Techniques d'analyse et de conception d'outils pour la gestion du processus de segmentation des abonnés des entreprises de télécommunication

R. Carlos Nana Mbinkeu[1,3], C. Tangha[1], A. Chomnoue[1], A. Kuete[2]

[1]Ecole Nationale Supérieure Polytechnique - Université de Yaoundé I, Cameroun

[2]Orange Cameroun SA – Groupe France Télécom

[3]DBGroupLab – Université de Modène, Italie

{nana_mbinkeu, nanacarlos2004}@yahoo.fr, ctangha@gmail.com

RÉSUMÉ. La plupart des entreprises de télécommunication sont en possession d'une remarquable quantité de données concernant leurs clients. Les connaissances que l'on voudrait découvrir dans leurs bases de données sont par exemple l'éventuel regroupement des abonnés au sein de sous-ensembles homogènes dont les membres ont des comportements de consommation similaires. On parle de segmentation de la clientèle. Cette segmentation va permettre d'identifier et de sélectionner les clients les plus susceptibles de réagir favorablement aux offres. Notre article propose des techniques d'analyse et de conception d'outils qui permettent de gérer le processus d'acquisition des données, de nettoyage des données et le choix de l'algorithme de segmentation.

ABSTRACT. Most telecommunications providers already possess a remarkable amount of data about their customers. The knowledge that we would discover in the database of telecommunications providers is vital to understanding the customers, and why high customer churn and declining brand loyalty continue to erode profit margins. We talk about customer segmentation. The segmentation will identify and select the customers most likely to respond favorably to offers. Our paper proposes a set of techniques to analyze and conceive tools that manages the process of data acquisition, data cleaning and selection of the segmentation algorithm. Our techniques help manage the process of segmentation of subscribers.

MOTS-CLÉS : Fouille de donnée, datawarehouse, télécommunication, marketing, segmentation.

KEYWORDS: Data Mining, datawarehouse, telecommunication, marketing, segmentation.

1. Introduction

L'industrie des télécommunications génère et stocke d'énormes quantités de données [1]. Ces données incluent les détails sur les abonnés, leurs utilisations du réseau, l'état des composantes matérielles et logicielles de ce réseau [1]. Mais ce potentiel énorme n'est pas toujours exploité de manière optimale. La fouille des données consiste tout d'abord à extraire les connaissances utiles de la masse d'informations disponibles. Les connaissances ainsi récoltées peuvent être employées à différentes fins: interpeller individuellement les clients, présenter des offres adaptées aux groupes cibles, prédire le comportement des clients. Ces connaissances permettent aussi d'identifier les clients susceptibles d'aller à la concurrence : on parle de churn [1]. La segmentation est très importante dans le domaine du marketing direct [2].

L'objectif de notre article est de proposer des techniques d'analyse et de conception d'outils capables de gérer tout le processus de segmentation de la clientèle d'une compagnie de téléphonie mobile. Ces techniques facilitent la gestion automatisée des phases d'acquisition et de nettoyage des données, le choix de l'algorithme de segmentation, la visualisation des segments recherchés et des statistiques [5].

Ce papier est organisé de la manière suivante : la section 2 nous présente les outils de fouille de données. Dans la section 3, nous présentons la gestion du processus d'acquisition des données. La section 4 fait une description du prototype réalisé. Nous terminons par une conclusion.

2. Les outils de fouille de données

La fouille de données n'existerait pas sans outils. Les logiciels de fouille de données sont des programmes spécialisés dans l'analyse et l'extraction des connaissances à partir des données informatisées. Actuellement, les outils les plus utilisés sont : SPSS, RapidMiner, SAS, Excel, R, KXEN, Weka, Matlab, Knime, Microsoft SQL Server, Oracle DM et STATISTICA [3]. La plupart de ces outils ne permettent pas une utilisation aisée et intuitive par des personnes non experte en fouille de données.

Nos techniques permettent d'automatiser la gestion du processus de segmentation des abonnés. L'implémentation de ces techniques dans la conception d'outils va permettre aux personnes non expertes de faire de la segmentation. Elles permettent la prise en charge de toute base de données relationnelle. Dans le cadre de notre étude et la réalisation d'un prototype, nous avons utilisé un datawarehouse sous le SGBD Oracle et le serveur de datamining JDM [4,10]. Les opérations liées à la segmentation et définies par l'expert. Nos techniques utilisent la méthode CRISP-DM [5,6].

3. Gestion du processus d'acquisition des données

3.1. Compréhension des données

La compréhension des données à partir d'une base de données existante peut nécessiter la définition d'une ontologie. Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance [7]. Le choix des méthodes à mettre en œuvre pour préparer les données, notamment pour discrétiser certaines, repose en grande partie sur cette étape, au cours de laquelle l'ontologie doit être construite. Cette ontologie est une ontologie d'application, c'est-à-dire que les concepts définis dépendent à la fois du domaine étudié mais aussi de la tâche à effectuer. L'ontologie permet l'établissement d'un niveau supérieur d'abstraction des données afin de simplifier le processus de découverte de connaissance en termes de réutilisabilité.

Dans notre cas, nous devons représenter de façon formelle les concepts issus de la compréhension du domaine et les mettre en relation avec les tables du data-warehouse. Euler et Scholz présentent une méthode utilisant des ontologies pour la modélisation des connaissances et leur interconnexion avec les données en base [7]. Nous nous sommes inspirés de leur travail pour construire notre modèle de données dans le formalisme que nous présentons à la section suivante.

3.2. Construction d'une ontologie applicative

La construction de l'ontologie permet de modéliser les données à deux niveaux. Au premier niveau, on retrouve les tables telles qu'existant dans le data-warehouse. Au second niveau, le concept et ses attributs nous permettent de décrire les données de manière plus abstraite. Cette approche nous permet d'obtenir une nouvelle vue de la base de données existante pouvant contenir une ou plusieurs tables virtuelles [7]. Nous appelons concept le sujet d'apprentissage faisant l'objet de la segmentation et nous avons dit que les instances d'un concept sont caractérisées par des attributs. Un avantage de cette modélisation à deux niveaux est qu'il permet sa réutilisation sur une nouvelle base de données juste en refaisant l'interconnexion entre les deux niveaux [7].

Considérons par exemple le concept *« abonné »* caractérisé par un numéro de téléphone et d'autres attributs significatifs pour le processus de segmentation. Les instances du concept sont obtenues en définissant pour chaque attribut une expression SQL sur une ou plusieurs colonnes des tables du data-warehouse. Le niveau conceptuel obtenu représente donc une ontologie du domaine et offre une vue simplifiée et plus compréhensible des données du data-warehouse. Nous proposons à la sous-section 3.2.2, une procédure générale pour le calcul des instances des attributs associés aux concepts définis au niveau conceptuel.

3.2.1. La dimension temps

Nous étudions le comportement des abonnés sur une certaine période. Elle peut être définie en jours, en semaines, en mois etc… Nous nous servirons donc des mois pour définir les périodes. Les tables du data warehouse possèdent en général chacune une colonne définissant la date des évènements qu'elle enregistre. Les tables les plus granulaires donnent le jour et l'heure de l'évènement, d'autres plus agrégées donnent plutôt la semaine ou le mois. Nous allons spécifier pour chaque table une expression donnant le mois d'occurrence des évènements.

3.2.2. Calcul des attributs du concept

Nous mettons en relation les attributs du concept et les tables à partir desquelles leurs valeurs sont calculées. Nous spécifions pour chaque attribut du concept :

- la table (les tables) à partir de laquelle (desquelles) sa valeur est calculée
- une expression qui donne la valeur de l'attribut à partir d'une ou de plusieurs colonnes de la table (des tables).

Plusieurs attributs du concept peuvent être liés à la même table du data warehouse. Dans ce cas une discrimination est incluse dans l'expression de calcul.

Le modèle de données défini dans le formalisme présenté ci-dessus a des attributs qui sont en relation avec les tables du data-warehouse. Nous proposons maintenant une procédure permettant de recueillir dans la table virtuelle la valeur de ses attributs pour chaque abonné. Elle s'exécute comme suit :

1. Regrouper les attributs en fonction des tables auxquelles ils sont liés;
2. Créer une requête sur chaque table en se servant des expressions de calcul de chaque attribut ;
3. Exécuter les requêtes ainsi obtenues. Notons que ces requêtes peuvent être exécutées en parallèle ;
4. Joindre les résultats pour obtenir l'ensemble des valeurs des attributs dans la table virtuelle créée ad hoc.

Le modèle de données défini dans le formalisme ci-dessus est évolutif. Des attributs peuvent être ajoutés ou retirés du concept, l'interconnexion entre les tables et les attributs peut être modifiée et cette modification sera automatiquement prise en compte lors de la prochaine exécution du processus de collecte des données. On obtient donc un faible couplage entre le niveau conceptuel et les tables du data warehouse. Ce faible couplage répond au besoin des utilisateurs de pouvoir étendre le concept (prendre en compte de nouveaux attributs) lors des éventuelles segmentations.

3.3. Traitement de la qualité des données

La qualité des données des tables virtuelles doit être analysée. Par conséquent, les données sont nettoyées, puis réduites avant d'être soumises au serveur de datamining JDM [4,10] qui se chargera de leur séparation en segments. La réduction du concept consiste à éliminer certains attributs du concept qui n'ont aucun impact sur le calcul des segments recherchés (la procédure de calcul est illustrée au 3.3.2). Des statistiques seront ensuite calculées sur les attributs restant pour obtenir les différents segments.

3.3.1. Nettoyage des données

3.3.1.1. Traitement des valeurs inexistantes

Les attributs peuvent avoir des valeurs nulles (absence de valeur). Les valeurs absentes dans ces cas sont dues en général à la non utilisation de certains services par les abonnés. Par défaut, nous remplaçons les valeurs inexistantes par 0, sauf dans le cas où l'utilisateur spécifiera l'un des traitements suivants pour un attribut:

- **suppression** : les instances n'ayant pas de valeur pour cet attribut seront écartées de l'ensemble pour la suite du processus ;
- **substitution** : les valeurs inexistantes seront remplacées par une valeur réelle spécifiée par l'utilisateur ;

3.3.1.2. Traitement des valeurs extrêmes

Nous avons choisi lors du processus de construction des segments d'éliminer les instances extérieures à un intervalle de confiance autour de la moyenne. Dans l'approche que nous proposons, l'utilisateur peut soit :

- Spécifier un intervalle de confiance pour chaque attribut ;
- Spécifier des valeurs maximales et minimales pour chaque attribut.

Les instances se retrouvant à l'extérieur de l'intervalle défini seront écartées.

3.3.2. Réduction du concept

L'opération de réduction du concept permet de choisir les attributs qui seront effectivement utilisés pour la construction des segments [9]. Pour cela le système exécutera automatiquement les tâches que nous avons décrites dans les phases antérieures. La procédure de réduction est la suivante :

Eliminer les attributs qui ne varient pas suffisamment

1. Calculer la variance des différents attributs après normalisation ;

2. Eliminer les attributs dont la variance est en dessous d'un seuil fixé par l'utilisateur.

Séparer les attributs fortement corrélés

1. Regrouper les attributs en paires ;
2. Calculer le coefficient de corrélation pour chaque paire ;
3. Ranger les paires par ordre de corrélation décroissante ;
4. Etablir un classement des attributs en fonction de la plus grande corrélation de chacun avec un autre attribut en dessous du seuil (0,9 par défaut) ;
5. Parcourir les couples ayant une corrélation au-dessus du seuil et éliminer dans chaque couple l'un des deux attributs (s'il n'est pas déjà éliminé) ;

Appliquer le principe de description de longueur minimale :

Dans le cas où le nombre d'attributs après les étapes d'élimination citées ci-dessus reste au-dessus d'un maximum fixé, le principe de description de longueur minimale peut être appliqué à l'aide de l'algorithme 'attribute importance' spécifié dans le standard JDM [4,10]. Cela implique des interactions avec le serveur JDM. Les attributs seront classés en fonction de leur importance dans la prédiction d'un attribut cible (par défaut le revenu généré joue ce rôle) et les moins importants seront écartés.

Les données obtenues après les étapes décrites plus haut sont prêtes à être élaborées par les différents algorithmes de segmentation [11]. Les données sont accessibles à partir du serveur de datamining.

4. Prototype : Fonctionnement

Notre prototype a pour objectif de faciliter la gestion du processus de segmentation par des utilisateurs non experts en s'inspirant de l'approche de Daniel C. Robbins [8] et en tenant compte de notre contexte de travail. Dans le cadre de notre collaboration avec l'opérateur téléphonique Orange, nous avons implémenté une application web qui permet de réaliser la segmentation des abonnés en utilisant l'approche décrite ci-dessus. Nous montrons dans la figure 1, les différents algorithmes [11] (K-means, O-cluster, etc.) qui peuvent être utilisés dans notre application web de fouille de données.

L'utilisateur pour collecter les données à analyser, définit une vue à la base de données et crée l'interconnexion avec les tables de la base de données (figure 2). Chaque vue est associée à un ensemble d'attributs. L'utilisateur spécifie pour chaque attribut : l'expression de la requête qui va sélectionner les données dans les tables existantes.

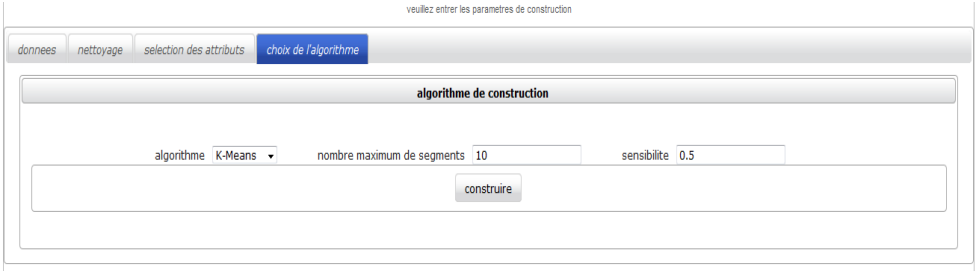

Figure 1. *Choix des algorithmes*

Le système guide l'utilisateur dans la phase de nettoyage des données (figure 3). L'utilisateur définit un ensemble d'opérations qui permettent de sélectionner les attributs pertinents (variance, corrélation). Toutes ces tâches de base définies par l'utilisateur sont stockées dans un fichier XML de l'application web. Cela a l'avantage d'automatiser le processus des futures segmentations car il ne sera plus nécessaire de redéfinir les tâches citées ci-dessus.

Figure 2. *Définition des attributs de la Vue (concept)*

L'application de fouille de données, que nous avons implémentée, a été utilisée par un operateur de téléphonie. Elle a été en mesure de gérer le processus de segmentation sur un ensemble de sept millions d'abonnés, à la grande satisfaction de l'operateur.

Figure 3. *Gestion de la qualité des données*

5. Conclusion

Nous avons proposé un ensemble de techniques qui aident à l'analyse et la conception d'outils pour la gestion du processus de segmentation des abonnés des entreprises de télécommunication. Ces techniques permettent aux personnes non expertes de faire de la segmentation grâce à une gestion automatisée du processus. Un prototype a été implémenté et est utilisé actuellement par Orange Cameroun.

Comme perspective, nous envisageons d'étendre ce travail en proposant l'ajout de nouvelles fonctionnalités suivant d'autres secteurs d'activités.

6. Bibliographie

1. Gary M. Weiss. Data Mining in the Telecommunications Industry. Data Mining and Knowledge Discovery Handbook 2005, Part 12, 1189-1201.
2. Edmunds Communications Group. The Importance of Data Mining and Segmentation in Direct Marketing, 2010.
3. K. Rexer, H. Allen, & P. Gearan, Data Miner Survey Summary, Oct. 2010.
4. Oracle Data Mining Application Developer's Guide 11g and Java API.
5. P. Chapman et al. CRISP-DM 1.0 Step-by-step data mining guide (2000).
6. A. Azevedo. KDD, SEMMA and CRISP-DM: A Parallel Overview, 2008.
7. Tim Euler and Martin Scholz. Using Ontologies in a KDD Workbench, In Workshop on Knowledge Discovery and Ontologies at ECML-PKDD 2004.
8. Daniel C. Robbins et al. ZoneZoom: map navigation for smartphones with recursive view segmentation, AVI'04, January 2004.
9. Igor Kononenko. Evaluating the Quality of Attribute, 2005.
10. Oracle Data Mining Concepts, 10g release 2. Part Number B14339-01.
11. Ayaz Ali, A. Bagherjeiran and Chen. Scalable Clustering Algorithms, 2004