Classification of Web Pages by Using Particle Swarm Optimization Algorithm

Muhammad Hamid Abdulraheem **Ghayda A.A. Al-Talib**
ghaydabdulaziz@uomosul.edu.iq
Computer and Internet Center *College of Computer Sciences and Mathematics*
University of Mosul, Mosul, Iraq

Received on: 29/06/2011 **Accepted on: 03/10/2011**

ABSTRACT

As the amount of information available on the internet grows so does the need for more effective data analysis methods. This paper utilizes the particle swarm optimization (PSO) algorithm in the field of web content classification, and used part of speech tagging algorithm to reduce the large numbers of attributes associated with web content mining. The proposed algorithm gave a good classification accuracy, which comparable to the accuracy of Ant-miner algorithm and acquire less training time.

Keywords: web content classification, PSO,

تصنيف صفحات الشبكة العنكبوتية باستخدام خوارزمية أمثَلَة عناصر السرب

غيداء عبد العزيز الطالب **محمد حامد عبد الرحيم**
كلية علوم الحاسوب والرياضيات *مركز الحاسوب والانترنت*
جامعة الموصل، الموصل، العراق *جامعة الموصل، الموصل، العراق*

تاريخ استلام البحث: 2011/06/29 **تاريخ قبول البحث: 2011/10/03**

الملخص

مع نمو المعلومات المتوفرة على الشبكة العالمية Internet ازدادت الحاجة إلى طرائق أكثر فاعلية في تحليل البيانات، في هذا البحث استخدمت خوارزمية أمثَلَة عناصر السرب (PSO) بأسلوب جديد في تصنيف المحتوى النصي لصفحات الشبكة العنكبوتية مع استخدام خوارزمية وسم أجزاء الكلام بوصفها وسيلة لتقليل السمات، واظهر أسلوب الخوارزمية المقترح دقة تصنيف جيدة ومقاربة لدقة تصنيف خوارزمية النمل في التنقيب Ant-Miner التي طبقت في المجال نفسه وبزمن تدريب اقل.

الكلمات المفتاحية: تصنيف المحتوى النصي لصفحات الشبكة العنكبوتية، خوارزمية أمثَلَة عناصر السرب

1. المقدمة

أصبحت المعلومات هي سمة العصر الذي نعيشه حتى أن البعض أطلق عليه عصر المعلومات، وتكتسب المعلومات أهميتها من انتشارها وقيمتها وسهولة تبادلها. ولا يمكن أن يتم تبادل هذه المعلومات دون وجود شبكات نقل المعلومات والتي تؤدي إلى سرعة وسهولة الحصول على المعلومة أو نشرها بغض النظر عن بعدها الجغرافي. فقد شاع استخدام الشبكة العنكبوتية في العقود الأخيرة ووفرت لنا منصة قوية لنشر واسترجاع وتحليل المعلومات، وتعد الشبكة العنكبوتية مخزن بيانات كبير يتألف من أنواع مختلفة من البيانات والتي تختفي فيها الكثير من المعرفة، وبسبب النمو السريع والكبير في كمية المعلومات، فأن المستخدمين غالباً ما يواجهون مشكلة فيض المعلومات. وصعوبة إيجاد معلومات مناسبة ودقيقة. وبالتالي أصبح الاهتمام والتحدي الكبير لباحثي الشبكة العنكبوتية هو تشخيص مسألة كفاءة وفعالية أنظمة إدارة واسترجاع المعلومات المبنية على الشبكة العنكبوتية [18]. وخلال هذا البحث سوف توضّح إمكانية استغلال تقنيات ذكاء الأسراب (swarm intelligence) في تصنيف صفحات الشبكة بناءً على موضوعاتها.

2. الدراسات السابقة

عام 2004 استخدم الباحثان (Holden and Freitas) [8]، خوارزمية (Ant-Miner) في مجال تصنيف محتوى الشبكة العنكبوتية، واظهر البحث أن خوارزمية (Ant-Miner) كانت أكثر فعالية من خوارزمية التصنيف (C5.0)، وأيضا تم التحقق من جدوى بعض تقنيات المعالجة المبنية على اللغة لتقليل عدد السمات. وكان أول تطبيق لخوارزمية أمثَلة عناصر السرب في التصنيف عام 2004 من قبل (Sousa et al) [17]، إذ اقترحوا استخدام خوارزمية أمثَلة عناصر الأسراب بوصفها أداة جديدة للتنقيب في البيانات ومثلوا ثلاثة نسخ لخوارزمية أمثَلة عناصر السرب وهم: خوارزمية أمثَلة عناصر الأسراب المتقطعة (discreet pso)، وخوارزمية أمثَلة عناصر السرب ذات الانحدار الخطي للأوزان (linear decreasing weight pso) وخوارزمية أمثَلة عناصر السرب المحددة (Constricted pso). وقورنت نتائجها مع الخوارزمية الجينية وخوارزمية شجرة الاستقراء، وتبين أن خوارزمية أمثَلة عناصر السرب كانت أداة مرشحة ومنافسة لمهام التصنيف.

3. التنقيب في الشبكة العنكبوتية (Web Mining)

يهدف التنقيب في الشبكة العنكبوتية إلى اكتشاف المعرفة من مصادر بيانات ضخمة متوفرة على الشبكة العنكبوتية من خلال استخدام أساليب التنقيب في البيانات [14]. فالتنقيب في الشبكة العنكبوتية هو تحويل الشبكة العنكبوتية إلى بيئة أكثر فائدة ويمكن لمستخدمها إيجاد المعلومات التي يحتاجها بسهولة وسرعة، وتتضمن كذلك اكتشاف وتحليل البيانات، والوثائق والوسائط المتعددة في الشبكة العنكبوتية [16]. ويمكن تصنيف التنقيب في بيانات الشبكة العنكبوتية إلى ثلاثة أنواع بناءً على أهداف التنقيب والذي يحدد التخصص الذي سوف يُنقب فيه من الشبكة العنكبوتية:

1. التنقيب في محتوى الشبكة العنكبوتية Web content mining
2. التنقيب في هيكل الشبكة العنكبوتية Web structure mining
3. التنقيب في استخدام الشبكة العنكبوتية Web usage mining

التنقيب في محتوى الشبكة العنكبوتية يحاول اكتشاف معلومات قيّمة من محتويات الشبكة العنكبوتية، وبصورة عامة محتوى الشبكة العنكبوتية يشير إلى المحتويات النصية وأيضا بوصفه مصطلحاً بديلاً يطلق عليه التنقيب في النص (text mining)، ومن ابرز مهام التنقيب في المحتوى هما التصنيف والعنقدة، أما التنقيب في هيكل الشبكة العنكبوتية فأنه يضم نمذجة مواقع الشبكة العنكبوتية من حيث هياكل الارتباطات التشعبية. إن معلومات الربط المتبادل التي يتم الحصول عليها يمكن أن تستخدم لإيجاد صفحات بناءً على التشابه أو الصلة بالموضوع بين الصفحات. أما التنقيب في استخدام الشبكة العنكبوتية فأنه يحاول الكشف عن نمط الوصول للشبكة العنكبوتية من خلال بيانات جلسة المستخدم (user data session) المسجلة في ملف أداء الشبكة العنكبوتية (Web log file) [11]،[18].

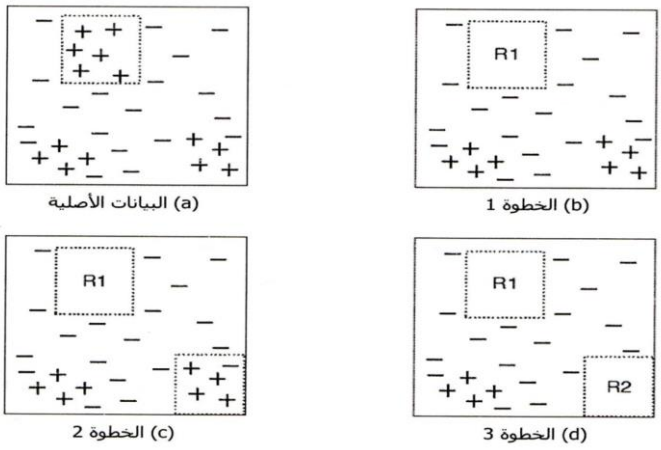
4. أهمية وتطبيقات تصنيف صفحات الشبكة العنكبوتية [14]

تصنيف صفحات الشبكة العنكبوتية، هي عملية تخصيص الصفحة إلى واحد أو أكثر من الأصناف المعرفة مسبقاً، والتصنيف غالبا ما يطرح بوصفه مسألة تعليم بالأشراف والذي فيه مجموعة من بيانات محددة تستخدم لتدريب المصنف والذي بدوره يطبق لتصنيف أمثلة غير محددة. وتبرز أهمية التصنيف للصفحات في الأمور الآتية:

1. إنشاء وإدامة وتوسيع أدلة البحث على الشبكة العنكبوتية.
2. تحسين نوعية نتائج البحث: إذ إن الغموض في الاستعلام هو من بين المشاكل التي تقوض نتائج عملية البحث.
3. الأنظمة المساعدة لإجابة الأسئلة (Helping question answering systems): في هذه الأنظمة ربما تستخدم تقنية التصنيف لتحسين جودة الجواب.
4. الباحث الموجه (focused crawling): إن الباحث الموجه حول موضوع محدد يهتم بالوثائق التي لها علاقة بمجموعة من المواضيع المحددة مسبقا.

5. الطرق المباشرة لاستخلاص قواعد التصنيف

تستخدم غالباً خوارزمية التغطية المتسلسلة (sequential covering algorithm) لاستخلاص قواعد تصنيف من البيانات مباشرة. تنمو القواعد بطريقة مفرطة (greedy) وفق مقياس تقييم محدد. فتستخلص خوارزمية القواعد صنفاً في كل مرة من مجموعة البيانات التي تحوي أكثر من صنفين. يعرض الشكل (1) آلية عمل خوارزمية التغطية المتسلسلة لمجموعة البيانات التي تتضمن مجموعة من الأمثلة الإيجابية والسلبية. تبدأ خوارزمية التغطية المتسلسلة بقائمة قواعد فارغة، ويتم استخلاص القاعدة R_1 (تظهر تغطيتها في الشكل (1-b)) أولاً لأنها تغطي أكبر جزء من الأمثلة الإيجابية، ومن ثم تتم إزالة كافة تسجيلات التدريب المشمولة بـ R_1 وتستمر الخوارزمية في البحث عن أفضل قاعدة تالية، وهي R_2. تكون القاعدة مقبولة إذا كانت تغطي معظم الأمثلة الإيجابية ولا تغطي أياً من أو القليل جداً من الأمثلة السلبية، وبمجرد إيجاد قاعدة كهذه تتم إزالة تسجيلات التدريب المشمولة بالقاعدة، وتتم إضافة القاعدة الجديدة إلى أسفل قائمة القواعد. يتكرر هذا الإجراء إلى أن يتحقق معيار توقف، تستمر الخوارزمية بعدها لتوليد قواعد من أجل الصنف التالي [1].



الشكل (1). مثال عن خوارزمية التغطية المتسلسلة [1].

6. استخدام المعالجة اللغوية في تقليل واختيار السمات

1-6 وسم أجزاء الكلام Part-of-speech-Tagging [10]

يعرف وسم أجزاء الكلام من خلال تخصيص وسم أو ملحق مفرد لكل كلمة في نص اللغة الانكليزية حسب الصنف اللغوي الذي تنتمي إليه الكلمة، وفي اللغة الانكليزية تصنف الكلمات إلى: الأسماء، الأفعال،

الضمائر، حروف الجر، الظروف، أدوات الربط، الصفات، أدوات التعريف. إن الإدخال لخوارزمية الوسم هي سلسلة من الكلمات ، ويكون الإخراج أفضل وسم مفرد لكل كلمة ، كما في المثال الآتي:

The/DT guys/NNS that/WDT make/VBP traditional/JJ hardware/NN are/VBP really/RB being/VBG obsolete/VBN by/IN microprocessor-based/JJ machines/NNS ,/, ”/” said/VBD Mr./NNP Benton/NNP ./.

2-6 إرجاع الكلمات إلى جذورها(stemming) [7]

وهي عملية اختزال صيغ مختلفة للكلمة إلى جذر مشترك، وخوارزمية التجذير هي الخوارزمية التي تزيل وتغير ملحقات الكلمات وكمثال الكلمتان bottles و bottle سوف تجذر إلى الجذر مشترك bottle والكلمات growled وgrows و growth سوف تجذر إلى grow.

7. استخدام معامل كسب المعلومات (information gain) في تقليل السمات [4]

في مسائل التصنيف يحسب معامل كسب المعلومات لسمات مسألة التصنيف بالاعتماد على معامل التباين (Entropy)، بوصفه مرحلة مسبقة لغرض تقليل تلك السمات. ويعرف معامل التباين بأنه مقياس نظري لدرجة التباين الموجودة في بيانات التدريب، بسبب وجود أكثر من تصنيف واحد للحالات. فإذا كان هنالك k من الأصناف يمكن الدلالة على نسبة الحالات أو الأمثلة بالصنف i إلى جميع الحالات بالنسبة Pi، وان معامل التباين لبيانات التدريب يرمز له E ويحسب من خلال المعادلة (1):

$$E = -\sum_{i=1}^{k} Pi \log_2 \quad \ldots(1)$$

إذ أن: i تمثل الصنف وتتغير من 1 إلى k من الأصناف.

Pi هي عدد مرات حدوث الصنف i مقسوماً على العدد الكلي للحالات والذي هو رقم بين 0 و 1.

إن القيمة ($-p_i \log_2 p_i$) تكون موجبة لقيم p_i تقع بين الصفر والواحد.

8. خوارزمية أمثَلَة عناصر السرب Particle Swarm Optimization

يقوم سرب الطيور على المبادئ الأساسية الآتية [3]:

1. التجانس Homogeneity: كل طائر في السرب له أنموذج التصرف نفسه، فالسرب يتحرك بدون قائد.
2. المحلية Locality: حركة كل طائر فقط تتأثر بحركة السرب الأقرب للطائر.
3. تجنب التصادم Collision Avoidance: البقاء بالقرب من رفاق السرب مع عدم التصادم.
4. تطابق السرعة Velocity Matching: المحاولة لتطابق السرعة مع الجار القريب في السرب.
5. انضمام السرب Flock Centering: محاولة البقاء بقرب الرفاق المجاورين في السرب.

إن خوارزمية أمثَلَة عناصر السرب طورت بناء على النموذج الآتي: [15]

1. عندما يحدد احد الطيور في السرب الهدف أو الطعام (أو اكبر قيمة لدالة الهدف) فان الطائر حالا ينقل المعلومات لباقي الطيور.
2. باقي الطيور تنجذب إلى الهدف أو الطعام تدريجيا.
3. هنالك إمكانيات لكل طائر منها التفكير الذاتي و تذكر أفضل موقع سابق.

إذاً فان هذا النموذج يحاكي عملية بحث عشوائي في فضاء بحث للوصول إلى أفضل قيمة لدالة الهدف، وبتكرار ذلك فان الطيور ستصل تدريجيا للهدف.

9. خوارزمية أمثَلَة عناصر السرب المتقطعة (DPSO: The discrete PSO Algorithm) [19،5،6]

إن خوارزمية (DPSO) تعالج المتغيرات المتقطعة (السمات) مباشرة، وتختلف عن خوارزمية (PSO) القياسية في أن مجموعة الحلول المرشحة تحوي على عناصر بأحجام مختلفة، هنالك N من العناصر في السرب، وحجم كل عنصر يتغير من 1 إلى n، إذ إن n هو عدد السمات في المسألة، وفي هذا السياق فأن حجم العنصر يرمز إلى عدد السمات المختلفة التي بمقدور العنصر تمثيلها. فمثلاً في خوارزمية (DPSO) ربما يحدث أن العنصر Z(i) في المجموعة له حجم 6 أي Z(i) = {*,*,*,*,*,*}، بينما عنصر آخر Z(j) في نفس المجموعة له الحجم Z(i) = {*,*} أي 2، وهكذا لباقي العناصر. وكل عنصر Z(i) يحتفظ بسجل لأفضل موقع زاره على الإطلاق، وتلك المعلومة تخزن في متجه آخر يسمى B(i)، كذلك يحتفظ السرب بسجل لأفضل موقع عمومي وتخزن تلك المعلومة في متجه مستقل يسمى G، وهو يمثل أفضل B(i) في السرب.

9-1 ترميز العنصر في خوارزمية (DPSO) :

تمثل كل سمة بفهرس، وتتغير تلك الأرقام من 1 إلى n، فالعنصر عبارة عن مجموعة فرعية من الفهارس غير مرتبة وبدون تكرار فمثلا العنصر k قد يمثل بالفهارس وكما يأتي:

$Z(k) = \{2,4,18,1\}, \quad k \in \{1,2,\ldots,N\}$.

9-2 توليد المجتمع الأولي في خوارزمية (DPSO)

لإنشاء المجتمع الأولي الذي تبدأ به الخوارزمية، فأن N من المتجهات بحجم n تولد عشوائياً، وفي خوارزمية أمثَلَة عناصر السرب الثنائية(BPSO: binary PSO)، تكون قيمة أبعاد المتجه إما 0 أو 1، كل متجه Y(i) يولد بصورة مستقلة، وكل بعد في المتجه Y(i) يحسب من خلال سحب رقم عشوائي **φ** في الفترة (0,1)، فإذا كان **φ** >0.5 عندها Y(i,d)=1 وإلا فإن Y(i,d)= 0.

إن طريقة توليد المجتمع الأولي في خوارزمية (DPSO) مطابقة لطريقة توليد المجتمع الأولي في خوارزمية (BPSO) وكما يلي:

إن فهرس كل سمة لها قيمة 1 تنسخ إلى الحل الجديد أو العنصر، فمثلا إذا كان العنصر الأولي المولد بطريقة (BPSO) يساوي Y(k) = (1,0,1,1,0)، فأنه يحول إلى المتجه Z(k) = {1,3,4} في خوارزمية (DPSO) كتهيئة أولية للعناصر، بسبب أن قيم السمات A1,A3,A4 في المتجه تساوي 1. إن ابتداء العناصر Z(i) بهذه الطريقة تجعل العناصر تأخذ أحجاماً مختلفة من أبعاد السمات، وليس حجماً واحداً يضم كل السمات. وعندما يأخذ العنصر هذا البعد في التهيئة الأولية يبقى الحجم ثابتاً خلال تنفيذ الخوارزمية، فمثلا العنصر Z(k)= (2,3,4,5) الذي ابتدأ بأربعة فهارس للسمات سوف يحمل دائماً أربعة فهارس، ومن المحتمل أن تتغير أرقام الفهارس في كل وقت يحدِّث العنصر موقعه.

3-9 تمثيل السرعة بواسطة الاحتمالية النسبية

إن خوارزمية (DPSO) لا تستخدم متجه السرعة كما في خوارزمية (PSO) القياسية، فهي تتعامل مع الاحتمالية النسبية. إن خوارزمية (DPSO) تستخدم المصفوفة M(i) لتمثيل الاحتمالية النسبية التي تقابل السمات المختارة. وكل عنصر في خوارزمية (DPSO) يترافق مع مصفوفة ذات بعد (2*n)، حيث يمثل 2 عدد الصفوف وn عدد الأعمدة. حيث إن عدد الأعمدة يعادل عدد سمات المسألة وتمثل قيم الصف الأول والثاني كما يأتي:

$$M(i) = \begin{Bmatrix} proportiord - likelihood - row \\ attribute \quad - \quad index \quad - \quad row \end{Bmatrix}$$

كل بعد في الصف الأول لهذه المصفوفة يمثل احتمالية اختيار السمة التي تقابلها في الصف الثاني، وهذا يعني وجود علاقة واحد إلى واحد بين الصف الأول والصف الثاني للمصفوفة، بدايةً فأن جميع القيم في الصف الأول للمصفوفة M(i) تأخذ قيمة 1، فمثلا إذا كان لدينا خمس سمات تكون المصفوفة M(i) كما يأتي:

$$M(i) = \begin{pmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

وبعد التوليد الأولي للعناصر، تُحدّث تلك المصفوفة دائماً قبل اشتقاق العنصر الجديد المرتبط مع تلك المصفوفة. إن تحديث الاحتمالية في المصفوفة يبنى على المتجهات الثلاثة Z(i)، B(i)، G وكذلك على ثلاثة أوزان تحديث $\propto, \beta, \gamma$ ، إن هذه الأوزان تحدد قوة مشاركة المتجهات المقابلة لها، في تعديل الاحتمالية في المصفوفة M(i)، وان قيم الأوزان الثلاثة تحدد من قبل المستخدم. إن مشاركة تلك الأوزان لتحديث الاحتمالية في المصفوفة M(i) تتم كما يأتي:

جميع الفهارس الموجودة في المتجه Z(i) تزيد الاحتمالية النسبية للفهارس المقابلة لها في المصفوفة M(i) للفهارس بمقدار $\propto$. إضافة إلى ذلك فأن جميع الفهارس الموجودة في المتجه B(i) تزيد الاحتمالية النسبية في المصفوفة M(i) للفهارس التي تقابلها في المصفوفة بمقدار β ، والشيء نفسه بالنسبة للمتجه G الذي يزيد الاحتمالية النسبية في المصفوفة M(i) للعناصر التي يحتويها G بمقدار γ . فمثلا:

لو كانت n=5، $\propto = 0.01, \beta = 0.12, \gamma = 0.14$ وان G={5,2} B(i)={3,5,2}, Z(i) = {2,3,4} والمصفوفة $M(i)$:

$$M(i) = \begin{pmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

فان تحديث الاحتمالية في المصفوفة يكون كالتالي:

$$M(i) = \begin{pmatrix} (1.0) & (1.0+\propto+\beta+\gamma) & (1.0+\propto+\beta) & (1.0+\propto) & (1.0+\beta+\gamma) \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

إن المصفوفة الناتجة الجديدة M(i) تحل محل المصفوفة القديمة، وتُستخدم لتوليد عنصر جديد، وهو ما يقابل تحديث الموقع في خوارزمية (PSO) القياسية.

4-9 تحديث موقع العناصر في خوارزمية (DPSO):

إن المصفوفة M(i) تستخدم لتحديث أو توليد حالة جديدة من العنصر Z(i) المرتبط مع المصفوفة M(i)، فهنالك سلسلة من العمليات تنجز على مصفوفة الاحتمالية، ففي البداية يضرب كل رقم احتمالية في الصف

الأول في المصفوفة برقم عشوائي بين(0,1)، حيث يسحب رقم عشوائي جديد لكل رقم في الصف الأول. ولتوضيح ذلك نفترض أن المصفوفة :

$$M(i) = \begin{pmatrix} 1.00 & 1.36 & 1.22 & 1.10 & 1.26 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

عندها يضرب الصف الأول بالأرقام العشوائية وكما يلي:

$$M(i) = \begin{pmatrix} (1.00 \times \varphi 1) & (1.36 \times \varphi 2) & (1.22 \times \varphi 3) & (1.10 \times \varphi 4) & (1.26 \times \varphi 5) \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

إذ إن (φ1 , …. , φ5) هي أرقام عشوائية منتظمة مسحوبة بصورة مستقلة في الفترة (0,1) وعلى افتراض أن المصفوفة الناتجة تكون:

$$M(i) = \begin{pmatrix} 0.11 & 0.86 & 0.57 & 0.62 & 1.09 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

وأن الموقع الجديد لقيم الفهارس المرتبطة مع قيم الاحتمالية يعرف من خلال ترتيب الاحتمالية النسبية في الصف الأول تنازلياً، وعندها تصبح المصفوفة M(i) كالتالي:

$$M(i) = \begin{pmatrix} 1.09 & 0.86 & 0.62 & 0.57 & 0.11 \\ 5 & 2 & 4 & 3 & 1 \end{pmatrix}$$

العملية التالية على المصفوفة هي اختيار الفهارس من الصف الثاني التي ستشكل موقع العنصر الجديد، بعد ترتيب المصفوفة يختار عدد Si من الفهارس من الصف الثاني من اليسار إلى اليمين لتمثل تحديث موقع العنصر الجديد، إذ إن Si يمثل حجم العنصر Z(i) المترافق مع المصفوفة M(i).

وعلى افتراض أن العنصر Z(i) المترافق مع المصفوفة M(i) له حجم {*,*,*}=Z(i) أي انه Si=3، فأن أول ثلاثة فهارس من الصف الثاني سوف تختار لتشكيل العنصر الجديد ويصبح: {5,2,4} = Z(i) حيث نلاحظ انه الفهارس التي لها احتمالية نسبية أعلى هي أكثر احتمالية لان تختار، وهكذا لكل عناصر السرب.

10. التحويرات والإضافات إلى خوارزمية (DPSO) مع أمثَلَة تطبيقية ونتائج من البحث

10-1 طريقة الابتداء للعناصر الأولية للسرب

إن جميع خوارزميات أمثلة عناصر السرب المطبقة تبتدئ المجتمع الأولي لها من خلال قيم عشوائية يأخذها عناصر السرب، أي أن القاعدة الواحدة التي تمثل عنصراً من عناصر السرب تولد أولاً عشوائياً، وهذا قد يجعل الوصول للحل الأمثل يستغرق وقتاً أو قد لا يصل إلى الحل أبداً في مسألة توليد قوانين التصنيف.

وعندما نلاحظ خوارزمية النمل Ant-miner نرى أن توليد قوانين الابتداء لا يتم عشوائياً، وإنما بالاعتماد على قيمة (η) التي تمثل دالة التخمين والتي تتحكم في مسألة اختيار السمة في أول جولة، حيث تكون قيمة (τ) التي تمثل إفراز النمل في البداية متساوية لجميع المسالك، ثم مع مرور الزمن تتحكم قيمة (τ) بالسيطرة على اختيار مسار بنود القاعدة للمسألة. إذاً كما في خوارزمية النمل من الممكن أن نبتدئ، خوارزمية أمثَلَة عناصر السرب المتقطعة بالاعتماد على قيمة دالة تخمين في اختيار تجمعات القواعد المولدة ابتداءً.

10-2 دالة التخمين المضافة إلى الخوارزمية (DPSO)

دالة التخمين التي استخدمت في هذا البحث تُحسب لكل بند (سمة=قيمة) في مسألة التصنيف، فهي تعبر عن أهمية البند للمسألة، ولإغراض التنقيب في البيانات فأن تلك الأهمية للسمة تقاس عادة بعدد الحالات التي

غُطّت بواسطة قيمة سمة معينة مع صنف ثابت، لذلك يمكن تعريف قيمة دالة التخمين لأهمية البند في المسألة كما في المعادلة رقم (1):

$$\eta_{i,k} = \frac{|\ V_i = Value_k\ \&\ class\ C\ |}{|\ V_i = Value_k\ |} \quad \ldots(1)\ [12]$$

إذ إن: $\eta_{i,k}$ تمثل قيمة الأهمية للسمة i عندما تأخذ قيمة معينة وعند ورودها مع صنف ثابت.

وهذا يعني انه كل بند في بيانات التدريب يأخذ عدد من قيم لـ η بقدر عدد الأصناف في المسألة.

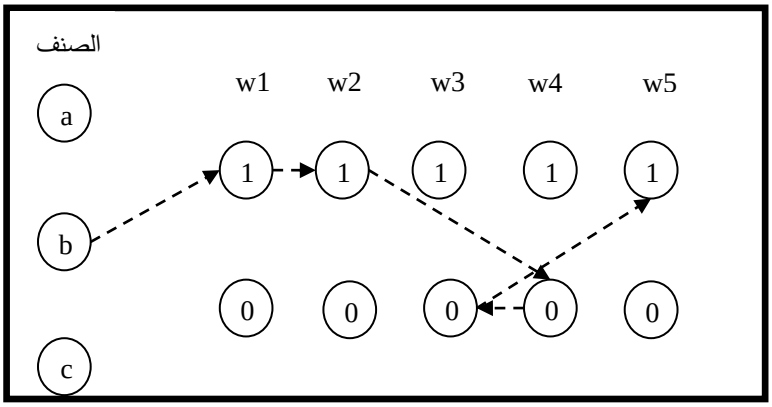
يتم حساب دالة التخمين مسبقاً لكل بند، وبقدر عدد الأصناف في المسألة كما موضح في المثال الآتي:

لو كان لدينا بيانات التدريب الممثلة بالجدول (1) إذ إن Wi تمثل الكلمة أو السمة والـ di تمثل صفحة شبكة. وقيم السمات في الجدول هي إما ورود الكلمة (1) أو عدم ورودها (0).

جدول (1). مصفوفة التدريب

	W1	W2	W3	W4	W5	W6	الصنف
d1	1	1	0	0	0	0	a
d2	0	1	1	1	0	0	b
d3	1	1	0	0	0	1	a
d4	1	0	1	0	1	1	c
d5	0	0	0	0	1	1	c
d6	0	0	1	1	0	0	b

ويمكن تمثيل فضاء أو مخطط البحث لهذا المثال كما في الشكل (2).



الشكل (2). فضاء (مخطط) البحث

وفي هذا المخطط الدائري المتجه، أي رأس في المخطط ممكن أن يرتبط مع أي رأس آخر وفي مسارات مخطط لا يتكرر المرور على نفس السمة أكثر من مرّة في المسار الواحد. ولحساب قيمة دالة التخمين η لرؤوس المخطط في هذا المثال، نلاحظ وجود ثلاثة أصناف في المسألة ولذلك هنالك ثلاثة قيم مع كل رأس وتحسب كما يأتي:

للرأس (W1=1) تحسب قيمة η مع الصنف a كما يأتي: $\eta = \frac{2}{3}$

إذ إن السمة W1 بقيمة تساوي 1 وردت مع الصنف a مرتين، وان السمة w1 بقيمة 1 تكرر ثلاث مرات بغض النظر عن الصنف. ولحساب قيمة دالة التخمين η للرأس (W1=1) مع الصنف b كما يأتي: $\eta = \frac{0}{3} = 0$

إذ إن السمة W1 بقيمة 1 لم ترد مع الصنف b، وان السمة W1 بالقيمة 1 تكررت ثلاث مرات.

ومع الصنف c أيضاً تحسب η للرأس (W1=1) $\eta = \frac{1}{3}$

إذ إن السمة W1 بالقيمة 1 وردت مع الصنف c مرة واحدة، وان السمة W1 بالقيمة 1 تكررت ثلاث مرات.

من خلال هذا المثال نلاحظ انه في حالة الصنف (a) لإنشاء القاعدة الابتدائية له، فان الرأس (W1=1) أو وجود الكلمة هي مرشحة لان تضاف لقاعدة اشتقاق الصنف a.

والجدول (2) يمثل قيم η لكل الرؤوس ولكل الأصناف في المثال السابق لثلاث سمات مع قيمتها.

الجدول (1). يمثل قيم دالة التخمين لبنود المسألة

السمة / الصنف	W1=1	W1=0	W2=1	W2=0	W3=1	W3=0
a	2/3	0	2/3	0	0	2/3
b	0	2/3	1/3	1/3	2/3	0
c	1/3	1/3	0	2/3	1/3	1/3
السمة / الصنف	W4=1	W4=0	W5=1	W5=0	W6=1	W6=0
a	0	2/4	0	2/4	1/3	1/3
b	1	0	0	2/4	0	2/3
c	0	2/4	1	0	2/3	0

عندما تبدأ الخوارزمية بإعطاء القيم الأولية لعناصر السرب، تبتدئ الخوارزمية بصنف معين مثلاً (a) ويتم أيجاد عدة قوانين بقدر عناصر السرب تخص الصنف (a) وفيها يتم التعامل مع قيم دالة التخمين η الخاصة بالصنف a فقط.

3-10 توليد قواعد التصنيف

بعد حساب قيمة دالة التخمين η لكل رأس في فضاء البحث للمسألة ، يتم حساب احتمالية اختيار كل تجمع أو رأس ($P_{i,k}$) وحسب المعادلة (2):

$$P_{i,k} = \frac{\eta_{i,k}}{\sum_{i=1}^{a} X_i * \sum_{k=1}^{bi} \eta_{i,k}} \quad \ldots(2)$$

إذ أن: i تمثل السمة.

k تمثل القيمة التي تأخذها تلك السمة.

$\eta i,k$ تمثل قيمة دالة التخمين للتجمع.

Xi تأخذ قيمة 1 إذا كانت السمة i لم تستخدم بعد في القاعدة وتأخذ قيمة 0 عدا ذلك.

a تمثل عدد السمات الكلية في المسألة.

bi تمثل عدد القيم التي تأخذها السمة i.

إن هذه المعادلة تكرر عند اختيار كل سمة يراد إضافتها للقاعدة، ولغرض تطبيق القاعدة على المثال السابق لإيجاد قاعدة تصنيف للصنف (a)، تحسب الاحتمالية بالاعتماد على قيم (η) للصنف (a) فقط وحسب الجدول (3):

جدول (3). قيم دالة التخمين والاحتمالية لبنود المسألة مع الصنف a

التجمع	W1=1	W1=0	W2=1	W2=0	W3=1	W3=0
قيم η *للصنف* a	2/3	0	2/3	0	0	2/3
$P_{i,k}$	0.182	0	0.182	0	0	0.182
التجمع	W4=1	W4=0	W5=1	W5=0	W6=1	W6=0
قيم η *للصنف* a	0	2/4	0	2/4	1/3	1.3
$P_{i,k}$	0	0.136	0	0.136	0.1	0.1

فمثلاً للتجمع (W1=1) يتم حساب Pi,k كما يأتي وبالاعتماد على المعادلة (2):

$$P_{w1=1} = \frac{\frac{2}{3}}{\frac{2}{3}+\frac{2}{3}+\frac{2}{3}+\frac{2}{4}+\frac{2}{4}+\frac{1}{3}+\frac{1}{3}} = (0.182)$$

وبعد حساب الاحتمالية لكل التجمعات يتم تطبيق طريقة عجلة الروليت (Roulette wheel) لاختيار إحدى التجمعات لأضافتها للقاعدة، وبعد اختيار تجمع معين، تعاد حسابات الاحتمالية من جديد لاستثناء السمة التي اختيرت سابقاً، إلى أن تكتمل القاعدة حسب دالة توليد عشوائية لطول القاعدة. ومن خلال الجدول السابق قد يولد القاعدة الآتية:

If W2=1 and W4=0 and W1=1 then a

إن هذه القاعدة تمثل احد عناصر السرب، وبعد توليد جميع عناصر السرب بهذه الطريقة يكون قد تكون لدينا سرب من الطيور يبحث عن القاعدة الأمثل لصنف معين، وان هذا السرب قد أخذ قيم أولية غير عشوائية.

4-10 تحوير في طريقة اختيار بنود القاعدة في خوارزمية (DPSO)

فمن خلال ملاحظة خوارزمية (DPSO) في المصادر [19, 6, 5] نجد أن الخوارزمية (DPSO) طُبّقت لاختيار سمات وليس تجمع (سمة =قيمة)، وهذا الاختيار للسمة يأتي كعملية مسبقة لخوارزميات التصنيف التقليدية.

إذاً يمكن توسعة هذا الأسلوب لخوارزمية (DPSO) عن طريق الأمور الآتية:

1. عدم الاعتماد على خوارزمية(BPSO) في توليد العناصر الأولية للسرب،وإنما تولد حسب الأسلوب السابق.
2. ترميز العنصر: في خوارزمية (BPSO) كل عنصر يمثل بمتجه من الأرقام الصحيحة التي تمثل فهرس السمة، ولكون إننا نهتم بالسمات والقيم التي تأخذها السمات، فيمكن إضافة متجه آخر يمثل القيم التي تأخذها فهرس تلك السمة، فمثلاً المتجه Z(i) يكون للفهارس والمتجه الآخر Z(ik) لقيم تلك الفهارس:

Z(i) = {2, 4, 18, 1}
Z(ik) = {0, 1, 1, 0}

هذا هو التمثيل البرمجي لعنصر السرب، وهو يكافئ التمثيل المنطقي للقاعدة الآتية:

If W2 = 0 and W4=1 and W18=1 and W1=0 Then a

وبعد توليد جميع العناصر في السرب، يقيم كل عنصر بواسطة دالة الجودة المستخدمة في مسائل التنقيب في البيانات. مثل تلك المستخدمة في مسائل التنقيب عن البيانات وحسب المعادلة الآتية.

$$Q = \frac{TP}{TP+FN} * \frac{TN}{FP+TN} \quad \ldots(3)[9]$$

إذ إن TP (true positive): عدد الحالات التي غطتها القاعدة في جزء الشرط والتي لها نفس الصنف في ناتج القاعدة.

FP (false positive): عدد الحالات التي غطتها القاعدة في جزء الشرط والتي لها صنف يختلف عن الصنف في ناتج القاعدة.

FN (false negative): عدد الحالات التي لم تغطها القاعدة في جزء الشرط ولكن لها نفس الصنف في ناتج القاعدة.

TN (true negative): عدد الحالات التي لم تغطَ من قبل شرط القاعدة وليس لها نفس الصنف في ناتج القاعدة.

ويعد العنصر الذي له أفضل قيمة لدالة الجودة هو أفضل عنصر في السرب ويرمز له G، وأيضاً كل عنصر يحتفظ بأفضل قاعدة أو موقع وَلدّه سابقاً وهو يمثل أفضل موقع محلي للعنصر B(i).

5-10 تمثيل السرعة بواسطة الاحتمالية النسبية:

كما مَرّ سابقاً في خوارزمية DPSO فان كل عنصر يأتي معه مصفوفة M(i) بأبعاد (2*n) حيث أن n هي عدد السمات في المسألة.

ولكوننا نهتم بالسمة وبالقيمة التي تأخذها السمة إذاً تكون المصفوفة بأبعاد (2*2n) بسبب أن كل سمة ممكن أن تأخذ قيمتين إما 0 أو 1، إذ إن كل سمة في المسألة عند تمثيلها في المصفوفة فإنها تأخذ فهرسين فهرس لكل قيمة فمثلاً السمة رقم 2 في العنصر Z(i) في المثال السابق تمثل في المصفوفة M(i) من خلال الفهرسين الجديدين (4,5) والسمة رقم 1 تمثل بـ (2,3) والسمة رقم 0 تمثل بـ (0,1) فمثلاً لو كان عدد السمات في المسألة أربعة تكون المصفوفة M(i) كالآتي:

$$M(i) = \begin{Bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{Bmatrix}$$

حيث يمثل الصف الأول الاحتمالية النسبية كما مرّ سابقاً، أيضاً تحدث المصفوفة M(i) التالية لكل عنصر من خلال قيم γ, β, α كما ذكر سابقاً في خوارزمية (DPSO)، وبعدها يتم اشتقاق القاعدة الجديد بعد ترتيب القيم تصاعدياً، والقيم التي تختار من المصفوفة لتمثل العنصر الجديد ترجع إلى قيم الفهارس الأصلية، فمثلاً إذا كانت المصفوفة M(i) بعد جمع قيم الاحتمالية النسبية بالمعاملات γ, β, α وضربها برقم عشوائي أصبحت كالآتي:

$$M(i) = \begin{Bmatrix} 1.40 & 1.30 & 1.2 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 \\ 6 & 3 & 5 & 1 & 4 & 7 & 2 & 0 \end{Bmatrix}$$

وعلى افتراض أن طول العنصر السابق كان 3. إذاً يعاد توليد عنصر جديد بنفس الطول ولكن بفهارس جديدة، ففي المثال السابق يتولد العنصر الآتي:

Z(i) = {3, 1, 2}
Z(ik) = {1, 0, 0}

إن توليد هذا العنصر يتم من خلال. قسمة الفهرس المأخوذة من المصفوفة M(i) على الرقم 2 ويؤخذ الجزء الصحيح من الناتج ويوضع في متجه الفهرس، أما متجه القيمة المقابل لمتجه الفهرس فتحسب فيما إذا كان متبقي القسمة السابق صفر أم لا فإذا كان صفر مثلاً فان القيمة 1، وإذا كان هنالك باقي للقسمة فالقيمة 0.

6-10 ملخص الخوارزمية المقترحة

يمكن تلخيص الخوارزمية (dpso) المقترحة كما في المخطط الانسيابي في الشكل (3)، والتي تكتشف قاعدة تصنيف واحدة في كل تنفيذ، وحسب الخطوات الآتية:

1.إدخالات الخوارزمية:وتشمل مصفوفة التدريب والمعامل(عدد الطيور) والأوزان(γ,β,α) والصنف الحالي(C).

2. توليد المجتمع الأولي للسرب بطريقة تخمينية: يولّد سرباً من القواعد (الطيور) وهذه القواعد تولدّ عن طريق الاحتمالية لقيم الدالة التخمينية η وحسب الصنف الحالي، وليست عشوائيا، ولضمان بساطة وشمولية القواعد الناتجة، تولّد أطوال القواعد عشوائيا إلى طول أقصى معين، وبواسطة دارة تتكرر بقدر الرقم العشوائي الناتج، وفي كل تكرار تحسب الاحتمالية لكل البنود التابعة للسمات التي لم تختر سابقا، ويتم اختيار بند للقاعدة بطريقة عجلة الروليت، تكرر العملية السابقة بقدر عدد طيور السرب N، وتحسب جودة كل قاعدة وتخزن القواعد في القائمة L.

3. اختيار أفضل موقع محلي لكل طائر: يحتفظ كل طائر بأفضل موقع مرّ به، وبداية يأخذ هذا المتغير (RL) قيمة القاعدة الحالية (R).

4. اختيار أفضل قاعدة في عامّة السرب(RG):بالاعتماد على أعلى قيمة لجودة القواعد المخزنة في القائمة(L).

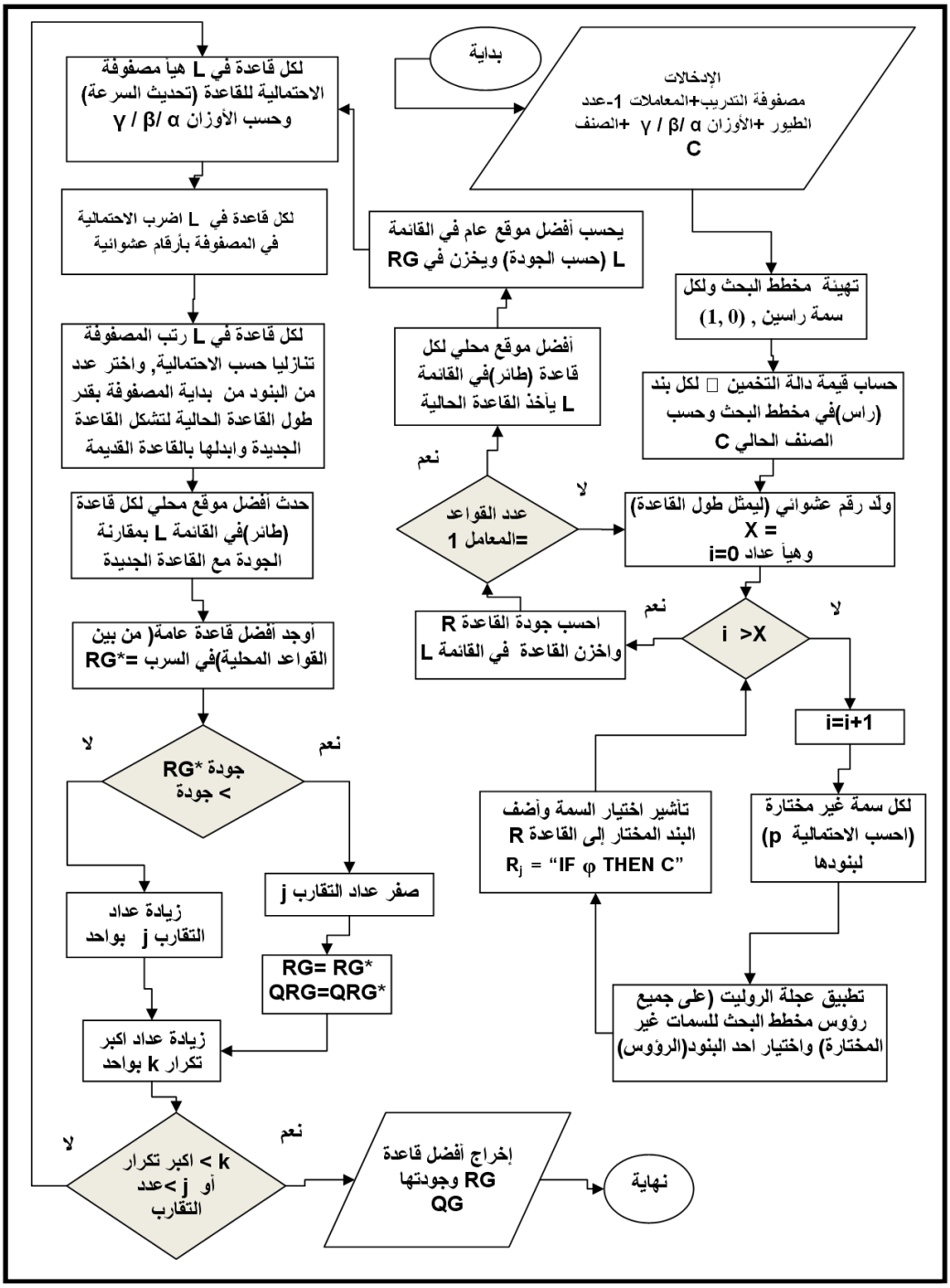
5. اعتماد مبدأ الطيور في البحث: ولكل قاعدة في القائمة L تحسب لها السرعة الجديدة ويحدّث موقعها أي توليد قاعدة جديدة، ويحدّث أفضل موقع محلي لكل طائر، إذ يتم تبديل أفضل موقع محلي بالموقع الجديد في حال كان أفضل جودة منه(RL=R).

6. توليد مجتمعات جديدة للسرب: بعد تحريك السرب في الخطوة السابقة (5) تكون القواعد قد تغيرت، إذ يتم حساب أفضل قاعدة في السرب من جديد (*RG) بالاعتماد على قيم دالة الجودة، وتقارن مع أفضل قاعدة من التشكيل السابق للسرب (RG)، فإذا كانت جودة (*RG) أفضل من (RG)، عندها يكون (RG =RG*) وكذلك الجودة (QRG=QRG*)، بحيث يحتفظ المتغير (RG) دائماً بأفضل قاعدة من بين القواعد العمومية لكل مجتمعات السرب، وفي حال لم تكن جودة (*RG) أفضل من (RG)، يتم فقط تحديث عداد التقارب ويستمر التكرار في دارة توليد مجتمعات السرب لجولة جديدة من تجديد السرعة وحساب الموقع الجديد لقواعد السرب، وتتوقف دارة توليد مجتمعات السرب في حال تحقق احد الشرطين، إما الوصول إلى أقصى عدد من التكرارات، أو حصول التقارب(عدم تحسن الجودة للقاعدة العامة في السرب لعدد من التكرارات المتعاقبة)، أي عندما يصل عداد التقارب إلى عدد محدد مسبقا للتقارب، وبعد الخروج من الدارة تُنتِج الخوارزمية أفضل قاعدة للصنف الحالي.

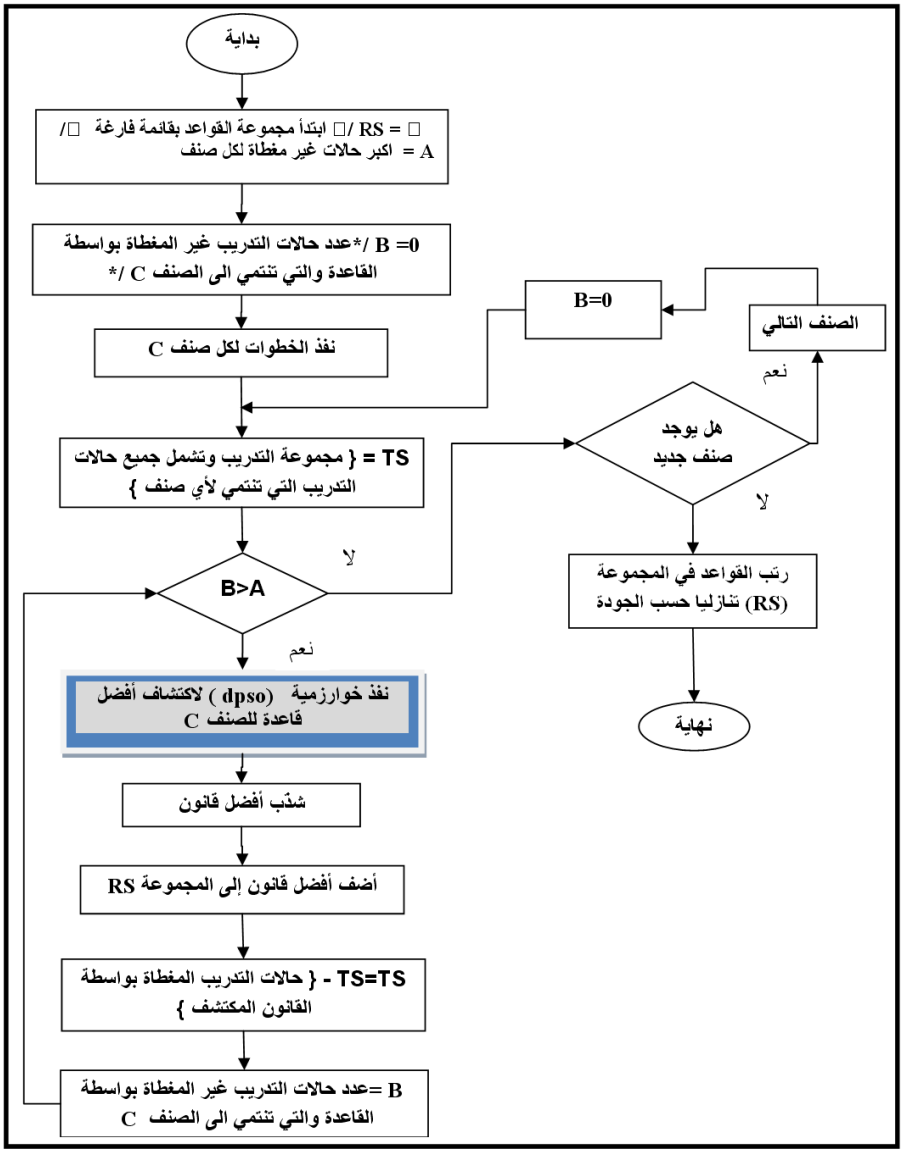
7-10 خوارزمية التغطية المتعاقبة المغلفة لخوارزمية (DPSO) [9]

يمكن استغلال خوارزمية التغطية المتعاقبة كما في المخطط الانسيابي للخوارزمية في الشكل (4)، لكي تكون الإطار الخارجي لخوارزمية (DPSO)، وفيها يمكن إيجاد قاعدة تصنيف واحدة في الزمن الواحد. حيث تبدأ خوارزمية التغطية من خلال تهيئة مجموعة القوانين (RS: rule set) بمجموعة فارغة، ثم لكل صنف تنجز الخوارزمية عدد من تكرارات الإيعاز (WHILE) وان كل تكرار ينفذ خوارزمية (DPSO) مرّة واحدة، والتي ترجّع أفضل قاعدة مكتشفة للصنف الحالي (C)، أما مجموعة التدريب (TS:training set) فتستخدم لخزن مجموعة بيانات التدريب والتي سوف تشتق منها قواعد التصنيف، بعد ذلك أفضل قاعدة مكتشفة تضاف إلى مجموعة

القواعد، بعد أن تمّ تشذيبها لإزالة البنود التي لأتشارك في تحسين جودة القاعدة، وأن الحالات التي غطتها القاعدة في مجموعة التدريب يتم إزالتها من مجموعة التدريب إذا طابقت جزء الشرط والناتج للقاعدة. يتم الاستمرار في التكرار (WHILE) ما دام عدد الحالات الغير مغطاة لذلك الصنف المحدد (C) هو اكبر من حد عتبة معين والمحدد من قبل المستخدم. وبعد الوصول إلى حد العتبة لذلك الصنف فأن مجموعة بيانات التدريب ترجع إلى حالتها الأولية التي تضم جميع حالات التدريب وتبدأ الخوارزمية بإيجاد قوانين التصنيف للصنف التالي، وبعد اكتمال جميع الأصناف في المسألة، ترتب قائمة القوانين المكتشفة حسب جودة القاعدة تنازليا.



الشكل (3). المخطط الانسيابي لخوارزمية DPSO المقترحة



الشكل (4). المخطط الانسيابي لخوارزمية التغطية المتعاقبة

11. نتائج التطبيق العملي

وتتضمن عدّة مراحل وهي:

11-1 تجميع وتهيئة بيانات التدريب

في مجال التنقيب في الشبكة العنكبوتية، لا توجد بيانات تعد قياسية لأغراض التدريب، لذلك ولغرض تهيئة بيانات التدريب، تم تجميع (300) صفحة شبكة في ثلاثة أصناف في المجالات (الصحة، والطعام، والبيئة) من دليلي البحث على الانترنت، http://dir.yahoo.com و http://www.dmoz.org مئة صفحة من كل صنف، وتم تخزين تلك الصفحات في ثلاث مجلدات، وكل مجلد باسم صنف. ويقوم البرنامج بقراءة الملفات من كل مجلد وبسحب النصوص داخل الصفحات من المناطق الآتية:

1. النص الموجود بين وسم العنوان في الصفحة <TITLE>……..</TITLE>

2. النص الذي يمثل الكلمات المفتاحية والموجود بعد كلمة المحتويات في الوسم الآتي:

<META NAME="keywords" CONTENT="……">

3. النص الذي يمثل الوصف لمحتويات الصفحة والموجود بعد كلمة المحتويات في الوسم الآتي:

<"META NAME="description" CONTENT="……">

حيث تشكل تلك المناطق الوصف الأدق لمحتوى صفحة الشبكة.

وبعد سحب النص من كل صفحة تجري العمليات التالية على النص:

1- تقطيع النص إلى كلمات (tokenized) حيث تزال جميع علامات التنقيط وتؤخذ فقط الكلمات بدون فراغات وأيضاً تسمى مصطلحات (terms).
2- جميع الحروف تحول إلى حالة الحروف الكبيرة أو حالة الحروف الصغيرة.
3- الكلمات ترجع إلى جذورها، وهذه العملية تسمى (stemming).
4- استخدام أحدى خوارزميات معالجة اللغة الطبيعية، والمسماة وسم أجزاء الكلام لعمل وسم لكلمات النص حسب تصنيف الكلمات، وبعد وسم الكلمات تسحب الأسماء فقط لأنها خير ما يعبر عن محتوى الوثائق كما في المصدر [8]، وترك باقي الكلمات.
5- حساب معامل كسب المعلومات للأسماء، وترتيبها تنازليا، واختيار عدد محدد من بداية القائمة لتمثيل سمات المسألة.

وبعد ذلك يتم تشكيل مصفوفة التدريب (الوثيقة–السمة)، والتي ستمثل فضاء البحث للمسالة، الطريقة التي استخدمت في البرنامج لتمثيل مصفوفة التدريب، هي عبارة عن مصفوفة ثنائية، وأن تقاطع (الصفحة – السمة) تمثل بـ 0 عند عدم ورود الاسم في تلك الصفحة أو 1 عند عدم ورود الاسم في الصفحة.

2-11 تهيئة المدخلات لخوارزمية (DPSO)

تحتاج خوارزمية (dpso) إلى قيم بعض المعاملات والأوزان التي تعد محددات في عملية توليد قواعد التصنيف وتشمل الآتي:

1. مصفوفة التدريب المولدة من مرحلة تهيئة البيانات.
2. الوزن α والذي يمثل وزن مشاركة موقع الطائر الحالي في تحديث سرعة الطائر= 0.1.
3. الوزن β والذي يمثل وزن مشاركة أفضل موقع محلي مرّ به الطائر في تحديث سرعة الطائر =0.13.
4. الوزن γ الذي يمثل وزن مشاركة أفضل موقع عمومي في السرب في تحديث سرعة الطائر =0.2.
5. أكبر حالات غير المغطاة لكل صنف في بيانات التدريب = 5.
6. عدد الطيور= 80.

اختيرت الأوزان (α، β، γ) بالاعتماد على القيم الواردة في المصدر [6]. ومن الطبيعي أن تكون الأوزان متدرجة، بحيث يأخذ الوزن γ أعلى القيم لأنة تابع لأفضل موقع في السرب، ثم يليه الوزن β التابع لأفضل موقع محلي للطائر، ثم اقل الأوزان α لأنها تمثل الموقع الحالي. أما المعاملان عدد الطيور وأكبر حالات غير المغطاة لكل صنف فتم اختيارهما بالتجربة وبالاعتماد على دقة التصنيف.

3-11 قياس دقة التصنيف والمقارنة مع خوارزمية Ant-miner

لغرض قياس دقة التصنيف تم استخدام طريقة الطيات العشرة (ten fold cross validation)[4]، وفيها تقسم بيانات التدريب إلى عشرة أقسام بنسبة توزيع متساوية بين الأصناف الثلاثة وتم تنفيذ الخوارزمية بصورة مفصولة لعشرة مرات وفي كل مرة يتم عزل 1/10 من البيانات لغرض الاختبار و9/10 من البيانات لغرض التدريب وفي كل تنفيذ يتم تبديل بيانات الاختبار بجزء آخر، وأيضا تم مقارنتها مع نتائج قياس دقة التصنيف لخوارزمية النمل في التنقيبAnt-miner والمنفذّة في المصدر [2]. والجدول (4) يبين نتائج قياس دقة التصنيف للخوارزميتين وزمن التنفيذ لعشرة مرات.

جدول (4). نتائج دقة التصنيف وزمن التدريب للخوارزميتين Ant-miner و DPSO

أجزاء التدريب والاختبار	خوارزمية النمل في التنقيب		خوارزمية أَمثَلَة عناصر السرب	
	الدقة %	زمن التدريب(ثواني)	الدقة %	زمن التدريب(ثواني)
1	93±4.554	6.640	93±4.554	9.348
2	93±4.279	9.921	93±4.279	9.542
3	93±4.150	13.100	90±5.004	7.142
4	93±4.554	14.300	90±5.004	9.202
5	96±2.984	10.450	93±4.554	7.155
6	90±5.004	10.984	87±5.681	10.531
7	93±4.554	14.158	93±4.554	10.545
8	93±4.150	12.742	90±5.004	10.560
9	84±6.241	19.840	84±6.241	5.468
10	77±6.928	9.318	77±6.928	9.938
المعدل	90.5±5.000	12.1453	89±5.150	8.9431

4-11 النتائج والاستنتاجات

من خلال ملاحظة معدل مرات التنفيذ للخوارزميتين في الجدول (4)، فان الخوارزمية الجديدة (DPSO) تعطي دقة تصنيف 89±5.150 وهي نسبة مقاربة لدقة تصنيف خوارزمية النمل البالغة 90.5±5.000، أما من ناحية معدل زمن التنفيذ، فكان زمن تنفيذ خوارزمية (DPSO) 8.9431 ثانية وهو زمن اقل من زمن تنفيذ خوارزمية النمل البالغ 12.1453 عندما كانت أمثلة التدريب 270 وعدد السمات 130، لذلك تعتبر هذه الخوارزمية منافسة في مجال التنقيب في محتوى الشبكة العنكبوتية.

المصادر

[1] الطويل، هالة، (2009). "المرجع التعليمي في التنقيب عن البيانات"، الطبعة الأولى، شعاع للنشر والعلوم، حلب، سورية.

[2] المشهداني، محمد حامد، (2011). تقصّي التنقيب في محتوى الشبكة العنكبوتية باستخدام خوارزمية أمثَلَة عناصر السرب وخوارزمية أمثَلَة مستعمرة النمل، رسالة ماجستير غير منشورة، كلية علوم الحاسوب والرياضيات، جامعة الموصل.

[3] Abraham A., Grosan C. and Ramos V., (2006). ”Swarm Intelligence in Data Mining”, Springer-Verlag, Berlin Heidelberg.

[4] Bramer M. (2007)."Principles of Data Mining", Springer-Verlag, London.

[5] Correa E. S., Freitas A. A., and Johnson C. G., (2006). A New Discrete Particle Swarm Algorithm Applied to Attribute Selection in a Bioinformatics Data Set, Genetic and Evolutionary Computation Conference - GECCO, pp. 35-42.

[6] Correa E. S., Freitas A. A., and Johnson C. G., (2008). Particle Swarm for Attribute Selection in Bayesian Classification: An Application to Protein Function Prediction, Journal of Artificial Evolution and Applications, Vol.2008, Article ID 876746, 12 pages.

[7] Hedlund N.,(2001). Automatic Construction of Stemming Rules, Master’s Thesis in Computer Science, Royal Institute of Technology, Stockholm, sweden.

[8] Holden N. and Freitas A.A., (2004). Web Page Classification with an Ant Colony Algorithm, Conference: Parallel Problem Solving from Nature - PPSN, pp.1092-1102.

[9] Holden N. P. and Freitas A. A.(2007):A Hybrid Pso/Aco Algorithm for Classification, GECCO Workshop on Particle Swarms, Second Decade, USA, pp. 2745–2750.

[10] Jurafsky D., Martin J. H. (2008). “Speech and Language Processing”, second edition, Prentice Hall.

[11] Liu B., (2007)."Web Data Mining Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag , Berlin Heidelberg, New York.

[12] Martens D, De Backer M, Haesen R, Snoeck M, Vanthienen J and Baesens B (2007). Classifcation with ant colony optimization, IEEE Transaction on Evolutionary Computation Vol. 11, No.5, pp.651–665.

[13] Nagel C. Evjen B. Glynn J. Skinner M. Watson K.,(2008)."Professional C# 2008", Wiley Publishing, Inc.,Indiana.

[14] Qi X., and Davison B.D. (2009). Web Page Classifcation: Features and Algorithms, ACM Computing Surveys - CSUR , vol. 41, no. 2, pp. 1-31.

[15] Rao S.S., (2009). "Engineering Optimization Theory and Practice", Fourth Edition, JOHN WILEY & SONS, INC., Canada.

[16] Scime A., (2005), Web Mining Applications and Techniques, Idea Group Inc, USA.

[17] Sousa T., Silva A., and Neves A. (2004). Particle Swarm based Data Mining Algorithms for classification tasks, Parallel Computing, Vol. 30, pp.767-783.

[18] Xu G., Zhang Z., and Li L. (2011). ”Web Mining and Social Networking”, Web Information Systems Engineering and Internet Technologies, Springer Science+Business Media, Australia.

[19] Yan Y., Kamath G. and Osadciw L. A., (2009). Feature Selection Optimized by Discrete Particle Swarm Optimization for Face Recognition, Proceedings of the SPIE, Volume 7306, pp. 73061W-73061W-11.