REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'Informations Avancé

Par: Benslimane Afaf

SUJET

Qualité des données dans le processus ECD

Soutenu publiquement le : 1 / 06 /2016 devant le jury composé de :

N.Amroun	**Université de M'sila**	**Président**
T.Mehenni	**Université de M'sila**	**Rapporteur**
M.Bounif	**Université de M'sila**	**Examinateur**

Promotion : 2016 /2017

Table de matière

INTRODUCTION GENERALE

L'Extraction de Connaissances à partir de Données (ECD) consiste à parcourir d'immenses volumes de données contenues dans une base, à la recherche de connaissances. C'est une discipline qui se situe à l'intersection de différents domaines tels que l'informatique, l'intelligence artificielle, l'analyse de données, les statistiques, la théorie des probabilités, l'optimisation, la reconnaissance de formes, les bases de données et l'interaction Homme-Machine..

L'ECD est le processus non trivial, interactif et itératif qui permet d'identifier des modèles valides, nouveaux, potentiellement utiles et compréhensibles à partir de bases de données massives. Le terme processus signifie que l'ECD se décompose en plusieurs opérations, allant de la phase de compréhension du domaine étudié jusqu'à l'interprétation des résultats, en passant par plusieurs étapes de sélection et de préparation des données qui s'avèrent très importantes pour garantir des résultats efficaces.

La phase de prétraitement est certainement l'une des phases de préparation des données la plus complexe. Elle consiste à nettoyer les données, les mettre en forme, traiter les données manquantes, échantillonner les individus, sélectionner et construire des variables, etc. On obtient ainsi un ensemble de données cibles. Cette phase a une place importante au sein du processus d'ECD car c'est elle qui va déterminer la qualité des modèles construits lors de la phase de fouille de données. Elle peut prendre jusqu'à 60% du temps dédié au processus d'ECD.

Notre travail consiste à mettre en évidence la phase de prétraitement des données, en présentant les méthodes et techniques y afférentes. Nous décrivons et implémentons un ensemble des ces méthodes que nous appliquons sur un extrait de données réelles.

Puisque la phase de prétraitement consiste à nettoyer les données pour assurer de bons résultats de fouille de données, nous abordons une évaluation des méthodes de prétraitement implémentées ; en comparant les résultats d'une méthode de fouille de données appliquées sur les données avant prétraitement et après prétraitement. Ceci pour démontrer que la qualité des données joue un rôle important sur la qualité des résultats de douille de données.

Le mémoire est composé de trois chapitres. Dans le premier nous présentons le processus ECD ainsi que les méthodes de datamining. Le deuxième chapitre est consacré à la description des différentes méthodes de prétraitement. Enfin le dernier chapitre décrit en détail les fonctionnalités de notre système de nettoyage, ainsi que la stratégie d'évaluation des méthodes de prétraitement implémentées dans le système.

BIBLIOGRAPHIE

[1] D. Kindjangu, l'informatisation de la gestion des abonnés de la SNEL, Société nationale d'électricité en RDC, 2012

[3] T. Mehenni, Data mining, cours de Master, Université Mohamed Boudiaf de M'sila, 2015/2016

[4] C. Bernard et P. Craveski, Classification de données, Cours de Master, Université Claude Bernard, Lyon, 2014

[5] R, Elamin, Data mining : techniques de DM pour la GRC dans les Banques, Thèse de doctorat ,Université de Biskra, 2015

[6] A. Djeffal , Fouille de données avancée, Cours de Master, Université Mohamed Khider Biskra, 2015/2016

[7] Google, https://simple-mail.fr/fonctionnalites_delivrer consulté avril 2016

[8] Wikipidia, https://fr.wikipedia.org/wiki/Nettoyage_de_donn%C3%A9es consulté mai 2016

[9] Comment ça marche, http://www.commentcamarche.net/contents/104-bases-de-donnees-introduction, consulté mai 2016

[10] T. Gatid , Mini-mémoire de BDA , dernière mise à jour le 22 Mars 2002

[11] F. Weber, Etude Statistique ,en 2005-2006

[12] D.Donsez, Intégration des données, Université Joseph Fournier, Cours de Doctorat, 2002

[13] M. Boufaida , Adaptation de technique de l'extraction des connaissance à partir de données , Thèse de doctorat, Université Mentouri Constantine, 2012

[14] D. Abdelkader et R. Rakotomalala, Extraction des connaissances à partir de données, Techniques de l'ingénieur, 2002

[15] B. Liaudet, Cours de data mining : Modélisation et présentation générale, Université de Finance, Septembre 2008

[16] D. Chami, La plate forme orientée agent pour le data mining, Mémoire de Magister, Université Hadj Lakhedar Batna, 2009/2010

[17] G. Calas, Etude des principaux algorithmes de data minig, Ecole ingeneur en informatique, Le Kremlin-Bicêtre, France, 2009

[18] F. Bash, K-means et théorie des graphes, Cours, Université Alexandre Boulch, 2010

[20] M.N. Mami, Extraction des connaissances dans l'environnement distribué, Mémoire de Master,_Ecole Nationale des Sciences de l'Informatique, 2013

[21] A. El Mhamdi, Gestion proactive du changement dans les projets de réingénierie des processus métiers, Thèse de Doctora , Université de Paris 8, 2009

[22] M. Jacques, Cours de Bases de données, Université du Sud Toulon-var, 2014

[23] M. Parmami, Mémoire fin d'études, 2013.

[24] S. Caron, Une introduction aux arbres de décision, Mémoire de License, France, 2011

[25] J.Han , M. Kamber ,Data Mining Concepts and Techniques deuxième édition, Stanley B. Zdonik et D.Maier . Etats-Unis d'Amerique,2006

[26] J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels, Exploratory Data Mining and Data Cleaning , J.Wiley ,Sons, Inc., Hoboken, N. Jersey, Canada., 2003

CONCLUSION GENERALE

Il va sans dire que la qualité d'un travail est l'objectif de tous. La qualité du résultat obtenu d'un algorithme est sans aucun doute la plus chère des recommandations de n'importe quel informaticien.

Notre travail s'est focalisé sur la qualité des données en vue d'obtenir une bonne qualité des résultats rendus d'un algorithme de fouille de données. Pour cela, nous nous sommes intéressés aux méthodes de prétraitement des données. Nous avons implémenté un certain nombre de ces méthodes et les ont appliquées sur un extrait de base de données d'une agence de location des véhicules.

Toutefois, avant d'aboutir à l'implémentation du système, nous avons présenté le processus ECD et les différentes méthodes de data mining, ceci pour montrer l'intérêt d'avoir une bonne qualité des données. Nous avons ensuite décrit les différentes étapes de prétraitement des données ainsi que la majorité des méthodes et techniques y afférentes.

Dans notre système de nettoyage, nous avons sélectionné trois types de prétraitement auxquels nous avons implémenté des algorithmes pour leur mise en œuvre. Ces types de prétraitement sont : traitement des valeurs manquantes, élimination des mots étrangers et traitement de l'incohérence des données.

Comme l'objectif de notre système de nettoyage est d'avoir des données plus propres pour avoir les meilleurs résultats de fouille de données, nous avons procédé à l'évaluation de l'ensemble des méthodes de nettoyage implémenté dans le système, en appliquant une méthode de data mining (classification bayésienne naïve) sur les données de l'agence de location avant prétraitement et après prétraitement par le système. Le but de cette évaluation est de comparer les résultats obtenus pour montrer que la qualité des données influe largement sur les résultats du data mining. Les évaluations ont montré qu'une partie importante des résultats ont été amélioré après application des méthodes de prétraitement.

Le système de nettoyage que nous avons implémenté est sujet à des améliorations. Plusieurs autres méthodes de prétraitement que nous n'avons pas implémentées peuvent être intégrées. Aussi, nos perspectives est d'avoir un système indépendant de la nature des données, ainsi qu'un ensemble de méthodes de data mining pour mieux évaluer la qualité des données.

ملخص: إن عملية استخراج المعلومات من خلال البيانات تتم عبر البحث في ثنايا الكم الهائل من البيانات عن المعلومات المختبئة فيها. وتتم هذه العملية عبر مراحل عديدة بدءا من الفهم الدقيق لميدان هذه المعلومات إلى مرحلة تفسير النتائج المتحصل عليها، مرورا بعدة مراحل ثانوية يتم فيها اختيار البيانات وتحضيرها والتي تبين أنها مهمة للغاية لأجل ضمان نتائج ذات جودة عالية. إن عملية تطهير البيانات هي حتما أصعب العمليات ضمن مرحلية التحضير. إن مشروعنا يتمثل في تسليط الضوء على هذه العملية وذلك بتقديم معظم الطرق والتقنيات المستخدمة لأجل ذلك. وقد أنجزنا تبعا لذلك نظاما لتطهير البيانات من خلال برمجة عدة خوارزميات وقمنا بعد ذلك بتطبيقها على مجموعة حقيقية من البيانات.

كلمات مفتاحية: استخراج المعلومات، جودة البيانات، المعالجة القبلية، التطهير، تصنيف بايز (Bayes)

Abstract : Knowledge extraction from data consists in browsing huge volumes of data contained in a database, in order to search knowledge. This process it composed of several operations, which begins by understanding the domain studied until the interpretation of the obtained results, while passing by several stages of selection and preprocessing of data that prove to be very important to guarantee efficient results. The data cleaning is certainly the most complex phase of data preparation. Our work consists in highlighting the cleaning phase of data, by presenting several methods and techniques that perform this task. We describe and develop a data cleaning system that performs a set of cleaning methods and apply them on real data.

Keywords: Knowledge extraction, data quality, preprocessing, cleaning, Bayesian classification.

Résumé : L'Extraction de Connaissances à partir de Données (ECD) consiste à parcourir d'immenses volumes de données contenues dans une base, à la recherche de connaissances. Il se décompose en plusieurs opérations, allant de la phase de compréhension du domaine étudié jusqu'à l'interprétation des résultats, en passant par plusieurs étapes de sélection et de préparation des données qui s'avèrent très importantes pour garantir des résultats efficaces. La phase de prétraitement est certainement l'une des phases de préparation des données la plus complexe. Notre travail consiste à mettre en évidence la phase de prétraitement des données, en présentant les méthodes et techniques y afférentes. Nous décrivons et implémentons un ensemble des ces méthodes que nous appliquons sur un extrait de données réelles.

Mots clés : Extraction des connaissances, qualité des données, prétraitement, nettoyage, classification bayésienne