ACO-FFDP : Approche de clustering incrémental basée sur la théorie des fourmis

Fadwa BOUHAFER, Anass EL HADDADI and Mohammed HEYOUNI
UMP, ENSA Al-Hoceima, Maroc

***Abstract*—Le projet XEW-WP1.S3 vise à intégrer et à mettre à disposition sur la plateforme XEW des outils d'analyse incrémental plus fine du contenu générer par le service de sourcing. L'objectif de l'incrémentalité est de nous faciliter l'interprétation et la visualisation de l'évolution d'un domaine scientifique ou technique, d'une alliance stratégique, d'une séquence temporelle, d'un réseau social ou sémantique. Nous proposons une approche de construction incrémentale des hypergraphes (grands graphes) par des fourmis artificielles qui s'inspire du comportement d'auto-assemblage de fourmis réelles.**

***Index Terms*—Big Data Mining, Clustering incrémental, Graph Clustering, théorie des fourmis.**

I. INTRODUCTION

Les données ont toujours été le moteur d'une réflexion profonde. De tout temps, les grandes organisations ont exploité les données pour identifier et saisir les opportunités de marché plus vite que leurs concurrents. Dans l'univers de Big Data Mining, elles ont également joué un rôle de premier plan de transformation des processus métier clés et la création de nouvelles opportunités de monétisation. Grâce à la richesse des sources de données du web, des réseaux sociaux, des objets connectés, des appareils mobiles, des capteurs et de la télémétrie, le Big Data Mining nous offre de nouvelles perspectives sur les clients, les produits, les opérations et les marchés. Les grandes organisations s'en servent pour réorganiser leurs processus de création de valeur, se différencier de la concurrence et proposer une expérience client pertinente et rentable.

II. BIG DATA MINING: CROISSANCE EXPLOSIVE DES DONNEES

La quantité d'informations, que les entreprises doivent gérer et surveiller, ne cesse pas d'augmenter, chaque seconde, 29.000 Giga-octets (Go) d'informations sont publiés dans le monde, soit 2,5 exaoctets par jour soit 912,5 exaoctets par an, 170 terabytes des pages web, 35.000.000.000/jour (400.000 Terabytes / an) des e-mails, 17,3 Exabytes / an de SMS, plus de 5000 banques de données professionnelles (Brevets, articles scientifiques et techniques, données économiques, etc.), l'émergence du web 2.0, 3.0 et 4.0. Un volume de « Big Data » qui croît à une vitesse vertigineuse et donne naissance à de nouveaux types de statistiques et d'approche de data mining. On ajoute à cela, la page web dédiée sur Internet en temps réel réalisé par Penny Stock Lab. Celle-ci reprend le trafic et le nombre de données qu'il y a sur de nombreux sites connus (Twitter, Facebook, Youtube, App Store, Play Store, Pinterest, Netflix…) et cela est assez impressionnant en 60 secondes, alors imaginez en une journée, un mois, un an !

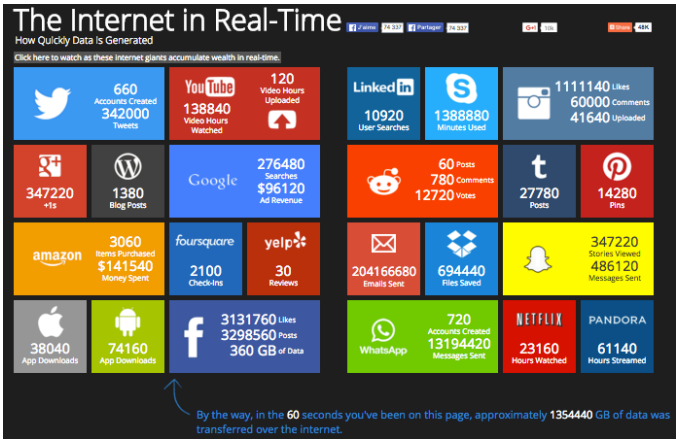

Fig. 1. Transfert de données sur Internet en temps réel (http://pennystocks.la/internet-in-real-time/ consulté le 29/02/2017 à 12 :18 GMT)

Ces données sont le moteur du mouvement Big Data, comme nous la montre la carte de l'histoire de Big Data (figure 2). L'objectif de cette carte est d'offrir une représentation graphique et métaphorique des meilleures pratiques nécessaires à la création d'une stratégie de Big Data Mining réussie.

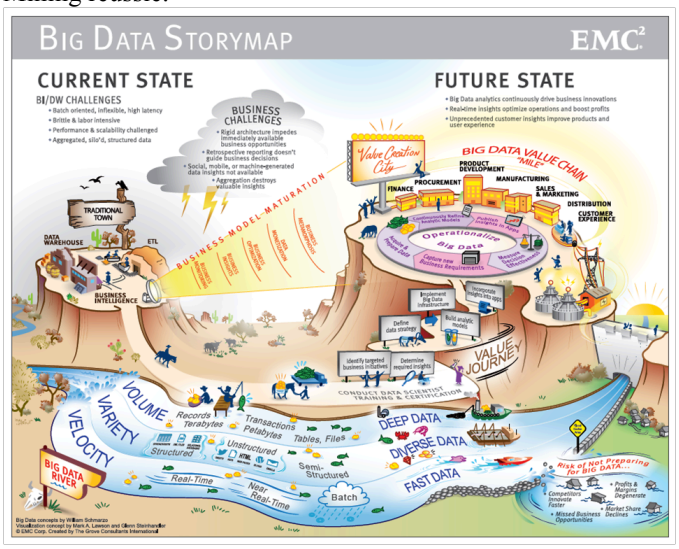

Fig. 2. Carte de l'histoire du Big Data [1]

Le Big Data Mining est profond et pertinent, rapide et puissant. Il nous offre de nouvelles perspectives grâce à sa capacité à :

- Fouiller les données des réseaux sociaux, des appareils mobiles et autre pour découvrir les centres d'intérêt, les passions, les associations et les affiliations des clients.
- Analyser les données provenant des machines, des capteurs et de la télémétrie pour prendre en charge la maintenance prévisionnelle, améliorer la performance des produits ou optimiser le fonctionnement de réseaux.
- Exploiter les perspectives comportementales pour créer une expérience utilisateur plus attrayante.

Les organisations commencent à apprécier les données, et à développer des processus pour les capturer, les gérer et les acquérir. Elles apprennent donc à les traiter comme un actif et non comme un coût. Peu à peu, elle saisissent l'avantage concurrentiel offert par leurs modèles analytiques et leurs perspectives, et gèrent ces analyses comme une propriété intellectuelle qui doit être capturée, affinée, réutilisée et, dans certains cas légalement protégés. Elles commencent à adopter et à cultiver une culture pilotée par les analyses ou les données en les laissant guider leurs prises de décisions au lieu de se fier à l'avis de la personne la plus expérimentée, comme c'est habituellement le cas (figure 3).

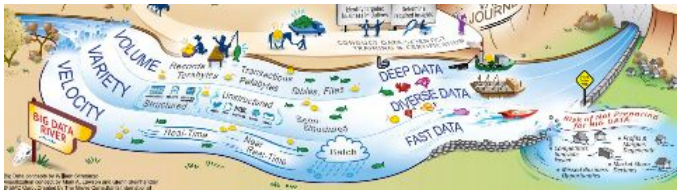

Fig. 3. Une croissance explosive des données et vectrice de données métier

III. LE CLUSTERING

A. *Définition du clustering*

Les méthodes de clustering permettent d'analyser un grand nombre d'objets ou d'individu qu'on cherche à répartir en catégories de façon non supervisées [2]. Elles visent des familles d'individus homogènes selon un critère donné. On peut utiliser ces méthodes dans plusieurs cas et sur plusieurs domaines.

Le clustering est typiquement un cas d'apprentissage non supervisé. La propriété des observations aux familles des données de sortie n'est pas connue à priori.

A savoir que le clustering chez les statisticiens est considéré comme des approches supervisées où l'on affecte des objets à des classes identifiées au préalable. Cependant on parle des techniques de classement qui sont très connotées.

Depuis les années 1970, le clustering est une branche extrêmement prolixe et diversifiée de l'analyse de donnes, qui n'a eu de cesser de se développer et de se perfectionner.

Le problème de base est classique : pour extraire de manière automatique des groupes nommé « clusters », on distingue deux primordiales approches que nous détaillerons par la suite: le clustering hiérarchique et le clustering non hiérarchique.

Dans les deux cas, un bon clustering devra permettre de produire des groupes avec :

- Une forte similarité intraclasse (ou faible variabilité intraclasse) ce qui signifie que les individus d'un groupe donné doivent se ressembler.
- Et une faible similarité interclasses (ou forte variabilité interclasse), ce qui signifie que les individus de groupes distincts ne doivent pas se rassembler.

B. *Le clustering hiérarchique*

Pour construire un clustering hiérarchique [3], on utilise deux principaux algorithmes :

- Les algorithmes ascendants construisent des classes par agglomérations successives des objets deux à deux.
- Les algorithmes descendants réalisent des dichotomies progressives de l'ensemble des objets.

Ces deux algorithmes permettent de construire un arbre appelé dendrogramme qui rassemble des individus de plus en plus dissemblables au fur et à mesure qu'on s'approche à la racine de l'arbre.

La méthode agglomérative est la plus populaire, les regroupements sont représentés par le dendrogramme. Celui-ci une hiérarchie indicée : chaque niveau a un indice qui représente de manière générale la distance entre ses éléments.

La distance la plus intuitive, communément utilisée, est la distance euclidienne standard soit une matrice X à n variables quantitatives. Les données étant centrées réduites.

D'autres distances peuvent être utilisées selon les spécificités des données et l'objectif de la classification. Par exemple la distance de Khi-deux, la distance de Manhattan…

Par rapport à la distance euclidienne usuelle, qui utilise le carré d'écarts, on utilise cette mesure pour minimiser l'influence des grands écarts.

Certaines circonstances impliquent l'utilisation de mesure qui n'est pas des distances au sens mathématique du terme, car elle ne respectant pas ce qu'on appelle l'intégralité triangulaire. On parle alors de dissimilarité plutôt que de distance/ ces mesure sont particulièrement utiles pour la dissimilarité entre des objets constitués d'attributs binaire. Pour cela on peut utiliser l'indice de similarité de Jaccard.

Ce type de distance est par exemple utilisé en botanique pour mettre en évidence des associations d'espèces végétales dans des milieux similaire. Il est aussi utilisé en génomique, pour mesurer la ressemblance entre les séquences des quatre baes constituant les gènes.

Il existe trois critères d'agrégations qui fonctionnent quelle que soit la mesure de distance ou dissimilarité :

- Le critère du lien minimum, qui va se baser sur les

deux éléments les plus proches

- Le critère du lien maximum, qui va ms baser sur les deux éléments les plus éloignés
- Et le critère du lien moyen, qui va utiliser la moyenne des distances entre les éléments de chaque classe pour effectuer les regroupements.

Lors de distances euclidiennes, d'autres critères peuvent être employés. Le plus utilisé est le critère de Ward, qui se base sur ce que l'on appelle l'augmentation de l'inertie. L'inertie totale de l'ensemble des individus est décomposée selon :

L'inertie intraclasse, qui représente l'écart entre chaque point et le centre de gravité de la classe à laquelle il appartient

L'inertie interclasse, qui représente l'écart entre chaque centre de gravité d'une classe et le centre de gravité général.

L'utilisation du critère de Ward va revenir à agréger deux classes de manière à ce que l'augmentation de l'inertie intraclasse soit la plus petite possible, pour que les classes restent les plus homogènes possibles.

Suivant le critère choisi, on peut aboutir à des arbres ayant des formes très différentes. Le saut minimum tend à construire des arbres aplatis, avec des accrochages successifs d'individus à un, alors que le maximum tend à former des groupes isolé et très compacts. Le critère de Ward tend quant à li à produire des classes d'effectifs similaires.

Pour définir le nombre de clusters le plus pertinent, on eut se baser sur les critères suivant :

L'allure générale de l'arbre : elle laisse souvent apparaitre un niveau de coupe « logique » indiqué par des sauts important dans la valeur des indices de niveau. Si ces sauts concernent les k derniers nœuds de l'arbre, alors un découpage en (k+1) classes sera pertinent.

Le nombre de clusters : éviter un nombre trop grand, auquel cas le clustering perd de son intérêt

La capacité à interpréter les clusters : inutile de chercher à retenir des clusters pour lesquels on n'arrive pas à donner de sens métier, privilégier les clustering qui ont du sens.

C. Le clustering non hiérarchique

La différence entre le clustering hiérarchique et le clustering non hiérarchique qu'on sait l'avance le nombre de classes à construire. L'objectif reste le même.

Pour une mesure de distance des individus donnée et pour un nombre de classes K connues, il faut utiliser des heuristiques et des algorithmes qui se basent sur la méthode des centres mobiles.

Pour k classes définies à l'avance, l'algorithme procède de faon itérative. Le processus se stabilise nécessairement, pour des raisons mathématiques. L'algorithme s'arrête soit si deux itérations successive aboutissent au même clustering, soit si un critère de contrôle choisi se stabilise, soit si on atteint un nombre d'itérations fixés.

L'algorithme n'aboutit pas nécessairement à un optimum global : le résultat final dépend des centres initiaux. Dans la pratique, on l'exécute souvent plusieurs fois. Ensuite, soit on conserve la meilleur solution, soit on cherche des regroupements stable pour identifier les individus qui appartiennent aux même partitions. Ces ensembles d'individus systématiquement rattaché à une même classe sont appelés « forme forte » ou « groupement stable ».

Les variantes les plus connus des centres mobiles sont nombreuses, voici quelque unes :

La méthode de k-means [4] : elle fonctionne exactement comme les centres mobiles, à une différence près, qui est le calcul des centres. Un recentrage est effectué dès qu'un individu change de cluster. On n'attend plus que tous les individus soient affectés à un cluster pour en calculer les centres de gravité, ces derniers sont modifiés au fur et à mesure des réaffectations.

La méthode des nuées dynamiques : elle favorise la recherche de groupement stable. C'est une généralisation des centres mobiles dont l'idée est associer à chaque classe un représentant différent de son centre de gravité. Dans la majorité des cas, on remplace le centre de gravité par un ensemble d'individus, qu'on appelle des « étalons » et qui constituent un « noyau », ce noyau est censé avoir un meilleur pouvoir descriptif que des centres ponctuels. A noter que d'autres représentants plus exotiques peuvent être utilisés : une droite, une loi de probabilité, etc.

- La méthode Isodata : Le principe du centre mobile est conservé, mais des contraintes vont permettre de contrôler l'élaboration du clustering. Ces contraintes servent à empêcher la formation de groupes à effectifs trop faibles ou de diamètre trop grand.
- La méthode de k-medoids : elle est similaire à la méthode des k-means, à une différence prés. En effet elle ne va plus définir une classe par une valeur moyenne, mais par son représentant le plus central. C'est donc un individu du cluster qui va représenter ce dernier. Cette méthode à l'avantage d'être plus robuste aux valeurs aberrantes.
- La méthode des cartes auto-organisées : cet algorithme diffère des centres mobiles par la mise à jour des clusters voisins, où les clusters deviennent des neurones activable. Cette méthode non linéaire permet de conserver toute la topologie des données, en partant le plus souvent d'une grille rectangulaire de neurones qui va se déformer.

D. Les approches mixtes

Nous avons exposé les approches hiérarchique et non hiérarchique un à un. Mais il est utile de les utiliser ensemble.

En mixant les deux approches [3], on peut tirer parti des principaux avantages des différentes méthodes :

- La capacité à analyser un grand nombre d'individus, point fort des méthodes non hiérarchiques
- Le choix d'un nombre de classes optimal, rendu possible par la classification hiérarchique.

Il existe des algorithmes mixtes ayant recours aux deux familles. On applique tout ou une partie de ces algorithmes en fonction des besoins.

IV. TOUR SUR LES ALGORITHMES DE CLUSTERING

La littérature regorge d'algorithmes de clustering [5, 6, 7, 8, 9, 10, 11, 12], cette section présente une catégorisation qui regroupe plusieurs algorithmes selon leurs types. Cette proposition de catégorisation est basée sur une perspective design d'algorithme qui se concentre sur les détails techniques des procédures généraux liés au processus de clustering. Sur ce, les processus des différents algorithmes de clustering peuvent être classifiés comme suit :

A. Algorithmes basés sur le Partitionnement

Dans cette catégorie d'algorithmes, tous les clusters sont déterminés rapidement. Des groupes initiaux sont spécifiés puis réalloués vers une union. En d'autres termes, les algorithmes de partitionnement divisent les nœuds par un nombre de partitions dont chacune représente un cluster. Ces clusters devraient répondre aux critères suivants : (1) Chaque groupe doit contenir au moins un nœud, et (2) chaque nœud doit appartenir à exactement un groupe. Si on prend l'algorithme K-means par exemple, les coordonnées d'un centre représentent la moyenne arithmétique des coordonnées de tous les points. Les nœuds qui sont proches du centre représentent les clusters. Il y a plusieurs algorithmes de partitionnement similaires au K-Means [4] [13] [14] comme par exemple K-modes, PAM, CLARA, CLARANS et FCM.

B. Algorithmes basés sur la Hiérarchie

Dans ce type d'algorithmes, les nœuds sont organisés de manière hiérarchique selon leur proximité du centre. Les proximités sont obtenues à travers des nœuds intermédiaires. Un dendrogramme représente les clusters et les nœuds individuels sont représentés par des feuilles. Le cluster initial se divise graduellement vers plusieurs clusters à travers la hiérarchie. Les méthodes basées sur le clustering hiérarchique peuvent être agglomérative (bottom-up) ou bien décisives (top-down). Un clustering agglomeratif commence par mettre un nœud dans chaque cluster puis continue à fusionner les nœuds dans les clusters appropriés. Un clustering divisif par contre, commence par un grand cluster contenant tous les nœuds du graphe puis continue à affecter les nœuds vers les clusters qui leurs sont les plus appropriés. Le processus continue jusqu'à ce qu'on atteigne le critère d'arrêt (souvent le nombre "k" de clusters désirés). Le plus grand défaut de cette méthode vient du fait que, une fois on atteint une étape (fusion ou division), on ne peut plus revenir en arrière. Parmi les algorithmes les plus connus de cette catégorie, on trouve BIRCH [15], CURE [16], ROCK [17] et Chameleon [18].

C. Algorithmes basés sur la Densité

Dans ce type d'algorithmes, les nœuds sont séparés dans des régions de densité, de connectivité et de frontières. Ils sont assez proches du principe des plus proches voisins. Un cluster est défini ici comme étant un composant dense connecté qui peut grandir dans n'importe sens où la densité pourrait le mener. Ceci permet aux algorithmes basés sur la densité de découvrir des clusters ayant des formes arbitraires. Ceci permet aussi de fournir une protection naturelle contre les outliers. La densité globale autour d'un nœud est donc étudiée pour déterminer quels groupes de nœuds influencent particulièrement le nœud en question. DBSCAN [19], OPTICS [20] et DENCLUE [21] sont des exemples d'algorithmes qui utilisent cette méthode pour filtrer le bruit (les ouliers) et découvrir des clusters ayant des formes arbitraires.

D. Algorithmes basés sur les Grilles

Pour ce type d'algorithmes, l'espace des nœuds est divisé en grilles. Le grand avantage de cette approche est le temps de processing, puisqu'elle parcoure tout le graphe une seule fois pour calculer les valeurs statistiques des grilles. Les données des grilles accumulées permettent aux techniques de clustering basées sur les grilles de travailler indépendamment du nombre de nœuds qui emploient une grille uniforme pour collectionner les données statistiques liées à la région en question puis lancent le clustering au niveau de cette région. La performance des algorithmes basés sur les grilles dépendent beaucoup plus de la taille de celles-ci que de la taille du graphe lui-même. Cependant, pour les larges graphes, l'utilisation d'une seule grille pourrait ne pas suffire pour atteindre la qualité et la précision désirées pour ces algorithmes. Wave-Cluster [22] et STING [23] sont exemples typiques de cette catégorie.

E. Algorithmes basés sur les Modèles

Ce genre d'algorithmes a pour objectif optimisé l'ajustement entre les nœuds du graphe et un certain modèle mathématique prédéfini. Ces algorithmes sont basés sur l'hypothèse que les nœuds sont générés par un mélange de distributions de probabilités. Ceci peut aussi permettre de déterminer le nombre de clusters en se basant sur des statistiques classiques tout en prenant le bruit (les outliers) en considération et donc cédant à une méthode robuste de clustering. Les algorithmes basés sur les modèles se voient divisés en deux catégories d'approches : les approches statistiques et les approches basées sur les réseaux de neurones. MCLUST [24] est probablement le plus connu des algorithmes basés sur les modèles. Cependant, il y a pas mal d'autres algorithmes qui relèvent de cette catégorie comme EM [25] (qui utilise un mélange de modèles basés sur la densité), clustering conceptuel (comme COBWEB |26]), et les approches basées sur les réseaux de neurones (comme les cartes de propriétés auto-organisables). L'approche statistique utilise des mesures probabilistes pour déterminer les clusters. L'approche basée sur les réseaux de neurones prend un ensemble de nœuds connectés par des liens comme entrées et sorties. A chaque lien est associé un poids. Les réseaux de neurones ont plusieurs propriétés qui leur font gagner en popularité quant au clustering. Premièrement, les réseaux de neurones ont une intrinsèquement une architecture distribuée qui permet de procéder à des calculs parallèles. Deuxièmement, les réseaux de neurones peuvent apprendre en ajustant les poids de leurs interconnections selon les nœuds. Les patterns de leur côté agissent en tant qu'extracteurs de propriétés (ou attributs) pour les différents clusters. Troisièmement, les réseaux de neurones agissent sur des vecteurs numériques et requièrent que les

patterns des nœuds soient représentés uniquement par des valeur quantitatives. Plusieurs algorithmes de clustering ne manipulent que des données numériques, donc pour les utiliser, il faut représenter les nœuds par des valeurs numériques (leurs coordonnées par exemple). L'approche par réseaux de neurones tend à représenter chaque cluster comme un exemplaire. Un exemplaire agit comme un prototype du cluster et n'a pas nécessairement besoin de lui ressembler à cent pour cent. Les nouveaux nœuds peuvent être affectés au cluster le plus similaire à l'exemplaire en se basant sur une sorte de mesure de distance.

V. BIG DATA MINING ET LE CLUSTURING INCREMENTAL

A. Points faibles des approches traditionnelles

Le Big Data va de pair avec une transformation de l'organisation. Au lieu de s'appuyer sur une vision rétrospective de l'activité en traitant par lots des sous-ensembles de données par lots pour surveiller la performance, elle va se transformer en une entreprise prédictive qui exploite en temps réel toutes les données disponibles pour l'optimiser. Malheureusement, les technologies et les approches de gestion des données traditionnelles font obstacle à cette transformation, car elles sont incapables de gérer le tsunami des données issues des réseaux sociaux, des appareils mobiles, des capteurs et de la télémétrie. Par conséquent, elles ne peuvent pas découvrir les perspectives enfouies dans ces sources de données en temps voulu. Comme le montre la figure 4, les technologies d'entreposage des données et de business intelligence conventionnelles freinent la croissance de l'activité pour plusieurs raisons :

- Elles ne peuvent pas stocker, gérer et fouiller les volumes de données massifs – qui se mesurent en pétaoctets – provenant à la fois de sources de données internes et externes.
- Elles sont incapables d'intégrer les données non-structurées – comme les commentaire des clients, les notes de maintenance, les données des réseaux sociaux, des appareils mobiles, des capteurs et celles générées par des machines – dans les infrastructures d'entreposage de données existantes.
- Elles emploient des techniques de gestion basées sur l'agrégation et l'échantillonnage des données, qui obscurcissent les précieuses nuances et perspectives enfouies dans les données.
- Elles sont incapables de proposer des analyses prédictives en temps réel capables de découvrir et de publier à temps des perspectives exploitables.
- Leurs architectures de traitement par lots peinent à découvrir à la demande les opportunités immédiates.
- Leur reporting rétrospectif n'offre pas les perspectives ou les recommandations nécessaires pour optimiser les processus métier.

Fig. 4. Les technologies et les approches traditionnelles sont insuffisantes [1]

Suite aux points faibles des approches traditionnelles nous avons proposé une nouvelle approche de clustering incrémental dans le Service Big Data Analytics de XEW (SBDA-XEW) [27] en se basant sur un système de fichiers distribué de Service de Sourcing de XEW (SS-XEW), comme la montre la figure 5. Avec le renforcement de la représentation multidimensionnelle, par les nouvelles base de données NoSQL comme : MongoDB, Neo4j, GraphDB et HBase.

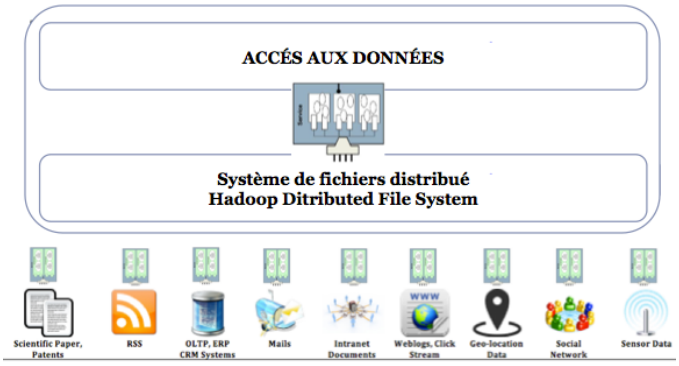

Fig. 5. Service de sourcing XEW

B. Le clustering incrémental dans SBDA-XEW

Le développement de méthodes d'analyse dynamique d'information, comme le clustering incrémental et les méthodes de détection de nouveauté [28] [29], devient une préoccupation centrale dans un grand nombre d'applications dont le but principal est de traiter de larges volumes d'information variant au cours du temps. Ces applications se rapportent à des domaines très variés et hautement stratégiques, tels que l'exploration du Web et la recherche d'information, l'analyse du comportement des utilisateurs et les systèmes de recommandation, la veille technologique et scientifique, ou encore, l'analyse de l'information génomique en bioinformatique. La majorité des méthodes d'apprentissage ont été initialement définis d'une façon non incrémentale.

Cependant, dans chacune des familles de méthodes, ont été développées des variantes incrémentales permettant de prendre en compte la composante temporelle d'un flux de données. D'une manière plus générale les algorithmes de clustering incrémental et les méthodes de détection de nouveauté sont soumis aux contraintes suivantes :

- Possibilité d'être appliqués sans connaître au préalable toutes les données à analyser ;
- Pris en compte d'une nouvelle donnée sans faire un usage intensif des données déjà prises en considération ;
- Résultat disponible après l'insertion de toute nouvelle donnée ;
- Les changements potentiels de l'espace de description des données doivent pouvoir être pris en considération.

L'accumulation de grands volumes de données issues de sources et de sites différents, nous oblige à mettre en œuvre des méthodes de classification collaboratives efficaces pour comprendre les mécanismes impliqués dans ce type de données. Actuellement nous rencontrons souvent des séquences de données structurées sous forme de graphes. Les graphes sont une représentation structurée de l'information contenue au sein d'un ensemble. Les méthodes de visualisation associées au « clustering » permettant d'aider à la compréhension des données volumineuses ainsi qu'à l'extraction de connaissance en proposant à la fois des attributs visuels facilement perceptibles.

Dans nos travaux de recherche, nous nous intéressons plus particulièrement aux séquences de graphes avec un formalisme d'apprentissage non supervisé incrémental. Pour ce faire, nous proposons la démarche suivante:

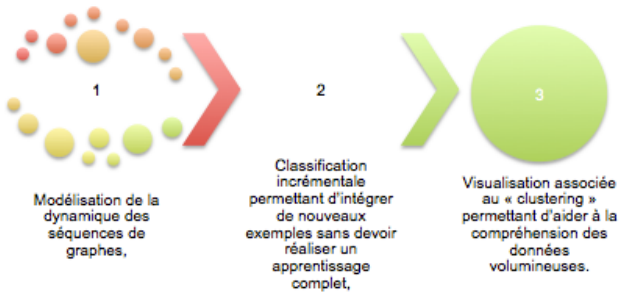

Fig. 5. Approche incrémentale de SBDA-XEW

Cette démarche on l'appliqué sur le projet XEW-WP1.S3 qui vise à intégrer et à mettre à disposition sur la plateforme XEW des outils d'analyse plus fine du contenu générer par le service de sourcing. L'objectif de l'incrémentalité est de nous facilité l'interprétation et la visualisation de l'évolution d'un domaine scientifique ou technique, d'une alliance stratégique, d'une séquence temporelle, d'un réseau social ou sémantique. Nous proposons une approche de construction incrémentale des hypergraphes (grands graphes) par des fourmis artificielles [30] qui s'inspire du comportement d'auto-assemblage de fourmis réelles se fixant progressivement à un support puis successivement aux fourmis déjà fixées afin de créer une structure vivante. La connexion entre fourmis (données) se fait à partir d'une mesure de similarité entre les données. Ce qui nous permet l'exploration visuelle et interactive des graphes en réponse aux besoins d'extraction de connaissance de l'expert du domaine. Ce dernier peut visualiser la forme globale (Macroscopique) et explorer localement les relations de voisinage et de permutation vers d'autre type de graphe.

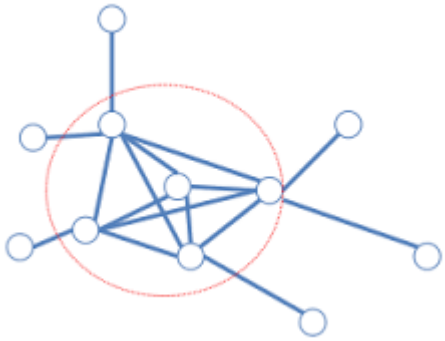

Fig. 6. Approche incrémentale par les fourmis artificielles

VI. Conclusion

L'article fournit une étude sur les algorithmes de clustering les plus connus de la littérature, leurs types, les critères auxquels ils devraient répondre. Afin d'avoir une vision globale sur les avantages et les inconvénients de ces algorithmes, on a procédé à une catégorisation de ces algorithmes selon leur typologie. Cette catégorisation a été mise en place d'un point de vue théorique qui nous a donné une idée sur l'approche à adopter pour notre proposition d'algorithme de clustering incrémental basé sur la théorie des fourmis.

References

Basic format for books:

[1] B. Schmarzo, "Big Data : Tirer parti des données massives pour développer l'entreprise," ISBN : 978-2-7540-5978-7, *2014.*
[2] E. Biernat and M. Lutz, "Data Science : fondamentaux et études de cas," in chapter Book, *Eyrolles, 2016.*
[3] JM. Bouroche and G. Saporta, "L'analyse de données," PUF, Que sais-je ?, *2005.*
[4] SS. Singh and NC. Chauchan, "K-means v/s K-medoids : A comparative study," , *In National Conference on Recent Trends in Engineering & Technology , 2011.*
[5] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay, "Clustering large graphs via the singular value decomposition", Machine Learning 56, 1, 2004, pp. 9-33.
[6] H. Zanghi, C. Ambroise, V. Miele, "Fast online graph clustering via Erdös-Rényi mixture", Pattern Recognition 41, 12, 2008, pp. 3592-3599.
[7] H. Zhang, T. Ma, G.B. Huang, Z. Wang, "Robust global exponential synchronization of uncertain chaotic delayed neural networks via dual-stage impulsive control", IEEE Transaction on Systems Man & Cybernetics Part B, Cybernetics, 40, 3, 2010, pp. 831-844.
[8] Y.J. Liu, S.C. Tong, D. Wang, T.S. Li, C.L.P. Chen, "Adaptive neural output feedback controller design with reduced-order observer for a class of uncertain nonlinear SISO systems", IEEE Transactions on Neural Networks, 22, 8, 2011, 1328-1334.
[9] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, "Analysis of topological characteristics of huge online social networking services", in: Proceedings of the 16th International Conference on WorldWideWeb,WWW'07,ACM, NewYork, NY, USA, 2007, pp. 835-844.

[10] H.F. Zhou, J. Guo, Y.H. Wang, "A feature selection approach based on term distributions", SpringerPlus 5, 1, 2016, pp. 1-14.

[11] X. Huang,W. Lai, "Clustering graphs for visualization via node similarities", Journal of Visual Languages & Computing 17, 3, 2006, pp. 225-253.

[12] M.E.J. Newman, "Fast algorithm for detecting community structure in networks", Physical Review E 69, 6, 2004, 066133.

[13] E. Zhou, S. Mao, M. Li and Z. Sun, "PAM spatial clustering algorithm research based on CUDA,", *24th International Confrence on Geoinformatics,* 2016, pp. 1–7.

[14] Garima, H. Gulati and P. K. singh, "Clustering techniques in data mining : A comparison,", *2nd International Conference on Computing for Sustainable Global Development (INDIACom),* 2015, pp. 410–415.

[15] Y. Yang, L. Wu, J. Guo and S. Liu, "Research on distributed Hilbert R tree spatial index based on BIRCH clustering", *20th International Conference on Geoinformatics,* 2012, pp. 1–5.

[16] P. Lathiya and R. Ranai, "Improved CURE clustering for big data using Hadoop and Mapreduce", *International Conference on Inventive Computation Technologies,* 2016, Volume 3, pp. 1–5.

[17] O. E. Baklanova and O. Y. Shvets, "Methods and algorithms of cluster analysis in the mining industry : Solution of tasks fot rocks recognition", *International Conference on Signal Processing and Multimedia Applications,* 2014, pp. 165–171.

[18] U. Gupta and N. Patil, "Recommender system based on Hierarchical clustering algorithm Chameleon", *IEEE Internationl Advanced Computing Conference,* 2015, pp. 1006–1010.

[19] T. R. Tuinstra, "Range and velocity disambiguation in medium PRF radar with the DBSCAN clustering algorithm", *IEEE Nationl Aerospace and Electronics Conferenceand Ohio Innovation Summit,* 2016, pp. 396- 400.

[20] Md. M. A. Patwary, D. Palsetia, A. Agrwal, WK Liao, F. Manne and A. Choudhary, "Scalable parrallel OPTICS data clustering using graph algorithmic techniques", *Internationl Conference on High Performance Computing Networking Storage and Analysis,* 2013, pp. 1- 12.

[21] A. Idrissi, H. Rehioui, A. Laghrissi and S. Retal, "An improvement of DENCLUE algorithm for the data clustering",*5th Internationl Conference on Information & Communication Technology and Accessibility,* 2015, pp. 1- 6.

[22] D. Jixue, "Data mining fo time series based on wave cluster", *Internationl Forum on Information Technology and Applications,* 2009, pp. 697- 699.

[23] I. A. Venkatkumar and S. J. K. Shardaben, "Comparative study of data mining clustering algorithms", *Internationl Conference on Data Science and Engineering,* 2016, pp. 1- 7.

[24] A. King, Z. Yang and Z. R.Yang, "Multivariate multi-scale Gaussian fot microarray unsupervised classification", *Internationl Joint Conference on Neural Networks,* 2014, pp.2458 - 2463.

[25] H. Ramadan and H. Tairi, "Collaborative Xmeans-EM clustering for automatic detection and segmentation of moving objects in video",*IEEE/ACS 12th Internationl Conference of Computer Systems and Applications,* 2015, pp. 1 - 2.

[26] A. Satyanarayana and V. Acquaviva, " Enhanced cobweb clustering for identifying analog galxies in astrophysics ", *IEEE 27th Canadian Conference on Electrical and Computer Enginneering,* 2014, pp. 1 - 4.

[27] A. El Haddadi,A. Fennan, A. El Haddadi, Z. Boulouard and L. Koutti, "Mining unstructured data for a competitive intelligence system XEW", *6th International Conference on Information Systems and Economic Intelligence,* 2015, pp. 146 - 149.

[28] C. Lei and W. Chong, "An Incremntal clustering algorithm based on sample selection", *9th Internationl Conference on Measuring Technology and Mechatronics Automation ,* 2017, pp. 158 - 163.

[29] D. Wang and A. Tan, "Self-regulated incremental clustering with focused preferences", *International Joint Conference on neural Networks,* 2016, pp. 1297 - 1304.

[30] R. Bhavani, G. S. Sadasivam and R.Kumaran, "A novel parrallel hybrid k-meanns-DE-ACO clustering approach for genomic clustering using MapReduce", *Worl Congress on Information and Communication Technologies,* 2011, pp. 132 - 137.