UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES
ET INFORMATIQUE APPLIQUÉES

PAR
ABDERRAOUF NOUASRIA

EXTRACTION D'ASSOCIATIONS LEXICALES FORTES DANS LES COMMENTAIRES

JUIN 2016

Résumé

L'augmentation du nombre de documents multimédia stockés sur les supports électroniques, issus de l'utilisation du Web et des réseaux sociaux, fait qu'il devient nécessaire de trouver un compromis entre l'utilisation du Web et le procédé de recherche d'information. La conception d'outils d'analyse et de traitement automatiques des contenus textuels est alors nécessaire, dans le but de d'assister l'utilisateur lors de la lecture et la compréhension des textes. Les outils d'exploration textuelle améliorent le processus de recherche de l'utilisateur en vue de déceler des informations pertinentes.

Les informations recueillies lors de l'étape exploration peuvent sembler parfois peu ou pas significatives. Mais elles ont une signification subjective au thème de recherche. Elles sont bien souvent nombreuses et très volumineuses. Malgré les adaptations apportées elles restent très bruitées. Cette ambiguïté constitue un obstacle majeur pour l'homme. En effet pour ce dernier une interprétation objective de ces connaissances extraites est très difficile dans un tel cas.

Ce projet de recherche a pour objectif principal la mise en place d'une conception, qui repose essentiellement sur le principe de représentation vectorielle en entrée. Les informations sont présentées au moyen d'une matrice des fréquences. Notre conception exploite principalement la force des techniques de fouille de données, pour chercher de l'information pertinente sur des données textuelles, essentiellement, les commentaires postés sur les réseaux sociaux, tel que Twitter, TripAdvisor et Facebook….etc.

Nos travaux démontrent que les algorithmes rapides offerts par les règles d'association facilitent l'extraction de l'information pertinente dans un temps acceptable, en permettant le traitement de gros corpus textuel allant jusqu'à 5000 items.

Mots clés :

Extraction de données, règles d'association, Data Minning, algorithme Apriori, Twitter, commentaires, réseaux sociaux, C #.

Remerciements

Je tiens à exprimer ma profonde reconnaissance et gratitude à Mr Ismail Biskri, pour avoir accepté de m'encadrer, et pour m'avoir soutenu, dirigé et orienté durant toute la période de ce projet.

Je remercie également les membres du jury pour avoir accepté d'évaluer mon travail.

Un remerciement spécial à ma petite famille pour leurs soutiens et leurs encouragements.

Un remerciement à toutes les personnes qui m'ont aidé de près ou de loin pour achever mon travail.

Abderraouf.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 - Introduction

La recherche d'information suscite de plus en plus l'intérêt des chercheurs dans le monde de la fouille de données. Un intérêt croissant et inévitable des chercheurs se porte sur cette thématique. Il est question de retrouver de l'information pertinente à partir de données fortement ambiguës et de sources non homogènes. Là où les algorithmes de recherche classiques ont échoué, les techniques d'intelligence artificielle offrent une meilleure réponse. En l'occurrence, adopter les méthodes de data Mining, plus précisément les règles d'association, pour extraire de l'information et des connaissances cachées souvent très pertinentes à partir d'un grand volume de données.

Effectivement, il existe plusieurs travaux de recherche qui se sont penchés sur l'application des règles d'association pour l'extraction d'information, comme l'équipe de recherche du professeur Ismail Biskri à l'Université du Québec à Trois-Rivières qui travaille sur ce concept depuis plusieurs années. Leurs travaux principaux se sont portés sur l'application de la classification textuelle pour l'analyse et le traitement d'information tel que la plateforme SATIM [1].

Les données de recherche sont représentées dans des matrices. Une matrice représente les occurrences et absences d'une information donnée dans un document donné. Cette matrice est souvent volumineuse, il est alors essentiel de doter notre système d'algorithmes d'optimisation. Ces algorithmes engendreront un moindre coût computationnel.

Face à la multitude d'informations diffusées dans les réseaux sociaux, la problématique qui se pose est la capacité d'extraire des règles d'association textuelle à partir de ces réseaux.

Notre défi est de trouver du savoir pertinent dans les informations véhiculées dans les réseaux sociaux, plus précisément dans les commentaires et les publications textuelles. On procédera selon une approche novatrice différente des autres travaux de recherche, du point de vue réalisation et conception.

Nous avons exploré dans cette voie, le potentiel des règles d'association fortes appliquées aux commentaires des réseaux sociaux.

Le présent mémoire est composé de six chapitres, Dans un premier temps, dans le chapitre 1 nous présentons la problématique et les objectifs de ce projet de recherche. Le second chapitre, le chapitre 2 traite du web, des technologies du web ainsi que des différentes évolutions du web. Le chapitre 3 traite des différents aspects théoriques et techniques des règles d'association régulières, ainsi que les règles d'association maximales qui permettent d'extraire des informations moins profitables, mais permettant de mieux interpréter l'information véhiculée dans un corpus.

Aussi nous définissons les mesures d'extraction, telle que les mesures basées sur le support et la confiance. Nous passons en revue les différents algorithmes d'extraction et leur principe de fonction. Nous faisons également une comparaison de ces algorithmes. Nous mettons la lumière sur l'algorithme APRIORI et donnons une présentation plus détaillée. Ce dernier représente l'algorithme principal utilisé dans notre chaîne de traitement. On fera une description des différentes notions et algorithmes qui interviennent dans le développement de notre projet.

Le chapitre 4, aborde les aspects théoriques et conceptuels de la chaîne de traitements et de l'architecture que prend notre système. Il décrit aussi l'analyse des besoins et la réalisation de notre plateforme. Dans ce même chapitre nous faisons une translation entre la méthodologie et la réalisation, et cela par une présentation détaillée du système développé, des différentes fonctionnalités de traitement et d'extraction qu'offre notre application, ainsi que des exemples commentés et consolidés par des prises d'écran et des figures descriptives. Le chapitre 5 est dédié à la présentation et à la discussion des différents résultats obtenus lors de la phase expérimentation. Nous présentons aussi la robustesse du système et l'efficience des résultats obtenus. Les tests émis sont les résultats d'une expérimentation sur un benchmark et des documents issus de la navigation, du flux d'information d'un compte twitter donné. L'application utilise des données d'actualité et de temps réel. Enfin le chapitre 6 donne une conclusion générale, qui se traduit par une vue globale et synthétisante de notre travail ainsi que la démarche adoptée pour résoudre la problématique de ce projet de recherche.

Nous présentons à la fin, une panoplie de perspectives et de propositions, qui feront peut être l'objet d'une continuité pour un travail de recherche ultérieur.

Chapitre 2 - Le Web

2.1 Introduction

On parcourt la toile et on scrute le web, en utilisant des outils de recherche et d'analyse tel que les moteurs de recherche, les bases de données, wiki, fils RSS...Etc. Dans un besoin d'information, on diffuse alors et on archive des quantités énormes et colossales de données de tout genre, textuel, pictural, vidéo, audio... Etc.

Il est plus que nécessaire de trouver un compromis entre utilisation et interprétation de tout ce flux, issu de diverses sources d'information telle que les Smartphones, tablettes et ordinateurs personnels.

La problématique ou l'enjeu ne réside pas juste en la façon d'interpréter ces différents types et catégories de savoir recueillis sur le réseau internet, mais de pourvoir les machines de capacité d'interprétation sémantique, de comprendre l'utilisation de l'homme de cette technologie du web. De plus avec les avancées technologiques, le problème computationnel et calculatoire ne se pose même pas. On dispose de supers calculateurs capables d'opérer de lourdes tâches et processus analytiques.

La complexité des machines et leur forte connexion, ainsi que l'interaction homme machine plus complexe et plus abstraite se rajoute à cette problématique, cette collection d'acteurs génèrent et échangent des quantités énormes de données, dans un environnement non homogène. Ce qui influence fortement et rend cette tâche encore plus difficile.

Dans ce chapitre on parlera de l'historique du web, des différents web existants, des plateformes d'échange, en particulier des réseaux sociaux, mais aussi de l'avenir et de l'enjeu du web sémantique.

2.2 Évolution du Web

Le Web représente une application d'internet qui permet d'accéder à des pages de sites éparpillées un peu partout dans le monde, grâce à un navigateur qui interprète ces données échangées en s'appuyant sur un protocole de donnés.

L'expansion du Web ne cesse d'accroître et est toujours en perpétuelle évolution, Le Web ou communément appelé la toile est construit de pages et d'applications qui contiennent en abondance des photos, des vidéos et du contenu interactif, par le biais de l'interaction entre les technologies Web et les navigateurs.

Les technologies Web ont permis aux développeurs de rendre le Web plus performant, utile et plus attrayant. La communauté Open Web ne ménage pas ses efforts nombreux dans la définition des technologies Web, telles que HTML5, CSS3, et font en sorte qu'elles soient prises en charge par tous les navigateurs.

Les interactions entre les technologies Web et les navigateurs sont à l'origine des puissantes applications Web d'usage actuel [1].

2.3 Le Web Passif

Le Web passif, appelé souvent Web statique ou Web 1.0 est axé sur la distribution d'informations. Aussi, les sites sont plus orientés produits. Ils présentent donc une interaction moindre avec l'utilisateur du Web, qui limite l'intervention des utilisateurs. Il se caractérise par le coût exorbitant et énorme des programmes et logiciels.

Le Web classique utilise la technologie hypertexte HTML, qui est un langage de structuration et de balisage spécifiant le contenu d'un document, sans spécifier le document original. Ce choix a été fait pour favoriser l'interportabilité. Un même document HTML peut être interprété par différents navigateurs. Ainsi, l'interprétation dépend du navigateur qui donne une transcription dépendamment du navigateur utilisé [2].

2.4 Le Web social

Avec l'arrivée du Web social, le web devient beaucoup plus axé sur le partage et l'échange d'informations. On retrouve du contenu divers tel que les images, textes, vidéos et encore plus. Le réseau internet est régi par le flux d'utilisation des nouvelles technologies mobiles des réseaux sociaux et des blogues. Il existe une relation intrinsèque entre l'avis de l'utilisateur et ses préférences, on parlera de socialisation virtuelle et de révolution technico-commerciale [3].

Web2.0 ou Web participatif et Web collaboratif sont les appellations données à un même et unique concept, un concept qui évoque marketing pour les autres, mais apriori qui représente une solution efficiente d'ouverture et de partage.

Ce n'est que vers la fin de l'année 2005 que l'appellation Web 2.0 fut adoptée. Et cela, lors d'une conférence organisée par l'éditeur de manuels informatiques O'Reilly. Le terme a

été inventé par Tim O'Reilly et John Battelle. Ils le définissent comme étant un ensemble de plates-formes logicielles en ligne indépendantes des systèmes d'exploitation et des données qu'elles utilisent. En effet une telle couche logicielle favorise les interactions entre internautes à partir des sites Web. Il représente un atout de forte marge qui propulse l'utilisation du Web et aussi le développement du Web [4].

Le Web 2.0 tire sa philosophie du modèle Peer to Peer, donc la performance du système dépend fortement du nombre et de l'interaction des utilisateurs et des données, le tout sous une plateforme participative. Il faut donner pour recevoir. Le système s'améliore au fur et à mesure que l'utilisation augmente. Comme exemple le crowdsourcing, la création et modification de contenus, les publications telles que les plates-formes d'échange d'information, de partage et enfin les réseaux sociaux [5].

2.5 Le Web sémantique

Les technologies de représentation du contenu représentent la plateforme du Web sémantique. Elles offrent aux programmes et logicielles l'accès et l'utilisation des ressources d'information. Elles représentent la couche qui s'ajoute au Web actuel. Communément ce type de système est appelé métadonnées formelles.

Les standards d'échange de données permettent l'interopérabilité, de ce fait elles contribuent à rendre le Web plus accessible.

La différence entre le Web actuel et le Web sémantique réside dans l'interprétation par machine versus l'interprétation humaine pour ce qui est du Web actuel. Ce type de Web utilise plus les technologies de formalisation et de représentation des données [7].

Dans une idée globale, on souhaite avoir un Web intelligent, où les informations ne seraient pas juste stockées dans des machines, mais arriveraient à avoir des informations comprises et interprétables par les ordinateurs. L'idée est donc de rendre l'utilisation du Web intelligente. De ce fait, si on fait une simple recherche sur un moteur proposant de la recherche en langage naturel, ce moteur de recherche transformera cette demande en langage compréhensible et cohérent pour la machine. Le Web pourrait ainsi devenir un guide intelligent, capable d'apporter des réponses complètes et immédiates à des requêtes en langage naturel [8].

Ainsi, Le Web sémantique permet d'avoir une définition transparente de chacune des ressources du Web, telles que les couches logicielles, les documents, les personnes, les objets.

Technologie du Web sémantique :

Le Web sémantique est l'ensemble des technologies qui œuvrent pour donner aux programmes et agents intelligents l'accès aux différentes ressources du Web et rendre leur utilisation plus efficiente. Le Web sémantique en sa représentation des données repose sur un système de métadonnées formelles. Dans ce qui suit nous donnons une liste non exhaustive des différentes technologies de représentation des données en Web sémantique développé par le consortium W3C [9].

Ce modèle de métadonnées fut avancé lors des premiers débuts du Web en 1994 par son inventeur Tim Berners Lee. Lors de la conférence du « World Wide Web 94 » qui coïncide avec la fondation du consortium W3C.

Ce modèle basé sur les métadonnées formelles, répond à la problématique de représentation des données par les machines de l'information, transcrites dans les documents

diffusés et traités sur le Web, comme par exemple les informations des employés d'une entreprise donnée [10].

Dans ce même ordre d'idée, on a assisté à l'aboutissement en 1999 de la publication de la première version de RDF (Resource Description Framework), qui est un langage permettant d'établir des critères dans un cadre général pour la standardisation des métadonnées des ressources que l'on trouve sur le Web. Par la suite, d'autres langages, technologies et vocabulaires spécifiques sont apparus sur la base de RDF, on retrouve le langage FOAF qui fait une description des relations entre personnes, puis des langages de structuration tel que RDFS et le langage d'ontologie OWL 2004 [11].

C'est le cas, dans les langages d'ontologie ou le langage SKOS, dans lequel chaque ressource Web est prise en considération comme une classe, paramètres propriétés… etc. [12].

Dans certain cas, on décrit ces langages et technologies du Web sémantique et on les assimile à des outils de représentation des connaissances adaptés à l'environnement Web, offrant une translation automatique des données en information, et des informations en savoir [8,9].

Exemple de quelques langages du Web sémantique :

Les technologies du Web sémantique sont multiples et sont en perpétuelle évolution, et leurs développements ne cessent d'accroître. Le but principal du Web sémantique est d'orienter l'évolution du Web pour permettre aux utilisateurs sans intermédiaires de trouver, partager et combiner l'information plus facilement.

XML

La technologie XML, (Xtensible Markup Language) représente le langage de balisage extensible. Elle s'inspire du langage de balisage HTML avec amélioration sur le point

balisage. Il est possible de définir de nouvelles balises et de formaliser d'autres objets ou documents [13].

Formellement, ce métalangage a été élaboré vers la fin des années 90, plus précisément 1997. Son but principal était de promouvoir l'échange de données sur le Web. En effet, il offre un codage de données et de documents par le biais de la structuration qu'offre un balisage d'éléments. La simplicité et la facilité d'utilisation du langage XML favorisent son application dans les applications d'échange Web. On citera le traitement automatique des langages « TAL » et la représentation des connaissances… Etc [14].

Structure d'un document XML

On retrouve dans la **Figure 2.1** une portion d'un document XML de base qui montre la racine ainsi que la formalisation de cinq éléments [15].

```
< ? XML version= » 1.0 »? >
<root>
     <items>
          <item nb= » 1">Premier élément</item>
          <item nb= » 2">autre chose</item>
          <item nb= » 3">Troisième élément</item>
               <item nb= » 4">Quatrième élément</item>
          <item nb= » 5">cinquième élément</item>
     </items>
</root>
```

Figure 2. 1. Exemple de document XML

Langage RDF

Dans la technologie RDF (Resource Description Framework), la définition des ressources offre la possibilité de stockages de ressources sur le Web. De plus, on peut utiliser et stocker d'autres formats, en écrivant des routines qui permettent par la suite d'exploiter ce genre de formats. Cet avantage est utilisé par les dernières générations de navigateurs Web, qui appliquent cette philosophie pour lire les courriels, les marques de pages et les utiliser [16].

Cette modélisation consiste en la représentation des données sous forme d'arbre.

Regardons l'exemple ci-dessous d'un gabarit RDF minimal :

On retrouve dans la **Figure 2.2** un document XML qui représente un fichier minimal, on note la version 1.0 et la balise de marquage de début et fin « RDF » [15].

```
< ? XML version= » 1.0 »? >
<RDF:RDF
  xmlns:RDF= » http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  ...
</RDF:RDF>
```

Figure 2. 2. Exemple de gabarit minimal RDF [17]

Fichier RDF

Un fichier RDF se doit de contenir des éléments sous forme de triplet : sujet, prédicat et objet. Le sujet désigne la ressource dont on souhaite faire la description, tandis que le prédicat constitue le niveau de priorité que l'on assigne à cette ressource. Finalement l'objet est la donnée qui est la valeur de cette ressource [18].

Dans l'exemple suivant, on retrouve une description des animaux, grâce à une balise RDF.

Cette balise représente un enregistrement dans lequel on retrouve les champs suivants : nom, espèce et classe. Chacun de ces derniers est assigné à un espace de nommage appelé « ANIMAUX », dont l'URL a été déclarée dans la balise RDF[19].

```
<RDF:RDF xmlns:RDF= » http://www.w3.org/1999/02/22-rdf-syntax-ns# »
         xmlns:ANIMALS= » http://www.some-fictitious-zoo.com/rdf#">

  <RDF:Seq about= » http://www.some-fictitious-zoo.com/all-animals">
    <RDF:li>
       <RDF:Description about= » http://www.some-fictitious-
zoo.com/mammals/lion">
         <ANIMALS:name>Chien</ANIMALS:name>
         <ANIMALS:species>Chat </ANIMALS:species>
         <ANIMALS:class>mammifère</ANIMALS:class>
       </RDF:Description>
    </RDF:li>
    <RDF:li>
       <RDF:Description about= » http://www.some-fictitious-
zoo.com/arachnids/tarantula">


         <ANIMALS:name>Tortue</ANIMALS:name>
         <ANIMALS:species>Amphibien</ANIMALS:species>
         <ANIMALS:class>Orignal</ANIMALS:class>
       </RDF:Description>
    </RDF:li>
    <RDF:li>
```

Figure 2. 3. Exemple de document RDF

Langage OWL

Les limites de RDF, qui sont principalement la description des ontologies ont poussé à l'expansion du langage de définition d'ontologies OWL. En effet, la problématique réside dans le fait de ne pas distinguer la nature des relations entre les ressources et les outils logiques. Ces limitations dans la technologie RDF ont encouragé grandement au développement du langage OWL [20].

Fichier OWL

Un document OWL se doit de contenir les rubriques suivantes : une déclaration d'espaces de nommage, l'en-tête et une description du contenu de l'ontologie, une définition des classes, des propriétés et enfin l'assertion de faits [21].

Une ontologie OWL est extensible, peut être étendue et peut être utilisée par d'autres extensions OWL déjà existantes [22].

Exemple de fichier OWL

Dans ce fichier OWL (voir figure 2.1), on définit la propriété de classe, pour définir un groupe de ressources ou individus ayant les mêmes caractéristiques.

Ici on retrouve les différents domaines d'études que propose une université. En l'occurrence : mathématiques, informatique, biologie, art, chiropratique.

```
<owl:Class>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#Mathématique"/>
    <owl:Thing rdf:about="#Informatique"/>
    <owl:Thing rdf:about="#Biologie"/>
    <owl:Thing rdf:about="#Arts"/>
    <owl:Thing rdf:about="#Chiropratique"/>
    </owl:oneOf>
</owl:Class>
```

Figure 2. 4. Exemple de Fichier OWL [23].

Les couches du Web sémantique :

Une application Web sémantique se compose de plusieurs couches logicielles, selon le **layer cake ou** couche en millefeuille [24]

Architecture d'une application Web sémantique :

Voici ci-après un exemple (voir figure 2.5) d'une application Web sémantique, composé de plusieurs couches logicielles selon le layer cake.

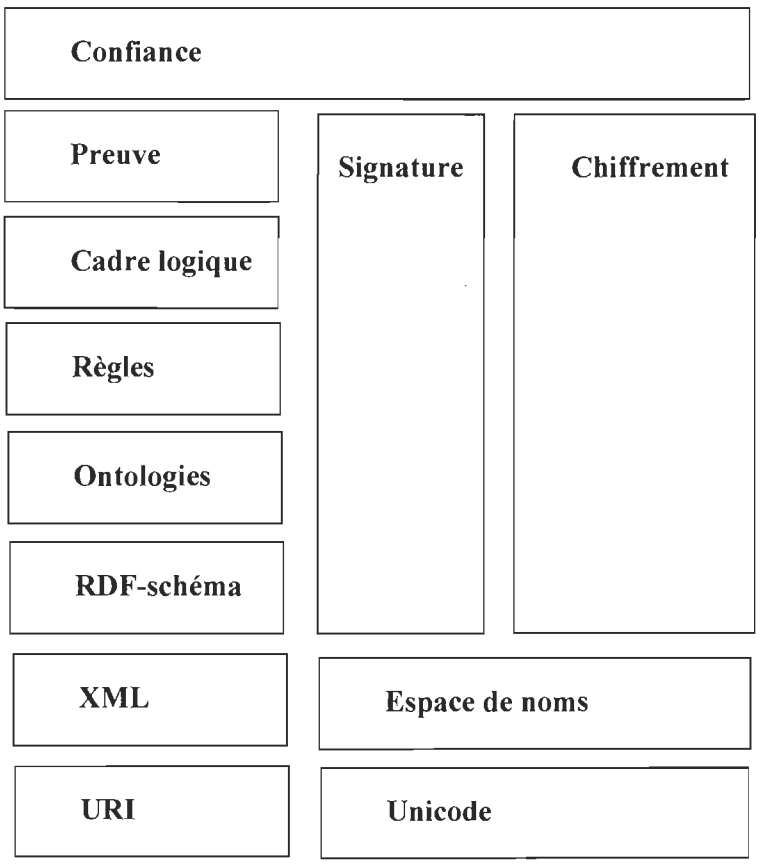


Figure 2. 5. Exemple d'architecture en Layer Cake [24].

Visualisation du Web sémantique :

Les données issues d'une application Web sémantique sont très abstraites. Ce qui peut confondre son utilisation et de ce fait s'éloigner de son but qui est de collecter, filtrer et déceler l'information pertinente de cette source d'information non homogène.

En effet il existe un nombre important d'outils de visualisation d'information Web sémantique. L'outil de visualisation le plus connu est l'outil GEPHI.

Voici ci-après un exemple de l'outil GEPHI présenté à la figure 2.6.

Figure 2. 6.Exemple de visualisation graphique avec GEPHI [25].

Dans cet exemple d'utilisation, on retrouve un graphe qui exprime les corrélations entre les nœuds par le biais d'arêtes.

Ici, à la figure 2.6 on retrouve un exemple d'utilisation du logiciel GEPHI, dans la partie droite se trouve la table des données, avec chaque nœud libellé, et à gauche le graphe en sortie.

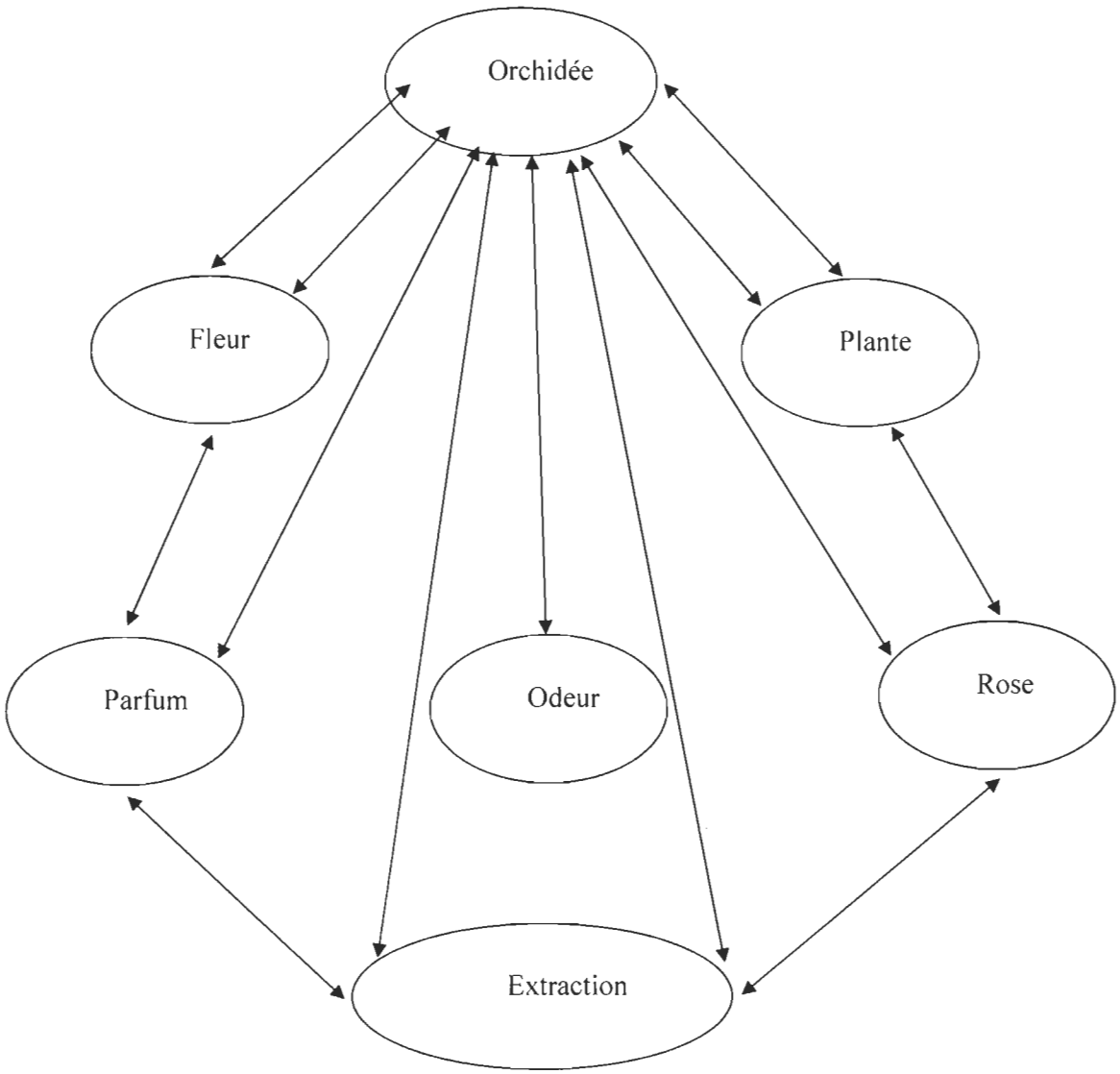


Figure 2. 7. Exemple de graphe sémantique [26].

Dans l'exemple exposé à la figure 2.7 de graphe sémantique, on retrouve les différents attributs d'une relation sémantique, les arêtes représentent les relations sémantiques entre ces attributs.

Exemple : Une orchidée est une plante, de couleur rose, qui émet une odeur et on peut extraire un parfum de cette fleur.

2.6 Les réseaux sociaux

Un réseau social représente une entité, la plupart du temps un site Internet ou blogue qui habilite des individus à communiquer, à échanger et à débattre sur un thème ou sujet d'une caractéristique commune comme la musique, le sport et la politique… etc [27].

Les réseaux sociaux sont extrêmement nombreux et variés, on citera : Twitter, Facebook, Google Plus, Tumblr, Pheed…etc [28].

Dans ce qui suit, nous passerons en revue les trois réseaux sociaux Twitter, TripAdvisor Facebook.

2.6.1 Twitter

Le réseau social twitter « **www.twitter.com** » crée en 2007 à San Francisco, initialement pour le partage d'informations restreintes à une catégorie de personnes. Maintenant, le but est le partage d'information avec le plus grand nombre de personnes à l'échelle mondiale [29].

Le réseau Twitter tire son avantage de la rapidité de diffusion de l'information et le partage du savoir [30].

Lors du tremblement de terre qui a frappé Haïti en 2010, des centaines de personnes se trouvant sur les lieux de la catastrophe ont pu informer le monde entier de l'ampleur des dégâts par la publication de photos et vidéos [31].

Maintenant, les grands médias d'information possèdent un consultant pour trouver de l'information sur les actualités sur twitter [32]. De même, plusieurs chefs d'états et hommes politiques utilisent cette plateforme pour informer et mener leurs campagnes. [33]

2.6.1.1 Les tweets

Les tweets sont des messages importants postés par un blogueur sur sa propre page Twitter [34]. Une spécificité est qu'un tweet ne peut contenir plus de 140 caractères. Mais, il est possible de publier un article plus long en y'ajoutant un lien URL pointant vers notre article ou publication [35].

2.6.1.2 Les retweets et les replies

Retweeter désigne le fait de rediffuser et de retransmettre l'information à un groupe de gens. L'action de retweeter permet la diffusion rapide d'information d'actualité.

Un retweet utilise le formatage suivant :

RT @ [auteur du tweet à retweeter] [contenu du retweet]

Les replies sont des réponses à des tweets lus auxquels on souhaite réagir, ou simplement des messages pour d'autres utilisateurs. Pour retweeter on utilise la syntaxe suivante :

@ [Auteur du tweet]

Il est possible d'envoyer des messages à de multiples utilisateurs, de créer des listes de ce fait. Quelqu'un qui suit nos actualités peut ainsi suivre les informations d'intérêt du groupe, et ainsi se retrouver à suivre les points d'intérêts du groupe et les préférences partagées entre les membres du même groupe. Ce groupe d'utilisateurs est appelé liste des followers ou liste des abonnés [36].

2.6.2 Facebook

Facebook « **www.Facebook.com** » représente le réseau social le plus utilisé au monde, ouvert aux utilisateurs en 2006 sous le nom Thefacebook.com en référence à un album photo sur lequel on peut retrouver les photos des anciens camarades de promotion. [37]

Facebook constitue de plus, en raison de son attrait divertissant, un intérêt commercial important et une valeur monétaire considérable. Il voit son intronisation en bourse en 2013. [38]

Exemple d'une page Facebook

Dans cette page exemple, un utilisateur peut poster des vidéos, des images, de la musique, des liens, mais surtout la principale force est de pouvoir commenter et interagir, donner son avis sur ce que l'on a publié.

Figure 2. 8. Exemple de page Facebook [39].

2.6.3 TripAdvisor

TripAdvisor « **www.tripadvisor.ca** » fut créé en 2000, il permet aux utilisateurs consommateurs de s'auto informer et d'échanger entre eux les informations et leur avis sur des produits touristiques.

TripAdvisor représente maintenant le plus important site d'avis de voyages. Il catalogue presque 35 millions d'avis et de points de vue positifs ou négatifs, sur des hôtels, des restaurants, des attractions, des musées et des destinations de voyage….. etc [40].

Exemple d'avis TripAdvisor

Dans cet exemple (figure 2.9), un avis sur un hôtel a été publié par son auteur, et d'autres utilisateurs expriment leur avis et interagissent avec l'auteur et entre eux en racontant leur propre expérience avec cet hôtel.

Figure 2. 9. Un avis utilisateur publié sur TripAdvisor [41].

2.7 Fonctionnalités dans les réseaux sociaux

Le Web social se démarque par la multitude de possibilités d'interaction entre les utilisateurs. Entre autres, cette technologie permet la création de pages, pour se présenter et résumer la représentation d'un individu, la possibilité de gérer une liste d'amis qui représente le cercle d'intérêt de la personne.

En plus d'avoir juste un contenu textuel, les sites des réseaux sociaux ne cessent d'offrir des possibilités technologiques énormes, dont la capacité d'afficher un contenu multimédia de photos, vidéos et audio visibles sur ce profil [42].

2.8 Utilisation des réseaux sociaux pour le data Mining

L'analyse des réseaux sociaux est précieuse si l'on cherche à contrôler les flux d'informations, améliorer ou simplement parfaire un marketing d'une étude de marché.

Les prospecteurs usent de la force du Data Mining. De telles techniques permettent de mieux connaître, cibler, approcher ses clients pour leur vendre plus, pour mieux innover, et ainsi se distinguer de la concurrence et développer un avantage concurrentiel [43].

L'importance d'une telle analyse pour les grandes entreprises réside dans la découverte des tendances et des penchants de leurs prospects essentiellement pour des fins de test d'amélioration. Effectivement, cela permettra de mieux cibler l'utilisateur, qui est le client. Et de mieux répondre et réagir face à un tel cas d'entrepreneuriat décisionnel [44].

Perspectives et avenir du Web

Le Web offre plusieurs fenêtres ouvrantes vers un avenir Web prometteur. Le Web sera doté de machines de calcul extrêmement rapides et avec des capacités de communication ultra rapides, ce qui mène à la nécessite d'avoir des technologies Web plus avancées.

2.9 Conclusion

Un réseau social permet à une communauté d'utilisateurs de se regrouper en fonction des centres d'intérêts communs telles que la politique la musique, la vie ou la cuisine... etc.

La plupart des sites de ces réseaux sociaux offrent plusieurs possibilités permettant des échanges et des réactivités entre utilisateurs. Mais de plus, ils offrent des fonctionnalités de divertissement et d'apprentissage.

L'interaction entre les utilisateurs des réseaux sociaux génère un énorme flux d'informations et une grande mine de données très intéressante.

De plus, avec les percées dans le domaine des technologies de communication et la diversité des sources d'information, il devient plus que nécessaire de se doter d'outils informatiques performants, afin de permettre aux machines de comprendre les données qui circulent sur la toile.

Dans ce travail de recherche, nous allons explorer cette voie. Concrètement, l'objectif principal est de recueillir ce flux d'information issue de l'interaction des réseaux sociaux et le traiter.

Chapitre 3 - Les règles d'association

3.1 Introduction

L'essor technologique ne cesse de croître, la multitude de sources de données est de plus en plus diverse et variée. La représentation et la présentation de l'information deviennent encore plus abstraites.

La nécessité de se munir d'outils d'analyse et d'extraction de ces colossaux recueils de données devient plus que vitale.

L'objectif est de découvrir des associations ou des corrélations intéressantes entre des éléments dans ces grandes collections et bases de données. Ces éléments peuvent être des: attributs, objets, individus, Items... etc. Prenons par exemple la transaction effectuée par un ensemble de clients d'une grande surface commerciale [45]. Un individu qui achète du café, du sucre et du lait représente une règle d'association entre les attributs café, sucre et lait.

Les règles d'association représentent un outil tangible, efficace et performant. Une règle d'association se présente de la forme $X \Rightarrow Y$, X en association avec Y ce qui veut dire que les transactions ou requêtes qui contiennent l'ensemble des objets X ont tendance à inclure les objets de l'ensemble Y [47].

La recherche des règles d'association est un procédé important dans le Data Mining. Plusieurs algorithmes de recherche des règles d'association existent et permettent de découvrir des relations d'intérêt entre deux ou plusieurs variables stockées dans de très grandes bases de données. Nous avons par exemple les algorithmes Apriori, FP-growth, Eclat, GUHA, OPUS et AprioriD [48].

La plupart des algorithmes d'extraction des règles d'association mettent en œuvre deux propriétés le support et la confiance. On parlera de ces critères et des mesures, mais aussi de l'algorithme d'exploration des données permettant d'extraire les règles d'association ultérieurement [49].

Dans ce chapitre on présentera un état de l'art sur les règles d'association, ainsi que les techniques d'extraction, les différents algorithmes et leurs variantes.

3.2 Notions et définitions sur les règles d'association

3.2.1 Représentation des données

Les données issues des différents documents et bases de données transactionnelles peuvent être représentées sous la forme d'une matrice booléenne à deux dimensions. Dans une telle base, chaque tuple représente une transaction tandis que les différents champs correspondent aux objets inclus dans la transaction. On note par N_e le nombre de transactions, par p le nombre d'articles, par 0 l'évènement d'absence de chaque article et par 1 sa présence dans la transaction. De ce fait on construira une matrice binaire de la base de données [50].

Transactions	Article1	Article2	Article3	Article4	Article5	Article6	Article7
T1	0	1	1	1	0	1	1
T2	0	0	0	1	1	1	0
T3	1	0	1	0	0	0	1
T4	0	0	0	0	1	1	1
T5	0	1	1	1	0	0	0

Tableau 3. 1. Base de données binaire [51].

Dans le tableau 3.1, on retrouve la représentation binaire d'une base de données. Par exemple la transaction T1, contient les articles 2, 3, 4, 6, 7.

Ce tableau représente une matrice creuse a deux dimensions N_e* P. Avec N_e = 5 (Le nombre de transactions), et P = 7. (Le nombre d'items)

Une transaction T représente un sous ensemble E. **Exemple** : soit un ensemble

E = {élément 1, élément 2,..... élément n}.

On aura T= {élément 1, élément 2}.

Notre exemple (**Tableau 3. 2**) donne ceci :

T_1 = {Article2, Article3, Article4, Article6, Article7}

3.2.2 Item et support [53]

Un Item est un objet, élément ou un article d'une base de données.

Exemple 1 :Article1 représente un item.

Exemple 2 :Article3 représente un item.

Un Itemset est un ensemble d'items, d'objets ou d'articles d'une base de données.

Exemple :{ item2, item3, item4, item6}

Un K-Itemset est un ensemble de k éléments, ou k-Items, il est aussi un Itemset.

Exemple 1: {item2, item3, item4, item6} représente un 4-Itemset.

Exemple 2 :{item2, item4, item6} représente un 3-Itemset.

3.2.3 Support d'un Itemset

Le support d'un Itemset représente le nombre total des transactions d'une base de données comportant cet Itemset divisé par le nombre total des observations de cette base de données. [54]. Par exemple, soit une base de données D et soit X un Itemset de n éléments. Dans une

base de données transactionnelle D, le support de l'itemset X est le nombre de transactions dans D incluant X, divisé par le nombre total des transactions de D (figure 3.1)

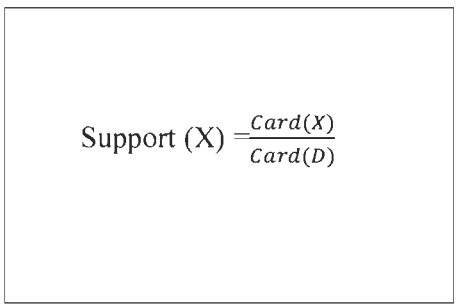


Figure 3. 1. Support d'un Itemset [55].

Exemple 1 : Soit X un Itemset, avec X= {Article2, Article3}.

Soit D la base des transactions présentées précédemment dans le tableau 3.1.

Card(X) est le nombre de transactions dans D, de tel que les Items Article2 et Article3 apparaissent simultanément dans chacune de ces transactions de D. Il est égal à 2

Card(D) est le nombre total des transactions. Il est égal à 5.

Alors Support (X) = $\frac{2}{5}$.

3.2.4 Itemset Fréquent

On dit qu'un Itemset X est un Itemset fréquent si et seulement si le support associé à cet Itemset est supérieur à un support minimum défini par l'utilisateur [56].

3.3 Règles d'association

3.3.1 Définition

Une règle d'association est une application de la forme X $\Longrightarrow$ Y, qui exprime une corrélation de cooccurrence [57].

Il existe deux mesures importantes, le support et la confiance, la robustesse d'une règle d'association est déterminée grâce à ces deux métriques [58]. Une règle d'association qui a un support faible va être observée rarement. La confiance mesure la pertinence de l'inférence dans une règle, par exemple plus grande est la mesure de confiance de la règle X $\Longrightarrow$ Y, plus cette règle sera pertinente [59].

3.3.2 Support et confiance d'une règle d'association

Les notions de support et de confiance ont été identifiées lors des premières études de recherche des règles d'association menées par Hajek, Havel et Chytil 1966 en l'occurrence la méthode GUHA [60].

3.3.2.1 Support d'une règle d'association

Le support d'une règle d'association s'exprime par le nombre de transactions qui contiennent les éléments de X et les éléments de Y divisé par le nombre total des transactions de la base des transactions.

Dans une base de données D, le support d'une règle d'association $X \Rightarrow Y$ est le nombre de transactions qui contiennent X et Y divisé par le nombre total des transactions [61].

$$\text{Support } (X \Rightarrow Y) = \frac{Card\ (X \cup Y)}{\text{Card } (D)}$$

Figure 3. 2. Support d'une règle d'association [62].

Exemple, la règle d'association " Lait ⇒ Pain " , littéralement, **Si** Lait **Alors** Pain.

Le support représente le nombre de transactions dans lesquelles on trouve les Items Lait et Pain, divisé par le nombre total des transactions.

3.3.2.2 Confiance d'une règle d'association :

La confiance d'une règle d'association s'exprime par le nombre de transactions qui contiennent la relation d'union entre la transaction X et la transaction Y divisé par le nombre des transactions qui contiennent la transaction X. [63]

$X \Rightarrow Y$ représente une règle d'association.

$X \cup Y$ représente l'ensemble union contenant les éléments de la transaction X et les éléments de la transaction Y.

La confiance d'une règle d'association est définie comme suit :

$$\text{Confiance } (X \Rightarrow Y) = \frac{Support\ (X \cup Y)}{Support\ (X)}$$

Figure 3. 3. Confiance d'une règle d'association [64]

Exemple, la confiance d'une règle d'association " Lait ⇒ Pain", est égale au support de la règle " Lait ⇒ Pain " divisé par le support de l'Item " Lait ".

3.4 Extraction des règles d'association

3.4.1 Préparation des données

Une étape importante avant de démarrer un processus d'exploration de données en data mining est la phase de préparation de données. En effet, pour pouvoir exploiter ces données il est souvent nécessaire d'appliquer un processus de nettoyage, les informations sont souvent bruitées et incomplètes. Il est alors important d'en tenir compte car la qualité des résultats est affectée fortement par les données utilisées, si on utilise une donnée bruitée on aura des relations entre des données intéressantes et des données peu significatives [65].

3.4.2 Algorithme d'extraction des règles d'association

Il existe plusieurs façons d'explorer les règles d'association, l'une de ces méthodes est la méthode naïve, on utilise alors toutes les combinaisons possibles des attributs et de leurs valeurs pour créer toutes les règles d'association possibles. Ce qui pose problème sur le plan complexité computationnelle du fait de l'explosion combinatoire. En effet, le nombre de règles générées est énorme. On peut optimiser cette méthode en gardant juste les règles avec un support et une confiance minimum. Cela reste insuffisant et les résultats sont insatisfaisants [67].

L'algorithme Apriori représente une approche révolutionnaire dans l'apprentissage et l'exploration des règles d'association.

Nous allons présenter deux différents algorithmes d'extraction des règles d'association : l'algorithme Apriori et l'algorithme Fp-Growth.

Ces deux algorithmes représentent une solution efficace. Le premier est très coûteux en termes d'accès à la base de transactions, le deuxième quant à lui optimise le coût d'accès [68].

3.4.2.1 Algorithme Apriori

L'algorithme Apriori créé par Agrawal et Srikant en 1994 [69], procède en deux temps. Il est basé sur le principe lié à l'approche de support et de confiance.

L'algorithme parcourt le treillis des itemsets pour rechercher les itemsets fréquents et en déduire les règles d'association dont la confiance dépasse le seuil de confiance min $_{conf}$ [69]. Le treillis des itemsets permet d'utiliser plus efficacement cet algorithme d'extraction en admettant les propriétés suivantes [70] :

Tout sous-ensemble d'un Itemset fréquent est fréquent.

Tout sur-ensemble d'un itemset non fréquent est non fréquent.

Exemple de treillis des itemsets

Le nombre d'itemsets fréquents qui peuvent être générés de n items est de 2^n, la génération des Itemsets fréquents est de complexité exponentielle, il est alors essentiel de trouver la méthode de recherche la plus optimale. Ces Itemsets représentent un treillis d'Itemsets représenté sous la forme d'un diagramme de Hasse présenté à la figure 3.4.

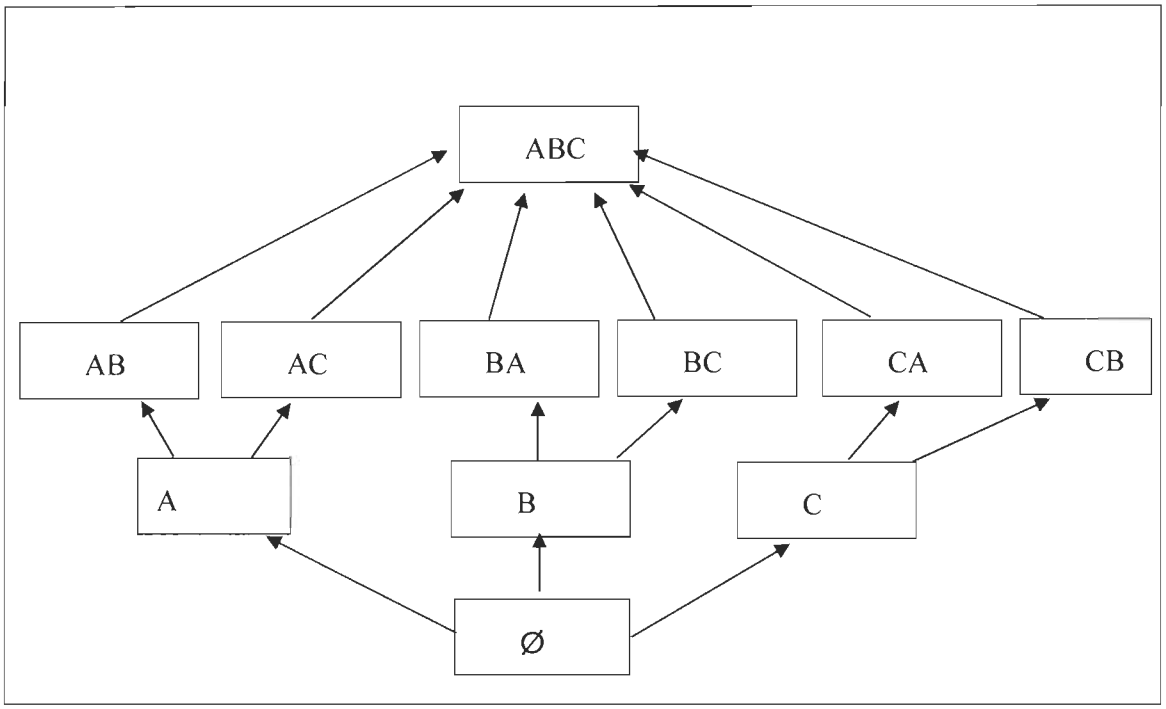


Figure 3. 4. Exemple de treillis d'Itemsets ou diagramme de Hasse

Fonctionnement de l'algorithme Apriori :

D'une manière plus concise, le déploiement de l'algorithme Apriori se fait comme suit [95, 50,62] :

1. Générer les Règles candidates.
2. Calculer le support pour chaque règle candidate.
3. Apparier les règles dont on a calculé le support avec le support choisi.
4. On rejette les candidats dont le support est inférieur au $supp_{min}$.

On termine en sortie avec toutes les règles dont le support est supérieur au support minimal.

En résumé le déroulement se fait comme suit [95] :

L'algorithme Apriori fonctionne en deux phases (voir la figure 3.5). La première consiste en la recherche des ensembles d'items fréquents notée **EIF** et la seconde utilise ces ensembles pour trouver les règles d'association dont la confiance est supérieure à un seuil prédéfini.

Le processus de découverte des ensembles d'items fréquents est itératif, on commence par la construction de EIF avec un seul item, ensuite on réitère pour construire des EIF avec deux items. Ceci va déterminer la taille de chaque ensemble d'items fréquents qui sera noté L_n.

L'ensemble fréquent avec un seul item fréquent sera noté L_1.

Les ensembles d'items candidats sont construits à partir des ensembles d'items fréquents de taille L_{n-1} et seront notés C_n. Par exemple C_2, l'ensemble d'items fréquents candidats de taille 2, sera produit par les ensembles fréquents de taille 1.

$\mathbf{L_n}$ est obtenue en ne gardant que les éléments de $\mathbf{C_n}$ dont le support est supérieur au seuil. On continue les itérations jusqu'à ce que les L_n soient vides.

Apriori (T, t)

Calcul de L_1

$N \leftarrow 2$

Tant que L_{N-1} ensemble vide

$C_N \leftarrow$ **Apriori** (L_{N-1})

Pour chaque transaction $t \in T$

$C_t \leftarrow$ **Sous-ensemble** (L_K , t)

Faire

Pour chaque candidate $c \in C$

Faire

$L_N \leftarrow$ count $\geq Supp_{Min}$

$N \leftarrow N+1$

Retourner L_N

Figure 3. 5 .Algorithme Apriori [81,72]

Dans cet algorithme les ensembles $\mathbf{L_n}$ **et** $\mathbf{C_n}$ sont des enregistrements dans lesquels on stocke les informations et les valeurs des variables.

La procédure **count** permet de calculer et de stocker la fréquence de chaque item de la base de données.

Exemple : l'utilisateur définit les seuils minimaux de support et confiance, on aura min_{supp} =0,30 et min_{conf} =0,30.

Soit la base de données D suivante :

Transactions	Items
T_1	1, 2 ,5
T_2	2 ,4
T_3	2 ,3
T_4	1 ,2 ,4
T_5	1 ,2 ,3
T_6	2 ,3 ,5
T_7	1 ,3
T_8	1 ,2 ,3 ,5
T_9	1 ,2 ,3
T_{10}	2 ,3

Dans un premier temps on calcule la fréquence d'apparitions de chaque item.

Tableau des fréquences d'apparitions de chaque Item:

Itemset	Fréquence
1	6
2	9
3	7
4	2
5	3

Calcul du support de chaque Item :

Itemset	Fréquence
1	6
2	9
3	7
5	3

Remarque, l'Item 4 est supprimé, car son support Support(4) = 0,20 est inférieur à Min_{Supp} =0,30.

Ensuite, on calcule les L_i et C_i, comme présenté dans l'algorithme Apriori de la **Figure 3.5.**

Génération des candidats $\mathbf{C_2}$:

Itemset	Fréquence
1,2	5
1,3	4
1,5	2
2,3	6
2,5	3
3,5	2

Génération des Itemsets fréquents $\mathbf{L_2}$:

Itemset	Fréquence
1,2	5
1,3	4
2,3	6
2,5	3

Génération des candidats $\mathbf{C_3}$:

Itemset	Fréquence
1, 2, 3	5
1, 2, 5	4
2, 3, 5	2

Génération des Itemsets fréquents $\mathbf{L_3}$:

Itemset	Fréquence
1, 2, 3	3

L'algorithme ne génère plus de candidats. De ce fait il s'arrête. Le dernier itemset candidat contient un seul élément.

Les règles retenues par l'algorithme Apriori sont celles formées par les Itemsets L_2 et L_3.

Après déroulement on aura les règles d'association suivantes :

Les combinaisons formées des Itemsets de L_2 donnent les règles :

Si 1 **Alors** 2

Si 1 **Alors** 3

Si 2 **Alors** 3

Si 2 **Alors** 5

Les combinaisons formées des Itemsets de L_3 donnent les règles :

Si 1, 2 **Alors** 3

Si 1, 3 **Alors** 2

Si 3, 2 **Alors** 1

3.4.2.2 Avantages et inconvénients de l'Algorithme Apriori

Avantages :

Il existe une multitude d'avantages dans l'utilisation de l'algorithme Apriori. On en énumère quelques-uns [72,73] :

La découverte rapide de règles d'association pertinentes entre objets.

La facilité d'interprétation des résultats lors de l'extraction des règles d'association, malgré le nombre important de ces dernières.

Inconvénients :

Les inconvénients auxquels on fait face lors d'une utilisation de l'algorithme Apriori sont les suivants [74] :

Les algorithmes d'extraction liés à l'approche support / confiance génèrent un grand nombre de règles d'association.

Un nombre important de configurations d'items ne peuvent pas engendrer de règles d'association.

La recherche de règles d'association impose un temps considérable qui peut s'avérer désavantageux si l'on fait face à une énorme base de données.

3.4.2.3 Algorithme Fp-Growth

L'algorithme Fp-growth permet la découverte des itemsets fréquents sans génération des itemsets candidats. Le processus se déroule en deux étapes, une étape de construction des arbres FP-tree et une étape d'extraction des itemsets fréquents directement de ces arbres [75].

La construction de l'arbre FP-tree s'effectue suivant les étapes ci-dessous [76] :

1. Calculer le support minimal.
2. Calculer chacune des occurrences d'un item constituant la base de transactions.
3. Établir un critère de priorité pour ces items.
4. Faire le tri des items en fonction de leur priorité.
5. Établir le nœud racine.
6. À partir de chaque nœud père insérer les enfants en partant du nœud racine
7. Valider la structure de l'arbre FP-Growth.

Exemple :

Soit la base de données D suivante :

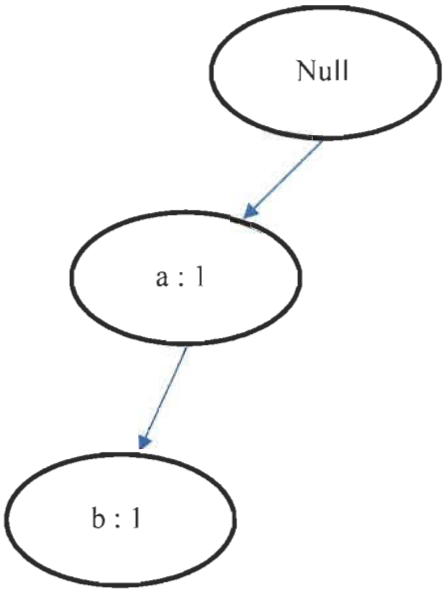
Transactions	Items
T_1	a, b
T_2	b, c, d

T_3	a, c, d, e
T_4	a, d, e
T_5	a, b, c
T_6	a, b, c, d
T_7	A
T_8	a, b, c
T_9	a, b, d
T_{10}	b, c, e

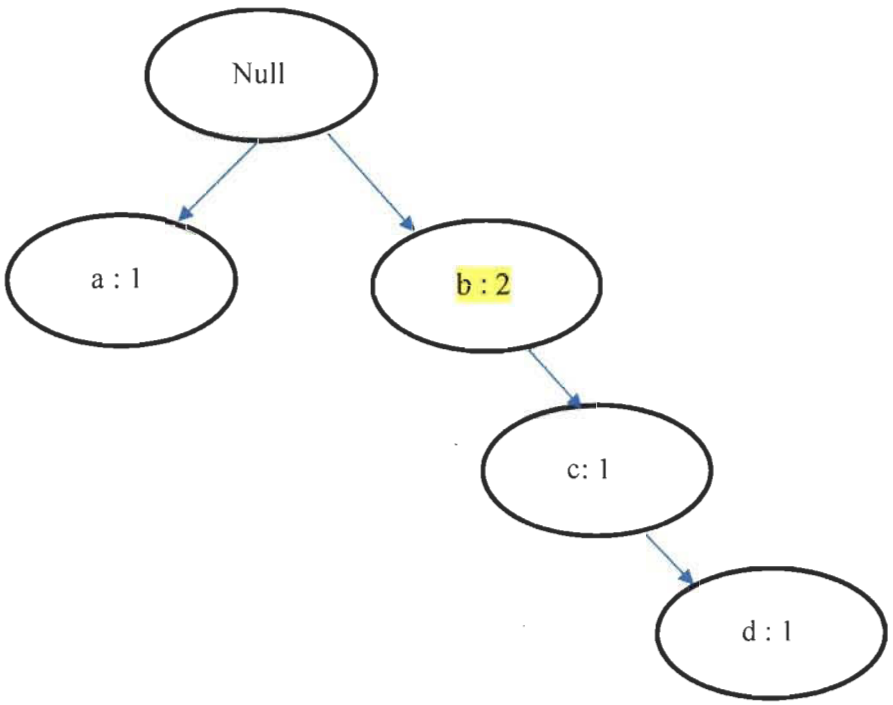
Tableau 3. 3. Exemple de base de données

Construction de l'arbre Fp-Tree

Après lecture de la transaction numéro T_1, on note que chaque nœud représente l'item et sa fréquence.



Le graphe après la lecture de la transaction T_2,



Le processus de construction de l'arbre Fp-tree s'arrête lorsque toute la base de transactions est parcourue.

3.4.2.4 Avantages et inconvénients de l'algorithme FP-growth

Avantages [77] :

La principale force de l'algorithme FP-Growth est le fait que l'algorithme ne fait que deux balayages de la base de transactions. Le premier balayage pour trouver les k-itemsets et pour construire la liste des items fréquents et le second pour construire le squelette de l'arbre FP-tree.

L'algorithme est considéré comme étant complet, la structure étant constituée de toutes les informations sur les éléments fréquents

La structure contient uniquement les objets fréquents classés par ordre de fréquence décroissante.

Inconvénients [78] :

Dans beaucoup de cas d'utilisation, on rencontre des bases de transactions beaucoup trop volumineuses, ce qui bloque le processus de recherche puisque les ressources en mémoire sont insuffisantes pour accueillir toute la structure.

La construction de l'arbre FP-Tree peut être très longue du fait de l'utilisation de beaucoup de ressources de calcul [79].

3.5 Les règles d'association maximales

3.5.1 Définition [80,72, 64]

Les règles d'association maximale parue dans le début des années 60, est un type de règles d'association représentant une méthode très pertinente pour extraire des relations existantes dans un ensemble de données.[81].

Une règle d'association maximale est notée comme suit : $X \overset{MAX}{\Longrightarrow} Y$

Catégorie et taxonomie d'item

Soit un ensemble d'items I = {Item1, Item2, Item3 ...Itemn}.

Une taxonomie T de I représente chaque sous-ensemble disjoint d'items de I. avec. T = {T_1 T_2,...., T_n }.un élément de T représente une catégorie d'items de I.

On notera T(i) la catégorie de i, pour tout Item de I, si X est un ensemble d'itemsets qui sont tous d'une seule catégorie, alors on notera cette catégorie par T(X).

Soit la base des transactions représentées dans le tableau suivant :

Transactions	Items
1	Paris, Madrid, Moscou, Pékin, anglais
2	Madrid, espagnol
3	Paris, français, anglais
4	Seattle, anglais
5	Pékin, Shanghai, mandarin

6	Alger, arabe, français

Tableau 3. 4. Exemple de base de données.

D'après le tableau 3.3 nous pouvons utiliser une taxinomie T qui compte deux catégories, avec T= {T_1 = "ville", T_2 = "langage"}, où ville = {Alger, Madrid, Moscou, Pékin, Paris, Seattle, Shanghai} et langage = {anglais, arabe, espagnol, français, mandarin}.

M-Support

On note le support maximal d'un item X par Smax (X).

Soit T_i une catégorie de I, avec X un itemset tel que $X \subseteq T_i$, On dit que X apparait seul dans une transaction t, si et seulement si $T_i = X$. Autrement dit, X est le plus grand sous-ensemble de T_i qui est dans la transaction t.

On dit qu'une règle $X \Longrightarrow Y$ M-supporte un item X, si X apparaît seul dans une transaction t alors l'item Y apparaît également.

Le M-support représente le support maximal de la règle $X \overset{MAX}{\Longrightarrow} Y$.

Le M-support d'une règle $X \overset{MAX}{\Longrightarrow} Y$ représente toutes les transactions de la base des transactions qui M-Supporte X et Supporte Y.

$$\text{M-Support}\,(X \overset{MAX}{\Longrightarrow} Y) = \{\, t \in T \text{ tel que } t : M - \text{supporte X et supporte Y}\}$$

Figure 3. 6. M-Support d'une règle d'association maximale.

Exemple : Soit la base de transactions suivante :

Transaction	Items

T_1	A, 2, V
T_2	M, L, 3
T_3	A, 1

Tableau 3. 5. Exemple de base de données.

D'après le tableau précèdent on pourra utiliser la taxonomie suivante :

T= {T_1= " lettres", T_2 ="chiffres" }, avec T_1 = {A, M, L, V}, T_2 = {1, 2,3}

Seule la transaction T_3 M-supporte A et M-supporte 1, Alors M-Support ($A \overset{MAX}{\Longrightarrow} 1$)=1.

M-Confiance

Le M-confiance représente la confiance maximale de la règle $X \overset{MAX}{\Longrightarrow} Y$.

Elle est notée par Cmax($X \overset{MAX}{\Longrightarrow} Y$).

Soit T(Y) la catégorie de Y, on note D(X, T(Y)) le sous-ensemble de la base de données D constitué de toutes les transactions qui M-supportent X et qui contiennent au moins un élément de T(Y).

La formule de la M-Confiance sera la suivante :

$$\text{M-Confiance}\ (X \overset{MAX}{\Longrightarrow} Y) = \frac{\text{M-Support}\ (X \overset{MAX}{\Longrightarrow} Y)}{D(X,T(Y))}$$

Figure 3. 7. M-Confiance d'une règle d'association maximale [82]

Exemple : Soit la base des transactions suivante :

Transactions	Items

T_1	B, 3
T_2	M, C, 3,4
T_3	B, 3,4

Tableau 3. 6. Exemple de base des transactions.

D'après le tableau précèdent on pourra utiliser la taxonomie suivante :

T= {T_1= " lettres", T_2 ="chiffres" }, avec T_1 = {B, M, C}, T_2 = {3,4}

Considérons la règle d'association maximale suivante :

$B \overset{MAX}{\Longrightarrow} 3$, M-Confiance= 0,5.

Le M-support est égal à 1, en effet seule la transaction T_1 **M-Supporte** B et **M-Supporte** 3.

T_1, contient X= {B} seul et il contient aussi l'élément 3 qui appartient à la catégorie de Y= {3}.

T_2 contient X= {B} seul et il contient aussi les deux éléments 3 et 4 qui appartiennent à la catégorie de Y= {3}.

Dans ce cas on a deux transactions T_1 et T_2 qui M-supportent X= {B}, et qui contiennent au moins un élément de la catégorie T(Y), On aura D (B, T(3)) = 2.

D'où, D (X, T(Y))= 2.

3.5.2 Algorithme des règles d'association maximales

Un attrait majeur qui distingue les règles d'association maximales est le calcul de règles qui s'effectue plus rapidement [83]. La complexité de ces algorithmes est étroitement liée au nombre de variables. La transformation des données vers des données binaires peut déboucher

rapidement vers un cas d'explosion combinatoire et de ce fait générer un nombre énorme de règles dont la plupart sont redondantes ou faiblement significatives [84].

Les règles d'associations maximales ont les mêmes propriétés que les règles d'associations régulières. La différence réside dans le calcul du support et de la confiance des règles d'association [85].

3.5.3 Avantages et inconvénients des règles d'association maximales

Les règles d'association maximales permettent de déceler des règles d'association dont la qualité d'information est jugée plus au moins importante, mais qui peut être importante pour comprendre un corpus donné. Quant à leurs inconvénients, ils résident dans le problème de la complexité computationnelle et l'explosion combinatoire [87].

La facilité d'utilisation et d'interprétation des résultats en sortie rend l'utilisation des règles d'association maximales plus avantageuse que celle des règles d'association régulières.

3.6 Conclusion

Beaucoup de champs d'application utilisent la puissance qu'offrent les règles d'association, citons l'analyse du panier de la ménagère.

D'après R.Feldman, les associations régulières ne sont pas capables d'identifier des relations solides entre les données. Il a introduit les associations maximales pour améliorer la qualité d'extraction des règles d'association [89].

Les règles d'association maximales offrent une solution assez efficace pour remédier au problème de perte d'informations dans des corrélations jugées moins importantes [90]. On

peut appliquer ces règles d'association pour détecter des groupes de profils à risque grâce à des mesures adaptées sur une caractéristique donnée.

Les règles d'association peuvent être appliquées à divers secteurs d'activité pour lesquels il est intéressant de rechercher des corrélations potentielles entre des objets de diverses catégories. Prenons par exemple le domaine médical, on pourrait chercher des complications dues à des associations de produits pharmaceutiques, ou dans le domaine financier pour trouver des fraudeurs en recherchant des relations de profils inhabituels [91].

L'exploration de ces données d'information à la recherche d'information de corrélation véhiculée dans ces règles d'association est une approche prometteuse à la problématique d'extraction des connaissances à partir d'énorme quantité d'information [93].

Dans le chapitre 4, nous présentons notre implémentation dans laquelle nous utiliserons les notions introduites dans ce chapitre.

Chapitre 4 - Implémentation

4.1 Introduction

Ce chapitre est essentiellement axé sur les grandes lignes qui nous ont permis de réaliser et de mettre en œuvre ce projet de recherche, et les outils exploités pour le développement du logiciel tels que l'environnement de programmation, le matériel utilisé et les principales fonctions de traitement à utiliser.

Aussi nous mettrons en évidence les différentes propriétés de notre implantation, et donnerons des captures-écrans, des principales interfaces et montrerons les spécifications et fonctionnalités qu'offre à l'utilisateur notre logiciel, mais aussi une présentation du mode de fonctionnement.

En réponse à la problématique énoncée dans les chapitres précédents et comme aboutissement, nous allons présenter les étapes de conception et de réalisation de notre système d'extraction des règles d'association.

Notre système se caractérise par l'efficacité, la rapidité d'exécution et les coûts réduits en mémoire et de l'unité arithmétique et logique, dû à l'utilisation la plus optimale des versions des algorithmes d'extraction et des différents artifices de programmation.

4.2 Environnement logiciel et matériel de développement

4.2.1 Langage de programmation du système d'extraction des règles d'association

À cause de la forte complexité calculatoire résultant de l'augmentation de combinaisons possibles entre les Items, l'efficacité doit être équivalente à celle d'un langage machine où le temps de réponse des algorithmes n'est pas soumis à des contraintes de compilation. En d'autres termes, la compilation doit être la plus optimale possible.

Le logiciel gère indépendamment son utilisation aux ressources de mémoire et de calcul, ce qui lui offre un comportement optimal à l'application générée sous cette IDE.

4.2.1 Choix du langage de programmation

Le système développé a été implémenté sous la plateforme Windows et avec le logiciel de développement intégré Microsoft Visual C#.

Le choix fut influencé par les critères de rapidité d'exécution et de la disponibilité de différentes bibliothèques de traitement de texte. En effet le processus d'extraction des règles d'association a besoin de ressources matérielles et logicielles optimales.

4.3 Architecture du système développé

Le système développé a une architecture modulaire composée des quatre modules principaux à savoir :

Le module de récupération de données textuelles (Commentaires)

Le module de traitement et de nettoyage des commentaires

Le module de génération des itemsets fréquents

Le module d'extraction des règles d'association fortes

Les quatre modules sont indépendants, mais indispensables. Comme illustré dans la figure 4.1, le processus suit un schéma en pipeline. Chaque module traite les données en sortie de la couche précédente pour ensuite les transmettre à la couche suivante jusqu'à aboutir à la sortie finale.

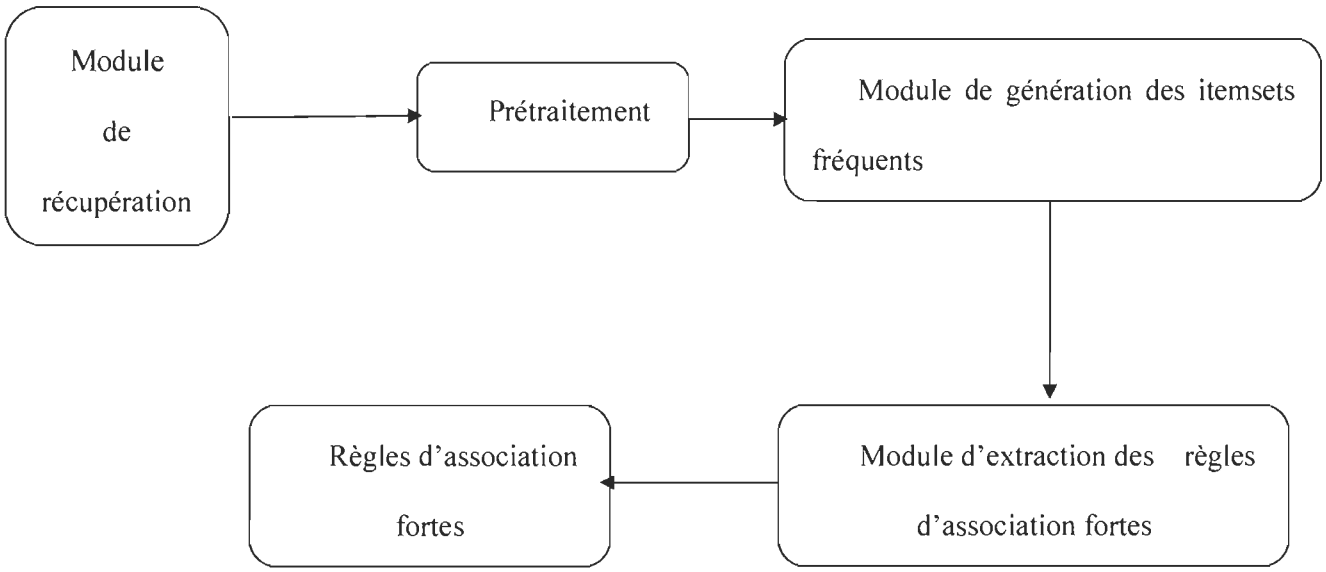


Figure 4. 1.Architecture modulaire du système développé.

4.4 Fonctionnement du système développé

Le système développé permet d'extraire les connaissances sous forme de règles d'association à partir d'informations recueillies des commentaires, en l'occurrence ici les tweets.

Récupération des tweets

La phase de récupération des commentaires se fait directement en scrutant dans les tweets diffusés par chaque utilisateur sur le site « www.twitter.com ».

Le système récupère la source d'information directement du site Web, ensuite crée un fichier texte avec l'extension « .TXT ». Pour cela on utilise la bibliothèque TwiterSHARP sous C #.

4.4.1 Création du corpus

La création de corpus se fait automatiquement et directement du site Web des tweets. Après récupération des commentaires on les stocke dans un fichier de base de données « texte ».

4.4.2 Prétraitement des tweets

Le prétraitement des commentaires se fait à l'aide d'expressions régulières appliquées à notre texte pour filtrer le texte en sortie des caractères indésirables, des sous-chaines qu'on souhaite retirer de façon à garder uniquement des données non bruitées.

Nous créons, ainsi, un motif de traitement à l'aide des expressions régulières pour nettoyer notre base de données textuelle.

Pour la création du motif ou pattern, on utilise la bibliothèque « Text.RegularExpressions » sous l'IDE C#.

Exemple 1 : « Armurie40 »

On souhaite retirer les chiffres d'un mot donné.

Exemple 2 : « #$Avocat »

On souhaite retirer les caractères peu significatifs tel que # ou £ ou $

Dans le tableau 4.1 on retrouve les caractères indésirables retirés :

→	]	[	<	;	=	{	?
□	“	»	>	.	+	}	:
$	%	«	—	,	)	|	&
*	'	_	§	/	(	¡	

Tableau 4.1 Liste des caractères indésirables.

Voici **les expressions régulières** et motifs utilisés :

Nettoyage des caractères spéciaux, par exemple « &»:

Le motif est : (?<!\w)&\w+

Exemple :&Charlie

Nettoyage des caractères spéciaux, par exemple « @»:

Le motif est : (?<!\w)@\w+

Exemple : @Obama

Nettoyage des liens web :

http [^\s] +

Exemple, le tweet suivant contient un lien vers l'article au complet, on souhaite récupérer juste le tweet sans le lien web :

Washington shootings: Wild chase ends at H Street. http://t.co/LwBvP02qv2
http://t.co/c1SKsWX1BO

Après application du motif, le tweet devient Washington shootings: Wild chase ends at H Street

Nettoyage des mots qui commence par un « # » :
^(#) [a-dA-Z] +

Avec ce motif, on nettoie un mot du caractère « # », et on garde le mot. Exemple dans #Liberté, on garde juste le mot Liberté.

Génération des itemsets fréquents

La génération d'un itemset se fait grâce à l'implémentation de l'algorithme APRIORI, l'application affichera une liste d'items, codés selon l'ordre alphanumérique et affichés aux utilisateurs.

4.4.3 Générations des règles fortes

La génération des règles fortes se fait en sortie finale, après l'application du processus en entier et s'affiche dans une fenêtre spécifiquement conçue pour afficher les règles d'association fortes au format codé et transcrit.

4.5 Notre logiciel

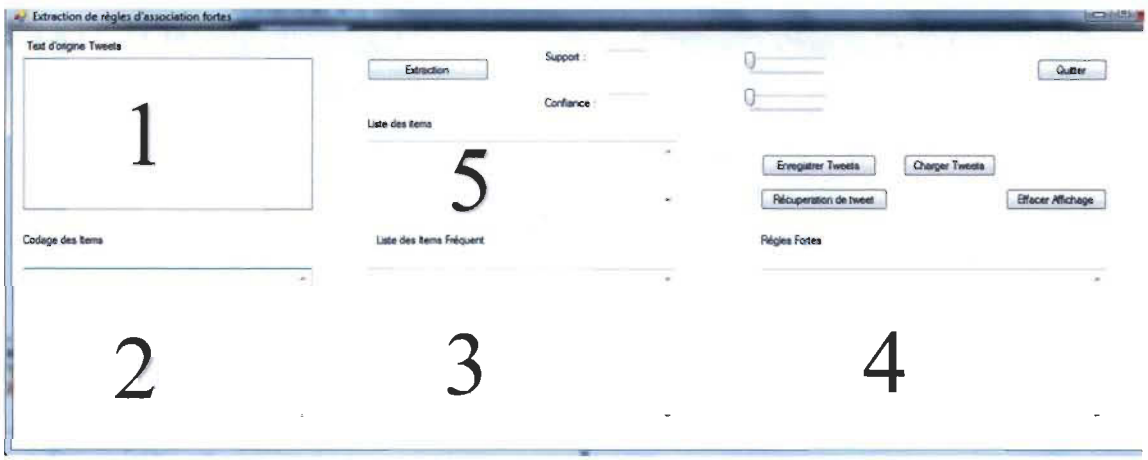


Figure 4. 2.Interface principale pour l'extraction des règles d'association fortes.

La figure 4.2 représente l'interface principale de notre logiciel. Celui-ci offre les fonctionnalités suivantes :

1- La récupération des tweets.

2- Le codage des tweets.

3- L'affichage de la liste des Itemset fréquents.

4- L'extraction des règles d'association fortes.

5- L'affichage de la liste des items.

Nous présenterons les détails du logiciel ultérieurement.

4.5.1 Paramétrage du système

Avant chaque processus d'investigation des règles d'association, il est utile de paramétrer notre système. Ce choix se répercutera directement sur les résultats en sortie.

Il existe deux types de paramètres; les paramètres de compte et les paramètres internes propres au logiciel.

4.5.1.1 Paramètres du compte

Le paramétrage du compte se fait directement sur le site de Twitter « www.twitter.com », et qui est propre à chaque utilisateur.

Un utilisateur doit avoir un nom d'utilisateur et un mot de passe unique, pour pouvoir se connecter à un compte sur le site.

4.5.1.2 Paramètres internes

Avant chaque extraction il faut choisir la valeur du seuil minimal de support èt de confiance.

Le paramétrage se fait par la modification des coefficients support et confiance

Support

Le choix du support se fait directement en introduisant le seuil de support désiré dans la case Support.

Confiance

Le choix du coefficient de confiance se fait directement en introduisant le seuil de confiance désiré dans la case Confiance.

Exemple :

Figure 4. 3. Choix du support et confiance

Dans ce cas, l'utilisateur choisit une expérimentation avec un support de 10% et une confiance de 10%

Remarque : on peut régler l'indice de support et de confiance par le pointeur à droite de la case.

4.6 Fonctionnement du système

Le système fonctionne de manière automatique et interactive. L'utilisateur navigue et consulte des pages et des profils de recherche d'information. Il récupère des listes de tweets ou commentaires publiés sur le réseau. Il a la possibilité d'indexer et de sauvegarder ces

commentaires pour construire son corpus de données textuelles composé des commentaires qui serviront de base de données d'information pour l'application.

Le système conçu appliquera exclusivement l'algorithme Apriori et les techniques d'extraction des règles d'association pour en extraire les règles d'information. Les résultats sont affichables et les paramètres d'expérimentation sont modifiables. L'utilisateur pourra modifier les paramètres support et confiance à sa guise, et ensuite appliquer l'algorithme et le réitérer le nombre de fois qu'il souhaite.

4.6.1 Récupération des Tweets

Le logiciel permet la récupération et l'affichage des commentaires originaux, en cliquant sur le bouton « Récupération de tweet », l'affichage des tweets se fera dans la case « Texte d'origine», comme l'illustre la figure 4.4.

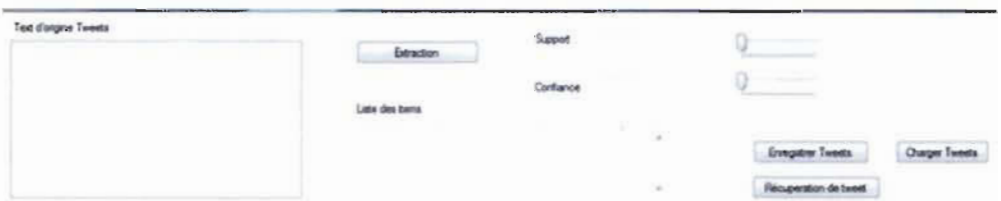


Figure 4. 4. Fenêtre de récupération des tweets.

Récupération des tweets :

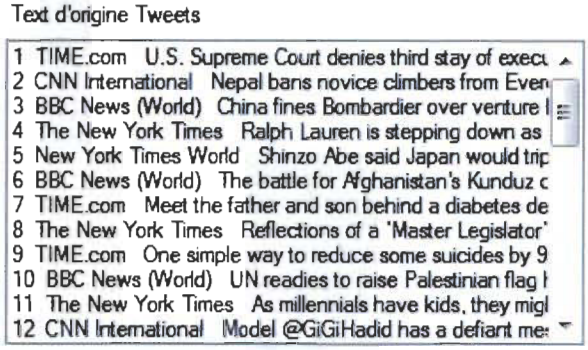


Figure 4. 5. Exemple des commentaires bruts.

La figure 4.5, illustre des tweets postés par les journaux «The New York Times», «CNN», «BBC News».

On a choisi ces medias comme sources d'information, pour la variété d'informations qu'on y trouve. En effet, ces medias traitent des informations internationales et d'actualités.

4.6.2 Traitement et nettoyage des commentaires

Le traitement des commentaires se fait principalement par le nettoyage du texte, comme présenté précédemment dans la partie prétraitement des données. Après le traitement des données le logiciel affichera une liste d'items triés par ordre alphabétique et leur fréquence d'apparitions, dans la case «Codage des Items».

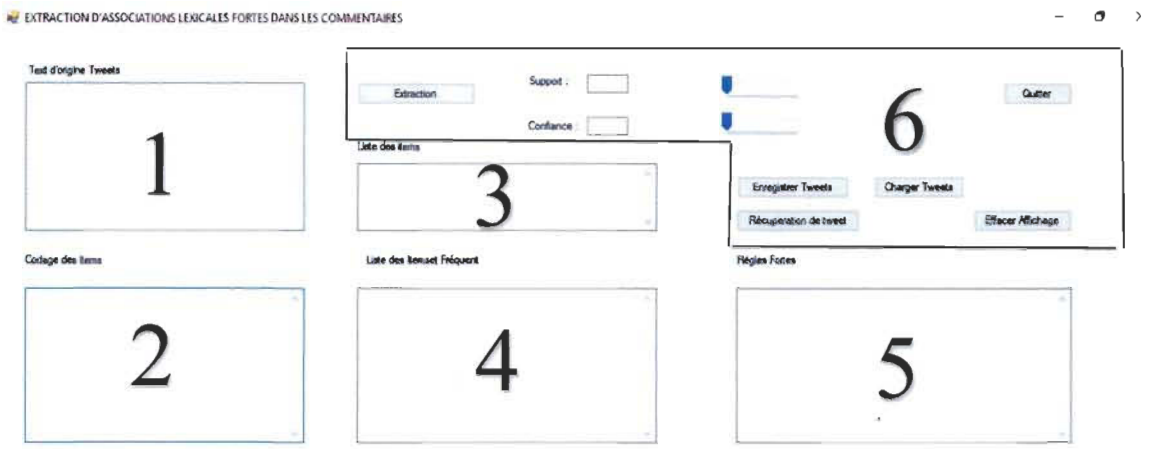


Figure 4. 6. Interface principale pour l'extraction des règles d'association fortes.

La figure 4.6, représente l'interface principale de l'application, on retrouve dans la zone « 6 » les fonctionnalités suivantes :

La récupération des tweets.

L'enregistrement et le chargement des tweets.

L'extraction des règles d'association.

Aussi, on retrouve les cases d'affichage des résultats :

1- L'affichage des tweets.

2- L'affichage du codage d'items.

3- L'affichage des items générés après traitement.

4- L'affichage des itemsets fréquents.

5- L'affichage des règles d'association.

Avant l'affichage on compte la fréquence d'apparitions de chaque item, on trie dans une liste les items, on assigne à chaque item un nombre selon sa position dans la liste triée, la liste des items est ordonnée selon la fréquence d'apparitions de chaque item.

Les items sont affichés dans la case « codage Items » selon leur code, on a adopté le codage numérique, c'est-à-dire chaque Item est représenté par le numéro de sa position dans la liste d'items. Exemple : l'item « Airstrike » apparait en 14^{éme} position dans la liste triée, son codage sera 14.

Génération des itemsets fréquents

La génération des itemsets fréquents se traduit successivement après l'application du processus de recherche. L'affichage se fait dans la fenêtre «Liste des Itemsets Fréquents», comme le montre la figure 4.7.

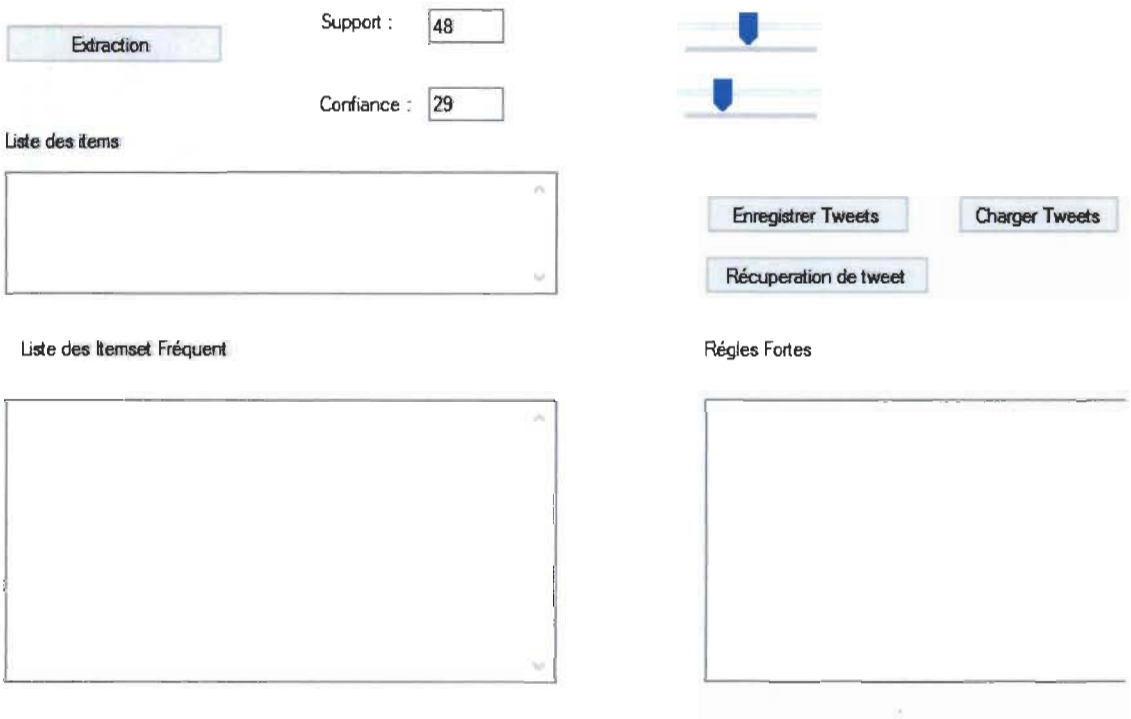


Figure 4. 7. Paramétrage du support et confiance pour l'extraction des règles d'association fortes.

Après le paramétrage des indices de support et de la confiance, on lance le processus d'extraction des règles d'association.

Dans cet exemple, on choisit le support à 48% et la confiance à 29%.

4.6.3 Extraction des règles d'association fortes

L'extraction des règles d'association fortes se fait automatiquement après le traitement du document source et affiche systématiquement le résultat final dans la case «Règles Fortes».

L'utilisateur peut réitérer le processus sans restrictions, il a la possibilité d'enregistrer les tweets qu'il juge intéressants, du point de vue de la pertinence de l'information. L'utilisateur a également la possibilité de charger des tweets issues d'une expérimentation antérieure.

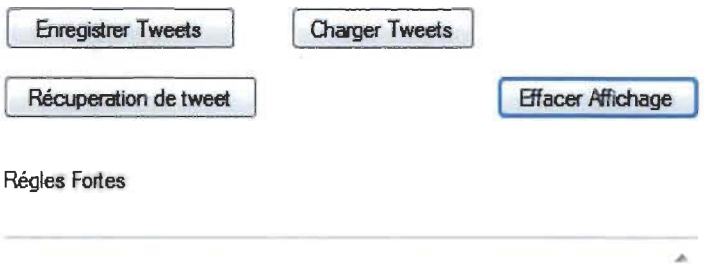


Figure 4. 8. Fenêtre de sorties d'affichage des règles d'association fortes.

La figure 4.8 montre la fenêtre d'affichage des résultats finaux. Les règles sont affichées sous la forme :

Si X **Alors** Y.

4.7 Conclusion

Dans ce chapitre nous avons exposé la chaîne de traitements appliquée dans notre programme et énoncé les différentes possibilités et fonctionnalités de notre système.

Le chapitre suivant traitera d'une phase importante qui est l'expérimentation de notre système sur des données réelles. Il traitera aussi de l'interprétation des résultats obtenus. Il expliquera pourquoi de tels résultats ont été obtenus.

Chapitre 5 - Expérimentations et discussions

5.1 Introduction

Dans cette partie nous exposerons les différents procédés de tests élaborés ainsi que le banc de test et les échantillons de commentaires utilisés lors de notre étape d'expérimentation. Nous présenterons aussi une partie discussion et interprétation des résultats.

Notre expérimentation s'est effectuée sur une base de données textuelles réelle. Voici à titre d'exemple un bout du texte source que nous avons entreposé dans notre base de données. On a exploré plus de 200 tweets, ce qui représente plus de 4731 items dans notre corpus.

La qualité des règles d'association est évaluée selon le critère de pertinence et de cohérence. L'utilisateur qui lit les commentaires va juger de la pertinence des résultats.

Texte d'origine des tweets (figure 5.1) :

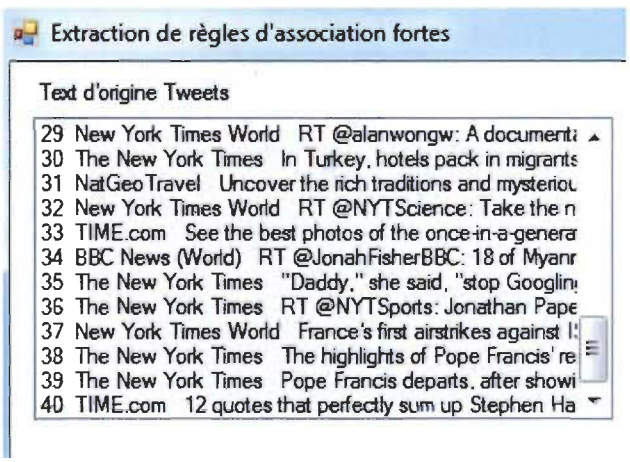


Figure 5. 1. Texte d'origine pour l'extraction des règles d'association fortes.

Dans cet exemple on récupère 40 tweets, chaque tweet est identifié par un numéro unique, ensuite l'identifiant de l'auteur du tweet et enfin le contenu du tweet. Par exemple on a le tweet numéro 39, posté par **l'auteur du tweet :** le journal «The New York Times», ensuite **le contenu du tweet** : " Pope Francis departs, after showing a deft touch ".

5.2 Résultats des expérimentations

L'expérimentation effectuée se déroule sur cinq périodes de temps. Dans notre cas une expérimentation se fait à chaque semaine, dans le but de prouver qu'à chaque période les résultats sont différents, mais aussi que les thèmes qui circulent sur twitter varient suivant l'actualité.

Avant de commencer la phase expérimentation, on définira les paramètres suivants qui serviront pour toute la phase expérimentation :

-Enlever les caractères spéciaux, exemple :#WhiteHouse.

-Supprimer les espaces blancs du texte.

-Supprimer les mots fonctionnels (exemple : for, or, avec, ainsi).

-Soit la règle $X \overset{MAX}{\Longrightarrow} Y$, le choix du X est fait sur la base : X a la plus grande fréquence d'apparitions dans les commentaires, ce qui justifiera par la suite l'étroite relation des règles résultantes avec le thème dominant dans les commentaires.

La qualité d'une règle d'association est la pertinence et la cohérence avec le sujet qui est traité dans les commentaires. Après chaque expérimentation nous mettrons en valeur chacune des règles, selon sa qualité et ce qu'elle apporte à la compréhension des commentaires.

Soit les résultats des cinq expérimentations suivantes :

5.2.1 Expérimentation 1 :

L'expérimentation effectuée dans cette période génère les règles fortes suivantes, sous la forme « $X \xrightarrow{max} Y$ » :

David $\xrightarrow{max}$ *Bowie*

David $\xrightarrow{max}$ *Perdu*

David $\xrightarrow{max}$ Rickman

David $\xrightarrow{max}$ *rebaptisé*

David $\xrightarrow{max}$ *décés*

Les résultats obtenus après l'analyse de ce texte sont (voir tableau 5.1) :

X	Y	M-Support	M-Confiance
	perdu	41 %	66 %
	vie	52 %	75 %
	Rickman	52 %	72.6 %
	rebaptisée	28 %	31 %

Bowie	An	14,68 %	20%
	David	100 %	100 %
	décès	100 %	87 %

Tableau 5. 1. Résultats de l'expérimentation 1.

On va lister dans le tableau suivant les tweets qui ont généré les règles d'association fortes précédentes, par exemple la règle : *David* $\xrightarrow{\max}$ *Bowie*

M-Confiance (%) =100	Tweets
	Une rue d'Austin au Texas rebaptisée "**David Bowie** Street"tp://
	Les gens en une semaine on a perdu **David Bowie** et Alan Rickman on arrête tout 2016 cancelled
	Michel Delpech, Michel Galabru, **David Bowie**, maintenant Alan Rickman..et on est qu'en janvier 2016 démarre très mal.
	Dis donc la vie, alors pourquoi **David Bowie** et Alan Rickman d'un coup ?
	Quand **David Bowie** et Alan Rickman décèdent la même semaine... et qu'on n'est que jeudi.
	Adam Lambert rend hommage à **David Bowie** avec le titre 'Let's Dance'

Tableau 5.2. Tweets de l'expérimentation 1.

Les résultats du tableau 5.2 démontrent la forte relation entre le prénom «David» et son nom «Bowie» avec une M-confiance de 100%, on retrouve aussi une relation forte entre le nom «Bowie» et le nom «Rickman» avec une M-confiance de 72,3%. On remarque que la relation entre le nom «Bowie» et l'évènement «Décès» a une M-confiance de 47%. Les commentaires parlent plus des hommages rendu que du décès.

L'utilisateur qui analyse les commentaires postés sur twitter la période du décès de David Bowie , remarquera que les tweets parlent du décès de l'artiste David Bowie, mais aussi de quelques hommages rendus à l'artiste, ce qui est en étroite cohérence avec les règles fortes générées par notre système.

Le système génère aussi des règles incohérentes qui représentent dans notre cas du bruit tel que la règle $Bowie \xrightarrow{max} An$.

Le lecteur comprendra facilement les thèmes importants durant cette période, plus rapidement en utilisant notre système.

Expérimentation 2 :

L'expérimentation effectuée dans cette période génère les règles fortes suivantes, sous la forme « $X \xrightarrow{max} Y$ »:

$Iran \xrightarrow{max} Nuclear$

$Iran \xrightarrow{max} Nucléaire$

$Iran \xrightarrow{max} Weapon$

$Iran \xrightarrow{max} Obama$

$Iran \xrightarrow{max} Plant$

$Iran \xrightarrow{max} USA$

Les résultats obtenus après l'analyse de ce texte sont (voir tableau 5.3) :

X	Y	M-Support	M-Confiance
Iran	obama	62 %	75 %
	weapon	74 %	80.4 %
	USA	28 %	31 %
	Plant	13 %	20%
	Nucléaire	100 %	100 %
	Nuclear	100 %	100 %

Tableau 5. 3. Résultats de l'expérimentation 2.

On va lister dans le tableau 5.4 les tweets qui ont généré les règles d'association fortes précédentes, par exemple la règle : $Iran \xrightarrow{max} Nuclear$

M-Confiance (%) =100	Tweets
	Each of the pathways **Iran** had toward enough fissile material for a **nuclear** weapon has been verifiably closed down.
	L'**Iran** prévoit l'entrée en vigueur de l'accord **nucléaire** "samedi ou dimanche"

	Feu vert à la mise en œuvre de l'accord **nucléaire iranien**
	L'AIEA annonce que l'**Iran** a respecté ses engagements liés à l'accord sur le **nucléaire**
	Laurent Fabius salue le début de la pleine mise en œuvre de l'accord **nucléaire** avec **Iran**

Tableau 5.4. Tweets de l'expérimentation 2.

Les résultats du tableau démontrent la forte relation entre «Iran», «nucléaire» et «Nuclear» avec une M-confiance de 100%. On retrouve aussi une relation forte entre le nom «Iran» et le nom «weapon» avec une M-confiance de 80,4%. On remarque que la relation entre le nom «Iran» et le nom «Obama» a une M-confiance de 75%. Le nom du président des États-Unis est toujours impliqué, quand on évoque la crise américano-iranienne.

L'utilisateur qui analyse les commentaires postés cette semaine sur twitter, comprendra que les tweets parlent de la crise américano-iranienne et de la levée des sanctions contre l'Iran, en collaborant avec le gouvernement Obama. Ce qui est cohérent avec les règles fortes générées par notre système.

Dans cette expérimentation on n'a pas décelé de bruit, mais de la redondance d'information telle que Nuclear et nucléaire.

5.2.2 Expérimentation 3 :

L'expérimentation effectuée dans cette période génère les règles fortes suivantes, sous la forme « $X \xrightarrow{max} Y$ » :

$Obama \xrightarrow{max} Whitehouse$

$Obama \xrightarrow{max} Barack$

$Obama \xrightarrow{max} Sanctions$

$Obama \xrightarrow{max} USA$

$Obama \xrightarrow{max} Crisis$

$Obama \xrightarrow{max} Iran$

Les résultats obtenus après l'analyse de ce texte sont (voir tableau 5.5) :

X	Y	M-Support	M-Confiance
Obama	whitehouse	53,2 %	66 %
	Barack	74 %	95 %
	sanctions	33 %	42.7 %
	USA	26 %	31 %
	Crisis	17 %	20%
	iran	100 %	100 %

Tableau 5. 5. Résultats de l'expérimentation 3.

On va lister dans le tableau 5.6 les tweets qui ont généré les règles d'association fortes précédentes, par exemple la règle : $Obama \xrightarrow{max} Iran$

M-Confiance (%) =100	Tweets
	Prisoner release helps **Obama** blunt **Iran** criticism

	Decret par lequel **Obama** lève les sanctions contre **Iran** / Executive order to revoke US #sanctions against **Iran**
	Obama pardons **Iran**-ians charged with sanctions violations
	Praises "God" for results of **Iran** diplomacy then calls BarackObama a fool for doing it. **Obama**
	Bonne nouvelle pour administration **Obama**,très critiquée pour son deal avec **Iran** qui faisait pas mention libération prisonniers. 4 libérés.
	Obama autorise l'exportation d'avion vers l'**Iran**

Tableau 5.6. Tweets de l'expérimentation 3.

Les résultats du tableau 5.6 démontrent la forte relation entre le prénom «Obama» et son nom «Barack» avec une M-confiance de 95%. On retrouve aussi une relation forte entre le nom «Obama» et le nom «Iran» avec une M-confiance de 100%. On remarque que la relation entre le nom «Iran» et l'évènement «Sanctions» a une M-confiance de 42%. On a l'évènement « Crisis » avec un M-confiance de 20%. Les commentaires parlent de l'implication du président et de la maison blanche dans la crise iranienne et des sanctions imposées à cette dernière.

L'utilisateur qui analyse les commentaires postés cette semaine sur twitter, comprendra que les tweets parlent de l'implication du président Obama et de la maison blanche dans le

processus de levée des sanctions contre l'Iran, mais aussi de la commande d'avions effectuée par l'Iran. Dans cette expérimentation on n'a pas décelé de bruit.

5.2.3 Expérimentation 4 :

L'expérimentation effectuée dans cette période génère les règles fortes suivantes, Sous la forme « $X \xrightarrow{max} Y$ » :

$Iran \xrightarrow{max} Obama$

$Iran \xrightarrow{max} Ussailor$

$Iran \xrightarrow{max} USA$

$Iran \xrightarrow{max} \text{home}$

$Iran \xrightarrow{max} deal$

Les résultats obtenus après l'analyse de ce texte sont (voir figure 5.7) :

X	Y	M-Support	M-Confiance
	obama	53 %	75 %
	home	33,4 %	51.6 %
	holds	38 %	43 %
	USA	14,78 %	20%
	deal	100 %	100 %

Iran	Ussailors	100 %	100 %

Tableau 5. 7. Résultats de l'expérimentation 4.

On va lister dans le tableau 5.8 les tweets qui ont généré les règles d'association fortes précédentes, par exemple la règle : $Iran \xrightarrow{max} Ussailor$

M-Confiance (%) =100	Tweets
	Iran holding 10 **USSailors**? The same #Iran someone cut a deal with? Deal with the devil sooner or later the devil shows horns
	Guide our sailors home safe **ussailors** from **Iran**
	Seeing how quickly/effectively **Iran** dealt w/**USSailors** illegal intrusion GCC "analysts" stopped discussing potential Saudi attack on Iran.
	Iran, Saved from the Brink of Crisis **USSailors**

Tableau 5.8. Tweets de l'expérimentation 4.

Les résultats du tableau 5.8 démontrent la forte relation entre «Iran» et «USsailor» avec une M-confiance de 100%. On retrouve aussi une relation forte entre «iran» et «deal» avec une M-confiance de 100%. Les commentaires parlent de la prise de soldats américains sur les côtes iraniennes et d'un arrangement entre le gouvernement Obama et l'Iran. Ceci démontre la cohérence des règles fortes générées par notre système.

L'utilisateur qui analyse les commentaires postés cette semaine sur twitter, comprendra que les tweets parlent de la capture de marines américains et d'un arrangement entre la maison blanche et l'Iran pour la libération des soldats.

5.2.4 Expérimentation 5 :

L'expérimentation effectuée dans cette période génère les règles fortes suivantes, sous la forme « $X \xrightarrow{max} Y$ »:

$Ebola \xrightarrow{max} Guinea$

$Ebola \xrightarrow{max} Liberia$

$Ebola \xrightarrow{max} SierraLeone$

$Ebola \xrightarrow{max} Guinea$

$Ebola \xrightarrow{max} ONU$

$Ebola \xrightarrow{max} Care$

$Ebola \xrightarrow{max} For$

Les résultats obtenus après l'analyse de ce texte sont (voir tableau 5.9) :

X	Y	M-Support	M-Confiance
	ONU	49,82 %	66 %
	For	28,98%	43%
	Care	58,3 %	75 %

Ebola	Child	56 %	72.6 %
	Healh	27%	31 %
	Liberia	100 %	92%
	Sierraleone	100 %	97 %
	Guinea	100 %	100 %

Tableau 5. 9. Résultats de l'expérimentation 5

On va lister dans le tableau 5.10 les tweets qui ont généré les règles d'association fortes précédentes, par exemple la règle : $Ebola \xrightarrow{max} Guinea$

M-Confiance (%) =100	Tweets
	Ebola "flare-ups": 4 in Liberia, 3 in each of **Guinea** & SierraLeone. 1st in MAR, last in mid-NOV
	The three countries-**Guinea**,Liberia,SierraLeone-remain at high risk of additional small **Ebola** outbreaks,also known as flare-ups
	A bigger question is why Liberia, **Guinea** SierraLeone had such broken #health systems BEFORE **Ebola** struck? cc ChathamHouse Telegraph
	A child who lost his family to **Ebola** plays at the Child Care Centre in **Guinea**

	The outbreak started Jan. 14, 2014 and claimed over 11,000 lives. **Guinea** Liberia SierraLeone are now **Ebola**-free

Tableau 5.10. Tweets de l'expérimentation 5.

Les résultats du tableau 5.10 démontrent la forte relation entre «Ebola» et «Guinea» avec une M-confiance de 100%. On retrouve aussi une relation forte entre «Ebola» et «Liberia» avec une M-confiance de 92%. L'ONU apporte son soutien aux peuples de Guinée, du Liberia et du Sierra Leone de peur que le virus ne réapparaisse.

L'utilisateur qui analyse les commentaires postés cette semaine sur twitter, comprendra que les tweets parlent du virus Ébola qui a fait ravage dans les pays comme la Guinée, le Liberia et le Sierra Leone, et des efforts de l'ONU et de l'organisation mondiale de la santé pour venir en aide aux enfants de ces pays-là.

Dans cette expérimentation on décèle du bruit, et de l'information incohérente avec la règle $Ebola \xrightarrow{max} For$.

5.3 Discussion et interprétation de résultats

Les résultats obtenus lors des expérimentations reflètent les sujets d'actualité qui font la une sur les réseaux sociaux.

Chaque règle d'association obtenue lors du processus d'extraction en sortie représente les associations les plus pertinentes qui circulent sur le réseau social twitter.

On retient d'une telle expérimentation les éléments suivants :

Les résultats obtenus sont en cohérence avec les commentaires postés, ils permettent de mieux comprendre le texte, dans notre cas ils représentent les tweets les plus importants.

Mais aussi les relations d'association fortes permettent de déceler de l'information pertinente jugée peu significative, mais qui est importante pour comprendre le texte, tel que vu dans les expérimentations précédentes (Expérimentations 1, 3 et 4).

Les bruits dans les résultats sont la conséquence d'un problème connu sur les réseaux sociaux. Les commentaires postés sur twitter et les réseaux sociaux en général ne respectent pas de règle grammaticale. Les informations postées sur le réseau twitter sont multi langues, c.à.d. un utilisateur va utiliser deux ou plusieurs langues dans une même phrase, les utilisateurs utilisent aussi pleins de caractères spéciaux on y retrouve des mots qui n'existent pas dans le dictionnaire…etc.

Le choix des données est important avant une expérimentation on pourrait envisager de faire une classification par les réseaux de neurones, ou le classifieur k-means afin d'obtenir des données de qualité qui améliorent nos résultats. Pour entrainer un réseau de neurones, on utilisera à titre d'exemple les algorithmes de rétropropagation, quant à la méthode k-means elle est facilement implémentable puisqu'on n'a pas besoin d'information sur les données.

Les résultats obtenus se traduisent en informations d'actualité utiles. Ces informations sont extrêmement importantes pour la compréhension des thèmes. Elles sont en cohérence avec les sujets d'actualité, de ce qui se passe sur le réseau, des sujets et des discussions à fort intérêt et qui suscitent une forte attention.

5.4 Conclusion

Les règles fortes sont un excellent outil si l'on souhaite extraire de l'information d'un grand nombre de sources de données variées et de volume assez important.

En effet, avec les algorithmes rapides qu'offrent les règles d'association, il est facile d'extraire de l'information pertinente. Les résultats obtenus sont assez satisfaisant compte tenu de la complexité et l'hétérogénéité des sources de données et le volume énorme du corpus textuel à traiter, en effet tel que démontré lors de nos expérimentations. Le système recouvre une quantité de règles dans un laps de temps assez bref. Il permet le traitement de gros corpus textuel allant jusqu'à 5000 items. Les tweets sont issues de sources d'informations variées : des medias politiques, sportifs, artistiques...Etc.

Le tableau 5.11 représente le nombre d'items extraits d'un corpus exprimé en temps d'exécution. Pour démontrer la rapidité d'exécution de notre système développé. Le temps de traitement de notre système dépend des opérations gourmandes en ressources de calcul, lié principalement au calcul de l'algorithme Apriori et des algorithmes de tris.

Nombre d'Items	Temps de traitement (min)
2328	7
3014	9
4731	12

Tableau 5.11. Temps de réponse par rapport au nombre d'items.

Pour ce qui est du critère du choix du support et de la confiance, on peut envisager une méthode adaptative, comme l'utilisation des réseaux de neurones ou les algorithmes génétiques, ou colonies de fourmi pour trouver un indice de support/confiance optimal.

Les résultats de la méthode d'extraction appliquée, aux commentaires sur twitter sont souvent bruités. Les commentaires postés ne suivent pas un modèle défini ou un format d'écriture tel qu'un langage structuré.

Le système représente une bonne alternative, pour pallier à la problématique liée au volume important de commentaires. De ce fait il est plus utile d'utiliser notre système, au lieu de lire les tweets par un utilisateur qui pourra omettre une information importante à la compréhension du corpus.

On peut alors généraliser ce constat. Il est de même sur tous les réseaux sociaux, il n'existe pas de motif, les réseaux sociaux ne respectent aucune syntaxe ou grammaire, par exemple on retrouve deux langues différentes dans une même phrase, une phrase qui manque de sujet ou de verbe, des mots qui n'existent pas…etc.

Chapitre 6 - Conclusion

Les règles d'associations jumelées au réseau social ont démontré leur intérêt applicatif aux entreprises. En effet, elles sont applicables dans leur projet directeur de prise de décision, de ciblage de clientèle et d'amélioration des produits selon l'intérêt porté par les utilisateurs. Essentiellement, il s'agira d'aller chercher et recueillir les intérêts des utilisateurs pour un tel produit et non un tel autre produit selon les tendances des clients et utilisateurs [99].

Dans ce mémoire, nous avons traité la problématique d'extraction d'associations lexicales fortes dans les commentaires. Les règles d'associations appliquées lors d'un processus de recherche des associations fortes présentent des résultats assez intéressants.

Ce travail a permis de déceler un potentiel important dans les règles d'association, ainsi, que faire ressortir l'information pertinente et de qualité que l'on peut retrouver dans les commentaires sur les réseaux sociaux. Cette expérimentation nous permet aussi d'évaluer le degré d'intérêt sur des sujets d'actualité et de voir réellement le lien étroit qui existe entre ces sujets d'actualité. L'impact des données utilisées est important, ces dernières doivent être recueillies et traitées de manière à ne pas avoir de données erronées ou d'informations manquantes, ce qui fausse le processus de recherche et donne des résultats altérés par du bruit, ce qui est le cas avec les données issues des réseaux sociaux tel que twitter.

Comme perspective nous proposons l'utilisation d'autres réseaux sociaux tels que Facebook, Instagram, et de généraliser cette solution pour couvrir un plus grand nombre

d'informations. Nous pourrons, Aussi appliquer les algorithmes mathématiques d'optimisation pour réduire les coûts en temps et mémoire.

Recueil bibliographique et Références

[1]. Biskri Ismail, Meunier Jean-Guy. "Systéme d'Analyse et de Traitement de l'Information Multidimensionnelle", JADT, 2002.

[2]. Sareh Aghaei, Mohammad Ali Nematbakhsh,Hadi Khosravi FarsanI. "Evolution of The World Wide WEB:From WEB 1.0 to WEB 4.0", International Journal of Web & Semantic Technology, 2012.

[3]. Berners-Lee, Tim. "Artificial Intelligence and the Semantic Web. Boston ", AAAI, 2006.

[4]. Chaimbault, Thomas. "Web 2.0 : l'avenir du Web? S.l. : école nationale supérieure des sciences de l'information", septembre 2007. p. 42.

[5]. Battelle, O'Reilly Tim. "Web Squared: Web 2.0 Five Years On", web2summit.com, 2009.

[6]. Brodeur, Jean. "Le Web sémantique et les documents liés", Natural Ressources Canada.

[7]. Budain Nathalie, Tedeschi Benoit, Vaubourg Stéphane. "Nouvelles Technologies Réseau", 2003 .

[8]. Gagnon, Michel. "Introduction au Web sémantique". UQAC. p. 94.

[9]. Troncy, Jérôme Euzenat — Raphaël. "Web sémantique et pratique documentaire". 2005.

[10]. Menon, Bruno. "L'architecture du Web sémantique". 2010.

[11]. Peccatte, Patrick. "Les métadonnées un élément clé de la gestion de contenu", ATICA, octobre 2002.

[12]. Fléty, Yann. "Vers une mise en observation des systèmes énergétiques pour territorialiser l'énergie", Université de Franche-Comté, 2014.

[13]. Zoghlami, Kaiser. "Création, partage et transfert d'ensembles de données terminologiques basés SKOS",UQAM, 2011.

[14]. AV Aho, R Sethi, JD Ullman. "Introduction To Web Services". pp. 140-118-105-174, 1986.

[15]. Carton, Olivier. "L'essentiel de XML", Université Paris Diderot, 2005.

[16]. SEPT, Équipe du pilotage du système. "Production de fichiers de données XML pour la transmission électronique de résultats d'analyse d'échantillons d'eau potable" , 2015.

[17]. Broekstra Jeen, Kampman Arjohn. "SeRQL: A Second Generation RDF Query Language. " 2003.

[18]. Lapalme, Guy. "Semantic Web for the Working Ontologist. ", UdeM. 2013.

[19]. Manolis, Koubarakis."An Introduction To RDF. ", Department of Computer Science ,University of Crete, 2003.

[20]. Michiel De Keyzer, Nikolaos Loutas ,Stijn Goedertier. "Introduction aux RDF & SPARQL", 2012

[21]. Lacot, Xavier. "Introduction au langage OWL, un langage XML pour ontologies Web", 2005.

[22]. Espinasse, Bernard. "Introduction au langage- Ontology Web Language- OWL", l'Université d'Aix-en-Provence, mars 2009.

[23]. Essaie, Amira. "BeliefOWL An Evidential Extension to OWL", 2008.

[24]. Lapiduq, Francis. "Le langage d'ontologie Web OWL", 2006.

[25]. Mestiri, Mohamed Amine. "Vers une approche Web sémantique dans les applications de gestion de conférences", Université de Laval, 2007.

[26]. Heymann, Sébastien. "Gephi : analyse et visualisation de données relationnelles, réseaux complexes". LIP6 (Université UPMC), 2012.

[27]. Larlet, David. "Le Web sémantique ou l'importance des données liées", 2010.

[28]. Teixeira, Manuela. "L'émergence de réseaux sociaux sur le Web comme nouveaux outils de marketing", Université d'Ottawa, 2009.

[28]. Mary Madden, Amanda Lenhar,Sandra Cortes,Urs Gasser. "Teens, Social Media,and Privacy", 2013.

[29]. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a Social Network or a News Media? ", Department of Computer Science, KAIST, 2010.

[30]. CSA. "Du 20 heures à Twitter : Les réseaux sociaux bousculent l'info", 2011.

[31]. Canadienne, Croix rouge. "Médias sociaux en situation d'urgence", 2012.

[32]. Maurel, Lionel. "Ce que Twitter fait aux bibliothèques",2003.

[33]. Levallois, Clement. "Débuter avec Twitter. ", Emylion Business School, 2015.

[34]. Karsenti, Thierry. "25 Usages éducatifs de Twitter. ", Udm, 2015.

[35]. Sysmos. "Inside Twitter: An In-Depth Look Inside the Twitter World", 2009.

[36]. Pillou, Jean-François. "Twitter — comment l'utiliser", 2015.

[37]. Fréchette, Élyse. "Élaboration d'un questionnaire portant sur l'usage", Université de Laval, 2015.

[38]. Dunay, Hoboken, Wiley John. "Facebook marketing for dummies", 2010.

[39]. Louvre, Musée dû. Musée du Louvre, https://fr-fr.facebook.com/museedulouvre/.

[40]. Vásquez, Camilla. Complaints online: "The case of TripAdvisor. ", Journal of Pragmatics, 2011,

[41]. TripAdvisor. Exemple Hotel : http://fr.tripadvisor.ca/.

[42]. Folgeman, Françoise . "Utilisation des réseaux sociaux pour le data mining", 2013.

[43]. Stattner, Erick. "Introduction à l'Analyse des Réseaux sociaux", Université des Antilles et de la Guyane, France, novembre 2012.

[44]. Wellhoff, Thierry. "Tout ce que vous avez toujours voulu savoir sur les médias sociaux" Wellcom, 2009.

[45]. Mouna Azzeddine. "Datamining Distribué dans les grilles : approche règle d'association", Université des sciences et de la technologie d'ORAN Mohamed Boudiaf, 2012.

[46]. Agrawal, R., R. Srikant. "Fast algorithm for mininig assoiation rules", 1994.

[47]. Wang, Xin ,Liu, Xiaodong , Pedrycz, Witold and Zhu, Xiaolei , Hu, Guangfei. "Mining axiomatic fuzzy set association rules for classification problems. ", European Journal of Operational Research, 2012.

[48]. Berlin, Heidelberg : Springer Berlin Heidelberg. "Association Rule Mining : Models and Algorithms", 2002.

[49]. Shroff, Neha M.,Gujar, G. V. "Mining Association Rules from XML Document", International Journal of Computer Science and Mobile Computing, 2014.

[50]. Boulanger, Alain. "Génération des règles d'association : Treillis de concepts denses", Université du Québec à Montréal, 2009.

[51]. Wa, Wan Aezwani. "Apriori and Eclat Algorithms in Association Rule Mining", 2014.

[52]. P. Tan, M. Steinbach, V. Kumar. "Association Analysis : Basic Concepts and.Algorithms", 2005.

[53]. Pasquier, Nicolas. "Extraction de base pour les Règles d'Association. ", Laboratoire d'Informatique (LIMOS) — Université Clermont-Ferrand II.

[54]. M. J. Zaki, S. Parthasarathy, M. Ogihara, W. Li. "New algorithms for fast discovery of association rules", 1997.

[55]. Day, Min-Yuh. "Data mining For business Intelligence. ", Departement of Information .Tamkang University, 2013.

[56]. M.Rajalakshmi, T.Purusothaman,. R.Nedunchezhian. "Maximal Frequent Itemset Generation Using Segmentation Approach", International Journal of Database Management Systems, 2011.

[57]. Plasse Marie, Ndeye Niang, Saporta Gilbert, Villeminot Alexandre, Leblon Laurent. "Méthodes de classification pour l'extraction de règles",CNAM Laboratoire CÉDRIC, 2008.

[58]. Sylvie Jami, Tao-Yan Jen, Dominique Laurent, George Loizou, Oumar Sy. "Extraction de règles d'association pour la prédiction de valeurs manquante. ", Université de Cergy-Pontoise Cergy-Pontoise — France, 2005.

[59]. Blanchard Julien, Kuntz Pascale, Guillet Fabrice, Gras Régis. "Mesure de la qualité des règles d'association par l'intensité d'implication entropique", IRIN – École polytechnique de l'université de Nantes, 2002.

[60]. Tihi, Adil. "Mise en place d'un projet de data mining pour identifier les règles d'association", HEC Montréal, 2007.

[61]. Mansoul Abdelhak, Baghdad Atmani. "Fouille de données biologiques : vers une représentation booléenne des règles d'association", Département informatique, Faculté des Sciences, Université d'ORAN, 2009.

[62]. Hilali, Hassane. "Application de la classification textuelle pour l'extraction des règles d'association maximales", Université du Québec à Trois-Rivières, 2011.

[63]. Benrhaiem, Morched. "La lecture assistée par ordinateur : une étude exploratoire", Université du Québec à Trois-Rivières, 2015.

[64]. Descôteaux, Steve. "Les règles d'association maximale au service de l'interprétation des résultats de la classification", Université du Québec à Trois-Rivières, 2014.

[65]. Hafida, Abbas. "Expansion de la représentation succincte des générateurs minimaux", Université du Québec à Montréal, 2013.

[66]. Idiri Bilal, Aldo Napoli. "Découverte de règles d'association pour l'aide à la prévision des accidents maritime", 2013.

[67]. Feno, Daniel Rajaonasy. "Mesures de qualité des règles d'association : normalisation et caractérisation des bases, 2010.

[68]. Godin, Robert. "Les entrepôts de données et l'analyse de données", 2009.

[69]. Rahmani Rabah,. "Découverte d'associations sémantiques dans les bases de données relationnelles par des méthodes de Data Mining", Université Mouloud Mammeri de Tizi Ouzou, 2004.

[70]. Mayers, André. "Concepts de base et algorithmes : Analyse d'association", Université de Sherbrooke, 2014.

[71]. Calas, Guillaume. "Études des principaux algorithmes de data mining", EPITA, 2009.

[72]. Achouri Abdelghani,. "Extraction de relations d'association maximales dans les textes : représentation graphique", Université du Québec a Trois-Rivières, 2012.

[73]. Pagé, Christian. "Bases de règles multi-niveaux", Université du Québec à Montréal, Fevrier 2008.

[74]. Bahri Emna, Lallich Stéphane. "Introduction de l'élagage pour l'extraction de règles d'association de classe sans génération de candidats", Laboratoire ÉRIC, Université de Lyon., QDC 2009.

[75]. Verhein, Florian. "An Introduction to Frequent Pattern Growth (FP-Growth) Algorithm. s.l. : School of Information Technologies", the University of Sydney, 2008.

[76]. Jatteau, Ganael. "Approximations du treillis de concepts pour la fouille de données", Université du Québec en Outaouais, 2005.

[77]. Abroche, Nicolas. "Méthodes d'apprentissage automatique pour l'analyse des interactions utilisateurs", Université Pierre et Marie Curie., décembre 2012.

[78]. Morgan Kaufmann. Bahri, E, S. Lallich. "FCP-Growth, une adaptation de FP-Growth pour générer des règles d'association de classe. Journées francophones d'Extraction et Gestion des Connaissances", EGC , Strasbourg, 2009.

[79]. Shashikumar G. Totad, Geeta R. B,Prassad Reedy. "Batch Processing for Incremental FP-tree Construction", international Journal of Computer Applications, 2010.

[80]. G. Gasmi, S. Ben Yahia,E. Mephu Nguifo,Y. Slimani. "IGB : une nouvelle base générique informative des règles d'association", Faculté des Sciences de Tunis, 2006.

[81]. Rakesh Agrawal, Tomasz Imieliński,run Swami. "Mining association rules between sets of items in large databases", ACM SIGMOD, 1993.

[82]. Yaxin Bi, Terry Anderson, Sally McClean. "A rA rough set model with ontologies for discovering maximal association rules in document collections. ", ES2002 Conference, 2003.

[83]. Morin, Vincent. "Étude comparative d'algorithme data mining dans le contexte de jeux vidéo", UQAC, 2014.

[84]. Rioult François, Crémillieux Bruno. "Optimisation d'Extration de Motifs : une nouvelle méthode fondée sur la transposition de données", Université de CAEN, 2003.

[85]. Allia Mohamed Rachid, Bouadi Tassadit. "Fouille de données séquentielle", Université Montpelier II, 2008.

[86]. Yves Bastide, Rafik Taouil, Nicolas Pasquie,Gerd Stumme,Lotfi Lakhal. "Pascal : un algorithme d'extraction des motifs fréquents", Technique et science informatiques. Volume 21, 2002.

[87]. Lallich Stéphane, Lenca Philippe. "Règles d'association et détection de profils à risque", Angers, 2006.

[88]. Martine Cadot,. "Extraire et valider les relations complexes en sciences humaines", École doctorale Langages, Espaces, Temps, Société, 2006.

[89]. Ronen Feldman, Amihood Amir, Yonatan Aumann, , Moshe Fresko. "Maximal Association Rules: A Tool for Mining Associations in Text", Journal of Intelligent Information Systems, 2005.

[90]. Gilleron Rémi, Tommasi Marc. "Découverte de connaissances à partir de données", 2000.

[91]. Chevrin Vincent, Couturier Olivier, Engelbert Mephunguifo, Roulard José. "Recherche anthropocentrée de règles d'association pour l'aide à la décision", Revue d'Interaction homme-machine, 2007.

[92]. Rakesh Agrawel, John Shafer ,Manish Meta. "The Quest Data Mining System", KDD, 1996.

[93]. Afonso, Filipe. "Méthode de suppression des règles d'association symboliques redondantes par la régression linéaire", Université Paris-Dauphine, 1992.

[94]. Pitarch, Yoann. "Résumé de Flots de Données : Motifs, Cubes et Hiérarchies", 2011.

[95]. Djeffal, Abdelhamid. "Cours fouille de données avancées", Université Mohamed Khider Biskra, 2015.

[96]. Brijs Tom, Swinnen Gilbert, Vanhoof Koen, Geert Wets. "Using association rules for product assortment decisions: a case study", KDD, 1999.

[97]. Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava. "Selecting the right interestingness measure for association patterns", KDD, 2002.

[98]. Gwenael Bothorel, Mathieu Serrurier, Christophe Hurter. Conférence francophone sur l'Interaction Homme-Machine, 2011.

[99]. Domingos Pedro. "Mining social networks for viral marketing", University of Washington, 1994.

Annexe A – Textes utilisés pour l'extraction à partir des tweets

The White House Correspondents Dinner -- the night when Washington decides to celebrate itself.

As the President said, "Somebody's got to do it."

Ever since the first one kicked off in 1920, the political press corps, members of the Administration, the President, and a few famous faces have gathered to celebrate the importance of journalism to our democracy -- and to face a good roast over the satirical fire.

Saturday Night Live's Cecily Strong summed up the night: "The White House Correspondents Dinner is a chance for all of you to unwind, relax, and laugh as soon as you notice someone slightly more powerful than you is laughing."

But first, President Obama took his place at the WHCD podium for the sixth time, and welcomed everyone to the fourth quarter of his presidency. Here are our top three favorite moments of the night:

The base was not much to behold when the American soldiers arrived in Kunduz in 2010. Nestled atop a vast plateau, it was little more than a collection of stucco buildings with chipping paint, a small airstrip on one side, a graveyard of rusting Soviet vehicles on the other. And everywhere was the Afghan dust, so fine it would puff like dry mountain snow with every step.

In the months to follow, the Americans greatly expanded the base. Seabees, members of the Navy's construction unit, used heavy equipment to build walls from containers of dirt that encircled an area large enough to hold a second airstrip.

A small city of yellow, air-conditioned tents, with a basketball court and a chapel, rose in the field of dust. So did a sprawling maintenance bay for the armada of armored trucks.

The soldiers of the First Battalion, 87th Infantry out of Fort Drum, N.Y., could never quite fathom the reason behind the expansion, and by 2013 American forces were being withdrawn from the province.

Looking back at photos of the base, it is hard not to think of it as anything but a monument to futility.

I had gone to Kunduz with the battalion to chronicle its experiences as part of the American troop increase in Afghanistan, Gen. Stanley A. McChrystal's strategy for quelling a resurgent

Taliban. Provinces like Kandahar and Helmand in the south were focal points for the American push.

But Kunduz in the north was nearly as important, the capital of a lush — by Afghan standards — region of rice and wheat fields, a gateway for trade in petroleum, manufactured goods, drugs and weapons to Tajikistan and other former Soviet republics to the north.

It was my second time in Kunduz. I had spent a week there in early 2002 after the American invasion of Afghanistan. Kunduz had been the Taliban's last stand. Just hours before the city fell to marauding militias commanded by local warlords and aided by American air power and Special Forces, Taliban commanders were airlifted to safe havens.

The city I saw then was battered and frightened, ringed by cratered roads that took hours to navigate.

But it was also hopeful: I watched as local teachers opened a school for girls, its buildings refurbished thanks to American money.

Eight years later, Kunduz was a new world. Small stores and coffeehouses lined its streets; the bazaars brimmed with locally grown produce and fresh meat. The city felt secure.

Yet the insurgency showed its face now and again. Girls at a local school were sickened by what the police said was poison gas. A hotel that was home to Western aid workers was briefly seized by militants. Late in the battalion's deployment, a suicide bomber killed the provincial governor.

Outside the city center, evidence abounded that the Taliban and their allies were waiting patiently in the shadows. The province of Kunduz is one of Afghanistan's most ethnically diverse, with nearly equal numbers of Tajik, Uzbek and Pashtun residents, and scatterings of Turkmens and Hazaras.

Yet loyalties to the Pashtun-led Taliban remained strong in enclaves like Archi, Chardara and Gor Tepa. For months, American patrols would not venture into those districts without mine-sweeping trucks, and when they did, firefights often ensued.

I met an American farming expert whose story said much about the region's simmering fears of its once and perhaps future overlords.

The expert, Eric Imerman, a child of Iowa, had gone to Kunduz to teach modern growing techniques. I met him at a meeting between Ministry of Agriculture officials and local farmers to discuss planting winter wheat. But when Mr. Imerman pulled up in a convoy of American armored trucks, the government officials became nervous and left.

He told me later that he was stunned by the backwardness of Afghan farmers; so much knowledge seemed to have been lost during decades of war.

Yet because he had to travel under the protection of American soldiers, and because local farmers were afraid the Taliban would kill them if they were seen with Americans, he was repeatedly frustrated.

He said that while serving in the Peace Corps in the Philippines in the 1980s, "Some of the best meetings I ever had with farmers was when you sit down under a shade tree and just talk with them. No agenda, just sit and talk. And you can't do that here. »

In the final months of 2010, soldiers from the battalion pushed deeper into Chardara, Gor Tepa and Taliban enclaves near the far northern city of Imam Sahib. By spring 2011, they declared much of the province cleared of insurgent fighters.

was dubious, but one experience made me believe that perhaps they were right.

Shortly before the battalion was to return home, a friendly officer took me on a day trip with a few of his soldiers. We piled into small green pickup trucks driven by Afghan police officers and sped off onto dirt back roads I had never traveled before.

Just as we crossed a small wooded stream that seemed perfect for an ambush, we entered a wide, rolling plain.

In the distance, we could see scores of men on horses, kicking up dust in a roiling scrum. They were playing buzkashi, the Afghan national sport, in which horsemen on competing teams vie for control of a headless goat.

The contest was being run by the local police commander, who the Americans believed profited handsomely from the region's thriving weapons and drug trade. But he was the host that day, and he graciously offered us horses and a place in the game.

The hillside was lined with families who had made a daylong picnic out of the event. As I watched two soldiers trot off — one at Sancho Panza pace, the other whipping and wheeling his animal with the skill of a Texas cowboy — I could not help thinking: Maybe, just maybe, this place might find its way to peace.

In the years since their deployment, the soldiers from the First Battalion, 87th Infantry have returned to American life.

One died in a shootout with the police outside a bar. Others went to graduate school or wrote books. A wisecracking private became a tough-guy sergeant. Another recently buried a child. Some have tried to ignore the news out of Kunduz. Many others are watching in dismay.

"It's difficult to not feel a sense of meaninglessness," one former soldier wrote to me on Facebook this week. « The feeling was already there regarding Iraq. My area of operations in Iraq is under ISIS control and now Kunduz via the Taliban. You wonder what all that effort and sacrifice was for. »

He paused, and then continued writing. "I always hated the GWOT/Vietnam comparison in the past," he said, referring to the Pentagon's official shorthand for the global war on terrorism. « But, now I can't help but draw parallels. I wonder if this is the sa

was dubious, but one experience made me believe that perhaps they were right.

Shortly before the battalion was to return home, a friendly officer took me on a day trip with a few of his soldiers. We piled into small green pickup trucks driven by Afghan police officers and sped off onto dirt back roads I had never traveled before.

Just as we crossed a small wooded stream that seemed perfect for an ambush, we entered a wide, rolling plain.

In the distance, we could see scores of men on horses, kicking up dust in a roiling scrum. They were playing buzkashi, the Afghan national sport, in which horsemen on competing teams vie for control of a headless goat.

The contest was being run by the local police commander, who the Americans believed profited handsomely from the region's thriving weapons and drug trade. But he was the host that day, and he graciously offered us horses and a place in the game.

The hillside was lined with families who had made a daylong picnic out of the event. As I watched two soldiers trot off — one at Sancho Panza pace, the other whipping and wheeling his animal with the skill of a Texas cowboy — I could not help thinking: Maybe, just maybe, this place might find its way to peace.

In the years since their deployment, the soldiers from the First Battalion, 87th Infantry have returned to American life.

One died in a shootout with the police outside a bar. Others went to graduate school or wrote books. A wisecracking private became a tough-guy sergeant. Another recently buried a child. Some have tried to ignore the news out of Kunduz. Many others are watching in dismay.

"It's difficult to not feel a sense of meaninglessness," one former soldier wrote to me on Facebook this week. "The feeling was already there regarding Iraq. My area of operations in

Annexe B – Bibliothèque d'extraction des tweets

TwiterSHARP bibliothèque sous C # Sharp.

La bibliothèque Twitter Sharp propose une boite d'outil varié, pour les développeurs qui souhaitent utiliser la plateforme twitter.

Parmi les fonctionnalités qu'offre cette bibliothèque, la possibilité de retrouver la liste des tweets publiés par un utilisateur donné, le Time Line... Etc.

Avant d'utiliser cette bibliothèque il est nécessaire d'avoir un accès à un compte twitter et de se procurer des codes API pour pouvoir utiliser les informations sous Stream Twitter.

Comment connecter notre application au compte Twitter :

1. **Authentification**

1.1 On rentre l'accès token.

2.2 On rentre le mot de passe.

Voici un bout de code utilisé pour pouvoir se loguer à la base de données twitter du compte :

```
requiert 'php/twitteroauth.php';
$connect_twitter = new TwitterOAuth(get_option['api-Key'],
get_option['api-secret'], get_option['acces-token'], get_option['accès
secret']);
$ tweets = $connect_twitter->get('statuses/user_timeline', arrêt [
    'count'=> get_option('count-get-tweet'),
]);
foreach ($ tweets as $ key => $ tweet) {
    ? >
    <p>< ? php echo $ tweet->text; ?></p>
    < ? php
}
```