditional FP tree، ويتم الحصول على النموذج بربط النموذج الابتدائي مع النماذج المتكررة المولدة من شجرة FP الشرطية.

يبين الشكل (1) المخطط الصندوقي لخطوات عمل خوارزمية FP Growth:



الشكل (1): المخطط الصندوقي لخوارزمية FP Growth

توضيح السبب الأساسي لتعديل خوارزمية FP Growth

إن قواعد الترابط الناتجة عن خوارزمية FP Growth هي قواعد تتجاهل تسلسل ورود الكلمات ضمن هذه القواعد، لأن هذه الخوارزمية تركز على ترتيب العناصر المتكررة ترتيباً تنازلياً بحسب قيمة تكرارها، هذا يعني أنه كلما كانت الكلمة أكثر تكراراً كلما كانت أقرب إلى جذر شجرة FP tree. سنوضح ذلك من خلال المثال البسيط التالي، ليكن لدينا جدول افتراضي وسنطبق عليه خوارزمية FP Growth من أجل min_sup =1.

الجدول الافتراضي: البيانات المخزنة في الجدول

ID	word1	word2
1	drink	tea
2	buy	tea

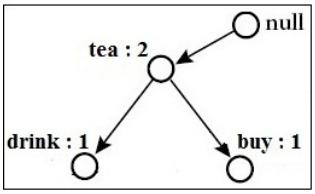
هذا الجدول يتضمن سجلين لتتالي فعل ومفعول به. عند المرور الأول على الجدول نحصل على العناصر المتكررة وتكراراتها، وبعد ترتيب العناصر المتكررة ترتيباً تنازلياً بحسب التكرار نحصل على اللائحة L الموضحة في الجدول (1):

الجدول (1)

لائحة العناصر المتكررة عند تطبيق خوارزمية FP Growth

item id	support
tea	2
Drink	1
Buy	1

بعد ذلك نحصل على شجرة العناصر المتكررة FP tree الموضحة بالشكل (2):



الشكل (2): شجرة FP tree بعد تطبيق خوارزمية FP Growth

من شجرة FP tree السابقة نحصل على قواعد الترابط التالية الممثلة بفروع الشجرة:

if word1='tea' then word2='drink'
if word1='tea' then word2='buy'

نلاحظ أن القواعد السابقة أظهرت الترابط بين الكلمات مع تجاهلها لتسلسل ورود هذه الكلمات، لكن موضوع ترتيب الكلمات يعتبر أساسياً في بحثنا لأنه يؤثر على إعطاء المعنى الدلالي الصحيح للجملة. هذا دفعنا إلى تعديل خوارزمية FP Growth كي نحصل على قواعد ترابط تعطي أهمية لتسلسل ورود الكلمات ضمنها.

التعديل على خوارزمية FP Growth كان من خلال تعديل الخطوة الأولى من الخوارزمية أثناء المرور الأول على قاعدة البيانات لإيجاد 1-itemsets مع تكراراتها، وتخزينها في اللائحة L، بحيث يتم ترتيب العناصر المخزنة في اللائحة L بحسب عناصر العمود الأول من الجدول فعناصر العمود الثاني فالثالث وهكذا (هذا الترتيب يضمن ترتيب العناصر بحسب تسلسل ورودها أثناء المرور الثاني على قاعدة البيانات)، كما شمل التعديل أيضاً حذف الخطوة 3 من الخوارزمية (ترتيب العناصر المتكررة ترتيباً تنازلياً بحسب قيمة تكرارها). وبالتالي فإن خطوات تنفيذ خوارزمية FP Growth المعدلة تتم وفق ما يلي:

1. المرور الأول على قاعدة البيانات لإيجاد 1-itemsets مع تكراراتها، وتخزينها في اللائحة L.
2. ترتيب العناصر المخزنة في اللائحة L بحسب عناصر العمود الأول من الجدول فعناصر العمود الثاني فالثالث وهكذا.
3. الحصول على العناصر المتكررة Frequent items (عناصر تكرارها >= min_sup).
4. المرور الثاني على قاعدة البيانات لبناء شجرة النموذج المتكرر FP tree.
5. التنقيب في شجرة FP tree للحصول على قواعد الترابط الممثلة بفروع شجرة FP tree (قواعد النموذج الشرطي للعقد الأوراق مع الأوراق)، حيث أن القواعد الناتجة عن هذه الفروع هي القواعد الدلالية المطلوب إيجادها.

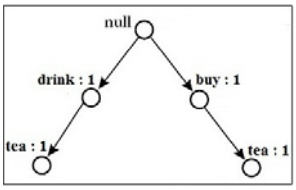
عند تطبيق خوارزمية FP Growth المعدلة على المثال السابق نحصل على اللائحة L الموضحة في الجدول (2):

الجدول (2)

لائحة العناصر المتكررة عند تطبيق خوارزمية FP Growth المعدلة

item id	support
Drink	1
Buy	1
Tea	2

ونحصل على شجرة العناصر المتكررة FP tree الموضحة بالشكل (3):



الشكل (3): شجرة FP tree بعد تطبيق خوارزمية FP Growth المعدلة

من شجرة FP tree السابقة نحصل على قواعد الترابط التالية:

if word1='drink' then word2='tea'

if word1='buy' then word2='tea'

نجد أن القواعد السابقة أعطت أهمية لتسلسل ورود الكلمات ضمنها، وهذا السبب الأساسي لتعديلنا لخوارزمية FP Growth بحيث تلائم وتناسب مسألة التحليل الدلالي لجمل اللغة الانكليزية.

آلية استثمار القواعد الدلالية الناتجة عن خوارزمية FP Growth المعدلة

سنقوم باستخدام القواعد الدلالية (قواعد الترابط الناتجة عن تطبيقنا لخوارزمية FP Growth المعدلة) في عملية التحقق من أن الجملة الصحيحة قواعدياً هي جملة صحيحة من الناحية الدلالية، وسنطبق ذلك على جمل اللغة الانكليزية التي يمكن ردها إلى إحدى الحالتين التاليتين:

- فاعل فعل مفعول به (Subject Verb Object).
- فاعل فعل تتمة الجملة (Subject Verb Complement).

ستكون تتمة الجملة إما عبارة اسمية أو حرف جر مع اسم مجرور، وقد اكتفينا بهاتين الحالتين لإثبات فعالية خوارزمية FP Growth المعدلة في مجال التحليل الدلالي، ولتوصيف آلية استثمار القواعد الناتجة عن الخوارزمية بشكل عام، ويمكن معالجة حالات إضافية لجمل اللغة الانكليزية بأسلوب مماثل لأسلوب معالجتنا للحالتين السابقتين.

إن قواعد الترابط الناتجة عن تطبيق خوارزمية FP Growth المعدلة ستمثل القواعد الدلالية، ويجب أن تحقق ما يلي:

- تتضمن كلمات مهمة مثل الأفعال والأسماء.
- لا تحوي كلمات غير مهمة مثل أدوات التعريف والتنكير.
- أن تكون الكلمات ضمن القواعد مرتبة حسب تسلسل ورودها.

لذلك سيكون نموذجنا للتحليل الدلالي مؤلفاً من خمس مجموعات لقواعد الترابط هي:

1. الصفة (واحدة أو أكثر) وما يليها من أسماء (adjective → nouns).
2. الاسم وما يليه من أفعال (noun → verbs).

3. الفعل وما يليه من أسماء (verb → nouns).
4. الفعل وما يليه من أحرف جر (verb → prepositions).
5. حرف جر وما يليه من أسماء (preposition → nouns).

تعتبر القواعد السابقة قواعد ترابط هامة لأنه تم حذف الكلمات غير المهمة من ناحية المعنى منها. من الممكن أن يتم اعتبار هذه القواعد عبارة عن قواعد نحوية، لكن بشكل فعلي ستمثل قواعداً دلاليةً لأنه ليس بالضرورة أن يكون لكل جملة صحيحة قواعدياً معنى دلالي صحيح. يوضح الجدول (3) أمثلة لجمل صحيحة قواعدياً وغير صحيحة دلالياً.

الجدول (3)

أمثلة لجمل غير صحيحة دلالياً وجمل صحيحة دلالياً

جمل صحيحة قواعدياً وغير صحيحة دلالياً			جمل صحيحة قواعدياً وصحيحة دلالياً		
م	الجملة	قواعد دلالية خاطئة	م	الجملة	قواعد دلالية صحيحة
1	I drink black milk	black → milk adjective → noun	1	I drink hot milk	hot → milk adjective → noun
2	The car eats the apple	car → eat noun → verb	2	The man eats the apple	man → eat noun → verb
3	I eat water	eat → water verb → noun	3	I eat bread	eat → bread verb → noun
4	I come in Syria	come → in verb → preposition	4	I come from Syria	come → from verb → preposition
5	I was playing football in Friday	in → Friday preposition → noun	5	I was playing football on Friday	on → Friday preposition → noun

نجد من الجدول السابق على سبيل المثال إن حالة adjective → nouns، تمثل قاعدة نحوية صحيحة لكن ليس بالضرورة أن تكون هذه القاعدة صحيحة دلالياً دائماً كما هو واضح في المثال الأول، حيث أن black milk قاعدة نحوية صحيحة لكنها غير صحيحة دلالياً. وكذلك من أجل الحالات الأخرى الموضحة في الجدول السابق. لذلك بحثنا عن قاعدة بيانات تتضمن الكلمات المترافقة (المتسلسلة) الأكثر تكراراً في اللغة الانكليزية التي سنتحدث عنها بشكل تفصيلي في الفقرة (4 - 5).

فيما يلي خطوات التحقق من أن الجملة الصحيحة قواعدياً هي جملة صحيحة من الناحية الدلالية:

أولاً: حذف أدوات الربط بين الجمل conjunction مثل (and، but، ...).

ثانياً: تقطيع الجملة الصحيحة قواعدياً إلى عناصرها الإعرابية (فاعل، فعل، مفعول به، ...).

ثالثاً: تحديد أجزاء الكلام لكل عنصر (الصنف الإعرابي لكل جزء).

رابعاً: رد العناصر الإعرابية إلى أصلها، ويتم ذلك من خلال:

- حذف أدوات التعريف والتنكير (the، a، an).
- حذف صفات الملكية (my، our، his، ...).
- رد الاسم إلى حالة المفرد إذا كان في حالة الجمع.
- رد الفعل إلى التصريف الأول (الجذر).

خامساً: رد الجمل إلى إحدى الحالتين:

- فاعل فعل مفعول به.
- فاعل فعل تتمة الجملة.

سادساً: مناقشة جميع حالات الكلمات المترابطة مع مراعاة تسلسل ورودها. هذه الحالات يمكن تلخيصها كما يلي:

- في حالة الفاعل أو المفعول به إذا كانا ممثلين بعبارة اسمية أو في حالة العبارة الاسمية في تتمة الجملة أو الاسم المجرور: عندئذ إذا كانت العبارة الاسمية تتضمن صفة (أو أكثر) يليها اسم نتحقق من وجود قاعدة دلالية ضمن المجموعة الأولى من القواعد الدلالية (adjective → nouns).
- في حالة الفاعل والفعل: نأخذ الاسم فقط من الفاعل ونتحقق من وجود قاعدة دلالية ضمن المجموعة الثانية من القواعد الدلالية (noun → verbs).
- في حالة الفعل والمفعول به أو الفعل وتتمة الجملة إذا كانت عبارة اسمية: نأخذ الاسم فقط من المفعول به أو تتمة الجملة ونتحقق من وجود قاعدة دلالية ضمن المجموعة الثالثة من القواعد الدلالية (verb → nouns).

- في حالة الفعل وحرف الجر نتحقق من وجود قاعدة دلالية ضمن المجموعة الرابعة من القواعد الدلالية (verb → prepositions).
- في حالة حرف الجر والاسم المجرور: نأخذ الاسم فقط من الاسم المجرور ونتحقق من وجود قاعدة دلالية ضمن المجموعة الخامسة من القواعد الدلالية (preposition → nouns).

سابعاً: إذا توفرت جميع القواعد الدلالية لجميع الحالات الموجودة في الجملة عندئذ تكون الجملة صحيحة من الناحية الدلالية.

يمكن توضيح الآلية السابقة من خلال المثال التالي:

لدينا الجملة التالية The tall men drank their hot tea، هذه الجملة صحيحة قواعدياً، وسنتحقق من أنها صحيحة من الناحية الدلالية كما يلي:

- نقطع الجملة إلى عناصرها الإعرابية: الفاعل the tall men والفعل drank والمفعول به their hot tea.
- نحدد أجزاء الكلام لكل عنصر:

subject: the (article) tall (adjective) men (plural noun)
verb: drank (verb in past form)
object: their (possessive adjective) hot (adjective) tea (noun)

- رد العناصر الإعرابية إلى أصلها:

subject: tall man
verb: drink
object: hot tea

- مناقشة جميع حالات الكلمات المترابطة:

1. نتحقق من وجود القاعدتين tall man وhot tea ضمن المجموعة الأولى من القواعد الدلالية (adjective → nouns).
2. نتحقق من وجود القاعدة man drink ضمن المجموعة الثانية من القواعد الدلالية (noun → verbs).
3. نتحقق من وجود القاعدة drink tea ضمن المجموعة الثالثة من القواعد الدلالية (verb → nouns)

- نحكم بأن الجملة صحيحة من الناحية الدلالية بسبب وجود القواعد السابقة ضمن مجموعات القواعد الدلالية لنموذجنا في التحليل الدلالي.

إعداد قاعدة البيانات التي سيتم تطبيق خوارزمية FP Growth عليها

قمنا بتطبيق خوارزمية FP Growth على قاعدة بيانات تم الحصول عليها من الموقع The Corpus of Contemporary American English لجامعة Brigham Young. يتضمن هذا الموقع ملفات نصية تحوي كلمات متسلسلة مع أجزاء الكلام (PoS) part of speech (الصنف الإعرابي مثل فعل، اسم، صفة، حرف جر، ...) الموافقة لكل كلمة. هذه الملفات تتضمن حوالي مليون سجل للكلمات المترافقة الأكثر تكراراً في اللغة الانكليزية لكل حالة (كلمتين متسلسلتين 2-grams، ثلاث كلمات متسلسلة 3-grams، أربع كلمات متسلسلة 4-grams، خمس كلمات متسلسلة 5-grams) (Daves, 2011). يوضح الجدول (4) عينة عشوائية لعشرة سجلات من الملف الذي يحوي ثلاث كلمات متسلسلة:

الجدول (4)

عينة عشوائية لعشرة سجلات من الملف الذي يتضمن ثلاث كلمات متسلسلة

frequency	word1	word2	word3	pos1	pos2	pos3
24	deep	blue	sea	jj	jj	nn1
30	eat	and	drink	vv0	cc	vv0
24	thin	young	man	jj	jj	nn1
98	when	you	eat	cs	ppy	vv0
40	new	position	of	jj	nn1	io
74	thin	red	line	jj	jj	nn1
55	good	in	my	jj	ii	appge
142	went	on	sale	vvd	ii	nn1
79	new	high	school	jj	jj	nn1
27	go	about	our	vv0	ii	appge

يحتوي الجدول السابق على الواصفات attributes التالية:

- التكرار frequency: عدد صحيح يعبر عن عدد مرات تكرار الكلمات، أصغر قيمة للتكرار ضمن البيانات هي القيمة 24.

- الأعمدة word1 وword2 وword3: تتضمن الكلمات الثلاثة متسلسلة حسب تسلسل ورودها.
- الأعمدة pos1 وpos2 وpos3: تخزن أجزاء الكلام للكلمات الأولى والثانية والثالثة على الترتيب. يمكن توضيح الاختصارات الممثلة لبعض أجزاء الكلام في الجدول (5) مرتبة حسب الترتيب الأبجدي:

الجدول (5)

الاختصارات الممثلة لبعض أجزاء الكلام مرتبة حسب الترتيب الأبجدي

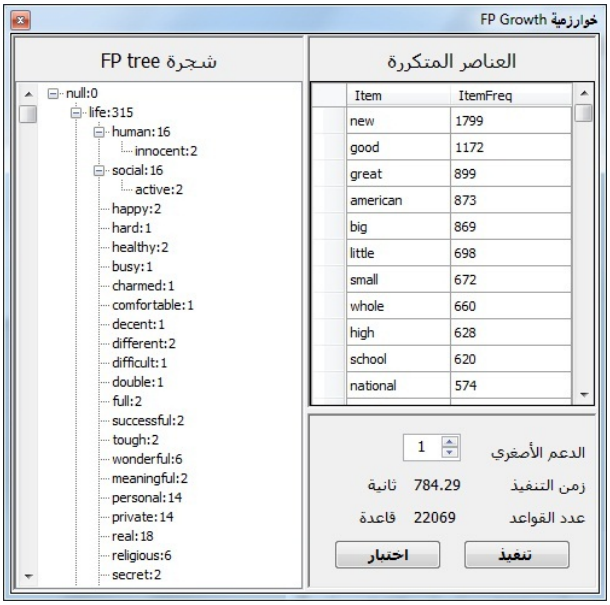
الرقم	الرمز	توصيفه
1	appge	صفة الملكية possessive pronoun
2	cc	أداة ربط coordinating conjunction
3	cs	أداة ربط تابعية subordinating conjunction
4	ii	حرف جر عام general preposition
5	io	of كحرف جر (as preposition) of
6	jj	صفة عامة general adjective
7	nn1	اسم مفرد singular noun
8	ppy	ضمير المخاطب 2nd person personal pronoun
9	vv0	الشكل الأساسي للفعل base form of lexical verb
10	vvd	الزمن الماضي للفعل past tense of lexical verb

قمنا بعد تحميل هذه الملفات من الموقع بتصديرها إلى قاعدة بيانات SQL Server 2008، وبعد ذلك أنشأنا قاعدة بيانات تتضمن جداول توافق المجموعات الخمس للقواعد الدلالية هي: الجدول adjective Noun يخزن سجلات للصفات وما يليها من أسماء، والجدول noun Verb يخزن سجلات للأسماء وما يليها من أفعال، والجدول verb Noun يخزن سجلات لأفعال وما يليها من أسماء، والجدول verb Preposition يخزن سجلات لأفعال وما يليها من أحرف جر، والجدول preposition Noun يخزن سجلات لأحرف جر وما يليها من أسماء. على سبيل المثال للحصول على سجلات

الجدول Verb noun قمنا باختيار select وإدراج insert جميع السجلات من جدول الكلمتين المترافقتين التي تحقق 'pos1='nn1 و'pos2='vv0، أما من جدول الكلمات الثلاثة المترافقة فقمنا باختيار وإدراج السجلات التي تحقق 'pos1='nn1 و'pos2='vv0 أو 'pos2='nn1 و'pos3='vv0 أو 'pos1='nn1 و'pos3='vv0 بشرط أن يكون pos2 مثلاً أداة تعريف أو تنكير أو صفة ملكية من أجل الحالة الأخيرة. وبأسلوب مماثل قمنا باختيار السجلات المناسبة للجداول الأربعة الأخرى وإدراجها فيها.

برمجة وتنفيذ خوارزمية FP Growth:

قمنا ببرمجة خوارزمية FP Growth باستخدام Microsoft Visual Studio 2010 (C#). يوضح الشكل (4) الواجهة الأساسية للتطبيق البرمجي المنجز:



الشكل (4): الواجهة الأساسية للتطبيق البرمجي

بعد الضغط على الزر تنفيذ تظهر العناصر المتكررة (بعد المرور الأول على قاعدة البيانات) مع تكرارها مرتبة ترتيباً تنازلياً بحسب قيم تكرارها في القسم الأيمن من الواجهة. كما تظهر في القسم الأيسر من الواجهة شجرة FP tree حيث أن فروع هذه الشجرة تمثل القواعد الدلالية الناتجة (قواعد الترابط بين الكلمات). يتيح التطبيق

للمستخدم إمكانية تغير قيمة الدعم الأصغري min_sup الذي يأخذ قيمة افتراضية تساوي 1. كما يتم عرض الزمن اللازم لتنفيذ الخوارزمية وعدد القواعد الدلالية الناتجة (فروع الشجرة) التي تم استخلاصها من شجرة FP tree. إضافة إلى ذلك يسمح الزر "اختبار" بإمكانية التحقق والتأكد من صحة القواعد الناتجة من خلال مقارنتها مع سجلات قاعدة البيانات الأصلية، حيث تظهر الواجهة الموضحة بالشكل (5) بعد الضغط على الزر المذكور:

الشكل (5) واجهة تبين نتائج اختبار خوارزمية FP Growth

توضح الواجهة السابقة عدد القواعد الصحيحة ونسبتها، بالإضافة إلى عدد القواعد الخاطئة ونسبة الخطأ، وتُظهر أيضاً نسبة ضغط قاعدة البيانات. حيث يتم حساب نسبة القواعد الصحيحة ونسبة القواعد الخاطئة ونسبة ضغط قاعدة البيانات بالقوانين الثلاثة التالية على الترتيب:

نسبة القواعد الصحيحة = [عدد القواعد الصحيحة/ عدد القواعد الناتجة]* 100

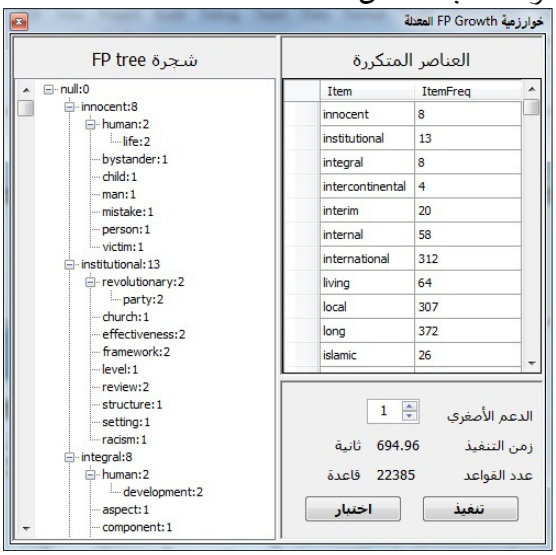
نسبة القواعد الخاطئة = [عدد القواعد الخاطئة/ عدد القواعد الناتجة]* 100

نسبة ضغط قاعدة البيانات = [(عدد سجلات قاعدة البيانات - عدد القواعد الناتجة)/ عدد سجلات قاعدة البيانات]* 100

تعديل خوارزمية FP Growth ودراسة النتائج بعد التعديل:

خوارزمية FP Growth تعطي قواعد الترابط ولكنها تتجاهل تسلسل ورود العناصر ضمنها، وبما أن نظامنا يركز على تسلسل ورود العناصر ضمن قواعد الترابط الناتجة ويعتبره مهماً لإنجاز التحليل الدلالي، لذلك تم تعديل خوارزمية FP Growth للحصول على قواعد ترابط تعطي أهمية لتسلسل ورود العناصر ضمنها هذه القواعد.

بعد إنجاز التعديلات على خوارزمية FP Growth الموضحة في الفقرة (4 - 3) وإعادة برمجتها من جديد ننفذ الخوارزمية المعدلة فنحصل على شجرة FP tree الموضحة بالواجهة الموضحة بالشكل (6):



الشكل (6) واجهة تبين شجرة القرار الناتجة بعد تنفيذ خوارزمية FP Growth المعدلة

لاختبار صحة القواعد الناتجة نضغط الزر اختبار كما هو موضح بالواجهة المبينة بالشكل (7):

الشكل (7) واجهة تبين نتائج اختبار تنفيذ خوارزمية FP Growth المعدلة

نلاحظ من الشكل (8) أن نسبة القواعد الصحيحة قبل إجراء التعديل على الخوارزمية 61.50% ونسبة القواعد الخاطئة 38.50%، كما نلاحظ من الشكل (8) أن نسبة القواعد الصحيحة بعد إجراء التعديل على الخوارزمية 97.00% ونسبة القواعد الخاطئة 3.00%. بالمقارنة بين هذه النسب نجد أن نسبة القواعد الصحيحة بعد التعديل أصبحت أكبر ونسبة القواعد الخاطئة أصبحت أقل.

النتائج والمناقشة

للتأكد من صحة النتائج التي توصلنا إليها قمنا بتطبيق خوارزمية FP Growth وخوارزمية FP Growth المعدلة على قاعدة البيانات عدة مرات من أجل قيم مختلفة للدعم الأصغري min_sup هي (1، 2، 5، 10، 12)، وتم ذلك على الجدول adjective Noun الذي يتضمن الصفات (صفة أو صفتين) مع الأسماء التي تليها وقد طبقنا ذلك على 50,000 سجل من الجدول. يوضح الجدول (6) نتائج تنفيذ خوارزمية FP Growth وخوارزمية FP Growth المعدلة:

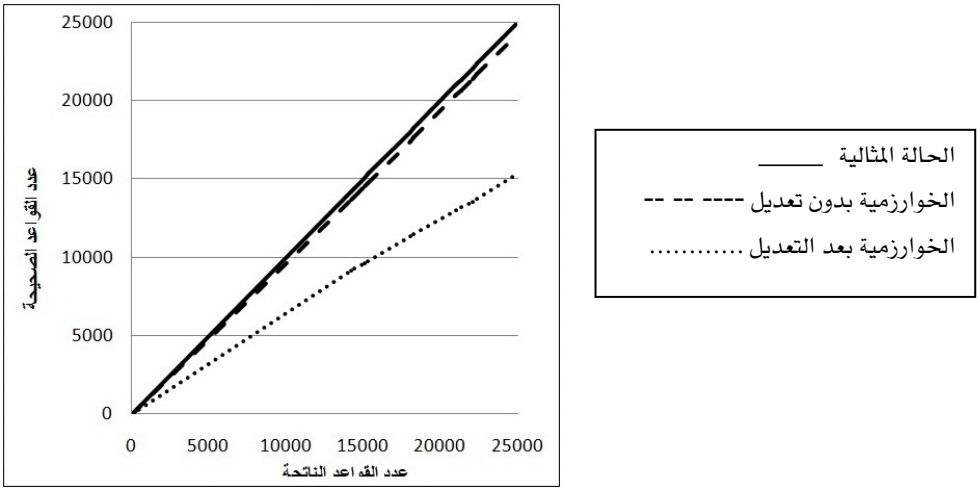
الجدول (6)

نتائج تنفيذ خوارزمية FP Growth وخوارزمية FP Growth المعدلة

	الدعم الأصغري	زمن التنفيذ (ثانية)	عدد القواعد الدلالية	عدد القواعد الصحيحة	عدد القواعد الخاطئة	الخطأ (%)	ضغط قاعدة البيانات (%)
خوارزمية FP Growth	1	784.29	22069	13572	8497	38.50	55.86
	2	870.51	21022	13029	7993	38.02	57.96
	5	765.43	18023	11359	6664	36.97	63.95
	10	757.61	15107	9646	5461	36.15	69.79
	12	755.62	14228	9136	5092	35.79	71.54
خوارزمية المعدلة FP Growth	1	694.96	22385	21713	672	3.00	55.23
	2	692.24	21338	20685	653	3.06	57.32
	5	674.90	18268	17642	626	3.89	63.46
	10	668.06	15321	14725	596	3.89	69.36
	12	667.45	14430	13856	574	3.98	71.14

عند دراسة النتائج الموضحة في الجدول السابق نجد أن نسبة القواعد الخاطئة بعد تعديل الخوارزمية أصبحت أقل بحوالي من 31.81% إلى 35.50% عن نسبة القواعد الخاطئة عند تنفيذ الخوارزمية بدون تعديل، كما نلاحظ أن زمن تنفيذ الخوارزمية المعدلة أقل من زمن تنفيذ الخوارزمية غير المعدلة، وأيضاً نلاحظ تقارب في نسبة ضغط قاعدة البيانات عند تنفيذ الخوارزميتين.

إذا أردنا أن نوضح العلاقة بين عدد القواعد الناتجة وعدد القواعد الصحيحة نحصل على الرسم البياني المعطى بالشكل (8):



الشكل (8): رسم بياني للعلاقة بين عدد القواعد الناتجة وعدد القواعد الصحيحة

حيث أن المحور الأفقي يمثل عدد القواعد الناتجة عند التنفيذ، بينما المحور العمودي يمثل عدد القواعد الصحيحة. الحالة المثالية (القواعد الناتجة جميعها قواعد صحيحة) تم رسمها بخط مستمر، أما خوارزمية FP Growth بدون تعديل فقد تم رسمها بخط منقط، وخوارزمية FP Growth المعدلة تم رسمها بخط متقطع. نستنتج من الرسم البياني السابق أن الخوارزمية المعدلة أقرب إلى الحالة المثالية.

نستنتج أيضاً من الجدول (10) أنه كلما كانت قيمة الدعم الأصغري أكبر كلما كان عدد القواعد الناتجة أقل ويكون زمن التنفيذ أقل، ويعود السبب في تناقص عدد القواعد الناتجة إلى أن عدد العناصر المتكررة أقل، وعندئذ نحصل على القواعد الأكثر أهمية لأنها تتضمن الكلمات الأكثر تكراراً، ولكن ذلك يؤدي إلى إهمال بعض القواعد. بينما إذا كانت قيمة الدعم الأصغري أصغر نحصل على عدد أكبر من القواعد لأنه لا يتم حذف عدد كبير من الكلمات عند تحديد العناصر المتكررة، وبالتالي من الأفضل أن تكون قيمة الدعم الأصغري صغيرة حتى نحصل على نتائج أكثر دقة لأننا سنشمل عدداً أكبر من القواعد.

حصلنا على نتائج مماثلة عند تنفيذ خوارزمية FP Growth المعدلة على الجداول الأربعة الأخرى في قاعدة البيانات، وبذلك قمنا ببناء نموذج للتحليل الدلالي مؤلف من القواعد المستخلصة من خمس أشجار FP tree بعد تنفيذ الخوارزمية المعدلة على الجداول الخمسة في قاعدة البيانات. إضافة إلى ذلك يعتبر النموذج الناتج عن تطبيق الخوارزمية المعدلة على قاعدة البيانات نظاماً متعلماً لأننا نستطيع تحديث وإعادة بناء هذا النموذج عند توفر بيانات جديدة مضافة إلى قاعدة البيانات.

مما سبق نجد أن خوارزمية FP Growth المعدلة أصبحت أكثر دقة ووثوقية في الحصول على قواعد ترابط تعطي أهمية لتسلسل ورود العناصر ضمنها، وبالتالي يمكن اعتبار الخوارزمية FP Growth المعدلة فعالة في مجال التحليل الدلالي وبناء القواعد الدلالية.

الخلاصة

قدمنا في بحثنا هذا نموذجاً لتوليد القواعد الدلالية لجمل اللغة الانكليزية يُستخدم في التحقق من معاني الجمل، ويوضح أسلوب بنائه اعتماداً على خوارزمية FP Growth التي تقوم بإيجاد النماذج المتكررة وقواعد الترابط، ويتم ذلك من خلال بناء شجرة FP tree ثم الحصول على القواعد الدلالية المتمثلة بفروع الشجرة. ونظراً لأن هذه الخوارزمية تعطي قواعد ترابط تتجاهل تسلسل ورود العناصر ضمنها تم إنجاز تعديل عليها للحصول على قواعد ترابط تعطي أهمية لتسلسل ورود العناصر ضمنها حيث أن هذا التعديل قادنا للحصول على خوارزمية أكثر دقة ووثوقية في الحصول على قواعد الترابط.

تم تطبيق الخوارزمية المعدلة في مجال التحليل الدلالي وذلك لبناء القواعد الدلالية لاستثمارها في التحقق من صحة الجمل من ناحية المعنى، وهذا يجعل من الحاسب أداة تحاكي الإنسان الخبير باللغة الإنكليزية. من خلال سياق عملنا سنقوم بتطبيق النتائج التي حصلنا عليها في بناء المدقق اللغوي المتكامل الذي ينطلق من مرحلة التحليل اللفظي والمنتهي بالتحليل الدلالي. من أهم النقاط التي يمكن لنا أن نفتح بها آفاق هذا البحث هي باستثماره في دعم اللغة العربية حيث أنه هو هدفنا المنشود مستقبلاً.

المراجع

Charniak, E. 1985. Introduction to Artificial Intelligence. Pearson Education.

Daves, M. 2011. N-grams data from the Corpus of Contemporary American English (COCA), Downloaded from http://www.ngrams.info on May 04, 2012.

Han, J. and Pei, J. 2000. Mining Frequent Patters by Pattern Growth: Methodology and Implications, ACM SIGKDD, United States of America.

Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques. 2nd ed, Elsevier Inc, United States of America.

Hayashi, Y. 1990, A neural expert system with automated extraction of fuzzy if-then rules, Advances in Neural Information Processing Systems 3: 578-584.

Kane, R. and Milgram, M. 1993, Extraction of semantic rules from trained multilayer neural networks, LRP-CNRS, Paris-VI Univ., France, IEEE International Conference on Neural Networks, 3: 1397 - 1401.

Manning, CH. and Schutze H. 1999. Foundations of Statistical Natural Language Processing. 2nd ed, Massachusetts Institute of Technology, London.

Masuika, R., Watanabe, N., Kawamura, A., Owada, Y. and K 1990, Neuro fuzzy system. Fuzzy inference using a structured neural network," in Proceeding of the international Conference on Fuzzy Logic €4, Asakawa, Neural Networks 173-177,.

Motoda, H. and Ohara, K. 2009. Apriori. In: (ed) Wu X., Kumar V. The Top Ten Algorithms in Data Mining, Taylor and Francis Group, LLC ,United States of America, pp: 61-92.

Witten, L., Frank, E. and Hall, M. 2011. Data Mining Practical Machine Learning Tools and Techniques, 3th ed, Elsevier Inc, United States of America.

Yamuna Devi, N., and Devi Shree J. 2013, A novel approach and comparative study of association rule algorithms in validation of semantics of sentences. International Journal of Computer Applications 62 (3): 0975 – 8887, January 2013.

Semantic Rules Extraction from a Database using Data Mining

Ahmad Badreddin Alkhodre[(1)], Mohammed Alhamid[(1)], Lina Baaj[(2)]

(1) Department of Software Engineering, Faculty of Informatics Engineering, University of Aleppo, Aleppo, Syria

(2) Department of Computer Engineering, Faculty of Electrical and Electronic Engineering, University of Aleppo, Aleppo, Syria

Abstract

Automatic English text semantic validation is rarely explored, in the recent decade. In this context the target of this paper is to demonstrate a model for generating semantic rules from a database containing the most frequent words in the English to validate the meaning of English sentences. To construct this model we used FP Growth algorithm, where the association rules between words in the database are extracted. The output of this algorithm ignores the sequence of words within the rules. But the proposed system focuses on the sequence of words within the rules and considers it necessary to achieve semantic analysis, so the FP Growth algorithm was modified to get association rules which give importance to the sequence of words. Since the natural languages processing is constantly evolving process, this research makes the achievement of semantic analysis applied by data mining, and it is considered as an important step to display the importance of Informatics researches in building systems which deal with natural languages and simulate the human expert in English.

Key Words: Association rules, database, data mining,, FP Growth algorithm, semantic analysis, semantic rules.