Sujet de thèse :

Étude des impacts respectifs entre le Big Data et le Datamining

UMI 209, UMMISCO. Université Cheikh Anta Diop à Dakar (UCAD)

2016-2019

1 Thème

1.1 Pistes à étudier.

La définition initiale donnée par le cabinet McKinsey and Company en 2011 s'orientait d'abord vers la question technologique, avec la célèbre règle des 3V : un grand Volume de données, une importante Variété de ces mêmes données et une Vitesse de traitement s'apparentant parfois à du temps réel. Puis, ces qualificatifs ont évolué, avec une vision davantage économique portée par le 4ème V de la définition, celui de Valeur, et une notion qualitative véhiculée par le 5e V, celui de Véracité des données [3]. Le datamining est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données [2]. L'objectif principale de ce travail est d'identifier les impacts mutuels entre le Big Data et le Datamining et de faire évoluer ces impacts dans le sens positif :

- Hadoop qui est l'un des standards informatiques de la gestion de données massives (i.e. Big Data). Apache Mahout est un projet de la fondation Apache visant à créer des implémentations d'algorithmes d'apprentissage automatique et de datamining. L'architecture d'Hadoop intègre dans sa partie accès aux données des aspect d'analyse incluant l'apprentissage automatique (i.e. le Datamining) à travers l'outil Mahout . En effet, Mahout est également un projet de la fondation Apache. C'est une Collection de méthodes programmées en java et destinées au machine learning sur des architectures avec mémoire distribuée, comme Hadoop. Dans cette partie vous allons étudier les outils et méthodes issues du Datamining utilisé pour tirer partie des grandes masses de données collectées dans le contexte du Big Data.
- L'étude de l'influence du Big Data sur le Datamining sera organisé selon les phases du processus de Datamining correspondant au modèle CRISP-DM [1] :
  1. Phases de compréhension et préparation des données :
     (a) Traitement des données manquantes
     (b) Traitement des données aberrantes
  2. Phase de modélisation (incluant validation de modèle)
     (a) Scalabilité (i.e. traitement d'une grande quantité de données)
     (b) Nécessite de Mise à jour du modèle
     (c) Possibilité de Mise à jour du modèle
     (d) Quantité minimum de données (case to IV [1] ratio),
     (e) Indépendance des observations,
     (f) Variance dans le calcul de l'erreur.

1.2 Partie pratique : Outil d'aide à la décision pour le diagnostic médicale basé sur l'historisation des données des patients

L'applicabilité de la méthodologie proposée sera testée sur une application d'aide à la décision pour le diagnostic médicale basé sur l'historisation des données des patients. Cette application devra être conçu et implémenté au cours de la thèse. En plus de la validation de la méthodologie, des objectifs pratiques sont attendues de l'application dans les deux axes suivants :

1. Historisation des données des patients. La disponibilité des données sur les patients consultable par le médecin permettra
   (a) valider les réponses à certaines questions du médecin concernant les informations qui sont sensés ne pas varier,
   (b) voir l'évolution de certaines indicateurs après les réponses à certaines questions du médecin sur des informations qui sont variant dans le temps,
   (c) gagner du temps car toutes les questions ne peuvent pas être posées faute de temps (e.g. antécédents personnels, antécédents familiaux, sensibilité particulière à certains composants d'un médicament car le patient y est allergique)

1. Ratio observations - Variables Indépendantes

(d) éviter la répétition de certains examens médicaux qui ne sont pas nécessaires. En effet, il peut arriver qu'un patient refasse un examen qu'il a déjà fait et dont le résultat est encore valide pour les raisons suivantes :

i. le patient ne sait pas que le résultat du premier examen est encore valide,

ii. le médecin n'a pas connaissance qu'un premier examen a été fait dans l'hôpital même ou dans une autre structure médicale.

2. Aide à la décision pour le diagnostic médicale. L'outil permettra de proposer ou de valider un diagnostic médical. La fonction du datamining qui sera privilégié est l'apprentissage supervisé avec comme variable de réponse la maladie.

2 Comité de suivi de thèse

2.1 Directeurs de thèse

1. Directeur de thèse : Dr Alassane BAH, Maître de Conférences Informatique (CAMES), Directeur du centre UCAD de l'UMI 209, UMMISCO.

2.2 Membres du comité

1. Dr Mamadou S. CAMARA, Maître-Assistant, Ecole Supérieure Polytechnique, UCAD.
2. Dr Mandicou BA, Assistant, Ecole Supérieure Polytechnique, UCAD.
3. Dr Ibrahima FALL, Maître-Assistant, Ecole Supérieure Polytechnique, UCAD.

3 Profil recherché

La candidat devra être titulaire d'un diplôme de Master 2 ou d'un diplôme d'ingénieur de conception dans une filière Informatique. Le candidat devra également posséder un excellent dossier scolaire et de bonnes connaissances en datamining, bases de données, développement logiciel et datawarehousing.

4 Planning

La candidat devra suivre le planning suivant qui est organisé selon les trois années prévues pour le déroulement de la thèse (2016 - 2019).

Phases	Etapes	Période
Etat de l'art	- Datamining - Relation entre Datamining et BigData - Datamining dans le diagnostic médical	1ère année
Proposition méthodologie	- Méthode pour une influence positive entre Datamining et BigData - Conception de l'outil de collecte et d'historisation des données des patients - Application du processus de Datamining	2ème année
Rédaction de la thèse et Soutenance	- Rédaction de la thèse sur la base des résultats de recherche - Procéder à la soutenance	3ème année

Références

[1] *CRISP-DM 1.0 : Step-by-step data mining guide.*

[2] *Data mining et statistique décisionnelle : l'intelligence des donnéees.* Editions Ophrys, 2007.

[3] YIN SHEN and OKYAY KAYNAK. Big data for modern industry : Challenges and trends. *Proceedings of the IEEE*, 2015.