Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'étude pour l'obtention du diplôme de

Master en informatique

OPTION : *Système D'information et de Connaissance (SIC)*

Thème

Estimation des donnés manquantes et catégorisation des identités sociaux

Réalisé par :

- Naim El Hosseyn

Sous la direction de **Mr Abdeldjalil KHELASSI**

Présenté le 24 Juin 2014 devant la commission d'examination composée de :

- Mr Y.Benziane (Président du jury)
- Mr A.Benamar (Examinateur)
- Mr H.Matallah (Examinateur)

Année universitaire: 2013-2014

Remerciements

Je remercie tout d'abord le bon Dieu, le tout puissant de m'avoir armé de force et de courage pour mener à terme ce projet.

Merci à tous ceux qui ont contribué de près ou de loin à ce que la réalisation de ce projet soit possible.

Ma profonde gratitude s'adresse à Monsieur ***Abdeldjalil KHELASSI****, encadreur de ce mémoire pour ses conseils fructueux et pour son aide précieuse qui ma conduits à concrétiser ce travail.*

Mes vifs remerciements s'adressent à monsieur ***Benziane*** *d'avoir honorer par sa présence en acceptant de présider le jury.*

Mes respectueux remerciements sont dédiés aux membres de jury, monsieur ***Benamar*** *et* ***Matallah*** *d'avoir accepté d'examiner et de porter leur jugement sur ce modeste travail.*

.

DIDICACE

C’est grâce à Allah seul que j’ai pu achever ce travail.

Je le dédie à :

***M**a très chère mère, qui a toujours été présente pour moi, dans les moments les plus difficiles et qui sans cesse veille sur moi avec ses prières, pour ses grands sacrifices et tout l’amour qu’elle me porte.*

***M**on très cher père, pour tous ses conseils et pour toute la confiance qu’il a mise en moi et pour son dévouement pour mon bonheur.*

Que dieu me le garde.

***M**es chères frères et sœurs*

***T**ous mes enseignants*

***T**ous mes amis*

***T**ous mes collègues*

***T**oute ma famille*

Naim El Hosseyn

Résumé

Résumé :

Plusieurs cheminements pourraient être impliqués dans le développement de la catégorisation, différents type de données utilisé pour faire la catégorisation. Le problème majeur de la catégorisation est la présence des données manquantes, tel que ceci se produit pour de nombreuses raisons : erreurs de saisie, rubriques non renseignées dans des enquêtes, valeurs aberrantes qu'on préfère supprimer, données recueillies difficilement, ainsi que les internautes ne décrivent pas toutes leurs propres informations pour inscrire dans un réseau social, et ça devenu un inconvénient pour faire une bonne catégorisation. Donc il faut estimer les valeurs manquante et les remplacer, pour cela il ya différentes méthodes de traitements des donnés manquante.

Dans le cadre de ce projet, j'ai présenté une application de catégorisation des identités sociaux à partir d'une base de données qui contient des informations de certain utilisateur de réseau sociale Facebook. Pour commencer, on a déjà fait la classification avec des méthodes non symbolique proposer par l'outil « Weka » avant d'essayé d'imputer les données manquantes, les résultats obtenue ne sont pas bien catégoriser tel que ils restent un nombre d'instances non regrouper. Donc il faut essayer après imputation, pour cela on a utilisé deux méthodes, la première avec la méthode de « Plus Proche Voisin : KNN », et la deuxième est : l'imputation par la moyenne. La seconde donne une bonne estimation et elle a nous aidé à faire une bonne catégorisation.

Ensuit la création d'une application JAVA avec une interface graphique facile a gérer pour faire la catégorisation selon certain attribut, pour cela il faut passer par la réalisation des base de connaissance à l'aides de la création des règles de production XML.

Mots clé :

Réseaux sociaux, Catégorisation, Données Manquante, Weka, JRuleengine, XMLRule

Abstract:

Several pathways may be involved in the development of categorization, different types of data used for categorization. The major problem of categorization is the presence of missing data, as this occurs for many reasons: input errors, items not indicated in surveys, it is preferred to remove outliers. And users do not disclose all information to their own part of a social network, and it becomes a disadvantage to make a good categorization. So we must estimate the missing value and replace it, why there are different methods of treatment missing data. In this project, I submitted an application for categorization of social identities from a data set which contains information of some social network "Facebook" user.

To start has already been made with the classification of non-symbolic methods found in the "Weka" tool before trying to impute missing data; the results obtained are not very good as they remain a number of instances not group. So try after charging for it we used two methods, the first with the method of "K-Nearest Neighbor: KNN," and the second is the mean imputation. The second gives a good estimate and has helped us make a good categorization.

Résumé

Following this, a Java application is developed with an easy graphical interface to manage categorization according to some attribute, for that you need to go through the achievement of the knowledge base to support the creation of XML production rules.

Keywords:
Social networks, Categorization, Missing Data, Weka, JRuleengine, XMLRule

ملخص:
عدة مسارات يمكن أن تستخدم في تطوير التصنيف، ايضا أنواع مختلفة من البيانات تستخدم للتصنيف. المشكلة الرئيسية من التصنيف هو وجود بيانات مفقودة، كما يحدث هذا لأسباب عديدة: أخطاء الإدخال، والبنود التى لم يشر اليها في الدراسات الاستقصائية، ويفضل إزالة القيم المتطرفة،كما أن معظم المستخدمين لا يكشفون عن كل المعلومات الخاصة بهم في الشبكات الاجتماعية، وأصبح الوضع غير مؤات لإجراء تصنيف جيد. لذلك لا بد لنا من تقدير القيم المفقودة واستبدالها، ولهذا هناك أساليب مختلفة من الطرق الخاصة لحل هذا المشكل.

في هذا المشروع، قدمت تطبيقا للحصول على تصنيف الهويات الاجتماعية من ملف يحتوي على معلومات من بعض مستخدمي الشبكة الاجتماعية الفيسبوك.

بدءنا التصنيف بطرق غير رمزية وجدت في أداة "ويكا" قبل محاولة تعويض البيانات المفقودة، والنتائج التي تم الحصول عليها ليست جيدة جدا لأنه لا يزال هناك عدد من الحالات الغير المجمعة. بعد ذلك قمنا بمحاولة استنتاج القيم الناقصة، ولأجل ذلك استخدمنا طريقتين، الأولى مع طريقة " ك الجار الأقرب"، والثاني هو احتساب المتوسط. فاعطت الاخيرة تقديرا جيدا كما ساعدتنا في جعل تصنيف جيد.

بعد ذلك انشاءنا تطبيق جافا مع واجهة سهلة لإدارة التصنيف وفقا لبعض المعلومات لذلك علينا أن ننشأ قاعدة المعرفة لدعم تطبيق خاصية التصنيف و هذا بمساعدة قواعد "اكس ام ال".

الكلمات الرئيسية :
الشبكات الاجتماعية، التصنيف، البيانات المفقودة ، ويكا،جي رول انجاين،اكس ام ال رول.

Table des matières

SOMMAIRE

Table des matières

Liste de figures

Liste des tableaux

INTRODUCTION GENERALE

I. Introduction générale

Aujourd'hui avec le développement d'internet nous sommes en présence d'une quantité énorme d'utilisateurs. Pour pouvoir gérer les données et tirer le plus d'information possible qui veulent chacun, il est nécessaire de catégoriser ces internautes.

L'analyse des catégories sociales peut devenir un domaine privilégié du lien entre les sciences de la cognition et les sciences sociales. Ici il y a un problème qui empêcher de faire une bonne catégorisation sociale, ce problème est la présence des données manquantes, donc il faut les imputer, ensuite les estimer.

Pour cela j'ai essayé de créer une application qui nous aide à faire classifier les identités sociales selon certain paramètres.

La classification professionnelle serait peut être construite à partir de processus de conceptualisation plus élémentaires des groupes sociaux.

C'est dans ce cadre que je fais mon modeste projet, ce mémoire est structuré en quatre chapitres :

- le premier chapitre présent les réseaux sociaux, comme leurs noms l'indiquent, les réseaux sociaux favorisent l'interaction entre personnes, ils ont fait leur apparition en mars 2003 sur internet avec le lancement du site Friendster, puis quelques mois plus tard, de MySpace aux Etats-Unis.
  Ainsi, dans ce chapitre j'ai présenté comment les internautes expriment différents aspects et de maintenir les diverses relations à l'intérieur du web social en fonction du contexte, avec une description de l'identité numérique, le profile, et les média sociaux. Parmi les travaux qui touchent les réseaux sociaux, il ya celle qui sert à rendre le réseautage social fédérée réel. Il s'agit de créer le « P2P du social », les plates-formes Web social fédérés, permettant aux utilisateurs de conserver leurs données où ils veulent même sur leur propre serveur tout en interagissant avec le reste du Web social.
- le deuxième chapitre concerne les données manquantes (aberrantes), ce sont des données pour lesquelles la valeur de certain attribut est inconnue. Suivi d'une classification de ces derniers tel qu'il existe trois types : Manquant complètement au hasard, Manquant au hasard, ne manquant pas au hasard.
  Le traitement des données manquantes a connu plusieurs techniques qui ont été développées dans les années 90, parmi ces techniques il y a les méthodes de suppression, de remplacement, et de modélisation. Pour analyser les données manquantes, il faut connait les informations sur ces données tel que le type, l'emplacement des valeurs manquante ; et après les estimer par la moyenne ou l'écart type, ou appliquer d'autre méthode d'estimation ; ensuite imputer les valeurs manquantes avec des valeurs estimées.

- Le troisième chapitre présente la catégorisation, elle consiste a affecté à chaque objet une classe existante, elle intervient dans la reconnaissance et l'identification des objets.

  La catégorisation est faite à l'aide des algorithmes de classification il y a : la classification Hiérarchique, Algorithmes par partitionnement, aussi que la classification par Modèles tel que chaque cluster est supposé suivre un modèle
- Le quatrième chapitre présente les étapes nécessaires pour la réalisation de l'application.

  Premièrement, Les données utilisées sont des informations de certains utilisateurs sur le réseau social Facebook stockées dans un fichier sous format **CSV,** collecté par Eugene Dubossarsky et Mark Norrie en 2004.

  Dans ce projet j'ai utilisé trois outils :
  - Weka : qui est un logiciel d'apprentissage automatique et d'exploration de données, ce dernier contient des méthodes non symboliques pour faire la catégorisation, pour cela j'ai appliqué trois méthodes de classification sur mon fichier (**Espérance-maximisation, Regroupement hiérarchique** et **SimpleKMeans)**
  - Langage R : R est un langage de programmation interactif interprété et orienté objet contenant une très large collection de méthodes statistiques, ce dernier à la capacité d'imputer les données manquantes a l'aide d'un package nommée « yaImpute ».

    Après cette imputation j'ai répété la catégorisation avec Weka pour comparer les résultats obtenue avec les premiers résultats.
  - NetBeans : est un environnement de développement intégré (EDI), permet de supporter différents langages, comme Python, C, C++, JavaScript, XML, PHP et HTML.

    Pour créer une application avec une interface facile a utilisé, il est préférable de choisir JAVA comme langage de programmation. J'ai choisi NetBeans pour que je puisse utiliser la programmation à base des règles (l'utilisation des règles XML).

    Alors j'ai utilisé deux bibliothèques JAVA : **JRuleEngine1.3** et **jsr94,** pour que je puisse créer un moteur d'inférence.
- Enfin, une conclusion pour comparer les résultats obtenue selon les méthodes symbolique (l'utilisation du JAVA) et les méthodes non symbolique (l'utilisation de WEKA).

  L'utilisation d'une interface qui affiche des résultats explicite (catégorie_1, catégorie_2, …) est plus claire que l'affichage des chiffres.

CHAPITRE 1 :
Réseaux sociaux

II. Chapitre 1 : Réseaux sociaux :

II.1 Introduction :

L'évolution d'Internet a permis la création de nouveaux outils de communication et de travail pour les entreprises ainsi que pour les particuliers.

Le Web 2.0 a amené l'évolution des réseaux sociaux. Ils sont de plus en plus utilisés par les internautes qui naviguent sur Internet, et ils touchent un public extrêmement large. Les étudiants et les adolescents ont été les premiers utilisateurs de ce genre de sites, faisant d'eux les précurseurs des réseaux sociaux actuels. Il faut également savoir qu'actuellement, pour beaucoup d'internautes, utiliser ces sites est considéré comme une activité sociale à part entière. Aujourd'hui, plus que 1.43 milliards d'utilisateurs des réseaux sociaux dans le monde en 2012 (19% de plus qu'en 2011), et on prévoit une croissance allant jusqu'à 1.85 milliards d'utilisateurs actifs au moins une fois par mois en 2014.

Les réseaux sociaux sont inspirés par la « théorie de 6 degrés » qui veut que chaque individu soit à six intermédiaires de toute autre personne sur la planète. Leur principe de fonctionnement s'inspire des sites de rencontre et les sites de retrouvailles de camarades de classe. Ainsi, ces réseaux permettent de mettre en relation des personnes entre elles et créent ce que l'on appelle des « graphes sociaux » qui représentent les liens qui s'établissent entre chaque individu d'une communauté.

II.2 Présentation :

Le Web social est un ensemble de relations qui lient les gens sur le Web. Alors que les sites de réseautage social actuels les plus connus sur le Web se limitent aux relations entre les personnes disposant de comptes sur un seul site, le Web social devrait permettre aux gens de créer des réseaux de relations à travers l'ensemble du Web, tout en donnant aux gens la possibilité de contrôler leur propre vie privée et des données.

Le Web social n'est pas seulement sur les relations, mais sur les applications et les innovations qui peuvent être construits au-dessus de ces relations. [1]

Les réseaux sociaux professionnels génèrent un maillage puissant et novateur dans l'organisation du travail. La connexion dans le monde virtuel est devenue un vecteur de création de valeur collective. [2]

Les réseaux sociaux ont complètement modifié la nature des interactions entre les marques et les consommateurs, ayant ainsi un impact direct sur son processus de décision actuel.

Plusieurs études tendent à montrer l'impact positif des réseaux sociaux sur une marque. En nous appuyant sur l'étude du Compete Institute (2011), 56% des personnes qui suivent une marque sur Twitter seraient plus enclins à acheter les produits de la marque. Pour ce qui est de Facebook, seulement 47% des « fans » de la page en question seraient prêts à acheter le produit, ce qui montre ici la plus solide relation créée entre la marque et les utilisateurs sur Twitter, Quand un consommateur a conforté son choix par l'intermédiaire d'un réseau social, il achètera le produit dans40% des cas.

Une autre étude a aussi montré que 56% des utilisateurs de Facebook qui sont devenus « fans » d'une marque sont plus enclins à la recommander ensuite auprès d'un ami. [3]

Une étude conduite par Socialbakers sur le marketing et les réseaux sociaux a été menée dans 82 pays, 20 secteurs d'activité et auprès de 500 professionnels du marketing.

Les entreprises interrogées étaient aussi bien petites que grandes et issues de multiples secteurs d'activité comme l'éducation, l'e-commerce, les organismes à buts non-lucratifs, les voyages etc. Et certains chiffres sont assez étonnants !

Et le grand gagnant des réseaux sociaux sur lesquels les entreprises vont mettre le paquet est Facebook (81,2 %) ! Suivi de Twitter (43,7 %) qui, en étant à la seconde place paraît quand même à la traîne, puis de YouTube (29,7%) à la troisième place et enfin Instagram (19.1%) [4]

II.3 Etat du web social en 2014 :

Le Web depuis sa création a été conçu pour inclure des connexions entre non seulement les documents hypertextes, mais les relations entre les gens.

Depuis les premiers jours du peuple Web qui ont maintenu leurs propres pages d'accueil ont enregistré des mises à jour sur les activités de leurs sites, ce qui a été poussé dans le grand public avec le développement de convivial logiciel de blogging, des innovations dans cet espace ont permis au grand public de plus en plus aptes à les blogs et les sites d'information indépendants comme Indymédia(1999) pionnier de la notion de gestion de contenu généré par l'utilisateur. Cependant, ces services sont restés assez expérimental jusqu'à après l'effondrement de la "dot-com" bulle initiale. Après cette éruption de sites de réseaux sociaux comme Friendster (2002), LinkedIn (2003), et Facebook (2004) ont décollé, et sont finalement devenus des sites les plus populaires sur le Web. À partir de Flickr (2004) et Youtube (2005), le contenu généré par l'utilisateur a pris au cours de cette nouvelle revigoré Web social. [5]

Qu'est ce qui a marqué **2013** sur le web ? Snapchat, par exemple, qui a décliné l'offre à 3 milliards de dollars de Facebook en novembre 2013. Snapchat, nouveau venu dans le milieu des réseaux sociaux, a bouleversé la donne. Au point qu'Instagram (propriété de Facebook) copie son système de messagerie.

Parmi les autres événements qui ont marqué l'année, la montée en puissance de Pinterest, scrapbook virtuel qui s'appuie sur le partage de photographies. Le réseau social s'est implanté en France et confirme l'intérêt grandissant pour les contenus visuels, qu'on avait déjà remarqué avec le succès d'Instagram et de Vine.

A quoi ressemblera **2014** ? On peut imaginer une résurrection de MySpace, lifté en juin dernier, penser que les artistes impliqueront davantage les fans dans leurs productions ou que Facebook servira davantage à suivre l'actualité qu'à s'enquérir de nouvelles de ses contacts. Dans la boule de cristal du "Nouvel Observateur", nous avons essayé de dégager quelques tendances sur les réseaux sociaux :

II.3.1 Le règne des contenus éphémères

Qu'on s'en serve ou pas, il faut reconnaître que Snapchat a révolutionné le monde des réseaux sociaux en se basant sur une seule idée : le contenu qui disparaît, au bout de 10 secondes maximum. Ainsi, Snapchat a rendu aux médias sociaux leur côté ludique et spontané - ce qu'on avait perdu depuis belle lurette sur Facebook, poussant probablement à l'exode des adolescents.

Meatspace s'est engouffré dans ce modèle de l'éphémère. Ce réseau social lancé il y a quelques mois par Jen Fong-Adwent, ingénieure chez Mozilla, propose un chat avec des messages limités à 250 caractères, qui disparaissent au bout de dix minutes.

Datant d'avril, Blink est une appli disponible sur iOS (système d'exploitation mobile développé par Apple) qui permet d'envoyer des contenus (photos, messages, sons) s'autodétruisant entre 1 seconde et 5 minutes après leur réception.

II.3.2 L'émergence de Google+

Après deux ans d'existence, Google+ veut devenir un grand des médias sociaux, même s'il doit pour cela y aller au forceps. Par exemple, il faut désormais un compte Google+ pour commenter les vidéos YouTube, ce qui a d'ailleurs fortement irrité les YouTubers.

De fait, l'outil gagne à être connu, principalement pour sa gestion des albums photo, avec de nombreuses fonctions de retouche et de création (automatique !) de gifs. Google a annoncé en octobre 300 millions d'utilisateurs actifs par mois, contre 190 millions en mai 2013.

II.3.3 Adieu SAV, bonjour Twitter

Les consommateurs ne s'embarrassent plus du téléphone pour se plaindre d'un produit. Selon une étude de Nielsen aux Etats-Unis en 2012, plus de la moitié des clients se tournent vers les réseaux sociaux pour une réclamation tandis que 81% d'entre eux attendent une réponse dans la journée.

Par le biais des médias sociaux, les gens ont un moyen tout trouvé de critiquer une marque et de déclencher un badbuzz. 2014 semble être le moment opportun pour que les entreprises renforcent leur service clients sur Twitter ou Facebook et appuient leur stratégie sur la conversation en temps réel.

II.3.4 De la publicité partout

La pub vidéo s'apprête à faire son entrée sur Facebook, directement dans le flux d'actualités et elle se lancera automatiquement. De quoi intéresser les annonceurs. Ils devraient aussi se saisir de la récente fonctionnalité de Twitter qui affiche directement les photos dans la timeline des internautes.

Google+ teste un service qui permet aux marques de se servir de ses contenus (photos, vidéos et même Hangouts) pour faire de la promotion. LinkedIn a lancé l'été dernier des postes sponsorisés. Les épingles sponsorisées sont aussi apparues chez Pinterest à la fin de l'année.

II.3.5 La mobile superstar

En 2014, le nombre de téléphones portables devrait dépasser la population mondiale. Les internautes vont de plus en plus surfer depuis leur smartphone. Aux entreprises donc de considérer la manière dont leur site est accessible : une version mobile est nécessaire ou un site en responsive design. Il faudra pousser les utilisateurs à agir vite, sans qu'ils aient à scroller ou cliquer. On peut imaginer que les sites proposent des méthodes de paiement adaptés au mobile ou encore des campagnes de publicité uniquement sur ce support. [6]

II.4 Cadres de Web sociaux :

II.4.1 Le problème de Walled Gardens

L'importance du Web a été limitée à l'hypertexte des pages web sans attention portée aux interactions sociales et les relations.

Cependant, ces types d'activités sont actuellement limités à certains sites de réseautage social, où l'identité d'un utilisateur et les données peut facilement être saisi, mais seulement accessibles et manipulées via des interfaces propriétaires, afin de créer un «mur» autour des raccords et des données personnelles, comme illustré dans l'image ci-dessous.

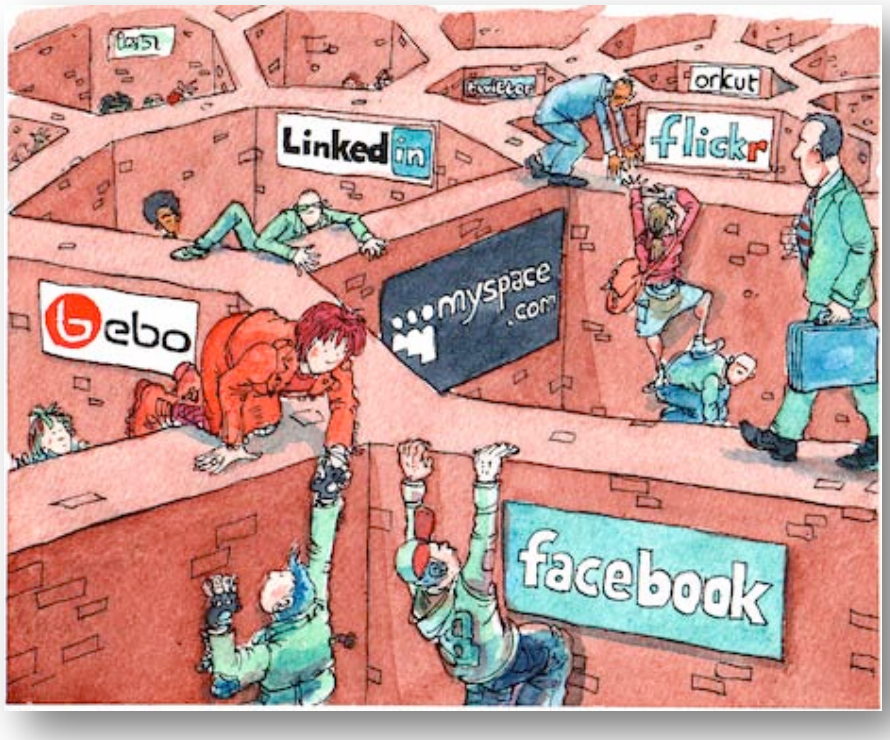


Figure 1: Problème de Walled Gardens [7]

Une architecture Web véritablement universel, ouvert et distribué social est nécessaire, L'absence d'une telle architecture profondément les impacts de l'expérience quotidienne du Web de nombreux utilisateurs. Il existe quatre grands problèmes rencontrés par l'utilisateur final :

Portabilité : Un utilisateur ordinaire ne peut pas télécharger leurs propres données et les partager comme il aime.

Identité : Chaque fois qu'un utilisateur accède à un nouveau site, ils doivent non seulement créer un nouveau nom d'utilisateur et mot de passe, mais retrouver leurs amis et inciter ses amis à déplacer des sites avec eux.

Linkability : Les utilisateurs n'ont aucun moyen d'être informé si elles sont mentionnées sur un site de réseautage social qui n'est pas membre.

Confidentialité : Un utilisateur ne peut pas contrôler la façon dont leurs informations sont consultées par d'autres dans des contextes différents par différentes applications sociales, même sur le même site de réseautage social. [7]

II.4.2 La Vision Web social

Les gens expriment différents aspects en fonction du contexte, se donnant ainsi plusieurs profils qui leur permettent de maintenir les diverses relations à l'intérieur et dans différents contextes : la famille, l'équipe sportive, l'environnement des affaires, et ainsi de suite. Même si, dans tous les contextes de certaines informations est généralement souhaité être gardé secret. Chez les personnes du «monde pré-Web» peut généralement soutenir cette multiplicité de profils comme ils sont physiquement limités à un petit ensemble de contextes sociaux et des possibilités d'interaction. À certains égards, la dynamique sociale sur le Web ressemble à ceux de l'extérieur sur le Web, mais les interactions sociales sur le Web diffèrent dans un certain nombre de points importants.

Toute personne doit être en mesure de créer et d'organiser un ou plusieurs profils différents en utilisant un site de réseautage social de confiance de choix, y compris l'hébergement de leur propre site sur lequel ils se sont lancés soit sur un serveur ou localement dans leur navigateur.

La vie privée est le contrôle de l'accessibilité de l'information sociale en général, y compris la sécurité comme un catalyseur (l'authentification de l'identité et de la propriété des données numériques). La vie privée doit être contrôlé par les utilisateurs eux-mêmes dans un contrat explicite avec les sites de réseautage social et des applications qui permet des contrôles de confidentialité faciles à utiliser et à comprendre. Comme dépositaire de leurs propres profils, les utilisateurs peuvent alors décider quelles applications social peuvent accéder aux détails de profil via exposer les données personnelles explicitement au fournisseur de l'application, et rétractant ainsi, à un niveau de granularité. [8]

II.4.3 La terminologie

Utilisateur : L'utilisateur est une personne, une organisation ou un autre agent qui participe dans les interactions sociales en ligne sur le Web.

Identité : Une représentation numérique unique d'un utilisateur. Ce sont potentiellement illimité et peuvent coïncider avec les différents personnages de l'utilisateur, tels que le profil personnel et le profil de travail. Il s'agit d'un «personne» dans le Lexique communes identité.

Attribut de Profil : Informations sur un utilisateur qui est un composant du profil tel que le nom, e-mail, statut, photo, téléphone professionnel, téléphone à la maison, adresse de blog…

Connexion social : Lien social sont des associations entre un profil et une ressource et peut inclure le type de la relation (par exemple, ami, collègue, etc.), La collecte de toutes les connexions d'un profil est appelé le **Social Graph** de ce profil.

Groupe social : Groupes sociaux sont des ensembles explicites nommés de liens sociaux entre les ressources. Par exemple, mon équipe de football, mes films préférés, etc.

Plates-formes sociales : Plates-formes sociales se réfèrent à un ensemble de fonctions dans lequel l'utilisateur peut interagir avec leurs liens sociaux et les médias sociaux, de publier les médias sociaux, et utiliser des applications sociales.

Social Graph Distribué : Un ensemble de profils et les liens sociaux entre les agents qui peuvent être hébergés sur différentes plateformes sociales

Applications sociales : Applications sociales sont des fonctions d'une plate-forme sociale telle que la messagerie en temps réel et les jeux sociaux. Applications sociales peuvent être liés à une plate-forme sociale particulière (Facebook et FBML (Facebook Markup Language), Twitter et Twitter OAuth (open protocol to allow secure authorization)) ou capable de fonctionner sur plusieurs plates-formes sociales (OpenSocial).

Profil Association : Une sorte de lien social. Un profil association est utilisé pour indiquer le lien entre un profil spécifique et une plate-forme sociale.

Interaction sociale : Une interaction sociale Liens utilisateur Web social et une plate-forme sociale en fournissant toutes les applications nécessaires et les informations de profil.

II.4.3.1 *Web social et profils utilisateur*

La figure ci-dessous montre comment un utilisateur unique (une personne) peut avoir plusieurs profils qui partagent des attributs communs. Un utilisateur peut alors associer son / ses profil au niveau du profil avec des applications particulières sociaux, leur contrôle en quelque sorte de vue agrégé que l'utilisateur peut avoir un accès aux applications de bureau via un agrégateur. Les profils sont exposés à et / ou synchronisés avec différentes plates-formes sociales. Dans la figure 2, un profil est associé à la "lumière bleue" et des applications sociales "rouges", un profil à la demande sociale "gris", et un profil à l'"bleu", "vert", et les applications sociales « orange».

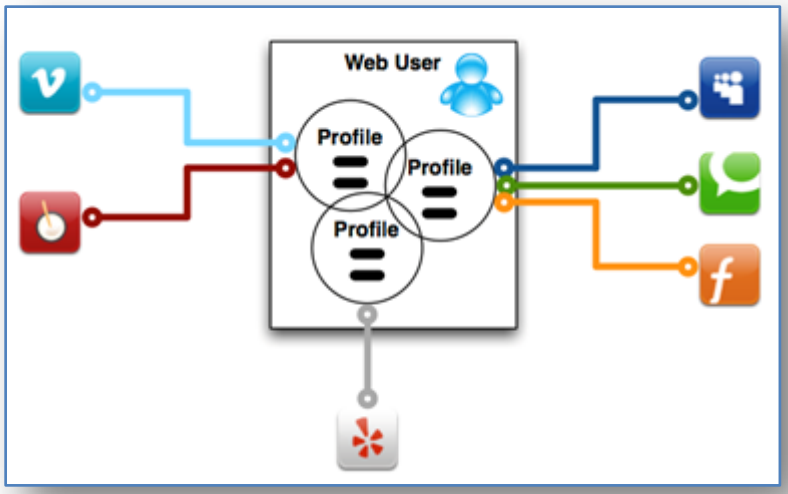


Figure 2:web social et profile d'utilisateur **[9]**

II.4.3.2 *Simple Graphe Sociale Distribuer*

Les attributs dans un profil, y compris les informations sur les connexions sociales, peuvent être distribués. Cela signifie que les attributs et les connexions concernées sociaux peuvent être stockés avec une application sociale pour une utilisation dans le contexte de cette application.

La figure 3 ci-dessous présente un profil qui comporte deux jeux de deux attributs dans des sites répartis chacune avec deux attributs locaux. L'utilisateur interagit avec le profil à travers la plate-forme sociale "bleu", qui pourrait être un nœud dans une plate-forme Web social décentralisée. Par exemple, un service de gestion de profil qui pourrait être couru dans le navigateur ou via un site Web tiers serait de garder trace des attributs distribués et plusieurs profils et permettre à l'utilisateur de modifier les attributs sur de multiples plateformes.

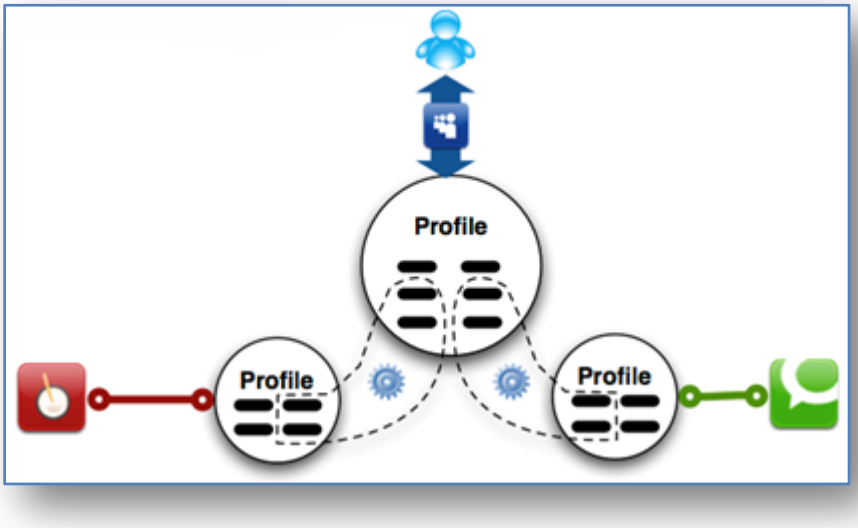


Figure 3: Graphe Sociale Distribuer **[9]**

II.4.3.3 *Multiples graphique sociale Distribuer*

Un profil est associé à un ou plusieurs plates-formes sociales dans lesquelles le graphe social de l'utilisateur est formé et nourri. La plate-forme sociale est le contexte de la façon dont un utilisateur est connecté aux profils des autres et soutiendra les types de connexion spécifiques (par exemple : ami, collègue, etc.) qui serviront généralement l'objet d'une certaine demande sociale. Une caractéristique de base ou d'un service d'une demande sociale est de faire, de maintenir et développer ces connexions.

Les connexions d'un utilisateur dans une plate-forme sociale particulière doivent être portables. L'utilisateur doit être en mesure de les prendre à l'autre plate-forme sociale, de sorte qu'il n'est pas nécessaire de rétablir toutes les connexions à nouveau dans une autre (nouveau) demande sociale. Notez qu'Amy (profil 1) dans la plate-forme sociale est reliée à deux reprises à Bob via son profil 1 et 2. Cela démontre que le même utilisateur peut se connecter via différentes plates-formes sociales qui pourraient avoir des liens à travers le Web ouvert. Les lignes entre les profils sont soit unidirectionnelle (suivant) comme **Twitter** ou bidirectionnelle (amitié) comme **Facebook**. [9]

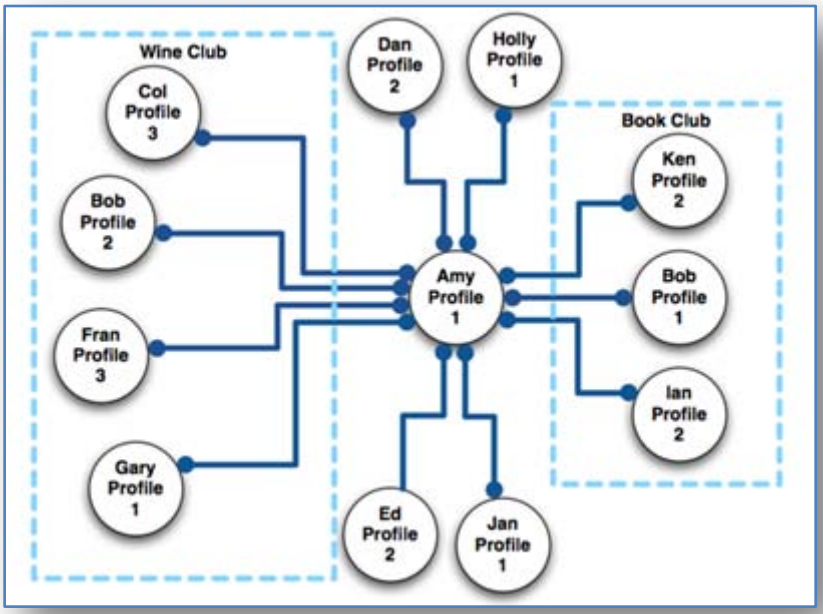


Figure 4: Multiples graphique sociale Distribuer **[9]**

II.5 Identité

L'identité est la liaison entre un profilé, un ensemble d'attributs, et un utilisateur. Certains titres de compétences ou la «preuve» de l'identité peuvent être nécessaires à l'utilisateur pour accéder ou créer un profil, qui est l'étape de **l'authentification**. En particulier, ces pouvoirs peuvent prendre de nombreuses formes comme un mot de passe, un certificat numérique signé, ou d'autres informations d'identification login.

II.5.1 Problème : Les noms d'utilisateur et mots de passe ne sont pas sûrs

Nom d'utilisateur et mot de passe combinaisons sont actuellement la technologie d'identification la plus répandue sur le Web. Ils sont faciles à comprendre, mais souffrent d'un certain nombre de désavantages techniques et économiques, y compris les menaces de phishing. Les internautes sont invités à créer des comptes soutenus par trop de mot de passe sur plusieurs sites Web, ce qui conduit à un mot de passe réutilisé à l'insécurité croissante de chaque compte. Les mots de passe qui sont générés manuellement sont souvent précaires, et celles générées automatiquement sont difficiles à retenir.

II.5.2 Normes d'identité

Cette section énumère un certain nombre de fournisseurs d'identité en ligne, ainsi quelque norme d'identité et l'authentification.

II.5.2.1 *Gestion de Mot de passe basée sur le navigateur :*

Il est maintenant plus facile pour les utilisateurs de créer des mots de passe différents pour chaque site par les rappelant à l'utilisateur, comme c'est actuellement mis en œuvre par Mozilla. Le projet Weave de Mozilla vise à faire de l'authentification par mot de passe plus intégrée dans le navigateur en permettant le navigateur pour créer et mettre à jour les mots de passe automatiquement sur le Web.

L'utilisateur final ne doit alors se souvenir de cette URL et le mot de passe pour l'un de ses contenus, pour être en mesure de le récupérer dans n'importe quel autre dispositif qui sait décrypter et lire le contenu.

II.5.2.2 *OpenID*

OpenID centralise l'étape de l'authentification à un fournisseur d'identité, de sorte que l'utilisateur peut s'identifier à un site (un fournisseur d'identité **OpenID**) et de partager leurs données de profil avec un autre site, la partie de confiance. Un utilisateur n'a qu'à me souviens d'une identité unique au monde, qui était un URI dans OpenID 1.0.

Une fois le fournisseur OpenID est découvert, un secret partagé est établi entre le fournisseur et la partie se fiant, leur permettant de partager des données. Principalement cela se fait via un protocole d'échange d'attribut **Attr OpenID**, qui permet à l'utilisateur de spécifier quelles données personnelles doivent être envoyées à la partie utilisatrice.

II.5.2.3 *WebID*

WebID, utilise TLS (Transport Layer Security) et les certificats côté client pour l'identification et Authentification. Pour authentifier un utilisateur demandant une ressource à accès contrôlé via HTTPS (HyperText Transfer Protocol Secure), l'agent vérifié le contrôle de la ressource doit demander un certificat X.509 du client. L'intérieur de ce certificat, en plus de la clé publique il y a un champ "Subject Alternative Name" qui contient un URI *(*Uniform Resource Identifier) identifiant l'utilisateur (le **WebID**). L'utilisation d'authentification mutuelle norme TLS, l'agent utilisateur confirme qu'il connaisse la clé privée correspondant à la clé publique du certificat. Une simple recherche de HTTPS cacheable sur le **WebID** doit récupérer un profil.

II.5.2.4 *XAuth*

XAuth (Extra Authentication) Permet plusieurs fournisseurs d'identités de mettre à jour un fournisseur de **XAuth** (actuellement seulement **Xauth.org**) afin que les tiers peuvent authentifier l'identité d'un utilisateur donné. Quand un utilisateur se connecte-ON pour un compte sur un fournisseur d'identité **XAuth** activé, le fournisseur d'identité avise **xauth.org**. Quand un site est rencontrée qui nécessite une authentification, le site peut utiliser quelques simples JavaScript intégré pour demander **xauth.org** où l'identité des fournisseurs l'utilisateur est connecté sur, puis utilise les cookies stockés localement sur le navigateur pour aider l'utilisateur à authentifier avec la troisième partie site.

II.5.2.5 *SAML*

SAML (Security Assertion MarkupLanguage) est une norme OASIS (Organization for the Advancement of Structured Information Standards) pour l'échange d'authentification et de confidentialité entre les fournisseurs d'identité et les fournisseurs de services qui utilisent un format de données basé sur XML, d'aborder le problème de la connexion unique parmi beaucoup d'autres. [10]

II.6 Profile

Le cadre de profil contient des applications qui peuvent être utilisés pour accéder à des attributs et l'accès à de telles informations distribuées. Les utilisateurs de cette étape devraient également être en mesure de trouver, découvrir, ajouter et supprimer des connexions afin de mettre à jour leur profil

II.6.1 Problème : Non Décrivez-vous

Aujourd'hui, lorsque les utilisateurs créer des profils, ils sont souvent contraints dans leur façon de se décrire et d'avoir à nouveau manuellement retrouvé leurs amis. Pire, certains sites de réseautage social limitent les préférences, telles que le sexe et les préférences de la religion, qui peut être très sensible.

II.6.2 Normes de profil

Un certain nombre de normes existent pour le profil et le rapport sur le Web. Une distinction entre eux est le format de données (texte en clair, XML, RDFa (Resource Description Framework dans des Attributs)), et si elles sont facilement extensible ou non. Plus important encore, il existe des différences dans la façon dont comment tenu de l'identité numérique, toute application particulière peut alors essayer de découvrir et d'accéder aux données de profil et d'autres capacités que l'identité numérique peut mettre en œuvre. Alors que certains profils mentionnent ces techniques de découverte et d'utilisation explicitement et d'autres pas, ces techniques de découverte commune ou normalisés seront mentionnées dans le contexte de chaque format de données de profil.

II.6.2.1 *XRD*

XRD (Extensible Description des ressources), anciennement Yadis et DRX-simple (XRD-S), est un format de fichier XML pour découvrir ce quel capacités d’un fournisseur de profil particulier peut avoir.

II.6.2.2 *VCard*

Le standard de l'IETF (Internet Engineering Task Force)**VCard** est le format le plus ancien et le plus répandu pour les données du carnet d'adresses personnel, le type d'information trouve habituellement sur une carte d'affaires, telles que le nom, l'adresse. Par conséquent, ce format sert en général comme le tronc commun de la plupart des données-formats. L'importation et l'exportation de **VCard** est pris en charge par la plupart des programmes de messagerie comme Thunderbird, Microsoft Exchange, et Apple Mail.

II.6.2.3 *FOAF*

Le premier projet utilisées les normes pour décrire les réseaux sociaux décentralisés distribués, a été le projet **FOAF** (friend of a friend, ami-de-un-ami). FOAF cependant que tente de relever les défis descriptifs, plutôt que l'espace de l'ensemble des problèmes. **FOAF** fournit une approche extensible et ouverte à la modélisation des informations sur des personnes, des groupes, des organisations et des entités associées, et est conçu pour être utilisé avec d'autres vocabulaires descriptifs.

FOAF lui-même ne prévoit pas de fonctionnalité "réseau social". Il suppose d'autres outils et techniques sera utilisé à côté de lui, et ne se précise pas les mécanismes d'authentification, syndication ou mise à jour. Aujourd'hui, la grande majorité des données exprimées en **FOAF** est exporté à partir de grands sites "de réseau social".

II.6.2.4 *PortableContacts*

Un profil standard plus en plus populaire est **PortableContacts**, qui est dérivé de **VCard**, et est sérialisé au format XML ou, plus couramment, JSON (*JavaScript Object Notation*). Il contient une grande quantité d'attributs de profil, telles que la propriété "relationshipStatus".

Plus d'une norme de profil, le profil régime **PortableContacts** est conçu pour donner aux utilisateurs un moyen sécurisé pour permettre aux applications d'accéder à leurs contacts, en fonction de XRDS pour la découverte de **PortableContacts** points terminaux et **OAuth** pour l'autorisation déléguée. Il fournit un modèle d'accès commun et système de contact ainsi que les exigences authentification et d'autorisation d'accès aux informations de contact privé. [11]

II.7 Médias sociaux

Le Web social n'est pas seulement les liens entre les gens, mais les liens entre les personnes et les ressources arbitraires, y compris les messages comme les messages de blog, audio, photos, vidéos, et autres ressources. Donc, les médias sociaux sont des ressources qui sont utilisées dans une relation sociale avec un utilisateur. Un utilisateur doit également être capable d'avoir des liens vers des ressources "non-Web" comme des endroits et des objets.

II.7.1 Problème : Amende pour consommation de médias sociaux

Les utilisateurs de plus en plus et les plateformes sociales se retrouvent consommation des médias sociaux, mais ne sachant pas si elles sont digne de confiance ou si oui ou non ils peuvent consommer tels médias sociaux sans une amende, à savoir si leur utilisation se casse le droit d'auteur du contenu !

Ne pas connaître cette information peut conduire à la catastrophe. Les gens qui sont souvent téléchargent et réutilisant les médias sociaux peuvent maintenant être condamnés à une amende d'énormes sommes d'argent, mais beaucoup d'entre eux ne sont pas conscients que les données étaient sous copyright en premier lieu. Ainsi de nombreux utilisateurs aimeraient avoir des mécanismes pour déterminer automatiquement si un document Web ou ressources peuvent être utilisées, en fonction de la source originale du contenu, les informations de licence associé à la ressource, et des restrictions d'utilisation sur ce contenu.

II.7.2 Normes de médias sociaux

II.7.2.1 *Teneur*

SIOC (SemanticallyInterlinked Online Communities) vise à développer un vocabulaire standard pour représenter le contenu généré par les utilisateurs sur le Web, en utilisant les technologies du Web sémantique. L'ontologie SIOC (une présentation de membre du W3C, toujours en évolution) consiste en un vocabulaire de base (avec des classes tels que SIOC : UserAccount et SIOC : article) et plusieurs modules.

II.7.2.2 *Microformats*

Les microformats sont un moyen simple d'intégrer la sémantique au format HTML ordinaire en réutilisant l'établissement des attributs HTML tels que 'rel', 'classe', et 'rev' avec une chaîne de valeurs donnée définition par un certain nombre de vocabulaires. Ces vocabulaires sont destinés à normaliser l'information commune (comme l'information de contact hCard) sur le Web social.

II.7.2.3 *Open Graph Protocol*

Facebook Open Graph Protocol est un vocabulaire de métadonnées pour décrire les documents et indirectement leurs sujets. Il est généralement publié en feuilleton dans RDFa en éléments <meta> dans les pages HTML. Une application d'Open Graph Protocol est le Javascript pour le bouton "Like" de Facebook. Cela permet aux développeurs d'ajouter un bouton «Like» à un élément décrit dans une page Web en ajoutant seulement une petite quantité de RDFa simplifiée à l'en-tête d'un site web. [12]

II.8 Confidentialité

Les politiques **centrée** sur la vie privée sont des règles qui peuvent capter les autorisations (contrôle d'accès), obligations (comme termes de service et de licences) et d'autres paramètres de traitement des données qui permettent à un utilisateur de contrôler leurs interactions avec les médias sociaux et d'autres utilisateurs. Les politiques appliquent les paramètres de confidentialité du profil et des cadres de médias sociaux pour gérer de manière cohérente les attentes des utilisateurs de la vie privée et d'autres obligations. Une plate-forme sociale qui gère la vie privée au nom d'un utilisateur sur plusieurs applications sociales et d'autres plates-formes est un **fournisseur de la vie privée.**

II.8.1 Problème : Violation de la vie privée

Les gens sont de plus en plus à trouver leur propagation des médias sociaux à travers de multiples plates-formes et accessibles par toutes sortes de gens, dont beaucoup ils n'ont pas l'intention à l'origine. Comme les médias sociaux est au centre de tout, de l'emploi, le recrutement, de relations personnelles, la possibilité d'accorder et de restreindre l'accès aux données personnelles de l'un devient une composante essentielle de nombreuses applications sociales. En outre, comme les médias sociaux sont partagés et regroupées dans différents sites, ce problème devient encore plus critique que généralement ce contrôle d'accès n'est pas "collant", c'est à dire après les données partout où il va.

II.8.2 Confidentialité et protection des normes

Les nouvelles technologies, l'ubiquité de l'Internet, et la quantité de temps que les gens passent interagir avec le monde numérique sont à la fois la promotion de nos libertés, alors que dans le même temps permettent aux nouveaux invasions de la vie privée.

Le **SWXG** (Incubateur Groupe Web social) a examiné un certain nombre de technologies / initiatives qui peuvent fournir un aperçu des méthodes de développement des langues politiques lisibles à la machine. Ces politiques lisibles à la machine aideraient permettent aux utilisateurs de définir des politiques sur leurs données, en indiquant comment ils ont l'intention d'avoir leurs données utilisés et partagés sur le Web social. Ces politiques lisibles par machine doivent être marée jusqu'à l'identité en ligne des utilisateurs. Voici quelques-unes des initiatives qui pourraient aider à éclairer la conception des politiques centrés sur l'utilisateur à la vie privée et le partage de données.

II.8.2.1 *P3P*

Exprimant la vie privée via des langues lisible à la machine, **P3P** (Platform for PrivacyPreferences) Recommandation a commencé avec le W3C, permet aux opérateurs de sites Web d'exprimer leur collecte de données, l'utilisation, le partage, et les pratiques de conservation dans un format lisible par machine.

II.8.2.2 *POWDER*

Le W3C **POWDER** (protocole pour la Description des Ressources du Web) langage fournit un mécanisme pour décrire des groupes de ressources en fournissant essentiellement un "globe" opérateur sur les URI et reliant ces groupes d'URI à un groupe de déclarations XML communes sur des sujets tels que l'authentification et RDF états. Bien plus générique que **P3P**, il visait les mêmes cas d'utilisation tels que les descriptions de la vie privée pour la protection des enfants.

II.8.2.3 *AIR*

Malgré le manque de déploiement de P3P, la recherche continue sur les langues pour exprimer les politiques de la vie privée et le traitement des données. **AIR** (AMORD In RDF) est une politique de la langue qui est représentée en tortue et dispose d'un niveau de preuve de base, ainsi que spéciale -usage des classes et des propriétés qui peuvent être utilisés pour définir des politiques de façon lisible à la machine. Cependant, l'AIR est limité par sa capacité à ne traiter les données RDF et dispose pas de mappage défini à FRR (Fonds de réserve pour les Retraites).

II.8.2.4 *Icônes de confidentialité de Mozilla*

Icônes de confidentialité de Mozilla prend une approche simple basée sur des icônes inspiré Créative Commons. Au lieu de spécifier chaque type possible la vie privée et le traitement des données de scénario, ils précisent que quelques scénarios de confidentialité communes que les utilisateurs peuvent rencontrer.

Les icônes sont conçus pour être faciles à utiliser et à comprendre par les utilisateurs finaux ordinaires. Comme il s'agit d'un pas d'incitation pour les sites qui violent la vie privée de l'utilisateur de se qualifier en tant que tel, il serait au navigateur d'étiqueter automatiquement ces sites. En outre, les utilisateurs ne remarquent pas habituellement une icône en son absence, mais seulement par sa présence, le navigateur utilise automatiquement l'icône à informer les utilisateurs qu'ils ont conclus un site où leur vie privée pourrait être violée. [13]

II.9 Activité :

La caractéristique la plus distinctive du Web social sur l'hypertexte Web précédente est la focalisation croissante sur le partage d'informations en temps réel. Contrairement à tirer des informations sur une base ponctuelle, les utilisateurs désirent avoir des informations qui peuvent être d'intérêt poussé pour les immédiatement. Les interactions sociales de l'utilisateur et des ressources, y compris d'autres utilisateurs, sont les **activités** de l'utilisateur. Chaque activité, comme le changement de statut, établir de nouvelles connexions, Création d'un billet de blog, et assister à des événements peut être considéré comme une mise à jour dans une activité, le total de toutes les activités d'un utilisateur est le courant de l'utilisateur. Les médias sociaux, comme un blog classique, peut avoir son propre flux d'activités telles que les commentaires, microblogs, les étiquettes et notes.

II.9.1 Problème : ne peut pas intégrer les conversations

Actuellement, les utilisateurs sont obligés non seulement de «silo» leurs informations de profil et les médias sociaux, mais aussi toute la mise à jour de temps sensible de cette information. Comme de plus en plus mises à jour, (allant de changements d'emplacement pour les commentaires de blog à "aimer") les médias sociaux sont distribués à travers de multiples plates-formes sociales et l'information est fragmentée sur le Web. Il n'y a pas de méthode standard pour mettre à jour et réintégrer d'autres commentaires attachés à une mise à jour de leur source d'origine.

II.9.2 Normes d'activité

On notera la capacité de la messagerie du réseau social à mettre en œuvre, à la fois asynchrones et en temps quasi réel, à être coordonnée par Atom, PubSubHubbub, et XMPP (Extensible Messaging and Presence Protocol), accordant une attention particulière aux mises à jour de flux d'activité.

Contrairement à paysage fracturé des profils portables, les normes utilisées pour décrire les activités sont à ce point nouveau et rapidement déployées. L'architecture de base suppose une capacité d'envoyer du contenu (des mises à jour de statut, les messages et autre contenu) en plus près en temps réel que possible.

Ceci est actuellement réalisé par deux architectures distinctes. La première basée sur **XMPP**. La deuxième architecture est basée sur **HTTP**, mais la substitution de son architecture traditionnelle **"pull"** avec une architecture sur la base de PubSubHubbub**"push".**

II.9.2.1 *XMPP*

XMPP (Extensible Messaging and Presence Protocol) est une RFC (requests for comments) de l'IETF pour le transfert en temps quasi réel de données XML.
XMPP dans sa forme la plus simple peut être considérée comme un protocole pour le passage des fragments XML entre les machines, mais dispose de sa propre méthode pour l'authentification de l'identité et de l'extensibilité.

II.9.2.2 *ActivityStreams*

ActivityStreams est une sérialisation Atom pour les flux d'activité tels que les mises à jour de statut sur des sites populaires de réseautage social. Bien que Atom est facile de travailler avec, il ne tient pas compte de la sémantique de l'activité d'origine de façon multiplateformes.

ActivityStreams standardise la façon d'incorporer la sémantique de mise à jour de statut en divisant l'activité dans une action qui a été effectuée (verbe) par un acteur sur une autre personne, lieu ou une chose (l'objet). Un objectif supplémentaire (comme un album photo) pourrait être impliqué.

II.9.2.3 *Protocole Salmon*

Comme le contenu commence à se déplacer à l'extérieur de sa plate-forme sociale originale. Comment les commentaires, notes, et les annotations qui se produisent sur une autre plate-forme sociale en plus de l'original en quelque sorte être renvoyés dans le message original ? Le projet de protocole Salmon aborde ce problème de " l'unification des conversations". Il suppose qu'il y aura du spam, mais utilise des signatures numériques pour garantir le contenu provient d'une identité légitime, de sorte que tout le contenu dont l'identité du créateur n'est pas authentifié disparaît tout simplement.

II.9.3 Cadres émergents :

De la plus haute importance est le fait que tout cadre devrait conduire à un ensemble de fonctionnalités de base qui permet aux développeurs interagissent facilement leurs technologies existantes tout en encourageant de nouvelles utilisations et donc conduit l'innovation plutôt que la retenant par l'optimisation prématurée. Le cadre que le SWXG propose également est modulaire, de sorte que de nouvelles applications et des cadres sociaux émergents peuvent être ajoutés.

Il est possible d'envisager un cadre d'analyse qui permet aux utilisateurs de bénéficier de la participation de l'application sociale active, en fournissant l'analyse dynamique du comportement des utilisateurs et de nourrir ce retour dans le profil de l'utilisateur via la création automatique et la mise à jour des informations de profil d'un utilisateur sur la base d'une analyse de leur activité.[14]

II.10 Préoccupations d'accessibilité

Accessibilité concerne recoupe tous les aspects du Web social. En ce qui concerne l'identité, d'avoir l'utilisateur soit en mesure de préciser leurs besoins en matière d'accessibilité dans le cadre de l'identification et de l'authentification auprès d'un fournisseur d'identité est extrêmement préoccupante. Dans le monde des médias sociaux, des outils de création devraient favoriser l'accessibilité.

Des questions spécifiques d'accessibilité des interfaces utilisateur Web social en général sont traités par ARIA **(**Accessible Rich Internet Applications) en HTML, de sorte que le W3C devrait encourager un recours de ARIA par existantes des sites de réseautage social.

De nombreux sites sociaux fournissent des API (Application Programming Interface), en plus de l'interface Web principal. Cela conduit à une possibilité de créer des interfaces substituts accessibles, mais dépend de l'API d'exposer toutes les fonctionnalités. Beaucoup d'utilisateurs et les communautés allaient travailler sur la création de ces produits si les API nécessaires à l'information. [15]

II.11 Projets de réseaux sociaux décentralisés

2010 a vu un grand nombre de travaux entrepris à rendre le réseautage social fédérée réel. Pour décrire plus en détail, afin de surmonter la nécessité pour les utilisateurs de remettre leurs données à un site de réseautage social tiers, un certain nombre de projets de codage concrètes ont commencé à construire des plates-formes Web social fédérés, qui permettent aux utilisateurs de gérer leur propre prestataire sociale Web, permettant aux utilisateurs de conserver leurs données où ils veulent même sur leur propre serveur tout en interagissant avec le reste du Web social.

Voici quelque projets actuellement en développement pour un Web social fédéré :

II.11.1 Status.net

Status.net est une plate-forme de micro-blogging de logiciel libre pour aider les gens dans une communauté, entreprise ou du groupe d'échanger des messages court (140 caractères par défaut) sur le Web. Les utilisateurs peuvent choisir les gens à "suivre" et ne recevoir que leurs amis ou collègues des messages d'état.

II.11.2 OneSocialWeb

Vodafone **OneSocialWeb** open source plate-forme web social décentralisée fédéré construit sur **XMPP**, OneSocialWeb a un plug-in Java pour les serveurs Web, les clients, et une application Android.

II.11.3 Projet de Higgin

Eclipse projet de **Higgin** est l'un des premières open-source tentatives de créer un réseau social décentralisé. Il est basé sur le modèle personnel du magasin de données et ses propres RDF / OWL (Web Ontology Language) Persona Data Model. Il comprend également un soutien pour les clients actifs et InfoCard OASIS IMI **(IdentityMetasystemInteroperability)** pour traiter des questions liées à l'approvisionnement de l'identité, des identités multiples, de multiples personnages, et plusieurs niveaux d'assurance.

II.11.4 Espaces de données OpenLink

Espaces de données OpenLink (ODS « Open Document Spreadsheet »*)* est un projet open source sur le serveur **OpenLink Virtuoso** avec plusieurs applications subsidiaires centré sur l'utilisateur prédéfinis. En plus d'OpenID et webID, Il supporte les technologies sémantiques Web, variantes Atom, oData et GData (communication via "pingback sémantique"). L'accent est mis sur la virtualisation et ACL (Access Control List) espace de données pour le stockage Web. [16]

II.12 Considérations commerciales

Le but d'un Web social distribué et décentralisé n'est pas de proposer ou promouvoir des solutions qui réduisent ou minent les entreprises existantes et viables. Il cherche à explorer la mise en place d'une toute nouvelle architecture du Web social que des entreprises nouvelles et existantes peuvent tirer profit dans l'avenir. [17]

II.13 Taille des réseaux sociaux

2,5 milliards d'internautes à travers le monde dont 1,9 milliard présents sur les réseaux sociaux. 68% des français sont présents sur les médias sociaux : 30 millions sur Facebook (dont 15 millions qui se connectent tous les jours). Les français utilisent les réseaux sociaux 1h30 par jour en moyenne.

Les principaux réseaux sociaux restent :

Facebook : 1,2 milliard d'utilisateurs (+20% par rapport à 2013)
Twitter : 900 millions de comptes dont 250 millions d'utilisateurs actifs
LinkedIn : 250 millions d'utilisateurs (+25% en un an)
Instagram : 150 millions d'utilisateurs (+50% en un an)

D'autres médias sociaux connaissent une croissance fulgurante :

Tumblr : 185 millions de visites par mois pour 112 millions de blogs
Pinterest : 70 millions d'utilisateurs (dont 80% de femmes) et une croissance de 300% en 6 mois
Google + : 300 millions d'utilisateurs actifs (+220% en un an) mais seulement 6 minutes en moyenne par mois contre 6h30 pour Facebook. [18]

II.14 Modèles d'affaires actuels de réseaux sociaux

Certains des plus grands réseaux sociaux aujourd'hui seulement utilisent la publicité pour générer des revenus. D'autres ont seulement des frais de service haut de gamme d'utilisateurs. Certaines entreprises combinent deux ou trois modèles d'affaires.

Facebook continue d'offrir son service sans frais pour les utilisateurs finaux et générer ses revenus entièrement sur la publicité, bien que les transactions dans la plate-forme sont activées et de générer des revenus de plus en plus à l'avenir. [19]

II.15 Les nouveaux modèles d'affaires

Dans l'avenir, les revenus peuvent être générés à partir du Web social par :

- De nouveaux types d'applications sociales qui permettent de nouvelles formes de collaboration doit avoir lieu que profiter de données sociales-mash-ups de multiples services.
- Les ventes de logiciels, de la maintenance ou de réglage de nouvelles applications Web sociales identité-aware et sécurité renforcée qui offrent de solides garanties aux utilisateurs.
- Un marché plus flexible et ouvert à faible coût des médias sociaux payés par les paiements sur le Web
- Une liquidité accrue des données sociales avec de fortes garanties d'un manque de responsabilité juridique.
- Intégrant des fonctionnalités sociales dans tous les aspects de l'informatique et des applications existantes en général. [20]

II.16 Conclusion :

Les réseaux sociaux se sont développés fortement ces dernières années. Nous avons assisté à une forte expansion de leur nombre mais aussi de leur type. Maintenant, chaque internaute peut, en théorie, trouver un réseau social qui lui correspond, qu'il soit à caractère général, thématique ou professionnel. Le développement rapide de ce phénomène a amené les entreprises à se demander si elles devaient participer à ce phénomène et si oui de quelle manière ?

Les réseaux sociaux sont encore peu utilisés par les compagnies, même si de manière générale, celles-ci prévoient de plus en plus de les utiliser dans leurs stratégies futures.

CHAPITRE 2 :
Données manquantes

III. Chapitre 2: Données manquantes :

III.1 Introduction :

Les observations ayant des valeurs manquantes représentent un défi important car les procédures de modélisation classiques éliminent tout simplement ces observations des analyses. Lorsque les valeurs manquantes sont peu nombreuses (très approximativement, moins de 5% du nombre total d'observations) et que ces valeurs peuvent être considérées comme aléatoirement manquantes, c'est-à-dire qu'une valeur manquante ne dépend pas des autres valeurs, alors la méthode traditionnelle d'élimination est relativement "sûre". L'option Valeurs manquantes peut vous aider à déterminer si l'élimination est suffisante et vous proposer des méthodes de traitement des valeurs manquantes lorsqu'elle ne suffit pas.

Pour éviter de supprimer ainsi les données, on peut remplacer une valeur manquante par la moyenne de la variable correspondante, mais cette moyenne peut être une très mauvaise approximation dans le cas où la variable présente une grande dispersion.

III.2 Données manquantes

Une donnée incomplète est une donnée pour laquelle la valeur de certain attribut est inconnue, parce qu'elles n'ont pas pu être observées ; elles ont été perdues ou elles n'étaient pas enregistrées. Ces valeurs sont dites manquantes.

Soit l'observation est un vecteur des valeurs de certains indicateurs ou attributs, les valeurs manquantes peuvent être de deux natures :

- Valeur manquante totale, c'est- à-dire que toute l'observation manque.
- Valeur manquante partielle, c'est-à-dire que l'observation est présente mais il manque certaines valeurs de cette observation. [21]

Exemple : Dans ce tableau, la valeur de l'attribut (a4) pour l'observation (w2) est manquante, la valeur n'est pas présentée et l'observation (w2) est dite incomplète.

Tableau 1: Valeur manquante et donnée incomplète

	Attributs			
observations	**A1**	**A2**	**A3**	**A4**
W1	56	98	1O	5
W2	40	87	21	?

Des valeurs manquent parce qu'elles n'ont pas pu être observées ; elles ont été perdues ou elles n'étaient pas enregistrées.

III.3 Classification des Données Manquantes

III.3.1 MCAR :

« Manquant complètement au hasard » La probabilité qu'une observation soit incomplète est une constante, i.e le fait de ne pas avoir la valeur pour une variable Xi est indépendant des autres variables $X_{\neq ij}$

Exemple : X1= âge ; X2= sexe ; X3= glycémie. La probabilité que l'âge soit une donnée Manquante ne dépend ni du sexe, ni des valeurs de glycémie, elle est la même pour tous les sujets

III.3.2 MAR :

« Manquant au hasard » La probabilité qu'une observation soit incomplète ne dépend que de valeurs observées (pas de valeurs manquantes), i.e le fait de ne pas avoir la valeur pour une variable Xi est dépendant d'une autre ou d'autres variables X j≠i observées
Exemple : X1= âge ; X2= sexe ; X3= glycémie, la probabilité que l'âge soit une donnée Manquante ne dépend que du sexe et des valeurs de la glycémie (valeurs observées), elle n'est pas la même pour tous les sujets

III.3.3 NMAR :

« ne manquant pas au hasard » (informative) La probabilité qu'une observation soit incomplète dépend de valeurs non observées, elle n'est pas aléatoire, i.e le fait de ne pas avoir la valeur pour une variable Xi observée est dépendant d'une autre ou d'autres valeurs non observées des variables X j≠i observée
Exemple : X1= âge ; X2= sexe ; X3= glycémie, la probabilité que l'âge soit une donnée Manquante dépend des valeurs manquantes pour le sexe et la glycémie (valeurs non observées), elle n'est pas la même pour tous les sujets. [22]

III.4 Conséquences de la présence de données manquantes

La présence de données manquantes peut avoir un impact à différents niveaux, au niveau de la validité interne, de la validité des associations entre variables, au niveau de la généralisation d'une association entre deux variables.
Les données manquantes peuvent menacer la validité interne à différents moments du processus de recherche : au moment de la sélection de l'échantillon, de l'assignation aléatoire au groupe expérimental ou au groupe contrôle, de la collecte (non-réponse complète ou partielle, attrition) et de l'analyse statistique des données.[23]

III.5 Revue de littérature des méthodes de traitement des données manquantes

De nombreuses techniques de traitement des données manquantes ont été développées dans les années 90, Hu et *al.* (2000), sans prétendre être exhaustifs, en identifiaient déjà plus d'une vingtaine, pour la plupart issues des recherches en statistique. Depuis, les chercheurs en intelligence artificielle et fouille de données (Data mining), se sont mis à étudier la question et à développer de nouvelles techniques. Recenser l'ensemble de ces techniques serait fastidieux.

Aussi avons-nous opté pour une mise en évidence des principales caractéristiques des différentes méthodes. Nous pouvons alors présenter les techniques les plus usitées et avoir ainsi une vue d'ensemble du domaine, ces méthodes s'appliquent selon la nature du processus et parfois compte tenu du nombre d'observations.

Selon Kline(1998), Song et Shepperd (2007), il y a trois stratégies possibles pour traiter des données manquantes :

- Utilisation des procédures de suppression.
- Utilisation des procédures de remplacement, (substitution) les données manquantes par les valeurs présentes.
- Utilisation des procédures de modélisation de la distribution des données manquantes et les estimés par certains paramètres. [24]

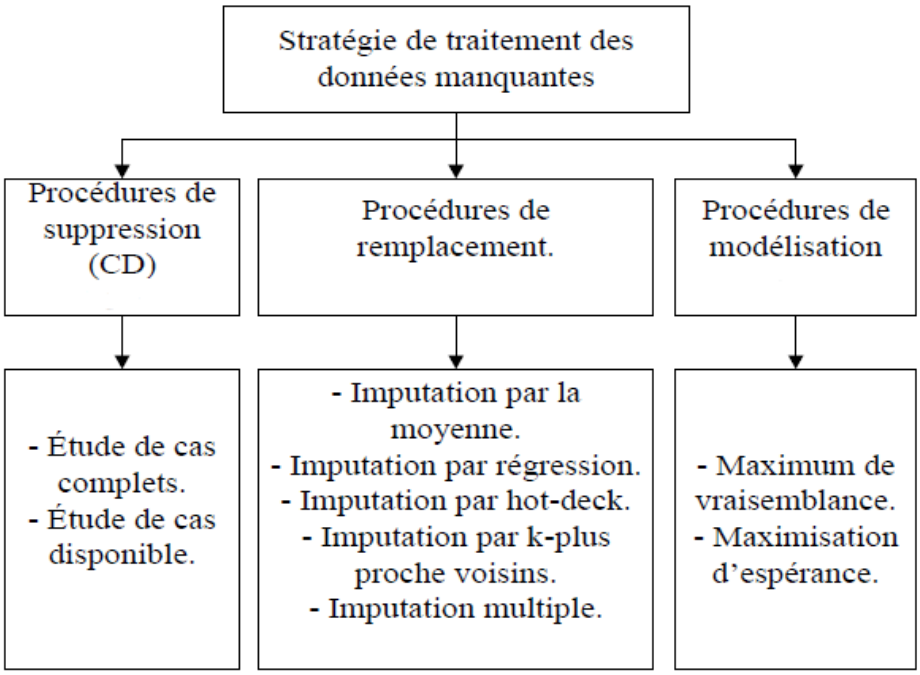


Figure 5: Les grandes catégories des méthodes pour le traitement des données[24]

III.6 Méthodes de Traitement

- Analyse de données complètes
- Indicateur de données manquantes
- Analyse pondérée
- Imputation simple
- Imputation multiple

III.6.1 Analyse de Données Complètes

Au lieu de travailler sur toute la table, l'analyse ne porte que sur les enregistrements complets

Tableau 2: Analyse de données complètes

	1	2	3
1			
2	NA		
3		NA	
4			
5			
6			
7		NA	
8			
9			NA
10			
11			NA
12			
13			
14			

	1	2	3
1			
4			
5			
6			
8			
10			
12			
13			
14			

- Stratégie la plus courante
- Généralement imposée par les logiciels
- Proportion d'observations complètes peut être faible même si, pour chaque variable, la probabilité qu'une donnée soit observée est grande
- Résultats non biaisés si les données sont MCAR, mais diminution de la précision et de la puissance
- Sinon biais importants

III.6.2 Indicateur de Données Manquantes

On ajoute une modalité à la variable catégorielle incomplètement observée et l'analyse portera sur tous les enregistrements

	1
1	1
2	NA
3	2
4	2
5	2
6	1
7	NA
8	2
9	NA
10	1
11	NA
12	1
13	2
14	2

Tableau 3: Indicateur de Données Manquantes

	1
1	1
2	9
3	2
4	2
5	2
6	1
7	9
8	2
9	9
10	1
11	9
12	1
13	2
14	2

- Suppose des données MCAR ou MAR
- Peut améliorer la précision de certains estimateurs
- Permet d'apprécier le risque de biais
  - ✓ Une interaction significative entre l'indicatrice de données manquantes et une variable explicative signale l'existence d'un problème
- Mais ne protège pas contre le risque de biais

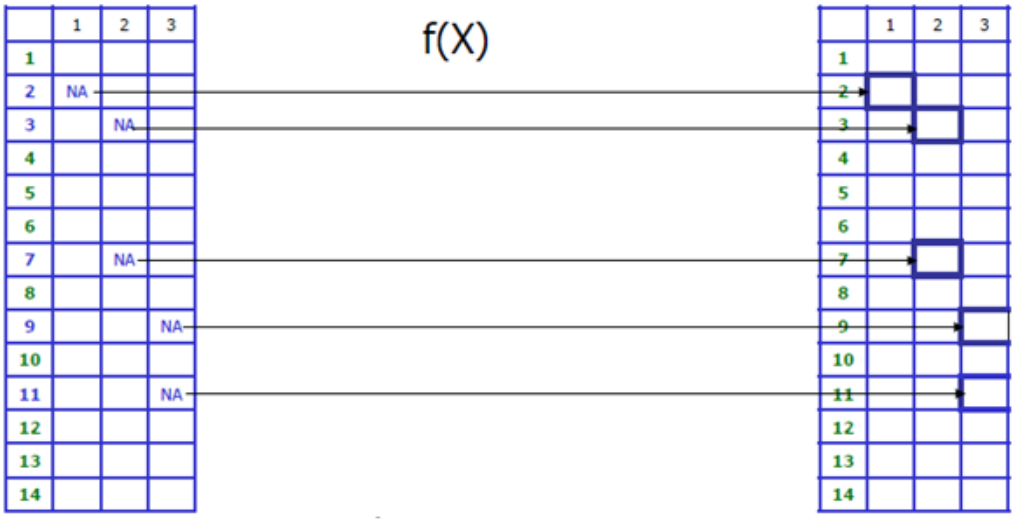
III.6.3 Analyse Pondérée

- Pour des données MAR
- Estimation de la probabilité qu'une observation soit complète pour chaque combinaison des variables influençant cette probabilité

III.6.4 Imputation Simple

On remplace chaque donnée manquante par une donnée prédite ou simulée, et l'analyse portera sur tous les enregistrements.

Tableau 4: Imputation simple



- L'imputation Simple est l'hypothèse d'un processus d'observation MAR
- Produit une valeur «artificielle» pour remplacer la valeur manquante
- Les informations disponibles sur les individus qui ne fournissent qu'une réponse partielle peuvent être utilisées comme variables auxiliaires pour améliorer la qualité des valeurs imputées

III.6.4.1 *Imputation Simple : Dernière Observation*

- Lors de mesures répétées, suppose que la vraie valeur reste inchangée depuis la dernière mesure
- Si pas de mesure disponible pendant le suivi, la valeur initiale est utilisée

III.6.4.2 *Imputation Simple : Hot-Deck et Cold-Deck*

- Hot-Deck : La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques
- Cold-Deck : La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques, mais provenant d'une autre source d'information

III.6.4.3 *Imputation Simple : par la Moyenne*

- Remplacement d'une valeur manquante par la moyenne des mesures disponibles
- La même pour toutes les valeurs manquantes d'une même variable
- Estimations non biaisées si les données sont MCAR

III.6.4.4 *Imputation Simple: par un modèle de Régression*

- Remplacement d'une valeur manquante Yi par une valeur prédite Y* obtenue par régression de Y sur X1, X2, …
- Possibilité d'ajouter un aléa à la prédiction
- Estimation ponctuelle correcte
- Variance sous-estimée

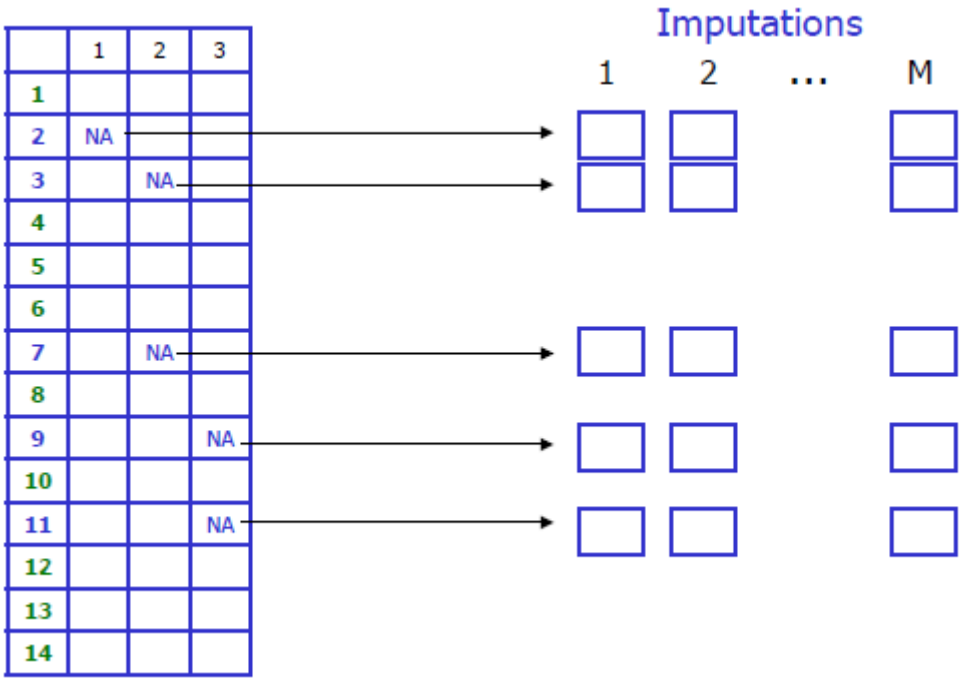
III.6.5 Imputation Multiple

- Méthode consistant à créer plusieurs valeurs possibles d'une valeur manquante
- Les buts sont :
  - ✓ De refléter correctement l'incertitude des Valeurs manquantes
  - ✓ De préserver les aspects importants des distributions
  - ✓ De préserver les relations importantes entre les variables
- Les buts ne sont pas :
  - ✓ De prédire les données manquantes avec la plus grande précision
  - ✓ De décrire les données de la meilleure façon possible

III.6.5.1 *Les Étapes de l'Imputation Multiple :*

Remplacer chaque valeur manquante par M > 1 valeurs tirées d'une distribution appropriée

Tableau 5: Imputations



Analyses indépendantes et avec la même méthode standard des M > 1 bases de données complètes

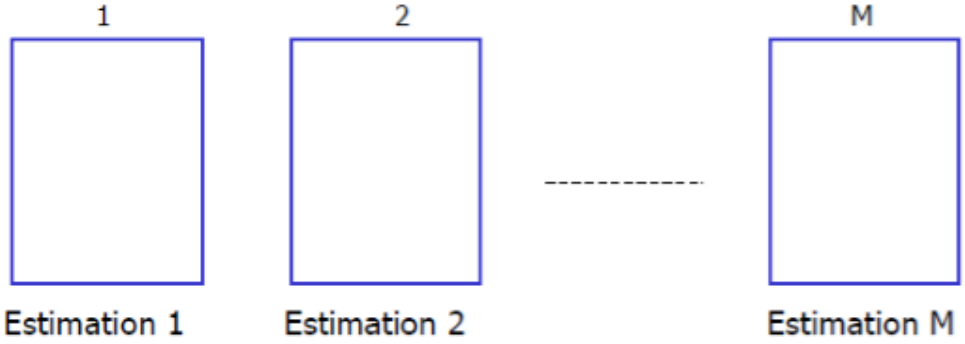


Figure 6 : Bases de données avec valeurs observées et imputées [25]

Combiner les résultats des analyses afin de refléter la variabilité supplémentaire due aux données manquantes

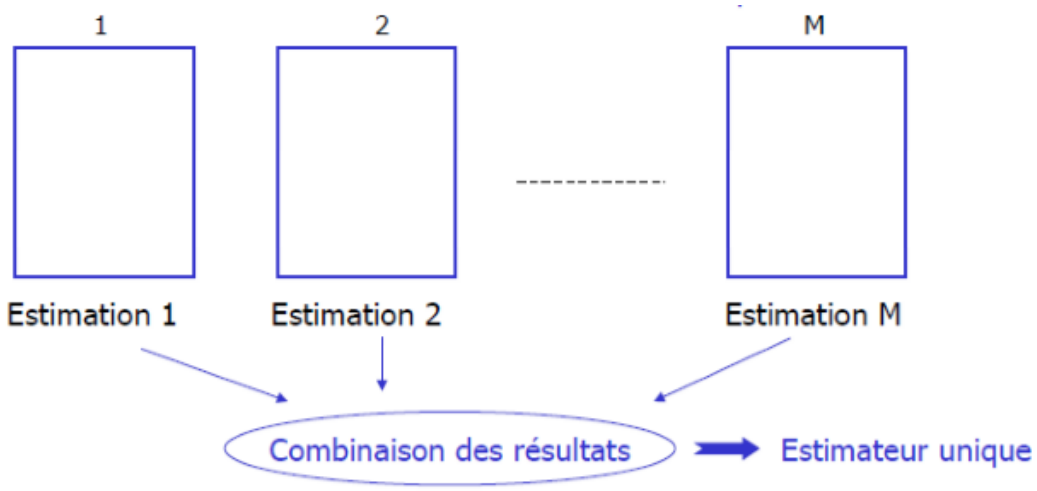


Figure 7: valeur estimée unique [25]

III.6.6 Une Synthèse

- **Analyse des données complètes** : Méthode la moins satisfaisantes en termes de biais et de précision
- **Indicateur de données manquantes** : Plus précis, permet d'identifier certains problèmes de biais, ne les traite pas de façon pleinement satisfaisante
- **Analyse pondérée** : Corrige le biais éventuel de données MAR, mais augmente l'imprécision
- **Imputation simple** : Exige souvent un processus d'observation MCAR
- **Imputation multiple** :
  - ✓ Prend en compte simultanément les problèmes de biais et de précision
  - ✓ Flexible
  - ✓ Adaptée pour des données qualitatives et quantitatives
  - ✓ Utilisable pour différents type d'analyse (régression logistique, …) [25]

III.7 Analyse des valeurs manquantes

La procédure d'analyse de la valeur manquante exécute trois fonctions principales :

- Elle décrit le type des données manquantes. Quel est l'emplacement des valeurs manquantes ? Quelle est l'importance de leur nombre ? Les paires de variables ont-elles tendance à contenir des valeurs manquantes dans les observations multiples ?
  Les données ont-elles des valeurs extrêmes ? Les valeurs manquent-elles de façon aléatoire ?
- Estime les moyennes, écarts-types, covariances et corrélations pour différentes méthodes relatives aux valeurs manquantes : par liste, par pair, régression ou EM (prévision-maximisation). La méthode concernant seulement les composantes non valides affiche également l'effectif des observations complètes par paires.
- Remplit (impute) les valeurs manquantes avec des valeurs estimées à l'aide de méthodes de régression ou EM ; mais les résultats de l'imputation multiple sont généralement considérés comme plus précis. [26]

III.8 Données manquantes et imputation

L'imputation regroupe les méthodes utilisées pour remplacer les données manquantes.

III.8.1 Imputation par la moyenne

Les méthodes d'imputation les plus simples consistent à remplacer les données manquantes par leur moyenne ou leur médiane. L'inconvénient de cette approche est qu'elle conduit à une sous-estimation parfois violente de la variance des estimateurs.

III.8.2 Imputation par tirage conditionnel

On peut améliorer l'idée de l'imputation par la moyenne en réalisant de l'imputation par tirage conditionnel. Le principe est d'utiliser l'information apportée par les variables renseignées.
Plusieurs approches sont possibles :

1. Estimer la loi jointe et générer conditionnellement une réalisation pseudo aléatoire de cette loi. Mais il est généralement difficile d'estimer une loi jointe au-delà de 2 ou 3 variables. Une alternative intéressante et plus facile à mettre en œuvre, bien qu'éventuellement couteuse, consiste à utiliser une méthode des plus proches voisins.
- Soit l'individu présentant un non réponse.
- Calculer la distance de à tous les individus ayant les mêmes variables renseignées.
- Retenir les plus proches voisins.
- Imputer la moyenne des plus proches voisins à la donnée manquante.
2. Réaliser une classification à partir des variables complètement renseignées et estimer la moyenne conditionnelle par classe. On peut voir cette méthode comme une sorte généralisation de la méthode des plus proches voisins. Dans les deux cas, l'imputation va se baser sur les observations les plus proches.
3. Construire un modèle de régression à partir des individus complètement renseignés et l'utiliser pour prédire les données correspondant aux données manquantes.

De façon générale, il est préférable de faire de l'imputation multiple. L'idée est de réaliser plusieurs tirages et de répéter les analyses pour prendre en compte et rétablir la variabilité sous-jacente à l'absence de données. L'usage est de faire 5 tirages...

III.8.3 Imputation par analyse factorielle

Considérons le cas de données issues de variables quantitatives. L'analyse en composante factorielle permet de 'reconstruire' des données par projection dans un espace de dimension réduite. Cette caractéristique peut-être exploité pour remplacer des données manquantes.
L'approche la plus naïve consiste à estimer la matrice de covariance à partir des individus renseignés puis d'estimer les paramètres de l'analyse en composante principale et enfin à reconstruire les données manquantes.[27]

III.8.4 Imputation par le plus proche voisin :

1. Calcul des **distances euclidiennes entre receveurs et donneurs**, pour chaque classe d'ajustement
2. **Recherche de la plus petite distance entre un receveur** (individu ayant des réponses manquantes) **et un donneur**
3. **Attributions des valeurs** des variables spécifiques du donneur au receveur

Elimination du receveur utilisé et du donneur. [28]

III.9 Dangers de l'imputation :

1. Même si l'imputation produit un fichier complet de données, l'inférence, en particulier l'estimation ponctuelle, n'est valide que si les hypothèses sous-jacentes sont satisfaites.
2. L'imputation modifie les relations entre les variables.

3. Si les valeurs imputées sont traitées comme des valeurs observées, la variance de l'estimateur risque d'être considérablement sous-estimée, surtout si la proportion de non-réponses est appréciable. [29]

III.10 L'estimation basée sur des modèles explicites

On se placera dans un cadre paramétrique puisque les méthodes utilisées vont tenir compte des données existantes, c'est à dire recueillies au cours des diverses enquêtes. En effet, l'estimation de paramètres d'une loi d'une variable présentant des données manquantes devra se référer à toute l'information existante sur cette variable et également sur les autres variables du fichier.

Chaque valeur manquante peut être estimée grâce à des techniques classiques de régression (régression linéaire, régression logistique, modèle linéaire généralisé). Chaque variable de $\mathbf{Y_0}$ est modélisée à partir des coordonnées de $\mathbf{X_0}$ qui joue le rôle de variables explicatives. Le modèle généré est alors appliqué au fichier receveur. Le principe de ces méthodes est de prédire les valeurs manquantes en utilisant un modèle de régression adapté aux variables observées. Il y a plusieurs possibilités et nous allons en citer quelques-unes :

- **une régression simple** en prenant la variable la plus corrélée.
- **une régression multiple** en prenant le meilleur sous-ensemble de variables explicatives, utilisant un modèle pas à pas, ou la méthode de Furnival et Wilson d'exploration optimisée de toutes les possibilités.
- **une analyse de variance**, cas particulier de la régression lorsque la variable explicative X est nominale et la variable à expliquer est quantitative.
- **une analyse discriminante,** lorsque la variable à expliquer est multi classe ou **une régression logistique** lorsque la variable expliquée est dichotomique. On impute alors par la catégorie la plus probable.

Toutes ces techniques, bien que simples présentent au moins deux inconvénients majeurs : les variables sont estimées une par une et non conjointement : ainsi ces techniques d'estimation ne prennent pas en compte les corrélations éventuelles entre les variables, ce qui peut induire des résultats incohérents. Si aucune vérification des résultats trouvés n'a été prévue à la fin du processus d'estimation, des résultats incohérents peuvent se produire, comme par exemple un jeune homme de 20 ans qui serait retraité…

On peut aussi appliquer **la méthode du maximum de vraisemblance**. Le principe de cette méthode, sous l'hypothèse que les données qualitatives proviennent d'un échantillon d'une variable aléatoire multinomiale, est le suivant :

Les paramètres de la loi multinomiale sont estimés par l'algorithme EM (Dempster, Laird, Rubin, 1977) que l'on décrit ici dans son principe. Partant d'une estimation des paramètres de la loi, il s'agit d'un algorithme itératif utilisant alternativement deux étapes. L'étape E (E comme espérance) consiste à déterminer l'espérance conditionnelle de chaque donnée manquante sachant les données observées et l'estimation courante des paramètres. L'étape M (M comme maximisation) consiste à calculer les estimateurs du maximum de vraisemblance des paramètres, les formules faisant usage des lois conditionnelles des données manquantes. De manière naturelle, à la convergence de l'algorithme EM, on attribue à chaque donnée manquante la valeur la plus probable pour l'estimation obtenue des paramètres de la loi multidimensionnelle.

De cette façon, tous les individus qui présentent des données manquantes pour les mêmes variables sont complétés de manière identique. Mais la méthode du maximum de vraisemblance ne prévient pas non plus des estimations incohérentes.

Un autre inconvénient de ces techniques d'estimation est le suivant : deux unités ayant les mêmes valeurs (lignes identiques dans $\mathbf{X_1}$) auront le même estimateur pour leur variable Y, d'où une variabilité insuffisante dans $\mathbf{Y_1}$ **;**

La technique d'imputation multiple (Rubin 1987), consiste à imputer chaque donnée par m (≥ 2) valeurs obtenues par tirage dans un ou plusieurs modèles d'estimation. Puis on fait l'analyse des données sur chacun des m jeux de données ainsi complété.
L'estimateur final d'un paramètre quelconque sera la moyenne des m estimations ainsi réalisées. L'imputation multiple sous un ou plusieurs modèles permet de simuler la distribution a posteriori des données manquantes sous ce ou ces modèles et d'obtenir des variances correctes. Les inconvénients majeurs sont la complexité des calculs sous un ou plusieurs modèles, le temps de calcul considérable et la quantité à stocker et à gérer. [30]

III.11 Filtrage de données sous Weka :

La section **Preprocess** permet aux filtres d'être définis ce qui a pour effet de transformer les données de diverses manières. La boîte **Filter** est utilisée pour mettre en place les filtres nécessaires. A gauche de la boîte **Filter** se trouve un bouton **Choose**. En cliquant sur ce bouton il est possible de choisir un des filtres dans WEKA. Une fois qu'un filtre a été choisi, son nom et ses options apparaissent dans le champ à côté du bouton **Choose**. [31]

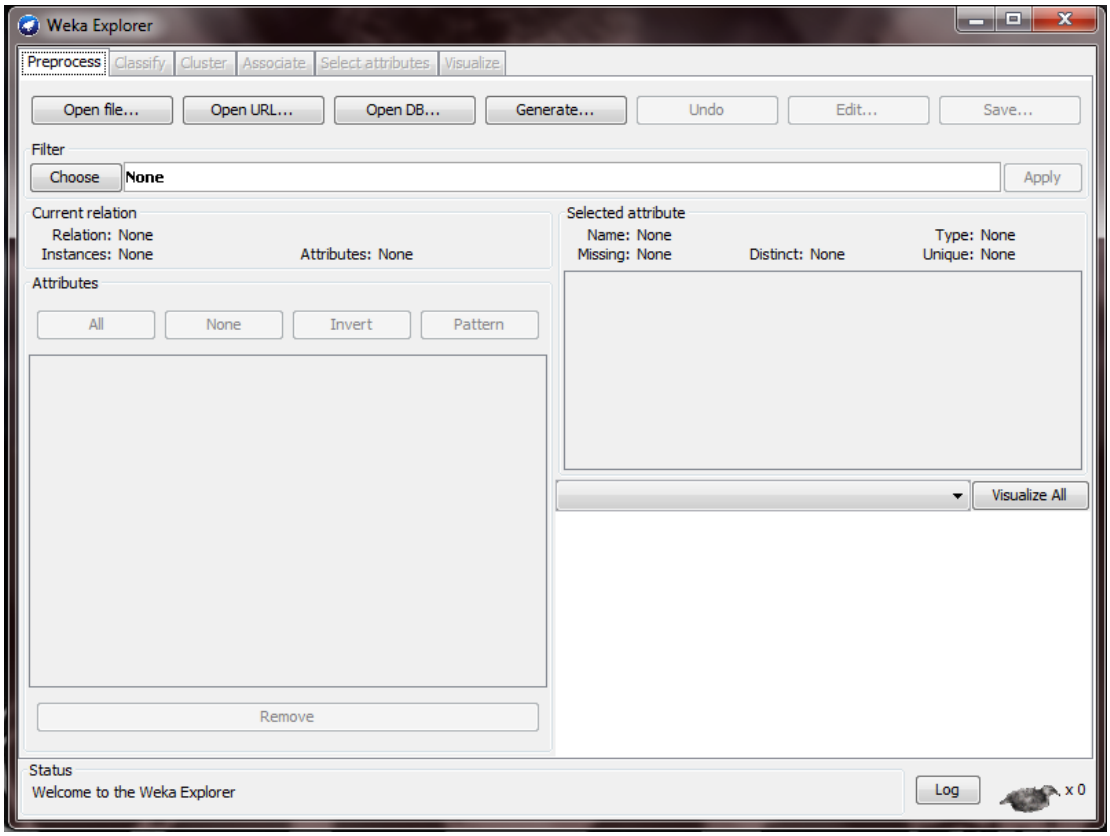


Figure 8: l'anglet Preprocess sous Weka

III.12 Conclusion :

On constate que les méthodes d'imputation de données manquantes sont nombreuses et qu'il n'existe pas de recettes définitives.
Il est souvent nécessaire d'opérer des va-et-vient entre les données brutes et les données corrigées ou imputées.
En effet il n'est pas toujours possible de définir *a priori* les contrôles susceptibles de détecter toutes les incohérences, et de prévoir à l'avance les méthodes d'imputations les plus pertinentes.

Il faut alors repartir des données brutes pour tester un autre mode de traitement, en veillant à ce qu'il n'interfère pas sur d'autres traitements déjà réalisés.

CHAPITRE 3 :
Catégorisation

IV. Chapitre 3 : Catégorisation

IV.1 Introduction :

La classification de données situées dans un espace de grande dimension est un problème délicat qui apparaît dans de nombreuses sciences telles que l'analyse des catégories sociales

La classification automatique des données consiste à diviser un jeu de données en sous-ensembles de données appelés classes pour que tous les individus dans même une classe soient similaires et les individus de classes distinctes soient dissimilaires.
Typiquement, chaque classe est représentée par un individu qui s'appelle le centre de la classe ou par certaines informations dérivées de tous les individus de la classe qui sont suffisantes de décrire la classe.
Il y a plusieurs algorithmes de classification des données. Ils diffèrent par la nature de données qu'ils traitent (données numériques ou données de catégorie, petit jeu de données ou gros jeu de données, données de dimension élevée ou moins élevée, sur un flux de données ou pas…), par les méthodes de distribution des données en classes, par la représentation des classes…

IV.2 État de l'art

La classification est une méthode d'analyse des données qui vise à regrouper en classes homogènes un ensemble d'observations. Ces dernières années, les besoins d'analyse de données et en particulier de classification ont augmenté significativement. En effet, de plus en plus de domaines scientifiques nécessitent de catégoriser leurs données dans un but descriptif ou décisionnel. [32]

Catégoriser c'est :

1. être capable d'opérer des rapprochements thématiques ou sémantiques entre des objets,
2. classer, ranger, trier les objets ou représentations d'objets selon des caractéristiques communes,
3. observé, distingué, discriminé, comparé,
4. trouver un critère de tri. [33]

IV.3 Principe de catégorisation

Il s'agit de stocker l'information en la structurant de manière mémorisable et opérante[1].

Selon l'approche logique, une catégorie est définie sur la base d'une relation d'appartenance permettant de dire si oui ou non un élément appartient à une catégorie[2]

La catégorisation se révèle être une activité cognitive consistant à regrouper des objets ou des événements non identiques dans des catégories[3]. [34]

IV.4 Catégoriser, a quoi ca sert ?

Les activités de catégorisation permettent de:

- Structurer et organiser la pensée,
- développer le langage, nommer avec précision, communiqué,
- aider a mémorisé,

[1] Ladwein, 1995.
[2] Piaget, 1972.
[3] Mervis et Rosch,1981.

- enrichir et préciser le lexique,
- développer la flexibilité pour mieux comprendre le monde, faire des relations entre les objets, relier l'inconnu au connu,
- développer l'esprit critique,
- construire et développer la pensée logique, mettre en place des stratégies, organiser le raisonnement,
- apprendre à justifier ses choix, a argumenté,
- travaillé sur le sens des consignes,
- donné des repères dans les apprentissages (méthodologie) [35]

IV.5 Algorithmes de classification

- **Algorithmes hiérarchiques :** décomposition/composition hiérarchique de clusters (CURE, BIRCH)
- **Algorithmes par partitionnement :** regroupement itérative avec amélioration par replacement des objets (k-means, k-médoides)
- **Fonctions de densité** (cluster avec forme arbitraire) : clusters grandissent aussi longtemps que la densité des objets dans leur voisinage est supérieure à une borne (DBSCAN, OPTICS)
- **Grilles :** l'espace est divisé en cellules qui forment une grille (STING, CLIQUE)
- **Modèles :** chaque cluster est supposé suivre un modèle : trouver la meilleure correspondance entre les modèles et les données (COBWEB) [36]

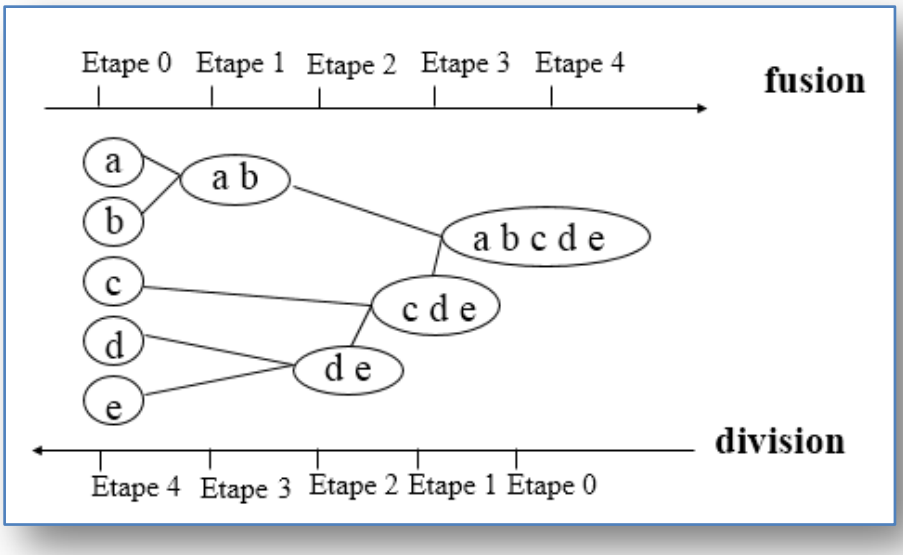


Figure 9: Exemple de classification hiérarchique [37]

IV.6 Taxinomie et catégorisation dans classement professionnel

Une des particularités du système de catégorisation professionnelle de l'INSEE (L'Institut national de la statistique et des études économiques) est de ranger les catégories socioprofessionnelles par rang hiérarchique de telle façon que chaque catégorie professionnelle se trouve à l'intérieur d'une catégorie dominante. Il fonctionne selon le principe d'une taxinomie par inclusion hiérarchique sous le modèle identique aux taxinomies des espèces basiques qui est logiquement subordonné à un rang supérieur (forme de vie ou niveau super-ordonné) (Berlin, 1976). Ce modèle suppose que les catégories, désignant les espèces naturelles, soient ordonnées hiérarchiquement autour d'un niveau fondamental de conceptualisation (générique ou niveau basique) qui est logiquement subordonné à un rang supérieur (forme de vie ou niveau super-ordonné).

L'application d'un modèle taxinomique a des justifications à des fins d'expertise pour obtenir une projection nationale des professions en France qui correspond à des niveaux différents de regroupement de population. Cependant la construction d'une représentation taxinomique des professions et des métiers présente une série de difficultés. Les professions ne sont pas institutionnalisées de la même façon, ainsi les critères se combinent mal (statut, catégorie, secteur d'activité) : "Cette coexistence (des modes de construction des catégories), qui suscite l'irritation des théoriciens…, est le reflet de ce que l'espace professionnel lui-même s'est historiquement construit sur la double base de la tradition des métiers (liée à une forme encore familiale et artisanale de la production), et des grilles d'emplois qualifiés, induites par une mise en relation logique entre des compétences codifiées par des diplômes et des postes résultant de l'organisation rationnelle du travail dans les grandes entreprises industrielles" (Desrosières, 1989). L'application d'une nomenclature fondée sur une projection nationale se heurte à des contraintes écologiques (démographique, organisation sociale des professions) qui ont été très bien analysées par Desrosières et Thévenot (1988). [38]

IV.7 Les différents types de données rencontrés

La classification intervient sur des données qui résultent d'une suite de choix qui vont influencer les résultats de l'analyse. Classiquement, les données sont décrites dans un tableau individus-variables par une valeur unique.
Dans les applications réelles, où le grand souci est de prendre en compte la variabilité et la richesse d'informations au sein des données, il est courant d'avoir affaire à des données complexes et hétérogènes (ou mixtes).

- ***Les variables quantitatives***

Une variable *quantitative* prend des valeurs ordonnées (comparable par la relation d'ordre≤) pour lesquelles des opérations arithmétiques telles que différence et moyenne aient un sens.
Une variable quantitative peut être *binaire*, *continue* ou *discrète*.

- ***Les variables qualitatives***

Une variable *qualitative* (ou aussi *catégorielle*) est une donnée dont l'ensemble des valeurs est fini. Elle prend des valeurs symboliques qui désignent en fait des *catégories* ou aussi *modalités* (*exemples* : le code de la ville, la couleur des cheveux). On ne peut effectuer aucune opération arithmétique sur ces variables. [39]

IV.8 Applications de la classification

La classification automatique est une technique utilisée dans plusieurs domaines. Sa capacité prédictive la rend rapide et efficace. Parmi les applications où la classification est utilisée, nous trouvons le filtrage de spam, en effet il s'agit de traiter les messages électroniques textuels, identifier leurs caractéristiques et les classer en deux groupes messages désirés ou non désirés.
Une autre application est la détermination automatique du sujet d'un texte pour le classer automatiquement afin de notifier des personnes intéressées par ce sujet de la présence d'un nouveau texte… [40]

IV.9 Classification automatique sur données mixtes

La classification automatique ou typologie (clustering en anglais) vise à regrouper les observations en classes : les individus ayant des caractéristiques similaires sont réunis dans la même catégorie ; les individus présentant des caractéristiques dissemblables sont situés dans des catégories distinctes. La notion de proximité est primordiale dans ce processus. Elle est quantifiée différemment selon le type des variables. La distance euclidienne est souvent utilisée (normalisée ou non) lorsqu'elles sont quantitatives, la distance du khi-2 lorsqu'elles sont qualitatives (les individus qui possèdent souvent les mêmes modalités sont réputées proches).

L'affaire se corse lorsque nous sommes en présence d'un mix de variables quantitatives et qualitatives. Certes il est toujours possible de définir une distance prenant en compte simultanément les deux types de variables (ex. la distance HEOM). Mais le problème de la normalisation est posé. Telle ou telle variable ne doit pas avoir une influence exagérée uniquement de par sa nature. [41]

IV.10 Etapes de la classification

La classification automatique de documents doit être obligatoirement précédée par une phase de préparation de données appelée « préprocessing ».

IV.10.1 Préprocessing

IV.10.1.1 *Extraction de données*

L'extraction de données est une tâche qui s'est développée dans le domaine de Traitement Automatique des Langues (TAL). Elle consiste à identifier et extraire d'un texte les éléments pertinents contenant des informations dont la nature est spécifiée à l'avance. Elle vise donc à transformer un texte de son format initial (une suite de chaines de caractères) à une représentation structurée et donc un format qui soit compréhensible par l'ordinateur. Elle se fait en reconnaissant dans le texte des unités lexicales particulières.
Il existe plusieurs techniques d'extraction de données telles que :

- **Les outils terminologiques** : nous nous intéressons aux termes présents dans un texte, ces outils peuvent être linguistiques, statistiques ou mixtes, les méthodes linguistiques se basent sur des patrons lexicaux-syntaxiques et sur le découpage des textes en unités syntaxiques. Les méthodes statistiques font des calculs sur les fréquences et les distributions des mots dans les textes.
- **Les méthodes d'extraction de relation terminologique** : ici nous nous intéressons plutôt à la relation entre les différents termes et structures des textes. Là aussi plusieurs approches peuvent être utilisées, on distingue les méthodes structurelles, contextuelles et distributionnelles
- **La reconnaissance des entités nommées** : le plus souvent il s'agit de noms de personnes, de lieux, de noms d'organisations, etc. Selon le domaine traité par le texte d'autres entités peuvent être ajoutées à la liste précédente, par exemple dans un texte médical il est important de reconnaître les noms de maladies ou de symptômes.

IV.10.1.2 *Calcul de poids*

Les données collectées du texte n'ont pas toutes la même valeur informative, Certaines données sont plus importantes que d'autres pour la classification car elles apparaissent de façon répétée dans le texte, d'autres le sont moins à cause de leur caractère commun dans la langue. Pour permettre une meilleure classification, il faut refléter ce degré d'importance dans l'algorithme. C'est ainsi que nous attribuons à chaque mot un poids.
IL existe plusieurs méthodes de calcul de poids, en voici deux des plus connues :

- **TF (Term Frequency) :** C'est la fréquence d'apparition d'un mot dans le texte, elle est égale aux nombre d'occurrences de ce mot divisé par le nombre total de mots du texte.
- **TF – IDF (term frequency, inverse document frequency):** C'est une mesure statistique qui permet d'évaluer l'importance d'un mot dans un texte relativement à tout un corpus. Le poids ne dépend pas seulement de la fréquence du mot dans le texte en question mais aussi de la fréquence du mot dans tout le corpus. Ainsi un mot qui se répète dans tous les documents devient inutile pour la classification.

$$TF\_IDF = TF \times IDF$$

IV.10.2 Classification

C'est à cette étape que se fait l'assignation du document à la classe à laquelle il appartient. La détermination de la classe se fait grâce à des algorithmes de classification qui exploitent les données extraites dans l'étape précédente et qui donnent en sortie la classe correspondante. Ces algorithmes se basent sur leur expérience passée où ils ont appris comment classer les textes, c'est de là que vient leur nom « Algorithmes d'apprentissage ». Il existe deux types de ces algorithmes :

- Les algorithmes d'apprentissage supervisé
- Les algorithmes d'apprentissage non supervisé. [42]

IV.11 Les Techniques de classification automatique

IV.11.1 Apprentissage non supervisé

IV.11.1.1 *Principe*

L'apprentissage non supervisé consiste à apprendre à classer sans supervision. Au début de processus nous ne disposons ni de la définition des classes, ni de leurs nombres. C'est l'algorithme de classification qui va déterminer ces informations. Nous ne disposons pas non plus de données en entrée qui sont déjà classées, c'est aussi à l'algorithme de découvrir par lui-même la structure plus ou moins cachée des données et de former des groupes d'individus dont les caractéristiques sont communes.
L'apprentissage non supervisé est utilisé dans plusieurs domaines tels que :

- Médecine : Découverte de classes de patients présentant des caractéristiques physiologiques communes.
- Le traitement de la parole : construction de système de reconnaissance de la voie humaine.
- Archéologie : regroupement des objets selon leurs époques.
- Traitement d'images
- Classification de documents

Dans la littérature il existe plusieurs types d'algorithmes d'apprentissage non supervisé tels que les algorithmes de partitionnements et les algorithmes de classification hiérarchique :

- **Le partitionnement**:

Consiste au regroupement des données suivant leur degré de similarité.
L'algorithme le plus célèbre appartenant à cette classe est K-means : c'est un algorithme qui permet de partitionner un ensemble de données automatiquement en K clusters. Il consiste tout d'abord à choisir k points qui représentent les centres des groupes à créer, puis à affecter les autres points aux centres les plus proches. Cette affectation est faite par le calcul de distance entre les points. Plusieurs distances peuvent être définies telles que la distance euclidienne ou la distance de Manhattan. Par la suite nous procédons à une étape de raffinement des groupes de façon itérative, le raffinement se fait par le recalcule des centres des groupes après chaque itération et par une réaffectation des points aux groupes. L'algorithme s'arrête quand aucun point ne bouge.

- **La classification hiérarchique** :

Il existe deux types de classification hiérarchique : Ascendante et descendent. La classification ascendante consiste à utiliser une matrice de similarité afin de partir d'une répartition fine vers un groupe unique. Donc, il s'agit de fusionner les groupes jusqu'à ce qu'on obtient un seul groupe englobant tous les autres. Cette classification peut être représentée par un arbre hiérarchique ou dendrogramme. La classification descendante se présente comme l'inverse de la classification ascendante. Donc il s'agit de décomposer un cluster unique en sous-groupes jusqu'à l'obtention des singletons.

IV.11.2 Apprentissage supervisé

IV.11.2.1 *Principe*

Contrairement à l'apprentissage non supervisé, nous commençons ici par un ensemble de classes connues et définies à l'avance. Nous disposons aussi d'une sélection initiale de données dont la classification est connue. Ces données sont supposées indépendantes et identiquement distribuées.

Elles nous servent pour l'apprentissage de l'algorithme. La classification se fait par l'algorithme selon le modèle qu'il a appris.
Il existe plusieurs algorithmes d'apprentissage supervisé, cette section présente quelques-uns des plus connus parmi eux, il s'agit de kNN, LLSF, les réseaux de neurones et Naive Bayes :

- **kNN : (k nearest neighbor),** c'est une approche statistique de classification très connue. Il a été prouvé que c'est une des méthodes les plus performantes après des tests réalisés sur le corpus de données Reuters. Le principe de l'algorithme kNN est le suivant : étant donné un texte à classer, l'algorithme cherche les k voisins les plus proches parmi les documents utilisés au cours de la phase d'apprentissage, les catégories de ces k voisins les plus proches serviront à donner des poids aux catégories candidates de classification. C'est le degré de similarité entre le document test et le document voisin qui est utilisé comme poids de la catégorie de ce dernier, si plusieurs voisins partagent la même catégorie alors le poids attribué à cette catégorie est égal à la somme des degrés de similarité entre le document test et chacun des voisins appartenant à cette catégorie. Par cette méthode on peut obtenir une liste des poids attribués à chaque catégorie, le document test est classé dans une catégorie si le poids attribuée à celle-ci est supérieur à un seuil fixé à l'avance.
- **LLSF (linear least square fit) :** c'est une approche de mapping développée par Yang, il s'agit d'écrire les données d'apprentissage sous la forme de paires de vecteurs entrée/sortie, le vecteur d'entrée est composé des mots du texte accompagnés de leurs poids respectifs, alors que le vecteur de sortie est composé des différentes catégories avec leur poids binaires (1 si le texte appartient à une catégorie et 0 sinon), la résolution de l'équation suivante permet d'obtenir la matrice des coefficients de régression mot-catégorie.
- **Les réseaux de neurones :** c'est une structure constituée de suite successive de couches de noeuds et qui permet de définir une fonction de transformation non linéaire des vecteurs d'entrées (composés dans le cas de classification des mots pondérés de leur poids) en vecteur de catégories. La disposition des neurones dans le réseau ainsi que le nombre de couches utilisées ont une influence sur le résultat de classification. Comparés aux autres méthodes de classification par apprentissage supervisé, les réseaux de neurones ont l'inconvénient que le coût d'apprentissage est assez élevé.
- **NB (Naive Bayes) :** c'est une méthode de classification probabiliste. Elle consiste à utiliser les probabilités jointes des mots et des catégories pour estimer la probabilité d'une catégorie sachant un texte à classer. Le caractère « naïf » de cette approche est dû au fait que les mots sont considérés indépendants, c'est-à-dire que la probabilité conditionnelle d'un mot sachant une catégorie est supposée indépendante des probabilités conditionnelles des autres mots sachant la même catégorie, cette assomption rend NB très efficace par rapport aux autres approches bayésiennes. Plusieurs versions de NB sont proposées dans la littérature, le model mixte multi nominal par exemple a permis d'avoir de bonnes performances. [43]

IV.12 Règles de Classification

IV.12.1 Stratégie « Separate-and-conquer »

La stratégie « separate-and-conquer » repose sur la séquence d'opérations suivante : construire une règle qui prédit au mieux sur un ensemble d'instance (conquête) ; retirer les exemples couverts par la règle de l'ensemble d'apprentissage (séparer) , cette opération est répéter jusqu'à ce qu'il ne reste plus d'observations.

La méthode produit ou bien des règles ordonnées, ou des règles indépendantes, ceci résulte de l'étape de séparation où nous pouvons retirer de la base toutes les observations couvertes par la règle construite, ou uniquement les observations qui sont correctement classées.

IV.12.2 Règles de classification

L'induction de règles de classification fait partie des approches « separateand-conquer » qui produisent de manière incrémentale un ensemble de règles de la forme : **Si** Condition **Alors** Conclusion ; où condition représente une suite de conjonctions de couples « attribut-valeur », et conclusion la classe d'affectation. Ces règles peuvent ou bien être ordonnées (appelé souvent liste de décision) ou indépendantes. [44]

IV.13WEKA :

IV.13.1 WEKA: c'est quoi?

Waikato Environment for Knowledge Analysis
Suite de logiciels d'apprentissage automatique et d'exploration de données, développée à l'université de Waikato en Nouvelle-Zélande

IV.13.2 Que contient le toolkit Weka ?

IV.13.2.1 *Outils de prétraitement des données (filtering)*

- Sélection, transformation, combinaison d'attributs
- Normalisation
- Re-échantillonnage, ...

IV.13.2.2 *Algorithmes pour l'exploration de données*

- classification non supervisée
- classification supervisée
  - Analyse de résultats
  - Comparaison d'algorithmes
  - Plusieurs interfaces graphiques

IV.13.2.3 *Format des fichiers d'entrée*

- CSV (Comma Separated Value)
- Bases de données SQL (avec JDBC)
- ARFF (Attribute-Relation File Format) :
  - les commentaires sont précédés de %
  - définition du nom de l'ensemble de données avec @relation
  - définition des descripteurs avec @attribute
    - ✓ attributs nominaux suivis des valeurs entre f, g
    - ✓ attributs numériques avec numeric
    - ✓ attributs chaines avec string
    - ✓ attributs dates avec date [45]

IV.13.3 Les algorithmes de classification

Les classifieurs dans Weka sont des modèles pour prédire des valeurs nominales ou numériques

- Algorithmes de classification inclus
  - Arbres de décision
  - Classification bayésienne naïve

- Machine à vecteurs de support (SVM)
- Perceptron multi-couche
- Réseau bayésien, etc.

- Des meta-classifieurs
  - Combinaison
  - Bagging
  - Boosting, etc. [46]

IV.13.4 L'onglet Classify dans l'Explorer

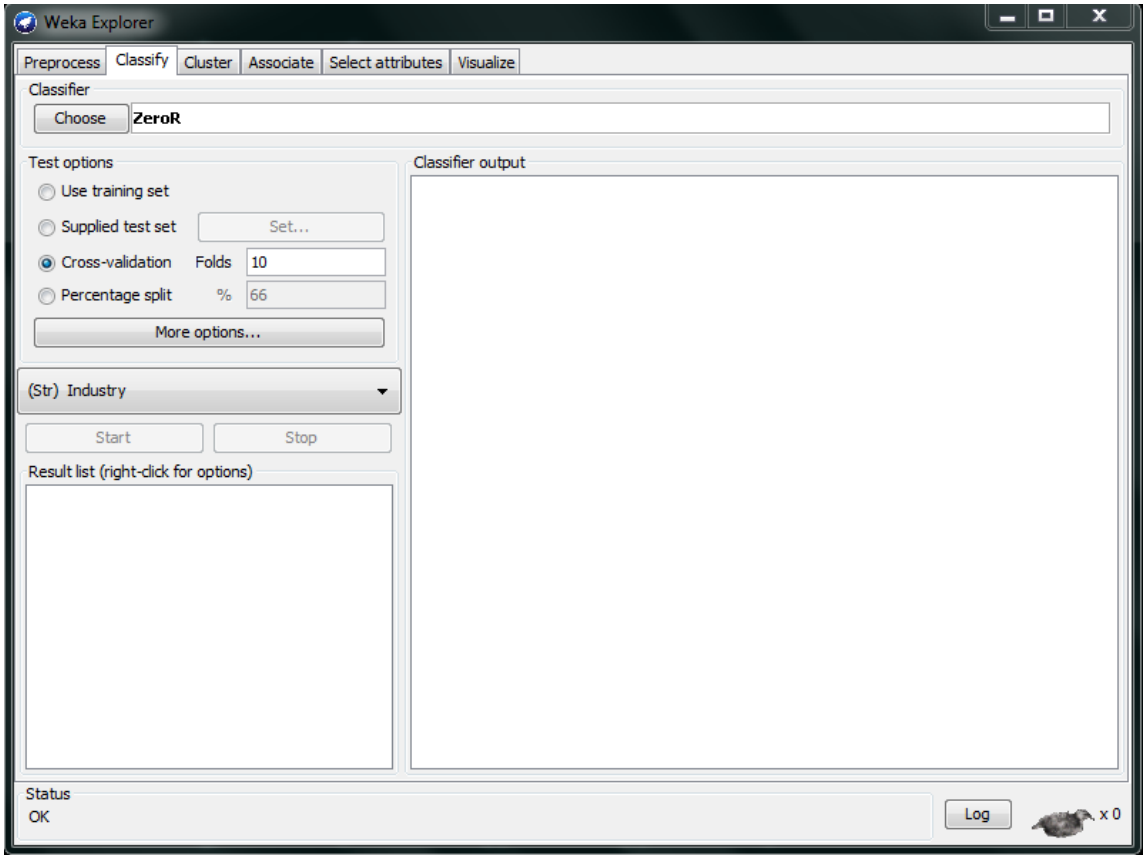


Figure 10: Classify dans l'Explorer

IV.13.5 Déduction de règles de classification

- Trouver des règles de classification simples (1R pour 1-rule)
- Idée
  - Une règle pour chaque attribut
  - Une branche pour chacune des valeurs des attributs
- 1 Pseudo code pour 1R

Pour chaque attribut

- Pour chaque valeur de cet attribut, créer une règle
  - Compter combien de fois chaque classe apparait
  - Trouver la classe la plus fréquente
  - Créer une règle : attribut-valeur -> classe
- Calculer le taux d'erreur de la règle

Choisir les règles avec le plus petit taux d'erreur [47]

IV.14 Conclusion

Le regroupement d'objets par rapport à leur similarité possède beaucoup d'applications Recherche d'information, traitement d'images, catégorisation de produits, analyse de données spatiales, nettoyage de données (recherche d'exceptions) .
Il est possible de définir des mesures de similarité pour différents types de données puisqu'il existe des algorithmes performants

CHAPITRE 4 :
Développement de l’application

V. Chapitre 4 : Développement de l'application

V.1 Introduction

Dans ce chapitre on va essayer d'expliquer le développement de l'application, le principe est de faire la catégorisation des identités sociaux, ces dernier sont des informations extraire a partir d'un fichier qui on va le décrire ci-dessous.

La catégorisation consiste à classifier les objets à l'aide de certains attributs, dans notre cas classifier les personnes selon le titre pour obtenir le sexe, la position et l'industrie, mais le problème qui nous a rencontré est la présence des donnés manquantes, puisque beaucoup des internautes ne déclarent pas tous ses informations personnel dans les réseaux sociaux.

Pour cela il est préférable de faire une imputation des donnés manquantes, cette dernier sert à remplit les valeurs manquantes avec des valeurs estimées à l'aide de méthodes de régression ou d'imputation avec la moyenne ou la médiane.

V.2 Présentation des données :

Les données utilisées sont des informations concernant des utilisateurs (275 utilisateurs) sur le réseau social Facebook stockées dans un fichier sous format **CSV,** collecté par Eugene Dubossarsky et Mark Norrie en 2004.ce fichier contient **17%** des données manquantes.

Attributs	**Type**
ID	real
LastName	string
Title	string
FirstName	string
Organisation	string
Position	string
Suburb	string
State	string
Postcode	real
WorkPhone	string
EmailAddress	string
Industry	string

Tableau 6 : les attributs qui decrit les données

V.3 Outils utilisées :

V.3.1 Weka:

Waikato Environment for Knowledge Analysis
Suite de logiciels d'apprentissage automatique et d'exploration de données, développée à l'université de Waikato en Nouvelle-Zélande

V.3.2 Logiciel R :

R est un langage de programmation interactif interprété et orienté objet contenant une très large collection de méthodes statistiques et des facilités graphiques importantes.

C'est un clone gratuit du logiciel S-Plus commercialisé par MathSoft et développé par Statistical Sciences autour du langage S (conçu par les laboratoires Bell). [48]

V.3.3 Netbans :

NetBeans est un environnement de développement intégré (EDI), placé en *open source* par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Développent and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).
NetBeans constitue par ailleurs une plate forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate forme.[49]

V.3.4 Moteur de règles « JRuleEngine1.3 » :

Est une bibliothèque Java basée sur **J**ava **S**pecification **R**equest 94. Un moteur de règles peut être considérée comme un « if / then » sophistiquée. Les états if / then qui sont interprétés sont appelés règles. Les parties 'if' de règles contiennent les conditions, Les parties 'then' de règles contiennent des actions [50]

V.3.5 jsr94-1.1 :

« Java Specification Request **»** pour une API Java Rules Engine, développé par le Java Community Process de programme (JCP), définit une API d'exécution Java pour les moteurs de règles en fournissant une API simple pour accéder à un moteur de règles à partir d'une plate-forme Java, Standard Edition (Java SE, anciennement connu sous le nom J2SE) ou une plate-forme Java, Enterprise Edition (Java EE, anciennement connu sous J2EE) client de la technologie Java. [51]

V.4 Préparation de données :

V.4.1 Filtrage de données :

Weka propose énormément de filtres. Parmi eux il ya des filtres qui permet de :

- Discrétiser des données (c'est-à-dire transformer l'ensemble de valeurs en un nombre fini d'éléments)
- Remplacer des valeurs manquantes
- Eliminer des attributs non pertinents
- Modifier le type d'attribut

Avant de commencer il faut élimine les attributs inutiles, puisque on n'a pas besoin de l'identifiant, le nom, le prénom ainsi que le Numéro de Téléphone et l'adresse email, parce ces derniers ne servent à rien dans la catégorisation.

La méthode utilisé pour le filtrage de donnés est la méthode trouver sous weka « **StringToWordVector »,** cette dernier sert à convertis des chaînes de caractères en un ensemble d'attributs représentant l'information sur les occurrences d'un mot, l'ensemble des mots (attributs) est déterminée par le premier lot filtré (formation généralement des données). [52]

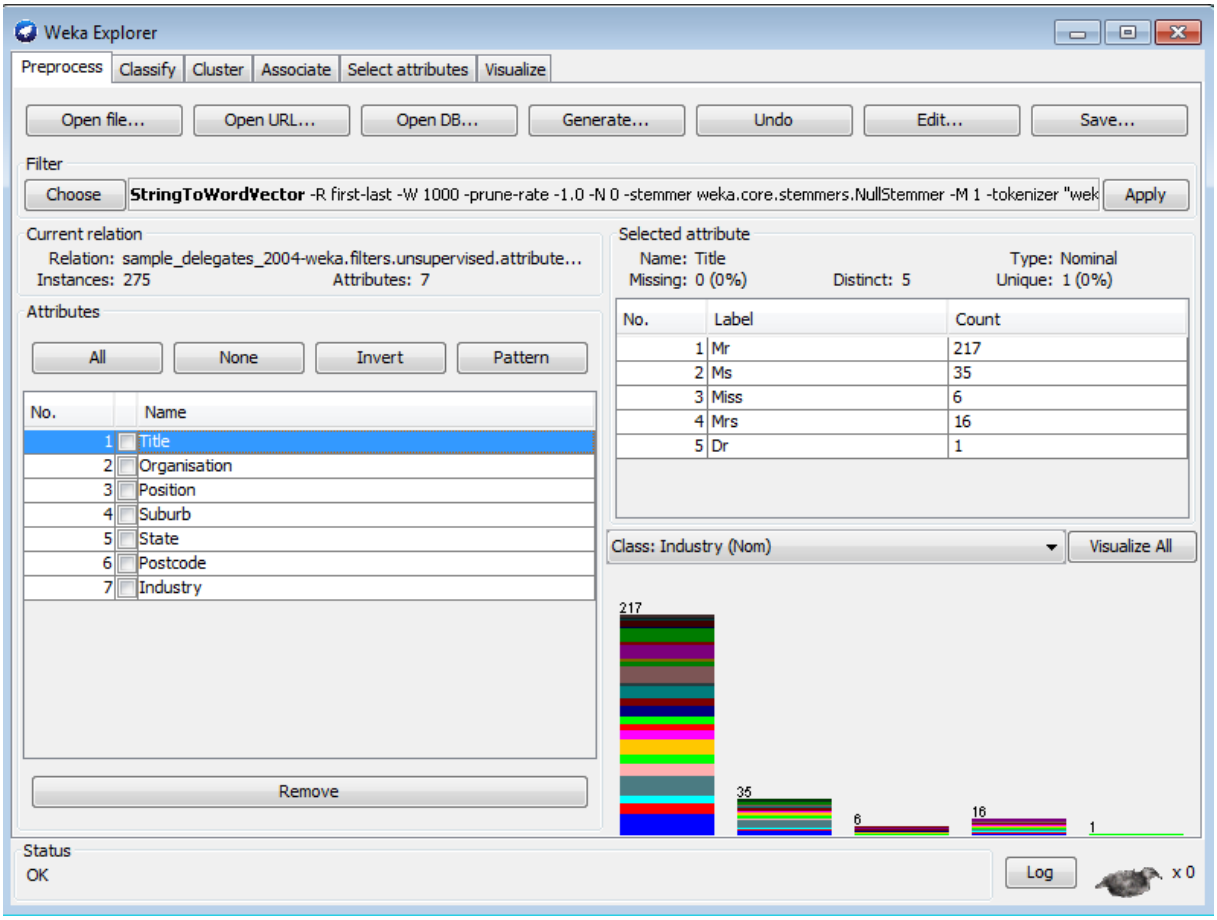


Figure 11 : filtrage de données

V.4.2 Imputation des données manquantes :

Rappelle sur les données manquantes :

Une donnée incomplète est une donnée pour laquelle la valeur de certain attribut est inconnue, ces valeurs sont dites manquantes.

Le traitement des données avec observations manquantes est un problème concret et toujours embarrassant lorsqu'il s'agit de données réelles.

V.4.2.1 *Avec K-NN*

Effectue l'imputation avec la méthode K-NN « le plus proche voisin » utilisant un ou plusieurs approches alternatives pour traiter des données multidimensionnelles. Il s'agit notamment de méthodes basées sur l'analyse de corrélation canonique, analyse canonique des correspondances, et une adaptation multi variée de la classification forêt aléatoires de et de techniques de régression de Leo Breiman et Adele Cutler.

D'autres méthodes sont également proposées. Le package « yaimput » comprend des fonctions pour comparer les résultats de l'exécution des techniques alternatives, la détection de cibles d'imputation qui sont notamment lointain à partir des observations de référence, la détection et la correction du biais, amorçage et de renforcement d'ensemble des imputations, et les résultats de la cartographie. [53]

Pour imputer les donnés manquante, j'ai utilisé ce petit code sous le langage R :

```
library(yaImpute)

xx=read.csv(file.choose())#le fichier qui contient uniquement les données complètes

yy=read.csv(file.choose())#le fichier qui contient uniquement les données manquantes

sdn=read.csv(file.choose())#la base de données brute

refs=sample(rownames(xx),47) #reference de la table

y=yy[refs,3:9]#les colonnes qui contient des données manquantes

x=xx #les observations pour l'imputation

mal <- yai(x=x,y=y,method="random") #construction de l'objet d'imputation

impute(mal)

malImp=impute(mal,ancillaryData=y)

plot(malImp)

write.csv(malImp,file="results.csv")
```

Tel que :
library(yaImpute) : pour appeler le package d'imputation **yaImpute**
impute () : la fonction d'imputation
plot : pour tracer les graphes
La dernière ligne pour sauvegarder le résultat dans un fichier **CSV.**

V.4.2.2 *Avec la moyenne*

Le remplacement par la moyenne du sujet (RMS), tels que les données manquantes pour un sujet sont remplacées par la moyenne des réponses de ce sujet aux autres items, et ce, pour tous les sujets de l'ensemble de données.

Le remplacement par la moyenne de l'item (RMI), les données manquantes sur un item sont remplacées par la moyenne des réponses à cet item, et ce, pour tous les items-cibles de l'échelle. [54]

Pour imputer les donnés manquante, on a utilisé la méthode trouver sous Weka **« ReplaceMissingValues »**

V.5 Catégorisation des identités :

V.5.1 Avec les méthodes non symboliques :

V.5.1.1 *EM (espérance-maximisation)*

L'algorithme espérance-maximisation (en anglais Expectation-maximisation algorithm, souvent abrégé **EM**), proposé par Dempster et al. (1977), est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. [55]

- **Avant imputation**

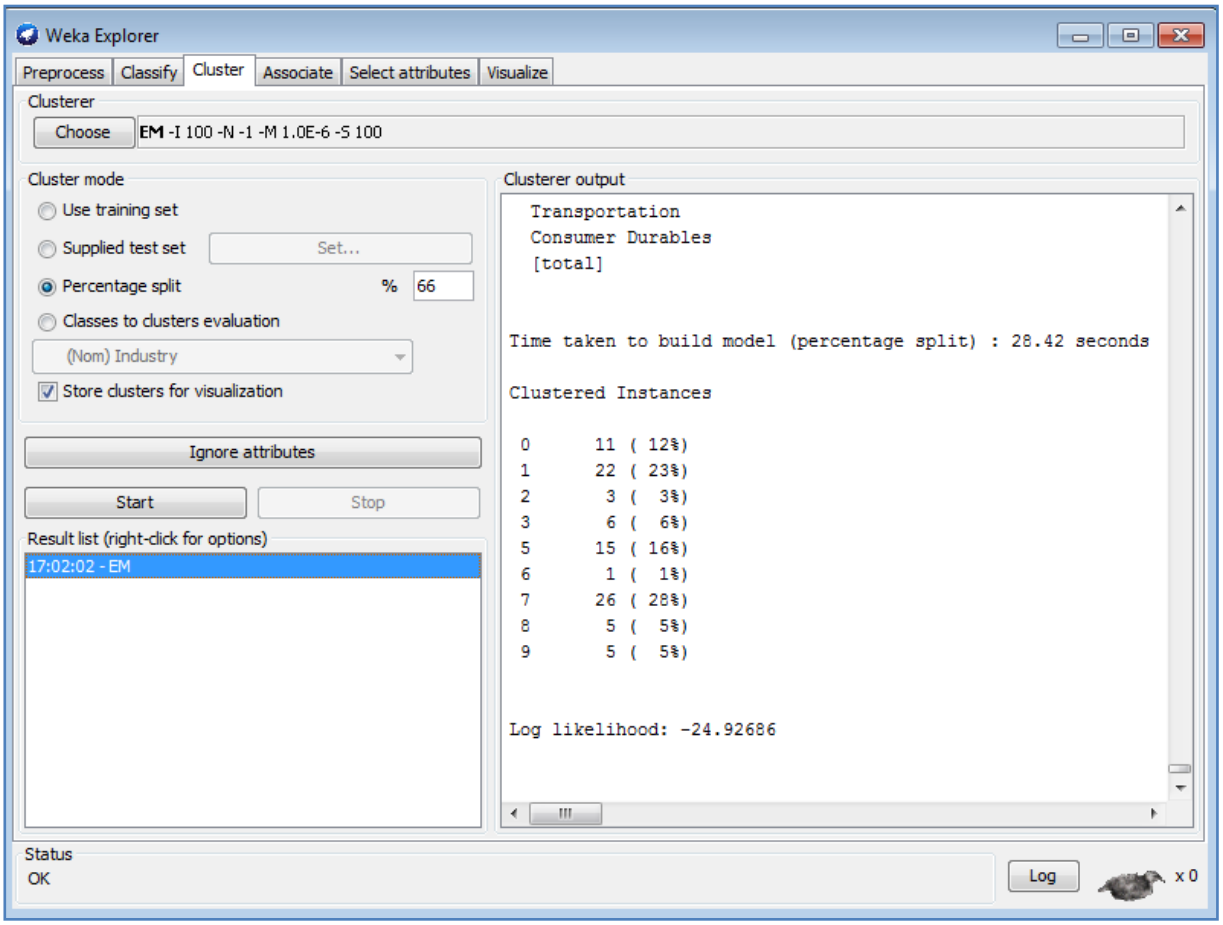


Figure 12:exécution de la méthode "EM" avant imputation

Il y a 181 instances non regroupé

- Le graphe associé

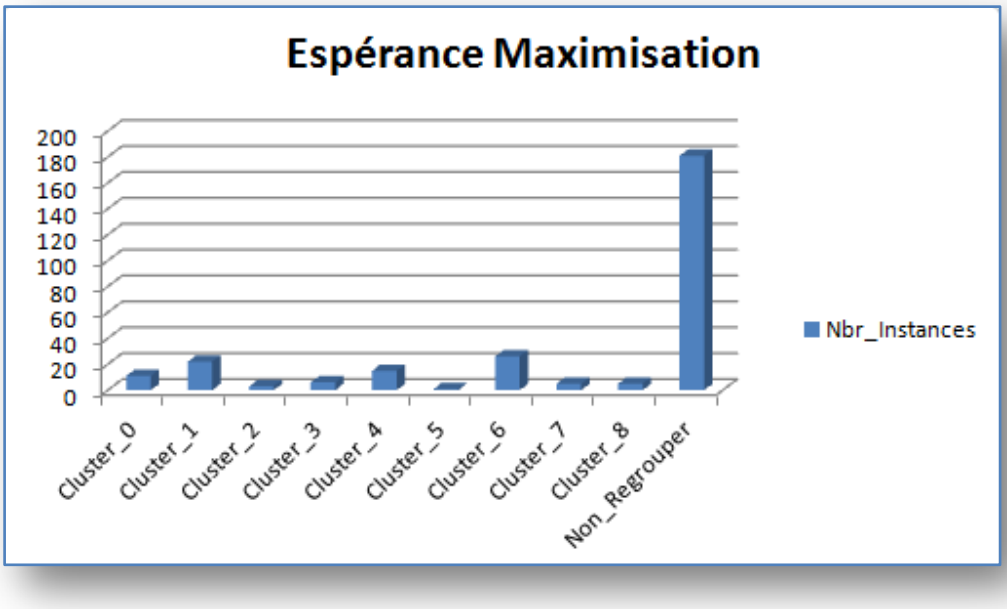


Figure 13: graphe EM avant imputation

- **Après imputation avec KNN**

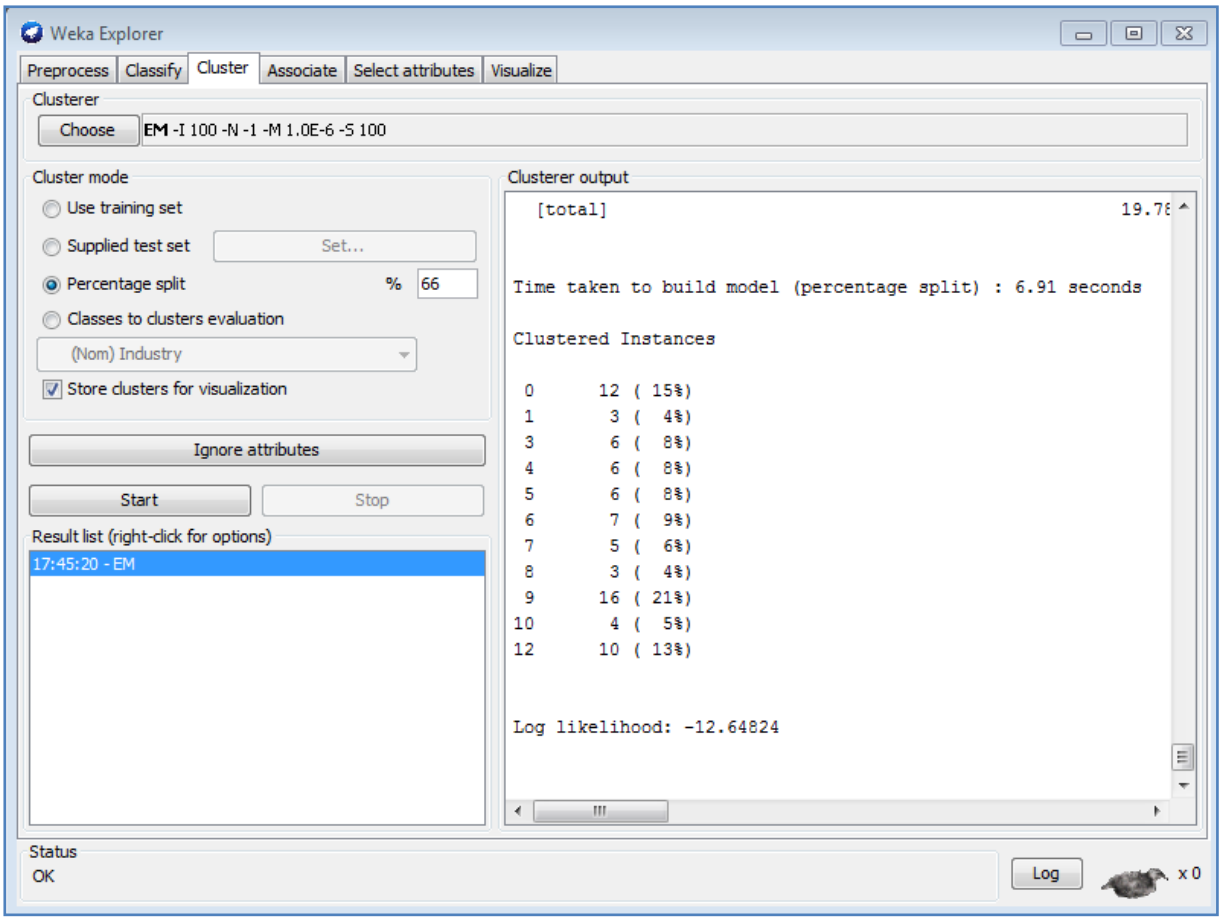


Figure 14: Exécution de la méthode "EM" après imputation avec KNN

Il y a 149 instances non regroupé

- Le graphe associé :

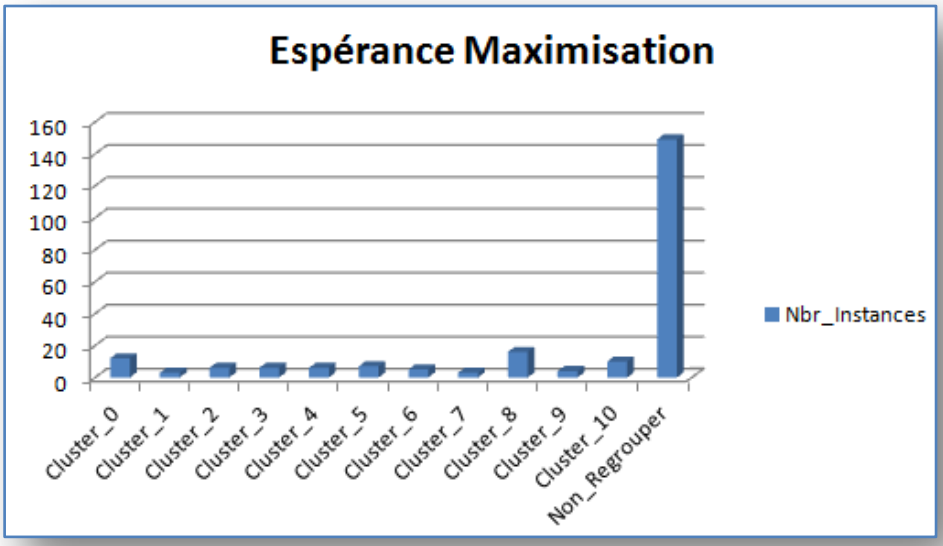


Figure 15:graphe EM après imputation avec KNN

On remarque qu'on obtient plus de clusters après imputation des données manquantes

- **Après imputation avec la moyenne**

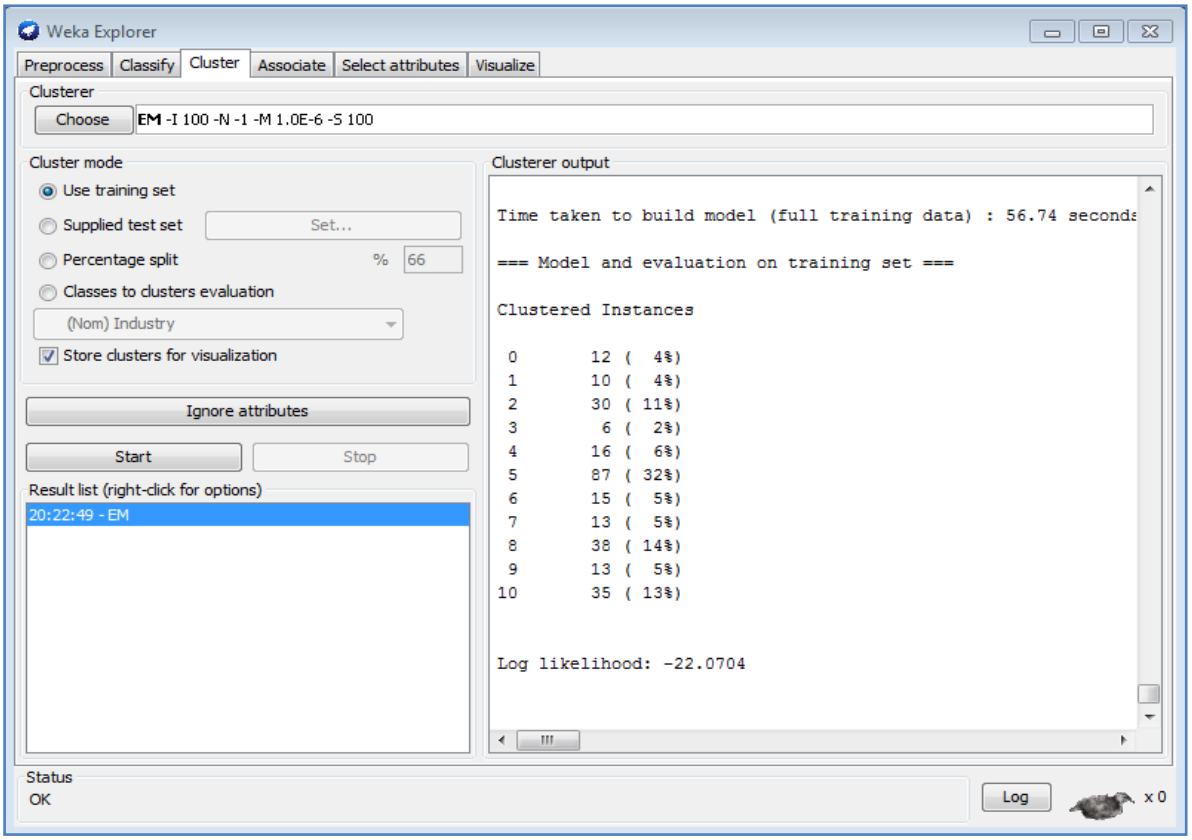


Figure 16: Exécution de la méthode "EM" après imputation par la moyenne

- Le graphe associé :

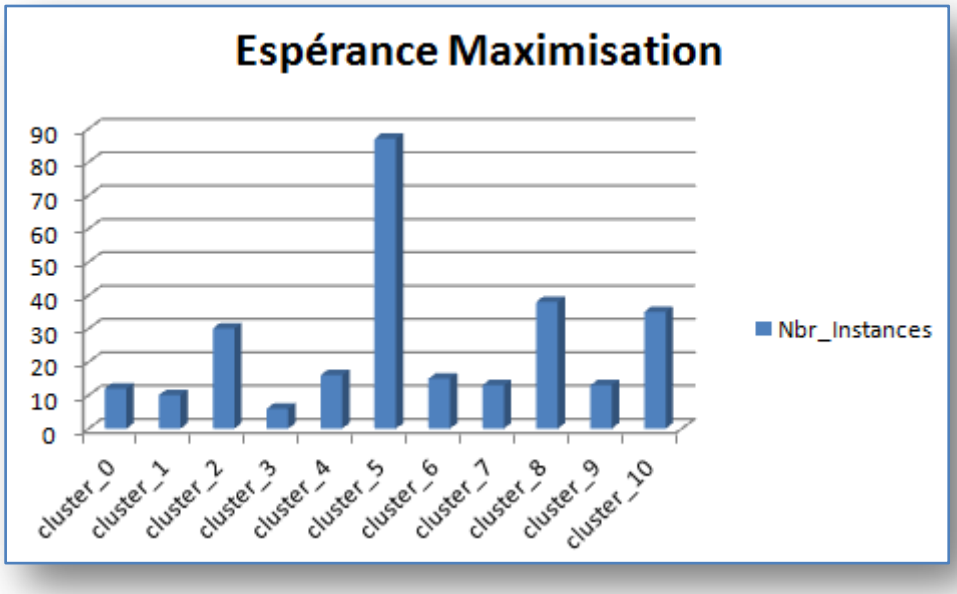


Figure 17 : graphe EM après imputation par la moyenne

On remarque que tous les instances sont regroupées

V.5.1.2 *HierarchicalClusterer (regroupement hiérarchique)*

Dans le domaine informatique, et plus précisément dans le domaine de l'analyse et de la classification automatique de données, la notion de **regroupement hiérarchique** recouvre différentes méthodes de clustering, c'est-à-dire de classification par algorithme de classification. [56]

- **Avant imputation**

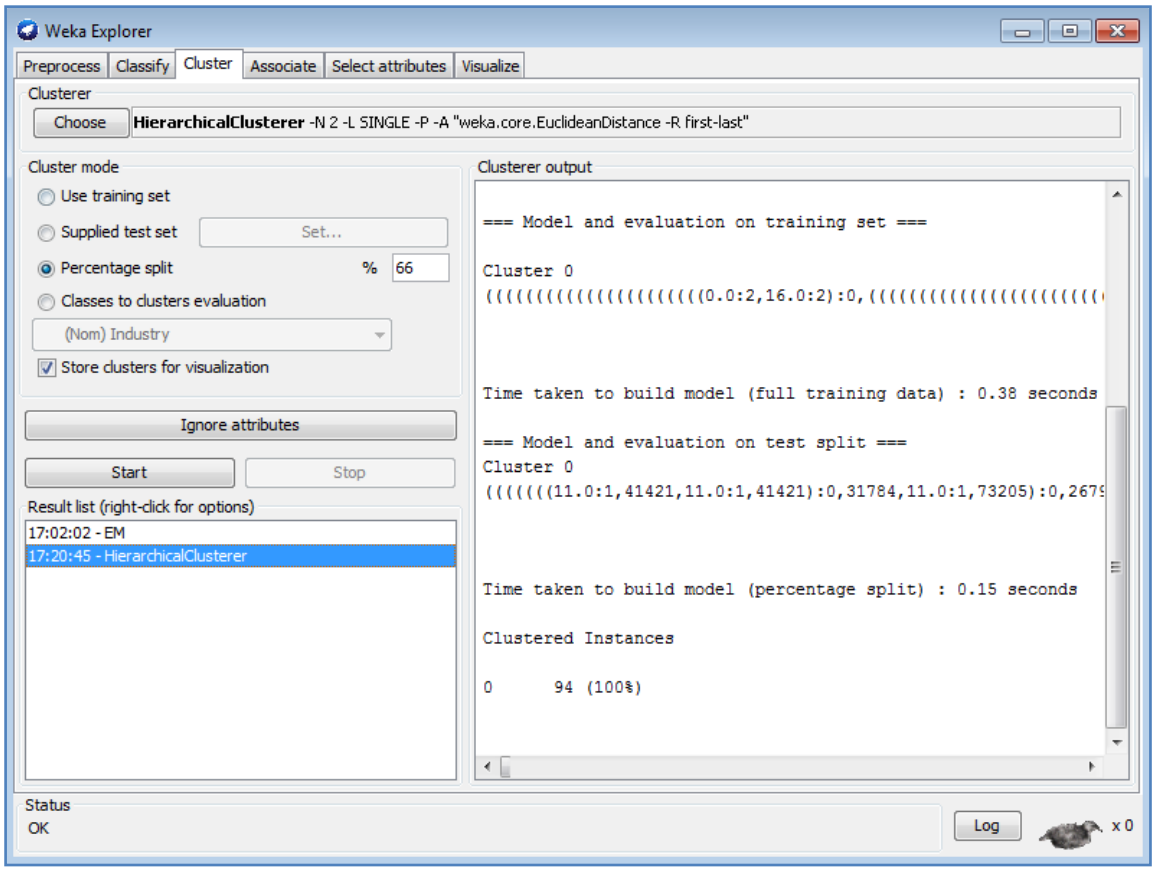


Figure 18 : exécution de la méthode "regroupement hiérarchique" avant imputation

- Le graphe associé

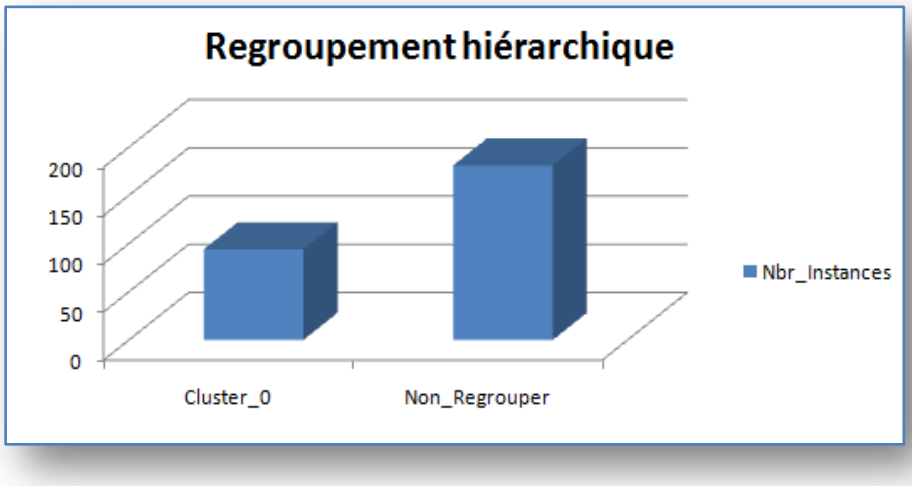


Figure 19 : graphe regroupement hiérarchique avant imputation

- **Après imputation avec KNN**

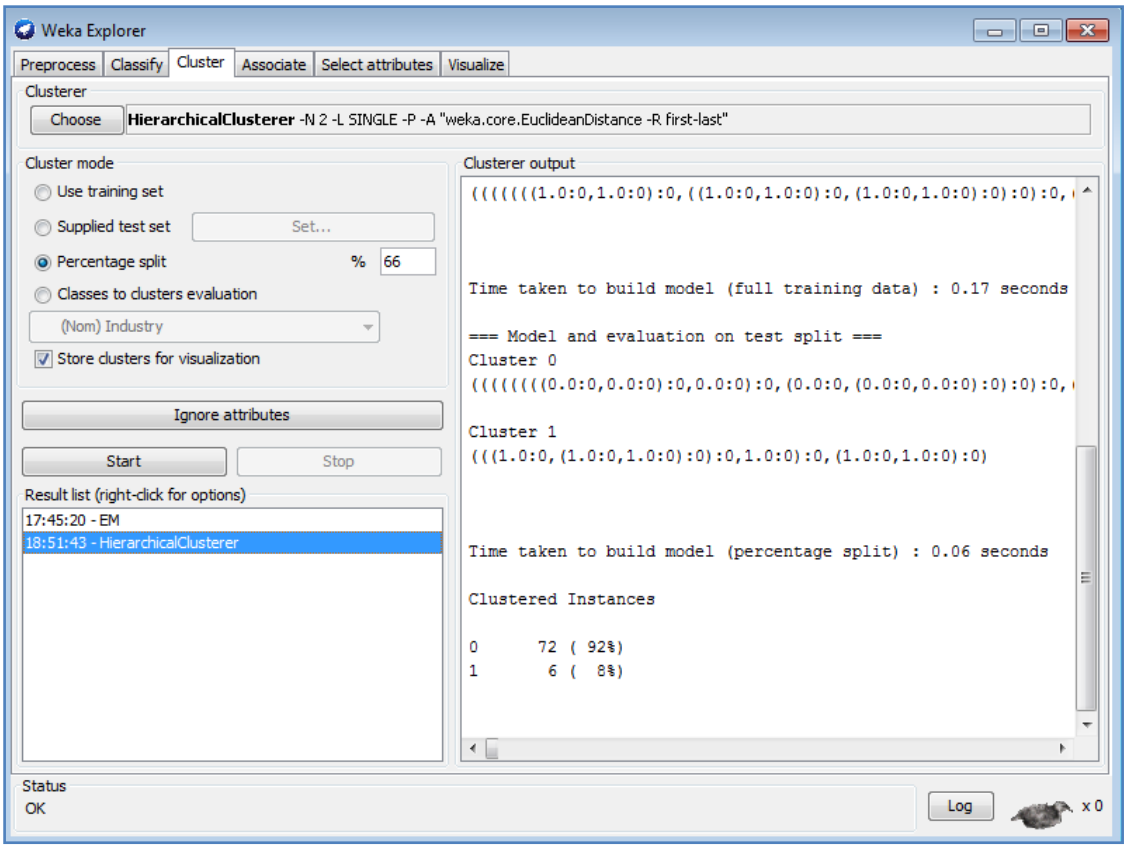


Figure 20 : Exécution de la méthode "regroupement Hiérarchique" après imputation avec KNN

- Le graphe associé :

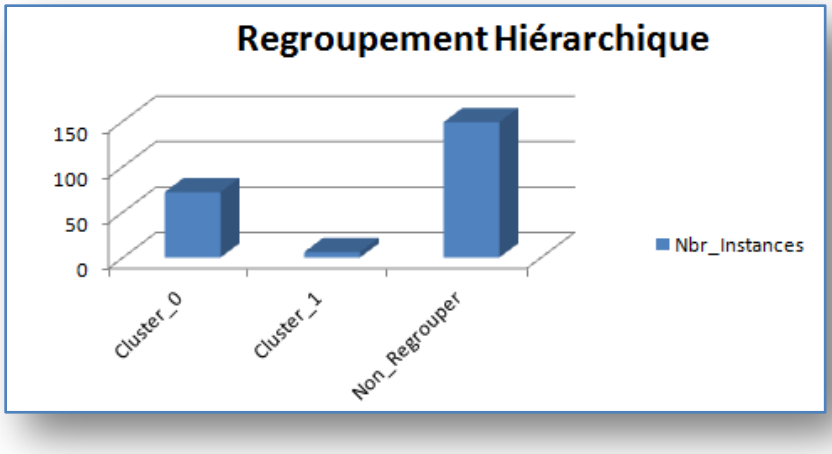


Figure 21 : graphe « Regroupement Hiérarchique » après imputation avec KNN

On remarque qu'on obtient plus de clusters après imputation des données manquantes

➢ **Après imputation avec la moyenne**

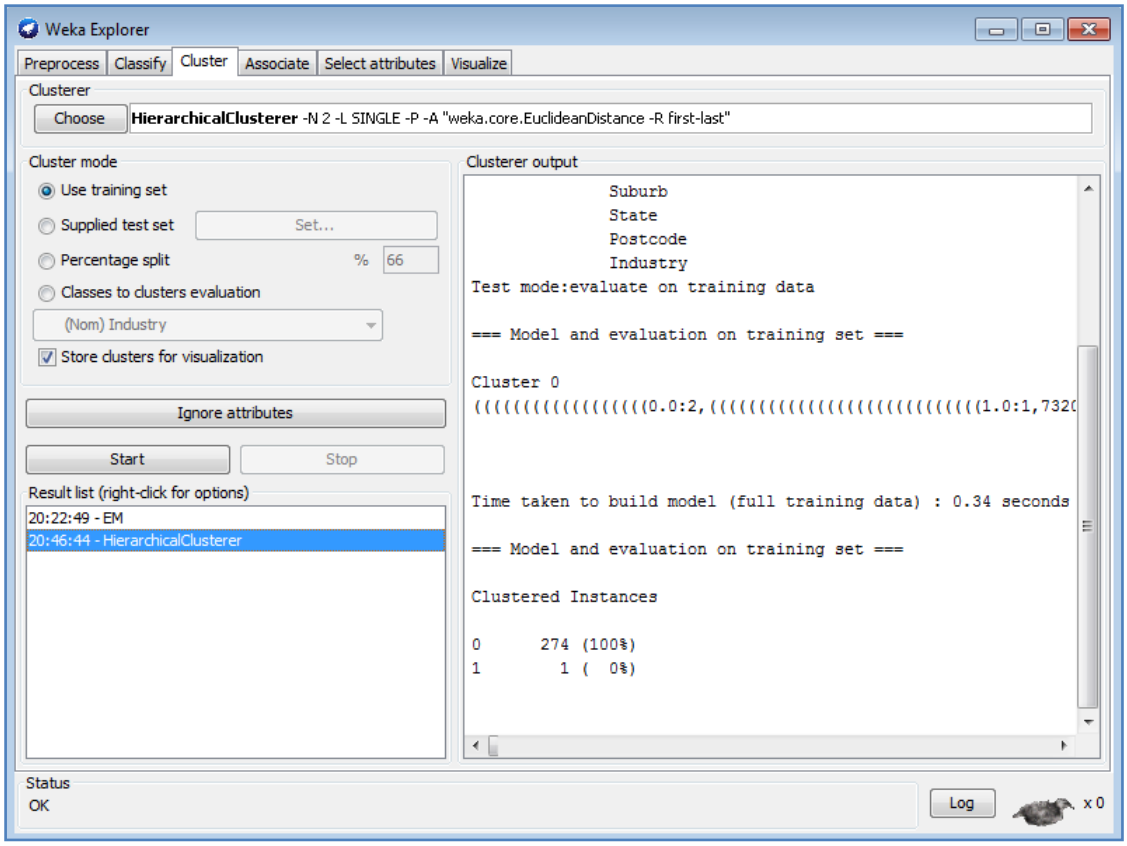


Figure 22 : Exécution de la méthode "regroupement Hiérarchique" après imputation par la moyenne

- Le graphe associé :

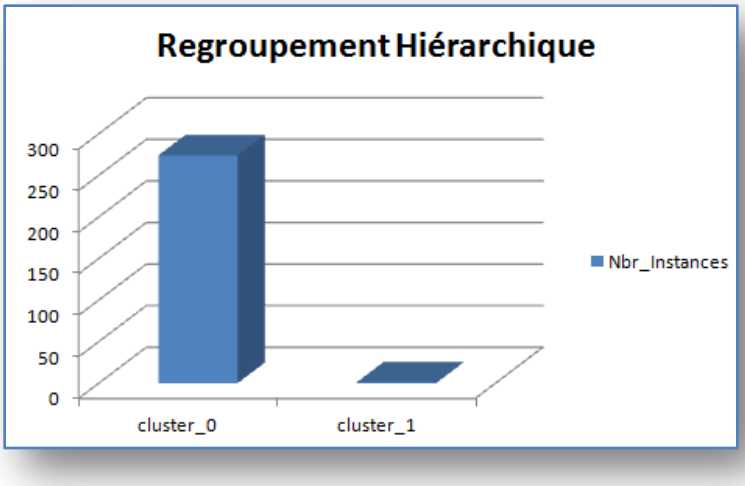


Figure 23: graphe de Regroupement Hiérarchique après imputation par la moyenne

V.5.1.3 *SimpleKMeans :*

Regrouper les données en utilisant le K-Means

L'**algorithme des k-moyennes** (ou **K-means** en anglais) est un algorithme de partitionnement de données relevant des statistiques et de l'apprentissage automatique (plus précisément de l'apprentissage non supervisé). C'est une méthode dont le but est de diviser des observations en *K* partitions (*clusters*) dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche. [57]

- **Avant imputation**

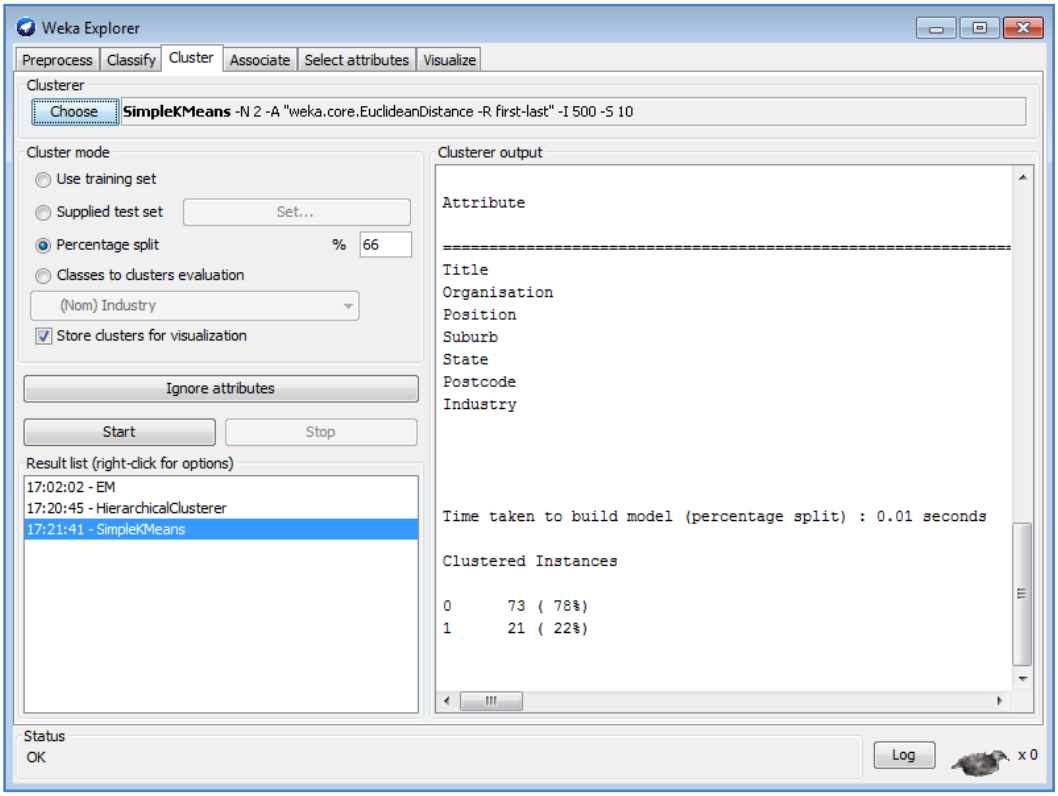


Figure 24: exécution de la méthode " SimpleKMeans " avant imputation

- Le graphe associé

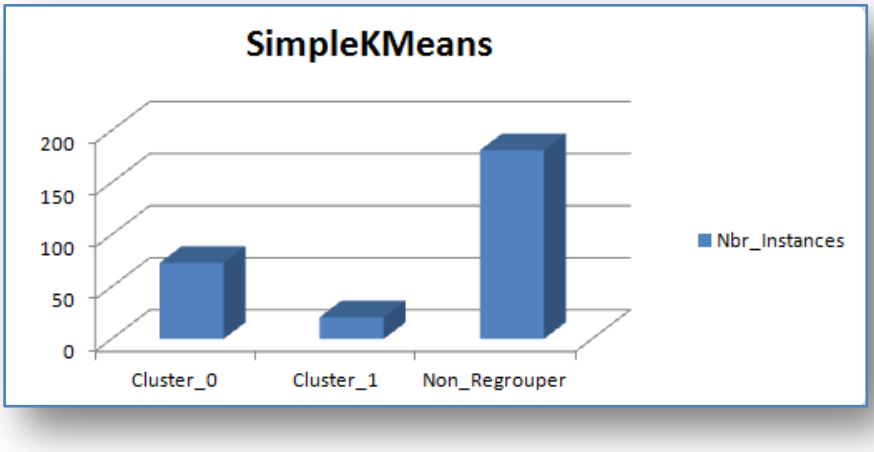


Figure 25 : graphe SimpleKMeans avant imputation

➢ **Après imputation avec KNN**

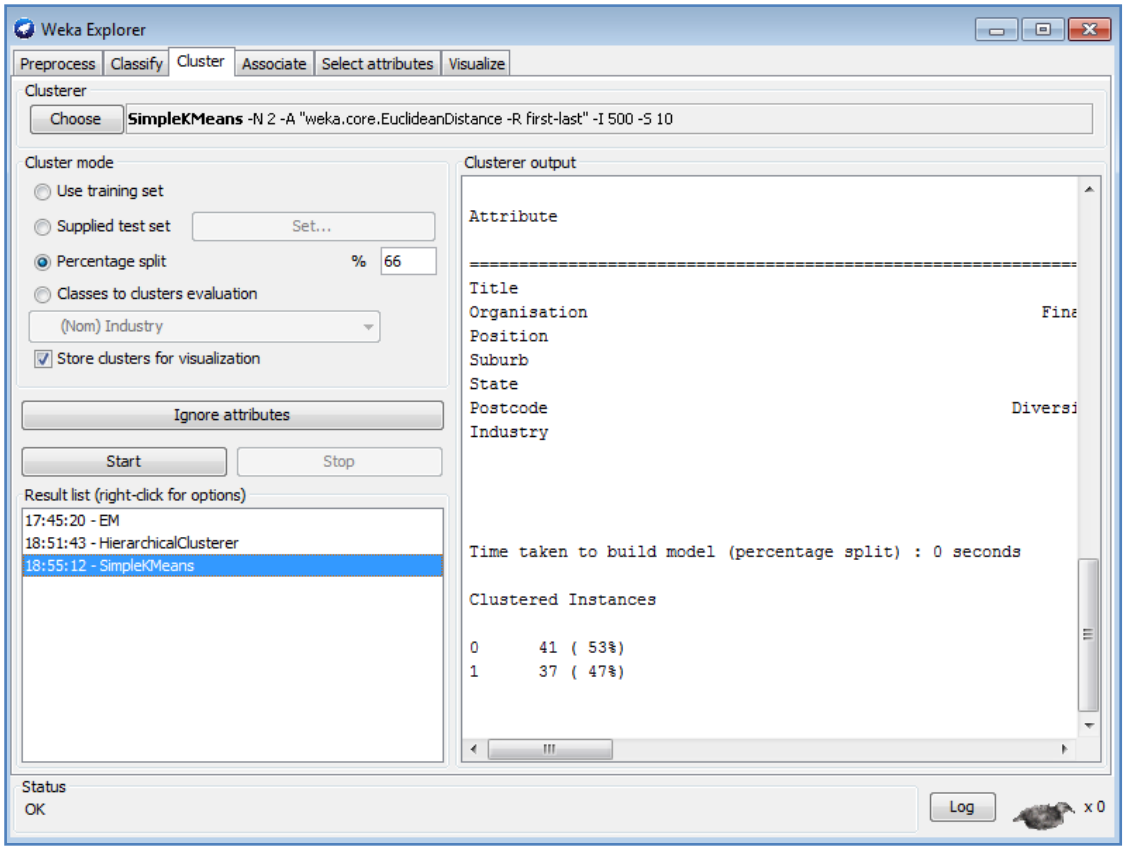


Figure 26 : Exécution de la méthode "SimpleKMeans" après imputation avec KNN

- Le graphe associé :

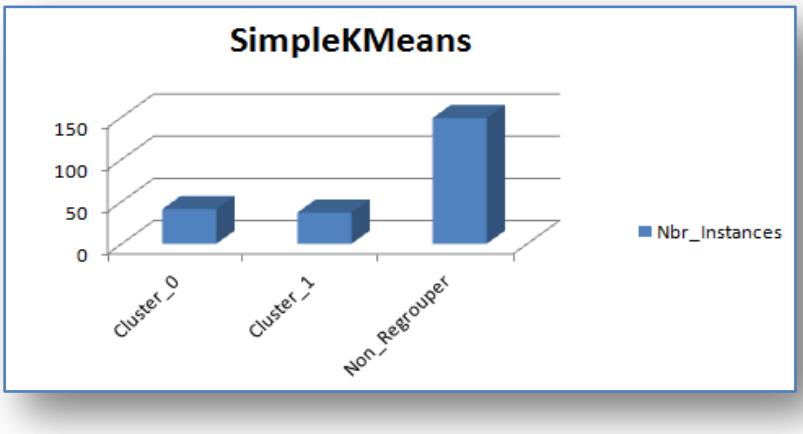


Figure 27: Graphe SimpleKMeans après imputation avec KNN

On remarque que le nombre de clusters après imputation des données manquantes est resté le même.

➢ **Après imputation avec la moyenne**

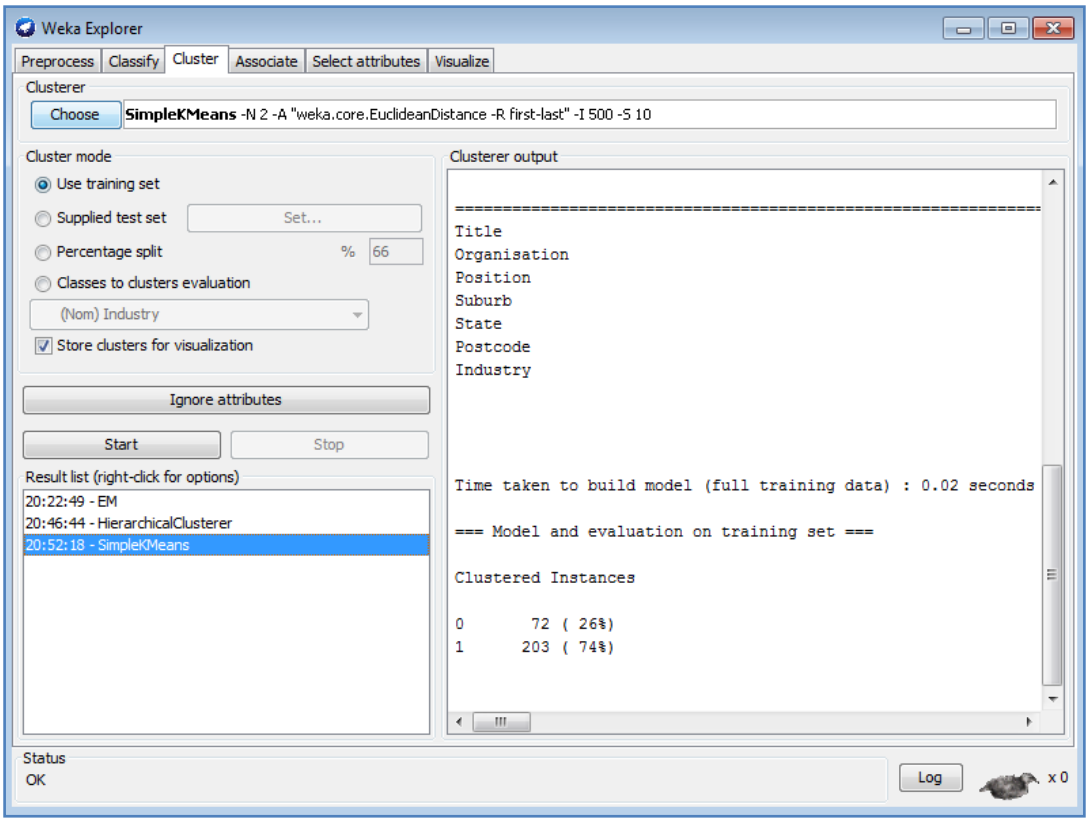


Figure 28: Exécution de la méthode "SimpleKMeans" après imputation par la moyenne

Le graphe associé :

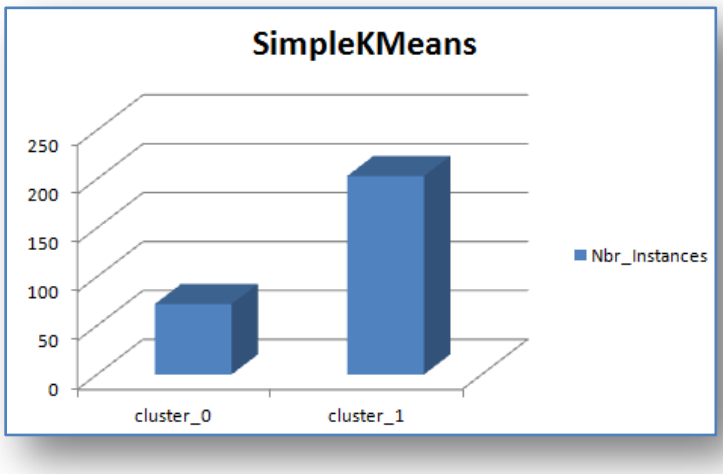


Figure 29 : Graphe SimpleKMeans après imputation par la moyenne

Après l'imputation des données manquantes avec les deux différentes méthodes, on obtient des résultats différents par rapport au premier résultat. Tels que au premier cas il ya 181 instances non regrouper, par contre dans le seconde où on a utilisé le KNN il ya 149 instances non regrouper, et le troisième cas où on a utilisé la moyenne on remarque que tous les instances sont regroupés.

V.5.2 Avec les méthodes symboliques :

V.5.2.1 *Déploiement de l'application :*

Après la création du notre projet Java, il faut ajouter les deux bibliothèques **« JRuleEngine1.3 »** et **« jsr94-1.1 »**, ainsi que les package « org.jruleengine » et « org.jruleengine.rule ». Ensuite la création d'une classe Java contient les méthodes nécessaires, plus l'interface JFrame avec des composantes pour facilité la manipulation de l'application, ainsi que la base des règles utilisées pour faire la catégorisation.

- **L'interface finale :**

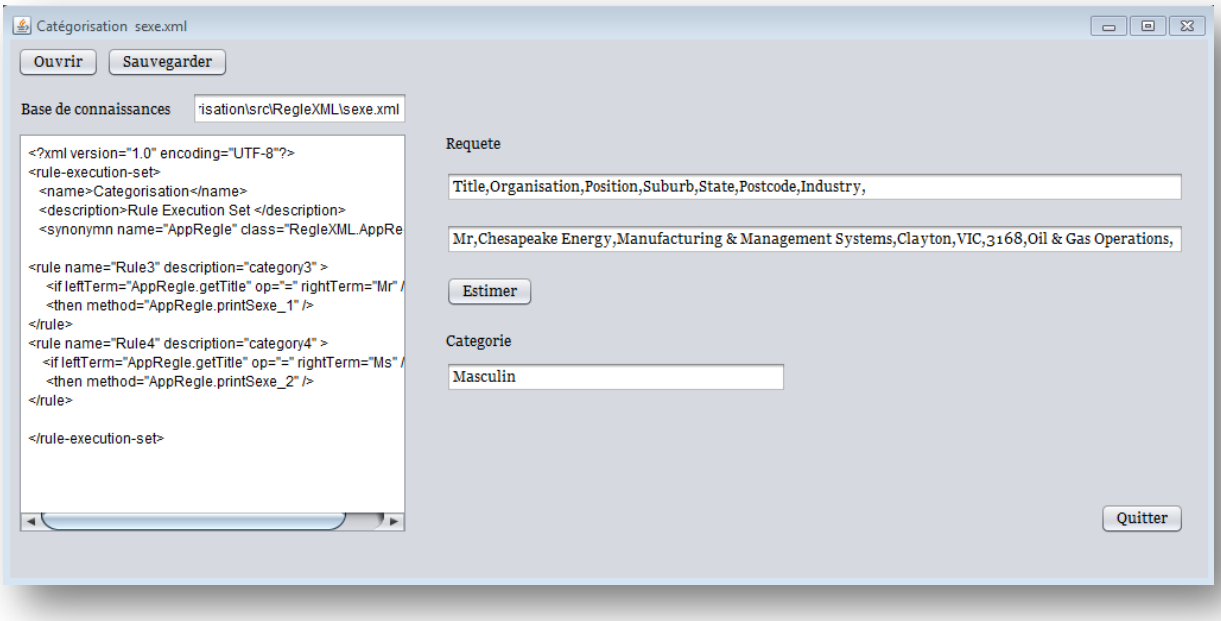


Figure 30 : l'exécution de l'application

V.6 Conclusion :

Dans ce chapitre on a réalisé une catégorisation des identités sociaux on utilisant des méthodes symbolique et non symbolique (règles XML ; EM, Regroupement Hiérarchique et SimpleKMeans).

Après ces expériences nous avons obtenus des résultats acceptable dont le nombre d'identité regrouper est augmenté après l'imputation des données manquante à l'aide de la méthode K-NN «le plus proche voisin » et la méthode d'imputation par la moyenne qui a été très fiable, elle nous a aidé de faire une bonne catégorisation des instances.

L'utilisation des méthodes symbolique (moteur de règle), à travers une interface graphique donne une bonne catégorisation, tels que l'utilisateur final de cette application obtient des résultats plus claire et interprétable par rapport aux méthodes non symboliques.

CONCLUSION GENERALE

VI. Conclusion générale

Dans ce PFE, on a conçu et réalisé une application swing contient un moteur d'inférence, cette application permet de faire la catégorisation selon certain paramètres, dans notre cas c'est la catégorisation des identités sociaux on se basant sur un fichier collecté par Eugene Dubossarsky et Mark Norrie qui contient des informations extraire a partir du réseau sociale Facebook. Et on a défini et évoqué les outils utilisés pour réaliser cette modeste application.

Le développement de l'application a vu trois étapes nécessaires

- la premier est la catégorisation des donnés avec trois méthodes non symbolique (Esperance Maximisation, Regroupement Hierarchique et Simple K-Means) sans imputation des valeurs manquant cette dernier elle donne des résultats, mais avec un grand nombre d'instances non regrouper.

- La deuxième étape elle est la même que la premier, sauf que ici on a essayé d'abord d'imputer les valeurs manquante à l'aide de deux méthodes distinct, et après faire la catégorisation :

  - La première consiste a utilisé la méthode d'imputation nommée « Plus proche voisin K-NN » à l'aide d'un package importer sous le langage de programmation « R », les résultats de la catégorisation obtenus sont différents par rapport au premier résultat et le nombre d'instances non regrouper est diminue.

  - La deuxième consiste a utilisé l'imputation de donnés par la moyenne, cette dernier se trouve sous l'onglet « Preprocess » de l'outil « weka », les donnés traiter par cette dernier donne une bonne catégorisation, tel que tous les instances sont regroupé.

- La troisième étape c'est l'utilisation des méthodes symbolique et de créer une interface graphique facile à gérer par n'import quel utilisateur. La catégorisation est faite à l'aide des Règles XML

Les résultats retournés dans la première et la deuxième étape sont difficilement interprétable, par contre l'utilisation des méthodes symbolique (moteur de règle), à travers une interface graphique donne une catégorisation compréhensible, mais on trouve une difficulté de créer des règles générales ainsi qu'on n'a pas eu assai de temps pour effectuer une catégorisation avec l'intégration des synonymes.

D'autre part, ce PFE il ma aidé à mieux comprendre le langage de programmation « JAVA », ainsi que l'utilisation de l'outil « Weka » et le langage « R » et en même temps d'avoir une idée globale sur les systèmes à base des règles et ses fonctionnalités.

Conclusion générale

En termes de perspective, je souhaite améliorer cette application telle que je vais essayé d'intègre d'autres fonctionnalités comme l'ajout d'un dictionnaire de synonyme.

Cette application est utile sur les sites de recommandation comme le e-commerce, pour cela j'essai aussi de la rendre plus efficace.

VII. Références Bibliographiques :

[1] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[2] http://lecercle.lesechos.fr/entreprises-marches/management/organisation/221172684/entreprise-20-survivra-sans-integrer-reseaux-s. Le 20/03/2014

[3]http://sites.essca.fr/blog/webmarketing/social-media-marketing/influence-des-reseaux-sociaux-sur-le-comportement-du-consommateur-852.htm. Le 20/03/2014

[4] http://blog.wacan.com/2014/03/06/actualites/tendances-2014-le-marketing-sur-les-reseaux-sociaux. Le 15/03/2014

[5] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[6] http://tempsreel.nouvelobs.com/vu-sur-le-web/20140108.OBS1730/ce-que-vous-ferez-sur-les-reseaux-sociaux-en-2014.html. LE 21/03/14

[7]http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[8] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[9] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[10] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[11] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[12] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[13] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[14] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[15] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[16] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[17] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[18]http://www.ludosln.net/web-et-medias-sociaux-etat-des-lieux-2014-video/.LE 21/03/2014

[19] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[20] http://www.w3.org/2005/Incubator/socialweb/wiki/FinalReport. Le 27/01/2014

[21]Bennane Abderrazak, «TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNEES À LA MAINTENANCE CONDITIONNELLE », document PDF, créer le 21/09/2010.

[22]Roch Giorgi, «Traitement des Données Manquantes», document PDF, créer le 04/02/2013.

[23]Stéphane Paquin, «Comparaison de quatre méthodes pour le traitement desdonnées manquantes au sein d'un modèle multiniveauparamétrique visant l'estimation de l'effet d'une intervention», document PDF, créer le 24/08/2010.

[24]Bennane Abderrazak, «TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNEES À LA MAINTENANCE CONDITIONNELLE », document PDF, créer le 21/09/2010.

[25]Roch Giorgi, «Traitement des Données Manquantes», document PDF, créer le 04/02/2013.

[26]IBM Corporation, «IBM SPSS Missing Values 20», document PDF, créer le 09/06/2011.

[27]Monbet, « Données manquantes», document PDF, créer le 13/03/2010.

[28] Flo, « Traitement des valeurs manquantes et des valeurs aberrantes», document PDF, créer le 20/09/2007.

[29]Flo, « Traitement des valeurs manquantes et des valeurs aberrantes», document PDF, créer le 20/09/2007.

[30]Gilbert Saporta et Nicolas Fischer, « fusion et greffes de Données», document Word, créer le 24/09/2008.

[31]http://www.information-mining.org/weka-data-mining/explorer. Le 15/03/2014

[32] Charles BOUVEYRON, « MODÉLISATION ET CLASSIFICATION DES DONNÉES DE GRANDE DIMENSION APPLICATION À L'ANALYSE D'IMAGES », document PDF, créer le 23/10/2006.

[33] CPC Combe de Savoie, « CATEGORISATION », document PDF, créer le 13/01/2012.

[34] http://fr.wikipedia.org/wiki/Cat%C3%A9gorisation. Le 18/02/2014.

[35] CPC Combe de Savoie, « CATEGORISATION », document PDF, créer le 13/01/2012.

[36] Bernd Amann, **«** Classification de données (partitionnement, clustering) », document PDF, créer le 27/01/2014.

[37] Bernd Amann, **«** Classification de données (partitionnement, clustering) », document PDF, créer le 27/01/2014.

[38] Bernard CONEIN, « CATEGORISATION PROFESSIONNELLE ET CLASSEMENTS SOCIAUX : UN OU DEUX SAVOIRS ? », document Word, créer le 12/10/2005.

[39] Haytham ELGHAZEL, **«** Classification et Prévision des Données Hétérogènes :

Application aux Trajectoires et Séjours Hospitaliers », document PDF, créer le 11/03/2008.

[40] Ameni Bouaziz, « Catégorisation automatique de news à l'aide de techniques d'apprentissage supervisé », document PDF, créer le 04/03/2013.

[41] http://tutoriels-data-mining.blogspot.com/2013/08/analyse-factorielle-de-donnees-mixtes.html. Le 08/03/2014.

[42] Ameni Bouaziz, « Catégorisation automatique de news à l'aide de techniques d'apprentissage supervisé », document PDF, créer le 04/03/2013.

[43] Ameni Bouaziz, « Catégorisation automatique de news à l'aide de techniques d'apprentissage supervisé », document PDF, créer le 04/03/2013.

[44] Somia RAHMOUN, « Méthodes d'apprentissage pour améliorer la QoS d'une flotte de logiciels embarqués », document PDF, créer le 15/09/2011.

[45] Brigitte Bigi, « WEKA : c'est quoi ? », document PDF, créer le 15/02/2011.

[46] Florian Boudin, « Machine Learning avec Weka », document PDF, créer le 08/01/2013.

[47] Florian Boudin, « Machine Learning avec Weka », document PDF, créer le 08/01/2013.

[48] Jean-Michel MARIN, « Initiation au logiciel R », document PDF, créer le 07/09/2005.

[49] http://fr.wikipedia.org/wiki/Netbeans. Le 29/05/2014.

[50] Alex Toussaint, BEA Systems, Inc, « Java Rule Engine API JSR-94», document PDF, créer le 10/10/2003.

[51] http://en.wikipedia.org/wiki/JSR_94. Le 03/06/2014.

[52]http://www.dbs.ifi.lmu.de/~zimek/diplomathesis/implementations/EHNDs/doc/weka/filters/unsupervised/attribute/StringToWordVector.html. Le 08/06/2014.

[53] http://cran.r-project.org/web/packages/yaImpute/index.html. Le 08/06/2014.

[54] http://theses.ulaval.ca/archimede/fichiers/23426/23426.html. Le 29/01/2014

[55] http://fr.wikipedia.org/wiki/Algorithme_esp%C3%A9rance-maximisation. Le 15/05/014

[56] http://fr.wikipedia.org/wiki/Regroupement_hi%C3%A9rarchique. Le 15/05/014

[57] http://fr.wikipedia.org/wiki/Algorithme_des_k-moyennes. Le 15/05/014

VIII. Annexe

VIII.1 Diagramme de classe associe à l'application:

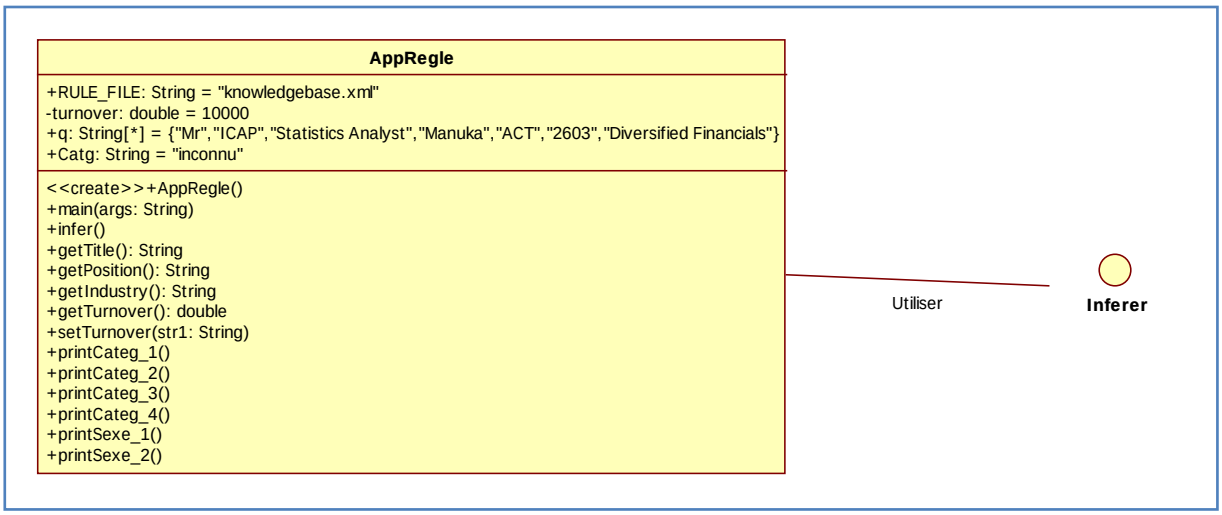


Figure 31: Diagramme de classe

VIII.2 Base de règles :

DTD pour une règle XML

Une règle peut être définir dans un fichier XML si elle respect la DTD suivantes:

```
<!ELEMENT rule-execution-set (name, description, synonymn*, rule*)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT synonymn>
<!ATTLIST synonymn name CDATA #REQUIRED>
<!ATTLIST synonymn class CDATA #REQUIRED>
<!ELEMENT rule (if*, then*)>
<!ATTLIST rule name CDATA #REQUIRED>
<!ATTLIST rule description CDATA #REQUIRED>
<!ELEMENT if >
<!ATTLIST if leftTerm CDATA #REQUIRED>
<!ATTLIST if op CDATA #IMPLIED>
<!ATTLIST if rightTerm CDATA #IMPLIED>
<!ELEMENT then >
<!ATTLIST then method CDATA #REQUIRED>
<!ATTLIST then arg1 CDATA #IMPLIED>
<!ATTLIST then arg2 CDATA #IMPLIED>
...
<!ATTLIST then argN CDATA #IMPLIED>
```

- **Knowledgebase.xml :**

```
<?xml version="1.0" encoding="UTF-8"?>
<rule-execution-set>
  <name>Categorisation</name>
  <description>Rule Execution Set </description>
  <synonymn name="AppRegle" class="RegleXML.AppRegle" />
  <rule name="Rule1" description="category1" >
    <if leftTerm="AppRegle.getPosition" op="=" rightTerm="Business
Analyst" />
    <if leftTerm="AppRegle.getIndustry" op="=" rightTerm="Aerospace et
Defense" />
    <then method="AppRegle.printCateg_1" />
  </rule>
  <rule name="Rule2" description="category2" >
    <if leftTerm="AppRegle.getPosition" op="=" rightTerm="Business
Analyst" />
    <if leftTerm="AppRegle.getIndustry" op="=" rightTerm="Media" />
    <then method="AppRegle.printCateg_2" />
  </rule>
<rule name="Rule3" description="category3" >
    <if leftTerm="AppRegle.getPosition" op="=" rightTerm="Director" />
    <if leftTerm="AppRegle.getIndustry" op="=" rightTerm="Construction" />
    <then method="AppRegle.printCateg_3" />
</rule>
<rule name="Rule4" description="category4" >

    <if leftTerm="AppRegle.getPosition" op="=" rightTerm="Director" />
    <if leftTerm="AppRegle.getIndustry" op="=" rightTerm="Restaurants et
Leisure" />
    <then method="AppRegle.printCateg_4" />
</rule>
</rule-execution-set>
```

sexe.xml :

```
<?xml version="1.0" encoding="UTF-8"?>
<rule-execution-set>
  <name>Categorisation</name>
  <description>Rule Execution Set </description>
  <synonymn name="AppRegle" class="RegleXML.AppRegle"
/>
<rule name="Rule3" description="category3" >
    <if leftTerm="AppRegle.getTitle" op="=" rightTerm="Mr" />
    <then method="AppRegle.printSexe_1" />
</rule>
<rule name="Rule4" description="category4" >
   <if leftTerm="AppRegle.getTitle" op="=" rightTerm="Ms" />
    <then method="AppRegle.printSexe_2" />
</rule>
</rule-execution-set>
```

IX. Acronyms

IOS : système d'exploitation mobile développé par Apple

FBML: Facebook Markup Language

OAuth: open protocol to allow secure authorization

TLS: Transport Layer Security

HTTPS: HyperText Transfer Protocol Secure

URI: Uniform Resource Identifier

XAuth: Extra Authentication

SAML: Security Assertion MarkupLanguage

OASIS: Organization for the Advancement of Structured Information Standards

RDFa : Resource Description Framework dans des Attributs

XRD : Extensible Description des ressources

IETF: Internet Engineering Task Force

FOAF: friend of a friend, ami-de-un-ami

SIOC: SemanticallyInterlinked Online Communities

SWXG: Incubateur Groupe Web social

P3P: Platform for PrivacyPreferences

POWDER : protocole pour la Description des Ressources du Web

AIR : AMORD In RDF

FRR *:* Fonds de réserve pour les Retraites

XMPP: Extensible Messaging and Presence Protocol

RFC: requests for comments

ARIA: Accessible Rich Internet Applications

API: Application Programming Interface

IMI: IdentityMetasystemInteroperability

OWL: Web Ontology Language

Acronymes

ODS: Open Document Spreadsheet

ACL : Access Control List

MAR : Manquant au hasard

MCAR : Manquant complètement au hasard

NMAR : ne manquant pas au hasard

EM: prévision-maximisation

Weka Waikato Environment for Knowledge Analysis

INSEE : Institut national de la statistique et des études économiques

TAL : Traitement Automatique des Langues

kNN : k nearest neighbor

LLSF: linear least square fit

NB: Naive Bayes

CSV: Comma Separated Value

ARFF: Attribute-Relation File Format