REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'enseignement supérieur et de la recherche scientifique

Université SAAD DAHLED- BLIDA

Département d'informatique

Mémoire de fin d'étude en vue de l'obtention d'un master en informatique

Option : ingénierie des logiciels.

Thème de recherche :

Elaboration d'un moteur de recherche par analyse de personnalité

Président jury : Chibani
Examinateurs : S. Ferfera
L. Ouhid

Elaboré par : SIDALI TEMKIT

Sous la direction de : M. REDHA SIDOUMOU

Année universitaire : 2011-2012

Liste des figures :

Liste des tableaux

Liste des tableaux

Liste des figures :

SOMMAIRE

Partie Pratique

Chapitre I : Implémentation et outils

Chapitre II Analyses et interprétations

INTRODUCTION GÉNÉRALE

« Cesse de chercher ta place dans la vie, ta place te cherche. »
Khalifa Ali. الخليفة علي رضي الله عنه

L'information est un enjeu des plus stratégiques dans un monde en accélération continue. Néanmoins, l'abondance de l'information et la multiplication des sources sont devenues au fil des dernières décennies une véritable « obstacle ». En effet, l'utilisateur lambda à la quête d'une information précise avec la contrainte de la profusion des données en même titre que celle du temps, se trouve face ne nécessité d'assistance. C'est pour cela que des moteurs de recherches ont fait leur appa n répondant ainsi à ce besoin d'*optimisation* et sont devenus des incontestables repères dans la jungle de l'information.

Par ailleurs, et paradoxalement, un effet *pygmalion1*s'est répandu dans les mécanismes de la recherche web. Dans une tentative de meilleurs rendements, les moteurs de recherche, généralisent et jumellent les différentes requêtes des utilisateurs en un seul résultat. Ce dernier qui faisant fi de la spécificité et l'individualité de chaque utilisateur est quasiment unique, appauvrissant de facto les résultats. Autrement dit, une requête émise par mille personnes donne un seul résultat tout en sachant que chacun des mille individus possèdent ses préférences personnelles.

Cette forme de carence que nous avons relevée, nous a incités à nous pencher sur ce volet qui constitue le point nodal de notre recherche. Le présent travail a pour but, donc, de proposer des résultats de recherche appropriés à chaque personne dont émane la requête pour qu'elle soit plus appropriée à chaque préférence.

Et nous aborderons ce thème en tentant de répondre à la problématique suivante :

Comment donner pour chaque personne un résultat répondant au mieux selon ses préférences ?

Pour mieux répondre à notre problématique nous avons divisé notre présent travail en deux parties composées de deux parties chacune. Dans un premier volet, nous avons abordé l'aspect théorique de notre étude par une recherche documentaire. Cette première partie se scinde en deux chapitre ; le premier est dédié à un balayage De 'histoire du développement

1 Théorie pédagogique, connue aussi sous le nom de l'effet Rosenthal, est une prophétie auto-réalisatrice qui consiste à influencer l'évolution d'un élève en émettant une hypothèse sur son devenir scolaire

des moteurs de recherche depuis leur apparition jusqu'à nos jours. Nous avons présenté les principaux moteurs de recherches en, mettant en exergue leur algorithme de fonctionnement. Le deuxième chapitre est consacré à l'exploration des différents paradigmes structurant ces moteurs de recherches tout en donnant une part léonine à l'aspect de personnalité qui est centrale pour notre recherche.

Dans la seconde partie, nous avons entamé notre recherche pratique. En premier lieu, nous avons présenté les différents outils d'implémentation utilisés dans l'élaboration de notre moteur e recherche. En second lieu, nous avons effectué des projections réalistes émanant des informations récoltées via l'étude statistique et la fouille de donné.

Nous avons tenté à travers notre recherche d'élaborer un moteur de recherche web prenant en compte les différentes préférences des utilisateurs par une dynamisation des données statiques.

Nous avons utilisé tout au long de ce travail le site web Casapro-dz.com2, comme exemple de référencement. Ce site web a été construit et référencé auprès des moteurs de recherche par nos soins dans un but commercial.

A noter que pour le mot « Laboratoire béton et sol « le site est classé premier dans la plupart des moteurs de recherche du monde avec un référencement naturel et sans préciser de pays, sachant aussi que l'un des géants mondiaux exerçants dans ce domaine a été déclassé par notre référencement.

2 http://www.casapro-dz.com

PARTIE THEORIQUE

CHAPITRE I

LES MOTEURS DE RECHERCHES : HISTOIRE ET FONCTIONNEMENT.

Avec un amas aussi considérable de données stockées et avec des toiles indénouables, l'accès à l'information ne trouve solution qu[illegible]a recherche des données.

1. La recherche en informatique :

La recherche de documents, d'informations présentes dans des documents ou dans tout corpus à caractère informatif remonte au 19iéme siècle. Cette discipline qui a l'origine de l'informatique s'appelle **Information retrieval (IR)**.

Les documents peuvent être structurés ou non structurés, cette discipline a pour but de les regrouper et de les représenter sous forme d'index3 pour mieux et rapidement les revisiter.

Les informaticiens parle de moteur de recherche en pensant à toutes implémentations informatiques qui a pour but de connaitre, de découvrir ou retrouver une informations dans un milieu riche en informations hétérogènes . Ce sont des algorithmes qui résolvent le problème de l'accès à une donnée le plus rapidement possible, en commençant par vérifier son existence au préalable, et en essayant de l'approcher selon la requête donnée[1].

1.1. Historique des moteurs de recherche :

avec l'ascension d'internet , le temps d'accès aux données ne cesse de croître avec la croissance exponentiel de ses documents , pour palier a se problème des développeur web vont mettre en place des annuaire structurant par thèmes les sites internet , avec le temps plusieurs annuaires virent le jour , certains meilleurs que d'autres , mais la croissance des données était telle que même ces annuaires devenait à leur tour difficile à consulter . C'est là

3 Un index est en toute généralité, une liste de descripteurs à chacun desquels est associée une liste des documents et/ou parties de documents auxquels ce descripteur renvoie. Ce renvoi peut être pondéré. Lors de la recherche d'information d'un usager, le système rapprochera la demande de l'index pour établir une liste de réponses. En amont, les méthodes utilisées pour constituer automatiquement un index pour un ensemble de documents varient considérablement avec la nature des contenus documentaires à indexer.

que certains développeurs pensaient a un moyen de *chercher l'information,* les annuaires sont mis à l'écart au profit des moteur chercheur qui pour une requête donnée, satisfaisait en un temps presque nulle, ces moteurs sont capables de chercher une aiguille dans une botte de foin tout en donnant des résultats probants.

Ces moteur de recherche on fait leur preuve et l'homme connu pour sa faim du besoin va aller encore plus loin et vouloir des moteurs personnalisés , où chaque personne aura son moteur de recherche qui lui proposera des résultats selon sa personnalité , son humeur du jour etc.

1.2. Historique du marché :

1. 1990 - *Archie* : Tous les moteurs de recherche descendent d'Archie, un logiciel conçu pour rechercher des documents sur internet à l'université McGill (Québec)[2].

2. 1993 - Wanderer: Le premier moteur de recherche digne de ce nom naît avec le web : il s'agit du **Wanderer** (*" le vagabond "*).

3. 1994 - Yahoo : Deux étudiants de l'université Stanford, ont une l'idée de sélectionner et recenser humainement les meilleurs sites dans un annuaire internet. **Yahoo** est né, il devient en quelques mois le portail le plus utilisé par les internautes [3].

4. 1995 - Lycos et Excite : Les machines à chercher se perfectionnent. Lycos, est mis en ligne en juin 1995[4].

5. 1996 – Altavista : Il propose de multiples fonctionnalités de recherche, notamment par langues. [5].

6. 1998 - Google : L'université de Stanford produit deux nouveaux petits génies : concepteurs d'un moteur de recherche baptisé Google. Sergei Brin et Larry Page vont révolutionner le secteur. À la fois pertinent et exhaustif, cet engin de nouvelle génération classe les résultats de recherche en fonction de leur popularité auprès des internautes. [6].

2. Fonctionnement

Les premiers moteurs (AltaVista ou Yahoo) ne faisaient qu'indexer, c'est à dire trouver toutes les pages contenant le ou les mots clés recherchés. On pouvait retrouver les pages contenant un mot clé donné mais les résultats n'étaient pas triés efficacement. Le nombre de fois où apparaissait le mot clé faisait apparaître la page en haut de la liste de résultat, ce qui n'est pas pertinent. Le nombre de répétions du mot clé n'est pas un critère intéressant car aisément falsifiable. Il faut faire une analyse plus fine du web pour être capable de mesurer automatiquement l'importance de chaque site. Sergey Brin et Lawrence Page, étudiants à Standford ont trouvé une solution aussi originale que simple : utiliser l'information des liens entre les pages pour mesurer l'importance des sites, et être alors capable de classer correctement les résultats d'une recherche de mots clés. Cet algorithme s'appelle le « PageRank ». Sergey et Lawrence ont créé l'entreprise Google qui est alors le meilleur moteur de recherche du web. Google est devenu très rapidement (et sans aucune publicité) un des moteurs les plus influant du web, les autres (Yahoo etc.) ont rapidement suivit, mais bien sur cette avancé technologique leur a permis d'être le « first to market » et donc de prendre une longueur d'avance sur la concurrence. Google essaye de garder cet avantage, en proposant toujours de réelles avancées technologiques, et bien sûr en améliorant l'algorithme PageRank qui est devenu un secret industriel aussi bien gardé que la fameuse recette exacte du coca-cola. La version que nous exposons plus bas reste une approche à la fois simple et instructive pour comprendre les bases du référencement vu que maintenant tous les moteurs de recherche utilisent des variantes de cet algorithme.

Ce qui fait la qualité d'un moteur de recherche n'est pas l'algorithme de son interface d'interrogation, mais sa maniéré de classé les données trouvé et de leur donné un poids de pertinence. Voici ici les algorithmes les plus répandu sur le marché :

2.1. **HITS :**

L'algorithme HITS s'appuie sur un principe simple : tous les sites web n'ont pas la même importance, et ne jouent pas le même rôle. Certains sites sont des "sites de référence", leurs pages sont souvent citées dans d'autres sites. Ces sites de référence sont appelés *"authorities"* dans HITS .Alors que les "authorities" sont les véritables sites qui contiennent de l'information, d'autres sites appelés "Hubs" jouent un rôle tout aussi important, bien qu'ils ne contiennent, pas, à proprement parler, de contenu informatif ; Il s'agit des sites qui contiennent des liens vers les "authorities", et qui permettent de "structurer" la Toile en indiquant où sont les pages intéressantes sur un sujet donné. [7]

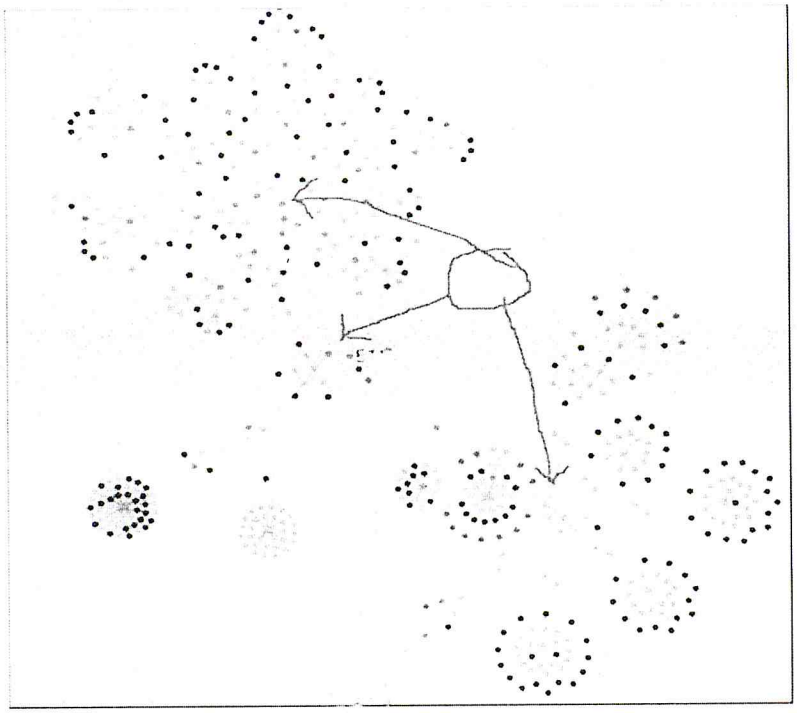


Figure 1 : illustration des authorities dans graphe web. [8]

Si l'on observe la structure des liens, un "Hub" se caractérise par la présence de nombreux liens sortants pointant vers des "authorities", tandis que les "authorities" montrent surtout des liens entrants "émanant" des hubs

L'analyse des "hubs" et "authorities" permet aussi de distinguer, sur la toile, l'existence de "communautés", c'est à dire de groupes de sites fortement liés entre eux. Un algorithme simple permet de repérer les sites concernés. C'est l'un des avantages de l'algorithme HITS

2.1.1.Algorithme :

Voici comment fonctionne l'algorithme HITS. Soit la requête d'un utilisateur notée REQ.

1. On fait d'abord une recherche classique (en utilisant par exemple un modèle vectoriel avec *TF.IDF*). On note les pages trouvées les plus pertinentes, PP.
2. À partir de l'ensemble des pages trouvées PP, on construit un plus grand ensemble LVPP qui contient :

   - les pages qui contiennent des liens vers PP.
   - les pages qu'on trouve à partir d'un lien sur une page se trouvant dans PP.

3. Une fois que LVPP et PP sont trouvés, on peut calculer la mesure « authority » A () ainsi que la mesure « hub » H () pour chaque page P appartenant a LVPP. La mesure « authority » quantifie la qualité de la page en tant que page qui reçoit des liens, alors que la mesure « hub » quantifie le statut de la page en tant que page de liens.

On peut ainsi trouver les pages qui sont des « autorités » quant à la requête de l'utilisateur. Il suffit d'offrir à l'utilisateur les pages ayant le meilleur score A().

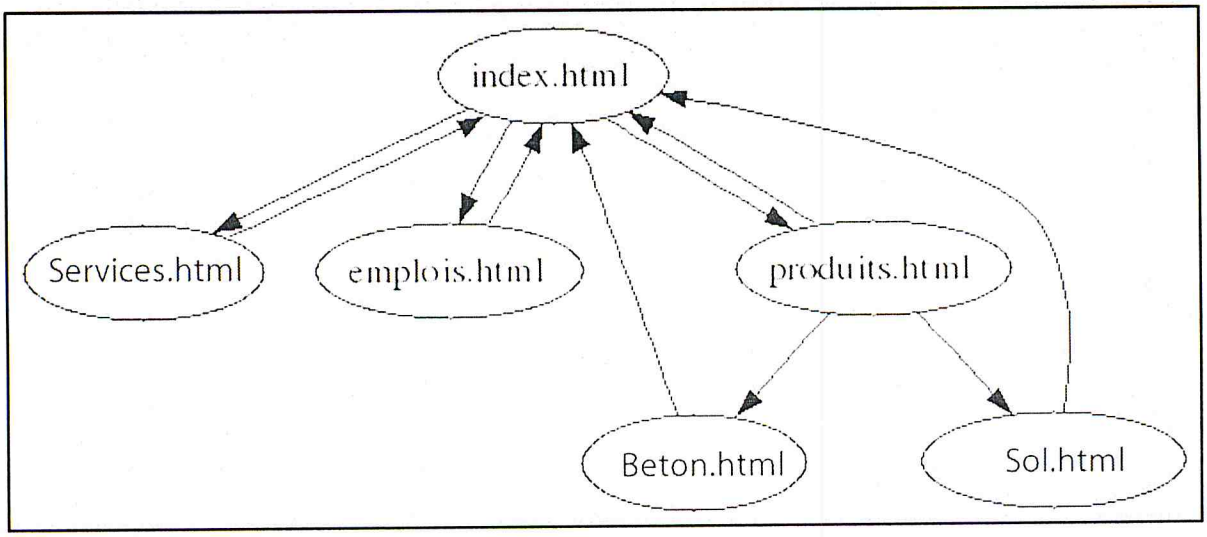


Figure 2 : schéma d'un site web pour illustration de l'exemple.

Soit les pages avec leur description :

1. index.html pointe vers 3 pages (Services.html, emplois.html, produits.html) ;
2. produits.html pointe vers 3 pages (Beton.html, Sol.html, index.html) ;
3. emplois.html pointe vers une page (index.html) ;
4. Services.html pointe vers une page (index.html) ;
5. Beton.html pointe vers une page (index.html) ;
6. Sol.html pointe vers une page (index.html).

L'algorithme HITS incitera notre utilisateur à visiter d'abord la page « index.html » et ensuite, seulement, la page « Beton.html ». La page « produits.html » ne serait pas recommandée. Pour expliquer ce résultat, on considère le sous-graphe suivant :

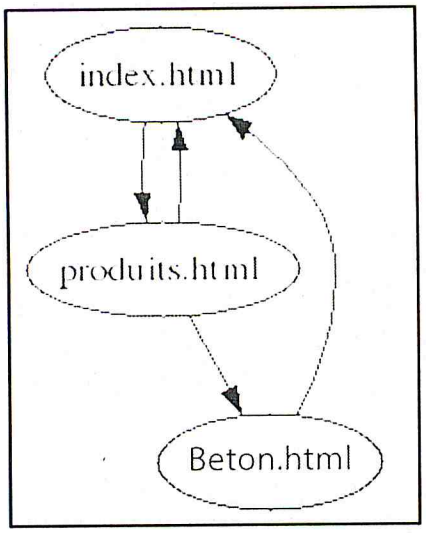


Figure 3 : Schéma du site web après exécution de l'algorithme

On voit que la page « index.html » reçoit plus de liens que toute autre page ce qui explique, en partie, qu'on la recommande d'abord.

2.1.2.Tester l'algorithme HITS :

Le moteur de recherche ask.com est basé sur l'algorithme HITS. Nous avons entré le mot clé « Laboratoires béton et sol »

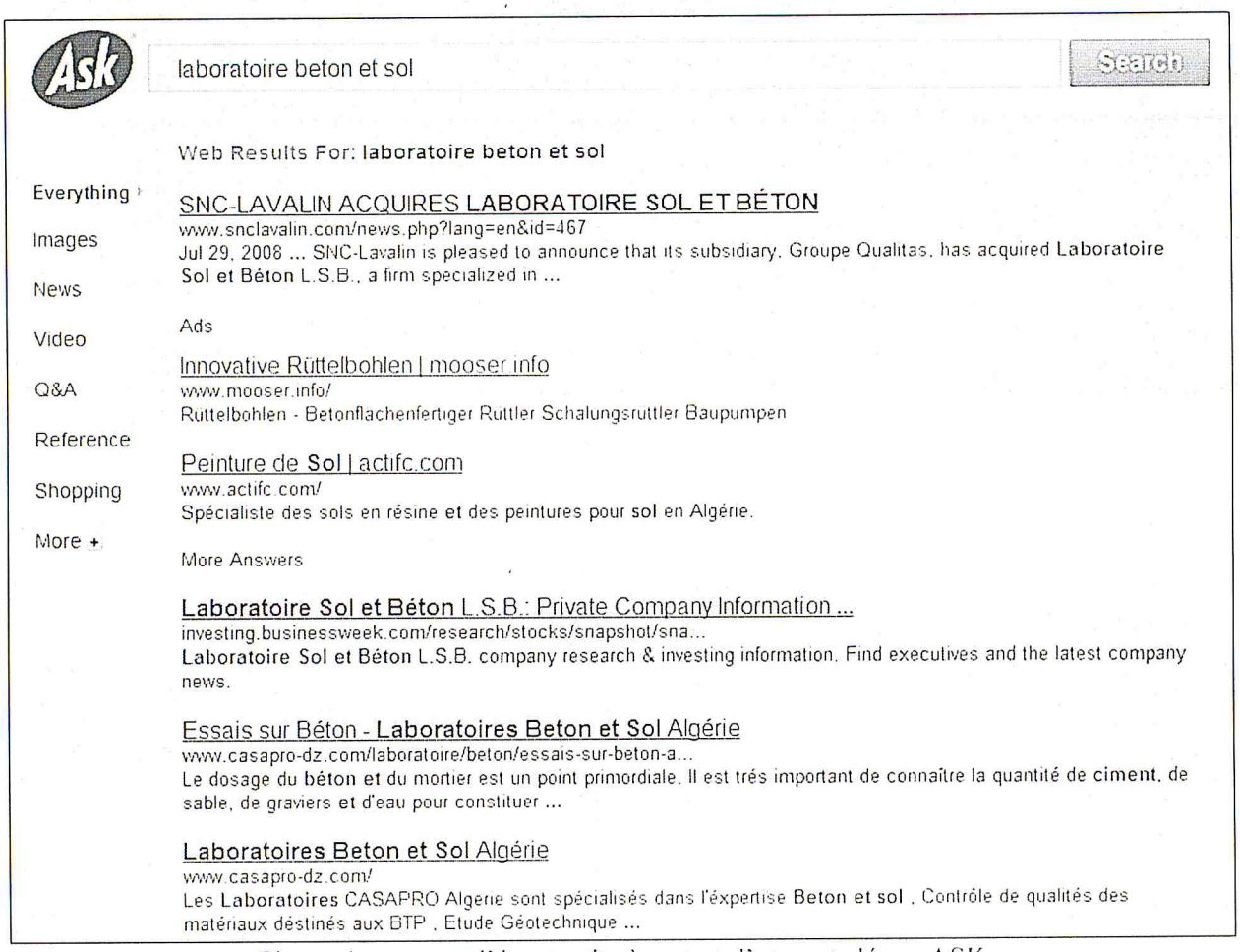


Figure 4 : capture d'écran suite à un test d'un mot clé sur ASK

On peut voir avec la figure 4, qu'ASK a commencé par nous donner le résultat de la page béton, en 5iéme position puis celui de l'index, la page béton dans notre site étant un authority plus importante que l'index.

2.2. PageRank

Nous noterons parfois PR dans la suite de ce document - est une formule mathématique. Cette méthode est utilisée par Google pour déterminer l'importance d'une page Web.

Un lien émis par une page A vers une page B est assimilé à un « vote » de A pour B. plus une page reçoit de « votes », plus cette page est considérée comme importante par Google, exactement comme le principe des élections.

Un vote émis par la page d'accueil d'un site majeur tel que Microsoft ou CNN pèse beaucoup plus lourd qu'un vote émis par la page perso de votre blog.

Retenons aussi que le PageRank est une mesure de l'importance d'une page, et non d'un site entier. Vous entendrez souvent parler de « site de rang n », il s'agit d'un abus de langage décrivant le rang de la page d'accueil du site.

Il n'y a pas, nous le verrons plus bas, de notion d'importance de site dans l'algorithme du PageRank. De même, l'importance d'une page est sans rapport aucun avec l'intérêt ou la pertinence de celle-ci, ces deux dernières notions étant totalement absentes de l'algorithme du PageRank [9].

2.2.1. **Algorithme** :

L'algorithme PageRank calcule un indice de popularité associé à chaque page web. C'est cet indice qui est utilisé pour trier le résultat d'une recherche de mot clé. L'indice est définit ainsi : L'indice de popularité d'une page est d'autant plus grand quelle à un grand nombre de pages populaires la référençant (ayant un lien vers elle). Cette définition est autoréférence car pour connaître le l'indice d'une page il faut d'abord connaître l'indice des pages ayant un lien vers elle... Il existe cependant un moyen assez simple d'approcher une valeur numérique de l'indice.

Tout d'abord il faut voir le web comme un graphe. Chaque page est un nœud du graphe, chaque lien entre page est un arc entre deux nœuds [10].

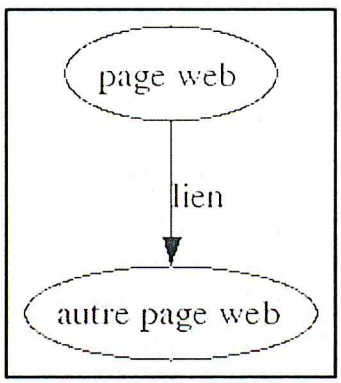


Figure 5 : illustration d'un lien entre deux pages

Les votes qu'une page A reçoit des liens émis par les pages T1...Tn. nous considérons la matrice M, matrice des transitions d'un site web, et V le vecteur de la probabilité qu'un utilisateur est sur une page.

2.2.2. Teste de l'algorithme

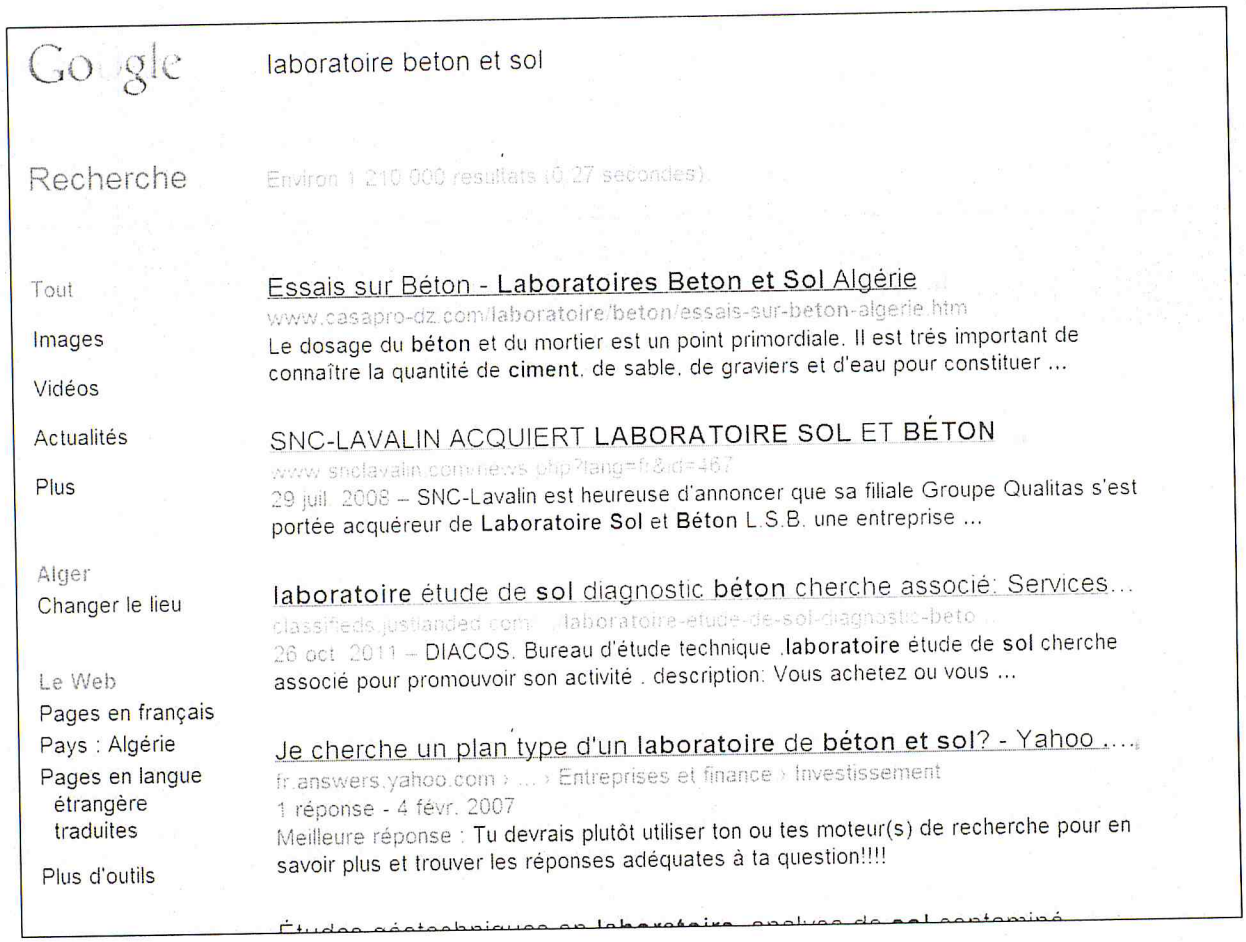


Figure 6 : Capture d'écran d'un test de requête sur Google

3. Architecture des moteurs de recherche :

3.1. Architecture générale des moteurs de recherche

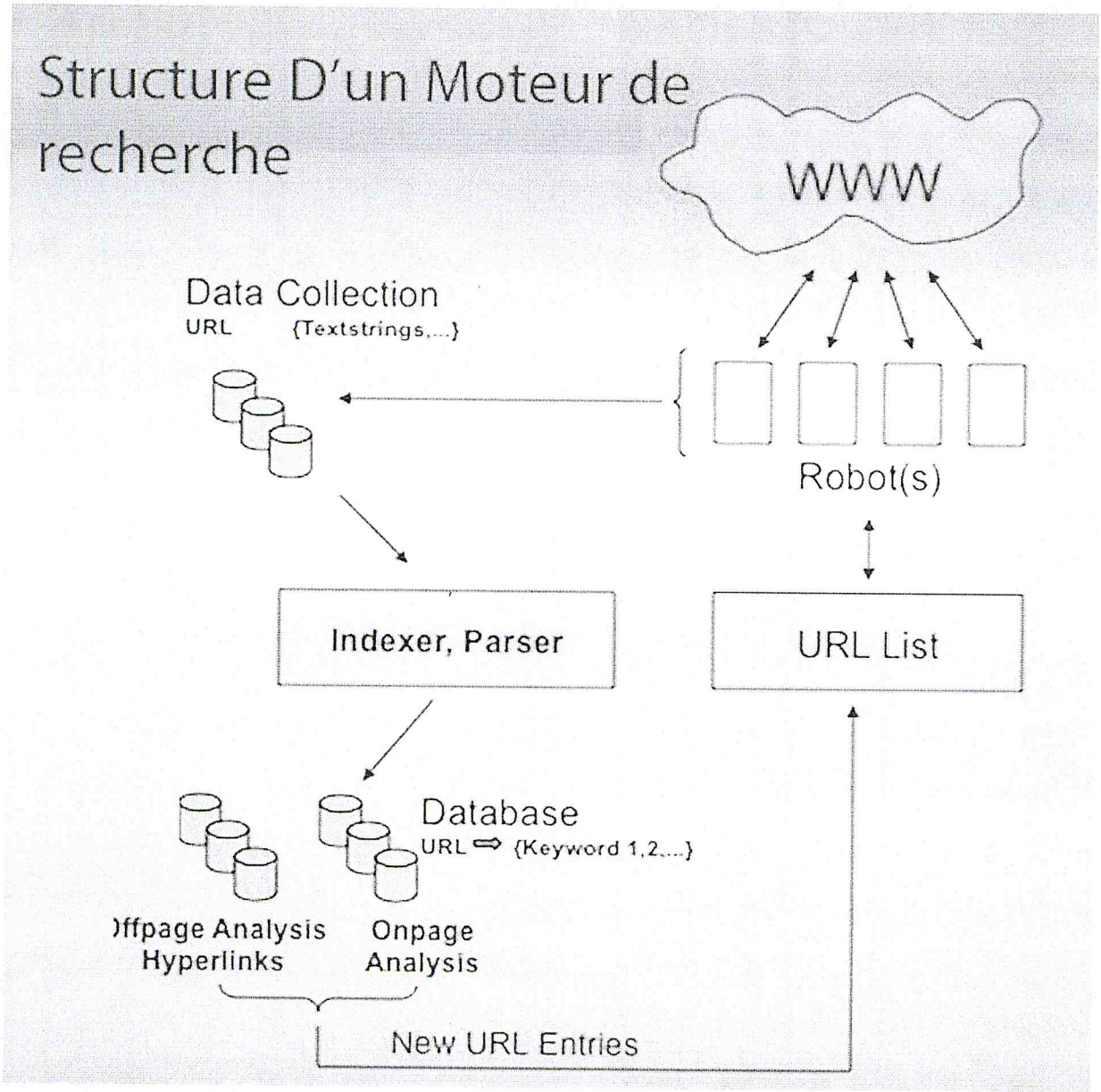


Figure 7 : Structure générale d'un moteur de recherche

Le fonctionnement de l'architecture d'un moteur de recherche est décrit dans la figure 7. Il s'agit d'une Architecture très simple qui se divise en deux parties distinctes. On retrouve d'une part, un les robots (crawlers) et d'autre part le système d'analyse des de traitement des données (Parser) récupéré par celui-ci. le crawler est considéré comme un robot chargé de rapatrier tous les documents Web contenus Sur Internet dans un index centralisé en suivant les

liens hypertextes rencontrés dans les pages analysées Ce système nécessite un matériel très avancé. Par exemple, Google utilisait un système redondant de data center éparpillé dans le monde et regroupant plus de 900 000 serveurs hyperpuissants [13].

❖ **Architectures des moteurs de recherche :**

9. Crawler

10. Indexeur

11. Parseur nettoyeur extracteur

12. Index liens

13. Interface d'interrogation

3.2. Architecture des Meta-moteurs

Les méta-moteurs présentent des stratégies de recherche beaucoup plus hétérogènes que ce que l'on peut trouver dans les moteurs de recherche classiques. Cependant, tous Ont pour point commun d'utiliser les résultats produits par ces mêmes moteurs de recherche. La plupart de ces outils ne fait que trier les liens récupérés de plusieurs sources en utilisant ses propres algorithmes de détection de pertinence. Dans ce cas, leur but est de tirer parti de la complémentarité et de la spécialisation de plusieurs outils de recherche. Dans ce cas ces outils se distinguent par les fonctions de tri utilisées qui sont propres à chaque méta-moteur. Cependant, il existe d'autres méta-moteurs qu'on peut qualifier de plus évolués. Ils ne se contentent pas de compiler les résultats d'autres moteurs de recherche mais parcourent également Internet à la recherche de documents pertinents. Il n'existe pas dans ce cas d'organisation type et chacun utilise ses propres stratégies afin de se distinguer des autres méta-moteurs [14].

CHAPITRE II

PARADIGMES DE RECHERCHE ET DE PERSONNALITÉ

Il existe plusieurs paradigmes de recherche d'informations, nous voulons dire par paradigmes la stratégie choisit pour indexer une information fiable intéressante.

1. Paradigmes de recherches :

Il existe plusieurs paradigmes de recherche appliqué au moteur chaque paradigme ayant sa propre approche.

1.1. Recherche sémantique :

La **recherche sémantique** a pour objectif d'améliorer la précision de recherche par la compréhension de l'objectif de recherche et la signification contextuelle des termes tels qu'ils apparaissent dans l'espace de données recherché, que ce soit sur le Web ou dans un système fermé, afin de générer des résultats plus pertinents.

1.1.1. La sémantique chez les géants de la recherche :

14. Bing lance une nouvelle version majeure de son moteur de recherche. Plus épurées, les pages de résultats de recherches deviennent sémantiques et sociales [15].

15. Google lui lance le Knowledge Graph, qui aidera à découvrir de nouvelles informations plus rapidement et plus facilement. [16].

1.1.2. contexte actuel de la sémantique dans les moteurs de recherche

Le web sémantique n'arrive pas vraiment à se lancer et l'intelligence artificielle rame toujours sur ce sujet car les but à atteindre n'est pas des moindres. En effet, par comparaison avec recherche syntaxique, de véritables recherches réalisées sur le web sémantique devraient être beaucoup plus conviviales pour l'utilisateur : contrairement à un moteur interrogé par une requête de mot clé appelant la fourniture de documents pertinents, un système sémantique n'impose pas à l'utilisateur de fournir les éléments de la réponse sous forme de mots-clés.

1. L'utilisateur d'un système sémantique doit pouvoir directement poser sa question en langue naturelle.

2. Un véritable moteur de recherche sémantique ne fournit pas de liste de pages répondant à une question mais la réponse précise [17].

1.2. Recherche syntaxique :

Consiste à mettre en évidence la structure d'un texte dans un document ou un dossier de document le programme informatique responsable est un analyseur syntaxique qui lui va chercher les portions de texte recherché dans un amas de texte.
Se paradigmes fut l'un des premiers paradigmes de recherche utilisé. Il reste à nos jour très intéressant a utilisé mais jumelé a d'autre paradigmes plus intelligent.

1.3. Annuaires :

Sur Internet, il n'y aucune centralisation, pas d'organisme chargé du "dépôt légal" des sites. Par conséquent, il n'existe pas d'"annuaire" général officiel ou exhaustif des sites internet ; les annuaires - ce terme désigne en réalité des répertoires de sites- "généralistes" existants sont constitués par des sociétés privées qui évaluent les sites qui leur sont soumis : il faut donc conscient lors d'une recherche d'information sur internet que les contenus des annuaires (ou répertoires de sites) ne peuvent donc qu'être partiels et subjectifs :

1. ils se limitent généralement à une zone géographique ou linguistique ;
2. le travail de classement et d'indexation est fait par des "cybers documentalistes" — ou par des internautes, dans le cas du répertoire collaboratif Open Directory qui se trouve sur DMOZ, sur la base d'une liste de propositions faites par les internautes ou les auteurs des sites ;
3. généralement, les sites inscrits sur une liste de soumission payante ont plus de chance d'être évalués (sans que cela leur garantisse d'être retenus)

Les annuaires, ou la partie "annuaire", "répertoire" ou "sélection de sites" des sites de recherche, se présentent sous forme d'une arborescence de rubriques aboutissant chacune à une liste de sites.

Sur la plupart des répertoires, on peut rechercher soit en parcourant l'arborescence jusqu'au thème cherché ou recherche par mots clés sur tout l'annuaire ou sur l'une de ses catégories. [18]

1.4. Recherche par analyse de personnalité :

Cette recherche a pour ambition de différencier entre les requetteurs et de leur donner chacun ce qu'il veut, on aura alors pour chaque requête émise autant de résultats que d'émetteurs.
La majorité des moteur de recherche ont pour une requête donné les même résultats quel que soit la personne son emplacement géographique sa culture sa langue ...etc., et c'est normale vous me dirai.
Car dans une bibliothèque pour chercher un livre, pour la requête d'un livre, tout le monde aura le même livre pour cette requête-là, on appelle cela la recherche statique et ou non intelligente, le monde d'aujourd'hui veut que pour une personne X et pour une requête R on aura X résultats. Un résultat pour chaque personne donc chaque personnalité.

1.4.1. Définition de la personnalité :

La Personnalité c'est l'ensemble des comportements qui constituent l'individualité d'une personne ; elle met en exergue l'originalité et la spécificité [19].

L'analyse de personnalité fait partir du domaine sentiment analysis en informatique.

1.4.1.1. Sentiment analysis

Sentiment analysis ou L'analyse des sentiments ou d'opinion se réfère à l'application de traitement et d'extraction des émotions et sentiment qui régissent la décision d'un individu, elle se sert de la linguistique computationnelle, et l'analyse de texte pour identifier et extraire l'information subjective dans les matériaux d'origine.

D'une manière générale, l'analyse des sentiments vise à déterminer l'attitude d'un client à l'égard de certain objet ou la polarité contextuelle globale d'un environnement. L'attitude peut être son jugement d'évaluation, l'état affectif par apport a des produits.

La montée des médias sociaux comme les blogs et les réseaux sociaux a alimenté l'intérêt pour l'analyse des sentiments. L'explosion du e-commerce a poussé les entrepreneurs aguerris à étudier l'opinion en ligne, laquelle s'est transformé en une sorte de monnaie virtuelle pour les entreprises qui cherchent à commercialiser leurs produits, identifier de nouvelles opportunités et de gérer leur réputation. Comme les entreprises cherchent à automatiser le processus d'étude de marché, la compréhension de l'environnement des clients potentiel [20].

1.4.1.2. La personnalité dans un problème de moteur de recherche ?

La connaissance de la personnalité est souvent un enjeu important en ce qu'elle permet de prévoir avec une marge d'erreur limitée le comportement de la personne dans des situations ordinaires et ou professionnelles , pouvoir connaître la personnalité d'une personne permettra de mieux la servir en prévoyant ses comportement dans divers situations ex : je suis cuisinier , et vous venez manger avec famille , tout le monde commande couscous -même requête- mais connaissant la famille , je donnerai un coucous leben au jeune , un coucous viande sans sel au parent etc.

1.4.1.3. Eléments extérieurs de la personnalité :

La première visite de l'utilisateur est la plus délicate, car il est inconnu par le système qui est dans l'impérative de le juger malgré cela. Pour mieux illustrer, nous revenons à notre exemple de serveur dans restaurant. Ce dernier, face à un nouveau client (profil), essaye à partir d'éléments préliminaires et extérieurs, de deviner ses goûts. Cette approche bien que limitée car se basant sur des a priori « trompeurs » peut avoir un résultat plus au moins satisfaisant.

Ce sont ces éléments extérieurs qui laissent entrevoir la personnalité sans la sonder. Nous dans notre contexte on pourra deviner des informations précieuse par rapport à sa région du monde, la langue de son navigateur, son système d'exploitation, tout cela nous dira le niveau d'étude, la tranche d'âge, on pourra émettre des suppositions selon sa région, etc.

1.4.1.4. Préférences :

Le monde tourne autour des préférences. Mais ces dernières sont très mal exprimées et rarement explicites. Elles ne sont bien exploitées que par les grandes industries par le biais d'un marketing puissant et manipulateur. Toutefois, il serait judicieux de se demander qu'est-ce que une préférence ? Comment se développent-elles ? Comment évoluent-elles ? Et pourquoi varient-elles d'une personne à une autre ?

Selon la définition retrouvée dans le Larousse 2010, une préférence est définie comme :

1. Considérer quelqu'un, quelque chose avec plus de faveur (que d'autres), le choisir plutôt que quelqu'un ou quelque chose d'autre : *Préférer le thé au café.*
2. Choisir comme étant mieux, meilleur, etc. : *Il préfère qu'on lui téléphone le matin.*
3. Aimer mieux quelque chose ainsi : *Je te préfère avec les cheveux longs.*

Les préférences sont intiment liées à la personnalité. Bien que ces deux notions suscitent nombre d'interactions et d'amalgames, nous précisons d'emblée que cela n'est pas le point centrale de notre étude, nous supposerons tout au long de ce mémoire que la préférence hérite de la personnalité, et par extension que la personnalité est donc une généralisation de plusieurs préférence.

La préférence se développe chez l'être humain lorsqu'il y a un choix. Lors de la présentation du choix, une décision dont une part est irrationnelle, pousserait à choisir. Ceci est communément appelé l'action de préférer, il en va de la personnalité de l'individu. Les décisions sont motivées de par l'ensemble de caractères et les tempéraments qu'englobe le terme générique de personnalité. Par exemple, une personne dite aventureuse ou hardie, par une poussée de curiosité, aura des tendances plus « osées » qu'une personne renfrognée ou intraverti qui par sa nature optera pour des décisions plus conventionnelles.

1.4.1.4.1. La concurrence Mère de la préférence :

La concurrence met au premier plan le marketing, l'intelligence économique et les choix. Ainsi, la préférence des clients va être l'élément le plus déterminant dans l'évolution de telle ou telle société.

Toutes décisions prises par un être humain portent en elle une part d'irrationalité et cela en dépit que cela de son aspect raisonné. Cette *extravagance intellectuelle* est intrinsèquement liée à la quête de l'homme de bonheur dont la ludicité fait de l'ombre à la lucidité.

Dans le cadre de notre projet (moteur de recherche) la préférence est primordiale car comme nous l'avons souligné, la préférence provient d'un sentiment irrationnel, alors que le but de naviguer dans un moteur de recherche est dans la plupart des cas rationnel. À court terme, son but est de trouver des résultats tout aussi rationnels. Notre travail consistera à ramener la courbe rationnelle à son apogée pour guider le client vers les résultats mieux probants.

1.4.1.4.2. Analyse des préférences :

1.4.1.4.2.1. Types de collecte de données

Nous pouvons recueillir les données à l'aide de trois grands types d'enquêtes : les recensements, les enquêtes-échantillon et les données administratives. Chacun présente à la fois des avantages et des inconvénients.

1) Recensement

Par recensement, on entend la collecte de données sur tous les membres d'un groupe ou d'une population. C'est ce qui se passe quand un pays veut connaitre son nombre d'habitants, des agents sillonnent le pays pour compter les personnes

Avantage :

3.1. **Variance d'échantillonnage** nulle : Il n'y a pas de variabilité d'échantillonnage à attribuer aux statistiques parce qu'elles sont calculées à partir de données sur la population tout entière.

3.2. Niveau de détail : On peut établir des données se rapportant à de petites sous-populations.

Inconvénients :

a) Coût : La tenue d'un recensement peut être dispendieuse si la population visée est nombreuse.
b) Temps : Un recensement prend plus de temps à réaliser.
c) Contrôle : La tenue d'un recensement d'une grande population est une telle démarche qu'il est difficile de conserver un niveau de minutie et de contrôle qui soit le même à chaque étape.

2) Enquêtes-échantillon

Dans une enquête-échantillon, on recueille les données auprès d'une partie seulement de la population. C'est le moyen utilisé par les chaines de télé pour comptabiliser d'audience.

Avantages

a) Coût : Une enquête-échantillon est moins coûteuse qu'un recensement puisque les données sont recueillies auprès d'une partie seulement d'un groupe de la population.
b) Temps : On obtient des résultats bien plus rapidement que dans un recensement. On communique avec moins d'unités, et il y a moins de données à traiter dans une enquête-échantillon que dans un recensement.
c) Contrôle : La plus petite envergure des activités facilite la gestion et le contrôle de la qualité.

Inconvénients

a) Variance d'échantillonnage non nulle : Il peut y avoir perte d'exactitude lorsque les données sont recueillies auprès d'une partie seulement d'un groupe plutôt qu'auprès de la population toute entière.

b) Niveau de détail : Il peut être impossible d'obtenir des renseignements sur une petite sous-population ou une petite région géographique.

C'est la manière choisit a l'aide d'outil informatique expliqué plus bas nous allons extraire les préférences des utilisateurs.

3) Données administratives

Les organismes recueillent des données administratives dans le cadre de leurs activités quotidiennes, que ces données portent sur les naissances, les décès, les mariages, les divorces, ou les immatriculations de véhicules automobiles. Dans notre pays, il existe un organisme qui est l'ONS (Office Nationale de Statistique) qui a pour but de collecter ses informations et qui les mets à la disposition du public.

Avantages

a) Variance d'échantillonnage nulle : Il n'y a pas de variabilité d'échantillonnage à attribuer aux statistiques parce qu'elles sont calculées à partir de données sur la population toute entière.
b) Séries chronologiques : La collecte de données est continue, d'où la possibilité d'analyser les tendances.
c) Simplicité : Avec des données administratives, il devient inutile de concevoir des activités de recensement ou d'enquête, ainsi que les travaux qui y sont liés[21].

Inconvénients

a) Souplesse : À la différence des données d'enquête, les données administratives peuvent se limiter à des renseignements administratifs essentiels.
b) Population : Les données se limitent à la population figurant dans les dossiers administratifs.
c) Évolution au fil du temps : la donnée administrative étant dans la plus par du temps daté [22].

1.4.1.4.2.2. Domaine étudié :

Notre moteur de recherche puise dans un intervalle de résultats bien limité : les petites annonces en Algérie. Nous avons fait le choix de faire un Meta-moteur qui sillonne les sites web d'annonce en Algérie. Cet échantillon du web est adéquat car des requêtes, des résultats et une forte affluence sont à profusion. Autre raison motivant notre choix est celle du volet matériel. En effet, il nous aurait été difficile d'atteindre notre but si nous avons ciblés la totalité du web à défaut de ressources matérielles et technique. Ce n'est qu'un géant à l'instar de Google qui a l'apanage de se permettre un projet d'une telle ampleur. On sait de ce dernier

qu'en 2011 possédait un parc de plus de 900 000 serveurs, avec des machines réparties sur 32 sites. Parallèlement, le moteur de recherche Google a indexé plus de 1 000 milliards de pages web en 2008[7]. En octobre 2010, Google représente 6,4 % du trafic Internet mondial.et enregistre une croissance supérieur à celle d'internet.

PARTIE PRATIQUE

CHAPITRE I

IMPLÉMENTATION ET OUTILS

Nous avons implémenté un aspirateur de sites web. Il s'agit d'un bot, un spider, un logiciel qui parcourt les pages web comme le ferait n'importe quel humain, mais munit d'algorithme de textmining et de datamining. Ce bot fonctionne sous le principe ETL, il extrait les données, les nettoie et les enregistre selon leur signification dans la base de données appropriées.Ce genre de moteur est connu dans le milieu anglo-saxon sous l'appellation de web-crawler, littéralement le *rampeur du web*.

1. **Outils :**

1.1. **Langage :** nous utilisons le langage JAVA pour faire notre moteur de recherche. Ce langage a été mis au point par l'entreprise Sun Microsystems (racheté depuis 2010 par Oracle). Il permet de produire des logiciels indépendants de toute architecture matérielle. c'est le langage le plus utilisé dans le monde selon TIOBE [24], avec plus de 30 milliards d'applicatif réparti dans le monde entre tout matériel électronique. Nous l'utilisons car outre son statut open source, le JAVA jouit d'une forte communauté sur la toile et son inter-compatibilité renforçant sa fiabilité

1.2. **base de données** : MySQL est un système de gestion de base de données (SGBD). Selon le type d'application, sa licence est libre ou propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public que par des professionnels.

1.3. **Framework Hibernate** : est un Framework open source gérant la persistance des objets en base de données relationnelle. Il facilite l'accès aux données en générant des *beans* qui amputent aux développeurs tout script MySQL [25].

1.4. **Api :** Une **interface de programmation** (*Application Programming Interface* ou *API*). Elle permet l'interaction entre machine algorithmique, et se base sur le paradigme de réutilisabilité .qui évite la répétition des opérations fastidieuses connues auprès des développeurs.

1.4.1. API Utilisé :

1.4.1.1. Http Client d'apache : connexion internet et récupération des pages à un niveau logiciel.

1.4.1.2. JDOM : manipulation des balise XML et HTML

1.4.1.3. Weka : l'api des algorithmes de datamining et génération des rapports

1.4.1.4. Osbcp-css-parser : pour la manipulation des CSS.

1.4.1.5. Meta data - extractor : extraction des metatdata des images web

2. Notre moteur de recherche :

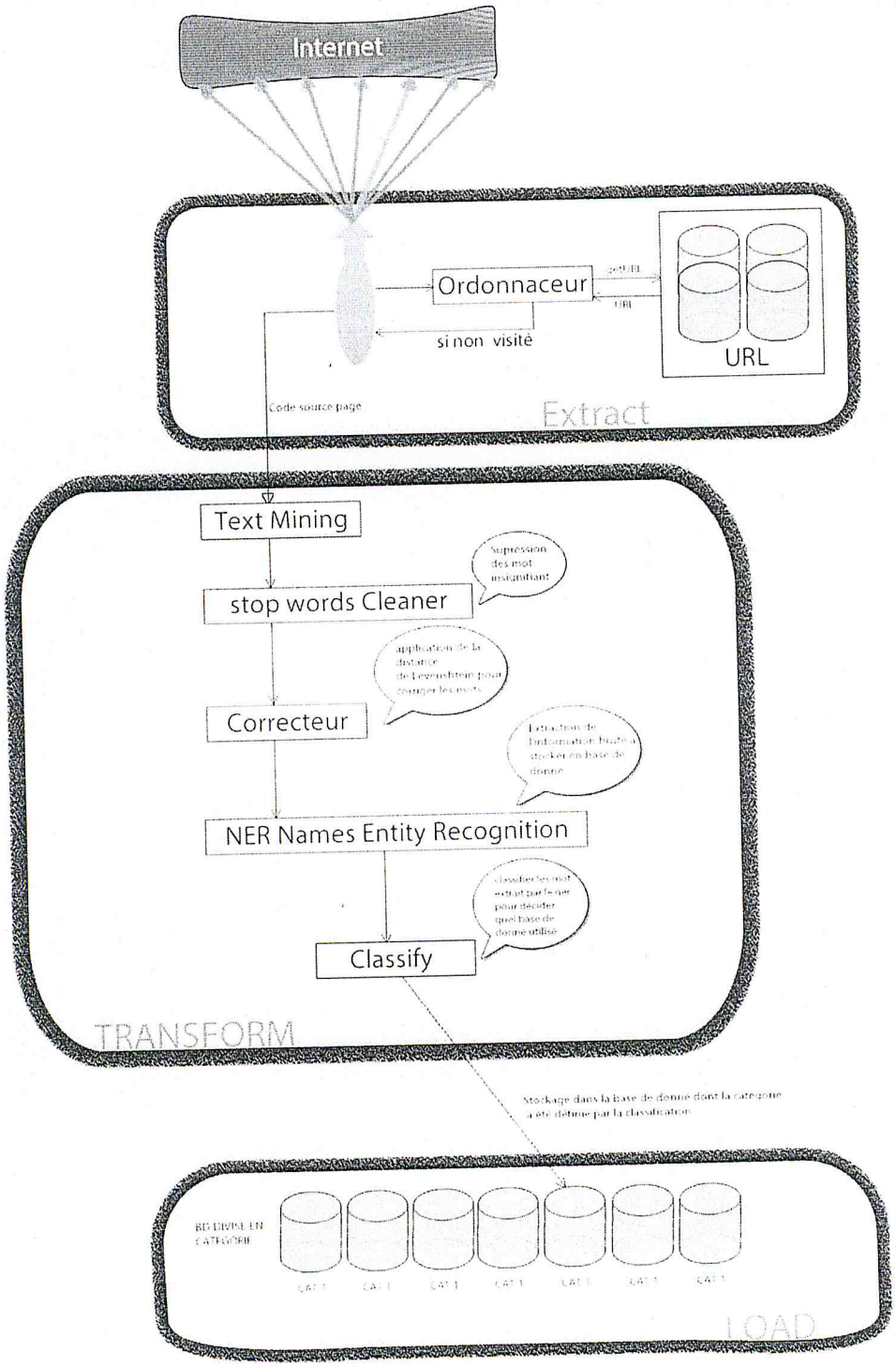


Figure 8 : Architecture générale de notre moteur de recherche ETL

ETL : sigle de **Extract, Transforme, Load :** c'est procédé que la plupart des moteurs de recherche utilise : extraction de données à partir d'une source (internet), les transformer pour pouvoir mieux l'apprécier et les stocker, puis les chargé en base de données en l'envoyant à l'index .L'ETL est un processus très utilisé pour la construction des *datawarhouse* dans les systèmes décisionnelles. Ceci n'en fait pas un paradigme dédié à cette discipline pour autant.

3. Textmining

C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. *Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.* [26]

3.1. L'apprentissage des mots : La fonction parse :

Cette fonction permet d'analyser le texte, d'en extraire les mots qui le constituent et de les mettre dans la base de donné temporaire de mots. Lorsque les mots présentent une forte redondance, ils deviennent importants dans la base de donné finale. En un mot cela est l'apprentissage. Si un mot reste à une fréquence faible après un certain temps, il sera supprimé. Nous nous sommes inspiré du fonctionnement humain afin d'établir cet algorithme. En effet, lors d'une conversation humaine, le contenu n'est retenu que par quelques mots, tout le reste des vocables, utilisés pour une meilleure compréhension en des liaisons syntaxiques et reformulations, se perd dans une mémoire à court terme.

idmots	mot	categorie	transaction	sousCategorie	freq
283630	cuisine terrasse jardin	immobilier	Location	Villa	1
283631	terrasse jardin équipé	immobilier	Location	Villa	1
283632	jardin équipé	immobilier	Location	Villa	1
283633	jardin équipé clim	immobilier	Location	Villa	1
283634	chauffage centrale li	immobilier	Location	Villa	2
283635	centrale ligne	immobilier	Location	Villa	2
283636	centrale ligne téléph	immobilier	Location	Villa	1
283637	ligne téléphonique a	immobilier	Location	Villa	2
283638	téléphonique adsl	immobilier	Location	Villa	2
283639	téléphonique adsl ré	immobilier	Location	Villa	1
283640	adsl réseau	immobilier	Location	Villa	1
283641	adsl réseau informat	immobilier	Location	Villa	1
283642	réseau informatique	immobilier	Location	Villa	3
283643	réseau informatique	immobilier	Location	Villa	1
283644	informatique acce	immobilier	Location	Villa	1
283645	informatique acce a	immobilier	Location	Villa	1
283646	acce auto	immobilier	Location	Villa	1
283647	acce auto route	immobilier	Location	Villa	1
283648	auto route prix	immobilier	Location	Villa	1
283649	route prix	immobilier	Location	Villa	4
283650	pins maritimes moha	immobilier	Location	Villa	1
283651	maritimes mohamma	immobilier	Location	Villa	1
283652	maritimes mohamma	immobilier	Location	Villa	1
283653	mohammadia alger	immobilier	Location	Villa	3

Figure 9 : Base de donné de mot

Exemple :

"L'**islam** est une religion monothéiste apparue en presqu'ile arabique au VII^{e} siècle. Le prophète de l'islam est Mohamed (SPSSL) .La religion de l'islam dite religion musulmane se veut une révélation en langue arabe de la religion originelle d'Adam, de Noé, et de tous les prophètes parmi lesquels elle place aussi le prophète Jésus (appelé Îsâ dans le Coran). Ainsi, Le livre sacré de l'islam est le **Coran**. Le Coran est le dogme de l'islam, le Coran contient le recueil de la révélation d'**Allah**, transmise oralement par son prophète Mohamed (slaws) .Le Coran reconnaît l'origine divine de l'ensemble des livres sacrés du judaïsme et du christianisme, tout en considérant qu'ils sont, dans leurs écritures actuelles, le résultat d'une falsification. "

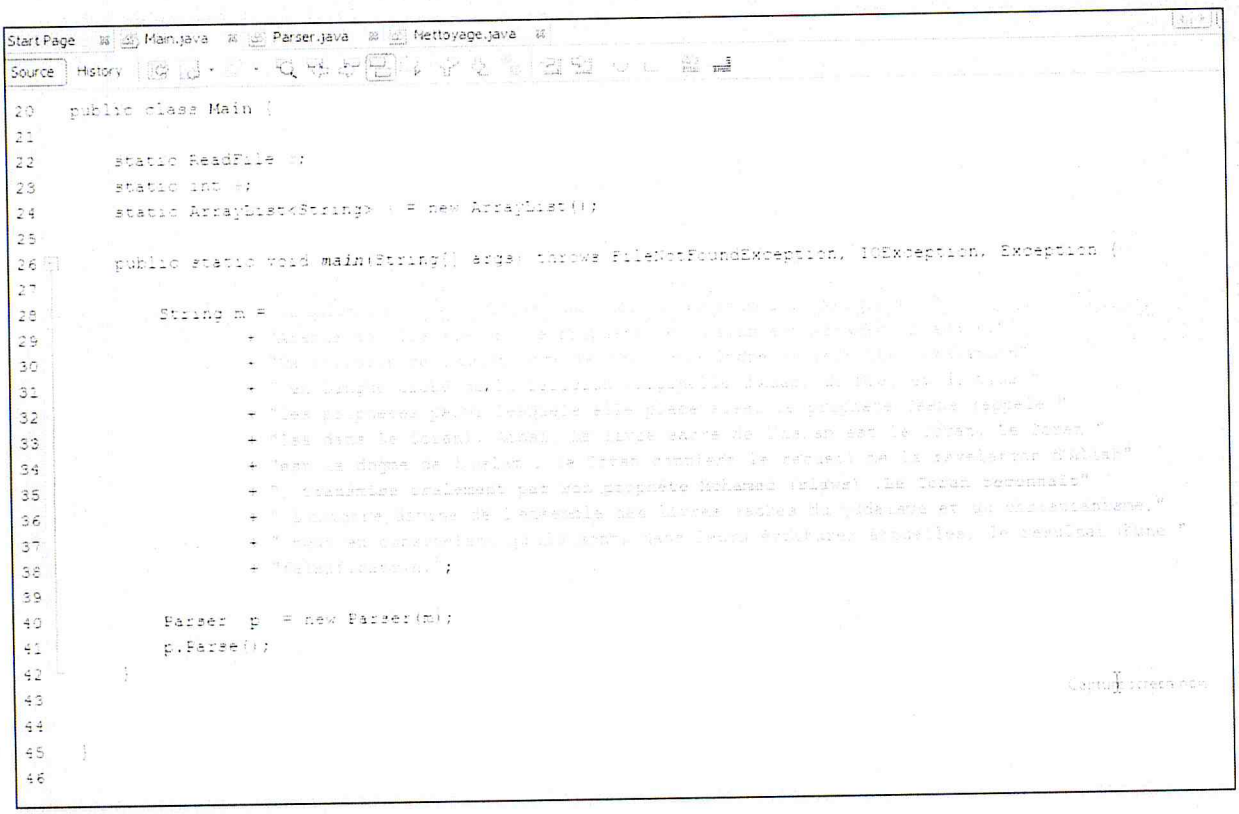


Figure 10: Exemple de classification de texte

Filter:

idwords	words	frequency
1	L'islam	5
2	est	4
4	religion	4
12	prophète	3
3	une	2
13	Mohamed	2
14	(islam)	2
18	révélation	2
31	dans	2
5	monothéiste	1
6	apparue	1
7	à	1
8	Mecque	1
9	Arabie	1
10	Ville	1
11	siècle.	1
15	La	1
16	dite	1
17	musulmane	1

Figure 11: Fréquence des mots d'un texte exemple.

Chaque mot apparaît comme une clef associée au nombre d'apparition de ce mot dans le texte analysé.

On va considérer par la suite que la valeur d'un mot correspond au nombre d'apparition dans un document ou dans une catégorie. Les paramètres de ces tableaux (clef-valeur) sont les mêmes qui seront utilisés pour la classification et l'apprentissage.

3.2. La fonction addwords

Cette fonction permet d'ajouter les différents mots ainsi que leurs valeurs associées à la catégorie correspondante. Dans un premier temps, elle fait appel à la fonction *parse* et ajoute le texte à la catégorie correspondante. Nous nous retrouvons dans ce cas avec une clef de type (catégorie-mot) et la valeur associée. Elle est liée à la phase d'entraînement du classificateur.

3.3. La fonction classify

Cette fonction permet de classer le document (après un entraînement préalable du système). Elle fait appel à la fonction *parse* le qui fournit une table hachée. Elle se charge également de tous les calculs de probabilité à effectuer pour associer le degré d'appartenance du document à la catégorie correspondante.

Avant de connaître la probabilité P(Ci), qui n'est rien d'autre que la probabilité d'une catégorie Ci, nous calculons le nombre total de mots dans la catégorie Ci et le nombre total de mots dans toutes les catégories. Le rapport entre ces deux valeurs nous donne la probabilité recherchée. Comme précisé plus haut, le calcul de P(D|Ci) se réduit au calcul suivant :

$$P(D|Ci) = P(W1|Ci) * P(W2|Ci) * ... * P(Wm|Ci)$$

Pour comparer deux motifs de texte nous nous ne sommes pas contentés de la célèbre fonction *compareTo* de la classe JAVA String. Mais nous implémentons un algorithme dit la distance de Levenshtein. Ce dernier reconnait les mêmes mots même s'ils diffèrent de graphie par exemple, variations de langue, français belge, canadien, algérien, d'évolution synchronique de la langue et la conjugaison des verbes, etc.

3.4. La distance de Levenshtein :

Elle nous permet de cerner le degré de similarité entre deux chaines de caractères. Elle est égale au nombre minimal de caractères qu'il faut modifié (*supprimer*, *insérer* ou *remplacer)* pour passer d'une chaîne à l'autre.

Nous allons établir une constante p = 2, qui est par proposition logique et statistique. Le nombre d'erreur en moyenne dans un même terme à condition que le terme contienne plus de 5 caractères. Si entre deux mots la distance de levenshtein est inférieure à 2 donc les deux mots sont similaires.

- **avantages:** Le principal réside dans son très faible coût d'entretien à l'égard au véritable service que cela peut apporter, à condition que le volume du corpus documentaire soit significatif, voire très important.
- **Les désavantages:** Le revers de la médaille, c'est qu'il n'y a pas de prise en compte des spécificités du corpus documentaire traité : textes médicaux, commerciaux, scientifiques ou autres, seront adressés de manière identique, grâce à la puissance du calcul statistique. Autre élément à prendre en compte, c'est la pertinence du traitement qui est non seulement difficilement prévisible, et en tout cas généralement moins élevée que l'approche sémantique.

4. ETL

Extract-Transform-Load est connu sous le terme ETL, (ou parfois : datapumping). Il s'agit d'une technologie informatique permettant d'effectuer des synchronisations massives d'information d'une source de donné de données vers une autre.

4.1. ETL : EXTRACT

Nous allons pour extraire nos données implémentées, un crawler, une entité -robot- pouvant surfer sur le web est ramasser le maximum voire la totalité de données consultées. Cette entité sera capable en un temps modeste sinon médiocre de consulter une base de données locale de sites et aller surfer sur chacun d'eux pour en extraire les données voulues et enfin les faire passer à la deuxième phase de traitement qui est la transformation

Pour ce faire nous avons utilisé l'api HTTPCLIENT d'apache.

Grâce à une boucle while (true), notre crawler sillonne le web sans interruption en récupérant les codes sources HTML des pages web.

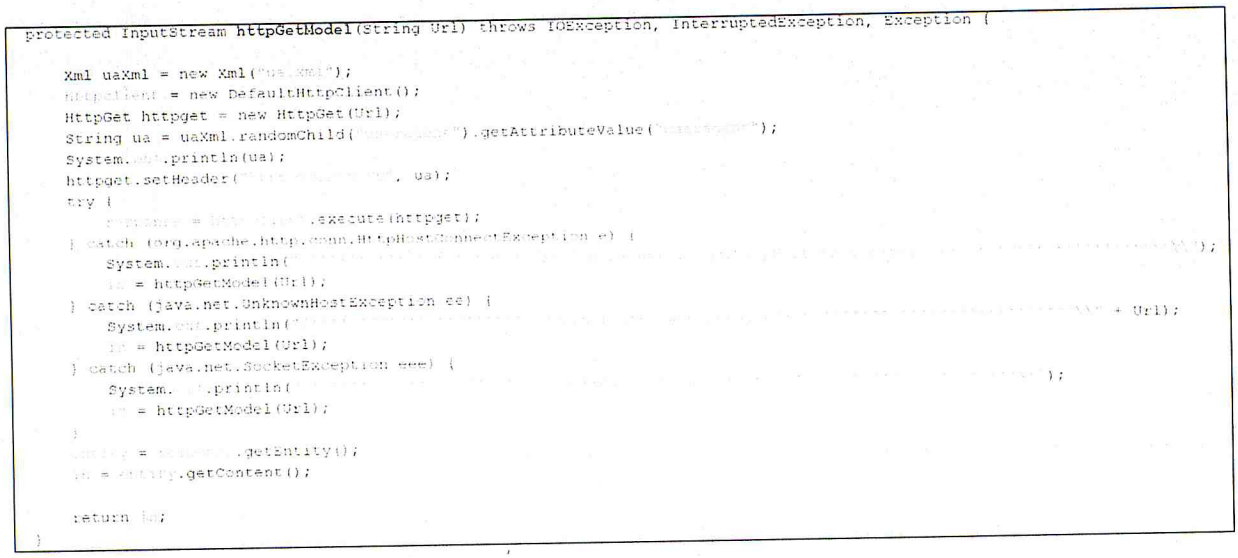


Figure 12 : code java de la fonction responsable de la récupération http.

4.2. ETL : TRANSFORM :

Au même moment où l'extraction se fait dès que la première donnée rapatriée -extraite-, les phases de transformation commencent. Il faut savoir que notre donnée arrive au tout début sous forme de page web, le nettoyage consistera à supprimer tous les balises et ne garder que le texte.

4.2.1. Nettoyage des mots vides:

En recherche d'information, les **mots vides** (ou *stop words*, en anglais) sont des mots qui sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche. En français, des mots vides évidents pourraient être « le », « la », « de », « du », « ce », « ça »..

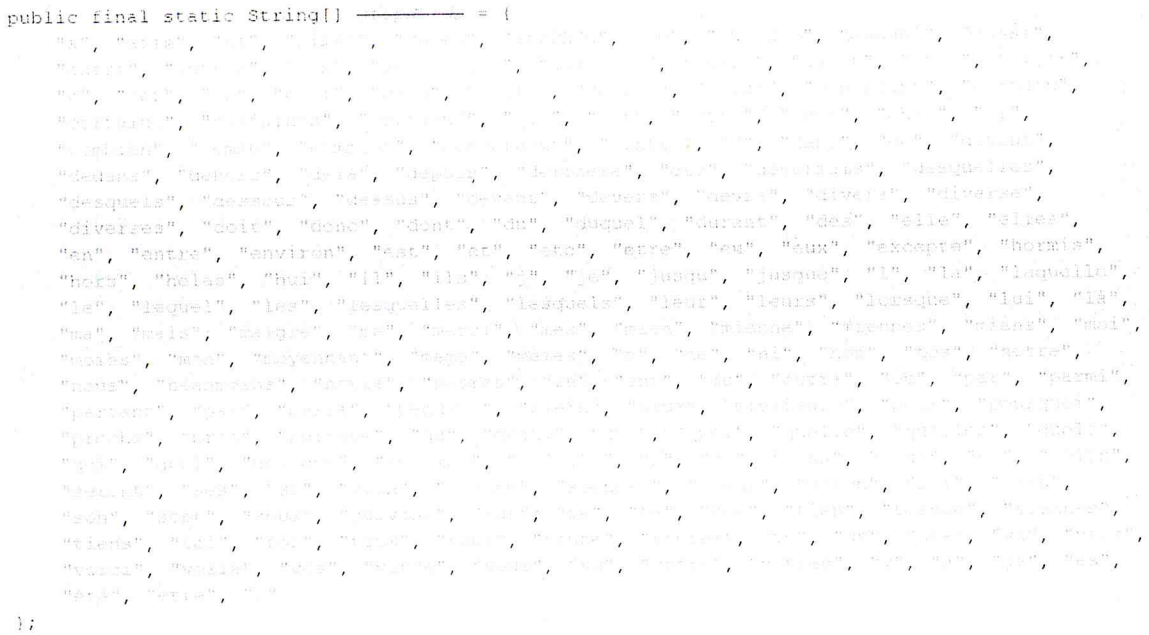


Figure 13 : liste des mots vides utilisés pour le nettoyage des textes.[25]

Un mot vide est un mot non significatif figurant dans un texte. La signification d'un mot s'évalue à partir de sa distribution (au sens statistique) dans une collection de textes. Un mot dont la distribution est uniforme sur les textes de la collection est dit « vide ». En d'autres termes, un mot qui apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant et ainsi il ne permet pas de distinguer les textes les uns par rapport aux autres.

Dans cette étape de suppression des mots très fréquent du français qui sont communs à tous les textes et dont l'importance est faible sera effectué.

4.2.2. Extraction de la connaissance :

A partir de nos pages web, maintenant devenu des textes nettoyés, nous allons entamer l'étape d'extraction des informations avec un algorithme nommé NER, **named entity recongnition** responsable d'extraire les connaissances du texte c'est-à-dire, les informations que nous voudrons stocker et qui valorise le document en soit

4.2.2.1. NER : Named Entity Recognition :

Le terme «entité nommée» (NER) est en cours d'utilisation dans l'extraction d'information (IE) des applications. Il a été inventé lors de la Conférence Message Understanding sixième (MUC-6) [26]

L'algorithme NER a pour fonction d'effectuer des tâches dans lesquelles des informations structurées sur la société et des activités liées sont extraites du texte non structuré, comme les articles de journaux.

Il est essentiel de reconnaître les unités d'informations telles que les noms, y compris de personne, les noms d'organisation, l'emplacement, des expressions numériques, y compris le temps, la date, l'argent, et pourcentages. Le terrain a été reconnu en 1996, d'importantes recherches ont été menées par l'extraction de noms propres à partir de textes. Reconnaissance des entités nommées,

La technique utilisée en textmining pour extraire des informations dont le format est déjà connu. Exemple d'entité nommé : les adresses de rue ont toujours le même format, un nom propre et un numéro de rue suivi d'un nom de localité plus nom de province plus nom de pays : Bend oued Ahmed 03 Hadjout Tipaza Algérie.

Il serait plus pratique grâce à des algorithmes de reconnaissance de forme (ici la forme = texte) de reconnaitre une adresse sans trop faire tourner un algorithme des heures.

Les NER sont utilisés pour extraire les numéros de téléphone, les adresses, les noms propres etc. tout entité ayant un format standard.

4.2.2.2. Technique d'implémentation d'un NER :

4.2.2.2.1. expressions régulières :

Technique très ancienne qui s'inspirent de la compilation, et de la manière dont les compilateurs qui eux même implémentent des NER.

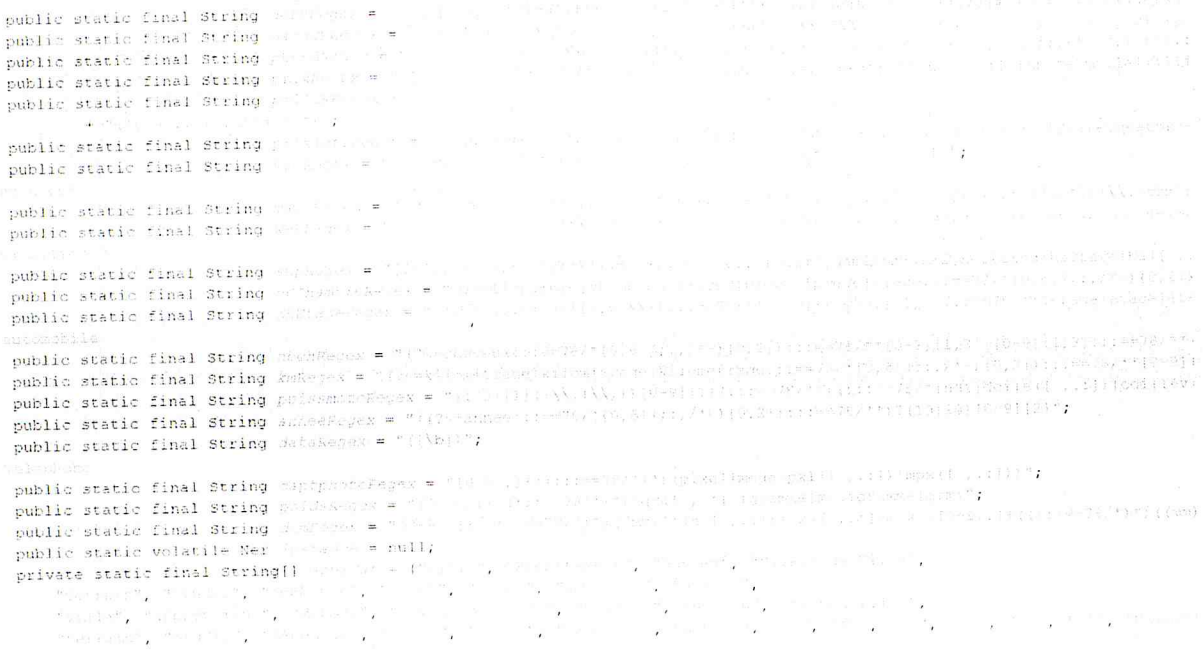


Figure 14 : Implémentation des expressions régulières du NER .

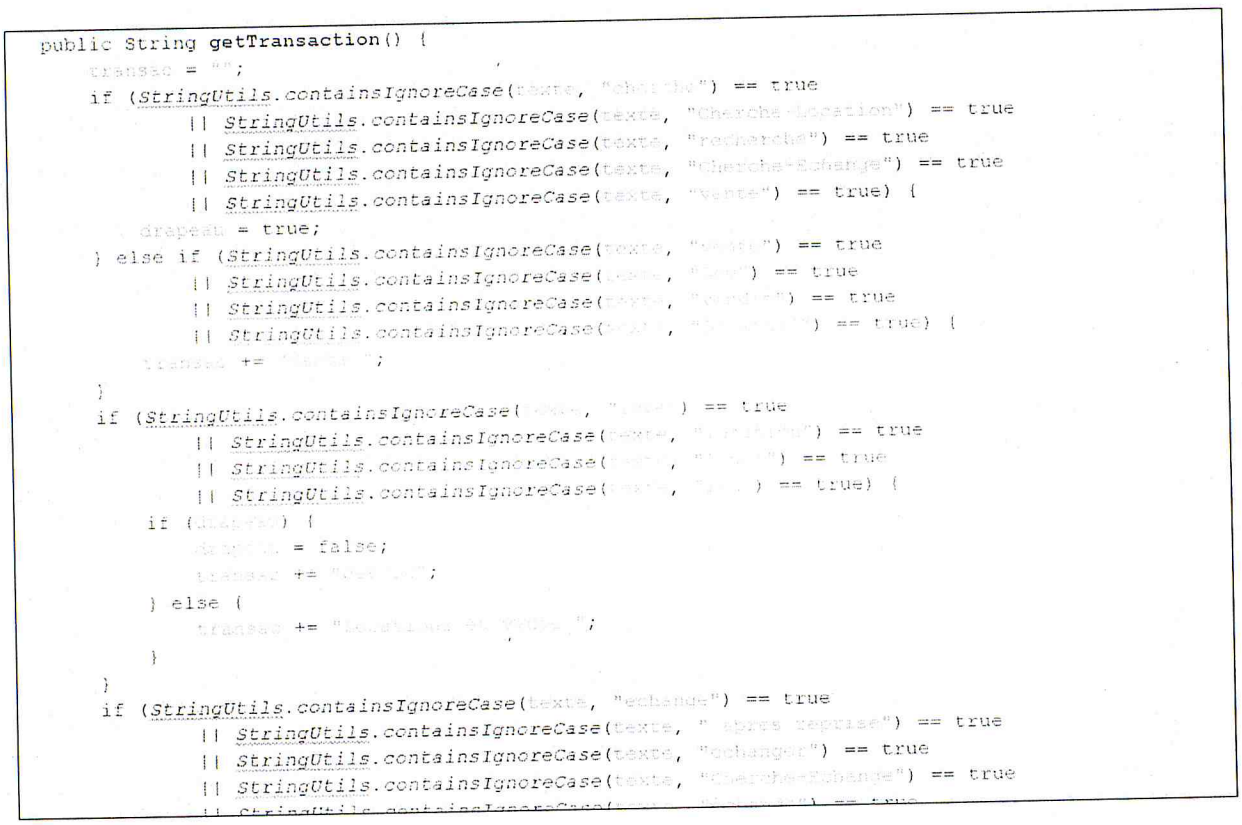


Figure 15 : Implémentation du NER par expressions régulières avec comparaison algorithmique

4.2.2.2.2. dictionnaires :

On utilise des dictionnaires de termes que l'on comparera aux termes suspecté.

Exemple :

Dictionnaire des commune, lorsqu'un terme est suspecter d'être une commune on le comparera au dictionnaire si il y est alors c'est une commune.

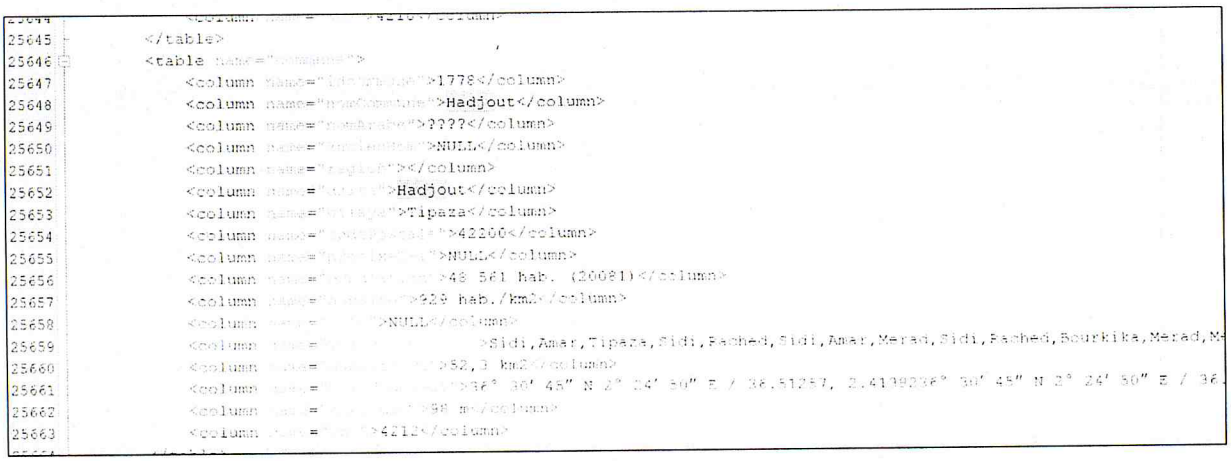


Figure 16 : Implémentation du NER par dictionnaires

4.3. ETL : Load :

Aucun stockage anarchique, on devra avant de passer a la base de donné, passer par un algorithme de classification pour savoir quel type de donné avant nous la, selon des thèmes bien précis limité par nos soin, la classe de donné démasqué ils seront alors inséré dans leur base de donné relatif.

La classification se fera par un algorithme très simple calculant la distance entre les mots des donné et les mots des classe existante (bd) la distance se calculé par le nombre de mot similaire entre donné et classe, plus le nombre est grand plus on a de chance que se soit la classe de donné.

4.3.1. Notre base de donné :

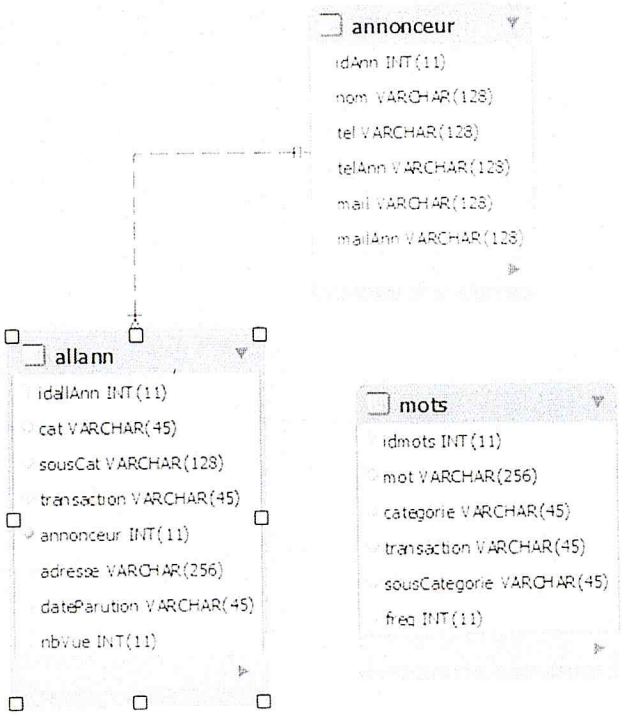
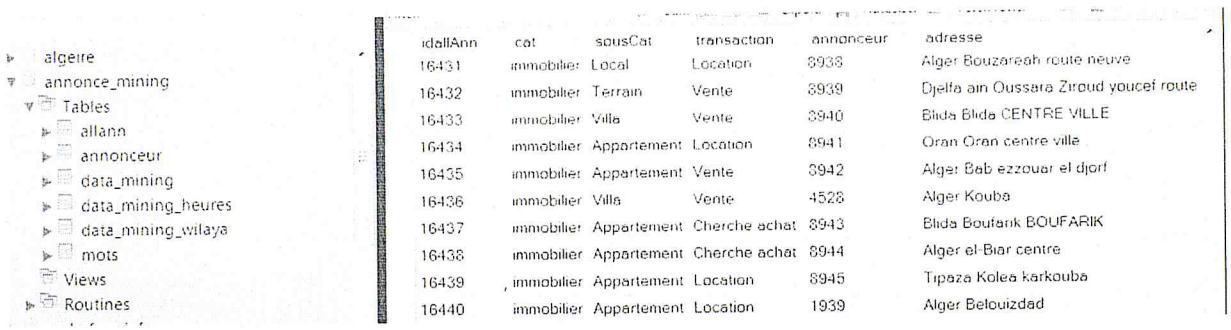


Figure 17 : Diagramme de classe de notre base de donné

4.3.1.1. Base de donné des annonces :

Elle contiendra toute les annonces nettoyées, classifié et indexé, notre algorithme NER aura pu déduire la sous-catégorie, la transaction et la ville de la transaction, puis le prix et l'adresse de l'annonceur qui a posté cette annonce il indexera aussi les informations sur l'annonceur pour une étude statistique plus tard.



idallAnn	cat	sousCat	transaction	annonceur	adresse
16431	immobilier	Local	Location	8938	Alger Bouzareah route neuve
16432	immobilier	Terrain	Vente	8939	Djelfa ain Oussara Ziroud youcef route
16433	immobilier	Villa	Vente	8940	Blida Blida CENTRE VILLE
16434	immobilier	Appartement	Location	8941	Oran Oran centre ville
16435	immobilier	Appartement	Vente	8942	Alger Bab ezzouar el djorf
16436	immobilier	Villa	Vente	4528	Alger Kouba
16437	immobilier	Appartement	Cherche achat	8943	Blida Boufarik BOUFARIK
16438	immobilier	Appartement	Cherche achat	8944	Alger el-Biar centre
16439	immobilier	Appartement	Location	8945	Tipaza Kolea karkouba
16440	immobilier	Appartement	Location	1939	Alger Belouizdad

Figure 18 : Base de donné après transformation pour les annonces

4.3.1.2. Base de donné des annonceurs :

nos algorithme de nettoyage et de récupération auront pu déduire les annonceurs avec leur numéros de téléphone et email puis chacun indexé avec les annonce qu'il a posté pour une étude statistique ultérieur sur les annonce aimé par cette annonceur.

Figure 19 : Base de donné après transformation pour les annonceurs

4.3.1.3. Base de donné des mots :

Utilisé pour posté une annonce, cette base de donné nous sera utile lots de la classification des sous-catégories.

idmots	mot	categorie	transaction	sousCategorie	freq
283630	cuisine terrasse jardin	immobilier	Location	Villa	1
283631	terrasse jardin équipé	immobilier	Location	Villa	1
283632	jardin équipé	immobilier	Location	Villa	1
283633	jardin équipé clim	immobilier	Location	Villa	1
283634	chauffage centrale li...	immobilier	Location	Villa	2
283635	centrale ligne	immobilier	Location	Villa	2
283636	centrale ligne téléph...	immobilier	Location	Villa	1
283637	ligne téléphonique a...	immobilier	Location	Villa	2
283638	téléphonique adsl	immobilier	Location	Villa	2
283639	téléphonique adsl ré...	immobilier	Location	Villa	1
283640	adsl réseau	immobilier	Location	Villa	1
283641	adsl réseau informat...	immobilier	Location	Villa	1
283642	réseau informatique	immobilier	Location	Villa	3
283643	réseau informatique ...	immobilier	Location	Villa	1
283644	informatique acce	immobilier	Location	Villa	1
283645	informatique acce a ...	immobilier	Location	Villa	1
283646	acce auto	immobilier	Location	Villa	1
283647	acce auto route	immobilier	Location	Villa	1
283648	auto route prix	immobilier	Location	Villa	1
283649	route prix	immobilier	Location	Villa	4
283650	pins maritimes moha ...	immobilier	Location	Villa	1
283651	maritimes mohamma...	immobilier	Location	Villa	1
283652	maritimes mohamma...	immobilier	Location	Villa	1
283653	mohammadia alger	immobilier	Location	Villa	3

Figure 20 : base de donné de mots pour la classification

CHAPITRE II

ANALYSES ET INTERPRÉTATIONS

1. Datamining

Apres avoir indexé une bonne partie du web annonce algérien, nous nous attaquons maintenant aux personnes qui ont posté ses annonces, nous allons les classifier pour crée des profils au préalable, comme cela dès sa première utilisation un requêteur sera classé selon sa catégorie et déjà là il aura un résultat selon la majorité de la personnalité de sa catégorie.

1.1. Définitions :

En Français Fouille de données. Terme récent (1995) représentant un mélange d'idées et d'outils provenant de la Statistique, l'Intelligence artificielle et l'Informatique. Le datamining est un processus de découverte de règle, relations, corrélations et/ou dépendances à travers une grande quantité de données, grâce à des méthodes statistiques, mathématiques et de reconnaissances de formes. [27]

1.2. Intérêt économique :

Du produit aux clients. Technologie de l'information : faible coût de stockage de données, saisie automatique de transaction (code bar, click, données de localisation GPS, internet) Augmentation de la puissance de calculs des ordinateurs (loi de Moore chaque 18 mois pour le même prix on obtient le double de performance) Extraire de la connaissance à partir des de grande bases de données devient possible

1.3. étude statistique :

Une bonne approche de fouille de donné repose sur une bonne étude statistique qui déjà va dévoiler quelque donné caché.

1.3.1. Définition de l'étude statistique :

C'est une discipline scientifique à part entière servant à décrire les caractéristiques scientifiques d'une situation. [28]

Objectif :

7. Lire une information chiffré et ou difficilement percevable a l'œil nue .
8. Expliquer des phénomènes
9. Comprendre des situations et mener des enquêtes dans un but commerciale et ou pédagogique
10. Exemple d'une statistique

1.3.2. Une étude statistique sur notre base de donné :

1.3.2.1. étude statistique des annonces :

voici une étude très importante dans notre cas , on pourra en mont de tout préoccupation des requêteur savoir quel sous domaine sont plus prisé que d'autres et pour quel transaction , par exemple l'on pourra constaté que la vente d'appartement est beaucoup plus prisé que la location de bungalow a cet saison de l'année (base de donné récolté en janvier 2012) E- la même chose ici une étude statistique qui reprend la première avec le critère requêteur et plus précisément sa localité, nous aurons quel sous-catégorie est plus prisé que d'autre dans chaque localité. A titre illustratif, nous pouvons voir que la recherche d'achat d'appartement à Alger est plus fréquente qu'à Annaba. () Donné réel. Tout cela nous permet de cibler les résultats selon la localité du requêteur si on la connait sinon nous ferons abstraction de cela est nous utiliserons la première étude. En d'autres termes des profile seront bien fait.

Légende des schémas qui vont suivre :
En bleu les annonce non signifiante n'ayant pas attiré l'attention.
En rouge les annonce ayant attiré l'attention.

1.3.2.1.1. Distribution des wilayas

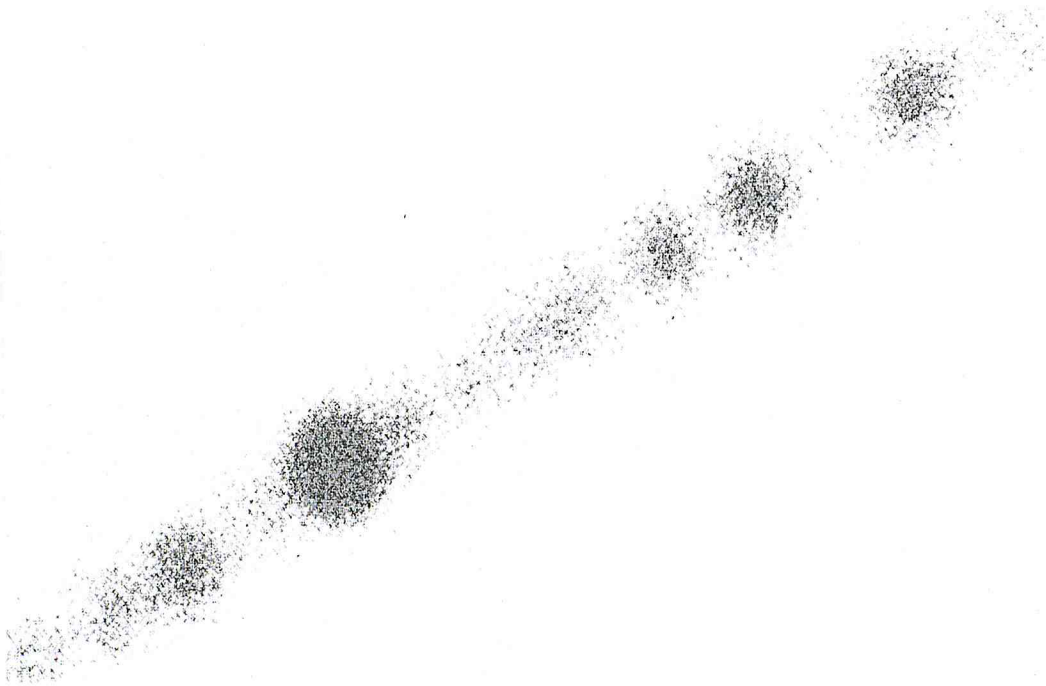


Figure 21 : distribution des wilayas

11. Abscisses : wilaya de 1 à 48

12. Ordonnés: wilaya de 1 à 48

Lecture :

Statistique sur les wilayas des annonceurs : on peut voir sur ce schéma que les wilayas 16 35 et 42 sont les wilayas les plus prisé

1.3.2.1.2. Distribution des catégories sur les wilayas

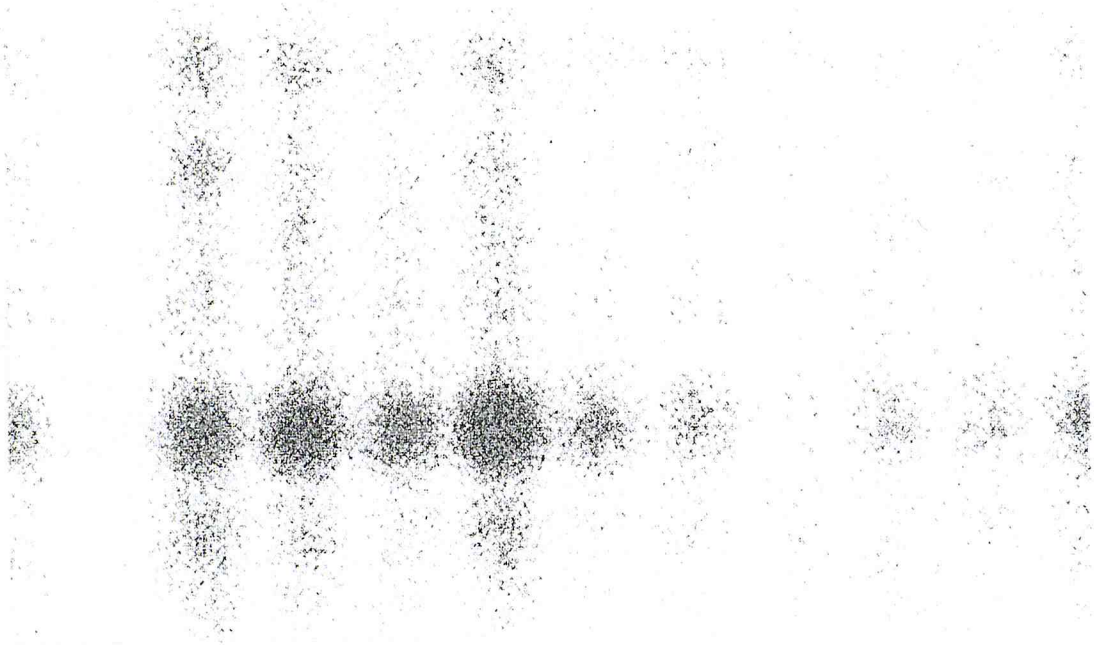


Figure 22 : distribution des catégorie par rapport au wilaya

- **Abscisses** : catégorie
- **Ordonnés** : wilaya

Lecture :

Ici le schéma des sous-catégories par rapport aux wilayas, on peut voir que la sous-catégorie 2, 3, 4,5 sont les sous-catégories les plus populaire avec 3,4 plus de déchets.

1.3.2.1.3. Distribution des transactions sur les wilayas

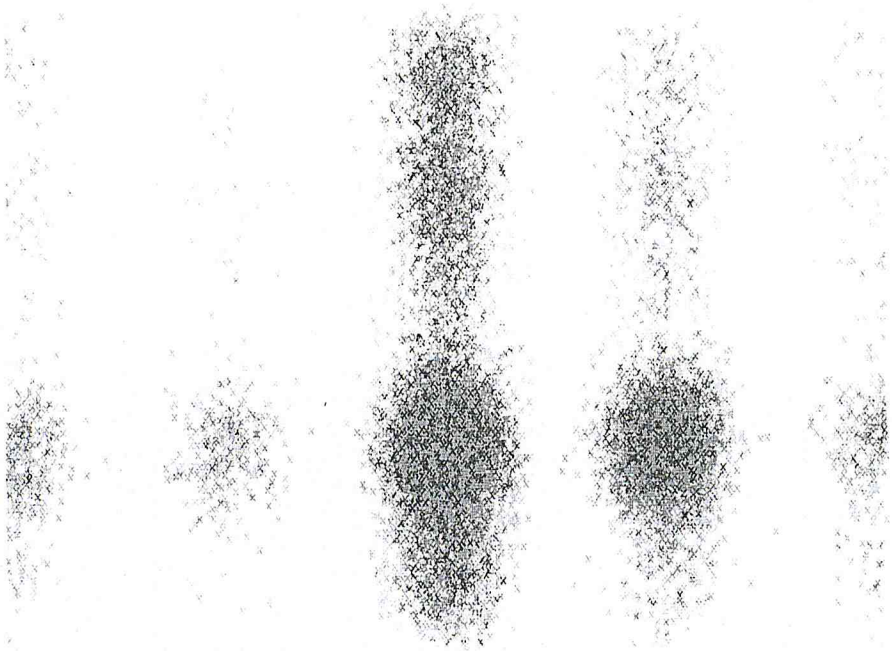


Figure 23 : distribution des transactions par rapport aux wilayas

- **Abscisses** : transaction
- **Ordonnés** : wilaya

Lecture :

Nous pouvons voir ici que les transactions 3 et 4 sont les transactions les plus prisées avec une concentration sur les wilayas 16.35.42. La transaction 3 prend visiblement le dessus.

1.3.2.1.4. Distribution des catégories

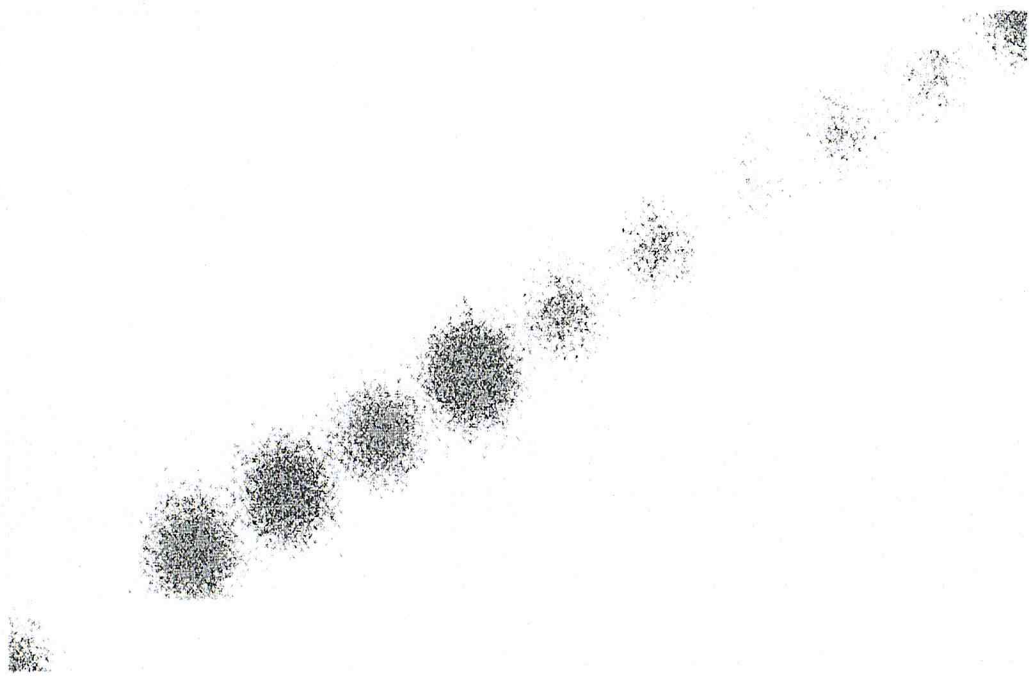


Figure 24 : distribution des catégories

- **Abscisses** : Catégorie.
- **Ordonnés** : Catégorie.

Lecture :

Sous-catégorie, 2, 3, 4,5 les plus prisées avec plus de déchets sur 2,4

1.3.2.1.5. Distribution des transactions sur les catégories

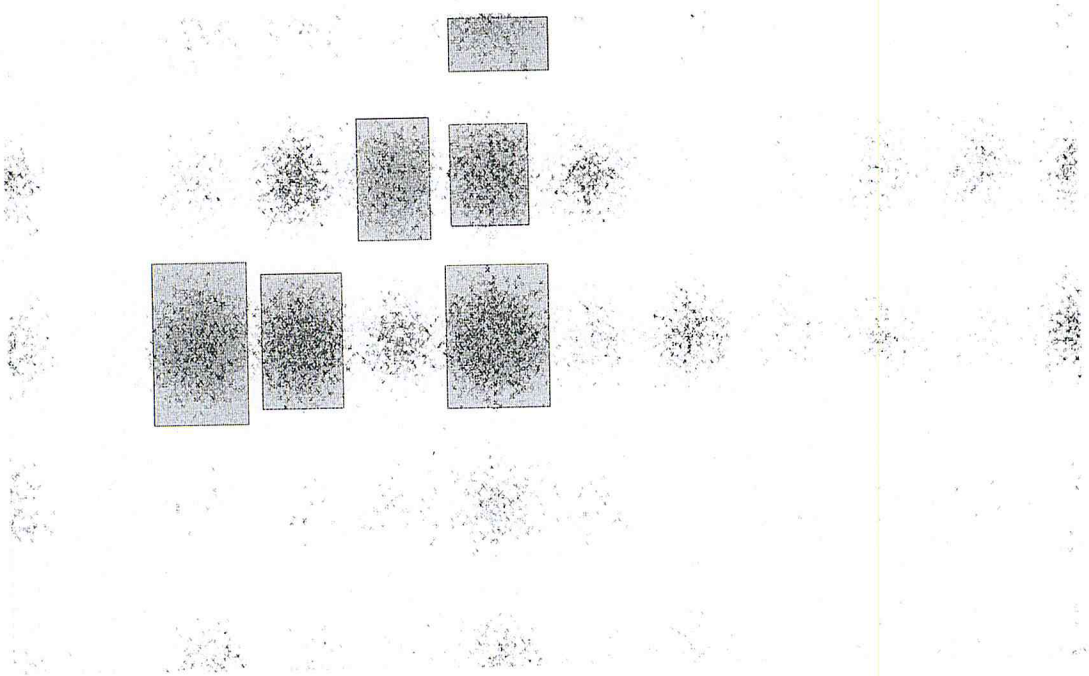


Figure 25 : distribution catégorie par rapport aux transactions

- **Abscisses** : Catégorie.
- **Ordonnés** : Transaction.

Lecture :

Les transactions 3, 4,5 avec plus de 2, 3,5 sous-catégories pour les 3 transactions

1.3.2.1.6. Colonne représentant l'analyse statistique générale sur nos donné après discrétisation

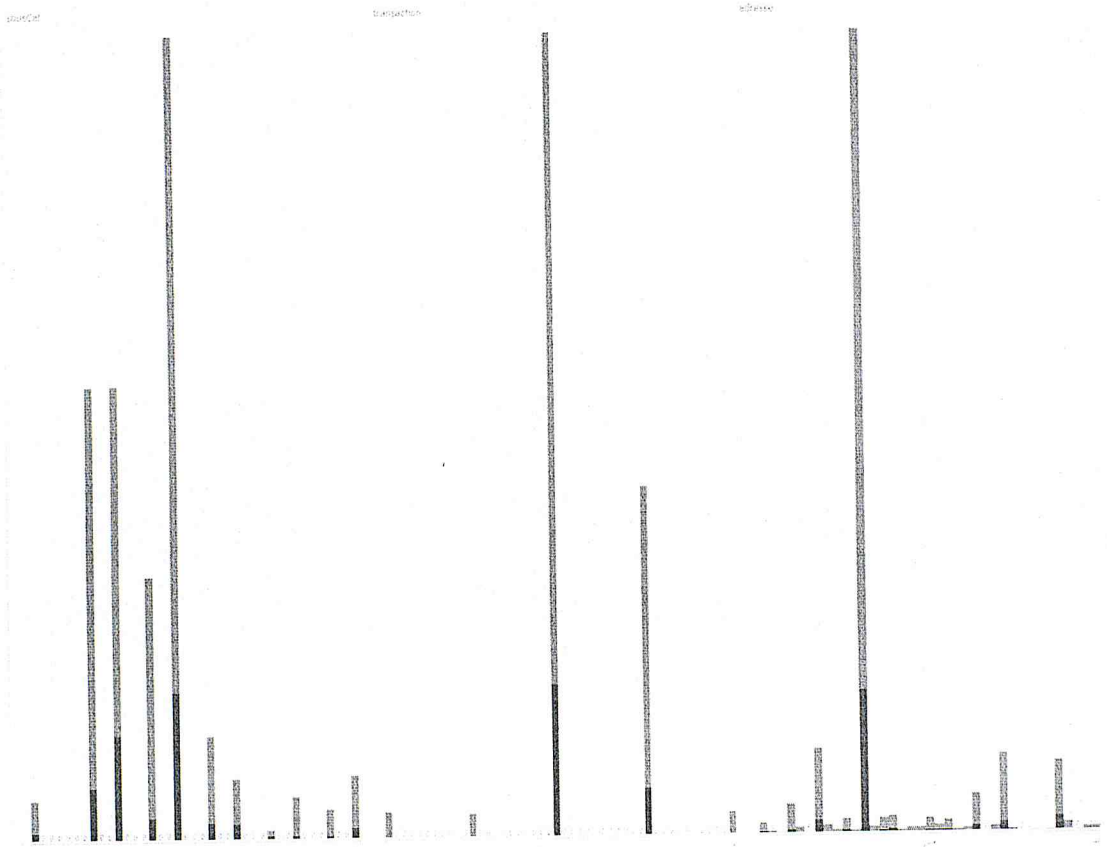


Figure 26 : confirmation

1.3.2.2. interprétations:

13. schéma 1 : les adresses les plus prisées sont les adresses correspondantes aux wilayas suivantes : ALGER, BOUMERDES, TIPAZA.

14. schéma 2 : on peut voir que la wilaya d'ALGER comprend le plus grand nombre d'affluence par rapport aux quatre sous-catégories parmi dix : appartement, villa, terrain, local.

15. schéma 3 : on peut constater que les transactions les plus fréquentes pour la wilaya la plus active (ALGER) sont la location et la vente.

Notre moteur de recherche se référa aux données collectées antécédemment afin d'orienter un nouveau client sur lequel nous n'avons d'informations que celles de sa localité.

1.4. fouille de donnés :

L'exploration de données, connue aussi sous diverse dénominations. Nous trouvons ainsi les expressions suivantes : de fouille de données, forage de données, prospection de données, *datamining*, ou encore extraction de connaissances à partir de données. Ou encore sous des sigles ECD (Extraction de connaissances à partir de données) ou KDD (Knowledge Data Discovring). I a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites web.

L'exploration de données[1] fait suite, dans l'escalade de l'exploitation des données de l'entreprise, à l'informatique décisionnelle. Celle-ci permet de constater un fait, tel que le chiffre d'affaires, et de l'expliquer comme par exemple le chiffre d'affaires décliné par produits, tandis que l'exploration de données permet de classer les faits et de les *prévoir* dans une certaine mesure [Note 2] ou encore de les éclairer en révélant par exemple les variables ou paramètres qui pourraient faire comprendre pourquoi le chiffre d'affaires de tel point de vente est supérieur à celui de tel autre.[28]

1.4.1. Outil et algorithmes :

Weka (acronyme pour Waikato Environnent for Knowledge Analysis, en français : « Environnement Waikato pour l'analyse de connaissances ») est une suite populaire de logiciels d'apprentissage automatique. Écrite en JAVA, développée à l'université de Waikato, Nouvelle-Zélande. Weka est un logiciel libre disponible sous la Licence publique générale GNU (GPL).

L'espace de travail Weka[1] contient une collection d'outils de visualisation et d'algorithmes pour l'analyse des données et la modélisation prédictive, allié à une interface graphique pour un accès facile de ses fonctionnalités.

1.4.1.1. **classification** :

La classification est un datamining (apprentissage automatique) technique utilisée pour prédire l'appartenance au groupe pour les instances de données. Par exemple, vous pouvez utiliser la classification de prédire si la météo sur un jour donné sera "ensoleillée", "des pluies" ou "trouble". Techniques de classification les plus populaires sont des arbres de décision et les réseaux neuronaux.

1.4.1.1.1. **La distance :**

Tout au long de nos algorithmes de datamining, il sera utilisé une approche de distance systématiquement pour chaque algorithme. En mathématiques, une **distance** est une application qui formalise l'idée intuitive de distance, c'est-à-dire la longueur qui sépare deux points. En datamining la distance entre deux attributs veut dire le degré de ressemblance quel que soit le type de ces attributs.

Propriétés de la distance:

d (A, A)=0

d (A,B)= d(B,A)

d (A,B)□ d(A,C) + d(B,C)

Type de distance selon éléments :

1. Distance numérique : d(x,y) = |x-y|
2. Euclidienne : Soit X = (x1,...., xn) et Y = (y1...., yn)

2.1.1. deux exemples, la distance euclidienne

2.1.2. entre X et Y est:

2.1.3. $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

3. Manhattan : La distance de Manhattan, appelée aussi taxi-distance, est la distance entre deux points parcourus par un taxi lorsqu'il se déplace dans une ville américaine où les rues sont agencées selon un réseau ou quadrillage. Un taxi-chemin est le trajet fait par un taxi lorsqu'il se déplace d'un nœud de réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau.

4. $\sum_{i=1}^{n}|x_i - y_i|$

1.4.1.1.2. Classification Supervisé :

La méthode d'apprentissage supervisé utilise cette base d'apprentissage pour déterminer une représentation compacte de *f* notée *g* et appelée *fonction de prédiction,* qui a une nouvelle entrée *x* associe une sortie *g(x)*.
Le but d'un algorithme d'apprentissage supervisé est donc de généraliser pour des entrées inconnues ce qu'il a pu « apprendre » grâce aux données déjà traitées par des experts, ceci de façon « raisonnable ».

1.4.1.1.2.1. Classification de Bayes :

La **classification naïve bayésienne** est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classificateur bayésien naïf, ou classificateur naïf de Bayes, appartenant à la famille des classificateurs Linéaires. Pour faire simple, un véhicule pour être catégorisé en voiture si il a quatre roues et fait moins deux mètres de hauteur et moins 10 mètres de longueur. Ce qui est vrai dans la plupart des cas mais peut tout aussi bien être faut Dans la classification Bayes tout attribut indépendant de l'autre

Ce classificateur s'appuie sur 3 lois statistiques très importantes qui sont :

L'Esperance : plus vulgairement appelé la moyenne

La variance : c'est une mesure de dispersion qui va donner le degré dans lequel les donné sont dispersées entre eux et ou désordonné. [29]

5. Exemple :

Pour le nombre 1, 2 et 3, par exemple, la moyenne est 2 et la variance, 0,667.

$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0{,}667$

6. Exemple Illustratif [30]:

Sexe	Taille (cm)	Poids (kg)	Pointure (cm)
masculin	182	81.6	30
masculin	180	86.2	28
masculin	170	77.1	30
masculin	180	74.8	25
féminin	152	45.4	15
féminin	168	68.0	20
féminin	165	59.0	18

féminin	175	68.0	23

Tableau 9 : base de donné taille sexe homme femme exemple Bayes

Le classificateur créé à partir de ces données d'entraînement, utilisant une hypothèse de distribution Gaussienne pour les lois de probabilités des caractéristiques, est le suivant :

Sexe	**Espérance (taille)**	**Variance (taille)**	**Espérance (poids)**	**Variance (poids)**	**Espérance (pointure)**	**Variance (pointure)**
masculin	178	2.9333e+01	79.92	2.5476e+01	28.25	5.5833e+00
féminin	165	9.2666e+01	60.1	1.1404e+02	19.00	1.1333e+01

Tableau 10 : résultat de Bayes sur base de donné exemple

On suppose pour des raisons pratiques que les classes sont équiprobables, à savoir P(masculin) = P(féminin) = 0.5 (selon le contexte, cette hypothèse peut être inappropriée). Si l'on détermine P(C) d'après la fréquence des échantillons par classe dans l'ensemble de données d'entraînement, on aboutit au même résultat.

Sexe	**Taille (cm)**	**Poids (kg)**	**Pointure (cm)**
inconnu	183	59	20

Tableau 11 : item méconnu a classifié exemple bayes

Nous souhaitons déterminer quelle probabilité *postérieure* est la plus grande, celle que l'échantillon soit de sexe masculin, ou celle qu'il soit de sexe féminin.

Nous avons la formule de bayes qui dit :

$$posterior(male) = \frac{P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)}{evidence}$$

$$P(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Ou δ est la variance obtenu après entrainement.

On aura

7. **p (masculin) = 1.3404e-10**
8. **P (féminin) = 1.5200e-05**

   Donc il est plus probable que l'échantillon **soit une femme**.

1.4.1.1.2.2. Bayes sur nos donné :

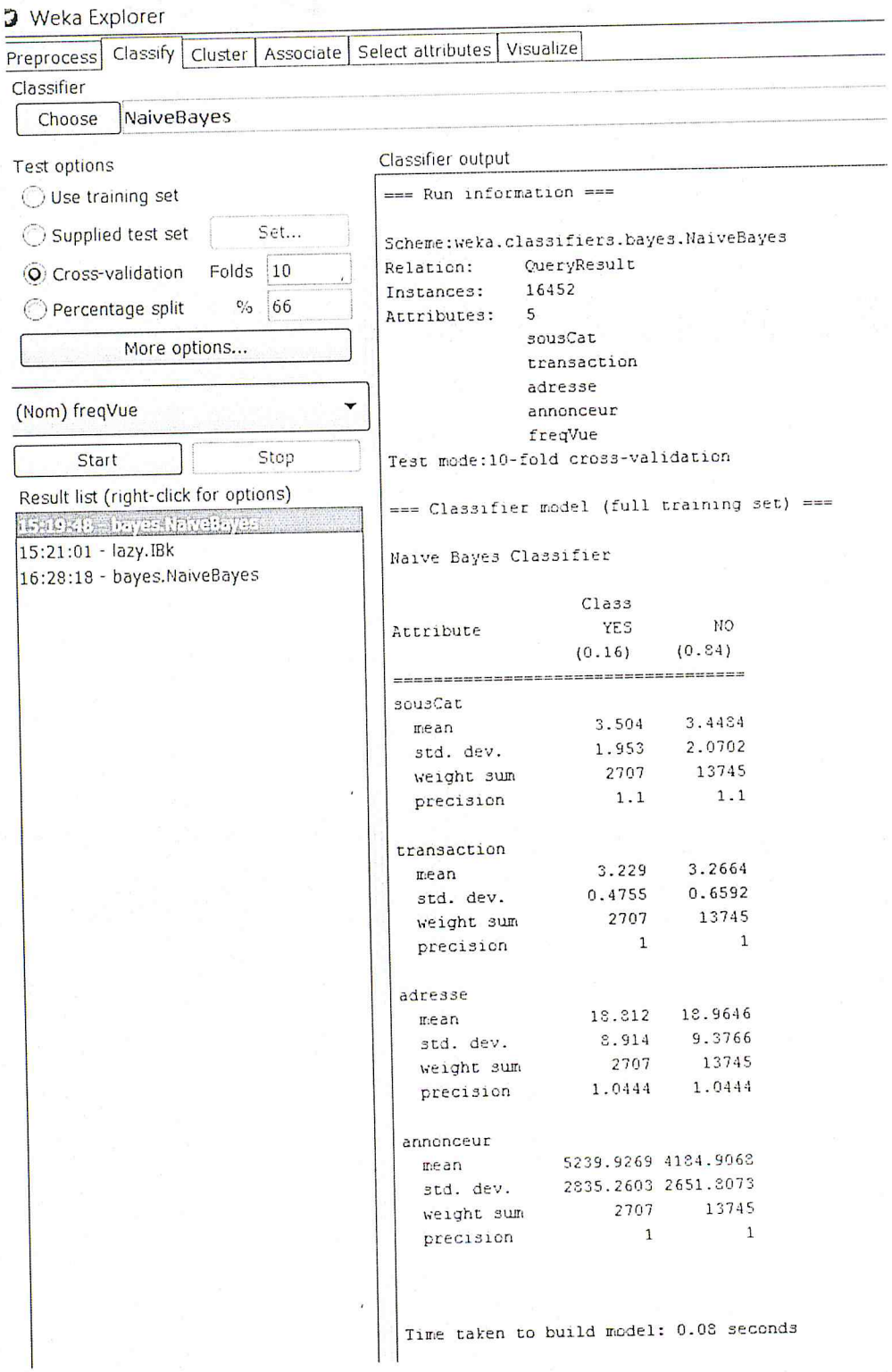


Figure 27 résultat bayes sur nos donné avec Weka

- **Explication des résultats bayésien :**

9. **Instances** : le nombre d'instance de la base de donné classifié qui est égale à 16452

10. « **Test mode 10 fold cross validation** » : nous avons paramétré 10 itérations de cross validation.

11. **Dernier tableau** : va nous donner les résultats de la classification après entrainement du classificateur sur les donné, avec *mean* qui est l'espérance et la variance.

Explication Seconde partie de la classification bayésienne

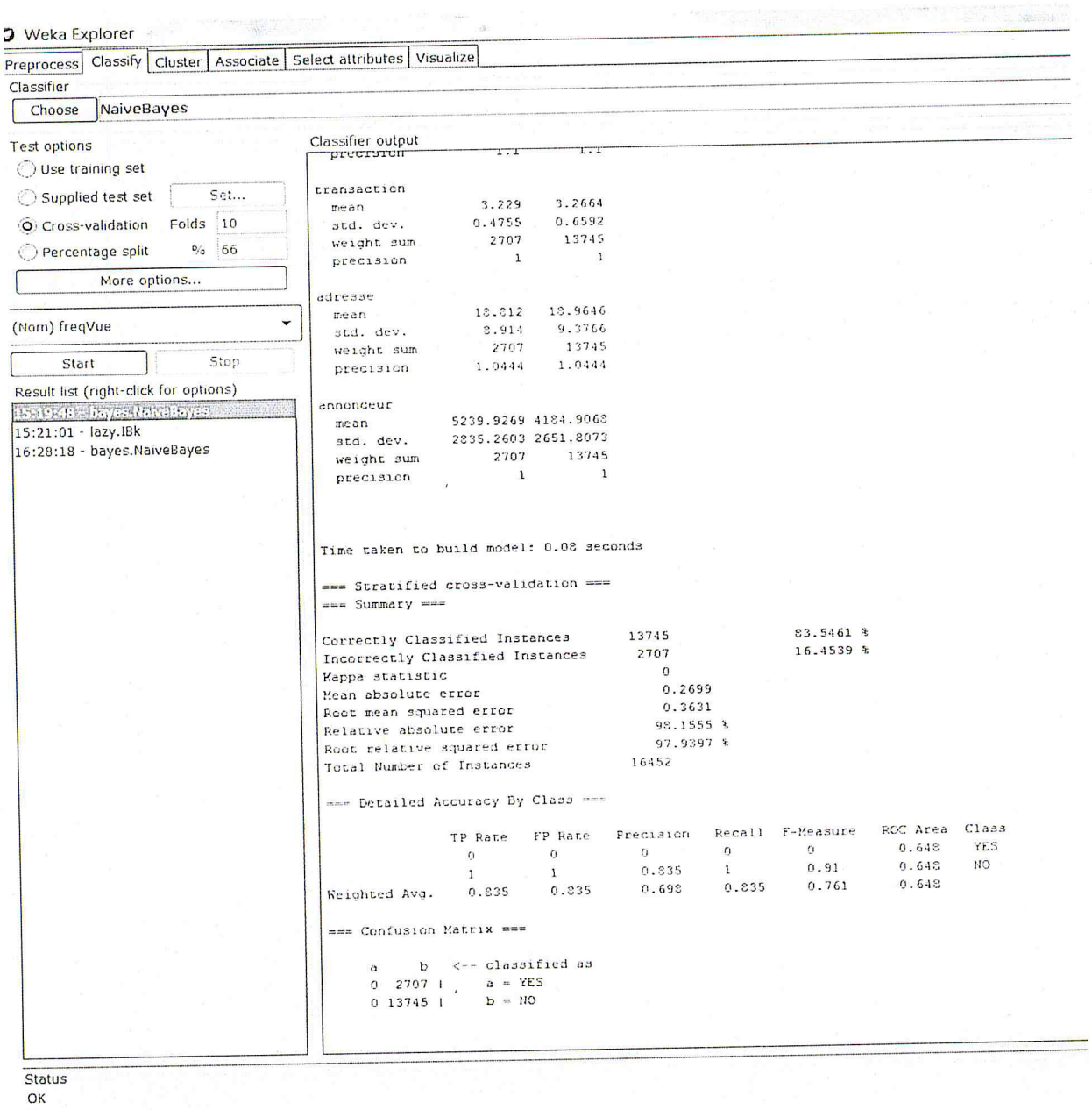


Figure 28 ; évaluation des résultats de bayes sur nos donné

Dans cette figure nous pouvons apprécier les résultats de l'évaluation de cette classification avec les 10 itérations de cross validation effectuées dessus. Nous pouvons voir que les résultats sont bons avec *correctly classified i*nstance 83.5461 % ce qui est un résultat presque parfait pour une classification informatique non assisté .

1.4.1.1.2.3. L'algorithme des *k*-plus proches voisins :

L'algorithme des *k*-plus proches voisins est un des algorithmes de classification les plus simples. Le seul outil dont on a besoin est une distance entre les éléments que l'on veut classifier. Si on représente ces éléments par des vecteurs de coordonnées, il y a en général pas mal de choix possibles pour ces distances, partant de la simple distance usuelle (euclidienne) en allant jusqu'à des mesures plus sophistiquées pour tenir compte si nécessaire de paramètres non numériques comme la couleur, la nationalité, etc.

Comment cela marche-t-il ? On considère que l'on dispose d'une base d'éléments dont on connaît la classe. On parle de base d'apprentissage, bien que cela soit de l'apprentissage simplifié. Dès que l'on reçoit un nouvel élément que l'on souhaite classifier, on calcule sa distance à tous les éléments de la base. Si cette base comporte 100 éléments, alors on calcule 100 distances et on obtient donc 100 nombres réels. Si $k = 25$ par exemple, on cherche alors les 25 plus petits nombres parmi ces 100 nombres. Ces 25 nombres correspondent donc aux 25 éléments de la base qui sont les plus proches de l'élément que l'on souhaite classifier. On décide d'attribuer à l'élément à classifier la classe majoritaire parmi ces 25 éléments. Aussi simple que cela. Bien sûr, on peut faire varier *k* selon ce que l'on veut faire, on peut aussi complexifier la méthode en considérant que les votes des voisins ne sont pas de même poids, etc. Mais l'idée reste la même. [30]

- **Exemple knearest neighboor sur homme et femme :**

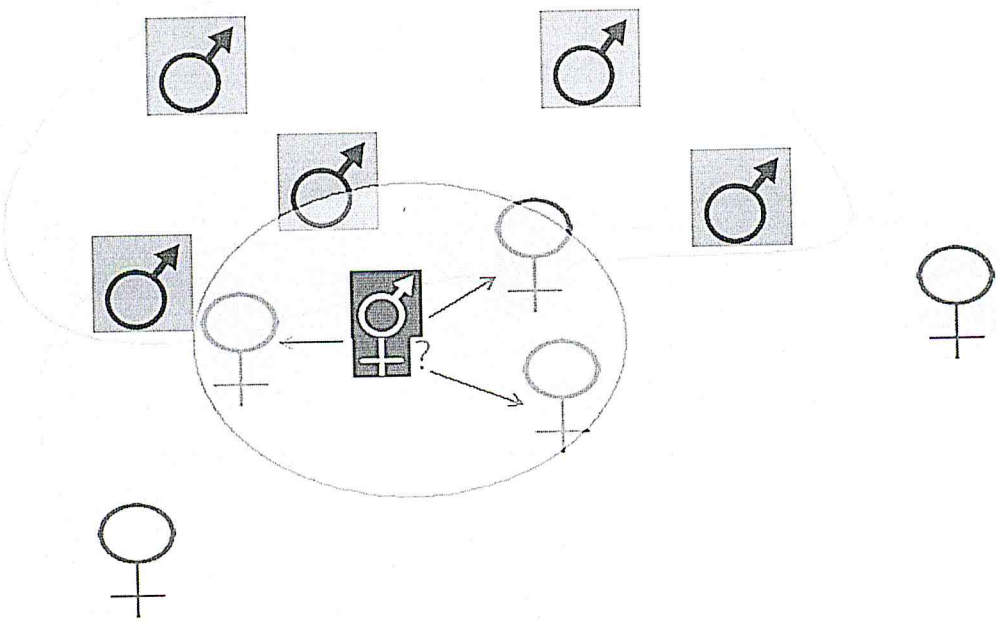


Figure 29 : plus proche voisin sur nos base de on exemple

1.4.1.1.2.4. k-plus proche voisin sur nos donné

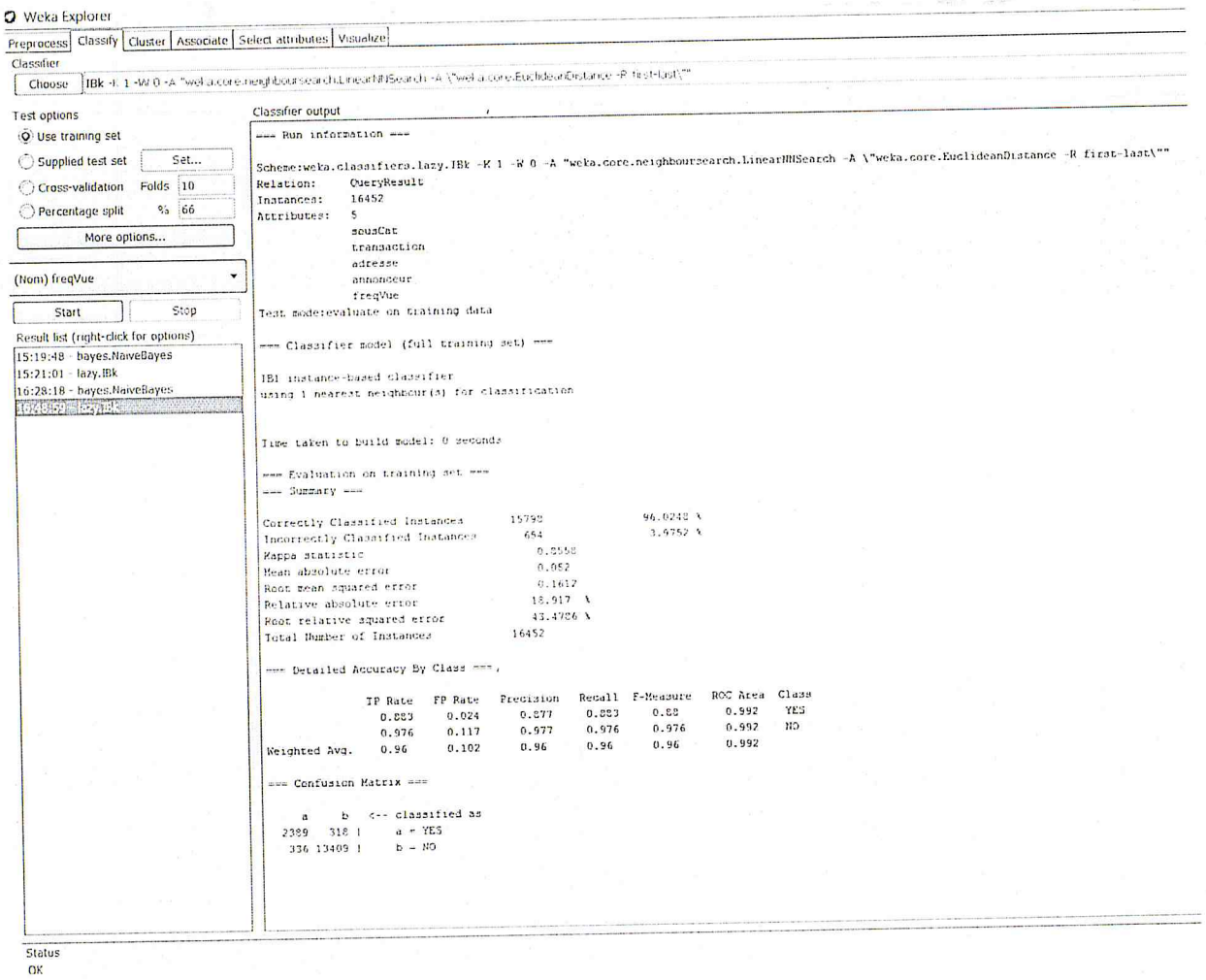


Figure 30 : résultats classification KNN sur notre base de donné avec évaluation

Nous pouvons voir ici que le résultat de test de l'évaluation est plus qu'impeccable avec plus *96%* d'item correctement classifié, ce qui nous pousse à choisir cet algorithme pour la classification sur notre moteur de recherche.

1.4.1.2. Évaluation de la classification

1.4.1.2.1. La transvaluation (cross validation):

Est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. En fait, il y a au moins trois techniques de validation croisée : « *testset validation* » ou « *holdout method* », « *k-fold cross-validation* » et « *leave-one-out cross validation* » (LOOCV).[31]

- La première méthode est très simple, il suffit de diviser l'échantillon de taille n en échantillon d'apprentissage (> 60 % de l'échantillon) et échantillon de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'erreur est estimée en calculant l'erreur quadratique moyenne.
- Dans la seconde, on divise k fois l'échantillon, puis on sélectionne un des k échantillons comme ensemble de validation et les (k-1) autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode l'erreur quadratique moyenne. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les (k-1) échantillons qui n'ont pas encore été utilisé pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.
- La troisième méthode est un cas particulier de la deuxième méthode où k=n, c'est-à-dire que l'on apprend sur (n-1) observations puis on valide le modèle sur la nième observation et l'on répète cette opération n fois[2].

Nous avons opté pour la deuxième manière d'évolution étant la plus utilisé dans les laboratoires de recherche. Et qui donne les résultats les plus probants. [32]

1.4.1.2.2. La matrice de confusion :

Le tableau de répartition (souvent en pourcentages) des individus observés, selon la valeur qu'on leur connaît pour la variable cible et selon la valeur prédite par le modèle. On repère ainsi le taux d'erreur, ou taux de confusion du modèle.

- **Exemple de matrice de confusion :**

Clients...	**... prédits bons**	**... prédits mauvais**
... réellement bons	20,60 %	4,44 %
... réellement mauvais	13,84 %	61,12 %

Tableau 12 : illustration de la matrice de confusion

Le modèle prédit à juste titre que 61,12 % des clients sont mauvais (il en "oublie" 13,84 % qu'il prédit comme étant bons). Et 20,60 % des clients sont réellement bons, et identifiés comme tels par le modèle 4,44 % des clients sont déclarés mauvais alors qu'ils sont bons). Ici, le taux de confusion s'établit à 13,84 + 4,44 = 18,28 % (somme des cases rouges). [33]

CONCLUSION

En guise de conclusion, nous pouvons dire que notre moteur de recherche fonctionne de manière correcte. Son bilan est des plus satisfaisantes car il rapporte des meilleurs résultats, avec un processus ETL qui frise la perfection.

Pour à bien mener notre travail, nous l'avons scindé en deux parties. En première lieu, nous avons effectué une recherche approfondie concernant le fonctionnement des moteurs de recherches de par le monde , nous avons illustré les deux algorithmes les plus utilisé PageRank et Hits dont se sont inspiré tous les autres algorithme , nous avons de même constaté leur référencement sur un site web dynamique implémenté et référencé par nous http://www.casapro-dz.com . Nous nous sommes penchés sur les paradigmes de recherche et leur mise en application via algorithmes parfois gardés secrets. De la syntaxe à la sémantique en passant par la personnalité et même en jumelage donnant des moteurs de recherches hybrides, nous avons passé en revue l'essentiel de la recherche actuelle afin d'élaborer le présent moteur de recherche.

Le volet pratique s'est déroulé après une sélection d'outils présentés précédemment au cours du mémoire. Pour récapitulatif, la méthode de fonctionnement se résume comme suit : L'Extraction s'effectue à partir d'api JAVA (HTTPCLIENT) qui implémenté dans un algorithme ordonnanceur, sillonne le web pour télécharger les pages HTML, après rapatriement des pages web en local, Une série d'algorithmes s'exécute pour nettoyer ces données et les rendre indexables. Cette partie qui transforme le textmining, prend place très importante dans cette phase de transformation avec le nettoyage des mots vides. Cela permet un épurement du texte pour ne garder que les mos susceptibles de faire osciller la classification, après nettoyage. Toujours dans la transformation, une autre phase de vérification utilisant une approche NER (Named Entity Recognition) qui est responsable de déterminer les parties importantes du texte et de les identifier en éliminant le reste. Par la suite, nous aurons quelques informations importantes à mettre en base de données.

Pour finir, l'indexation s'effectue grâce à des algorithmes de classification pour regrouper chaque donnée dans la case appropriée. Les bases de données de mot déjà existantes selon catégorie, vont servir à classifier un texte en calculant la distance avec chaque base données. La distance la plus réduite identifiera la base de donné où il sera chargé est le dernier texte avant d'aboutir à la dernière étape celle Load (chargement).

L'analyse de personnalité aura été des plus enrichissante, après deux phase d'analyse sur les donné. L'une purement statistique, celle qui nous a permis de baliser les différents profils susceptibles d'interroger notre moteur ; l'autre plus intelligente en datamining afin de interpréter en projections futures les résultats.

En utilisant les données récoltées au cours de notre étude datamining, nous pouvons diriger un client encore méconnu de notre moteur selon la wilaya d'où il se connecte. Ce qui est déjà une personnalisation selon la majorité de sa wilaya.

Une solution adaptative est complémentaire à un bon moteur de recherche suivant les démarches personnelles et selon les besoins de l'utilisateur. Une personne étant classifiée comme agent immobilier se verra proposé que des résultats ayant une relation avec l'immobilier, alors que cette personne un jour voudra surement explorer d'autre catégorie. C'est pour cela qu'une approche adaptative selon clic est primordiale pour garder un résultat cohérant. Dans le cas où cette même personne clique sur un sujet autre que l'immobilier, toute les informations utilisées sur cette pérsonne, se verront oublié pour proposer des résultats généralistes selon le profile générale auquel elle appartient. Toutefois, le clic sera enregistré et la classification le prendra en compte pour la prochaine visite.

Les écueils que nous avons dû faire face lors de l'élaboration de ce genre de moteurs sont divers. Des fautes orthographiques récurrentes faussent souvent les résultats des analyses. Nous avons dû faire un recours à un algorithme correcteur afin d'atténuer les erreurs.

Egalement, les sites miroirs qui par leur courte durée de vie surchargent l'index et l'alourdissent, induisant ainsi des erreurs malencontreuses.

Toutefois, ces obstacles ont pu être obviés par divers méthodes et solutions proposées au fil de l'élaboration de notre moteur de recherche qui, au demeurant, offre un ration de succès supérieur à beaucoup d'autres de ses émules.

Pour conclure, nous recommandons comme perspective une approche jumelé d'analyse et de projection en même temps avec les algorithmes récent de datamining en direct (Online Datamining) les moteur s'adapte en temps réel à la personnalité des gens et sans se faire une étude en préalable.

BIBLIOGRAPHIE

Document ET sites Web:

[1] AmitSinghal.ModernInformationRetrieval:ABriefOverview.insinghal.info/ieee2001.pdf consulté en Novembre 2011.

[2] The first searchenginehttp://people.lis.illinois.edu/~chip/projects/timeline/1990archie.htm. Consulté en Février 2012.

[3] Guide complet des moteurs de recherches Yahoo : http://www.lesmoteursderecherche.com/yahoo.htm consulté en Mars 2012

[4] le journal du net : Lycos : http://www.journaldunet.com/offline/lycosoffline.shtml consulté en Mars 2012.

[5] Description Altavista : Wikipédia France : http://fr.wikipedia.org/wiki/AltaVista . Consulté en Septembre 2011.

[7] L'algorithme du PageRank expliqué. ingénieur informatique DAN sur http://www.webmaster-hub.com/publication/L-algorithme-du-PageRank-explique.html consulté en Octobre 2011.

[8] Pourquoi le site est de 50 backlinks grimpe mieux que votre site avec 1000 backlinks: http://www.marketingfan.at/backlink-qualitaet#ixzz1ycYJXiyF n traduit de l'allemand par Google translate consulté en Mai 2012.

[9] Explication du PageRank : http://www.infowebmaster.fr/18,news-explication-pagerank.html consulté en Mars 2012 consulté en Janvier 2012.

[11] The PageRank Citation Ranking: Bringing Order to the Web Janvier 29, 1998 ilpubs.stanford.edu:8090/422/1/1999-66.pdf. Consulté en Mai2012.

[12] PageRank comment ça marche ? http://professeurs.esiea.fr/wassner/?2007/06/03/74-l-algorithme-pagerank-comment-a-marche. Consulté en Mars 2012.

[13] Report: Google Uses About 900,000 Servers : http://www.datacenterknowledge.com/archives/2011/08/01/report-google-uses-about-900000-servers/. Consulté en Janvier 2012.

[14] Typologie et mode de fonctionnement des outils de recherche d'information sur internet en Biologie/Médecine sur :http://hal.archives-ouvertes.fr/docs/00/59/56/39/PDF/boudry_typologie_MS.pdf consulté en Mars 2012.

[15] Annonce de bing sur son blog de son intéressement a l'analyse de personnalité sur http://www.bing.com/community/site_blogs/b/search/archive/2012/05/10/spend-less-time-searching-more-time-doing-introducing-the-new-bing.aspx consulté en Juin 2012.

[16] Introducing the Knowledge Graph: things, not strings sur : http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html. consulté en Juin 2012.

[17] Moteur de recherche http://fr.wikipedia.org/wiki/Moteur_de_recherche#.C3.89volution_vers_le_web_s.C3.A9mantique [18]. Consulté en Février 2012

[19] Recherched'informationsurinternet/Annuairesinternet.(2008,avril14).Workbooks,Retrieved10:54,juin22,2012http://www.passion-psycho.fr/psychometrie/la-personnalite/142-definition-et-evaluation-de-la-personnalite Consulté en Juin 2012

[20] http://www.passion-psycho.fr/psychometrie/la-personnalite/142-definition-et-evaluation-de-la-personnalite Consulté en Mars 2012.

[21] http://www.cs.uic.edu/~liub/FBS/TOC-MC_Liu.pdf Consulté en Février 2012.

[22] http://www.fao.org/DOCREP/004/X2465F/x2465f08.htm Consulté en Juin 2012.

[23] http://www.statcan.gc.ca/edu/power-pouvoir/ch2/types/5214777-fra.htm Consulté en Mai 2012.

[24] http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html Consulté en Juin 2012.

[25]DécouvertedeMySQL,PostgreSQLetOracle.(2012,mars1).Wikibooks,Retrieved13:58,juin22,2012depuisdisponiblesurhttp://fr.wikibooks.org/w/index.php?title=D%C3%A9couverte_de_MySQL,_PostgreSQL_et_Oracle&oldid=357314 Consulté en Juin 2012.

[26] http://en.wikipedia.org/wiki/Message_Understanding_Conference. Consulté en Mars 2012.

[27] http://www.thearling.com/text/dmwhite/dmwhite.htm . Consulté en Février 2012.

[28] http://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9esv. Consulté en Janier 2012.

[29] http://www.statcan.gc.ca/edu/power-pouvoir/toc-tdm/5214718-fra.htm. Consulté en Avril 2012.

[30] http://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne Consulté en Avril 2012.

[30] http://interstices.info/encart.jsp?id=c_41867&encart=3&size=600,500 . Consulté en Mail 2012.

[31] http://en.wikipedia.org/wiki/Cross-validation_(statistics) Consulté en Juin 2012.

[32] http://web.archive.org/web/20060623055814/http://decisiontrees.net/node/36 Consulté en Juin 2012.

[33] http://en.wikipedia.org/wiki/Confusion_matrix Consulté en Juin 2012.

Livre :

[6] Tara Calishainet Rael Dornfest, **Google à 100%, 200 trucs, secrets et techniques**, Édition O'Reilly, 2011Mathématique a l'usage des informaticiens Thierry brugére alain mollard, ellipse.

http://fr.wikibooks.org/w/index.php?title=Recherche_d%27information_sur_internet/Annuaires_internet&oldid=166789.

Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)

Api :

Apache httpclient.

Weka.

Jdom.

Jgrapht.

/