نموذج مقترح لعملية التنقيب في البيانات واكتشاف المعرفة في جامعة الطائف

A Model for Data Mining and Knowledge Discovery Process in Taif University

د/ رزق السيد حامد الوزير
مدرس الإحصاء التطبيقي بقسم الإحصاء التطبيقي والتأمين بكلية التجارة جامعة المنصورة
أستاذ مساعد بقسم اقتصاديات وإدارة المشروعات بكلية العلوم الإدارية والمالية جامعة الطائف

ملخص:

التنقيب في البيانات تخصص حديث يُعرّف على أنه العملية الشاملة والمستمرة التي تهدف لاكتشاف المعرفة والأنماط المخفية من مجموعات البيانات الكبيرة وفق خطوات محددة باستخدام الطرق الإحصائية وطرق تعليم الآلة والذكاء الاصطناعي للحصول على نتائج ذات معنى قادرة على دعم اتخاذ القرار. كما يمكن تعريف التنقيب في البيانات باختصار بأنه: عملية ذكاء المنشأة Business Intelligence، أو بأنه عملية اكتشاف المعرفة، أو طرق إيجاد معنى للبيانات الكبيرة بغرض دعم اتخاذ القرار.

ويهدف هذا البحث إلى صياغة عملية محددة للتنقيب في البيانات في جامعة الطائف، وبعبارة أخرى تحديد مكتوب للخطوات التي يجب أن نسير عليها عند التعامل مع حجم كبير من البيانات –ابتداءً من اختيار البيانات التي تناسب هذا النوع من التحليل وانتهاءً بتقييم النموذج ونشره واكتشاف المعرفة، وإعادة الكرة مرة أخرى.

وقد توصل البحث إلى نموذج مقترح لعملية التنقيب في جامعة الطائف يتكون من 6 خطوات هي: فهم الجامعة، وفهم البيانات، وتحضير البيانات للنمذجة، والنمذجة، والتقييم، والنشر. كما عرض البحث عدة أمثلة مقترحة لمهام التنقيب في البيانات في جامعة الطائف.

الكلمات الدالة: التنقيب في البيانات - اكتشاف المعرفة - النمذجة

Summary:

Data mining is a new discipline. It is known that the overall and ongoing process aims to discover the hidden knowledge and patterns from large data sets according to specific steps using statistical methods, machine learning methods and artificial intelligence to get meaningful results that are able to support decision-making. It can also define the data mining briefly as: business Intelligence process, or that the knowledge discovery process, or how to find the meaning of large data to support decision-making.

This research aims to formulate a data mining at the University of Taif, in other words identifying written steps that must be walked upon when dealing with large amounts of data - from selection of data that fit a particular type of analysis to the end of model assessment, deployment and knowledge discovery, and return the ball again.

The search had reached a proposed model for the data mining process in Taif University consists of 6 steps: understanding the university, understand the data, preparing data for modeling, modeling, evaluation, and deployment. It also presented a suggested several examples of data mining tasks in Taif University.

Key words: data mining - knowledge discovery- modeling

1. الإطار النظري

(1-1) مقدمة

يقتضي العمل الروتيني في معظم المنشآت جمع بيانات عن العملاء وحفظها في سجلات إلكترونية في قواعد بيانات عملاقة. وقد يتم تسجيل البيانات بشكل لحظي كما هو الحال في أسواق الأسهم، أو على فترات متفاوتة (كلما أدخل العميل بطاقة الصراف إلى الماكينة أو تم التحويل من/إلى حسابه). ويندرج ذلك أيضاً على الجامعات حيث يتم تسجيل مجموعات كبيرة من البيانات على فترات منتظمة (أول ومنتصف وآخر كل فصل دراسي) فيما يخص الشئون التعليمية، أو بشكل يومي كما هو الحال في معظم إدارات الجامعة.

وقد ساهم تطور الحاسبات وزيادة سعة وسائل التخزين ودقة أجهزة الالتقاط الآلي للبيانات في ظهور كم هائل جداً من البيانات يصعب التعامل معه بالطرق التقليدية. وتحول هم المنشآت من مجرد جمع البيانات والاحتفاظ بها بصورة مفهرسة –ليسهل استدعائها فيما بعد– إلى تحليل تلك البيانات بهدف دعم اتخاذ القرار أو التعرف على العملاء المهمين (المربحين) أو زيادة الإيرادات أو تقليل التكاليف أو حتى أداء الخدمة بشكل مرضي. ومن أجل ذلك كان علم التنقيب في البيانات Data Mining الذي يمكن تعريفه ببساطه على أنه عملية الاختيار والاستكشاف والنمذجة لقواعد البيانات الكبيرة بغرض اكتشاف النماذج والأنماط غير المعروفة مسبقاً.

ويُعد علم التنقيب في البيانات تخصصاً حديثاً، وقد استُخدم هذا المصطلح للمرة الأولى رسمياً بواسطة Usama Fayyad et al في المؤتمر الدولي الأول لاكتشاف المعرفة والتنقيب في البيانات الذي أُقيم في مونتريال في عام 1995 والذي كان ولا يزال أحد أهم المؤتمرات في هذا الموضوع. وقد نُشر في هذا الفرع –في خلال تلك الفترة الوجيزة– كم هائل من البحوث، وخُصص له العديد من الدوريات، كما تسارعت معظم المنشآت العملاقة على تطبيقه واتخاذه منهاجا لها. إن التنقيب في البيانات هو صيحة القرن 21. فإذا كانت المنشآت التجارية العملاقة تتسابق الآن في تطبيقه، فيُتوقع أن يمتد ذلك أيضاً إلى الوزارات والمصالح الحكومية (ومنها الجامعات) خلال السنوات القادمة باعتبار أن البيانات هي الركيزة الأساسية التي تدعم القرار.

ويهدف هذا البحث إلى صياغة عملية التنقيب في البيانات في جامعة الطائف، وبعبارة أخرى تحديد مكتوب للخطوات التي يجب أن نسير عليها عند التعامل مع حجم كبير من البيانات –ابتداءً من

اختيار البيانات التي تناسب نوع بعينه من التحليل وانتهاءً بتقييم النموذج ونشره واكتشاف المعرفة، وإعادة الكرة مرة أخرى.

(2-1) تعريف التنقيب في البيانات

هناك تعريفات كثيرة للتنقيب في البيانات. فقد عرفه Usama Fayyad et al [1996] على أنه:

الأساليب التحليلية المتكاملة التي تنقسم إلى عدة مراحل والتي تهدف لاكتشاف المعرفة غير المعلومة من مجموعات البيانات الكبيرة التي لا تُظهر في سلوكها انتظاماً واضحاً.

ويرى Jiudici [2003] أن تعريف Fayyad et al أهمل جانب مهم –هو الهدف النهائي من التنقيب في البيانات– وهو إمداد مالك البيانات بالنتائج التي تم التوصل إليها، واقترح التعريف التالي:

التنقيب في البيانات هو عملية اختيار، واستكشاف، ونمذجة كميات كبيرة من البيانات لاكتشاف العمليات المنتظمة والعلاقات التي لم تكن معروفة من قبل بهدف الحصول على نتائج واضحة ومفيدة لمالك قاعدة البيانات.

كما عرف Newton's Telecom Dictionary التنقيب في البيانات بأنه:

القدرة على البحث المحكم في البيانات باستخدام الخوارزميات الإحصائية لاكتشاف الأنماط والعلاقات في البيانات.

ولفهم مصطلح التنقيب في البيانات أكثر، يمكن الرجوع للمعنى القاموسي للفعل "ينقب mine" الذي يعني: يستخرج؛ أي عمليات استخراج المعادن المخفية من الموارد الهائلة الموجودة في باطن الأرض. وعندما تقترن كلمة "البيانات Data" مع كلمة "التنقيب Mining"، فإن هذا يعني: البحث المتعمق لإيجاد المعلومات الإضافية التي لم تكن معلومة مسبقاً من الكم الكبير للبيانات المتاحة.

وفي ضوء التعريفات السابقة يمكن تعريف التنقيب في البيانات بأنه: *العملية الشاملة والمستمرة التي تهدف لاكتشاف المعرفة والأنماط المخفية من مجموعات البيانات الكبيرة وفق خطوات محددة باستخدام الطرق الإحصائية وطرق تعليم الآلة والذكاء الاصطناعي للحصول على نتائج ذات معنى قادرة على دعم اتخاذ القرار*. كما يمكن تعريف التنقيب في البيانات باختصار بأنه: *عملية ذكاء المنشأة Business Intelligence*، أو بأنه *عملية اكتشاف المعرفة*، أو *طرق إيجاد معنى للبيانات الكبيرة* بغرض دعم اتخاذ القرار.

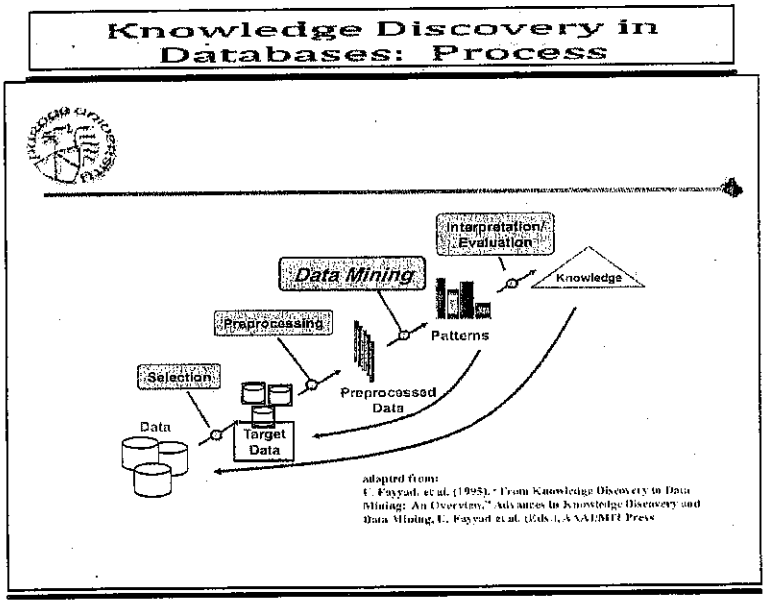
(3-1) أدبيات البحث

يمكن تصنيف دراسات التنقيب في البيانات إلى ثلاثة أقسام: الدراسات التي اهتمت بعملية التنقيب في البيانات، والدراسات البحتة التي تبحث في أساليب التنقيب في البيانات سواء من الناحية الإحصائية أو الحاسبية، والدراسات التطبيقية التي تهتم بتطبيق عملية وأساليب التنقيب في البيانات على بيانات فعلية (أو حالات عمليه).

أ) دراسات عملية التنقيب في البيانات:

اهتمت العديد من الدراسات الأولى للتنقيب في البيانات بوضع عمليه تحدد خطوات التنقيب في البيانات، وقد تم ذلك بشكل مكثف في الفترة من 1996-2003. ثم تحول اهتمام البحوث في العشر سنوات الأخيرة من تطوير العمليات الشهيرة الموجودة إلى إيجاد نماذج جديدة. ومن أشهر هذه الدراسات:

دراسات Fayyad et al [1996a-e]: هي أول وأشهر الدراسات التي تعاملت مع عملية التنقيب في البيانات واكتشاف المعرفة. وقد لخصت مفهومها لتلك العملية –كما هو مبين بشكل (1) في 9 خطوات متتالية ودائرية تبدأ بتحديد مصادر البيانات وتنتهي باكتشاف المعرفة.

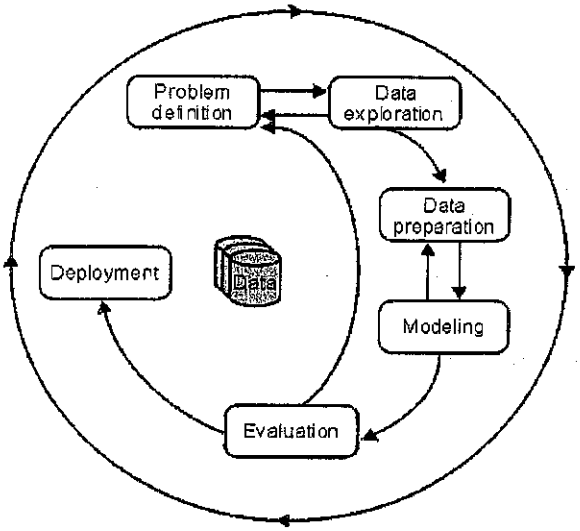


شكل (1): عملية Fayyad للتنقيب في البيانات واكتشاف المعرفة

دراسة SEMMA [1997]: هي عملية مكونة من 5 خطوات أنتجتها شركة SAS المتخصصة في حزم البرامج الإحصائية بعد خبرتها الطويلة في عمل التحاليل الإحصائية للعديد من الشركات الكبرى. وقد

أُطلق على تلك العملية اسم SEMMA وهو اختصار الأحرف الأولى للأفعال التي تتم في الخطوات الخمس، وهي: اسحب عينة Sample، استكشف Explore، عدل Modify، ضع نموذج Model، وقيّم الوضع Assess. وقد استند معهد SAS على هذه الخطوات الخمس كأساس لحزمة برامجها الجديدة للتنقيب في البيانات التي أطلقت عليها اسم "مُنقب المشروع SAS Enterprise Miner.

دراسة IBM: هي عبارة عن عملية دائرية مكونة من 6 خطوات متتالية تبدأ بتعريف المشكلة وتنتهي بالنشر مع مراعاة العودة من الخطوة الثانية (استكشاف البيانات) للأولى، ومن الرابعة (النمذجة) للثالثة (تحضير البيانات)، ومن الخامسة (التقييم) للأولى. ويوضح شكل (2) تسلسل هذه الخطوات.



Source: publib.boulder.ibm.com/.../c_dm_process.html

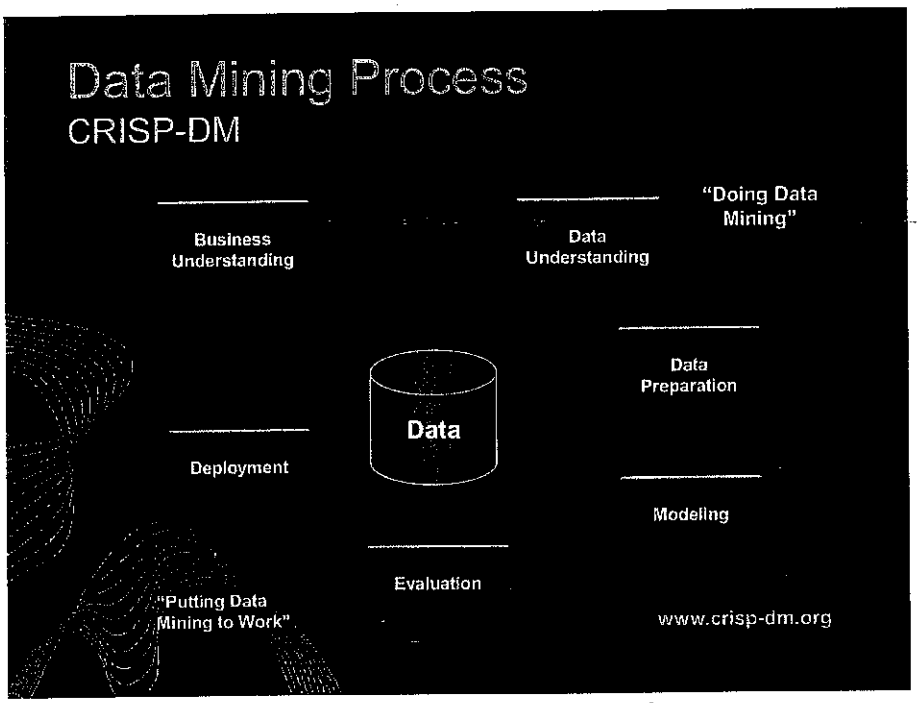
شكل (2): عملية IBM للتنقيب في البيانات واكتشاف المعرفة

دراسة Statistica: هي عبارة عن عملية مكونة من 3 خطوات هي: الاستكشاف exploration، وبناء النماذج والتحقق من مصداقيتها model building and validation، والنشر deployment باختيار أفضل نموذج وتطبيقه على بيانات جديدة لتوليد التنبؤات.

دراسة SPSS: هي عبارة عن عملية مكونة من 5 خطوات أيضاً، وهي معروفة باسم 5A's وهو اختصار الأحرف الأولى للأفعال التي تتم في الخطوات الخمس، وهي: قيّم الوضع Assess، أحضر البيانات Access، حلل البيانات Analyze، تفاعل مع الخطوات السابقة Act، ودع العمل يسير بتلقائية Automate. وقد أصدرت الشركة لهذا الغرض برنامجاً –تم إلحاقه بـ . SPSS– أطلقت علية اسم Clementine.

دراسة CRISP-DM [2003]: هي مجهود تعاون 4 شركات أوربية: SPSS (a provider of commercial DM solutions), NCR (a database provider), Daimler Chrysler, and OHRA (an insurance company) من أجل توحيد خطوات التنقيب في البيانات في الشركات الصناعية. ويعبر مصطلح CRISP-DM عن الأحرف الأولى للكلمات اللاتينية (CRoss-Industry Standard Process for DM).

وقد صدر الإصدار الأول من هذه العملية في عام 2000 ولاقى قبولا واسعاً في الشركات الصناعية الأوربية (حيث قامت باعتماده حتى الآن أكثر من 300 شركة). ويوضح شكل (3) أن عملية CRISP-DM هي عبارة عن نموذج دائري مبني على البيانات ويحتوي على 6 خطوات تبدأ بفهم البيانات في ضوء احتياجات المنشأة، ثم تحضير البيانات، ثم النمذجة، ثم التقييم، ثم النشر، وتنتهي (مؤقتاً) من حيث بدأت. ويُلاحظ على العملية إمكانية العودة من الخطوة الثانية للأولى، ومن الخطوة الرابعة للثالثة، ومن الخطوة الخامسة للأولى.



شكل (3): عملية CRISP-DM للتنقيب في البيانات واكتشاف المعرفة

دراسة Tsui [2009]: هي عبارة عن نموذج مكون من 4 خطوات متتالية يبدأ بتحديد أهداف المنشأة وينتهي بالتعزيز والتطبيق مع مراعاة العودة من الخطوة الرابعة للخطوات الثلاث السابقة. ويوضح شكل (4) تسلسل هذه الخطوات.

Determine
Business Objectives

Data Preparation

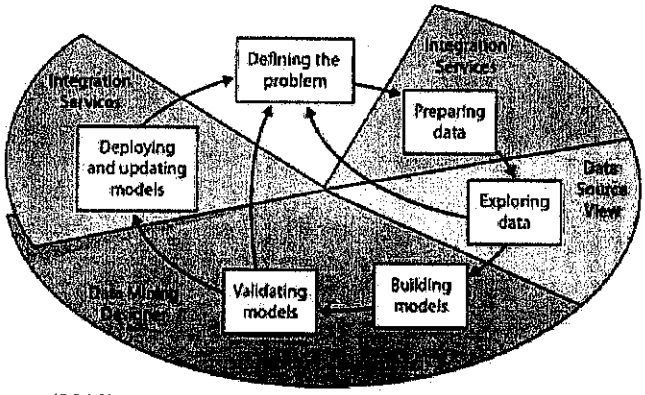
Mining & Modeling

Consolidation & Application

Source: Tsui (2009)

شكل (4): عملية Tsui للتنقيب في البيانات واكتشاف المعرفة

دراسة Microsoft [2010]: تعاملت مع عملية التنقيب في البيانات واكتشاف المعرفة على أنها عملية دائرية مكونة من 6 خطوات تبدأ بتعريف المشكلة وتنتهي أيضاً عندها. وقد لخصت الدراسة مفهومها لعملية التنقيب في البيانات في شكل (5) التالي:



Source: Microsoft SQL Server (2010)

شكل (5): عملية مايكروسوفت للتنقيب في البيانات واكتشاف المعرفة

(ب) دراسات أساليب التنقيب في البيانات:

ويندرج تحت هذا النوع من الدراسات البحوث التي تعاملت مع مصطلحات وأساليب وخوارزميات التنقيب في البيانات. ومنها:

دراسة Giudici [2003]: عرضت الطرق الإحصائية والحاسبية للتنقيب في البيانات، حيث تناولت بالتفصيل أساليب الاستكشاف ومقاييس المسافة والتحليل العنقودي والانحدار الخطي والانحدار اللوجستي وشجرة القرارات وتحليل التمايز والشبكات العصبية ونماذج أقرب جار والنماذج الخطية المعممة.

دراسة Tan, Steinbach and Kumar [2006]: تعرضت لطرق ومفاهيم التصنيف Classification وتحليل الاقتران Association Analysis والتحليل العنقودي Cluster Analysis.

دراسة Thearling [2009]: تعرضت للتطور التاريخي للتنقيب في البيانات في ضوء احتياجات المنشأة. كما تناولت بالتعريف بعض مفاهيم التنقيب في البيانات مثل الشبكات العصبية وشجرة التصنيفات وتنقيح البيانات والتحليل التشغيلي على الهواء OLAP والتحليل الاستكشافي والنماذج الخطية وغير الخطية.

(ج) دراسات تطبيقات التنقيب في البيانات:

امتدت تطبيقات التنقيب في البيانات من التسويق إلى المنشآت الصناعية والخدمية والبنوك وأسواق المال والرعاية الصحية. ومن بعض هذه التطبيقات:

دراسة Lucas [2006]: قدم دراسة بعنوان "التحليل البيزي، وتحليل الأنماط، والتنقيب في البيانات في مجال العناية بالصحة". وقد كان غرضه استخدام الأساليب المذكورة لحل مشاكل الطب الحيوي والعناية بالصحة biomedical and health-care problems. وقد توصلت الدراسة إلى قاعدة قرار (باستخدام شجرة التصنيفات classification trees) لبقاء المريض على قيد الحياة استناداً إلى ضغط دمه الانقباضي Systolic Pressure وعمره.

دراسة Rygielski et al [2002]: قدم دراسة بعنوان "أساليب التنقيب في البيانات لإدارة العلاقة بالمستهلك بغرض استخلاص المعلومات الخفية من قواعد البيانات الكبيرة بتحديد العملاء المهمين والتنبؤ بسلوكهم في المستقبل. وقد استخدمت الدراسة للوصول لذلك أسلوبي الشبكات العصبية و CHAID.

دراسة Giudici [2003]: قدم العديد من المشاكل العملية التي تواجه منشآت الأعمال، وقام بحلها باستخدام أساليب التنقيب في البيانات. ومن هذه الحالات: تحليل سلة السوق، وإدارة العلاقات بالمستهلك، ومشكلة الجدارة الائتمانية.

(1-4) مشكلة البحث وأهميته

يتكدس لدى المؤسسات –ومنها الجامعات– نتيجة العمل اليومي أحجام كبيرة جداً من البيانات. إذ يتراوح حجم البيانات بين ($10^{12} - 10^{8}$) بايت بالنسبة للسجلات، وبين ($10^{4} - 10^{2}$) بايت بالنسبة للحقول أو المتغيرات المدروسة Bajcsy [2010]. فكيف يمكن اكتشاف الأنماط المفيدة من هذا الكم من البيانات باستخدام مداخل التحليل التقليدية؟ وبدون تحديد عملية واضحة للتنقيب في البيانات: تتبعثر البيانات داخل الجامعة، ويتم الحصول على بيانات مختلفة من أكثر من مصدر عند إجراء الاستعلام الواحد، كما يخضع منها للتحليل نسبة ضئيلة جداً لا تتجاوز 5%. وتستمر الجامعة في تجميع البيانات (لأن عملها الروتيني يوجب ذلك)، وتتزاكم البيانات دون خضوع معظمها للتحليل ودون الوقوف على المؤشرات الهامة ودون الاستفادة منها للتخطيط للمستقبل.

فمن المعروف أن الغرض من أي بيانات هو تحديد المتغيرات (أو المؤشرات) الهامة ثم التنبؤ بالاتجاه العام لها، ولكن حتى الآن فإن هذه الإمكانية لم يتم استغلالها بالكامل في منشآت الأعمال عموماً وفي الجامعات خصوصاً. ويرجع ذلك إلى وجود المشاكل التالية عند التعامل مع البيانات:

1- تشتت (أو تبعثر) البيانات داخل أنظمة أرشيف مختلفة غير متصلة ببعضها البعض، وهو ما يؤدي إلى عدم كفاءة تنظيم البيانات.
2- عدم الإلمام الكافي بالأدوات الإحصائية ومنهج التنقيب في البيانات وطرق توظيفها لاستخلاص المعلومات، وإهمال الدور الذي يمكن أن يلعبه أساتذة الجامعة المتخصصين في الإحصاء التطبيقي ونظم المعلومات في دعم اتخاذ القرار.
3- نقص البيانات. بمعنى عدم وجود الحقول من أصله أو وجودها مع بعض القيم المفقودة.
4- صعوبة الوصول للبيانات الموجودة في المنظومة بحجة السرية.

ويمكن تلخيص الجوانب الأربعة السابقة للمشكلة في عبارة واحدة هي عدم وجود عملية (خطوات مقننة) للتنقيب في البيانات واكتشاف المعرفة تدعم اتخاذ القرار في جامعة الطائف.

وتعود أهمية هذا البحث إلى أن إيجاد عملية للتنقيب في البيانات في جامعة الطائف سوف يُدخل الجامعة إلى عالم جديد هو عالم التنقيب في البيانات واكتشاف المعرفة. وسوف يُسجل أن جامعة الطائف أول جامعة عربية تقوم بانتهاج التنقيب في البيانات واكتشاف المعرفة كأسلوب عمل لها، وهو ما سوف يؤهلها للتميز العلمي في هذا التخصص في المنطقة واجتياز اختبارات الجودة والاعتماد الأكاديمي بشكل كبير.

(1-5) تصميم البحث

يهدف البحث إلى تصميم عمليه أو خطوات محددة أو نموذج للتنقيب في البيانات واكتشاف المعرفة في جامعة الطائف. وقد اعتمد أسلوب البحث –عند صياغة تلك العملية– على المقارنة المتوازية بين العمليات الشهيرة التي أفرزتها الدراسات السابقة. ولأن هذه العمليات قد تم تصميمها أساساً في مجال الصناعات ومنشآت الأعمال، فقد تم تصميم عملية جديدة تتوافق مع ظروف واحتياجات جامعة الطائف. وتأتي خطة هذا البحث في خمسة أجزاء: خُصص أولها للإطار النظري، وثانيها للمقارنة بين عمليات التنقيب في البيانات، وثالثها للنموذج المقترح للتنقيب في البيانات، ورابعها لدراسة مدى تأهل جامعة الطائف لتطبيق أدوات وأساليب التنقيب في البيانات واكتشاف المعرفة، وخامسها لعرض الخلاصة. وقد استمد البحث بياناته من الدراسات السابقة، ومن المنظومة الجامعية لجامعة الطائف.

(1-6) الاستفادة من البحث

يقدم هذا البحث خطوات محددة في التعامل مع البيانات ابتداء من جمعها حتى اكتشاف المعرفة منها بغية إيجاد عملية للتنقيب في البيانات في جامعة الطائف يُدخل الجامعة إلى عالم جديد هو عالم التنقيب في البيانات واكتشاف المعرفة. ولا شك أن ذلك سيُسجل للجامعة على أنها أول جامعة عربية تقوم بانتهاج التنقيب في البيانات واكتشاف المعرفة كأسلوب عمل لها، وهو ما يؤهلها للتميز العلمي في هذا التخصص في المنطقة واجتياز اختبارات الجودة والاعتماد الأكاديمي بشكل كبير.

كما يمكن الاستفادة من البحث في تصميم مقرر عام للتنقيب في البيانات يتم تدريسه في المستوى الخامس في الكليات التي تقرر إحصاء 1 وإحصاء 2، وعدة مقررات متخصصة تُدرَس في المستويات الأعلى مثل الشبكات العصبية وشجرة القرارات وتحليل الاقتران. ويمكن أيضاً الاستفادة من البحث في تقديم الاستشارات للشركات التي تريد أن تنتهج هذا المدخل الجديد في اكتشاف المعرفة.

2. المقارنة بين عمليات التنقيب في البيانات

تناول هذا الجزء المفاهيم الأساسية للتنقيب في البيانات. ثم انتقل لتسليط الضوء على الفرق بين التنقيب في البيانات واكتشاف المعرفة وذكاء المنشأة. وانتهى بعمل مقارنة بين العمليات المختلفة للتنقيب في البيانات، تمهيداً لتقديم النموذج المقترح في الجزء الثالث.

(2-1) المفاهيم الأساسية للتنقيب في البيانات

عملية التنقيب في البيانات: هي دستور المنشأة في التعامل مع البيانات. أي المراحل المختلفة (المكتوبة والمفهومة) التي تمر عليها البيانات، وعلاقة تلك المراحل ببعضها (التغذية الأمامية والخلفية) ابتداءً من فهم المنشأة حتى اكتشاف المعرفة.

التشغيل التحليلي على الهواء OLAP: لا يجب الخلط بين التنقيب في البيانات والطرق المستخدمة في إنشاء أدوات إعداد التقارير متعددة الأبعاد؛ أي التشغيل التحليلي على الهواء On Line Analytical Processing (OLAP). إذ أن OLAP عادةً ما يكون أداة رسومية تُستخدم لتسليط الضوء على العلاقة بين كل زوج من المتغيرات المتاحة في رسم بياني ذو بعدين. أما في حالة التنقيب في البيانات، فيتم دراسة كافة المتغيرات المتاحة ببناء نماذج مفيدة واستخدامها في التنبؤ. إن التنقيب في البيانات لا يقتصر على تحليل البيانات، بل أنه عملية أكثر تعقيداً بكثير يكون فيها تحليل البيانات أحد مراحلها. وإذا كان الاستعلام وإعداد التقارير أداتين لوصف محتويات قاعدة البيانات، فإن OLAP يُستخدم لشرح سبب وجود العلاقات بين المتغيرات. ففيه يضع الباحث فروضه عن العلاقات الممكنة بين المتغيرات ويتطلع لتأكيد رأيه بمشاهدة البيانات، على عكس الحال في التنقيب في البيانات حيث تكشف البيانات عن الفروض ولا تُصنع مقدماً.

التنقيب في البيانات: مداخل التحليل: قدم Linoff and Berry (2011) مدخلين تحليليين للتنقيب في البيانات. الأول هو التحليل من أعلى إلى أسفل (التحليل التأكيدي confirmative analysis)، والثاني هو التحليل من أسفل إلى أعلى (التحليل الاستكشافي explorative analysis). ويهدف التحليل التأكيدي إلى تأكيد أو رفض فروض البحث المصنوعة مقدماً ويحاول توسيع معرفتنا بالظواهر المفهومة جزئياً من خلال الأساليب الإحصائية التقليدية. بينما يبدأ التحليل الاستكشافي بالبحث في البيانات بغية اكتشاف المعلومات المفيدة التي لم تُلاحظ من قبل وينتهي هذا البحث بصياغة الفروض، وهو التحليل النموذجي للتنقيب في البيانات. وفي الواقع، فإن كلا المدخلين يكملان بعضهما البعض. إذ أن المعلومات الناجمة عن عن التحليل الاستكشافي –التي تحدد العلاقات الهامة والاتجاهات العامة– لا يمكن أن تجيب على السؤالين التاليين: 1) لماذا تُعتبر هذه الاكتشافات مفيدة؟ 2) وما هو مدى صدقها؟ وهما سؤالان يستطيع التحليل التأكيدي الإجابة عليهما ببساطة، حيث تمكّن أدواته من التأكد من هذه الاكتشافات وتقييم جودة القرارات.

مفاتيح نجاح التنقيب في البيانات: إن المنشأة التي تريد أن تنجح في اتخاذ التنقيب في البيانات كمنهاج عمل لها، يجب أن ترتكز على دعامتين. الأولى هي الصياغة الدقيقة للمشكلة التي تحاول حلها في إطار

ما يسمى بفهم المنشأة. والثانية هي استخدام البيانات الصحيحة (والنموذج الصحيح) من بين البيانات والنماذج المتاحة لها. وبتعبير أدق، يجب أن تلتزم المنشأة بعملية محددة للتنقيب في البيانات، وهو ما يهدف إليه هذا البحث.

الأنماط: إذا كان الهدف الأساسي للتنقيب في البيانات هو اكتشاف الأنماط التي لم تكن معلومة من قبل، فإن النمط pattern or model هو عينة موثوق فيها من الصفات المشتركة أو الأفعال أو الاتجاهات أو أي خصائص أخرى مشاهده عن شخص ما أو مجموعة أو مؤسسة. أو هو العلاقة أو الملخص الناتج عن تطبيق التنقيب في البيانات. وتُستخدم الأنماط القابلة للفهم في: تفسير البيانات الموجودة، وتلخيص قاعدة البيانات الكبيرة لدعم صنع القرار، واستكشاف أنماط أعمق، والتنبؤ بالبيانات الجديدة أو تصنيفها.

تعليم الآلة Machine Learning: هو استنساخ عملية توليد البيانات، بما يسمح للمحللين بالتعميم من البيانات المشاهدة إلى الحالات الجديدة غير المشاهدة. وتتفق علوم الإحصاء التطبيقي وتعليم الآلة والحاسب والذكاء الاصطناعي (وهم أدوات التنقيب في البيانات) على هدف واحد هو البحث في البيانات لإيجاد العلاقات وتحديد العمليات التي تتكرر بصفة دورية للوصول إلى حقائق عامة. وقد قدم Rosenblatt [1962] أول نموذج لتعليم الآلة، أطلق عليه اسم "الفاهم The perceptron". وتلى ذلك الشبكات العصبية neural network في النصف الثاني من الثمانينات. كما اهتم بعض الباحثين –خلال نفس الفترة– بتطوير نظرية شجرة القرارات decision trees التي تُستخدم للتعامل مع مشاكل التصنيف.

التصور Visualization: هو العرض البياني للعلاقة بين المتغيرات وفحصه بغرض الحصول على استنتاج مبدئي يُستخدم فيما بعد في مرحلة التحليل، ويستخدم بعض الباحثين مصطلح آخر بديل هو الاستكشاف exploration الذي يعني استخدام أدوات الإحصاء الوصفي (جداول، ورسوم، ومقاييس) لتحقيق نفس الغرض. ويسمح التصور للباحث بأن يرى غابة كبيرة مع أخذ نظرة مكبرة لجزء معين فيها.

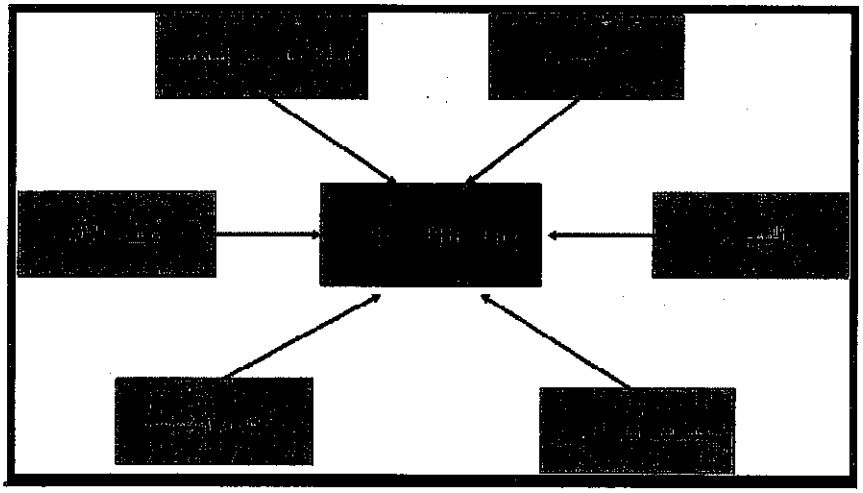
التصنيف classification: هو تقسيم أو ترتيب البيانات في مجموعات مُعرّفة مسبقاً. ومن أشهر خوارزمياته: الشبكات العصبية، ومصنف بايز، والجار الأقرب.

العنقدة clustering: هي تقسيم البيانات إلى مجموعات تختلف بشدة من الخارج لكنها متجانسة من الداخل. وعلى عكس التصنيف، فلا يعرف الباحث مقدماً ما هي العناقيد التي سيبدأ بها ولا الخصائص التي ستُجمع على أساسها العناقيد. وبعد إيجاد العناقيد، يمكن استخدامها في تصنيف بيانات جديدة. وتعد Kohnen maps و k-means من أشهر الخوارزميات المستخدمة في إجراء عملية العنقدة.

تحليل الروابط link analysis: هو مدخل وصفي لاستكشاف البيانات يساعد في تحديد العلاقات بين القيم في قاعدة البيانات. ومن أشهر أساليبه: اكتشاف الاقتران association discovery، واكتشاف التسلسل sequence discovery. إذ يختص تحليل الاقتران بتحديد العناصر التي يمكن أن تقع معا في حدث ما مثل عملية شراء معينة (الذي يشتري مطرقة، يشتري مسامير). ويُعد تحليل سلة السوق أحد الأمثلة الشهيرة لاكتشاف الاقتران. بينما ينظر اكتشاف التسلسل إلى السلسلة على أنها اقتران متعلق بالزمن.

(2-2) علاقة التنقيب في البيانات بالعلوم الأخرى

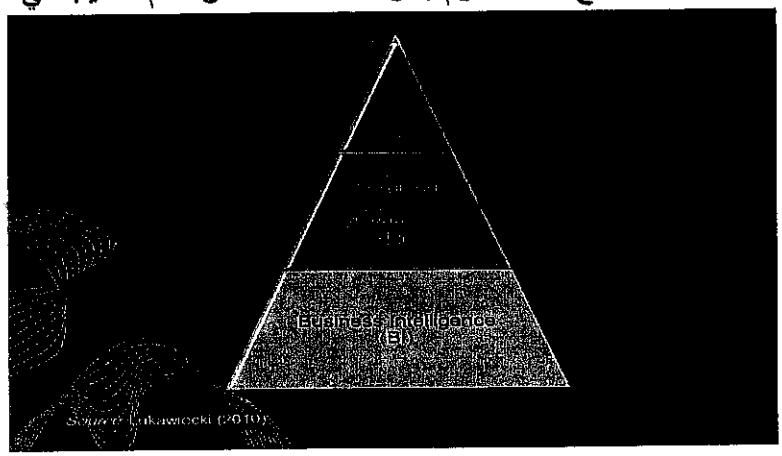
استفادت تقنية التنقيب في البيانات من التقدم الكبير الذي حدث في علمي الإحصاء والذكاء الاصطناعي AI، حيث عمل كلاهما على مشاكل التعرف على الأنماط والتصنيف وساهما في فهم وتطبيق الشبكات العصبية وأشجار القرارات. ويبين شكل (6) أن التنقيب في البيانات يُعد توليفة أو ملتقى لعلوم الإحصاء التطبيقي Applied Statistics (وهو جسم التنقيب في البيانات)، وتعليم الآلة Machine Learning، وأنظمة قواعد البيانات data base systems، والتبصر Visualization، والخوارزميات Algorithms، بالإضافة لتخصصات أخرى (مثل التَعرُّف على الأنماط Pattern Recognition، وأساليب الأمثلية Optimization Techniques).



شكل (6): التخصصات المكونة لعلم التنقيب في البيانات

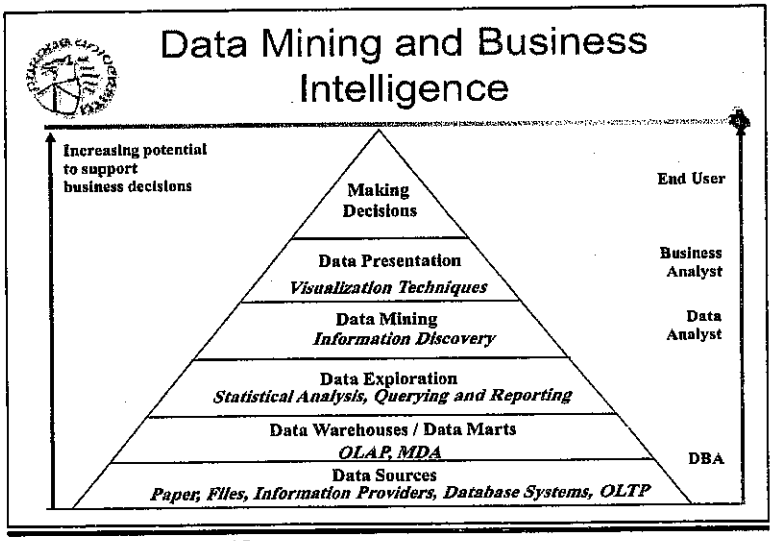
(2-3) التنقيب في البيانات واكتشاف المعرفة وذكاء المنشأة

يرى Lukawiecki (2010) أن ذكاء المنشأة هو مجموعة من التطبيقات والتكنولوجيا تُستخدم: لتجميع، وتخزين، وتحليل، والمشاركة في البيانات والوصول إليها بهدف اتخاذ قرارات أفضل. ويبين شكل (7) أن عملية ذكاء المنشأة تقع أسفل الهرم يعلوها اكتشاف المعرفة ثم التنقيب في البيانات.



شكل (7): عملية ذكاء المنشأة واكتشاف المعرفة والتنقيب في البيانات

ويرى Bajcsy [2010] أن تفصيل تلك العملية يأتي في 6 خطوات تبدأ بتحديد مصادر البيانات ويتخللها التنقيب في البيانات وتنتهي بصنع القرارات (شكل 8). كما يرى Fayyad [1996] أن التنقيب في البيانات جزء من عملية أكبر هي اكتشاف المعرفة.



Source: Bajcsy [2010]

شكل (8): عملية ذكاء المنشأة

وعلى الرغم من اختلاف تعريف المصطلحات الثلاث، إلا أن معظم بحوث التنقيب في البيانات التي جاءت فيما بعد تستخدم المصطلحات الثلاثة بالتبادل للإشارة لذات المعنى وإن اختلفت في تحديد عدد ومحتوى هذه الخطوات. فبمقارنة الشكل (1) مع الشكل (6)، يتبين أن الأول"عملية اكتشاف المعرفة" يحتوي على نفس مصطلحات وخطوات الشكل الثاني "عملية ذكاء المنشأة". وعلى الرغم من أن التنقيب في البيانات يأتي في المرحلة الوسطى في كلا الشكلين، إلا أنه استُخدم فيما بعد كبديل للمصطلحين.

وقد ذكر Kurgan & Musilek (2006) أن المصطلح المناسب والدقيق الذي يعبر عن العملية الشاملة لاستخلاص أو اكتشاف المعرفة هو مصطلح اكتشاف المعرفة والتنقيب في البيانات Knowledge Discovery and Data Mining (KDDM). ويشمل ذلك: تخزين البيانات والوصول إليها واختيارها وترميزها وتنقيحها وتحويلها، وتقديم خوارزميات كفؤة وقابلة للقياس لتحليل مجموعات البيانات الكبيرة، وبناء النماذج وتقييمها ونشرها، وتفسير وعرض النتائج.

(2-4) المقارنة بين عمليات التنقيب في البيانات

بفحص دراسات عمليات التنقيب في البيانات واكتشاف المعرفة (الثمانية المذكورة في الجزء الأول)، اتضح أن كل عملية لابد أن تمر بثلاث مراحل على الأقل من إجمالي سبع. وتشير السبع مراحل إلى جميع المراحل الممكنة التي يمكن تمر بها عملية التنقيب في البيانات. وهذه المراحل هي: تعريف فهم المنشأة ، سحب عينة، الاستكشاف، تحضير البيانات للنمذجة، النمذجة، التقييم، والنشر. ويوضح جدول (1) ملخصاً بالمقارنة بين العمليات الثمانية بحسب المراحل السبع للوقوف على أوجه القصور والشمول لكل عملية.

جدول (1): ملخص بالمقارنة بين عمليات التنقيب في البيانات

عملية DM	فهم المنشأة	سحب العينة	الاستكشاف	تحضير البيانات	النمذجة	التقييم	النشر
Fayyad et al	×	✓	×	✓	✓	✓	✓
SEMMA	×	✓	✓	✓	✓	✓	×
IBM	✓	×	✓	✓	✓	✓	✓
Statistica	×	×	✓	×	✓		✓
SPSS	تقييم الوضع			✓	✓	✓	
CRISP	✓	فهم البيانات		✓	✓	✓	✓
Tsui	✓	×	×	✓	✓	✓	
Microsoft	✓	×	✓	✓	✓	✓	✓

ويتضح من الجدول السابق أن عملية Fayyad et al قد أهملت مرحلتين مهمتين من عملية التنقيب في البيانات؛ هما: فهم المنشأة والتحليل الاستكشافي للبيانات. كما أهملت عملية SEMMA مرحلتي فهم المنشأة ونشر النموذج. ولم تتعرض عملية IBM لمرحلة سحب العينة. أما عملية Statistica فقد أهملت مراحل فهم المنشأة وسحب العينة وتحضير البيانات، ودمجت النمذجة والتقييم في مرحلة واحدة. كما استبدلت عملية SPSS المراحل الثلاثة الأول في الجدول السابق بمرحلة واحدة تحت مسمى تقييم الوضع، ودمجت مرحلتي التقييم والنشر. كما وضعت عملية CRISP مصطلح "فهم البيانات" كبديل لمرحلتي سحب العينة والاستكشاف. وأهملت عملية Tsui مرحلتي سحب العينة والاستكشاف، ودمجت أيضاً مرحلتي التقييم والنشر. في حين أهملت عملية Microsoft مرحلة سحب العينة فقط.

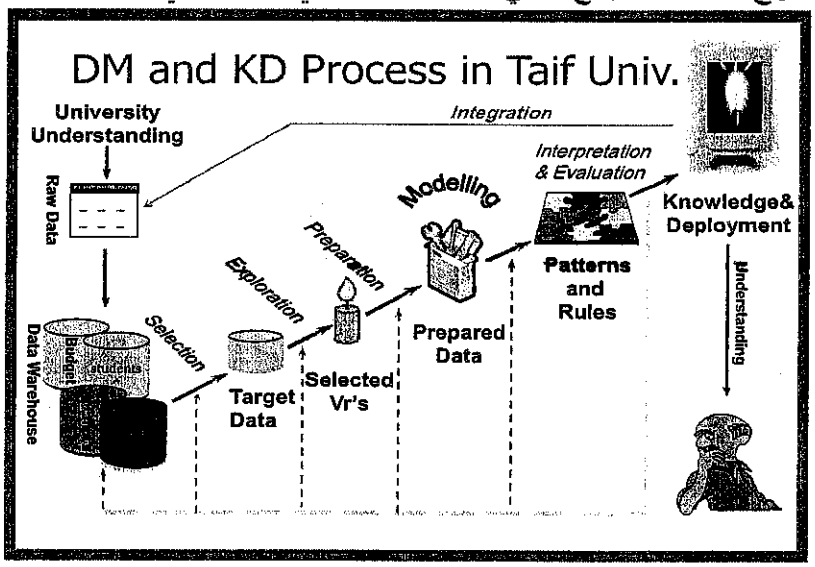
وهكذا، فلم تجمع أي من العمليات الثمانية على المراحل السبع المذكورة في جدول (1). إذ سقطت مرحلة أو أكثر من حسابات كل عملية، وهو ما أظهر الحاجة إلى عملية جديدة تأخذ في حسبانها المراحل الضرورية التي ظهرت في كافة العمليات، وهو الموضوع التالي.

3. النموذج المقترح لعملية التنقيب في البيانات

نبين هنا النموذج المقترح لعملية التنقيب في البيانات في جامعة الطائف، والخطوات التفصيلية لتلك العملية.

(3-1) النموذج المقترح لعملية التنقيب في البيانات

بعد الفحص الدقيق للعمليات الشهيرة للتنقيب في البيانات (في الجزء الأول) والمقارنة بينها في (الجزء الثاني)، اقترح الباحث النموذج التالي كعملية للتنقيب في البيانات في جامعة الطائف.



شكل (9): النموذج المقترح لعملية التنقيب في البيانات في جامعة الطائف

ويبين النموذج أن العملية المقترحة للتنقيب في البيانات هي عملية دائرية ومرتدة ومستمرة تبدأ بفهم الجامعة فهما دقيقاً وتنتهي (لكنها لا تنقطع) باكتشاف المعرفة. إذ يؤدي فهم الجامعة إلى تكوين أفكار عن المشاكل التي تواجهها، ومن ثم جمع البيانات الخام التي قد تساهم في حل تلك المشاكل. وتُخزن البيانات الخام في مخزن البيانات (المنظومة الجامعية)، وتُصنف إلى بيانات تخص شئون الطلاب وأخرى تخص أعضاء هيئة التدريس وثالثة تخص الموظفين ورابعة تخص كافة الأمور المالية في الجامعة. ويتم اختيار البيانات التي تخص المشكلة محل الدراسة من قاعدة البيانات الفرعية، ثم تُنقح للوصول إلى ما يسمى ببيانات الهدف. ويتم عمل التحليل الاستكشافي على بيانات الهدف بغية تحديد المتغيرات الهامة التي ستدخل إلى التحليل متعدد المتغيرات في مرحلة النمذجة بعد تحضيرها لذلك. وبعد توفيق النموذج، يتم تقييمه وتفسيره للوصول إلى الأنماط والقواعد الحاكمة للمشكلة. ويتم تطبيق النموذج الموفق على مجموعة بيانات جديدة بغرض التنبؤ بغية اكتشاف المعرفة.

ويؤكد النموذج المقترح على نقطتين هامتين: الأولى هي أن العملية قد تدور عدة دورات للوصول إلى حل مرضي للمشكلة قيد البحث، وهو ما يبينه السهم الهابط قبل الأخير على الرسم الذي يشير إلى إمكانية ارتداد العملية قبل مرحلة اكتشاف المعرفة إلى أيٍ من المراحل الخمس السابقة عليها. والثانية هي

أن عملية اكتشاف المعرفة ذاتها عملية غير منتهية، فهي قابلة لإعادة التشغيل من مرحلة البيانات الخام بواسطة السهم الأفقي المنكسر. ويبين القسم التالي مزيداً من التفصيل للنموذج المقترح.

(3-2) الخطوات التفصيلية لعملية التنقيب في البيانات

على الرغم من علانية أهداف الجامعة بشكل واضح ومحدد، إلا أن تحديد المشاكل الأساسية التي تواجه أنشطة الجامعة المختلفة وترجمتها إلى أهداف تفصيلية قابلة للتحليل ليس معلناً بدرجة كافية. ولاشك أن ذلك التحديد الواضح للمشاكل بطريقة قابلة للقياس هو ما تنطلق منه عملية التنقيب في البيانات، وهو أيضاً من أصعب وأهم مراحلها. إذ أن كل المراحل التالية هي مجرد إجراءات تنفيذية تستند على التصور الذي تم في هذه المرحلة؛ فإذا كان هذا التصور خاطئاً، فإن عملية التنقيب في البيانات سوف تُهدم برمتها مهما كانت درجة التدقيق المصاحبة لكل مرحلة.

وتستمد الجامعة بياناتها من المنظومة الجامعية (مستودع البيانات data warehouse)؛ وهو مخزن إلكتروني للبيانات التاريخية التي لم تعد تخضع لأي تغييرات. ويتم تحضير البيانات للنمذجة ببناء مصفوفة البيانات الخاصة بمشكلة معينة في ضوء الاحتياجات والأهداف المحددة في المرحلة السابقة. وبعد عمل التحليل الاستكشافي يتم تنقيح مصفوفة البيانات بحذف المتغيرات التي لن تخدم عملية التحليل وتدقيق النظر في محتوى كل متغير لمعالجة القيم المفقودة والبيانات غير الصحيحة. كما يُعد اختيار المتغيرات (الحقول) وحجم العينة (الصفوف) وإنشاء متغيرات جديدة (تحويل المتغيرات) وتكويد المتغيرات من أهم محطات مرحلة تحضير البيانات للنمذجة.

وإذا لم نحتاج للعودة للخلف، ننتقل لمرحلة النمذجة باختيار نوع التنبؤ الذي يتوافق مع الأهداف التفصيلية. إذ يُعتمد على التصنيف في التنبؤ بالفئة التي ستقع فيها حالة معينة، كما يُعتمد على الانحدار للتنبؤ بالقيمة العددية التي سيأخذها المتغير التابع في ظل متغير مستقل أو أكثر؛ وإذا كان المتغير التابع يعتمد على الزمن، يتم التنبؤ باستخدام السلاسل الزمنية. وأي كان الأسلوب المنتهج، يجب مراعاة بناء عدة نماذج بديلة للتعبير عن المشكلة محل الدراسة. بل أن الأمر يتعدى ذلك بالمقارنة بين منهجي تحليل مختلفين في تناول المشكلة، مثلما هو الحال عند المقارنة بين الانحدار اللوجستي وشجرة القرارات.

وتتطلب عملية بناء النماذج التنبؤية بروتوكول مصادقة محدد لضمان الحصول على تنبؤات أكثر دقة وقوة؛ وهو ما يُعرف باسم التعلم المراقب supervised learning. ويعتمد هذا البروتوكول على تقسيم مجموعة البيانات الكلية إلى ثلاث مجموعات، ليتم تقدير النموذج من المجموعة الأولى واختباره من المجموعة الثانية والتحقق من مصداقيته بتطبيقه على المجموعة الثالثة.

ومن الضروري جداً –وقبل انتاج القرار النهائي– استخدام أكثر من معيار لتقييم النماذج البديلة لتحديد أفضلها. فإذا تحقق ذلك، ننتقل إلى تفسير النموذج وإنتاج محرك القرار. وإذا لم يحقق أي منها أهداف التحليل، يكون من الضروري العودة للخلف للبحث عن نموذج جديد يناسب التحليل أكثر. وتُختتم عملية التنقيب في البيانات بالنشر؛ أي تطبيق أفضل نموذج على البيانات الجديدة لتوليد التنبؤات أو التقديرات عن النواتج المتوقعة. ويوضح شكل (10) سلسلة الأنشطة المصاحبة لخطوات التنقيب في البيانات في الجامعة وتفاصيل كل خطوة.

4. جامعة الطائف والتنقيب في البيانات

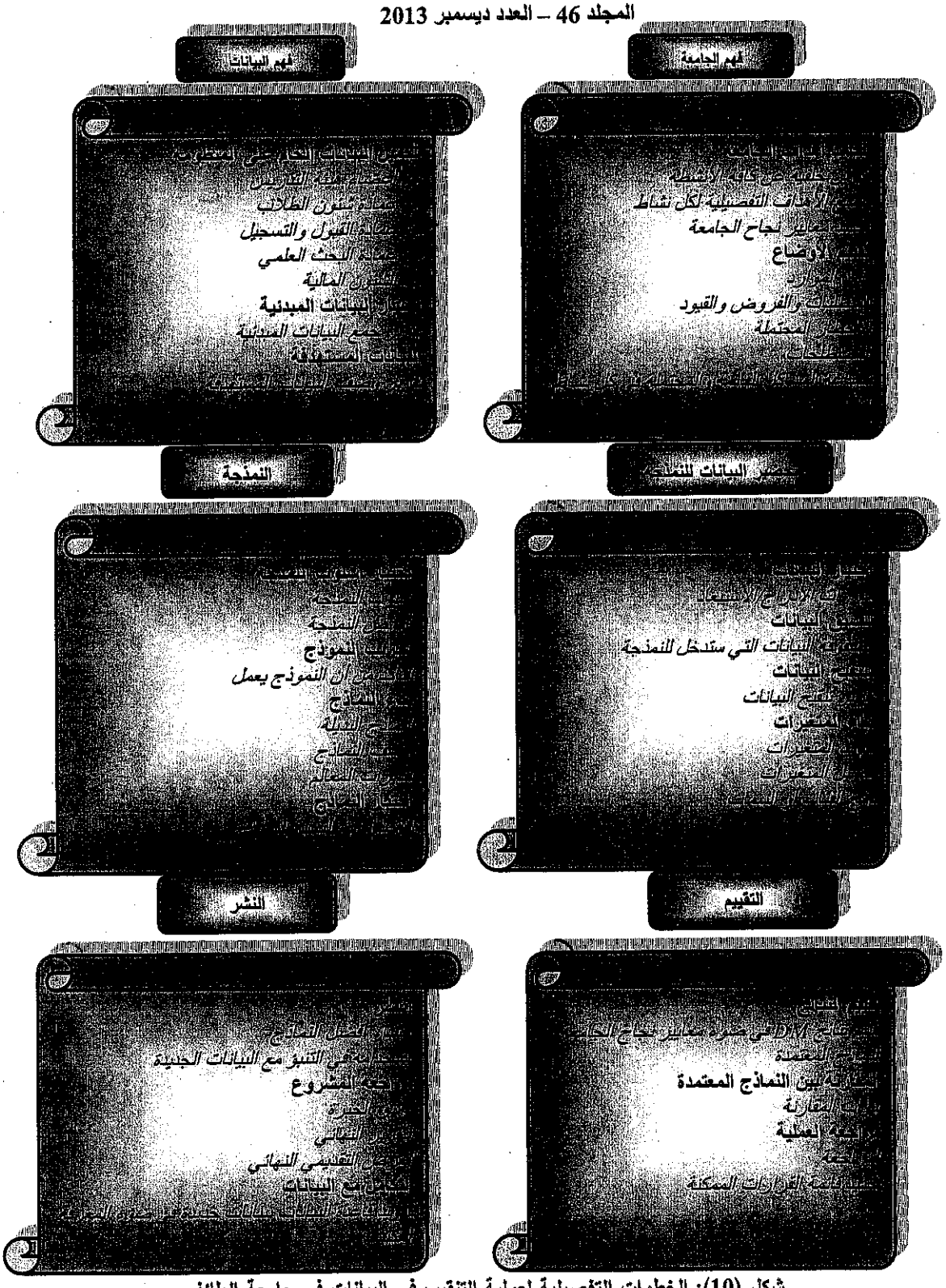
يتناول هذا الفصل متوسط الوقت المستغرق بالفعل للمراحل المختلفة لعملية التنقيب في البيانات في المشاريع المختلفة، ومهام التنقيب في البيانات في جامعة الطائف، وأسماء البرامج التي اعتمدت عليها التطبيقات الحديثة في إنجاز مشاريعها.

(4–1) توزيع الوقت على مراحل التنقيب في البيانات

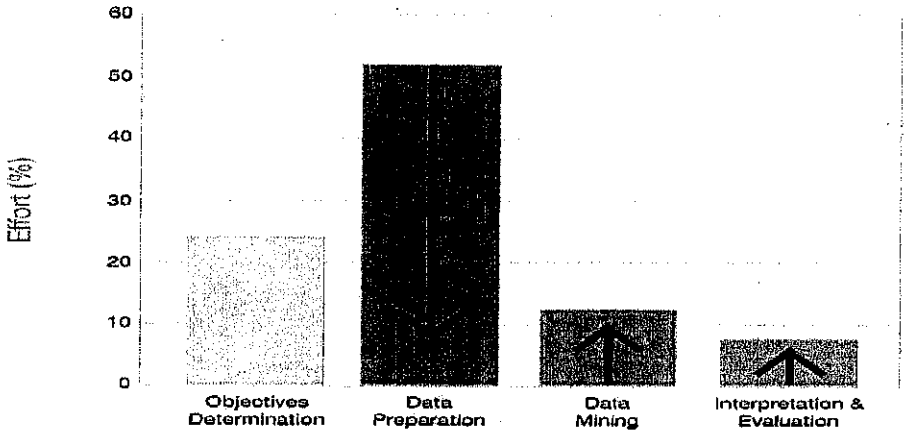
تتفق معظم دراسات التنقيب في البيانات على تخصيص ربع الوقت المتاح للمشروع لتحديد الأهداف، وترى أنها نسبة معقولة. أما تحضير البيانات للنمذجة فيستحوذ على حوالي نصف الوقت المخصص للمشروع، وهي نسبة يجب العمل على تقليلها. كما تستحوذ مرحلتي التنقيب في البيانات (الاستكشاف والنمذجة والاختبار) والتفسير والتقييم على نسبة 12% و 9% على التوالي، وهو ما يجب العمل على زيادته على حساب مرحلة التحضير. ويبين شكل (11) متوسط توزيع الوقت على مراحل مشاريع التنقيب في البيانات التي أُنجزت بالفعل، حيث تبين الأسهم الآمال التي نطمح لتحقيقها في المستقبل بالنسب للجهود المبذولة في المراحل المختلفة لتلك المشاريع.

(4–2) مهام التنقيب في البيانات في جامعة الطائف

استحدثت الجمعية العالمية للتنقيب في البيانات التعليمية (EDM) في عام 2008 تخصصاً جديداً يهتم بتطوير أساليب استكشاف البيانات التي تأتي من البيئات التعليمية، واستخدام هذه الأساليب من أجل فهم أفضل للطلاب، والأوضاع التي يتعلمون فيها. واتفقت الجمعية مع البحوث السابقة للتنقيب في بيانات منشآت الأعمال بأن مهام التنقيب في بيانات المؤسسات التعليمية أيضاً تنحصر في: التنبؤ (التصنيف والانحدار وتقدير دالة الكثافة الاحتمالية)، والعنقدة، والتعلم من خلال قواعد الاقتران. ويوضح جدول (2) بعض الأمثلة التي يمكن اعتبارها مهاماً للتنقيب في البيانات في جامعة الطائف، كما يوضح شكل (12) بعض التطبيقات المقترحة للتحليل التنبؤي.



شكل (10): الخطوات التفصيلية لعملية التنقيب في البيانات في جامعة الطائف

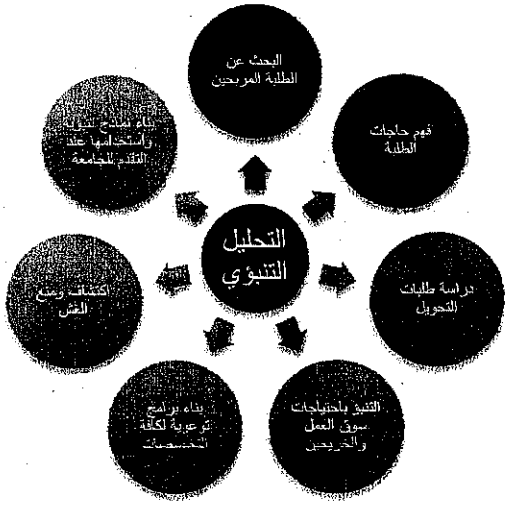


Source: Bajcsy; P. (2010)

شكل (11): متوسط الوقت الذي تستغرقه مراحل التنقيب في البيانات في المشاريع المختلفة

جدول (2): الأمثلة المقترحة لمهام التنقيب في البيانات في جامعة الطائف

المهمة	الأسلوب	الخوارزمية	المثال
التنبؤ	الانحدار	الانحدار الخطي، وغير الخطي، واللامعلمي، والمتدرج، واللوجستي	جميع المتغيرات المستمرة التي تأخذ قيما عددية مثل: التنبؤ بمعدلات الطلاب
	التصنيف	المصنفات الخطية (التمايز الخطي – بييز)، وآلات دعم المتجهات، وشجرة القرارات، ونماذج ماركوف الخفية، والشبكات العصبية، والشبكات البييزية، وأقرب الجيران	النماذج التي تحتوي على خليط من المتغيرات الكمية والتصنيفية مثل: نموذج تحديد بدل التميز والرضا الوظيفي وقرار تجديد العقود وتقييم المتقدمين للوظائف
	تقدير دالة الكثافة الاحتمالية	قاعدة بييز لتقدير الاحتمالات البعدية بمعلومية الاحتمالات القبلية	تقدير تعثر الطلاب
العنقدة	التحليل العنقودي	العنقدة المتدرجة، والمبنية على التمركز، والمبنية على التوزيع، والمبنية على الكثافة	تقسيم الطلاب إلى شعب حسب خصائصهم المتشابهة (كالمعدل)، وهو ما يسمح بمتابعة أدق لكل من الطلاب المتفوقين والمتعثرين
التنقيب في العلاقات	قواعد الاقتران	Apriori, Eclat, FP-growth, OPUS search, GUHA	تحديد الطلاب الذين سيختارون مواد اختيارية معينة معاً



شكل (12): قيمة التحليل التنبؤي في جامعة الطائف: التطبيقات النموذجية

(4-3) برمجيات التنقيب في البيانات

هناك فئتان من برامج التنقيب في البيانات؛ البرامج المفتوحة والبرامج التجارية. ويبين جدول (3) بعض الأمثلة على تلك البرامج:

جدول (3) بعض الأمثلة على برامج التنقيب في البيانات

الفئة	اسم البرنامج	الوظيفة
المفتوحة	R	لغة برمجة وبيئة برامجية للتحليل الإحصائي، والتنقيب في البيانات، والرسم البياني كجزء من GNU project
	Weka	مجموعة من تطبيقات تعليم الآلة كُتبت بلغة Java
	Carrot2	التنقيب عن النصوص والبحث عن النتائج في إطار العنقدة
	Orange	مجموعة برامج للتنقيب في البيانات وتعليم الآلة مكتوبة بلغة Python
	ELKI	برنامج للتعامل مع التحليل العنقودي المتقدم واكتشاف القيم الشاذة بواسطة البرمجة
	RapidMiner	بيئة لتعليم الآلة وتجارب التنقيب في البيانات
	UIMA	إطار لتحليل المحتوى مثل النصوص والأصوات والفيديو، قُدم في الأصل بواسطة IBM
التجارية	IBM SPSS Modeler	برنامج تنقيب في البيانات بواسطة IBM
	SAS: Enterprise Miner	برنامج تنقيب في البيانات بواسطة SAS Institute
	Oracle Data Mining	برنامج تنقيب في البيانات بواسطة Oracle
	STATISTICA	برنامج تنقيب في البيانات بواسطة StatSoft
	Microsoft Analysis Services	برنامج تنقيب في البيانات بواسطة Microsoft
	LIONsolver	تطبيق برامجي متكامل للتنقيب في البيانات، وذكاء المنشأة، والنمذجة
	DTREG	برنامج للنمذجة التنبؤية والتوقع

وعلى الرغم من أن البحوث السابقة لم تتفق على أفضلية مطلقة لأحد هذه البرامج، إلا أن Rexer et al (2010) توصلوا في مسحهم الميداني لآراء منقبي البيانات إلى أن (43%) منهم يستخدمون برنامج R أكثر من البرامج الأخرى بالنسبة للبرامج المفتوحة، كما احتلت برامج IBM SPSS Modeler، STATISTICA مقدمة التقييم بالنسبة للبرامج التجارية.

الخلاصة:

فى ضوء ما سبق عرضه:

- خلص البحث إلى تصميم عمليه أو خطوات محددة أو نموذج مقترح للتنقيب في البيانات واكتشاف المعرفة في جامعة الطائف.
- اقترح البحث المهام المتوقعة للتنقيب في بيانات الجامعة.
- اقترح البحث الخطوات التفصيلية داخل كل مرحلة من مراحل العملية المقترحة للتنقيب في البيانات.
- عرض البحث كافة المفاهيم والمصطلحات التي تيسر انتهاج تقنيات التنقيب في البيانات في حالة إقراره.
- التعرف على المشاكل الأساسية التي تواجه أنشطة الجامعة المختلفة وترجمتها إلى أهداف تفصيلية قابلة للتحليل.
- توسيع قاعدة بيانات الجامعة لتشمل بيانات سوق العمل والبيانات الاجتماعية الاقتصادية للطلاب، وإتاحة الفرصة للمحللين للوصول إلى نسخ منها. ففي ظل الوضع الراهن لن تستطيع المنظومة الإجابة على الأسئلة التالية:
  - هل هناك توازن بين عرض الخريجين والطلب عليهم؟ لأن ذلك يتطلب إدخال بيانات طلب سوق العمل على التخصصات المختلفة والتوزيع السكاني إلى المنظومة.
  - هل توجد فجوة بين الدراسة الأكاديمية وتطبيقها في الواقع العملي؟ لأن ذلك يتطلب إدخال بيانات من أرباب العمل عن المهارات النوعية التي يجب توافرها في الخريجين. وهل يحقق التدريب الهدف المرجو منه في تقليص تلك الفجوة؟
  - من هم الطلاب الذين تعثروا بفعل الظروف الاجتماعية؟ لأن ذلك يتطلب إدخال بيانات عن العلاقات الأسرية وظروف السكن وحجم الأسرة.

وفى الختام، نرى أن جامعة الطائف مؤهلة بشكل كبير لتطبيق التنقيب في البيانات لأنها تمتلك قاعدة بيانات عملاقة وتتعامل مع كم كبير من البيانات ولديها عدد معقول من الكوادر المتخصصة في كل من الإحصاء التطبيقي ونظم المعلومات والبرمجة. وينقصها انتهاجها لهذا المدخل صراحةً، وتوسعيها لقاعدة بياناتها لتسمح بالإجابة على الأسئلة المطروحة في التوصيات، والسماح للإحصائيين بالوصول إلى تلك البيانات.

المراجع

[1] Anand; S, Patrick; A, Hughes; J and Bell; D (1998), A data mining methodology for cross-sales. *Knowledge Based Systems Journal*, 10, pp. 449–461.
[2] Bajcsy; P. (2010), Introduction to Data Mining, Automated Learning Group, National Center for Supercomputing Applications, University of Illinois, http://www.NCSa.com.
[3] Cabena; P, Hadjinian; P, Stadler; R, Verhees, ; and Zanasi; A (1998), Discovering Data Mining: from concepts to implementation. *Prentice Hall.*
[4] CRISP-DM (2003), CRoss Industry Standard Process for Data Mining, http://www.crisp-dm.org.
[5] Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds) (1996a), Advances in Knowledge Discovery and Data Mining, *AAAI Press*.
[6] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996b), From Data Mining to Knowledge Discovery: An Overview. In Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds), Advances in Knowledge Discovery and Data Mining , *AI, DDM, AAAI/MIT Press*, pp. 1-34.
[7] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996c), The KDD process for extracting useful knowledge from volumes of data, ***Communications of the ACM*, 39 (11), pp. 27-34.**
[8] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996d), Knowledge Discovery and Data Mining: Towards a unifying framework, *AI, DDM, AAAI/MIT Press*, pp. 82-88.
[9] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996e), From data mining to knowledge discovery in databases, AI Magazine,17, (3), pp. 37-54.
[10] Giudici; P. (2003), Applied Data Mining: Statistical Methods for Business and Industry, *John Wiley & Sons Ltd.*
[11] http://en.wikipedia.org/wiki/Data_mining (2010).
[12] IBM, The data mining process, publib.boulder.ibm.com/.../c_dm_process.html.
[13] International Educational Data Mining Society, http://www.educationaldatamining.org/
[14] Kurgan; L.A. and Musilek; p. (2006), A survey of Knowledge Discovery and Data Mining process models, *The Knowledge Engineering Review*, Vol. 21:1, pp. 1–24, Cambridge University Press.
[15] Linoff and Berry (2011), Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd ed., Wiley.
[16] Lucas; P. (2004), Bayesian Analysis, Pattern Analysis and Data Mining in Health Care, *Current Opinion in Critical Care*, 10:pp. 399-403.
[17] Lukawiecki; R.(2010), Introduction to Data Mining Project Botticelli Overview, download.microsoft.com/.../DATAMIN/IntroductiontoDataMining.ppt.
[18] Microsoft SQL Server (2010), http://technet.microsoft.com/en-us/library/ms174949.aspx.
[19] Newton's Telecom Dictionary (2010), Harry Newton, CMP Books, http://www.cmpbooks.com.
[20] Rexer; K., Allen; H., & Gearan; P. (2010), Data Miner Survey Summary, presented at Predictive Analytics World.
[21] Rygielski; C., Wang; J. and Yen; C. (2002), Data mining techniques for customer relationship management, *Technology in Society*, 24, pp. 483–502.
[22] Rosenblatt; F. (1962), Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism. Spartan, Washington DC.
[23] SAS (1997), From Data to Business Advantage: Data Mining, SEMMA Methodology and the SAS System, *SAS Institute Inc*, White Paper.
[24] SAS Institute (2001), SAS Enterprise Miner Reference Manual, *SAS Institute Inc.*, Cary NC.
[25] SPSS (2009), Clementine 16.0, *SPSS*, Inc. http://www.spss.com/spssbi/clementine/.
[26] Statistica, Data Mining Techniques, *Statsoft Electronic Statistics Textbook* www.statsoft.com/textbook/data-mining-techniques/.
[27] Tan; P.N., Steinbach; M. and Kumar;V. (2006), Introduction to Data Mining, ***Addison-Wesley Companion Book Site*,** www.cs.umn.edu/~kumar/dmbook.
[28] Thearling; K. (2009), An Introduction to Data Mining: Discovering hidden value in your data warehouse, www.thearling.com/text/dmwhite/dmwhite.htm.
[29] Tsui; K. (2009), Introduction to Data Mining, *Industrial & Systems Engineering*, Georgia Institute of Technology, www2.isye.gatech.edu/~shan/.../Introduction_to_Data_Mining.pdf.