N° D'ORDRE : . . . . . .

République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

Université Djillali Liabes - Sidi Bel Abbes
Faculte Des Sciences Exactes
Departement D'Informatique

THÈSE
DE DOCTORAT EN SCIENCE

Présentée par : Ahmed Chaouki LOKBANI

Option : Informatique

Intitulée

Le problème de sécurité par le Data Mining

Soutenue, le .../.../2017
devant le jury composé de :

Président	**M. FARAOUN Kamel Mohamed**	Professeur à l'UDL Sidi Bel Abbes
Examinateurs	**M. AMINE Abdelmalek**	Professeur à l'UTM Saida
	M. ATHMANI Baghdad	Professeur à l'U Oran 1
	M. ADJOUJ Reda	Maître de conférences "A" à UDL Sidi Bel Abbes
	M. HAMOU Reda Mohamed	Maître de conférences "A" à l'UTM Saida
Directeur de Thèse	**M. LEHIRECHE Ahmed**	Professeur à l'UDL Sidi Bel Abbes

Année universitaire : 2016/2017

Résumé

De nos jours et dans le monde dans lequel nous vivons l'ordinateur est omniprésent de plus en plus. Il contient un tas d'informations qui peuvent être banales comme confidentielles.

Initialement isolés les uns des autres, ces ordinateurs sont à présent interconnectés et le nombre de points d'accès ne cessent de croître. Ce phénomène a été catalysé par l'essor de l'Internet qui attire de plus en plus d'internautes par les nombreux avantages et la diversité des services rendus accessibles. Mais cet accroissement du nombre d'utilisateurs peut causer des dégâts sur le système informatiques, par l'exploiter les vulnérabilités des réseaux et des systèmes pour essayer d'accéder à des informations sensibles dans le but de les lire,les modifier ou les détruire, portant atteinte au bon fonctionnement du système.

L'importance de la sécurité des systèmes informatiques motive divers angles de la recherche dont l'objective est de fournir de nouvelles solutions prometteuses qui ne pourraient être assurées par des méthodes classiques. Les systèmes de détection d'intrusions sont l'une de ces solutions qui permettent la détection des utilisations non autorisées et des anomalies, les mauvaises utilisations et les abus dans un système informatique par les utilisateurs externes ainsi que les utilisateurs internes.

Depuis l'apparution des premiers modèles de détection d'intrusion par Denning [30] plusieurs systèmes de détection d'intrusion plus performant et précis ont vu le jours et qui sont basés sur les connaissances des experts de sécurité ou les méthodes statistiques et les approches de l'intelligence artificielle qui ont montré beaucoup de limites, ce qui a pousser les chercheur à s'orienté vers d'autres techniques et en particulier, les techniques du data mining, sur les quelles des modèles de détection d'intrusion plus précis et plus rapides ont été développés.

Le défi dans le domaine de la sécurité informatique et plus précisément dans les systèmes de détection d'intrusions est de pouvoir déterminer la différence entre un fonctionnement normal et un fonctionnement anormal. Cependant, les systèmes et les réseaux à protéger sont devenus de plus en plus complexes comme pour la nature des intrusions courantes et futures ce qui nous incite à développer des outils de défense automatiques et surtout adaptatifs. Une solution prometteuse est d'utiliser les systèmes bio inspirés appelés : les systèmes bio- informatique.

Mots clefs : Sécurité Informatique, système de détection d'intrusion, Data Mining, approche par scénario, la protection des abeilles sociales, Kdd Cup'99.

Abstract

Today and in the world we live in is pervasive computer more. It contains a lot of information that can be mundane as confidential.

Initially isolated from each other, these computers are now interconnected and the number of access points continues to grow.This phenomenon has been catalyzed the growth of the Internet attracts more and more users from the many advantages and diversity of the services available. These increase in the number of users who are not necessarily full of good intentions vis-à-vis these computer systems. They can exploit vulnerabilities in networks and systems to try to gain access to sensitive information in order to read them, modify them or destroy them, impairing the functioning of the system.

The importance of the security of computer systems motivates different angles of research whose objective is to provide promising new solutions that could not be met by conventional methods. Intrusion detection systems are one of these solutions to detect unauthorized use and anomalies, misuse and abuse in a computer system by external users and internal users.

Since apparution first intrusion detection models by Denning [30] more efficient and more accurate intrusion detection systems have seen the days that are knowledge-based security expert or statistical methods and approaches of artificial intelligence that showed any limits, which push the researcher is facing other techniques and in particular, data mining techniques on which intrusion detection models more precise and faster have been developed.

The challenge in the field of computer security and specifically in the intrusion detection system is able to determine the difference between a normal and abnormal operation. However, systems and networks to protect have become increasingly complex as to the nature of current and future intrusions prompting us to develop automatic defense tools and particularly adaptive. A promising solution is to use the systems inspired by biology called : the bioinformatics systems.

Keywords : Computer Security, Intrusion detection system, Data mining, scenario approach, order Recognition, social bees protection, Kddcup'99.

ملخص

اليوم وفي العالم الذي نعيش فيه أصبح الكمبيوتر أكثر انتشارا و يحتوي على الكثير من المعلومات التي يمكن أن تكون عادية أو سرية. لهاذ أنظمة الكمبيوتر تحتل مكانة مهمة في جميع القطاعات المهنية بالإضافة إلى العمل الخاص اليومي.

في البداية كل هذه الشبكات كانت معزولة عن بعضها البعض و الآن مترابطة و مما حفزت هذه الظاهرة نمو الإنترنت الذي يجدب العديد من المستخدمين لما يقدمه من العديد من المزايا وتنوع الخدمات المتاحة.

هذه الزيادة في عدد المستخدمين الذين ليست لهم بالضرورة نوايا حسنة بالنسبة لأنظمة الكمبيوتر و يمكن أن تستغل الثغرات الموجودة في الشبكات والأنظمة للمحاولة للوصول إلى المعلومات الحساسة من أجل قراءتها، أو تعديلها، أو تدميرها، مما قد يعرقل السير الحسن للنظام. وهذا يثير حتما تساؤلات حول سلامة أنظمة الكمبيوتر هذه وأمن المعلومات الموكلة إليهم بأكثر فعالية ضد التهديدات والهجمات.

إن أهمية أمن أنظمة الكمبيوتر تحفز زوايا مختلفة من البحوث التي تهدف لتقديم حلول جديدة واعدة التي لا يمكن أن تتحقق من خلال الطرق التقليدية . إن أنظمة كشف التسلل هي واحدة من هذه الحلول للكشف عن استخدام والشذوذ غير المصرح به وسوء الاستخدام وسوء المعاملة في أنظمة الكمبيوتر بالنسبة للمستخدمين الخارجيين والمستخدمين الداخليين.

منذ ظهور أول نماذج الكشف التسلل (دينينغ ١٩٨٧) عدة أنضمة كشف التسلل أكثر كفاءة و دقة صهرت للعيان قائمة علي معرفة خبراء الأمن المعلوماتية أو الأساليب الإحصائية و النهج الذكاء الاصطناعي التي أظهرت العديد من النقائص مما دفع الباحث للتوجه إلي تقنيات أخرى، وعلى وجه الخصوص تقنيات التنقيب عن البيانات للكشف عن التسلل بأكثر وأسرع دقة تم تطويرها.

يتمثل التحدي في مجال أمن تكنولوجيا المعلومات، وتحديدا في نظام كشف التسلل ، في إمكانية تحديد بين عملية طبيعية وغير طبيعية.

لهاذا الأمر أصبحت حماية أنظمة وشبكات تزداد تعقيدا و ذالك لطبيعة التدخلات الحالية والمستقبلية مما دفعنا إلى تطوير أدوات الدفاع الآلية وقابلة للتكيف بشكل خاص . حلا واعدا هو استخدام أنظمة مستوحاة من البيولوجيا يسمى النظم المعلوماتية الحيوية

الكلمات المفتاحية الأمن المعلوماتية، نضام كشف الإختراقات، تنقيب البيانات، نهج سيناريو، حماية النحل الإجتماعية، ك دي دي ٩٩

Dédicace

***A** ma très chère défunte grand-mère zaza,*

***A** ma très chère défunte mère,*

***A** mon défunt père,*

***A** ma chère épouse,*

***A** mes deux petites prunelles*
et ma raison de vivre :Rouaia et Nouha,

***A** ma chère et unique sœur Zhor,*
A mon oncle Halim , mes tantes, cousins et cousines.

***J**e vous aime!*

Remerciement adressés aux membres du jury

Avant tout, le grand et le vrai merci revient à Allah qui m'a donné la force, la foi et la vie pour accomplir cette tâche, qui au début paraissait une mission difficile.

Je tiens à remercier mon directeur de thèse Mr LEHIRECHE Ahmed qui a dirigé ce travail de recherche de doctorat, pour sa sagesse, son expérience, ses conseils et encouragements, sa patience, pour m'avoir fait confiance et laissé la liberté nécessaire à l'accomplissement de mes travaux. Merci Cheikh.

Je suis très honorée par la présence de Mr FERAOUN Mohamed Kamel, qui a accepté de présider le jury de ma thèse, je suis également très honorée par la présence de Mr ATHMANI Baghdad, Mr AMINE Aabdelmalek, Mr HAMOU Reda Mohamed et Mr ADJOUJ Réda qui ont accepté d'être les rapporteurs de cette thèse. Qu'ils trouvent ici mes plus vifs remerciements pour l'effort qu'ils ont fait pour lire mon manuscrit et l'intérêt qu'ils ont porté à mon travail.

Remerciement

A la mémoire de ma défunte mère "Cherifa", sans qui je ne serais absolument pas où j'en suis aujourd'hui. Je remercie Dieu de m'avoir donné courage et patience pour pouvoir achever ce modeste travail qui était son souhait le plus cher.

Enfin une grande reconnaissance de tout mon cœur à ma petite famille surtout à ma chère épouse FATIMA, mes deux prunelles RAOUAIA et NOUHA et mon unique sœur ZHOR qui m'ont toujours soutenu durant ces dernières années en comblant le vide qu'a laissé ma défunte mère en moi.

A ma grade famille, ma défunte grand-mère maternelle ZAZA qui m'était très chère, mon oncle HALIM qui m'a élevé et tout apprit dans cette vie, mes tantes qui étaient toujours présentes pour moi.

Je tiens aussi, à exprimer ici tous mes remerciements ainsi que toute ma gratitude à toutes les personnes sans lesquelles ce travail n'aurait pas été mené à terme, et surtout pour leurs conseils et remarques constructives qui m'ont permis d'améliorer grandement la qualité de mes travaux et du présent mémoire :

— *A leur tête en particulier mon ami et frère Reda pour tous ses efforts consentis et les nombreuses discussions fructueuses qui ont animé ces années de thèse, ses conseils, sa disponibilité et surtout son appui moral qu'il a consenti de près ou de loin quant à l'aboutissement de ce travail surtout durant les moments difficiles que j'ai passé.*

— *A nos futures docteurs Boudia et Rahmani qui me boostaient aux moments ou ça n'allait pas vraiment et surtout durant ces deux derniers mois lors de la finalisation de la rédaction de la thèse, un grand merci à eux.*

A Mes amis Amine, Belahcene, Bensadek, Boumedienne, Bouhmidi, Chikouche, Kies, Mabrouk, Medjadji, Ould kada, Rahmani, et à tous ceux que j'ai oublié de mentionner, pour leurs soutiens moral.

A mes collègues de travail : Khobzaoui, Benyahia, Bendaoud, Yahlali, Rabotte, Meziane, Bahloul Et Raimes et autres. . . .

Table des matières

Liste des figures

Liste des tableaux

Liste des Algorithmes

Table d'abréviations

BDD : **B**ase **D**e **D**onnées.
CPU : **C**entral **P**rocessing **U**nit.
CSI :**C**omputer **S**ecurity **I**nstitue.
DDoS : **D**istributed **D**enial **o**f **S**ervice attack.
DNS : **D**omain **N**ame **S**ystem
ECD : **E**xtraction des **C**onnaissances à partir des **D**onnées.
FN : **F**aux **N**égative.
FP : **F**aux **P**ositive.
HIDS : **H**ote **I**ntrusion **D**etection **S**ystem.
IA : **I**ntelligence **A**rtificielle.
ICMP : L'**H**yper**T**ext **M**arkup **L**anguage.
IR : **I**nformation **R**etrieval.
IDS : **I**ntrusion **D**etection **S**ystem.
IDWG : **I**ntrusion **D**etection **W**orking **G**roup.
IEC : **I**nternational **E**lectrotechnical **C**ommission
IEEE : **I**nstitute of **E**lectrical and **E**lectronics **E**ngineers.
IETF : **I**nternet **E**ngineering **T**ask **F**orce.
IP : **I**nternet **P**rotocole
IPS : **I**ntrusion **P**revesion **S**ystem.
IPv4 : **I**nternet **P**rotocole **v**ersion 4
IPv6 : **I**nternet **P**rotocole **v**ersion 6
ISO : **I**nternational **S**tandardization **O**rganization.
KDD : **K**nowledge **D**iscovery in **D**atabases.
LDAP : **L**ightweight **D**irectory **A**ccess **P**rotocol
NIDS : **N**etwork **I**ntrusion **D**etection **S**ystem.
RAM : **R**andom **A**ccess **M**emory
SIA : **S**ystème **I**mmunitaire **A**rtificiel.
SNMP : **S**ample **N**etwork **M**anagement **P**rotocol.
SSH : **S**ecure **S**hell.
SSL : **S**ecure **S**ockets **L**ayer.
TCB : **T**rusted **C**omputing **B**ase.
TCP : **T**ransport **C**ontrol **P**rotocol.
Tf-IDF : **T**erm **F**requency-**I**nverse **D**ocument **F**requency.
TN : **T**rue **N**egative.
TP : **T**rue **P**ositive.
UDP : **U**ser **D**atagram **P**rotocol.
USA : États-Unis.
VN : **V**rai **N**égative.
VLBD : **V**ery **L**arge **D**ata **B**ase.
VP : **V**rai **P**ositive.

Introduction générale

Centre d'intérêt

De nos jours et dans le monde dans lequel nous vivons l'ordinateur est omniprésent de plus en plus. Il contient un tas d'informations qui peuvent être banales comme confidentielles comme par exemple numéros de carte de crédit, les listes des contactes ...etc. ce qui permet aux systèmes informatiques d'occuper une place prédominante dans tous les secteurs professionnels : les entreprises, les administrations, les banques, les assurances, la médecine ou encore le domaine militaire, et plus précisément dans le quotidien des particuliers.

Initialement isolés les uns des autres ces équipements informatiques sont à présent interconnectés et le nombre de points d'accès ne cessent de croître, et ce phénomène a été catalysé par l'essor de l'internet qui attire de plus en plus d'internautes par les nombreux avantages et la diversité des services rendus accessibles. Ils peuvent ainsi bénéficier de communication rapides à moindre coût, partager des ressources de traitement et de stockage de grandes capacité (Cloud Computing), faciliter les échanges commerciaux et financiers (e-Commerce, e-Banking) et, plus généralement, partager et accéder à l'information [33].

Cet accroissement du nombre d'utilisateurs qui ne sont pas forcément pleins de bonnes intentions vis-à vis de ces systèmes informatiques. Ils peuvent exploiter les vulnérabilités des réseaux et des systèmes pour essayer d'accéder à des informations sensibles dans le but de les lire, les modifier ou les détruire, portant atteinte donc au bon fonctionnement du système. Les risques, causés par les attaques sur les systèmes informatiques, deviennent un réel problème pour les entreprises et les organisations.

Notre dépendance croissante aux systèmes informatiques dans notre quotidien soulève inévitablement la question quant à leur sécurité et à la sécurité de l'information qui leurs sont confiées le plus efficacement possible contre les menaces et attaques.

En dépit des efforts consentis dans le domaine de la sécurité depuis un certain nombre d'années pour essayer d'endiguer les problèmes de la sécurité, force est de constater que le nombre de vulnérabilités dans les systèmes informatique et les activités malveillantes se sont multipliées. Les données statistiques fournies par Kaspersky Lab [82], spécialisé dans la sécurité des systèmes d'information tendent à confirmer ce phénomène. Notamment, il ne dénombrait pas moins de 946393693 attaques réussies au travers de l'Internet en 2011 contre seulement 580371937 en 2010 représentant ainsi une progression de 39% en une seule année. Les attaques les plus récentes profitent des failles de sécurité des services ou des systèmes informatiques qui sont plus vulnérables.

Pour cerner le problème de la sécurité informatique, divers moyens ont été mis en place pour prévenir quelconques attaques comme les pare-feux, authentification, les Proxy...etc. Malheureusement, ces moyens ne suffisent pas à bloquer tous les types d'attaques qui nuisent à la confidentialité, l'intégrité ou la disponibilité. Pour pallier à ce problème, un nouveau moyen de sécurité appelé systèmes de Détection d'intrusions (SDI ou IDS en anglais) ont fait leur apparition. Ce nouveau concept introduit par James Anderson en 1980 à pour objectif l'analyse du trafic des requêtes et la détection des comportements malveillants [5].

Contexte et Problématique

Cette thèse s'inscrit dans les domaines de la sécurité informatique, système de détection d'intrusion, le Datamining et les méthodes bio-inspirées tel que les essaims d'abeilles.

L'importance de la sécurité des systèmes informatiques motive divers angles de la recherche dont l'objective est de fournir de nouvelles solutions prometteuses qui ne pourraient être assurées par des méthodes classiques. Les systèmes de détection d'intrusions sont l'une de ces solutions qui permettent la détection des utilisations non autorisées et des anomalies, les mauvaises utilisations et les abus dans un système informatique par les utilisateurs externes ainsi que les utilisateurs internes [30].

Depuis l'apparution des premiers modèles de détection d'intrusion par Denning, plusieurs systèmes de détection d'intrusion plus appropriés et plus précis ont vu le jour. La première génération d'IDS qui été basée sur les connaissances des experts de sécurité, ou les méthodes statistiques et les approches de l'intelligence artificielle ont été utilisées, pour créer les noyaux des modèles de détection d'intrusion. Mais ces techniques de l'IA ont montré beaucoup de limites à cause des différents problèmes tel que le grand volume de trafic réseau, la distribution déséquilibrée des données, la prise de décision difficile entre un comportement normal et anormal en plus de l'exigence de l'adaptation permanente pour les environnement en constante évolution, ce qui a pousser les chercheur à s'orienté vers d'autres techniques et en particulier les techniques du data mining, sur les quelles des modèles de détection d'intrusion plus clair et plus expéditifs ont été développés c'est ce qu'on appelle la deuxième génération des systèmes de détection d'intrusion. Cependant, les systèmes et les réseaux à protéger sont devenus de plus en plus complexes ainsi que la nature des intrusions courantes et futures.

Cette thèse a comme objectif de mettre en évidence de nouvelles méthodes de sécurité informatiques, ce qui nous incite à développer des outils de défense automatiques et surtout adaptée à la nature. Une solution prometteuse est d'utiliser les systèmes inspirés de la biologie appelés : les systèmes bios-informatique et spécialement des algorithmes calqués des essaims d'abeilles (Bee swarm) [74].

Organisation de la thèse

Les chapitres qui composent cette thèse sont organisés en deux grandes parties :

La première partie est un état de l'art présentant respectivement le domaine dans lequel la problématique de la thèse est posée et les outils que nous avons utilisé comme support dans notre contribution. La deuxième partie met en relief notre contribution.

1. **La sécurité informatique** : Nous présentons dans ce chapitre les notions générales de la sécurité informatique en débutant par donner un historique, des définitions de da sécurité informatique et les points essentiels que touche la notion de sécurité tel que : les réseaux, les systèmes et applications.
2. **Les systèmes de détection d'intrusion** : nous présentons dans ce chapitre ce que la littérature nous offre en essayant de ce focaliser sur la notion d'intrusion, la détection d'intrusion et enfin les SDI en passant par le comportement des ces derniers c'est-à-dire passifs ou actifs. En fin de ce chapitre nous nous somme penchés sur les différentes techniques existantes et bien entendu celles du Data mining que nous utilisons dans notre contribution.
3. **La recherche d'information et Data mining** : Nous avons présenté dans ce chapitre les concepts de base de la recherche d'information. Nous commençons par donner une définition de la RI et nous décrivons les différents modèles servant de cadre théorique. Nous illustrons également le processus de RI en présentant les étapes d'indexation, fonction de correspondance, ainsi que les processus d'évaluation. A la fin du chapitre nous la notion du Data mining et le processus d'extraction des connaissances.
4. **Expérimentation et résultats** : Dans ce chapitre nous avons donné en détail la vie d'une abeille et son comportement du point de vue sécurité ainsi que la stratégie adoptée par celle ci. Ensuite nous avons modélisé un système de détection d'intrusion inspiré du système de protection des abeilles sociales ayant une défense active dont elles utilisent pour défendre la ruche. Nous avons aussi présenté le corpus KDD, qui est utilisé pour l'évaluation des anomalies et la détection d'intrusion. Enfin de ce chapitre nous avons argumenté les résultats fournis par notre algorithme.
5. **Conclusion et perspective** : Nous avons présenté dans cette partie les avantages important de notre approche et ce que l'on prévoie de faire à l'avenir avec le même système (IDS Bees) mais

avec d'autres algorithmes du data mining pour pouvoir déduire la ou la robustesse du système au maximale.

Chapitre 1

La sécurité informatique

Sommaire

Introduction

Dans le domaine de l'informatique, le mot "sécurité" peut couvrir plusieurs acceptions. La première correspond à la sécurité-innocuité (en anglais safety) et concerne la prévention de catastrophes : dans ce sens, un système informatique aura une sécurité satisfaisante si aucune de ses défaillances éventuelles ne peut provoquer de dégâts importants, ou si celles de ses défaillances qui peuvent provoquer des dégâts importants sont suffisamment peu probables. Ce type de sécurité est bien évidemment une exigence majeure lorsque le bon fonctionnement du système informatique est nécessaire pour la sauvegarde de vies humaines ou de l'environnement, ou encore d'intérêts financiers importants. C'est en particulier le cas des systèmes tels que les systèmes de transport ou de contrôle des centrales nucléaires [32].

Une seconde signification du terme de sécurité correspond au mot anglais "security" et concerne la capacité du système informatique à résister à des agressions externes physiques (incendie, inondation, bombes, etc.) ou logiques (erreurs de saisie, intrusions, piratages, logique, malicieuse, etc.). C'est généralement le sens choisi par les spécialistes de l'audit de sécurité, lorsqu'ils doivent, pour une entreprise donnée, évaluer les risques liés à l'informatique [32].

Mais plutôt que de définir la sécurité vis-à-vis des conséquences de la non-sécurité (au sens safety) ou vis-à-vis des agressions contre la sécurité (au sens "security"), il semble préférable, à l'instar des IT-SEC, de considérer la sécurité comme la combinaison de trois propriétés : la confidentialité, l'intégrité et la disponibilité de l'information [92].

Notons que ces trois propriétés se rapportent à l'information, et le terme d'information doit être pris ici dans son sens le plus large, couvrant non seulement les données et les programmes, mais aussi les flux d'information, les traitements et la connaissance de l'existence de données, de programmes, de traitements, de communications, etc. Cette notion d'information doit aller jusqu'à couvrir le système informatique lui-même, dont parfois l'existence doit être tenue secrète [32].

La sécurité, telle qu'elle est ici appréhendée, implique d'empêcher la réalisation d'opérations illégitimes contribuant à mettre en défaut les propriétés de confidentialité, d'intégrité et de disponibilité, mais aussi de garantir la possibilité de réaliser les opérations légitimes dans le système. Assurer la sécurité du système, c'est assurer que les propriétés retenues sont vérifiées, autrement dit, garantir la non-occurrence de défaillances vis-à-vis de ces propriétés [32].

Aujourd'hui Internet est devenu un élément primordial dans le monde informatique qu'on utilise dans notre quotidien, le commerce, la communication...etc.

En parallèle de cette fulgurante montée technologique il y'a eu expansion du nombre d'utilisateur d'Internet, on compte aujourd'hui plus de 2,5 milliards d'utilisateurs.

Par conséquent les informations circulant sur le net peuvent être très importantes, cruciales, secrètes et confidentielles. Chose que les concepteurs d'Internet n'ont pas prévue la sécurité de ces informations car leur but initial était de relier les différents réseaux informatique entre eux.

Tout système lors de sa conception présente des faiblesses que certains utilisateurs mal intentionnés peuvent exploités les vulnérabilités d'Internet pour accéder à des informations puis les lire, les modifier et même les détruire. Dès lors que ce réseau est devenu comme cible potentielle d'attaque, leur sécurité est devenue un point primordial et même crucial.

Pour cela le concept de sécurité informatique a été apparu afin de trouver des méthodes, des outils et des mécanismes de protection.

De nombreux algorithmes et techniques ont été introduits dans le domaine de la sécurité tels que les algorithmes de classification (data mining) et les algorithmes mathématiques complexes, mais le problème de la sécurité n'a pas été totalement résolu, comme le dit Donald L. Pipkin : « Il n'existe pas de réponse simple aux problèmes de sécurité. Malheureusement, les gens sont trop souvent convaincus que les seules choses dont ils aient besoin pour assurer leur sécurité, c'est d'installer un firewall, d'améliorer leur méthode d'authentification ou de définir un règlement de sécurité. En vérité, toutes ces mesures contribuent à l'amélioration de la sécurité, mais aucune d'entre elles n'est une solution complète. », auteur du livre « Sécurité des systèmes d'information » [88].

Une attaque informatique est devenue une arme destructrice pour les différents secteurs tel que l'industrie, les gouvernements, militaires... etc. comme ce fut le cas de l'attaque informatique qui a paralysé tout un pays comme l'ESTONIE en avril 2007, ou le sabotage du centre nucléaire en IRAN en septembre

2010 ou bien l'espionnage industriel ou la CHINE occupe le premier rang mondial. Donc l'enjeu de ces attaque est devenu un projet militaire stratégique et non pas un jeu de plaisir et de loisir d'où la notion de cyber guerre.

1.1 Les origines de l'insécurité

Les origines de l'insécurité d'un réseau informatique peuvent être énumérer comme suit [40] :

— **Les problèmes physiques :** L'accès aux matériels informatiques d'une entreprise ou une administration peut causer des dégâts catastrophiques pour une entreprise après l'exploitation de cet accès physique pour voler un mot de passe, effacer des données, usurper l'identité d'un autre ou injecter des programmes malicieux.

— **Les failles réseaux :** Depuis l'apparition des réseaux informatiques, d'énormes efforts fondés sur des normes et standards ont été consentis de plusieurs organismes collaborant entre eux pour pouvoir sécuriser au maximum les réseaux. Mais malheureusement, Il y'a toujours certaines failles de fonctionnement des standards exploitables. Le problème avec les failles réseau c'est la complexité de leurs corrections qui varie d'après la taille du réseau. Par exemple, la correction des failles réseau d'Internet est utopique, c'est la raison pour laquelle on se contente de faire des améliorations comme le passage vers IPV6 ou IPSec.

— **Les problèmes systèmes :** Différents mécanismes de sécurité, comme les mots de passe, les logs, séparation des privilèges...etc., sont intégrés dans les systèmes d'exploitation sophistiqués d'aujourd'hui. Mais leurs complexités, la mauvaise configuration ainsi que les faiblesses de certains de ces mécanismes des systèmes d'exploitation représentent un danger pour les utilisateurs. Par exemple la complexité d'un mécanisme de sécurité pousse les utilisateurs à le désactiver, de plus la mauvaise configuration peut engendrer l'arrêt ou la saturation du système.

— **Les problèmes applicatives :** Les problèmes applicatives sont très connues et très répondues qui peuvent engendrer beaucoup de problèmes influençant ainsi le fonctionnement du système et qui peuvent être causés par la mauvaise conception, non-traitement des exceptions, faille dans le langage de programmation.

— **Les problèmes Web :** Les problèmes web peuvent être causées par l'une des failles précédemment citées ou par des failles qui résident au niveau des protocoles et des standards du fonctionnement du web qui n'est autre que la combinaison de différents protocoles de : réseaux, système et application.

1.2 Les différents types d'attaque informatique

Selon la littérature toute tentative de destruction, de modification, de voler ou obtenir un accès non autorisé à une information ou à un élément physique d'un réseau comme un serveur est considéré comme une attaque informatique selon la norme ISO/IEC 27000 [34].

Il existe cinq formes d'attaque que nous détaillerons comme suit [25] :

— **Attaque passive :**Tout acte qui nous permet de faire, la surveillance, l'analyse et le décryptage des communications ainsi que la capture des informations d'authentification du trafic réseau, représente une attaque passive. Cette dernière peut entraîner la divulgation des informations ou des données à un attaquant sans que la victime soit consciente. À d'exemple l'interception du mot de passe, numéros de carte de crédit, des emails représente tous des attaques passives.

— **Attaque active :**Toute tentative ayant pour but de contourner ou arrêter les fonctions de protection, introduire un code malveillant et de voler ou modifier des informations représente une attaque active.

— **Attaque externe :**Elle représente l'utilisation de la proximité physique du réseau ou du système qui a été obtenue grâce à l'entrée clandestine ou un accès ouvert afin de modifier, collecter ou refuser l'accès à l'information.

— **Attaque interne :**Les attaques internes sont de deux genres :
  — Intentionnelles : représentent les tentatives d'espionner, de voler ou d'endommager des informations, utiliser l'information de manière frauduleuse, ou interdire l'accès à d'autres utilisateurs autorisés.
  — non intentionnelles : représentent le résultat d'une mauvaise manipulation, la négligence ou le manque de connaissances.
— **Attaque de distribution :**Toute modification malveillante du matériel ou du logiciel en usine ou lors de la distribution représente une attaque de distribution, qui consiste à introduire un code malveillant dans un produit comme un port dérobé pour obtenir un accès non autorisé à des informations ou une fonction système.

1.2.1 Exemples des attaques informatiques

Dans ce sous chapitre nous présenteront trois types d'attaques informatique très connues par leurs dangers et les préjudices qu'ils peuvent causer [2].

— **L'attaque man-in-the-middle :**L'attaquant s'introduit entre deux systèmes sans que l'un d'entre eux aperçoive l'existence d'un troisième système qui fait passer les échanges réseau. Pour réussir une telle attaque, il faut que la machine de l'attaquant soit physiquement entre les deux machines victimes ou que l'attaquant arrive à modifier le routage réseau afin que sa machine devienne un des points de passage [76]. Le schéma suivant illustre le fonctionnement de l'attaque man-in-the-middle (Voir Figure 1.1).

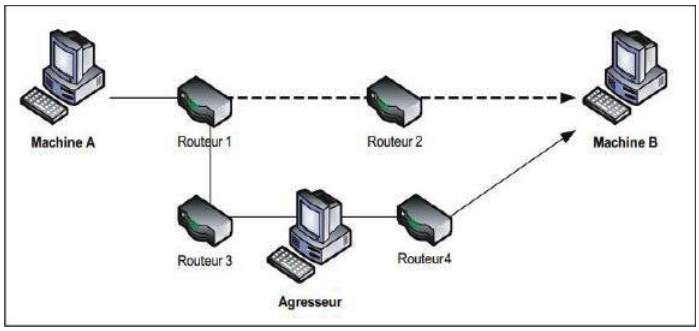


FIGURE 1.1 – L'attaque man-in-the-middle

— **L'attaque de déni de service distribué (DDoS) :**Elle représente la version distribuée de l'attaque de déni de service. Le but de cette variante de l'attaque DoS est que la victime n'arrive pas à isoler les attaquants vu le nombre important des machines utilisées pour réaliser cette attaque. Pour réaliser cette attaque, il faut premièrement pénétrer par diverses méthodes des systèmes dits "handlers" et agents. Où l'attaquant contrôle un ensemble de systèmes "handlers" qui contrôlent eux-mêmes un ensemble de systèmes agents. Le hacker lance l'attaque en ordonnant les systèmes "handlers", qui eux-mêmes ordonnent les agents [76]. Le schéma suivant illustre le fonctionnement de l'attaque DDoS(Voir Figure 1.2).

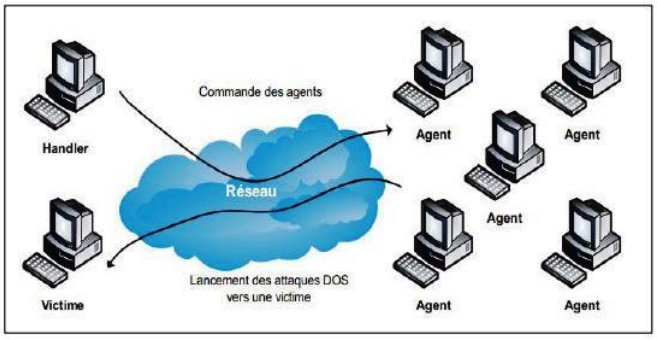


FIGURE 1.2 – L'attaque DDoS

— **Attaque par virus :**Un virus informatique est tout programme capable de se reproduire par lui-même, c'est le type d'attaque le plus fréquent. Il peut prendre la forme d'une routine ou d'un programme une fois activé il utilise tous les moyens pour empoisonner la vie de l'utilisateur. Plusieurs types de virus peuvent êtres cités [76] :

  — virus de secteur d'amorçage.
  — virus d'infection des fichiers (parasites).
  — virus non-résidents mémoire.
  — virus résidents mémoire.
  — virus polymorphes (mutants).
  — virus réseau et vers (worms).
  — bombes logiques. chevaux de Troie.

1.2.2 Les effets des attaques informatiques

Le nombre des attaques informatiques était très limité et était mené par des experts au début de l'histoire des réseaux informatiques. Contrairement aujourd'hui ou les outils de piratage informatiques sont disponibles aux amateurs avec quelques dollars à l'opposé des pertes qui peuvent être engendrées et qui sont de plus en plus graves. On parle aujourd'hui des milliards de dollars de perte, des pays paralysés, des projets stratégiques sabotés, des programmes présidentiels divulgués tout ça à cause des attaques informatiques qui varient dans le but, l'ampleur et la dangerosité [2].

La figure suivante (Voir Figure **??**) montre les pertes causées par des attaques informatiques dans une étude munie par le CSI (Computer Security Institue) sur 194 organisations. Le montant total de perte durant l'année 2007 est de 66.930.950 dollars American. Les fraudes financières occupent la première place avec 21.124.750$ USA, la deuxième place est occupée par les virus avec 8.391.800$ USA [94].

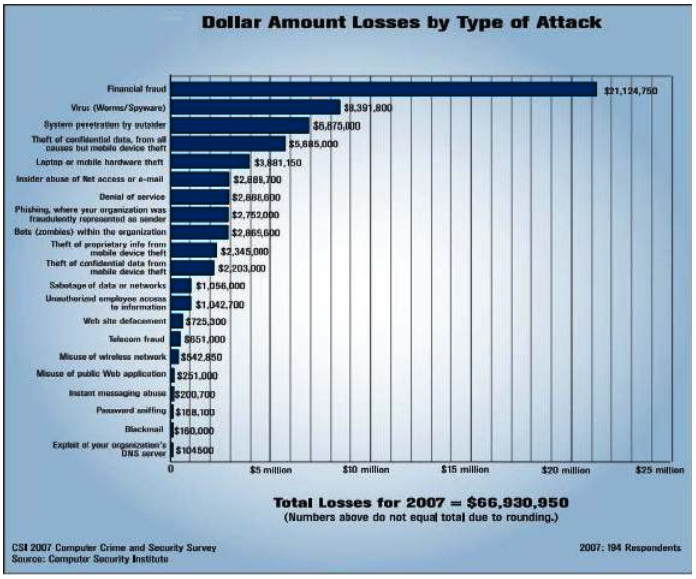


FIGURE 1.3 – Le rapport des pertes causées par des attaques informatiques sur 194 organisations durant l'année 2007

1.3 La sécurité informatique

1.3.1 Définitions de la sécurité informatique

Les diverses définitions de la sécurité informatique que relate la littérature spécifient que c'est l'ensemble des techniques et outils collaboratifs permettant de garantir trois objectifs essentiels de la sécurité (confidentialité, intégrité, et disponibilité) , ces outils peuvent êtres, organisationnels, matériels,

logiciels, ou juridiques dont le but est de protéger les informations et les systèmes d'information contre l'accès, l'utilisation malveillante ou non autorisée, la modification, la divulgation, et la destruction des données et connaissances [68].

1.3.2 Objectifs de la sécurité

L'objectif de la sécurité est d'assurer les cinq principes clés suivants [14] [25] :

— **La confidentialité**La confidentialité est définie par l'organisation internationale de normalisation (ISO) comme « le fait de s'assurer que l'information n'est seulement accessible qu'à ceux dont l'accès est autorisé », elle consiste à préserver la révélation non autorisée d'information sensible. La révélation pourrait être intentionnelle comme les attaques qui visent de casser le chiffrement des données et lire les informations, ou involontaire dû au manque de vigilance ou de l'incompétence des individus qui manient les informations.

— **La disponibilité**La disponibilité assure la pérennité du service opportun aux utilisateurs autorisés qui ont un accès non interrompu aux informations dans le système et le réseau.

— **L'intégrité**L'intégrité est la propriété d'une information de ne pas être altérée. Donc le système informatique doit :

  — Empêcher une modification par une personne non autorisée ou une modification incorrecte par une personne autorisée.

  — Faire en sorte qu'aucun utilisateur ne puisse empêcher une modification légitime de l'information.

  En plus, il faut se prémunir contre les fautes affectant l'intégrité des données, en intégrants dans le système des mécanismes permettant de détecter les modifications des informations d'une part et de contrôler l'accès à ces dernières d'autre part (en gérant les droits d'accès des programmes et utilisateurs). De plus, une validation en amont peut également être réalisée pour prévenir les fautes accidentelles.

— **L'authentification :**L'authentification est tout simplement le contrôle d'accès, ce service signifie que celle-ci les personnes autorisées peuvent accéder aux informations, des simples moyens comme la gestion des mots de passe permet de garantir les services d'authentification.

— **La non répudiation :** Cette propriété garantie qu'un sujet ayant réalisé une action dans le système ne puisse nier l'avoir réalisée. Assurer la non répudiation d'une transmission signifie assurer que les extrémités d'une transmission (émetteur et récepteur) sont bien les seules personnes autorisées à envoyer ou réceptionner les informations sans aucune remise en cause, qui est d'ailleurs généralement garanti par le moyen d'un fichier électronique appelé certificat numérique qui assure l'identité de l'émetteur et le récepteur. Les certificats eux mêmes sont protégés par le moyen des signatures des utilisateurs, dans certains cas la signature d'un utilisateur est son identité.

L'importance de ces principes diffère selon le contexte de l'application, par exemple :

— la confidentialité est la plus importante dans le cadre d'une transmission des messages secrets entre deux agences de sécurité nationale ou internationale, si quelqu'un arrive à décrypter le message transmis la sécurité sera compromise et l'information sera divulguée.

— Par contre la disponibilité est la plus importante pour les sites de e-commerce, la non-disponibilité est catastrophique pour des sites comme amazone et eBay [25].

1.3.3 La gestion des risques

La gestion des risques comprend trois processus : l'estimation des risques, la réduction des risques, et l'évaluation et l'estimation [76].

Le processus de l'estimation des risques comprend :

— L'identification et évaluation des risques.

— L'identification et évaluation de l'impact des risques.

— La recommandation des mesures pour la réduction des risques.

L'évaluation du risque est donnée par l'équation suivante :

$$\text{R}isque = \frac{menace * vulnérabilité}{contremenace} \tag{1.1}$$

Le processus de la réduction des risques comprend les tâches suivantes :

— Donner la Priorité aux mesures de réduction des risques recommandés par le processus de l'estimation des risques.

— Implémenter les mesures de réduction des risques recommandés par le processus de l'estimation des risques.

— Maintenir les mesures de réduction des risques recommandés par le processus de l'estimation des risques.

Le processus d'évaluation et d'estimation inclut un processus d'évaluation continu. Par exemple, l'autorité approbatrice désignée des États-Unis d'Amérique (DAA) détermine si un risque résiduel dans le système est acceptable ou que des mesures de contrôle et de protection supplémentaires devraient être implémentées pour accomplir l'accréditation d'un système informatique [25].

1.3.4 La politique de sécurité

Une politique de sécurité guide un système informatique selon des buts et des objectifs en fournissant un cadre pour la sélection et la mise en œuvre des contre-mesures contre les menaces, sans elle ce dernier est susceptible d'avoir un désordre de contre-mesures.

Une bonne politique est toujours adaptée aux menaces, car elle devrait préciser qui est responsable de quoi (mise en œuvre, exécution, vérification, examen), la nature de cette politique de sécurité du réseau et pourquoi elle est de cette nature. La réponse à ces questions est très importante parce qu'une politique claire, concise, cohérente et constante est plus susceptible d'être suivie.

La politique de sécurité est de savoir comment vous déterminez quelles contre-mesures il faut utiliser. Par exemple : Avez-vous besoin d'un pare-feu? Comment devez-vous configurer votre pare-feu? Avez-vous besoin d'un jeton d'accès, ou un mot de passe est suffisant? Les utilisateurs sont autorisés à accéder à la vidéo en streaming à partir de leurs navigateurs Web? S'il n'y a pas de politique, il n'y aura pas de base pour répondre systématiquement à ces questions. Malheureusement, la plupart des organisations n'ont pas une politique de sécurité réseau. Ou bien ils le font, mais personne ne la suit [107].

Les politiques de sécurité sont différentes d'une organisation à l'autre, mais dans le cas le plus simple une politique de sécurité devrait comprendre les éléments suivants [84] :

— Une explication claire et simple du but de la politique de la sécurité et les objectifs et l'importance de cette stratégie de sécurité pour l'entreprise.

— Une déclaration de soutien pour cette politique de sécurité de la part des cadres supérieurs de l'organisation, ce qui prouve leurs engagements.

— Offrir des formations afin d'aider les employés à comprendre la sécurité de l'information et les dégâts qui peuvent être causés si on enfreint la politique de sécurité.

— Une explication simple et claire des normes minimales de sécurité en mettant l'accent sur les procédures à suivre dans des domaines qui relève une importance particulière pour l'entreprise. Par exemple, toute politique de sécurité doit préciser les précautions élémentaires concernant les virus informatiques, des consignes pour la mise en place des mots de passe...etc.

— Définir les rôles et les responsabilités de la sécurité des informations au sein de l'organisation.

— Exiger des rapports, réponses, résolutions pour n'importe quel type d'incident de sécurité au sien de l'organisation.

— La nécessité d'un plan de continuité des activités, ce qui explique comment l'entreprise va continuer à fonctionner en cas d'une défaillance catastrophique, comme un incendie ou une inondation.

— Avoir un support bien documenté pour le référencement au sien de l'organisation comme la politique, les guides, les procédures et les standards de sécurité.
  — Les politiques : c'est les documents non techniques qui décrivent d'une manière formelle les principes ou les règles auxquelles se conforment les personnes qui reçoivent un droit d'accès au capital technologique et informatif de l'organisation.
  — Les guides : c'est les documents qui complètent les documents de la politique où ils détaillent comment implémenter cette politique de sécurité.
  — Les standards : c'est les documents de standardisation des normes et des méthodes provenant d'organismes internationaux tels que l'ISO, (International Standardization Organization), l'IETF (Internet Engineering Task Force), l'IEEE (Institute of Electrical and Electronics Engineers)... etc.
  → Les procédures : c'est les documents techniques qui décrivent d'une manière claire et précise les étapes à suivre pour atteindre un objectif de sécurité donné[76] [2].

1.3.5 Les techniques de sécurité

La mise en œuvre d'une politique de sécurité consiste à déployer les différents moyens et dispositifs visant la sécurisation du système d'information ainsi que l'application des règles définit dans la politique de sécurité adoptée. Ce qui signifie, faire le bon choix de l'ensemble des mécanismes et des techniques les plus simples possible permettant de protéger les ressources d'une manière très efficace avec un faible coût. Il existe différentes techniques utilisées contre les attaques informatiques, ces techniques sont classées en cinq catégories qui sont [76] [2] :

La sécurisation des accès réseau

La maîtrise du flux réseau à l'aide des pare-feux assure un niveau de confidentialité des données grâce aux protocoles de sécurité tel que l'IPSec, ce qui permet la sécurisation des accès réseau [76].

Superviser les connexions réseau La vérification du trafic réseau consiste à ne laisser passer que les connexions autorisées. Cela est possible par [76] :

— la création d'un périmètre de sécurité,
— limiter le nombre de points d'accès pour rendre la gestion de la sécurité plus facile
— et disposer de trace des systèmes en cas d'incident de sécurité.

Nous citons certains dispositifs de contrôle et de filtrage de connexion :

— Le pare-feu : C'est le système qui permet de mettre en œuvre la politique du filtrage au sien de l'organisation, selon plusieurs principes de filtrage :
  — le filtrage des paquets au niveau réseau (IP, etc.),
  — le filtrage à mémoire des paquets de manière dynamique,
  — la passerelle de niveau transport filtrant les paquets en gérant le concept de session
  — la passerelle de niveau applicatif filtrant les paquets du niveau applicatif.
— Contrôle de l'accès réseau : C'est un nouveau concept développé par Cisco, et ayant pour but le contrôle des accès les plus près à leurs sources où il permet de vérifier un certain nombre de points de sécurité avant d'autoriser un système à se connecter au réseau local.

Assurer la confidentialité des connexions La confidentialité des données est assurée au sien d'un réseau informatique par l'utilisation du chiffrement, par un cryptage des données avant leur envoi et un décryptage à leur réception. Le schéma suivant (Voir Figure 1.4) montre ce le chiffrement dans l'architecture de communication TCP/IP [76].

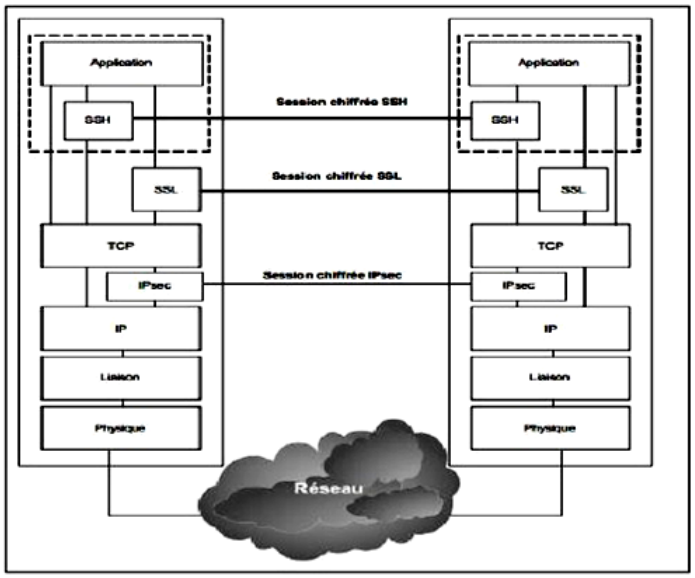


FIGURE 1.4 – La représentation en couches des protocoles de sécurité

— Les algorithmes cryptographiques : le chiffrement-déchiffrement des données est effectué par des algorithmes cryptographiques qui reposent sur des problèmes mathématiques difficiles à résoudre. Il existe deux grandes catégories d'algorithme de cryptographie :
  — les algorithmes cryptographiques à clé secrète ou symétrique qui se basent sur une même clé qui chiffre et déchiffre. Cette clé est partagée par les deux communicants.
  — Les algorithmes cryptographiques à clé publique ou asymétrique qui se basent sur une clé publique de chiffrement et une clé secrète de déchiffrement [76].

Il existe aussi les algorithmes de hachage qui nous permettent d'obtenir une signature numérique à partir des données comme :

— IPSec : il est créé pour faire face aux problèmes d'authentification et de confidentialité du protocole IP. IPSec opère au niveau IP et il encapsule nativement tous ces protocoles (TCP, UDP, ICMP, etc.). IPSec offre des services de contrôle d'accès, d'intégrité, d'authentification, de confidentialité de plus il fait face aux attaques de type paquets replay [76].

— SSL (Secure Sockets Layer) : opère au-dessus de la couche TCP et offre aux navigateurs internet la possibilité d'établir des sessions authentifiées et chiffrées. Le protocole SSL a été standardisé par le groupe de travail TLS (Transport Layer Security) formé au sein de l'IETF [76].

— SSH (Secure Shell) : il opère au niveau application et permet d'obtenir un interprète des commandes (Shell) à distance d'une manière sécurisé.

La protection des équipements réseau

Protéger un réseau informatique c'est assurer la protection des équipements qui le composent et qui recouvre les trois domaines suivants [31] :

— **La protection physique :** c'est la sécurité physique des équipements face aux menaces physiques externes comme le feu, l'inondation, le survoltage, l'accès illégal à la salle informatique... etc.

— **La protection du système d'exploitation :** c'est la sécurité des systèmes d'exploitation contre les faiblesses de sécurité ou les bugs.

— **La protection logique :** la mise en œuvre d'une politique de sécurité passe par une configuration de l'équipement réseau.

La sécurité des équipements réseau nous permet de se protéger contre les attaques suivantes [31] :

— Les attaques par déni de service visant à exploiter des faiblesses de configuration.

— Les attaques permettant d'obtenir un accès non autorisé à un équipement réseau suite à des faiblesses de configuration.

— Les attaques exploitant un bug référencé du système d'exploitation Cisco, Microsoft, RedHat...

La sécurité des applications réseau et systèmes

Un réseau informatique nous offre un ensemble de services sur des systèmes dédiés. Ainsi la protection de ces systèmes ainsi que les applications qui nous offrent ces services passe par l'implémentation de ces services, mais uniquement ces services. Pour cela, il faut séparer les plates-formes, sécuriser les systèmes d'exploitation, configurer les pare-feux, sécuriser le contrôle d'intégrité, maîtriser la sécurité des applications [119].

Séparation des plates-formes Elle consistant à déployer des services de nature différente sur des plates-formes distinctes. Cette approche implique un surcoût de déploiement et d'administration, mais elle offre aussi une facilité de déploiement et de gestion et une grande résistance contre les attaques.

Protection des systèmes d'exploitation La protection d'un système d'exploitation passe par le déshabillage (strip-down) et le blindage (hardening) du système [76].

— **Le déshabillage :** Déshabiller un système d'exploitation c'est désactiver tous les services réseau inutiles ou dangereux. Cette étape est très importante, car un système qui n'entend pas certaines requêtes est complètement immunisé contre les attaques ciblant ces services. Par exemple, les services tels que Berkeley (rsh, rexec, rlogin) et Telnet doivent être désactivés sur les systèmes Unix.

— **Le blindage :** ça consiste à appliquer systématiquement la règle du privilège minimal. Par exemple :

  — rétrogradation ou redéfinition des privilèges sur les processus,

  — synchronisation de l'horloge du système sur au moins deux sources fiables,

  — installation d'un système de vérification de l'intégrité des répertoires et des fichiers stables.

Les pare-feux Deux grandes catégories des pare-feux sont utilisées :

— ceux qui visent la protection d'une zone en coupure de ligne,

— ceux qui contrôlent uniquement les accès au système hôte.

Le pare-feu zonal comme le pare-feu embarqué peuvent être "stateless", "stateful" ou "proxy" au niveau applicatif.

Il existe plusieurs facteurs qui interviennent dans le choix du type de pare-feu, comme l'isolement topologique du serveur, la différence du type de contrôle nécessaire, une autorité différente...etc [76].

Protection des droits d'accès Elle consiste à authentifier les accès sur une base individuelle et chaque profil doit respecter la règle du plus bas niveau des privilèges, où le niveau des privilèges monte sur une base temporaire pour effectuer une tâche bien précise, puis il redescend à son niveau initial. Il faut distinguer la gestion des droits d'accès à une plate-forme donnée de la gestion des droits d'accès à une application implémentée par un programme. Parmi les gestionnaires des droits d'accès, on peut citer :

— L'annuaire LDAP qui représente une structure centrale qui peut supporter tout type de table, y compris des tables d'authentification.

— Kerberos est un système qui gère les droits d'accès des systèmes distribués. Il est basé sur la notion des tickets. On peut le déployer sur Windows comme l'Unix [76].

Protection du contrôle d'intégrité Une politique de sécurité repose sur une partie principale qu'est le contrôle d'intégrité. Pour n'importe quelle politique de sécurité, on doit vérifier l'intégrité de l'implémentation et celle de la configuration du système. Il existe deux méthodes de vérification d'intégrité [119] :

— faire une copie de tous les fichiers système et l'archiver puis comparer la version actuelle avec la version archivée.

— créer une signature numérique et l'archiver puis comparer la signature numérique de la version actuelle avec la signature archivée.

Protection des applications Malgré le nombre important des vulnérabilités qui existent dans les applications, on peut toujours créer une suite de logiciel robuste comme l'a démontré l'équipe de développement du système OpenBSD. Pour cela, il faut respecter ces quatre pratiques [76] :

— Codage défensif : il faut appliquer certaines règles simples, mais très importantes comme la validation des entrées, le contrôle de la gestion de la mémoire dynamique, l'application des privilèges minimaux...etc.

— Environnements d'exécution sécurisés : c'est la recompilation du code source d'un programme afin d'utiliser d'une manière transparente des environnements d'exécution sécurisés.

— Environnements cloisonnés : c'est installer un programme relatif à l'application dans une zone cloisonnée selon deux techniques de cloisonnement :

  — les cloisonnements système de type "prison" et "parking" comme la primitive "chroot" de l'Unix et la machine virtuelle.

— Tests de validation : faire les tests de logiciel comme :

  — les tests d'endurance aux entrées illégitimes,

  — les tests d'endurance avec des entrées aléatoires,

  — analyse rétrospective du code source...etc.

1.3.6 Processus de sécurité :

La sécurité est un processus garanti au moyen de plusieurs outils, Mr. DONALD Pepkin a défini cinq mécanismes importants dans son livre « **sécurité des systèmes d'information** » pour élaborer un bon plan de sécurité. [88]

Inspection :

C'est un mécanisme d'identification des fonctionnalités de base pour l'évaluation des besoins sécuritaires de l'entreprise, l'inspection est applicable en six étapes :

— Inventaire des ressources.

— Estimation de la menace.

— Analyse des pertes potentielles.

— Identification des vulnérabilités.

— Organisation de la protection.

— Évaluation de l'état actuel.

Protection :

C'est un mécanisme assuré via un ensemble de moyens au but de réduire dynamiquement les risques, parmi les moyens utilisé :

— Le logiciel antivirus.

— Le contrôle d'accès (Authentification).

— Le pare-feu.
— L'établissement de procédure de sécurité.
— Le chiffrement des données.
— Les mécanismes de sécurité physique.
— Les sauvegardes.
— Les mécanismes de redondance de l'information.

Détection :

Est un mécanisme de réaction contre les risques, la détection est basée généralement sur la prévention des attaques avant quelles seraient exploitées, ce mécanisme se compose de trois processus de base :

— **Analyse de signature :** Ce processus permet la récupération des informations concernant l'ouverture d'une session afin de la comparer avec une base de traces des attaques pré connues pour le système ainsi permet le sauvegarde de nouvelles attaques dans la base (les attaques sont identifiés par leur signatures)
— **Analyse statique :** Ce processus permet la détection des failles sur les produits afin de définir le niveau de risque et prévenir les dégâts pouvant être causés par ces vulnérabilités.
— **Analyse dynamique :** Ce processus permet la détection des attaques au moment d'interception du comportement normale du produit en basant sur le résultat de l'analyse de signature.

Réaction :

Ce mécanisme consiste à définir un plan de secours en cas de détection d'une attaque, le plan de secours se roule généralement sur trois composants importants :

— **Surveiller et avertir :** Ce composant consiste à envoyer des messages d'alertes à l'administrateur une fois un fait suspect est détecté alors que l'administrateur à son tour doit lancer une analyse du fait.
— **Réparer et signaler :** Certaines attaques s'agissent d'être remédier le plus vite possible, dans ce cas l'intervention du système de sécurité ne doit pas par la perte du temps d'avertir l'administrateur alors qu'il essaye de régler le problème tous seul puis envoyer un rapport à l'administrateur pour l'informer sur l'état de comportement avant, lors et après la correction.
— **Poursuivre en justice :** Dans le cas des dégâts importants et catastrophiques sur le produit, l'attaquant doit être poursuit en justice, c'est pour cela que le composant de poursuivre doit informer le service juridique de l'entreprise.

Les messages d'alertes ainsi que les rapports d'information sont de différents types :

— Documentation
— Détermination
— Notification
— Appréciation
— Éradication
— Remise en état

Réflexion :

Ce mécanisme est exécuté après la remise du système à son état normale, ce mécanisme permet l'étude de l'évènement afin de réaliser un plan d'intervention et de protection en basant sur le résultat des dégâts causés par l'attaque, le mécanisme de réflexion se roule en quatre étapes :

— Documentation de l'incident

— Évaluation de l'incident
— Relations publiques
— Suites judiciaires

Conclusion

La sécurité des systèmes informatiques nécessite beaucoup de surveillance et de précaution car il existe une multitude de vulnérabilités auxquelles il faut faire face en utilisant différents outils et techniques de sécurité informatique. Par conséquent, une bonne configuration, protection du réseau, système et applications est la solution de sécurité adéquate au sien d'une organisation. Nous devons donc toujours prévoir des solutions à toute sorte d'attaques en suivant les différentes étapes nécessaires à la sécurisation des données de l'organisation des attaques irréversibles, et comme on dit : « mieux vaut prévenir que guérir ».

Chapitre 2

Les systèmes de détection d'inturusion

Sommaire

Introduction

Pour faire face aux problèmes de sécurité que les systèmes et réseaux informatiques enduraient, une multitude de solutions sont proposées et mis en place pour prévenir toute sorte d'attaque comme les pare-feux, l'authentification, les proxys... etc. Néanmoins ces mécanismes ont des limites où certains types d'attaques peuvent les contourner pour nuire à la confidentialité, l'intégrité ou la disponibilité. Par conséquent, un nouveau concept appelé système de détection d'intrusion a été introduit comme une seconde ligne de défense afin de renforcer la sécurité des systèmes informatiques [5], où l'importance des traces d'audit a été souligné pour relever toute violation potentielle de la politique de sécurité. Cette idée fut concrétisée par Denning en 1987 qui a proposé le premier modèle de détection d'intrusions dans son article « An intrusion detection model » [30].

En 1988, il existait au moins trois prototypes : Haystack [112], NIDX [10] et [108]. Mais ce nombre de prototypes s'est considérablement augmenté par la suite, surtout aux états unis ou le gouvernement qui était le précurseur dans le domaine de la sécurité a investit plusieurs millions de dollars dans ces recherches pour renforcer au maximum la sécurité des machines et réseaux.

2.1 La vérification (audit) de sécurité

Un événement est une activité système produite suite à un ensemble d'action effectué par un utilisateur, processus ou application, à un moment donné. Le journal d'audit de sécurité c'est le fichier qui enregistre chronologiquement tout ou une partie des événements produits dans un système donné. Généralement le journal d'audite nous permet de connaitre l'opération faite, l'utilisateur qui la fait, quand il la fait, les ressources système affectés par cette opération, l'utilisateur a pu terminer l'opération sinon pourquoi l'opération a échoué [3].

L'accomplissement de audite de sécurité doit répondre aux questions suivantes :

— Quoi vérifier ?

— Comment le collecter ?

— Comment l'analysé ?

2.1.1 La vérification des activités du système

Selon la politique de sécurité et le niveau de sécurité désiré, l'administrateur peut définir des événements vérifiables correspondants à certaines informations qui peuvent être [3] :

— **Un accès au système :** Ce sont les informations nous permettant de détecter toute violation de sécurité comme : qui a accédé au système ? Quand ? Où ? et comment ?

— **Un usage des ressources système :** Ce sont les informations relatives à l'utilisation des ressources système tel que : les commandes système, l'utilisation de CPU, RAM, les entrées/sorties.

— **Un usage des fichiers :** Ce sont les informations concernant l'accès aux fichiers comme l'horodatage de l'accès, type d'accès, source d'accès...etc.

— **Des événements liés aux applications :** Ce sont les informations relatives aux événements engendrer par des applications qui peuvent influencer la sécurité du système comme le lancement et l'arrêt des applications, les entrées utilisés et les sorties produites...etc.

— **Les violations éventuelles de la sécurité :** Ce sont les informations relatives aux tentatives d'accès non autorisé à des ressources système comme l'exécution d'une application ou d'une commande en mode privilégie, changement des droits d'accès...etc.

— **Les statistiques du système :** représentent les informations statistiques nous permettant à repérer toute activité anormale comme les statistiques sur le nombre de tentatives d'accès refusés.

2.1.2 La collecte des événements

Les systèmes d'exploitation actuels possèdent un mécanisme d'audit capable de générer certains types d'événement. Où le noyau assure la génération et la collecte de ces événements. Concernant les applications on doit fournir aux développeurs un ensemble de primitives de génération et de collecte des événements. Par conséquent nous aurons l'audit système et application [3].

2.1.3 L'analyse du journal d'audit

L'analyse est la détection de toutes les violations potentielles de la politique de sécurité qui peuvent atteindre la confidentialité, l'intégrité ou la disponibilité. La fréquence d'analyse peut être faite soit en temps réel ou en temps différé. Pour minimiser les dégâts en cas de violation de la politique de sécurité, il faut surveiller le système en temps quasi réel. Dans le cas d'un réseau, il faut créer un fichier d'audit global qui regroupe les différents audits collectés des différentes machines du réseau [64].

Il est primordial de protéger le journal d'audit contre toute tentative de modification par certains utilisateurs non autorisés et voulant modifie le journal d'audit afin d'effacer toute trace qui peut révéler leur intrusion. Pour remédier à ce problème, dans le cas d'un réseau informatique, une protection simultanée du fichier global d'audit et le transfert des informations auditées s'impose.

2.2 Les systèmes de détection d'intrusion

2.2.1 L'intrusion

L'intrusion est toute utilisation d'un système informatique d'une manière frauduleuse et bien entendu à des fins autres que celles prévues. Comme interrompre le fonctionnement d'un système, usurper l'identité d'un utilisateur ou modifier des informations. Cette définition comprend toutes les tentatives qui réussissent ou celles qui échouent [41].

Selon Heady and al [56], une intrusion dans un système informatique est définie comme : « N'importe quel ensemble d'actions essayant de compromettre l'intégrité, la confidentialité ou l'accessibilité d'une ressource ». Les intrusions sont regroupées en deux catégories :

— Les intrusions connues : se sont des attaques répertoriées exploitant les faiblesses identifiées du système cible.

— Les intrusions inconnues ou anomalies : elles sont détectées dès qu'il y a une déviation par rapport du profil normale du système.

2.2.2 Détection d'intrusions

La détection d'intrusions c'est l'analyse des informations collectées par les mécanismes d'audit de sécurité, à la recherche d'éventuelles attaques sur les systèmes informatiques. Les méthodes de détection d'intrusion diffèrent sur la manière d'analyser le journal d'audits [13].

2.2.3 Définition d'un IDS

Un système de détection d'intrusion (IDS) est tout équipement, méthode ou ressource nous permettant de surveiller un réseau ou un hôte donné afin de prévoir ou identifier toute action suspecte et non autorisée et éventuellement réagir à cette action.

Les systèmes de détection d'intrusion actuels détectent les activités réseau qui peuvent être une intrusion ou non et non pas l'intrusion. La détection d'intrusion est précisément une partie d'un système de protection total installé autour d'un système ou appareil. Il n'est pas une mesure de protection autonome [41].

Un système de détection d'intrusion standard

Un modèle général qui englobe et standardise la structure d'un système de détection d'intrusion a été le centre d'intérêt du groupe IDWG (Intrusion Detection Working Group) de l'IETF, qui a proposé le modèle général des systèmes de détection d'intrusion comprenant :

— un senseur (collecteur),

— analyseur,

— manager (administrateur).

La figure suivante (Voir Figure 2.1) montre en détail les composants d'un système de détection d'intrusion [121].

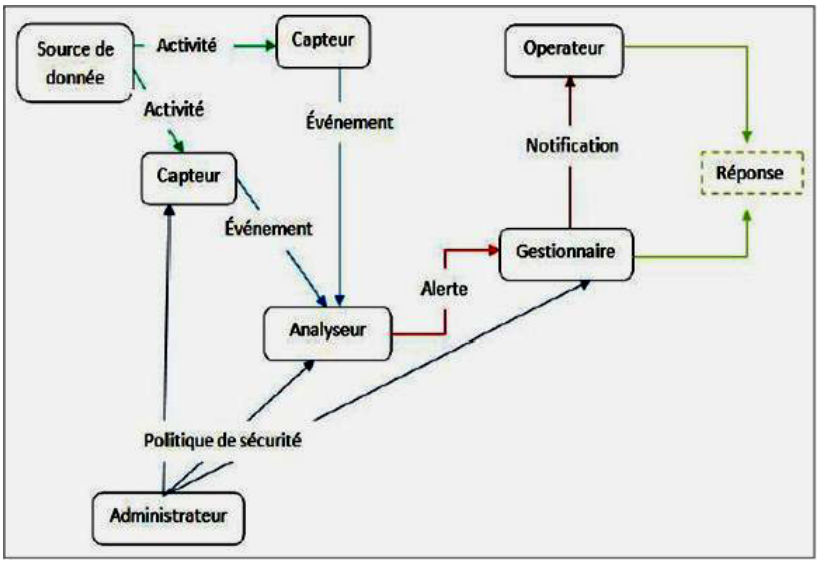


FIGURE 2.1 – Le modèle générique de la détection d'intrusions proposé par l'IDWG

— **L'activité :** c'est les éléments de la source de données qui sont identifiés par le capteur ou l'analyseur et ayant un intérêt pour l'opérateur. Par exemple les entrées des fichiers journaux du système d'exploitation montrant un utilisateur essayant d'accéder à des fichiers qui ne lui sont pas autorisés, les fichiers journaux d'application montrant des échecs de connexion persistants...etc.

— **L'administrateur :** c'est l'architecte de l'élaboration de la politique de sécurité d'une entreprise ou organisation, il déploie et configure l'IDS. Dans certaines organisations l'administrateur est associé à un réseau ou à des groupes d'administration de système. Dans d'autres organisations, c'est une position indépendante.

— **L'alerte :** c'est un message qui passe de l'analyseur au gestionnaire pour lui informer qu'un événement d'intérêt a été détecté. Une alerte contient généralement des informations sur l'activité inhabituelle qui a été détectée ainsi que ces détails.

— **L'analyseur :** Il signale les activités non autorisées ou indésirables ou les événements qui pourraient avoir un intérêt pour l'administrateur de sécurité en analysant les données recueillies par le capteur, c'est le composant clé. Dans la plupart des IDSs existants, le capteur et l'analyseur font partie d'un même composant.

— **La source de données :** Elle représente les informations brutes utilisées par le système de détection d'intrusion pour détecter les activités non autorisées. Les sources de données communes incluent les paquets bruts du réseau, les journaux d'audit du système d'exploitation, les journaux d'audit d'applications et les données de contrôle générées par le système.

— **L'événement :** c'est toute occurrence détectée dans la source des données par un capteur et qui peut donner lieu à une alerte. Par exemple une attaque.

— **Le gestionnaire :** à partir de cet élément clé que l'opérateur gère les différents composants du système. Les fonctions du gestionnaire comprennent généralement :

- la configuration du capteur,
- la configuration de l'analyseur,
- la gestion de la notification d'événements,
- la consolidation des données et la gestion des rapports.

— **La notification :** c'est la méthode avec laquelle le gestionnaire de l'IDS informe l'opérateur de la survenance d'une alerte. Généralement la notification se fait via :
- l'affichage d'une icône colorée sur l'écran du gestionnaire de l'IDS,
- la transmission d'un e-mail ou un message,
- ou la transmission d'un Sample Network Management Protocol (SNMP) trap... etc.

— **L'opérateur :** c'est l'utilisateur principal du gestionnaire de l'IDS. L'opérateur surveille souvent la sortie du système de détection d'intrusion et déclenche ou recommande d'autres actions.

— **La réponse :** c'est les mesures prises contre un événement, et effectuées automatiquement par une entité dans l'architecture de l'IDS ou initiées par un humain. L'envoi d'une notification à l'opérateur est une réponse très commune, d'autres réponses incluent :
- la journalisation de l'activité,
- l'enregistrement des données brutes (à partir de la source de données) qui ont caractérisé l'événement,
- l'arrêt du réseau ou de l'utilisateur ou la session de l'application,
- la modification des contrôles d'accès réseau ou système.

— **Le capteur :** Il collecte les données à partir de la source de données selon une fréquence de la collecte des données qui varie selon la configuration de l'IDS, et il est mis en place pour transférer des événements à l'analyseur.

Modèle simplifié d'un IDS

Selon Hervé Debar, on a simplifié un IDS en un détecteur qui analyse les informations en provenance du système surveillé (Voir Figure 2.2) [29].

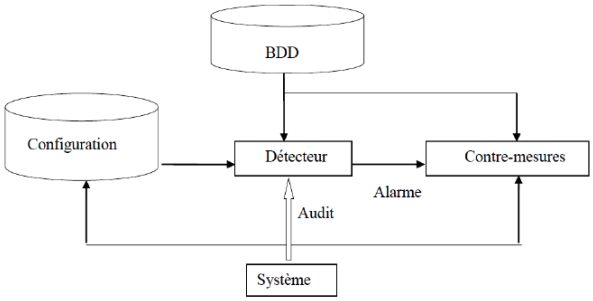


FIGURE 2.2 – Modèle simplifié d'un système de détection d'intrusions

Un détecteur, qui est l'élément principal de cette architecture, analyse trois types d'informations :

1. Les informations relatives aux techniques utilisées dans la détection (Base de donnée de signatures).
2. Les informations relatives à la configuration du système déterminant son état actuel.
3. Les informations relatives à l'audit décrivant les évènements survenus dans le système.

Pour définir l'efficacité d'un IDS, trois critères de base sont définis [90] :

1. **L'exactitude :** c'est lorsqu'une activité légale est déclarée comme une intrusion, ce qui correspond aux faux positif.

2. **La performance :** c'est le taux de traitement des évènements. S'il est faible, la détection est impossible.
3. **La complétude :** c'est lorsqu'on rate la détection de l'attaque, c'est le plus difficile critère car on ne connaît pas l'ensemble des attaques, ce qui correspond aux faux négatifs.

   On rajoute à ceux-ci deux autres critères [29] :
4. **La tolérance aux fautes :** les IDS doivent eux-mêmes résisté aux attaques, spécialement au déni de service, car généralement ils s'exécutent sur des matériels ou logiciels connus vulnérables aux attaques.
5. **La réaction en temps réel :** la diffusion des résultats de l'analyse doit se faire en temps réel pour permettre au manager de la sécurité de prendre la décision adéquate afin d'éviter tout dommages graves du système et en même temps réagir à cet évènement.

2.2.4 Les Caractéristiques d'un système de détection d'intrusions

Tout IDS doit présenter les caractéristiques suivantes :

— Pouvoir effectuer une surveillance permanente et émettre une alarme en cas d'intrusion.

— Fournir beaucoup d'information pour pouvoir réparer le système et la responsabilité de l'intrus.

— S'adapter aux différentes plates formes et architectures réseaux par sa modulation et configuration.

— Assurer sa propre défense en supportant qu'une partie ou la totalité du système soit hors service.

— Avoir un taux de faux positifs faible.

— Avoir une réponse automatique en cas d'attaques coordonnées ou distribuées.

— Être en mesure de repérer les premiers événements de corruption pour réparer correctement le système.

— Ne pas créer de vulnérabilités supplémentaires.

Les systèmes de détection d'intrusion offrent beaucoup d'avantages comme :

— Beaucoup plus efficace qu'une détection manuelle des intrusions.

— La prédiction des intrusions par l'utilisation d'une base de connaissance plus grande.

— La capacité de traiter un large volume de données.

— La réaction de l'IDS par une alerte en temps réel réduit le dommage important des attaques.

— Des mesures de contre-attaque automatique sont prises comme la fermeture des sessions, désactivation des comptes utilisateur...

2.3 Classification des systèmes de détection d'intrusion

Les Systèmes de détection d'intrusion existants peuvent être classifiés d'après plusieurs critères. Nous citons dans la figure suivante (Voir Figure 2.3) les cinq critères de classification des IDS introduits par Hervé Debar, Marc Dacier et Andreas Wespi [29].

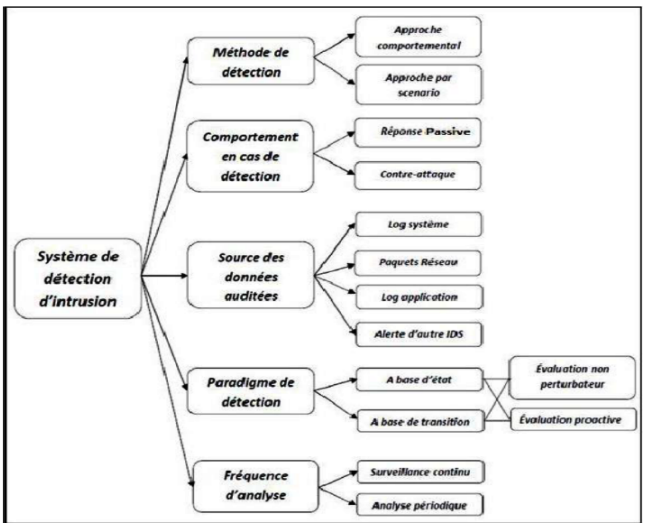


FIGURE 2.3 – Les critères de classification des IDSs

2.3.1 Les méthodes de détection d'intrusion

A ce jour deux méthodes de détection sont proposées, la première est appelée approche Comportementale (anomaly detection) qui consiste à créer un modèle basé sur le comportement normal du système et toute déviation par rapport à celui-ci est considérée comme suspecte. La seconde est appelée approche par scénario (misuse detection ou knowledge based detection) qui consiste à utiliser des connaissances accumulées sur les attaques puis on en tire des scénarios d'attaques et on recherche dans les traces d'audit leur éventuelle survenue [29], [83].

L'approche comportementale

Cette approche a été proposée par Anderson [5] puis développée par Denning [30], et se base sur l'hypothèse que l'exploitation d'une faille du système nécessite une utilisation anormale du système, donc un comportement inhabituel de l'utilisateur. Elle repose sur l'observation du système et toute déviation par rapport au comportement normal prévu du système est considérer comme intrusion.

Cette approche consiste, dans une première phase, à définir un modèle de comportement du système, des utilisateurs des applications, etc. qui sera considéré comme « normal ». Dans une seconde phase, l'activité actuelle du système est confrontée avec le modèle établi dans la phase une par l'IDS. En cas où une déviation est détectée, une alerte sera déclenchée.

En plus, cette approche considère comme intrusion, tout comportement qui n'est pas précédemment enregistré. Par conséquent, la précision reste son plus grand souci [5] [30].

Avantages :

— La détection de nouvelles formes d'attaques exploitant de nouvelles formes de vulnérabilités non connues auparavant.

— Elle est moins dépendante du système d'exploitation par rapport à l'approche par scénario.

— Elle détecte les attaques d'abus de privilège qui n'exploite aucune vulnérabilité.

Inconvénients :

— Un taux de fausses alarmes très élevé parce que l'ensemble du périmètre du comportement d'un système d'information ne peut pas être complètement couvert pendant la phase d'apprentissage. En outre, le comportement peut changer au fil du temps. Ce qui nous oblige à refaire l'apprentissage du comportement normal, ce qui cause soit l'indisponibilité temporaire du système de détection d'intrusion ou des fausses alarmes supplémentaires.

— De plus le système d'information peut subir des attaques au moment d'apprentissage. Par conséquent, le profil de comportement normal contiendra des comportements intrusifs qui ne seront pas détectés comme anormale.

— En cas de profondes modifications de l'environnement du système cible, le modèle statistique déclenche un flot d'alarmes ininterrompu, du moins pendant la phase de transition du système.

— Il est difficile d'affirmer que les observations faites sont des activités à prohiber.

Les systèmes de détection d'intrusion basés sur cette approche sont implémentés à l'aide de différentes techniques tel que : les systèmes experts, la méthode statistique, les réseaux de neurones [29].

L'approche par scénario

Elle vise à détecter des signes d'attaques connues selon une BDD d'attaques connues (les connaissances accumulées sur des attaques spécifiques et les vulnérabilités du système). Le système de détection d'intrusion contient les informations sur les vulnérabilités et cherche toute tentative de les exploiter. L'IDS confronte le comportement observé du système à la base de données d'attaques connues, Si ce comportement correspond à l'une des signatures de la base une alerte est déclenchée. En d'autres termes, toute action qui n'est pas explicitement reconnue comme une attaque est considérée comme acceptable. Par conséquent, la précision des systèmes de détection d'intrusion basée sur l'approche par scénario est bonne. Cependant, cette précision dépend toujours de la mise à jour des connaissances sur les attaques qui doit être régulière [83], [29], [126].

Avantages : les avantages les plus importants que l'on peut retenir de cette approche sont :

— le taux très faible de fausse alarme.

— l'analyse contextuelle très détaillée.

Cette approche a permis à l'administrateur de sécurité, grâce à la compréhension des problèmes de sécurité et la prise des mesures préventives ou correctives qui sont devenues plus faciles, de sécuriser convenablement son système.

Inconvénients : les inconvénients se résument en :

— Les bases de signatures sont difficiles à construire.

— La difficulté de mettre à jour la base des attaques connues avec les nouvelles vulnérabilités.

— Il faut maintenir la base de connaissance du système de détection d'intrusion par une analyse approfondie de chaque vulnérabilité, et qui n'est pas un travail facile à réaliser.

— Un IDS est lié à son environnement par le fait que la connaissance des attaques est très liée au système d'exploitation, la version, la plateforme et l'application.

— Les attaques par abus de privilèges (interne) sont difficiles à détectées, car il n'y a aucune vulnérabilité exploitée par l'attaquant.

— Il n'y a pas de détection d'attaques non connues.

Actuellement, de nombreux systèmes de détection d'intrusion dites par scénario ont été implémentés à l'aide de multiples techniques tel que : les systèmes experts, l'analyse de la signature, les réseaux de pétri, l'analyse de l'état transition.

Les systèmes de détection d'intrusion commerciaux actuels se penchent vers l'approche par scénario pour les raisons suivantes :

— L'approche par scénario est plus facile que l'approche comportementale dans l'implémentation.

— le taux élevé de fausse alarme pour l'approche comportementale la rend inapproprié pour des IDSs commerciaux.

— La vitesse de traitement des audits est un facteur très important c'est la raison pour laquelle les signatures sont utilisées à la place des règles.

2.3.2 Le comportement d'un système de détection d'intrusion

Le comportement d'un système de détection d'intrusions décrit sa réponse, qui peut être passive ou active [86].

Réponse passive

Dans ce cas, aucune contre-mesure ne sera appliquer activement pour arrêter ou limiter les dégâts d'une attaque, tout simplement des informations nécessaires sont fournies par l'IDS aux administrateurs réseau et aux responsables de la sécurité pour aider à prendre des mesures basées sur ces informations tel que (les réactions ou alarmes sont générées par affichage d'un message d'alerte concernant des informations détaillées de l'intrusion sur l'écran du responsable de la sécurité réseau, générer un son ou alarme particulière, envoyer des emails, archiver dans un fichier ou base de données permettant aux analystes de faire des analyses avancées et de faire une corrélation avec l'historique des évènements produits auparavant, envoyer des rapports aux systèmes de gestion de réseaux (Network Management System) utilisant le protocole SNMP (Sample Network Management Protocol) dédié à la gestion du réseau, etc.).

Réponse active

Dans ce cas, des actions correctives, ou proactives sont engendrées (Changer l'environnement tel que les règles de filtrage d'un Firewall des connexions TCP, Génération de scripts qui peuvent supprimer la vulnérabilité par changement des permissions sur un système de fichiers par exemple, Restaurer le système à son état antérieur, ou bien attaquer l'intrus par ces propres moyens ou d'autres moyens plus dévastateurs dans le but d'obtenir plus d'informations sur l'attaquant et son emplacement ou bien de l'anéantir mais ceci nous ramène à la question suivante : « est ce que notre réponse active peut être illégale ? » chose qu'il faut prendre avec prudence, etc.).

Avec l'arrivée de nouveaux produits de détection d'intrusion, l'élément de contre-mesure est devenu de plus en plus prépondérant où plusieurs IDSs inclurent [29] [83] :

— la capacité de couper les connexions qui transportent les attaques,

— bloquant la connexion des hôtes à partir desquels les attaques proviennent,

— ou la reconfiguration des autres équipements tels que les pare-feux ou les routeurs. Grâce à des stratégies de sécurité proactives.

2.3.3 Les types de système de détection d'intrusion

Les systèmes de détection d'intrusions sont regroupés selon leurs sources d'informations (sondes). Il y'a ceux qui sont basés sur les informations sur le réseau, ceux basés sur les informations produites par le système d'exploitation et d'autres par les informations des applications [9] [86].

Les systèmes de détection d'intrusion à base de données système : HIDS (Host-Based System)

Les premiers IDSs qui ont vu le jour sont ceux basés système, vu que les interactions avec l'extérieur du système ont été vraiment rares, ce qui a rendu la tâche du système de détection d'intrusion plus facile. Le système de détection d'intrusion analyse les informations d'audit fourni par l'ordinateur central, soit localement ou sur une machine séparée, et signale tous les événements suspects. Ce type d'IDS, analyse l'information concernant uniquement cet hôte, donc il est plus précis sur le type d'attaques. Cette source de données basé système est également vulnérable à des altérations dans le cas d'une attaque réussie, ce qui crée une contrainte importante amenant le système de détection d'intrusion à analyser l'audite de sécurité en temps réel et générer des alarmes avant que l'attaquant essaye de modifier l'audite ou arrêter le système de détection d'intrusion [86]. Parmi les données d'audit réseau Nous citons :

Les données statistiques relatives aux systèmes Parmi les sources d'information qui reflètent le comportement du système, on cite Les données statistiques relatives aux systèmes. Elles fournissent des renseignements sur la consommation des ressources (le temps processeur, mémoire, l'utilisation du disque ou du réseau, les applications lancées...etc.) partagées par les utilisateurs du système. Ceci a conduit certains concepteurs des prototypes de détection d'intrusion d'utiliser ces statistiques comme source d'audit. Mais vu que ces statistiques relatives aux systèmes possèdent un certain nombre d'inconvénients qui les rend peu fiables comme source d'audit. Par exemple, le manque de paramétrage, absence

d'identification précise de commande, le retard pour obtenir des informations, ils n'ont jamais été utilisées dans l'approche de détection par scénario, et rarement utilisées pour la détection d'intrusion basée sur le comportement [29].

Les commandes systèmes On peut obtenir l'état du système au moment de son exécution par des primitives du système d'exploitation. Comme le cas sur Unix, on a les primitifs ps, pstat, vmstat, getrlimit. Ces dernières fournissent des informations précises et ciblées sur les événements, car ils examinent directement la mémoire noyau du système. Par contre, il est très difficile d'utiliser ces commandes pour une collecte d'information continue, car ils n'offrent pas des données bien structurées [29].

Le service d'audit Syslog Toute application reçoit du système d'exploitation un service d'audit appelé syslog, qui contient une chaîne de caractère spécifiant le temps et le nom du système sur lequel s'exécute l'application. L'utilisation facile du syslog a incité les développeurs d'applications à l'utiliser comme piste de vérification. Un certain nombre d'applications et de services réseau utilisent ce service, tels que login, sendmail, nfs, http, et cela inclut également des outils liés à la sécurité tels que sudo, klaxon ou TCP wrappers. Par conséquent, des outils de détection d'intrusion qui utilisent les informations fournies par le Syslog ont été développés, par exemple Swatch. Vu la quantité d'information d'audit minime et nécessaire à la sécurité, que le Syslog génère, le rend peu fiable dans le développement des IDSs [29].

Les traces d'audit de sécurité C2 (Commande et contrôle) Le principe de base de toutes les traces d'audit de sécurité est d'enregistrer le passage des instructions exécutées par le processeur dans l'espace utilisateur et les instructions exécutées dans l'espace « Trusted Computing Base » (TCB) considéré comme fiable, et que les actions dans l'espace utilisateur ne peuvent pas nuire la sécurité du système. Ce model est la principale source d'informations d'audit pour la majorité des prototypes liés à la détection d'intrusion basés système, vu leur fiabilité pour recueillir des informations détaillées sur les actions prises sur un système d'information.

Beaucoup de travaux ont été menés par plusieurs groupes de recherche pour définir les informations qui devraient être figurées dans le journal de trace d'audit de sécurité ainsi que le format commun de ces fichiers [29].

Les avantages et inconvénients des HIDS Ce type de système de détection d'intrusions présente des avantages et des inconvénients que nous citerons comme suit :

Avantages :

— Une meilleur réaction, vu la possibilité d'un constat immédiat d'une attaque.

— Une meilleure optimisation du système, l'observation avec précision de l'activité sur l'hôte.

— Contrairement aux systèmes basés réseau (NIDS), il détecte facilement les attaque de type « cheval de Troie ».

— Les attaques faisant partie du trafic réseau crypté et impossible à détecter par les NIDS, sont détectées par le HIDS.

Inconvénients :

— L'HIDS peut être facilement reconnu et mis hors d'état de service par un attaquant.

— Il est sensible aux attaques Déni de service.

— Aucune alerte n'est donnée si les entrées des journaux d'évènement ne correspondent pas aux signatures ou des règles prédéfinies.

— Il est gourmand en CPU et peuvent modifier les performances d'un hôte.

Les systèmes de détection d'intrusion à base de données réseau : NIDS (Network-Based IDS)

Dès l'apparition des réseaux informatiques, de nouveaux types d'IDSs ont vu le jour pour répondre à ces nouvelles exigences des systèmes distribués, car les attaquants lancent leurs attaques sur plusieurs systèmes en passant d'une machine à l'autre en changeant éventuellement leurs identités lors de leurs déplacements. Il est donc primordial que le NIDS local doit disposer d'un échange d'informations (enregistrements d'audit brut sur le réseau) avec ses égaux.

Comme, ils doivent faire face à des attaques contre le réseau lui-même. Leurs rôle est d'analyser et d'interpréter les paquets circulant sur le réseau en temps réel et rechercher les attaques réseau de type (DNS Spoofing, TCP détournement, balayage des ports, etc.) qui ne sont pas détectés par l'analyse d'audit de sécurité. Ces outils sont souvent intéressants pour les administrateurs système et sont installés à des endroits stratégiques du réseau à laide de capteurs en mode furtif (stealth mode), et donc invisibles aux autres machines, qui génèrent des alertes en cas de détection d'attaque, qui sont envoyées à leur tour à une console sécurisée située sur un réseau isolé , qui relie uniquement les capteurs et console pour une analyse et un éventuel traitement [29] [86].

Parmi les données d'audit réseau Nous citons :

Les paquets réseau La plupart des accès aux ordinateurs sensibles se font via les réseaux informatiques, il faut donc capturer les paquets avant qu'ils entrent au serveur pour mieux les contrôler. Les paquets réseau est une source de donnée très importante représentant les informations relatives aux évènements qui se produisent sur le réseau. L'analyse des paquets représente le moyen le plus efficace pour détecter les attaques de type Déni de Service (DoS) qui provient dans la plupart du temps du réseau et qui ne sont pas détecter par les HIDS [86].

Les informations SNMP La gestion d'un réseau se fait par le biais du répertoire d'information « Simple Network Management Protocol (SNMP) », qui contient les informations de configuration (table de routage, adresses, noms) ainsi que les informations liées à la performance du réseau et les conteurs qui mesurent le trafic sur les différentes interfaces réseau et aux différentes couches. Le SNMP représente une intéressante source d'audit pour les systèmes de détection d'intrusion [29].

Les avantages et inconvénients des NIDS **Avantages**

— Un intrus ne sait jamais qu'il est contrôlé du fait que le système est totalement invisible sur le réseau.

— Il est capable de capturer tous les paquets envoyés à la cible.

Inconvénients

— Aucune alarme n'est donnée que s'il y'a correspondance avec les règles et signatures préconfigurées.

— Aucune certitude en cas d'attaque réussie.

— Le trafic chiffré n'est pas examiné.

— Pour que les NIDS voient tout le trafic, il faut des configurations spéciales sur les réseaux commutés.

Les systèmes de détection d'intrusion à base de données d'application

Vu l'augmentation de l'utilisation des serveurs d'application, les IDSs basés sur les applications on vu le jour et qui ne sont qu'un sous groupe des HIDS. Par conséquent, les fichiers logs des applications sont devenus une source d'information pour ces IDS, par les informations qu'ile contiennent sur les activités d'une application particulière. Il se situe donc au niveau de la communication entre un utilisateur et l'application surveillée [86]. La comparaison entre cette source de données et les traces d'audit de sécurité C2 ou les paquets réseau montre trois avantages [29].

— du fait que l'information est directement reçue du log ce qui prouve sa précision contrairement aux traces d'audite de sécurité C2 ou les paquets réseau qui nécessitent un traitement avant que le système de détection d'intrusion arrive à comprendre quelle information est actuellement reçue par l'application.

— Des informations pertinentes et complètes sont contenues dans le journal des applications et qui ne se présentent ni dans les audits de sécurité ni dans les paquets réseau.

— Du fait que la sélection des informations pertinentes pour la sécurité est laissée à la charge de l'application, ce qui réduit la surcharge induite par ce mécanisme de collecte par rapport à l'audit de sécurité C2 et donc améliore la performance.

— Il est possible de détecter et empêcher les commandes particulières dont l'utilisateur pourrait s'en servir avec le programme.

— Il est possible de surveiller chaque transaction entre utilisateur et application.

— Pour ce système, les données sont décodées dans un contexte connu ce qui rend son analyse plus fine et précise.

Né au moins il existe quelques inconvénients pour l'utilisation des fichiers logs d'application pour la détection d'intrusion [29] :

— L'absence d'informations requises par ce système de détection d'intrusion, car Les attaques ne sont détectées que lorsque le log d'application est écrit. Et si l'attaque arrive à empêcher l'écriture dans le log de l'application (ce qui est le cas dans des nombreuses attaques par déni de service), elle causera des dégâts.

— Certaines attaques comme les attaques par déni de service visent les niveaux bas du système, tels que les pilotes de réseau. Comme ces attaques n'exécutent pas du code d'application, ils ne peuvent pas être visibles dans les logs d'application.

— Ils présentent une sécurité plus faible du fait qu'ils n'agissent pas au niveau du noyau. Comme c'est le cas des attaques de types « cheval de Troie ».

Ce type d'IDS sera utile pour la surveillance d'une application très sensible, mais dans le cas ou il est associé à un HIDS et surtout il faudra contrôler le taux d'utilisation CPU des IDS pour ne pas compromettre les performances de la machine.

Les IDSs hybrides

Les approches hybrides ont également été développées. Elles rassemblent généralement les caractéristiques des systèmes de détection d'intrusions basés réseau (NIDS) et système (HIDS). Ils permettent en un seul outil de surveiller le réseau et l'hôte en même temps [29].

2.4 L'efficacité des IDSs

Les paramètres qui permettent d'évaluer l'efficacité des systèmes de détection d'intrusion sont cités ci-dessous [29].

— **La précision :** On dit qu'un système de détection d'intrusion est précis, s'il détecte les attaques sans faire des fausses alarmes (c'est-à-dire une action légitime est déclarée comme anormale ou intrusive).

— **La performance de traitement :** Elle est mesurée par la vitesse avec laquelle les événements d'audit sont traités. Si la performance de traitement de l'IDS est faible, alors la détection en temps réel est impossible.

— **La complétude :** C'est capacité d'un IDS de détecter toutes les attaques. Elle est beaucoup plus difficile à évaluer par rapport aux autres mesures, car il est impossible d'avoir une connaissance globale sur les attaques ou les abus des privilèges.

— **La tolérance aux pannes :** Un IDS devrait conçu avec l'objectif d'être résistant aux attaques en particulier les attaques de déni de service. Ceci est particulièrement important car les IDSs s'exécutent dans des systèmes d'exploitation ou des matériels qui sont connus pour être vulnérables aux attaques.

— **La rapidité :** la réaction du responsable de sécurité dépend de la rapidité de l'exécution et propagation de son analyse, afin de minimiser les dégâts possibles, et aussi pour empêcher l'attaquant d'altérer la source de vérification ou interrompre le fonctionnement du système de détection d'intrusion.

2.5 Les techniques de détection d'intrusion

De nombreux travaux concernant la construction des IDSs performant et précis ont vu le jour, Depuis que Denning a publié dans son article le premier concept d'un modèle de détection d'intrusion [30]. Cette première génération se basait sur les connaissances des experts de sécurité où les méthodes statistiques et les approches de l'intelligence artificielle pour construire les noyaux (moteurs) des modèles. Face à des problèmes tels que le grand volume du trafic réseau, la distribution des données très déséquilibrée, la difficulté de prendre une décision entre le comportement normal et anormal et l'adaptation permanente pour des environnements en constante évolution, les techniques de l'intelligence artificielle ont montré beaucoup de limites. Pour remédier à ces problèmes on s'est penché vers les techniques de data mining et qui Grâce à ces dernières, des modèles de détection d'intrusion dite de deuxième génération, plus rapides et plus précis ont été développés [122].

2.5.1 Les systèmes de détection d'intrusion dite de première génération

Les techniques ou les méthodes utilisées dans la première génération des systèmes de détection d'intrusion dépendent de l'approche adoptée qui peut être comportementale ou par scénario.

Les méthodes utilisées dans l'approche comportementale

Plusieurs méthodes sont utilisées pour implémenter l'approche comportementale qui consiste comme nous l'avons décrit précédemment comme étant la déviation par rapport au comportement normal du système. Parmi les principales méthodes utilisées on cite : la méthode statistique, les systèmes experts, les réseaux de neurones, l'immunologie [29].

La méthode statistique C'est la méthode la plus utilisée dans l'approche comportementale. Elle consiste à mesurer le comportement de l'utilisateur ou du système par un nombre de variables échantillonnées dans le temps tel que le temps de connexion et de déconnexion de chaque session, l'utilisation de la mémoire, l'occupation du processeur, l'accès aux fichiers, etc. Le modèle de base conserve les moyennes de toutes ces variables, puis il les compare avec les valeurs des variables, où il détecte si les seuils sont dépassés. Ce modèle de base ne pouvait représenter de façon crédible les données vu sa simplicité, c'est la raison pour laquelle on a développé un modèle plus compliqué qui compare les profils des activités des utilisateurs à long terme et à court terme. Une mise à jour régulière de ces profils s'impose chaque fois que le comportement des utilisateurs change. Ce modèle statistique est utilisé dans de nombreux prototypes de systèmes de détection d'intrusion [66] [58] [57] .

Le système expert Un système expert est un programme conçu pour simuler le comportement d'un humain qui est un spécialiste ou un expert dans un domaine très restreint. Il se compose d'une base des faits, une base des règles et un moteur d'inférence. Dans ce cas de système, la base des règles peut être conçue de deux façons [30] :

— La première s'appuie sur un ensemble de règles qui décrivent statistiquement le comportement des utilisateurs, où elle utilise les enregistrements de leurs activités sur une période de temps donnée. Ensuite l'activité courante est comparée à ces règles afin de détecter un comportement inco-

hérent. La base des règles est mis à jour régulièrement pour tenir compte des nouvelles utilisations. Wisdom et Sense sont des IDSs basés sur ce mode de fonctionnement [118].

— Dans la deuxième approche, ou les actions des utilisateurs sont vérifiés en fonction d'un ensemble de règles qui décrivent la politique du bon usage, et on signale toute action qui ne correspond pas aux modèles acceptables. AT& T's Computer Watch est un IDS basé sur ce mode de fonctionnement [35].

Cette approche est utile pour des profils d'utilisation fondés sur les politiques de sécurité, mais elle est moins efficace que l'approche statistique pour le traitement des grandes quantités d'informations d'audit.

Il représente l'une des méthodes les plus utilisés dans les systèmes de détection d'intrusion basés scénario [42], surtout pour les IDS commerciaux. Son principe est le suivant :

— Le système expert contient un ensemble de règles qui décrivent les attaques.

— Les événements de l'audit de sécurité sont ensuite traduits en des faits portant leurs significations sémantiques pour le système expert.

— Le moteur d'inférence tire des conclusions à l'aide de ces règles et des faits.

Cette approche nous permet de naviguer systématiquement dans les traces d'audit à la recherche des preuves d'une tentative d'exploitation des vulnérabilités connues.

Les réseaux de neurones C'est est un modèle de calcul inspiré du mode de fonctionnement des neurones biologiques. Il est utilisé pour instruire la relation entre deux ensembles d'informations, puis généraliser cette relation. Dans la détection d'intrusion, ce modèle a été principalement utilisé pour apprendre le comportement des utilisateurs du système.

Certaines concordances entre les modèles des réseaux de neurones et statistiques ont été illustrées, mais l'avantage d'utiliser des réseaux de neurones par rapport à l'approche statistique réside dans le moyen simple d'exprimer des relations non linéaires entre les variables, et dans le fait que l'apprentissage / réapprentissage du réseau de neurones est automatique [47] [104].

Des expériences pour la prédiction du comportement des utilisateurs ont été effectuées en utilisant un réseau de neurones, ont montré que le comportement de la plupart des utilisateurs est prévisible, et qu'il y a qu'une très petite partie des utilisateurs dont le comportement est imprévisible. Les réseaux de neurones ne sont pas très utilisés par la communauté de détection d'intrusion [28].

L'immunologie Un système immunitaire artificiel (SIA) est inspiré des principes du fonctionnement du système immunitaire naturel qui apprend et mémorise pour résoudre les problèmes. Son principe est de construire un modèle de comportement normal des services réseau UNIX, plutôt que le comportement des utilisateurs.

Ce modèle utilise des courtes séquences d'appels système effectués par les processus. L'outil rassemble tout d'abord un ensemble de traces d'audit de référence qui représente le comportement approprié du service, puis extrait un tableau de référence contenant toutes les bonnes séquences connues des appels système. Ces modèles (Pattern) sont ensuite utilisés pour la surveillance en temps réel pour vérifier si les séquences générées sont répertoriées dans le tableau, sinon, l'IDS génère une alarme. Cette technique a un taux de fausses alarmes potentiellement très faible si la table de référence est suffisamment exhaustive.

L'inconvénient de cette méthode réside dans ses faiblesses pour les erreurs de configuration dans un service, c'est-à-dire quand les attaques utilisent des mesures légitimes prises par le service pour accéder sans autorisation [44].

Les méthodes utilisées dans l'approche par scénario

Elle se base sur l'utilisation des connaissances regroupées sur des attaques spécifiques et les vulnérabilités du système. Les principales méthodes utilisées pour cette approche sont : les systèmes experts, l'analyse de la signature, les réseaux de Pétri, l'analyse de l'état transition et les algorithmes génétique, etc.

Le système expert Il représente l'une des méthodes les plus utilisés dans les systèmes de détection d'intrusion basés scénario [42], surtout pour les IDS commerciaux. Son principe est le suivant :

— Le système expert contient un ensemble de règles qui décrivent les attaques.

— Les événements de l'audit de sécurité sont ensuite traduits en des faits portant leurs significations sémantiques pour le système expert.

— Le moteur d'inférence tire des conclusions à l'aide de ces règles et des faits.

Cette approche nous permet de naviguer systématiquement dans les traces d'audit à la recherche des preuves d'une tentative d'exploitation des vulnérabilités connues.

L'analyse de la signature L'analyse de la signature suit le même principe d'acquisition des connaissances du système expert, mais les connaissances sont exploitées d'une manière différente. Par exemple, des scénarios d'attaque pourraient être traduits en séquences d'événements d'audit qu'ils génèrent ou en des modèles de données qui peuvent être recherchés dans la trace d'audit généré par le système.Cette technique permet une mise en œuvre très efficace et applicable pour les produits commerciaux de détection d'intrusion [73], [114].

Son inconvénient c'est :

— de faire des mises à jour fréquentes comme toutes les approches basées sur l'approche par scénario.

— l'obligation de représenter toutes les facettes possibles des attaques par des signatures.

— Ce qui nous conduit à représenter une attaque par un certain nombre de signatures, au moins une pour chaque système d'exploitation pour que le système de détection d'intrusion devient portable.

Les réseaux de Pétri Ils ont été utilisés pour représenter les signatures des intrusions. Par exemple, IDIOT est un IDS qui utilise un réseau de Pétri coloré développé par l'université de Purdue. Ses avantages résident dans :

— leurs capacités de généralisation,

— leurs simplicités conceptuelles,

— et leurs représentations graphiques.

En raison de la grande capacité de généralisation du réseau de Pétri coloré, même les signatures complexes peuvent facilement être écrites. Néanmoins, la comparaison entre cette signature complexe et les traces d'audit peut devenir très coûteuse en termes de calcul [72].

Les algorithmes génétiques Les algorithmes génétiques simulent la théorie de darwinienne pour le processus de l'évolution naturelle, dont le but est de trouver une solution proche de la solution optimale d'un problème donnée [61]. Leur utilisation dans l'approche par scénario est de trouver les signatures d'attaques prédéfinies dans les traces d'audit de sécurité, par la traduction du problème de la recherche de la signature des attaques dans l'audit de sécurité en un problème d'optimisation, où l'on recherche la meilleure solution qui maximise le nombre des attaques détectées toute en respectant la contrainte de la correspondance entre le nombre des événements de chaque signature d'attaque et le nombre des événements audités [78].

2.5.2 Les systèmes de détection d'intrusion de deuxième génération

Pour contrer, le nombre très élevé des personnes malintentionnées voulant accéder au volume important de données des cybers infrastructures, et aux systèmes par les différents moyens de piratage qui ne cesse d'évoluer, les chercheurs en sécurité informatique ont utilisé les différentes techniques d'apprentissage automatique, de statistique, et de data mining, afin de relever les défis de la cyber sécurité. L'utilisation des techniques de data mining pour la détection d'intrusion est passé par différents chemins où on trouve l'utilisation des classificateurs simple, hybride ect... [117].

Conclusion

Nous avons présenté dans ce chapitre une idée générale sur les systèmes de détection d'intrusion. Il est clair que ces systèmes sont plus qu'indispensable à toute entreprise en ce qui concerne sa sécurité informatique.

Les systèmes de détections d'intrusion représentent une seconde ligne de défense contre les attaques qui peuvent contourner les mécanismes de sécurité classique comme les pare-feux, l'authentification, le proxy...etc. Ce mécanisme représente un atout dans la guerre contre la cybercriminalité.

A l'heure actuelle ces systèmes sont loin d'être infaillibles, malgré leur apport complémentaire en la sécurité du système d'information, car ils présentent quelques imperfections. La plus part des IDSs sont construits dans une architecture hiérarchique dont la masse de données transférées à travers le réseau peut congestionner ce dernier et sont même susceptibles d'être attaqués.

Comme nous l'avons décrit dans ce chapitre, le développement des IDSs est passé par deux générations. La première ou les méthodes statistiques et les techniques de l'IA ont été utilisées, et qui ont monté des limites importantes comme la performance, l'intervention des experts humains,... etc. une seconde génération d'IDSs a vu le jour pour combler les imperfections de la première, dans la quelle les techniques de Data mining ont été utilisés. Ce qui nous a amené à investiguer dans le domaine du data mining dans le chapitre suivant.

Chapitre 3

Recherche d'information et Data Mining

Sommaire

Introduction du sous chapitre 1 : « Recherche d'Information »

Historiquement, la croissance du volume de données textuelles comme les livres et les articles dans les bibliothèques durant des siècles a imposé de définir des mécanismes efficaces pour les localiser. D'où la naissance de la recherche d'information qui est un ancien domaine datant des années 1940 [70] [89]. Une des premières définitions de la Recherche d'Information (RI) a été a été introduite par Calvin Mooers en 1950 ou il définit la recherche d'information comme une discipline informatique traitant le problème d'accès à l'information pertinente dans une masse de données importante [85].

Au début on s'intéressait à la recherche documentaire par contre maintenant on s'intéresse à d'autres tâches tel que : le filtrage d'information, l'extraction d'information, la recherche d'information multilingue, la recherche d'information sur le web, les questions réponses, etc. [17] L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI) qui est un mécanisme de gestion qui joue l'intermédiaire entre un utilisateur et une collection d'informations dans le but de lui satisfaire le besoin en information [36].

Ce chapitre est organisé en cinq grandes parties comme suit :

La première présente la définition, l'historique et les concepts de base de la recherche d'information ainsi que les domaines d'application. La deuxième partie expose les modèles de recherche d'information existants qui ont été proposés pour fournir un cadre théorique pour la modélisation du processus de recherche d'information.

La troisième partie décrit en détails le processus de recherche d'information, à savoir les étapes d'indexation, d'interrogation et de mise en correspondance, ainsi que les techniques de reformulation des requêtes.

La quatrième partie sera consacrée à la recherche d'information sur le web en présentant la différence entre la recherche d'information classique et la recherche sur WEB, la source d'information, les outils de recherche d'informations sur le web : notamment les algorithmes des moteurs de recherche d'information et l'architecture générale des moteurs de recherche.

Enfin en cinquième partie, nous nous achevons le chapitre par la description du mécanisme et les techniques d'évaluation utilisées pour évaluer les performances d'un Système de Recherche d'Information ainsi que la présentation des différentes mesures d'évaluation de la pertinence.

3.1 Définitions, histoire et concepts de base de la Recherche d'Information

3.1.1 Définitions

Plusieurs définitions de la recherche d'information ont été introduites depuis son apparition en 1940. Nous citons quatre définitions que nous jugeons globales et qui se complètent entre elles :

Définition 1 « généraliste » : *"la Recherche d'information est une discipline informatique traitant le problème d'accès à l'information pertinente dans une masse de données importante"*[85].

Définition 2 « selon le but » : *"La Recherche d'Information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin d'information"*[59] [89].

Définition 3 : « fonctionnelle » : *"La Recherche d'Information est un ensemble d'opérations, méthodes et procédures qui permettent de trouver à partir d'une collection de documents de volume important, l'information pouvant répondre à une question sur un sujet précis"*[85] [59].

Définition 4 : *" la Recherche d'Information est un domaine qui a pour objectif, la représentation, l'analyse, l'organisation, le stockage et l'accès à l'information "* [52] [89].

Par ces définitions, on comprend que le défit principale de la recherche d'information est de pouvoir, trouver les documents qui répondent à l'attente de l'utilisateur parmi le volume important de documents disponibles en appliquant un processus de recherche en vue d'exploiter l'information contenue dans ces documents (son ,texte, image) par rapport à une requête formulée par l'utilisateur, en retournant le moins possible de documents non pertinent sachant que les contenus des documents peuvent être non structurés ou semi structurés [89].

Le Web qui est la source d'information numéro un dans le monde représente un espace de recherche ayant une taille importante diminuant ainsi la possibilité de retrouver l'information désirée facilement.

Même si la recherche dans le domaine de la Recherche d'Information sur le Web a évoluée en apportant de grandes rénovations aux techniques et modèles, l'augmentation exponentielle de la quantité d'information diverses et hétérogènes reste un problème qui incite constamment à de nouvelles recherches. Autrement dit, le web a remis la recherche d'information face à de nouveaux défis : de retrouver une information pertinente dans un espace hétérogène et de taille considérable.

3.1.2 Historique

La méconnaissance de l'histoire du domaine de la recherche d'information (RI) est une des causes de confusion des concepts et définition de ce domaine. La recherche d'information a une longue histoire de plus de 50 ans. Dans ce qui suit, nous allons donner très brièvement l'histoire et les principaux évènements de la recherche d'information comme suite :

— 1950 : après la deuxième guerre mondiale, le monde assiste à la naissance de recherche d'information avec le début de petites expérimentations en utilisant des petites collections de documents.

— 1960-1970 : expérimentations plus larges ont été menées, ou on a développé une méthodologie d'évaluation du système qui est aussi utilisée maintenant dans d'autres domaines (des corpus de test ont été conçus pour évaluer des systèmes différents). L'utilisation du modèle booléen.

— 1970 : développement du système SMART dirigé par G. Salton[100], avec l'introduction du modèle vectoriel et la technique de relevance feedback. Avec l'utilisation du modèle probabiliste.

— 1980 : l'intelligence artificielle influe sur les travaux consacrés à la Recherche d'Information[60] ;

— 1990 : Internet à propulser la RI en avant de la scène avec beaucoup d'applications avec modification de la RI. Cependant, les techniques de base utilisées dans les moteurs de recherche sur le web restent identiques. [87].

3.1.3 Les types de tâche de la Recherche d'Information

Le tableau ci-dessous (Voir Tableau 3.1) présente quelques types de tâches qui peuvent être considérer dans le domaine de la recherche d'information.

TABLEAU 3.1 – Description des tâches de la Recherche d'Information

Tâche	**Description**
Ad hoc Recherche	Retrouver les documents pertinents dans une collection fixe
Question / Réponse	Extraire des réponses dans les documents récupérés
Diffusion sélective d'information	Contrôlez un flot de documents correspondant à un profil
Classification de documents	Regroupement automatique de documents
Catégorisation de documents	Affecter un document à une catégorie prédéfinie

3.1.4 Concepts de base

Les définitions proposées pour la Recherche d'Information ainsi que son histoire nous mènent à définir plusieurs concepts clés que nous jugeons utile d'expliquer.

Système de Recherche d'Information : un système de recherche d'information est défini par un langage de représentation des documents et des requêtes qui expriment un besoin de l'utilisateur, et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information [17].

Document : le document constitue l'information élémentaire d'une collection de documents, appelée aussi granule de document qui peut représenter tout ou une partie d'un document. C'est l'élément essentiel dans un système de recherche d'information ou il est considéré comme un support physique de l'information pouvant être du texte, une page web ou du multimédia [70].

Dans le cas d'un document texte on peut le représenter selon deux vues [46] :

— La vue sémantique : elle se concentre sur l'information véhiculée dans le document.

— La vue logique : elle définit la structure logique du document.

Collection de documents : (ou corpus ou fond documentaire) elle est constituée d'un ensemble de documents constituant l'ensemble des informations exploitables et accessibles.

Besoin d'information : la notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Deux types de besoin utilisateur ont été définis [63] :

— *Besoin vérificatif :* l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.

— *Besoin thématique :* l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connus ou inconnu pour l'utilisateur. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble.

Requête : la requête constitue l'expression du besoin en information de l'utilisateur selon le formalisme d'interrogation d'un système de recherche d'information. Elle représente l'interface entre le système de recherche d'information et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique [18].

Modèle de représentation : un modèle de représentation est un processus permettant d'extraire d'un document ou d'une requête une représentation qui couvre au mieux son contenu sémantique [18].

Modèle de recherche : il représente le modèle du noyau d'un système de recherche d'information. Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer [18].

Pertinence : est une notion cruciale qui mesure le degré de ressemblance entre la requête et les documents renvoyés en se référant aux deux concepts : bruit et silence. Tel que, le silence correspond aux documents pertinents qui n'apparaissent pas dans le résultat de la recherche, alors que le bruit correspond aux documents ramenés en réponse, mais qui ne sont pas pertinents par rapport à la question posée [20].

3.2 Modèles de Recherche d'Information

Le domaine de la RI ne cesse d'évoluer, ainsi que l'exigence des utilisateurs en matière de qualité de réponse des systèmes de recherche d'information. Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information et un cadre théorique pour la modélisation de la mesure de pertinence autrement dit une interprétation théorique de la notion de pertinence. Il permet aussi de créer une représentation interne d'un document ou d'une requête, afin de pouvoir créer une correspondance entre les deux [18].

Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requête-document. Le vocabulaire d'indexation V = $t_i, i \in \{1, ..., n\}$ est constitué de n termes qui apparaissent dans les documents. Selon Baeza-Yates[8], un modèle de Recherche d'Information est défini par un quadruplet $(D, Q, F, R(q, d))$:
où

— D est l'ensemble de documents ;

— Q est l'ensemble de requêtes ;

— F est le schéma du modèle théorique de représentation des documents et des requêtes ;

— $R(q, d)$ est la fonction de pertinence du document d à la requête q.

Dans la littérature, on trouve un grand nombre de modèles de RI qui ont été développé. Mais dans ce qui suit nous pouvons distinguer les trois grandes familles de modèles dans la RI : la famille des modèles ensemblistes, la famille des modèles algébriques, la famille des modèles probabilistes.

Actuellement, la notion de Recherche d'Information sémantique devient un sujet incontournable car elle doit permettre d'obtenir un meilleur résultat dans la mesure où elle s'intéresse davantage à la compréhension du message véhiculé dans les documents et dans la requête.

3.2.1 Classe des modèles ensemblistes

Cette classe englobe deux types de modèles à savoir : le modèle booléen et le modèle flou. Tous les deux se basent sur la théorie des ensembles.

Modèle booléen

Les premiers systèmes de recherche d'information développés sont basé sur le modèle booléen, même aujourd'hui beaucoup de systèmes commerciaux notamment les moteurs de recherche utilisent le modèle booléen, cela est du à la simplicité et à la rapidité de sa mise en œuvre. Ce modèle a été proposé par Salton en 1971 [100], représente le premier modèle utilisé en recherche d'information. Il est basé sur la théorie des ensembles et sur l'algèbre de Boole. Dans ce modèle, chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document. La requête est représentée sous forme d'une équation logique. Les termes sont reliés par des connecteurs logiques ET, OU et NON. Le moteur de recherche retrouve les documents qui correspondent exactement à la requête, compte tenu de la présence ou de l'absence des termes de celle-ci dans la représentation des documents et de l'expression booléenne.

Ainsi, la correspondance entre un document d et une requête q est définie par :

$$\mathrm{R}(q,d) = \begin{cases} 1 & si\ d\ appartient\ \text{à}\ l'ensemble\ décrit\ par\ q \\ 0 \end{cases} \tag{3.1}$$

Cependant les principaux inconvénients de ce modèle sont [19] :

— les formules de requêtes sont complexes, non accessibles à un large public. Par conséquence l'utilisateur a du mal à formuler de bonnes requêtes.

— La représentation binaire est peu informative, elle ne donne pas de renseignement ni sur la fréquence du terme dans le document ni sur la longueur de document, notant que ces renseignement sont très importante pour la recherche d'information.

— la réponse du système dépend de l'ordre de traitement des opérateurs de la requête, autrement dit, les documents retournés ne sont pas ordonnée selon leur pertinence.

— Le modèle est inadapté à une recherche progressive : il ne supporte pas la réinjection de pertinence.

Une extension de ce modèle a été effectuée par Salton et al. [102], puis amélioré par Cater et Kraft [19], Cooper [26] pour remédier à quelques faiblesses du modèle booléen.

Modèle flou (modèle booléen étendu)

Pour remédier aux inconvénients du modèle précédent, Salton et al [102]a proposé une extension du modèle booléen où Il s'appuie sur la théorie des ensembles flous proposé par Zadeh en 1965 [123] [45]. C'est un modèle booléen étendu qui complète le modèle de base en intégrant des poids dans l'expression de la requête et des documents. Ceci induit la sélection de documents sur la base d'un appariement rapproché par une fonction d'ordre. Dans ce modèle, le poids affecté à un terme reflète la mesure dans laquelle ce terme décrit le contenu du document où il apparaît. L'intégration de cette théorie dans la RI a permis de traiter l'ambigüité des requêtes, l'imprécision qui caractérise le processus d'indexation ainsi que la divergence de pertinence entre les documents résultats. Ce modèle a été amélioré ensuite par Cater et Kraft [19] et Cooper [26].

La réponse du système de recherche d'information dépend de l'ordre de traitement des opérateurs de la requête, autrement dit, les documents retournés ne sont pas ordonnée selon leur pertinence, et demeure l'inconvénient principale du modèle flou.

3.2.2 Classe des modèles algébriques

Cette classe est représentée par deux modèles importants en l'occurrence le modèle vectoriel et le modèle d'indexation sémantique latente, qui sont fondés sur l'algèbre.

Modèle vectoriel

Ce modèle est issu des travaux de Salton en 1971 [99] dans le domaine de l'Information Retrieval. Le modelé vectoriel a été implémenté la première fois dans le cadre du système SMART, il a un retentissement majeur, puisqu'il inspire la plupart des moteurs de recherche par index sur Internet (comme Alta Vista).

Le modèle vectoriel est une représentation mathématique du contenu d'un document et des requêtes, selon une approche algébrique sous forme des vecteurs dans un espace multidimensionnel, où chaque dimension correspond à un terme d'indexation. L'ensemble de représentation des documents est un vocabulaire comprenant des termes d'indexation. Éventuellement ils peuvent être des constructions plus élaborées comme des expressions ou des entités sémantiques. À chaque élément du vocabulaire est associé un indexe unique arbitraire.

Chaque document d est ainsi représenté par un vecteur v, dont la dimension correspond à la taille du vocabulaire. Chaque élément v_i du vecteur v consiste en un poids associé au terme d'indice i dans le document d. Un exemple simple est d'identifier v_i au nombre d'occurrences du terme i dans l'échantillon de texte. La composante du vecteur représente donc le poids du terme i dans le document.

Formellement, un document d_i est représenté par un vecteur v de dimension n.

$d_i = (w_{i1}, w_{i2}, ..., w_{in})$ pour $i = 1, 2, ..., n$. Où w_{ij} est le poids du terme t_j dans le document d_i, n est le nombre de termes d'indexation.

Une requête Q est aussi représentée par un vecteur de mots-clés défini dans le même espace vectoriel que le document. $Q = (w_{Q1}, w_{Q2}, ..., w_{Qn})$ Où w_{Qj} est le poids de terme t_j dans la requête Q. La pertinence du document d_i pour la requête Q est mesurée comme le degré de corrélation des vecteurs correspondants.

La figure suivante (Voir Figure 3.1) illustre la représentation des documents dans l'espace d'indexation. Dans cette figure les t_i sont les termes, les d_j sont les documents et les w_{ij} sont les pondérations des descripteurs dans le document d_j.

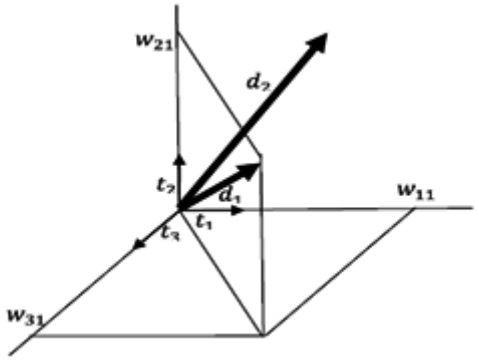


FIGURE 3.1 – La représention des documents dans l'espace d'indexation vectoriel

Dans le modèle booléen, la pondération t_i est 1 si le terme apparait dans le document 0 si non. Cette pondération uniforme ne permet pas de distinguer deux documents qui sont indexés par les mêmes termes. Ainsi, il est impossible de présenter à l'utilisateur une liste triée selon l'ordre de pertinence. Dans le modèle vectoriel la pondération des termes a été prise en compte. Elle a été étudiée dans de nombreux travaux [101] [111]. Elle consiste à affecter un poids à chaque terme d'indexation, ce poids détermine l'importance du terme dans la représentation du document.

Dans la littérature, plusieurs mesures de pondération ont été proposées. La majorité de ces mesures prennent en compte la pondération locale et la pondération globale. La pondération locale traite des informations locales. Ces informations sont spécifiques au document dans lequel le terme d'indexation

ti apparait. En général, la pondération locale d'un terme dans un document, est exprimée en fonction son nombre d'apparition ou sa fréquence de dans le document. Le modèle vectoriel est relativement simple à appréhender puisqu'il se base sur l'algèbre linéaire, ainsi il est facile à implémenter. Il permet de retrouver assez efficacement des documents pertinents dans un corpus non structuré en les ordonnant en fonction de leurs pertinences, son efficacité dépendant pour une grande part à la qualité de la représentation : vocabulaire utilisé et schéma de pondération. Il comporte également plusieurs limitations qui furent, pour certaines, corrigées par des affinements du modèle. En particulier, ce modèle suppose que les termes représentatifs sont indépendants. Ainsi, dans un texte, l'ordre des mots n'est pas pris en compte. Dans sa version la plus simple, il ne prend pas non plus en compte les synonymes ou la morphologie des contenus [101].

Ce modèle de recherche présente plusieurs avantages : le langage de requête est plus simple car c'est une liste de termes, les performances sont meilleures grâce à la pondération des termes, la restitution de documents à pertinence partielle est possible, la fonction d'appariement permet de trier les documents résultats. Cependant, quelques inconvénients sont constatés sur ce modèle recherche :

— Le modèle ne considère pas les éventuels liens qui peuvent exister entre les termes,

— le langage de requête est moins expressif,

— l'utilisateur voit moins pourquoi un document lui est renvoyé.

— Le traitement de chaque terme indépendamment du reste représente sa limite majeure.

A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante[99].

L'indexation sémantique latente -LSI-

Parmi les variantes d'approches algébrique qui on été proposées pour remédier aux limitations du modèle vectoriel, afin de prendre en compte la dépendance entre les termes d'indexation, on cite le modèle LSI. L'objectif fondamental du modèle LSI [38] est d'aboutir à une représentation conceptuelle des documents, en apportant une solution au problème d'indépendance de termes repéré dans le modèle vectoriel par la réduction de l'espace dimensionnel de représentation des documents en reliant les termes afin de traiter les "concepts" plutôt que les mots simples.

Dans ces documents les effets dus à la variation d'usage des termes dans la collection sont nettement atténués. Ainsi, des documents partageant des termes co-occurents ont des représentations proches dans l'espace défini par le modèle. Ceci permet de sélectionner des documents pertinents même s'ils ne contiennent aucun mot de la requête.

Comparativement au modèle vectoriel, la technique LSI réduit la dimension de l'espace de représentation aux seuls vecteurs de représentation de l'information sémantique et ce en réduisant l'effet de variations d'utilisation des termes [38].

3.2.3 Classe des modèles probabilistes

Modèle probabiliste classique

Le modèle fondé sur les probabilités date depuis (1976) en se basant sur le principe de classement des probabilités (probability ranking principle). Il est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête en estimant qu'il y a une incertitude dans la représentation de la requête et des documents de la collection [96]. Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Les documents et la requête sont représentés par des vecteurs dans l'espace d'indexation comme dans le modèle vectoriel. Dans ces vecteurs les pondérations des index sont binaires. Pour une requête q l'ensemble des documents disponibles est divisé en deux sous ensembles : l'ensemble R des documents pertinents et l'ensemble NR des documents non pertinents. A chaque document d on associe deux probabilités :

— P(R|d) : la probabilité que le document d soit pertinent pour la requête q ;

— P(NR|d) : la probabilité que le document d soit non pertinent pour la requête q.

Ainsi le modèle probabilistes trie les documents selon leur probabilité vis-à-vis une requêtes.

L'idée principale et de calculer les probabilités de pertinence et non pertinence d'un document par rapport à une requête et de sélectionné les documents ayant à la fois une forte probabilité de pertinence et une faible probabilité de non-pertinence. Pour ce modèle, des mesures de similarité entre le document d et la requête q sont alors calculées en fonction de ces deux probabilités P(R/d) et P(NR/d). Contrairement au modèle booléen qui fait une recherche restrictive par appariement exact [96].

L'idée générale de cette approche est d'ordonner l'ensemble des documents de la collection selon leurs probabilités de pertinence. Si un système de recherche ordonne les documents selon leurs probabilités de pertinence, où les probabilités sont estimées en fonction des données disponibles au système pour cet objectif, alors la performance du système sera améliorée. Selon Robertson [96], deux hypothèses doivent être vérifiées pour garantir l'optimalité d'ordonnancement d'une collection de documents :

— la pertinence des documents doit être une variable aléatoire binaire prenant l'une des valeurs vrai ou faux,

— la pertinence d'un document ne doit pas influencer la pertinence d'un autre document.

Modèle Réseaux Inférentiels Bayésiens

Ce modèle est fondé sur l'approche probabiliste en tenant compte de la dépendance des termes d'indexation. Un réseau Inférentiel bayésien est un graphe de dépendance acyclique et orienté ; En associant des probabilités initiales pour les racines du graphe, et en calculant, le degré de croyance associé à chacun des nœuds restants de proche en proche. Les réseaux Inférentiels bayésiens sont utilisés en recherche d'information pour représenter les dépendances entre termes d'indexation permettant ainsi de déterminer l'importance des termes à partir des liens de dépendance avec d'autres termes d'indexation. Le réseau est constitué d'un ensemble de nœuds référents des concepts, assortis de deux valeurs : degré de pertinence et degré de non pertinence. Les liens de dépendances entre nœuds sont issus du calcul des coefficients de corrélation entre les termes.

Dans le contexte de la recherche d'information les et les arcs sont définis comme suit :

— Les noeuds : représentant des concepts, des groupes de termes ou des documents.

— Les arcs : représentant les dépendances entre termes et entre termes et documents.

3.3 Processus de Recherche d'Information

Le but de la recherche d'information et de donner l'illusion que l'accès à l'information est une tâche simple qui se réalise en quelque clics, Sauf que derrière cette simplicité se cache tout un processus compliqué, minutieux et robuste. Nous présentons ci-dessous les trois grandes fonctions du modèle basic de ce processus de recherche d'information :

1. fonction de l'indexation des documents : transpose le corpus manipulé par le système de rechercher d'information en corpus indexé.
2. Fonction de correspondance (similarité).
3. fonction d'interrogation : représentation, analyse et reformulation des requêtes.

Le système de recherche d'information (SRI) adopte un modèle de représentation afin de synchroniser le modèle du document avec le modèle de la requête (Voir Figure 3.2). C'est-à-dire que ce système repose sur la définition d'un modèle de recherche d'information qui effectue ces deux transpositions et fait correspondre les documents aux requêtes. Ainsi la transposition d'un document en un document indexé repose sur un modèle document. De même, la transposition du besoin utilisateur en requête repose sur un modèle requête. Enfin, la correspondance entre une requête et des documents s'effectue par une relation de pertinence [81].

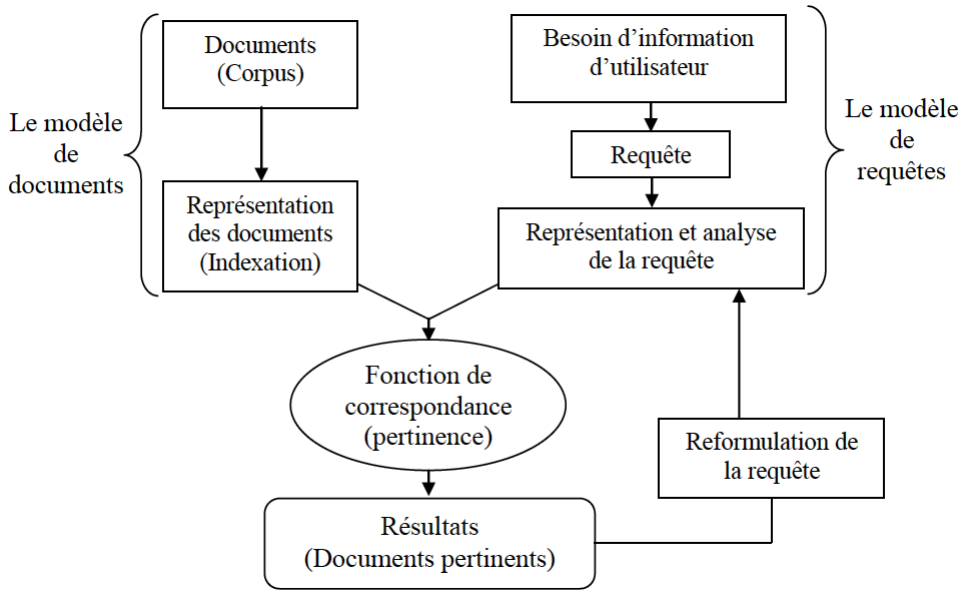


FIGURE 3.2 – Modèle basique de processus de la Recherche d'Information (Modèle en U)

3.3.1 Fonction d'indexation

L'indexation représente une étape cruciale dans le processus globale de la recherche d'information, son objectif est de représenté les documents dans un format exploitable par le système de recherche d'information, et consiste à extraire des documents les termes les plus discriminants appelé index. Cette fonction peut être assez longue temporellement selon le nombre de documents dans le corpus ainsi que la taille des documents. L'indexation peut se faire de trois manières différentes selon Salton [99] :

— Manuelle : faites pas un spécialiste du domaine, elle assure une meilleur précision dans les documents restitué par le SRI, Néanmoins, elle présente plusieurs inconvénients comme : L'effort humain, la perte du temps, et la subjectivité.

— Semi-automatique : la tâche de l'indexation est réalisée conjointement par un programme informatique et un spécialiste du domaine, le choix final revient généralement à l'indexeur humain. Ce type d'indexation utilise dans la majorité des cas un vocabulaire contrôlé.

— Automatique : dans ce cas, l'indexation est entièrement automatisé, elle est réalisé par un programme informatique et passe par un ensemble d'étapes que nous détaillons dans ce sous chapitre.

L'indexation est le processus qui consiste à décrire et à caractériser un document selon une représentation afin de permettre une recherche efficace des informations contenues dans une collection de documents sans avoir à analyser chaque texte de document à chaque interrogation ou recherche.

Mathématiquement, un index est une relation qui relie chaque document à l'ensemble des mots clés ou descripteurs décrivant le thème qu'il traite. La relation inverse permet de capturer, pour chaque mot clé, le document qu'il décrit.

Plusieurs critères sont imposés afin d'avoir une indexation de bonne qualité et utile pour la recherche d'information. En effet, les résultats d'un système de recherche d'information dépendent de la qualité du processus d'indexation des données qui doit respecter :

— L'exhaustivité : tous les termes du document doivent être représentés dans l'indexation.

— La spécificité : l'indexation ne doit pas être ni trop générale, ni trop spécifique ou particulière.

— La sélectivité : le degré d'intérêt des informations retenues pour les utilisateurs,

— L'uniformité : tous les documents doivent être indexés de la même manière.

— La cohérence : deux textes traitant un même sujet, sans utiliser le même vocabulaire, sont indexés avec les mêmes descripteurs.

— Adéquation entre les représentations : il faut que l'indexation de la requête et du document partagent le même vocabulaire.

Il existe deux façons contradictoires de décrire le problème de l'indexation en recherche d'information. L'une est "la représentation sans discrimination' qui caractérise le document par la représentation de son contenu, sans tenir compte des autres documents. L'autre manière appelée " la discrimination sans représentation" consiste à s'assurer que lors de la caractérisation d'un document, on le distingue complètement ou potentiellement du reste des documents. Notant qu'aucune de ces deux approches extrêmes n'est appliquée en pratique. Bien que l'indexation se base sur des techniques bien définies, il y a plusieurs indexations différentes possibles et valables d'un même texte du faite où 80% des documents existant a présent sont des textes. Dans ce qui suit nous détaillerons les étapes de l'indexation des documents textuelle [4].

Choix de terme

Cette étape consiste à transformer chaque document d_i en un vecteur $v_i(w_{1j}, w_{2j}, \ldots, w_{|\mathrm{T}|j})$ où T est le nombre de tout les termes qui apparaissent au moins une fois dans le corpus. Le poids w_{ki} indique l'importance du terme t_k dans le document d_i.

Avant de commencer cette étape, il est fortement conseiller de faire un prétraitement de texte afin d'écarter les termes non informatifs. Par exemple la suppression de terme qui appartient a un stop-liste (une liste des terme non informatifs dresser pour chaque langue) est une opération de prétraitement, la suppression des caractère spéciaux et une autre opération. On peut faire beaucoup d'opérations dans le prétraitement mais il est important de faire attention au sens, autrement dit, minimiser la perte d'information. Généralement le prétraitement permet de réduire significativement la taille des vecteurs de 30% selon les statistiques [100].

Représentation en « sac de mot » La plus simple représentation de documents textuels, a été introduite dans le cadre du modèle vectoriel. Son principe consiste simplement à transformer chaque texte en un vecteur dont chaque composante représente un mot [103].

Avantage : dans ce cas le mot possède un sens explicite mais il faut tout d'abord définir ce que c'est un mot ; par exemple, dans les langues française ou anglaise les mots sont séparés par des espaces ou des signes de ponctuations.

Inconvénient : on trouve des sigles ou ce que l'on appelle les abréviations par exemple le mot « ANP » qui signifie : Armée Nationale Populaire, ainsi que les mots composés, par exemple le mot «clou de girofle» ; Ces exceptions nécessitent un prétraitement linguistique.

Avec cette approche les documents sont représentés par des vecteurs de très grande dimension. Même des textes de taille moyenne peuvent contenir de nombreux mot différent. Notons que les algorithmes d'apprentissage deviennent difficile a l'utiliser a cause de la grande taille des vecteurs, et en plus de ça cette représentation des documents textuelle est typiquement creuse.

Représentation en « sac de phrase » Certains chercheurs optent pour l'utilisation des phrases comme unité. Les phrases sont plus informatives que les mots seuls, par exemple : « recherche d'information» ou « world wide web » car les phrases ont l'avantage de conserver l'information relative à la position du mot dans la phrase [103].

Représentation avec des racines lexicales (Stemming) Dans la description du modèle « sac de mot », on a le problème de dimension des vecteurs, ce problème est causé en particulier par les différentes formes de conjugaison d'un verbe qui donne autant de mots. Pour résoudre ce problème il faut considérer uniquement la racine des mots plutôt que les mots entiers (stem en anglais). On ne manque pas d'algorithmes pour substituer les mots par leur racine ; l'un des plus connus pour la langue anglaise est l'algorithme de "porter" [103].

Représentation avec des lemmes La lemmatisation est l'utilisation de l'analyse grammaticale pour remplacer le verbe par son infinitif et les noms par leur forme au singulier, ce qui est plus difficile et

plus compliqué par rapport à la détection de racine. Son objectif est d'associer à chaque mot une entrée dans le lexique (ensemble de lemme). Un algorithme efficace, nommé TreeTagger [106], a été développé pour les langues anglaise, française, allemande et italienne. Cet algorithme utilise des arbres de décision pour effectuer l'analyse grammaticale, puis des fichiers de paramètres spécifiques à chaque langue. Cette représentation est simple mais elle a ces faiblesses :

— La perte d'information par le contexte syntagmatique, nécessaire à la distinction des lemmes polysémiques.

— Les présences de synonymes, considérés comme des lemmes différents même s'ils font référence au même concept.

Représentation basée sur les n-grammes La notion de n-grammes et plus particulièrement bigramme et trigrammes (c'est-à-dire avec respectivement $n = 2$ et $n = 3$) est apparue à l'origine dans Pratt [91] selon Shannon [110]. Ce dernier a introduit la notion de n-grammes le cadre de systèmes de prédiction de caractères en fonction des autres caractères précédemment entrés. La notion de n-grammes de X se définit comme une séquence de n X consécutifs. De là X peut être un mot ou un caractère.

Les n-grammes de caractère prennent en considération les espaces. Car la non prise des espaces introduit du bruit. Les n-grammes de caractères sont les premiers à avoir été utilisés pour une tâche utilisant des données textuelles et ils sont très utilisés dans l'identification de la langue ou encore la recherche documentaire. De nombreux travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes. Cette technique présente pas mal de points forts, par rapport à d'auteurs technique :

— les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales.

— l'indépendance de langue comme le montre le travail de Dunning[39] ;

— les n-grammes de caractères sont tolérants aux fautes d'orthographes et au bruit pouvant être causés lors de l'utilisation de lecteurs optiques.

Représentation conceptuelle Nommée aussi : représentation basée sur ontologie, elle se base aussi sur le formalisme vectoriel pour représenter les documents : les éléments du vecteur ne sont plus associés à des termes d'indexation mais plutôt à des concepts. Pour permettre une telle représentation des documents, il est nécessaire de pouvoir projeter les termes sur un thesaurus comme *Wordnet*.

Le seul handicape de cette représentation est la difficulté de la construction et la mise en œuvre d'une ontologie, de plus, quelques termes peuvent être ambigus et dénoter plusieurs concepts. Ainsi les noms propres du document ne sont pas pris en compte. En effet les noms propres, étant sémantiquement vides par définition, ne possèdent pas de représentation [103].

Pondération des termes

Le calcul de la pertinence d'un document en réponse à une requête utilisateur est lié aux poids attribués aux termes communs au document et à la requête. La pondération des termes ("term weighting") est l'élément principal dans tout modèle ou processus de Recherche d'Information. Une fois le choix des composantes du vecteur représentant un texte d_i fait, il faut procéder à une codification pour chaque coordonnée de son vecteur v_i. Et pour cela il faut calculer le poids w_{ki} de chaque terme, pour cela, il existe différentes méthodes pour calculer ce poids ; les plus simples sont :

— Binaire : « 0 » pour indiquer que le terme n'existe pas dans le document et « 1 » pour indiquer que le terme existe dans le document ;

— Par occurrence : w_{ki} égale au nombre de fois ou le terme apparaît dans le document.

Les méthodes de codage sont basées sur les deux observations suivantes [79] :

— Plus le terme est fréquent dans un texte, plus il est en rapport avec le sujet de ce texte ;

— Plus le terme est fréquent dans un corpus, moins il sera utilisé pour faire la différence entre les documents.

Lorsqu'un auteur écrit un texte, il répète certains termes pour développer un aspect du sujet.

Quelque années plus tard une autre observation phare a été faite par : "la fréquence d'apparition des mots dans les textes en langage naturel est significative de l'importance de ces mots dans le seul but de représenter le contenu de ces textes". Delà la naissance du principe de pondération.

La pondération des termes est une mesure statistique qui peut être modifiée par différents critères et s'appuie sur plusieurs méthodes statistique, sémantique ou probabilistes, qui sont proposées dans la littérature pour mesurer les termes "importants".

Loi Zipf Les travaux de Zipf [48] ont marqué le début d'un courant : l'analyse linguistique par des méthodes quantitatives. En particulier l'analyse de l'indexation des bases de données et la recherche des lois de distributions des termes indexés. Il constate, en étudiant des corpus de données textuelles, des régularités sur la fréquence d'apparition des termes.

La loi de Zipf décrit la répartition statistique des fréquences d'apparition des différents éléments d'un ensemble, comme les termes d'un texte. Concernant le texte, selon Zipf, les termes dans les documents ne s'organisent pas de manière aléatoire mais suivant une loi inversement proportionnelle à leur rang. Le rang d'un terme est sa position dans la liste décroissante des fréquences des termes du corpus. Ainsi, la fréquence du second terme le plus fréquent dans le corpus est la moitié de celle du premier, la fréquence du troisième terme le plus fréquent, son tiers, etc. La loi de Zipf est à la base de la conjecture de Luhn [79]. Si les termes sont rangés par ordre décroissant de leur fréquence d'apparition, le produit de cette fréquence par le rang est quasiment constant.

Cette loi est connue sous le nom de " loi de Zipf " :

$$Frequence \times \mathrm{R}ang \approx Constante \tag{3.2}$$

Tf-IDF La fonction Tf-IDF (acronyme pour « term frequency inverse document frequency ») est la pondération la plus utilisée dans la littérature [103]. Sa force réside dans le faite qu'elle implémente en même temps : l'Exhaustivité et Spécificité.

Le poids de terme t_k appartenant au document d_i égale à :

$$\mathrm{T}f-\mathrm{IDF}(t_k,d_i) = \mathrm{N} \times \log(\frac{\mathrm{A}}{\mathrm{B}}) \tag{3.3}$$

Où :

— N : le nombre d'occurrences du terme t_k dans le texte d ;

— A : le nombre total de textes du corpus ;

— B : le nombre de textes dans les quels le terme t_k apparaît au moins une fois.

Si on désire avoir des poids entre 0 et 1, on peut la normaliser. La fonction $\mathrm{T}f-\mathrm{IDF}$ a deux points forts : efficacité dans des tâches de catégorisation de textes et simplicité de calcul. Cette pondération issue du domaine de Recherche d'Information est inspiré de la loi de Zipf.

TFC Contrairement au codage $\mathrm{T}f-\mathrm{IDF}$, le codage TFC est similaire à celui de $\mathrm{T}f-\mathrm{IDF}$ mais il corrige les longueurs des textes par la normalisation en cosinus, pour ne pas favoriser les documents les plus longs [103].

$$\mathrm{TFC}(t_k,d_i) = \frac{\mathrm{T}f-\mathrm{IDF}(t_k,d_i)}{\sqrt{\sum_{k=0}^{|\mathrm{TOKEN}|} \mathrm{T}f-\mathrm{IDF}(t_k,d_i)^2}} \tag{3.4}$$

Modèle de poisson En 1994, Robertson et Walker [97] ont proposé le modèle probabiliste basé sur la distribution de Poisson. Dans ce modèle, l'appariement entre un document et une requête est obtenu par le produit des poids des termes de la requête communs à ceux du document. La pondération d'un terme est représentée dans ce modèle par $\mathrm{P}_{tf(t_k,d_j)}$ signifiant la probabilité qu'un terme k apparaisse avec la fréquence t_f dans un document pertinent j ($\overline{\mathrm{P}_{tf(t_k,d_j)}}$ respectivement non pertinent). P0 et $\overline{\mathrm{P0}}$

représentent l'absence du terme dans un document pertinent et non pertinent respectivement. Le poids est donné par :

$$poisson(t_k, d_i) = \log \frac{\mathrm{P}_{tf(t_k,d_i)} \times \mathrm{P0}}{\mathrm{P}_{tf(t_k,d_j)} \times \overline{\mathrm{P0}}} \tag{3.5}$$

Réduction de dimension

La taille de données est calculée à partir de nombre de variables (termes en données textuelles) et le nombre d'exemples (textes en données textuelles). Ces deux dimensions peuvent être très grandes : en Text Mining les variables c'est-à-dire les termes peuvent atteindre des centaines de milliers, ce qui peut poser un problème lors de l'exploration et l'analyse. La réduction des dimensions est une approche ancienne qui permet d'apporter une solution à ce problème[23].

Elle a pour objectif de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes selon un critère fixé. La sélection de ce sous-ensemble de caractéristiques entraîne l'élimination des informations non pertinentes et redondantes et par conséquent, elle permet la réduction de la dimension[115].

En effet, les principaux objectifs de la réduction de dimension sont de :

— Faciliter la visualisation et la compréhension des données ;

— Réduire l'espace de stockage nécessaire ;

— Réduire le temps d'apprentissage et d'utilisation ;

— Identifier les facteurs pertinents.

Les techniques mathématiques de réduction de dimension sont classées en deux grandes catégories [93] :

— La sélection d'attributs « feature selection » ;

— L'extraction d'attributs « feature extraction ».

Sélection des termes La sélection de termes présente trois intérêts :

— Elle permet d'écarter les termes non pertinents d'un point de vue statistique ;

— Elle permet d'éviter le sur-apprentissage (overfitting) ;

— Elle permet d'améliorer l'efficacité des algorithmes d'apprentissage ayant des difficultés à gérer un espace de représentation important.

Des techniques de sélection d'attributs ont été développées pour réduire la dimension de l'espace vectoriel. Chacune de ces techniques utilise ses propres critères qui lui permettent de rejeter les attributs jugés inutiles.

On obtient alors un vocabulaire réduit, des textes représentés par des vecteurs de dimension plus réduite que celle d'origine. Le rayon d'action de la réduction s'étend même sur le temps de calcul qui sera plus court et dans certains cas sur la précision de classification qui augmente.

Il existe plusieurs métriques de sélection de terme, citant [93] :

— Information mutuelle ;

— Chi_2.

Extraction d'attributs La réduction de dimensionnalité par extraction de termes a un effet positif préservation de sens, en « fondant » les unes dans les autres des dimensions associées à des mêmes concepts (synonymie).

Le processus de réduction par l'extraction d'attributs consiste à créer un ensemble de nouveaux attributs à partir des attributs originaux pour optimiser la performance du système de Recherche d'Information.

Les principales méthodes sont [93] :

— LSA et ses dérivés ou LPSA ;

— Le regroupement des termes (term clustering).

3.3.2 Fonction de correspondance (ou mesure de similarité)

La fonction de correspondance est la pièce maitrise d'un système de Recherche d'Information, elle établit une relation d'égalité ou ressemblance entre les termes d'un document avec ceux d'une requête, elle aboutit à calculer la pertinence du document vis-à-vis d'une requête par le biais des mesures de similarité [18]. Elle permet ensuite au système de Recherche d'Information d'ordonner les documents renvoyés à l'utilisateur.

Évaluer des similarités entre entités est l'un des problèmes centraux dans plusieurs disciplines comme l'analyse de données notamment textuelles, la recherche documentaire.

Afin de produire les structures qui vont être utilisées pour représenter les textes lors du calcul de similarité, les données textuelles doivent tout d'abord être décomposées en unités lexicales plus simples. L'objectif ici est de voir si deux documents sont proches l'un de l'autre, s'ils se ressemblent. Par document, on entend : texte, suite de mots, énoncés, dialogues...

On dit que deux documents sont proches l'un de l'autre si la distance entre eux est petite. On dit que deux documents se ressemblent s'ils sont similaires. Si la distance entre le document D1 et le document D2 est petite alors leur similarité est grande.

Nous allons voir ci-après les distances de similarité les plus utilisés.

Indice de Jaccard

L'indice de Jaccard est le nombre de mots communs divisé par le nombre total de mots moins le nombre de mots communs :

$$Similarité(D_1, D_2) = \frac{m_c}{(m_1 + m_2) - m_c} \tag{3.6}$$

m_c = nombre de mots en commun = lexique partagé ;
m_1 = taille du lexique du document D_1 (nombre de mots différents dans D_1) ;
m_2 = taille du lexique du document D_2 .

Les vecteurs utilisés dans ce calcul de la similarité avec l'indice de Jaccard se fondent sur la présence/absence des mots. Ils n'utilisent pas les valeurs numériques et les fréquences, seulement l'absence ou la présence de mots, qu'on peut caractériser par un 0 (absence) ou par un 1 (présence).

Cosinus

Cette mesure, nécessite l'utilisation de la représentation vectorielle complète, c'est-à-dire avec la fréquence des mots.

Deux documents sont similaires si leurs vecteurs sont confondus. Si deux documents ne sont pas similaires, leurs vecteurs forment un angle α dont le cosinus vaut :

$$Similarité(D_1, D_2) = \cos\alpha = \cos(V_1, V_2) = \frac{V_1.V_2}{\|V_1\|.\|V_2\|} \tag{3.7}$$

Se lit : produit scalaire V_1. V_2 divisé par le produit de la norme de V_1 multiplié par la norme de V_2 ; la norme de V_1 = la longueur de V_1 (Pythagore). Si le cosinus de α est égale à 1 c'est-à-dire que les deux documents sont identiques, tant que la valeur de cosinus et petite on déduit que les deux documents sont moins similaires.

$$Similarité(D_1, D_2) = \cos\alpha = \cos(V_1, V_2) = \frac{a_1.b_1 + a_2.b_2 + ... + a_n.b_n}{\sqrt{{a_1}^2 + {a_2}^2 + ... + {a_n}^2} + \sqrt{{b_1}^2 + {b_2}^2 + ... + {b_n}^2}} \tag{3.8}$$

$V1 = a_1, a_2, \ldots, a_n$, $V2 = b_1, b_2, \ldots, b_n$

Minkowski

Cette distance est une généralisation de la distance Euclidienne et de la distance de Manhattan. Elle se calcule de la façon suivante :

$$Similarité(D_1, D_2) = \left(\sum_{i=1}^{n} |a_i + b_i|^r \right)^{\frac{1}{r}} \tag{3.9}$$

— Si $r = 1$ alors il s'agit de la distance de Manhattan ;

— Si $r = 2$ alors il s'agit de la distance Euclidienne.

3.4 Évaluation d'un système de Recherche d'Information

Dès l'apparition des premiers systèmes de recherche d'information, la tâche d'évaluation dédiée à ces systèmes est apparue. L'évaluation est une opération très nécessaire et fait partie du processus de la recherche car elle permet de contrôler et d'évaluer les opérations et la performance du système, de caractériser le modèle et de fournir des éléments de comparaison entre modèles de recherche d'information et peut être même le déclencheur de l'optimisation : reformulation des requête par exemple. Elle se base sur plusieurs critères : le temps de réponse, la facilité d'utilisation, la pertinence des résultats, la qualité de la présentation... etc. Cette évaluation des systèmes s'étale sur deux aspects [16] :

— d'une part l'aspect d'efficience qui est lié au rendement du système de recherche d'information, elle consiste à mesurer la qualité de la recherche en termes de critères liés au déroulement pratique d'une session de recherche.

— Et d'autre part l'efficacité qui est lié à la qualité du résultat, elle mesure la pertinence de la recherche par la comparaison des documents retournée par le système avec des documents attendus par l'utilisateur en utilisant une collection de test qui englobe une liste de réponses idéales établies par des experts ayant une grande connaissance du corpus et du domaine des documents.

Par conséquent, plusieurs questions se posent au sujet de ces outils, notamment au sujet de leur performance et de la pertinence des résultats qu'ils retournent.

D'autre communauté de chercheurs propose une autre vue théorique d'évaluation des systèmes de recherche d'information, en proposant deux paradigmes d'évaluation :

— la première est appelée « paradigme système », elle est quantitative et représente le modèle actuel dominant, elle évalue les performances du système en termes de pertinence des documents retournés par le système vis-à-vis des besoins en information des utilisateurs [21].

— La deuxième est dite « paradigme usager », qui est centré sur la satisfaction de l'utilisateur, et non sur les performances intrinsèques du système, par la prise en compte du comportement des utilisateurs en situation de recherche[21].

Plusieurs travaux qui se concerne les collections de tests, des protocoles d'évaluations et des mesures de jugement de pertinence. Nous abordons le protocole d'évaluation des SRI et les fondements basic de ces protocoles y compris les différentes compagnes et mesures d'évaluation qui existent dans la littérature. Enfin nous cernons les limites des approches classiques pour l'évaluation et nous présentons les nouvelles métriques adaptées aux systèmes de recherche contextuelle et sémantique.

3.4.1 Notion de pertinence

En recherche d'information, la pertinence représente un critère crucial d'évaluation. Elle est liée essentiellement au centre d'intérêt de l'utilisateur et du domaine d'application de système de recherche d'information, On peut distinguer deux sortes de pertinence :

Pertinence système

La pertinence système est valeur numérique calculé pour chaque document par le système de recherche d'information selon modèle de pertinence qui généralement est le score de ressemblance supposée pour l'utilisateur entre la requête et le document.

Pertinence utilisateur

La pertinence utilisateur est l'évaluation exprimée par l'utilisateur subjectivement de son satisfaction des documents retournée par le système de recherche d'information. Le jugement de pertinence est donné par un degré de pertinence des documents, puisque les besoins des utilisateurs est différent, le même besoin est exprimé de manière différente par les utilisateurs, ainsi la diversité d'évaluation entre utilisateurs qui dépend sur leurs connaissances personnelles et de leur expérience, ainsi que du contexte dans lequel s'effectue leur recherche, mène au désaccord entre les utilisateurs.

Nous pouvons classer la pertinence selon quatre types :

— **Pertinence algorithmique** : c'est une classe indépendante du contexte, elle dépend des caractéristiques des requêtes d'une part et des documents d'autre part [18] ;

— **Pertinence thématique** : elle évalue le degré d'adéquation de l'information retournée au thème la requête et non au contenu [124] ;

— **Pertinence cognitive** : c'est une amélioration de la pertinence thématique par la perception ou les connaissances de l'utilisateur sur ce même thème [124] ;

— **Pertinence situationnelle** : évalue l'utilité de l'information relativement au but de recherche de l'utilisateur [18].

3.4.2 Compagnes d'évaluation

L'histoire de la Recherche d'Information a été marquée par l'apparition de plusieurs compagnes d'évaluation. Les campagnes d'évaluation permettent d'évaluer plusieurs systèmes de Recherche d'Information par des collections différentes, afin de valider les différents modèles mis en œuvre, et de comparer les systèmes. Leurs objectifs essentiels sont :

— Encourager la Recherche d'Information sur de grandes collections fermées ;

— Cordonner entre l'industrie, l'académie et le gouvernement par la mise en place d'une aire ouverte d'échange d'idées sur la recherche ;

— Rendre accessibles les nouvelles techniques d'évaluation pour les industriels et les académiciens.

Compagne TREC

Cette équipe a pris un grand élan dans l'évaluation des systèmes de recherche d'information, elle organise des conférences annuelles qui ont pour but de mettre en évidence les différentes méthodes et techniques évoluées dans la recherche d'information sur des collections de test volumineuses, TREC est financée par la DARPA (Defense Advanced Research Projects Agency) et le NIST (National Institute of Standards and Technology). À chaque session, TREC met à disposition des chercheurs un ensemble de documents étiqueté (pertinent ou non pertinent) pour un ensemble de requêtes déterminé par des juges humains, ainsi qu'un programme nommé trec-eval qui permet de calculer, pour un ensemble de requêtes, les performances des systèmes selon plusieurs critères et mesures[71].

La première campagne TREC-1 été en 1992 avec la participation de 25 intervenants issus du monde académique et industriel, les premiers thèmes traités étaient le routage et la recherche ad hoc. Par la suite il y a eu, le filtrage de l'information, la recherche d'information non anglaise, la question/réponse, la recherche d'information multimédia, la recherche d'information sur le Web ainsi que la recherche d'information contextuelle à travers les workshops HARD track et Contextual Suggestion Track. Certaines tâches peuvent être considérées comme principales ou plus importantes. Cet ensemble évolue chaque année en fonction des demandes et des propositions.

Compagne CLEF

L'équipe CLEF (Cross-Language Evaluation Forum) s'est spécialisée aux systèmes multilingues et vu le jour en l'an 2000, c'est un programme européen développée pour l'évaluation de système de recherche d'information en multilingues (Cross-Langage) et la recherche inter-langues, plus précisément en langues européennes. Par la suite, il a été élargi pour couvrir d'autres tâches comme la recherche

d'information sur le Web, la recherche d'information géographique (RIG), la recherche vidéo et image et récemment la tâche d'évaluation de la recherche XML[17].

CLEF met a la disposition des chercheurs des collections de test pour les langues européennes, et fournis les infrastructure pour tester et évaluer les systèmes de recherche d'information multilingues. Ces collections multilingues sont constituées d'environ 1.8 millions de documents issus de la même période et dans 10 langues. Ce programme intègre plusieurs tâches et protocole des conférences TREC et les adapte pour le multilingues : la tâche adhoc, questions-réponses...

Compagne NTCIR

L'équipe NTCIR (NII-NACSIS Test Collection for IR Systems) a été orienté pour l'évaluation des systèmes de recherche d'information pour les langues asiatiques. Depuis 1997, Cette équipe organise une série d'ateliers dans les thèmes de question/réponse, le résumé de texte, la recherche d'information chinois et japonais, la recherche d'information sur le web, le multilingue. La compagne « NII Test Collection for IR Systems » (NTCIR) est dédiée aux langues asiatiques ainsi que la recherche inter-langues entre ces langues et l'anglais. Elle fournit des collections de test utilisables pour les expérimentations, elle offre aussi une infrastructure d'évaluation commune permettant des comparaisons entre systèmes [17].

3.4.3 Processus d'évaluation

Dès les années 1970, dans le cadre du projet Smart de Gérald Salton a conçu et mis en place un protocole complet d'évaluation de systèmes d'informations documentaires. Un protocole est constitué de la description claire et détaillé des conditions ainsi que les étapes bien défini de déroulement d'une expérience, dans notre cas l'évaluation des systèmes de recherche d'information fera l'objet d'une analyse pour détecter les limites afin de l'évoluer[7].

Un protocole d'évaluation repose généralement sur des collections de tests et des jugements de pertinence. L'évaluation d'un système de recherche d'information consiste à soumettre sous un protocole d'évaluation qui exécute des requêtes test sur une collection test, et compare les réponses du système aux réponses attendues pour obtenir une mesure de qualité sur les performances du système documentaire [27].

Collection de test

Pour évaluer un système de recherche d'information vis-à-vis une requête, il est nécessaire d'avoir un ensemble des documents pertinents pour une requête donnée. Si on veut comparer deux systèmes de recherche d'information, il faut les tester avec le même corpus de test[16].

Donc une collection test est constituée d'un ensemble de documents donnée, d'un ensemble de requête-tests associé avec leurs "réponses idéales" qui sert à évaluer la qualité des réponses retournée par les systèmes soumis à l'évaluation. Il faut savoir que la construction d'une collection test nécessite au moins 18 mois à deux ans, et consomme un investissement important de la part des experts. C'est à cette fin que des collections de tests ont été élaborées comprenant :

— Un corpus de documents : ensemble de documents accessibles, exploitables, homogènes et cohérents élaboré par des experts couvrant un domaine.

— Requête : c'est une représentation d'un besoin d'information. De préférence, les requêtes sont à la forme requise par système, afin d'éviter une analyse des résultats sur l'indexation de requête en cas ou elle est en forme libre.

— Des jugements de pertinence : la liste des documents pertinents pour chaque requête, cette liste est établie par l'utilisateur ou par un observateur expert.

Les collections de test sont le résultat de projets d'évaluation qui se sont multipliés depuis les années 1970, on peut citer la collection CACM, la collection CISI, la campagne CLEF et la campagne TREC. Cette dernière est la campagne de référence dans le cadre de l'évaluation des systèmes de recherche d'information et cela depuis son lancement en 1992, son objectif est de réunir des collections de test, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche.

Mesures d'évaluation

La pertinence des résultats d'une recherche représente la comparaison entre les réponses du système avec les réponses idéales que l'utilisateur s'attend à recevoir en utilisant certaines mesures. L'étude présentée par Baccini et son équipe de recherche [7] définit plus de 20 mesures d'évaluation. Dans ce qui suit, nous présentons les mesures les plus utilisées.

Rappel et silence Le rappel mesure la couverture du système, il représente la capacité du système à retourner tous les documents pertinents à une requête. Il est égal au ratio entre le nombre de documents pertinents retournés par le système p_t et le nombre de documents pertinents présents dans le corpus (collection de test) p , plus le rappel est proche de 1, meilleure est la réponse du système de Recherche d'Information.

$$\mathrm{R}appel = \frac{p_t}{p} \tag{3.10}$$

La notion de silence est le point de vue opposé du rappel,

$$Silence = 1 - \mathrm{R}appel \tag{3.11}$$

Précision et bruit La précision évalue la capacité du système de Recherche d'Information à retrouver uniquement les documents pertinents. La précision permet de mesurer la fraction des documents pertinents parmi ceux qui ont été retournés par le système. Elle est égale au ratio entre le nombre de documents pertinents retournés p_t et le nombre total de documents retournés par le système D, et plus la précision est proche de 100%, meilleure est la réponse du système de Recherche d'Information.

$$\mathrm{P}récision = \frac{p_t}{\mathrm{D}} \tag{3.12}$$

On utilise aussi la notion du bruit qui présente le problème selon le point de vue opposé de la précision,

$$\mathrm{B}ruit = 1 - \mathrm{P}récision \tag{3.13}$$

Courbe rappel/précision (ROC) Un système de Recherche d'Information optimal est celui qui retourne tous les documents pertinents (rappel = 1), et tous les documents qu'il retourne sont pertinents (précision =1) pour une requête de l'utilisateur. En pratique, une grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel. Inversement un grand rappel risque de diminuer la précision en retournant aussi plus de documents non pertinents, le système idéal n'est jamais atteint puisque ces deux quantités évoluent en sens inverse.

La courbe ROC permet d'évaluer les performances du système pour la requête considérée, dans un système optimal, le taux de précision est égal au taux de rappel, la courbe prend la forme d'une droite.

Mesure harmonique (F-mesure) Les mesures basiques telles que rappel et précision évaluent un aspect restreint, d'où l'idée de les combiner et de créer de nouvelles mesures.

Plusieurs mesures ont été développées afin de synthétiser les indicateurs rappel et précision qui sont une estimation courante de la performance d'un système de Recherche d'Information. Nous retiendrons la mesure décrite par Van Rijsbergen en 1979 [95] en occurrence, le F-mesure qui est la mesure de synthèse représentant la moyenne harmonique entre le rappel et la précision. On combine la précision et le rappel avec la pondération β afin de calculer la F-Mesure, où β traduit l'importance relative du rappel et de la précision.

$$\mathrm{F} - mesure = \frac{(1+\beta^2) \times \mathrm{P}récision \times \mathrm{R}appel}{\beta^2 \times \mathrm{P}récision + \mathrm{R}appel} \tag{3.14}$$

β est une valeur réelle positive, généralement initialisée à 1 donnant la formule connue de F-Mesure :

$$\mathrm{F} - mesure = \frac{2 \times \mathrm{P}récision \times \mathrm{R}appel}{\mathrm{P}récision + \mathrm{R}appel} \tag{3.15}$$

La valeur maximale de la F-mesure est 1 et la valeur minimale de la F-Mesure est 0. Tant que la valeur F-mesure tend vers la valeur maximale, cela reflète que le système de Recherche d'Information trouve le meilleur compromis entre le rappel et la précision. Et se traduit opérationnellement que tous les documents retournés sont pertinents, et tous les documents pertinents ont été retournés [18].

3.4.4 Limites et problèmes d'évaluation classique

Les approches d'évaluation utilisées dans la Recherche d'Information sont mises en place depuis plus de 20 ans, cela ne les empêchent pas toutefois de présenter certaines limites. De nos jours on ne peut pas introduire l'utilisateur au processus d'évaluation sans le séparer de sa subjectivité. Même sans l'introduction de l'utilisateur, l'évaluation rencontre un problème majeur qui est la constitution de la collection de test et également de l'évaluation elle-même [109]. Les travaux de recherche dans l'axe des mesures d'évaluation semblent stagner et non attirante; actuellement, le modèle d'évaluation le plus utilisé a été proposé en 1967 à l'issue du projet Cranfield de Cyril Cleverdon qui a proposé un modèle d'évaluation utilisant les mesures du rappel et/ou de précision et une des collections test standard. Ce modèle fournit un certain degré de standardisation en expérimentation et permet la comparaison des résultats parmi des groupes de recherche, cependant cette Recherche d'Information ne peut jamais avoir simultanément une bonne précision et un bon rappel.

Divers travaux ont montré les limites de l'évaluation des systèmes de Recherche d'Information qui varie selon les auteurs et selon le but poursuivi. D'autres remettent en doute en totalité l'approche d'évaluation des systèmes de Recherche d'Information. Dans notre thèse nous proposons de regrouper ces limites comme suit : l'absence de l'usager dans le processus d'évaluation, la spécificité des collections de test, la taille des collections de test, l'interaction de l'utilisateur avec le système [70].

L'absence de l'utilisateur dans le processus d'évaluation des systèmes de Recherche d'Information reste le plus important des défauts de modèle d'évaluation classique qui sont artificiels et arbitraires. Les utilisateurs utilisent des critères innés et implicites autres que le rappel et la précision lorsqu'ils initient ou terminent une session de recherche en tenant compte du contexte dans lequel se fait la recherche en situation d'utilisation réelle et en impliquant leurs connaissances personnelles, leurs expériences et leurs capacités mentales [70].

La subjectivité et l'ambigüité de la notion « pertinence » peuvent constituer une autre limite du modèle d'évaluation. Le modèle d'évaluation opère de façon binaire : un document peut être soit pertinent, soit non pertinent. Sauf que le degré de pertinence n'est pas évoqué, certains documents sont plus pertinents que d'autres qui le sont quand même. Le degré de pertinence dépend du jugement de l'utilisateur ayant réellement besoin de ces documents. La pertinence considérée dans l'évaluation classique des systèmes de Recherche d'Information est thématique, indépendante du contexte et des centres d'intérêt des utilisateurs.

Un autre problème réside dans la construction des collections de test elles-mêmes, leur taille, leur description, leur degré de confiance et d'utilité. La construction des collections test nécessite des choix précis des documents et requêtes à sélectionner, ainsi que les règles des jugements de pertinence. Ces choix reflètent souvent le but initial d'une collection.

Depuis la généralisation de l'architecture client/serveur, les interfaces des systèmes de Recherche d'Information sont graphiques. L'évaluation complète nécessite donc l'évaluation de l'interaction de l'utilisateur avec le système.

Conclusion sur le sous chapitre 1 : « Recherche d'Information »

Dans ce sous-chapitre nous avons présenté les principales notions de base et concepts de la Recherche d'Information, des systèmes de Recherche d'Information. Nous avons détaillé le processus de recherche et après avoir préparé le terrain théorique en exposant les différents modèles de la Recherche d'Information existant dans la littérature, nous avons discuté et donné brièvement les avantages et les limites de chacun. Nous avons aussi évoqué les paramètres et les méthodes d'évaluation des systèmes de Recherche d'Information. Avec l'avènement du Web, nous avons survolé la Recherche d'Information sur le Web.

La Recherche d'Information vise à assurer un accès simple, facile, rapide et pertinents à un ensemble de documents se trouvant dans des bases documentaires ou encore sur le Web par la définition des modèles et des systèmes et des protocoles d'évaluation de ces derniers dans le but répondre aux besoin en information des utilisateurs. Le point faible de la Recherche d'Information et l'inexistence de définitions formellement la notion de pertinence, elle représente également le point de départ pour de nouveaux paradigmes de recherche.

Introduction du sous chapitre 2 : « Data Mining »

Aujourd'hui, l'exploration de données (Data mining) est la vogue de l'industrie de base de données permettant ; Vue l'explosion des données (milliard d'instances qui doublent tous les vingt mois) et les très larges bases de données (VLDB), et leur multi dimensionnalité inexploitable par les méthodes d'analyse classiques ainsi que leurs besoins de traitement en temps réel,) ; d'en extraire de façon automatique ou semi-automatique des informations pertinentes et cachées en vue d'une utilisation industrielle ou opérationnelle de se savoir. Une fois cette information essentielle découverte, elle peut être utilisée de manière prédictive pour une variété d'applications telles que les banques, supermarchés, etc.. . . , afin de mieux comprendre le comportement des sujets étudiés.

Le data mining, qui est apparu au début des années 90, est la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques. Il est vu comme une exigence des entreprises pour valoriser les données importantes qu'elles accumulent dans leurs bases.

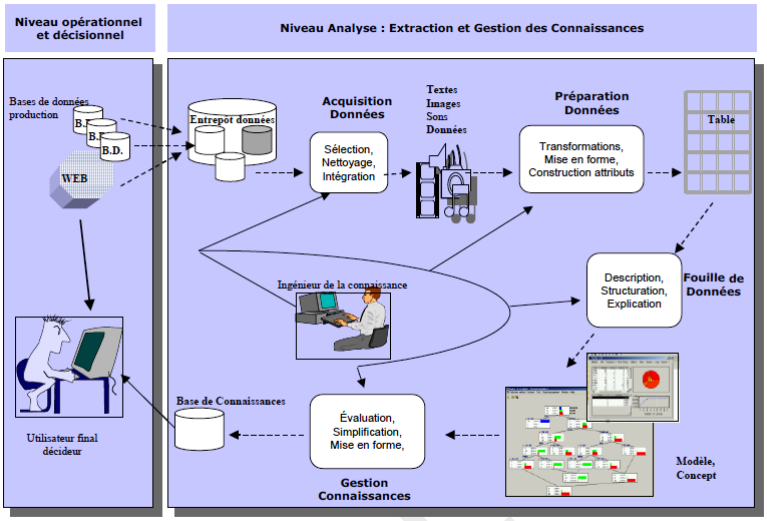
La question qui se posait est que doit-on faire des données coûteuses à collecter et à conserver. D'où la notion de l'exploration de données (Data mining), et ses effets prodigieux sur la recherche d'information pertinente dans un tas de données gigantesques, est apparue. Les techniques du Data mining représentent les nouvelles tendances du comportement, qui jusqu'à certain temps passaient inaperçu pour les entreprises [69].

Une confusion subsiste encore entre data mining (fouille de données) et knowledge discovery in databases (KDD) (extraction des connaissances à partir des données) (ECD). Le data mining est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données. Sous forme imagée, nous pourrions dire que l'ECD est un véhicule dont le data mining est le moteur.

Le data mining est l'art d'extraire des connaissances à partir des données stockées dans des entrepôts (data warehouse), dans des bases de données distribuées ou sur Internet. Il ne se limite pas au traitement des données structurées sous forme de tables numériques, au contraire il aborde les corpus en langage naturel (text mining), les images (image mining), le son (sound mining) ou la vidéo (multimedia mining). L'ECD, par le biais du data mining, est alors vue comme une ingénierie pour extraire des connaissances à partir des données, donnant leurs fruits dans la gestion des connaissances, knowledge management, ou l'indexation de documents, et récemment dans la sécurité informatique. Aucun domaine d'application n'est a priori exclu car dès que nous sommes en présence de données empiriques, le data mining peut rendre de nombreux services [65].

3.5 Extraction des connaissances à partir des données

L'ECD est un processus ou les connaissances extraites doivent être les plus intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et ordonnées. Nous allons détailler toutes ces notions et les situer dans le processus général de l'ECD. L'introduction de l'ECD dans les entreprises est récente [65]. Il convient de distinguer le niveau opérationnel et le niveau « analyse » décrit dans le schéma de la figure suivante (Voir Figure 3.3).



• Figure 5 : Processus général d'ECD

FIGURE 3.3 – Processus Général d'ECD

3.5.1 Niveau opérationnel et décisionnel

Toutes les actions sont très souvent le résultat d'une décision prise pour répondre à une demande de l'environnement. Les décisions importantes nécessitent une évaluation qui repose sur des connaissances ou des modèles préétablis. Par exemple, pour un service de vente en ligne, dès qu'un client se connecte sur le site Internet, on cherche à le profiler selon des modèles préétablis pour lui proposer l'offre de service qui est susceptible de l'intéresser.

3.5.2 Niveau d'analyse

C'est le centre des opérations d'extraction des connaissances à partir des données. Les données issues des bases de données de production alimentent les entrepôts de données qui seront utilisées en ECD.

3.6 Le processus d'Extraction Connaissance de Données

Généralement, le processus d'ECD, sous la supervision d'un spécialiste, se déroule en quatre phases : acquisition des données ; pré-traitement et mise en forme ; fouille de données (data mining dans un sens restrictif) et analyse ; validation et mise en forme des connaissances [125].

3.6.1 Phase d'acquisition des données

Les données peuvent être stockées selon des architectures variées : dans des bases de données relationnelles, dans des entrepôts de données, sur le web ou dans des banques de données spécialisées (images, bibliothèques ou librairies numériques, base de données génomiques). Elles peuvent être structurées ou non selon différents types : données tabulaires ou textuelles, images, sons ou séquences vidéo. En ECD, l'analyste ne se lance pas sans avoir une certaine idée des objectifs de son opération et des moyens informationnels et technologiques dont il dispose. La phase d'acquisition vise ainsi à cibler l'espace des données qui va être exploré, le spécialiste du data mining agit ainsi un peu à l'image du géologue qui définit des zones de prospection, étant persuadé que certaines régions seront probablement vite abandonnées car elles ne recèlent aucun ou peu de minerais. Le processus d'ECD n'est pas linéaire car il arrive aussi que l'on revienne, après analyse, rechercher de nouvelles données.

Cette phase sert généralement à nettoyer les données qui sont rapatriées. On peut également explicitement chercher à limiter le nombre d'enregistrements que l'on souhaite traiter. A l'issue de cette phase, l'analyste est, a priori, en possession d'un stock de données contenant potentiellement l'information ou

la connaissance recherchée. Il convient de préciser que l'ECD s'effectue toujours sur un échantillon de données relative à des événements passés du monde réel même s'il atteint plusieurs téra-octets [125].

3.6.2 Pré-traitement des données

Les données acquises depuis l'entrepôt peuvent être de types différents. On peut y trouver des textes de longueur variables, des images, des enregistrements quantitatifs ou des séquences vidéo.

La préparation consiste à homogénéiser les données et à les disposer en tableau lignes/colonne, car il s'agit presque toujours de la structure la mieux adaptée à l'exploitation des données. Formellement, chaque ligne-colonne peut être considérée comme un objet vecteur ayant un nombre fixe de composantes qui sera vu comme un objet mathématique que l'on pourra [24].

Précisons que dans certains cas les données arrivent sous une forme appropriée et que ça ne nécessite pas de modification alors que dans d'autres cas elles sont dans une structure qui exige des transformations telles qu'un recentrage par rapport à la moyenne ou une normalisation. En fait, le prétraitement est un acte de modélisation d'expert. Ce dernier devra définir les bonnes transformations ou les bons attributs et éventuellement effectuer une série de transformations pour obtenir des données adaptées aux méthodes d'exploitation [55].

3.6.3 Phase de fouille de données

La fouille de données est au cœur du processus d'ECD, elle fait appel à de multiples méthodes issues de la statistique, de l'apprentissage automatique, de la reconnaissance de formes ou de la visualisation. Les méthodes de data mining permettent de découvrir ce que contiennent les données comme informations ou modèles utiles. Trois catégories se distinguent pour classifier les méthodes de fouille de données utilisées. Le phénomène de l'extraction des connaissances est une préoccupation constante de l'être humain aboutissant à son évolution. Dans sa forme actuelle, le data mining est né d'un besoin de valoriser les bases de données dont la taille croît de manière exponentielle afin de mieux maîtriser la compétitivité [55].

3.6.4 Phase de visualisation et évaluation

Cette phase est constituée de l'évaluation, qui mesure l'intérêt des motifs extraits, et de la présentation des résultats à l'utilisateur grâce à différentes techniques de visualisation. Cette étape est dépendante de la tâche de fouille de données employée. En effet, bien que l'interaction avec l'expert soit importante quelle que soit cette tâche, les techniques ne sont pas les mêmes. Ce n'est qu'à partir de la phase de présentation que l'on peut employer le terme de connaissance à condition que ces motifs soient validés par les experts du domaine. Il y a principalement deux techniques de validation qui sont la technique de validation statistique et la technique de validation par expertise [12].

3.7 Définitions du Data Mining

"Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

U.M.Fayyad, G.Piatetski-Shapiro

Le Data Mining est un ensemble de méthodes et techniques qui permettent la prise de décisions, à travers la découverte, rapide et efficace, de schémas d'informations inconnus ou cachés à l'intérieur de grandes bases de données. Selon David hand, « Le datamining est l'analyse d'un ensemble d'observations qui a pour but de trouver des relations insoupçonnées et résumer les données d'une nouvelle manière, de façon qu'elles soient plus compréhensibles et utiles pour leurs détenteurs » [55]. En d'autre terme, le data mining est un procédé d'exploration et d'analyse de grands volumes de données en vue d'une part de les rendre plus compréhensibles et d'autre part de découvrir des corrélations significatives, c'est-à-dire des règles de classement et de prédiction dont la finalité ultime la plus courante est l'aide à

la décision. Ces définitions nous poussent à définir quelques termes usuels dans le domaine de la fouille de données, c'est-à-dire les données, l'information et la connaissance. La distinction entre donnée et information est que cette dernière est porteuse de connaissance[98].

3.7.1 Donnée

Parmi les définitions existantes, nous retenons qui la plus pertinente pour nous :

Une donnée est un entrant présenté sous conventionnelle pour être traiter. Il ne diminue pas l'incertitude, à l'inverse de l'information. Elle peut aussi être définie comme un « enregistrement selon un code convenu par un groupe social de la mesure ou du repérage de certains attributs d'un objet ou d'un événement » [98].

Une donnée est un enregistrement au sens des bases de données, que l'on nomme aussi "individu" (terminologie issue des statistiques) ou " instance " (terminologie orientée objet en informatique) ou même "tuple" (terminologie base de données) et "point" ou "vecteur" parce que finalement, d'un point de vue abstrait, une donnée est un point dans un espace euclidien ou un vecteur dans un espace vectoriel. Une donnée est caractérisée par un ensemble de "champs", de "caractères", ou encore d' "attributs" (en suivant les 3 terminologies précédemment évoquées : bases de données, statistiques et conception orientée objet). Un attribut peut être de nature qualitative ou quantitative en fonction de l'ensemble des valeurs qu'il peut prendre. Un attribut est qualitatif si on ne peut pas en faire une moyenne ; sa valeur est d'un type défini en extension. Sinon, l'attribut est de nature quantitative : un entier, un réel, ... ; il peut représenter un salaire, une surface, un nombre d'habitants, ... On peut donc appliquer les opérateurs arithmétiques habituels sur les attributs quantitatifs, ce qui n'est pas le cas des attributs qualitatifs [98].

Il n'est pas inutile ici de consacrer quelques lignes à ce qu'est la valeur d'un attribut. Naturellement, cette valeur est censée représenter une certaine mesure d'une quantité dans le monde. Ainsi, quand on dit qu'une couleur est " bleue ", cela signifie que nous en avons une certaine perception visuelle qui est associée à ce que, en français, on désigne par le mot " bleu " ; elle aurait pu être verte et on l'aurait appelée verte. Il est a priori impossible de comparer bleu et vert ; ce sont deux couleurs, un point c'est tout : la couleur est un attribut nominal. Si on dit qu'aujourd'hui, la température est de 20°C et qu'hier, il faisait 18°C, on peut dire que la température est plus élevée aujourd'hui qu'hier : cette fois-ci, on peut comparer les deux valeurs d'attributs, cela a un sens. Mais, si on se rappelle ses cours de physique, on sait bien que ce 20 et ce 18 sont aussi arbitraires que les mots " bleu " et " vert " : ces valeurs dépendent, notamment, des unités de mesure : la température est un attribut ordinal. Maintenant, si on dit que le nombre d'enfants de Paul est 2 et que Jacques a 3 enfants, d'une part on peut bien affirmer que Jacques a plus d'enfants que Paul et, de plus, ces deux nombres 2 et 3 ne sont pas arbitraires : le nombre d'enfants est un attribut absolu.

Au-delà de la distinction qualitatif/quantitatif, on voit donc apparaître des distinctions plus subtiles entre des attributs dont les valeurs sont arbitraires et incomparables (attribut nominal), des attributs dont la valeur est arbitraire mais que l'on peut comparer (attribut ordinal) et des attributs dont la valeur n'est pas arbitraire (attribut absolu).Ces différentes natures entraînent le fait que les opérations que l'on peut faire sur ces attributs ne sont pas les mêmes. En statistiques ou de la fouille de données, les opérations arithmétiques effectuées selon la nature de l'attribut sont licites ou non. Il importe donc de ne pas faire n'importe quel calcul, d'appliquer n'importe quel algorithme sans prendre garde aux attributs sur lesquels on les effectue. Par ailleurs, nous devons garder un principe d'indépendance du résultat des calculs par rapport aux unités de mesure dans lesquelles sont exprimées les valeurs des attributs, car Il n'y a aucune raison que l'information extraite d'une base de données change selon qu'une longueur est exprimée en millimètres, mètres ou années lumières.

En général, il faut donc gérer des données dont certains attributs ont une valeur inconnue ou invalide ; on dit que les données sont "bruitées". La simple élimination des données ayant un attribut dont la valeur est inconnue ou invalide pourrait vider complètement la base de données ! On touche le problème de la collecte de données fiables qui est un problème pratique très difficile à résoudre. En fouille de données, il faut faire avec les données dont on dispose sans faire comme si on disposait des valeurs de tous les attributs de tous les individus.

3.7.2 Information

Une information est un signal, un message qui produit un effet sur le comportement d'un vivant ou sur son état cognitif, en modifiant la représentation qu'il se fait d'un phénomène selon la définition de Mélèse en 1992. Elle est le contenu de la communication qui prend un sens. Mais la définition qui nous paraît la plus pertinente reste celle posée par J. Link-Pezet : « dans un environnement donné, l'information peut être considérée comme un ensemble de données, qui présentées d'une certaine manière (structurée) et au bon moment vont améliorer la connaissance d'une personne qui en la recevant va pouvoir réaliser une tache ou prendre une décision particulière en fonction de ses intentions originelles ». Généralement on distingue deux définitions de l'information l'une objective et l'autre subjective.

Définition objective

C'est l'ensemble de données ayant un sens particulier, pour un utilisateur. C'est-à-dire toute donnée porteuse de sens sera qualifiée d'information. Ce type de définition connaît ses limites, c'est la raison pour la quelle aujourd'hui nous adoptons une définition plus subjective [80].

Définition subjective

Une autre approche plus féconde, consiste à considérer que tout objet peut être comme une information, mais que c'est uniquement le regard porté sur un objet qui le rend porteur d'information. Dans ce cas l'objet en lui-même n'est plus porteur d'information, mais c'est le regard qui est créateur d'information, ou plutôt de sens. Cette approche est plus riche de conséquences et plus englobante. Dans ce cas, n'est information pour moi que ce à quoi je m'intéresse (Éric Sutter et Jean Michel) [54].

3.7.3 Connaissance

La connaissance traduit un processus d'acquisition d'information. Ce processus traduit un travail de synthèse de l'individu, qui va intégrer les informations qu'il reçoit, les assimiler et les critiquer. Ce processus terminé, sa « connaissance » sera enrichie. Linguistiquement on définit la connaissance comme le fait d'avoir une compétence dans un domaine donnée et comme synonyme c'est le savoir et l'acquisition. En Informatique c'est un ensemble de règles utilisant les données pour en déduire d'autres.

3.8 Le processus itératif du Data mining

Les techniques de data mining sont utilisés pour aider à élaborer des modèles prédictifs qui permettent une réponse en temps réel après une séquence de processus qui comprennent l'échantillonnage des données en temps réel, la sélection, l'analyse et la recherche [36]. Le processus du data mining est établit comme suit :

1. Pendant le nettoyage des données, le bruit et les données non pertinentes sont supprimés de la collection.
2. L'intégration des données consiste à combiner les données provenant des sources multiples et hétérogènes dans une base de données.
3. Les techniques de sélection des données permettent à l'utilisateur d'obtenir une représentation réduite de l'ensemble des données afin de maintenir l'intégrité de l'ensemble des données d'origine dans un volume réduit.
4. Dans la transformation des données, les données sélectionnées sont transformées en un format souhaitable.
5. Le data mining est l'étape dans laquelle les outils d'analyse sont appliqués pour découvrir des modèles qui pourraient être utiles.
6. L'évaluation du modèle consiste à identifier des modèles intéressants et utiles en utilisant des mesures de validation des données.
7. La représentation des connaissances est la phase finale du processus de découverte des connaissances où le savoir découvert est présenté aux utilisateurs dans des formes visuelles.

3.9 Les tâches du Data Mining

Les deux principaux objectifs de l'exploration de données (data mining) dans la pratique sont la prédiction et la description.

— **La prédiction** implique l'utilisation de certaines variables ou de champs dans la base de données pour prédire des valeurs inconnues ou d'autres futures variables d'intérêt.

— **La description** se focalise sur la recherche de modèles interprétables décrivant les données

L'importance relative de la prédiction et la description pour des applications particulières de data mining peuvent varier considérablement. Toutefois, dans le cadre de KDD (Découverte des connaissances), la description tend à être plus importante que la prédiction. Ceci est en contraste avec la reconnaissance de formes et les applications du machine learning (comme la reconnaissance vocale) où la prédiction est souvent le but principal du processus de KDD. Le Data Mining n'est pas comme on le croit le traitement prodige capable de résoudre toutes les difficultés ou besoins de l'entreprise. Cependant, une multitude de problèmes d'ordre intellectuel, économique ou commercial peuvent être regroupés, dans leur formalisation, dans l'une des tâches suivantes [75] Classification, Estimation, Prédiction, Groupement par similitudes, Segmentation (ou clusterisation), Description, Optimisation.

Alors définissons ces derniers pour lever toute ambiguïté sur les termes qui peuvent paraître similaires.

3.9.1 Classification

La classification sans qu'on se rende compte existe depuis une belle lurette pour comprendre et communiquer notre vision du monde (par exemple les espèces animales, minérales ou végétales). « La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini. » [11].

Dans le cadre informatique, les éléments sont représentés par un enregistrement et le résultat de la classification viendra alimenter un champ supplémentaire et permet de créer des classes d'individus. Les techniques les plus appropriées à la classification sont :

— Les arbres de décision,

— Le raisonnement basé sur la mémoire,

— Eventuellement l'analyse des liens.

3.9.2 L'estimation

L'estimation permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce à un barème. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type de véhicule et nombre, profession ou catégorie socioprofessionnelle, type d'habitation, etc.). Il sera ensuite possible de définir des tranches de revenus pour classifier les individus. Cette technique est souvent utilisée en marketing pour proposer des offres aux meilleurs clients potentiels. La technique la plus appropriée à l'estimation est : le réseau de neurones [11].

3.9.3 La prédiction

La prédiction ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé. La seule méthode pour mesurer la qualité de la prédiction est d'attendre![11]

Les techniques les plus appropriées à la prédiction sont :

— L'analyse du panier de la ménagère.

— Le raisonnement basé sur la mémoire.

— Les arbres de décision.

— les réseaux de neurones.

3.9.4 Le regroupement par similitudes

Le regroupement par similitudes consiste à grouper les éléments qui vont naturellement ensembles. La technique la plus appropriée au regroupement par similitudes est l'analyse du panier de la ménagère [11].

3.9.5 L'analyse des clusters (segmentation)

L'analyse des clusters consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis. La technique la plus appropriée à la clusterisation est l'analyse des clusters [11].

3.9.6 La description

C'est souvent l'une des premières tâches demandées à un outil de Data Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications. La technique la plus appropriée à la description est l'analyse du panier de la ménagère [11].

3.9.7 L'optimisation

Pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associer une fonction d'évaluation. Le but de l'optimisation est de maximiser ou minimiser cette fonction. Quelques spécialistes considèrent que ce type de problème ne relève pas du Data Mining. La technique la plus appropriée à l'optimisation est le réseau de neurones.

3.9.8 Le cercle vertueux

On ne met pas en œuvre une technique de Data Mining pour faire une simple exploration. Il faut l'inscrire dans un contexte plus global, appelé le cercle vertueux. Celui-ci est composé de quatre étapes :

— Identifier le domaine d'étude : Il faut répondre aux questions : de quoi parlons nous et que voulons nous faire ? A ce stade, on définit un objectif général.

— Préparer les données : il faut recenser les données relatives au domaine, puis les regrouper pour en faciliter l'exploration. Nous parlons de regroupement logique, ce qui inclus le client / serveur, même si ce n'est pas recommandé.

— Agir sur la base de données : étape consiste à mettre en œuvre une ou plusieurs techniques de Data Mining pour une première analyse.

— Evaluer les actions : Après étude des résultats, des actions sont mises en œuvre. Cette étape consistera à évaluer ces actions, et par-là même la performance du Data Mining, voire le retour sur investissements. L'achèvement du premier cycle débouche souvent sur l'expression de nouveaux objectifs affinés, ce qui nous ramène à la première étape.

3.10 Motivation

Une approche statistique consiste à formuler des hypothèses théoriques qui sont confirmées ou infirmées à l'aide de tests statistique, si l'étude est correctement menée, on formule les hypothèses, on en déduit le plan d'expérience, on recueille les données, puis on les analyses. En data mining, la recherche d'information est moins guidée. Cependant, le data mining repose sur un ensemble de méthode issues de statistique et de l'intelligence artificielle. Nous citons Quelques repères :

— Statistique :
  — Quelques centaines d'individus.
  — Quelque variable recueillie avec un protocole spécial (échantillonnage, plan d'expérience).
  — Fortes hypothèses sur les lois statistiques suivies.

    - Les modèles sont issus de la théorie et confrontés aux données.
    - Méthodes probabilistes et statistique.
    - Utilisation en laboratoire.
- Analyse des données.
    - Méthode essentiellement descriptive.
    - Quelques dizaines de milliers d'individus.
    - Quelques dizaines de variables.
    - Importance du calcul et la représentation visuelle.
- Data Mining.
    - Plusieurs millions d'individus.
    - Plusieurs centaines de variables.
    - Nombreuses variables non numériques, parfois textuelles.
    - Données recueillies avant l'étude, et souvent à d'autres fins.
    - Données imparfaites, avec des erreurs de saisie, de codification, des valeurs manquantes aberrantes.
    - Population constamment évolutive (difficulté d'échantillonner).
    - Nécessité de calculs rapides, parfois en temps réel.
    - On ne recherche pas toujours l'optimum mathématique, mais le modèle le plus facile à appréhender par des utilisateurs non statisticiens.
    - Faible hypothèse sur les lois statistiques suivies.
    - Les modèles sont issus des données et on en tire des éléments théoriques.
    - Méthodes statistiques, d'intelligence artificielle et de théorie de l'apprentissage (machine learning).
    - Utilisation en entreprise
- La fouille de données est motivée par les arguments suivants :
    - Les données traitées concernent à la fois des attributs qualitatifs et quantitatifs, ce qui justifie les étapes de discrétisations pour obtenir des contextes booléens ;
    - Les données sont volumineuses : des objets par millions, des attributs par milliers. Ces caractéristiques posent de nombreux problèmes algorithmiques ;
    - La fouille de données poursuit un but d'exhaustivité des connaissances découvertes. À la différence des techniques statistiques, ce ne sont pas seulement les tendances globales des données qui sont recherchées mais également des propriétés locales qui concernent un petit nombre d'objets ;
    - Dans l'optique des méthodes d'exploration qui permettent d'aider l'expert dans sa prise de décision, il est souhaitable que l'aide fournie soit clairement justifiée, expliquée et compréhensible.

3.11 Objectif du Data Mining

Dans un monde de concurrence commerciale sévère, l'une des grandes difficultés est de savoir comment chercher un profil typique d'un client dans un si grand amas de données (BDD de plusieurs tera-octets). Le data mining offre les moyens d'analyse pour cerner un tel profil comportemental. Par exemple, chercher des clients qui changent de fournisseurs, afin de tenter par des actions commerciales ad hoc, de les garder. Le data mining est un processus faisant intervenir des méthodes et des outils issus de différents domaines de l'informatique, de la statistique ou de l'intelligence artificielle en vue de découvrir des connaissances utiles qui peuvent prendre la forme d'un rapport ou d'un graphique, ou s'exprimer comme un modèle mathématique ou logique pour la prise de décision. Ensuite, elles peuvent

alimenter un système à base de connaissances ou un système expert, comme le montre la figure suivante (Voir Figure 3.4) [24].

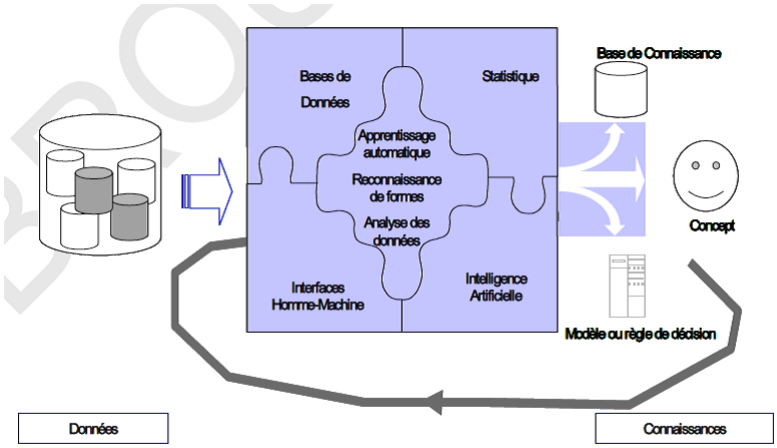


FIGURE 3.4 – Techniques et modéle d'ECD

3.12 Technique de visualisation des résultats de Data Mining

3.12.1 Les procédés de visualisation et de description

Il s'agit donc principalement d'outils de synthèse d'information. Cette synthèse peut s'exprimer par des indicateurs statistiques. Par exemple, pour des attributs quantitatifs, les indicateurs les plus utilisés sont la moyenne, l'écart-type, le mode et la médiane. Pour des attributs qualitatifs, on associe généralement la distribution selon les modalités de l'attribut. Ces indicateurs sont généralement représentés par des graphiques, car plus faciles à interpréter, comme les boites de Tuckey, les distributions (densités ou fonctions de répartition), les nuages de points. On trouve dans les logiciels de data mining une kyrielle de formes géométriques et de styles de présentation de ces concepts. La description et la visualisation peuvent être mono ou multidimensionnelles. Pour l'essentiel, il s'agit de rendre visible des objets ou des concepts qui se trouvent dans des espaces de description trop riches. Le schéma suivant (Voir Figure 3.5) représente un processus de data mining orienté vers la visualisation et la description. Pour simplifier la présentation, le tableau des données (CREDITS) ne contient que 1000 clients. Les traitements qui figurent aux extrémités des flèches synthétiseront les données des clients d'une banque selon différentes caractéristiques, qu'elles soient numériques ou graphiques .

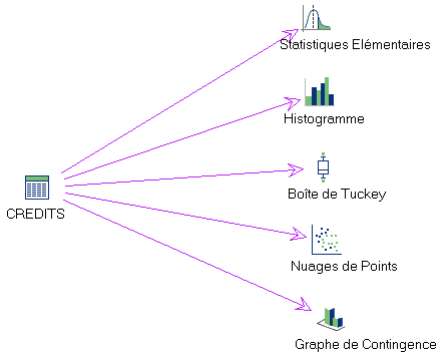


FIGURE 3.5 – Traitement de description et de visualisation

3.12.2 Les procédés de structuration et classification

En ECD, nous avons affaire à une profusion de données. Décrire ces données s'avère parfois difficile à cause de cette volumétrie. L'utilisateur cherche souvent à identifier des groupes d'objets semblables au sens d'une métrique donnée. Ces groupes peuvent par exemple correspondre à une réalité ou à des concepts particuliers. Les lignes ou les colonnes du tableau sont vues comme des points d'un espace multidimensionnel qui n'a pas obligatoirement une structure d'espace vectoriel. Les méthodes de structuration ont pour objet de repérer ces structures de groupe invisibles à l'oeil nu. Les techniques employées pour des opérations de classification relèvent de ce que nous appelons l'apprentissage non supervisé, car l'utilisateur ne sait pas a priori quelles classes, groupes ou catégories il va obtenir. Ce mode d'apprentissage est également appelé « apprentissage sans professeur » [69]. Les techniques employées sont des méthodes de classification automatique (cluster analysis), appelées aussi « classifications conceptuelles » ou « méthodes de taxinomie ». Les principales techniques se répartissent en trois groupes :

— Les méthodes nomothétiques.

— Les méthodes polythétiques.

— Les méthodes basées sur les réseaux de neurones.

3.12.3 Les procédés d'explication et de prédiction

L'objectif est de rechercher à partir des données disponibles un modèle explicatif ou prédictif entre, d'une part, un attribut particulier à prédire et, d'autre part, des attributs prédictifs. Dans le cas où on produit un tel modèle valide, il pourrait alors être utilisé à des fins de prédiction. Dans ce contexte on parle « d'apprentissage supervisé »car l'attribut à prédire est déjà préétabli. Il s'agit alors de mettre au point un processus permettant de le reconstituer de façon automatique à partir des autres attributs. En apprentissage supervisé, il y a, d'une part, une phase « inductive » consistant à développer les règles d'identification à partir d'exemples particuliers et, d'autre part, une phase « prédictive » visant à utiliser ces règles pour identifier de nouvelles instances. [37] Il existe une multitude de méthodes d'explication et ou de prédiction développée dans différents contextes, nous citons :

— **Les graphes d'induction** :

Les graphes d'induction, dont les modèles les plus populaires sont les arbres de décision, connaissent un grand succès du fait de leur facilité à mettre en œuvre, les résultats qu'ils fournissent sont aisés à interpréter et les modèles qui en sont déduits sont performants. Ils appréhendent des bases de données de grandes tailles et applicables sans restriction sur des données de n'importe quel type (qualitatives, quantitatives, ou un mélange des deux). Toutes les méthodes à base de graphes d'induction sont décrites par l'algorithme général suivant :

On part de la partition grossière formée de tous les individus de l'échantillon d'apprentissage, on recherche ensuite, parmi les p variables exogènes $(X_1, X_2, \ldots, X_p)$, celle qui permet d'engendrer la meilleure partition au sens d'un critère donné. Celui-ci devra être d'autant meilleur que les classes de la partition sont homogènes. Nous obtenons un arbre à deux niveaux dont la racine représente la partition grossière ω et dont les feuilles représentent les modalités de la variable exogène.

— **Les réseaux de neurones**

Les réseaux de neurones sont parmi les outils de modélisation les plus utilisés, en particulier pour les problèmes difficiles où le prédicateur que l'on cherche à construire repose sur de nombreuses interactions complexes entre les attributs exogènes. La structure générale d'un réseau de neurone se présente comme suit (Voir Figure 3.6) :

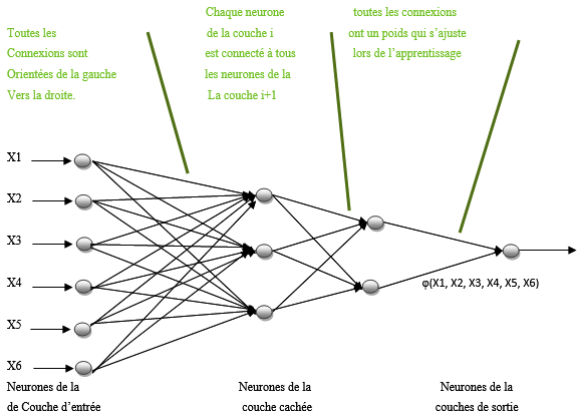


FIGURE 3.6 – réseau multicouche

— **Les méthodes de régression** :

Il s'agit d'expliciter une relation de type linéaire ou non entre un ensemble de variables exogènes et une variable endogène. Généralement, dans le cadre de la régression, toutes les variables sont considérées continues. La régression linaire est dite multiple quand le nombre de variables exogènes est supérieur ou égal à deux.

— **L'analyse discriminante** :

L'analyse discriminante est l'une des plus anciennes techniques de discrimination, proposée par Fischer en 1936. Elle concerne la discrimination de trois classes de la famille des iris (versicolor, setosa et verginica). Pour cela, il a pris un échantillon de 150 iris répartis sur les trois classes et, pour chaque fleur, il a mesuré la longueur et la largeur des pétales et des sépales. Le fichier des ces données se trouve dans la plus part des ouvrages consacrés à cette méthode et sur les sites web des communautés d'apprentissage automatique comme celui de l'Université de Californie à Irvine ou celui de la communauté de data mining comme Kdnuggets [62].

— **Les réseaux bayésiens** :

Les réseaux bayésiens sont apparus au début des années 1980. Rendus populaires par le groupe de recherche de la firme Microsoft qui les introduits dans les systèmes d'aide contextuelle d'Office, ils sont maintenant très utilisés dans la modélisation des processus complexes de décision. L'idée de base des réseaux bayésiens repose sur le célèbre théorème de Bayes mais s'est considérablement enrichie ces vingt dernières années. Un réseau bayésien est représenté par un graphe acyclique dans lequel les sommets sont des variables booléennes et les arcs les relations de dépendance. L'architecture du réseau est généralement fournie par l'expert [67].

— **Les règles d'association** :

La recherche de règles d'association dans une base de données est probablement le problème qui a le plus fortement contribué à l'émergence du data mining en tant que domaine scientifique à part entière. La grande distribution, les télécommunications et plein d'autres secteurs de la grande consommation enregistrent, dans un but de facturation, l'ensemble des transactions commerciales avec leurs clients. La recherche de règles d'associations dans un ensemble de transaction s'opère en deux temps :

  — On cherche les ensembles d'items fréquents, c'est-à-dire ceux qui apparaissent un nombre minimum de fois dans l'ensemble des transactions.

  — On génère les règles d'associations pertinentes, c'est-à-dire celles qui vérifient simultanément la contrainte minimale sur le support et la confiance.

— **Autres approches en prédiction** :

Il existe de nombreuses autres techniques destinées à la prédiction parmi lesquelles on peut citer :

  — CN2 qui opère sur des données discrètes et qui génère des règles logiques de la forme « si condition alors conclusion »,

— les SVM (support vector machine) issus des travaux sur la théorie de l'apprentissage ou les méthodes à base de voisinage comme les « k plus proches voisins » qui, pour un point dont on souhaite déterminer la classe d'appartenance, utilisent l'information sur les classes dans le voisinage du point en question.

— Nous pouvons également citer les algorithmes génétiques pour modéliser un apprentissage de règles, ou le raisonnement à partir de cas [120].

3.13 Évaluation en Data Mining

Nous devons évaluer, notamment pour les méthodes d'apprentissage supervisé, Les modèles extraits, c'est-à-dire les soumettre à l'épreuve de la réalité et apprécier leur justesse. La manière habituelle consiste à estimer le taux d'erreur du modèle, pour que l'utilisateur décide ou non d'appliquer le modèle de prédiction. Il y a plusieurs protocoles d'évaluation d'une méthode d'apprentissage, nous survolerons par la suite les deux protocole les plus utilisées : la validation croisée, leave one out.

La validation croisée, ou cross validation, consiste à répartir l'échantillon d'apprentissage aléatoirement en K paquets d'effectifs identiques. Si on note ω $k, k = 1, ..., K$ les différents sous-échantillons, le taux d'erreur en validation croisée est calculé en réservant, à tour de rôle, un échantillon ωk qui servira à mesurer le taux d'erreur en validation, l'apprentissage étant réalisé sur la totalité des individus restants : k ω-ωk . On obtient ainsi K taux d'erreurs E_k, $k = 1, ...K$. Le taux d'erreur en validation croisée est alors la moyenne arithmétique des taux d'erreurs partiels :

$$E = \sum_{i=1}^{n} Ek \tag{3.16}$$

Généralement on lui associe la variance :

$$\sigma^2 K = \frac{1}{K} \sum_{i=1}^{n} (Ek - E)^2 \tag{3.17}$$

Dans le cas où le découpage de l'échantillon conduit à avoir un individu par paquet, c'est-à-dire $K = n$ la méthode est connue sous le nom « leave one out ».

L'échantillon de validation est tiré aléatoirement à chaque itération. Contrairement à la validation croisée, un même individu statistique peut être pris dans plusieurs échantillon, ou jamais. Les paramètres fournis à l'utilisateur sont le nombre de répétitions du tirage et la taille de l'échantillon de validation. De façon analogue à la validation croisée, l'erreur du modèle est estimée par l'erreur moyenne réalisée sur les différents échantillons de validation. Les méthodes de validation citées en haut fonctionnent après apprentissage, et elles sont globales à chaque modèle. La mise en forme de ces modèles comporte différents aspects, allant de la visualisation des connaissances pour les rendre intelligibles jusqu'à l'agrégation des modèles, en passant par leur simplification [116].

Après la définition du protocole, l'évaluation revient à générer la matrice de confusion et de calculer les différentes métriques d'évaluation : rappel, précision, F-mesure ...etc.

3.14 Les défis du data mining en sécurité informatique

Les défis du data mining en termes de cyber sécurité sont classés en quatre domaines d'application qui sont [36] :

3.14.1 La modélisation des réseaux à grande échelle

La modélisation d'un cyber infrastructure est difficile, car de nombreuses mesures des graphes communs sont difficiles à calculer pour des réseaux sous- jacents. Il est difficile de construire le modèle explicatif des réseaux en raison des exigences en matière de précision d'apprentissage et de prédiction.

Un modèle de réseau peut être extrait en partie et avec attention pour l'analyse avancée, et un réseau peut être construit dans un monde réel de façon significative, mais il ne peut pas suivre l'hypothèse de

variables aléatoires. De plus, il y a les difficultés résidantes dans le calcul des mesures graphiques du modèle de réseau. Des exemples de ces modèles graphiques ont la dynamique du réseau de télécommunications, des réseaux de communications électroniques de courrier par lequel les virus se propagent, ou du réseau de liens hypertextes entre les sites Web. Le diamètre de graphe, la distance maximale entre deux nœuds dans un graphe sont des exemples d'une mesure graphique. Les difficultés de calcul nous poussent à appeler les modèles de data mining qui peuvent découvrir la nature des données réelles en utilisant un modèle plus simple.

3.14.2 La découverte des menaces

L'utilisation de data mining dans des cybers infrastructure pour la découverte de menaces souffre du volume et des données hétérogènes du réseau, le changement dynamique des menaces et les graves déséquilibres des classes de comportements normaux et anormaux. Ces défis nous poussent à appeler des méthodes qui peuvent agréger les informations des réseaux dynamiquement et localement pour détecter les attaques complexes en plusieurs étapes et prévoir les menaces potentielles et rares sur la base de l'analyse du comportement des données et des événements du réseau. Les méthodes les plus employées pour détecter le code ou le comportement malveillant sont les modèles à base de règles ou statistiques pour identifier les menaces en temps réel en utilisant la détection adaptative de la menace avec la modélisation des données temporelles et les données manquantes.

3.14.3 Le dynamisme du réseau et les cyberattaques

Beaucoup de cyber attaques propagent des programmes malveillants pour les ordinateurs vulnérables. En raison des conditions de déclenchement inconnu des logiciels malveillants qui peuvent infecter les ordinateurs dans un réseau à des degrés divers. Une fois que l'administrateur réseau les détecte, la propagation des infections doit être étudiée pour construire un système de protection. Les nouvelles méthodes de data mining sont nécessaires pour prévoir les futures attaques en se basant sur l'évolution des logiciels malveillants. Cependant, la structure détaillée du réseau est inconnue, ce qui limite les connaissances de l'évolution de l'infection.

3.14.4 La préservation de la vie privée en data mining

L'extraction de données peut être utilisée avec malveillance dans les cybers infrastructure pour violer la vie privée. En principe, plus les données complètes sont disponibles pour l'exploration des données, plus le résultat obtenu sera précis. Toutefois, les données complètes et précises peuvent également soulever des questions d'atteinte à la vie privée. En outre, le résultat d'exploration des données peut potentiellement révéler des informations privées. Le concept de la préservation de la vie privée dans le data mining (Privacy preserving data mining PPDM) protège les données privées d'être volées ou mal utilisées par des utilisateurs malveillants, tout en permettant d'extraire les données pour être utilisées [36].

Conclusion du Chapitre III

Dans ce chapitre nous avons présenté les principales notions de base et concepts de la recherche d'information. Nous avons détaillé le processus de recherche et après avoir préparer le terrain théorique en exposant les différents modèles de la recherche d'information existant dans la littérature en montrant brièvement les avantages et les limites de chacun. Nous avons aussi évoqué les paramètres et les méthodes d'évaluations des systèmes de recherche d'information. A travers les différentes sections que nous avons présentées, nous concluons que la recherche d'information, s'attache à définir des modèles et des systèmes afin de faciliter l'accès à un ensemble de documents se trouvant dans des bases documentaires ou encore sur le web. Le but est de permettre aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information, il s'agit donc de retourner l'ensemble de documents pertinents. Cependant, nous constatons que la notion de pertinence dépend de la satisfaction de l'utilisateur d'une part, et des différents sens portés par les termes de la requête d'une autre part. Cette

constatation constitue le point faible de la recherche d'information classique, elle représente également le point de départ pour de nouveaux paradigmes de recherche.

Chapitre 4

Expérimentation et Résultats

Sommaire

Introduction

Le biomimétisme est une approche scientifique révolutionnaire qui consiste à imiter les plus belles inventions de la nature (la capacité énergétique de la photosynthèse, solidité du corail, résistance du fil de soie de l'araignée...) pour les adapter au service de l'homme [6]. Mais le biomimétisme de Janine Benyus, scientifique américaine et consultante en innovation auprès de grandes entreprises américaines, va au delà de ce biomimétisme pratiqué « au premier degré », où l'on ne ferait que copier les formes naturelles (la pointe du TGV japonais est copiée du bec du martin-pêcheur ; Léonard de Vinci s'inspirait des ailes des oiseaux pour imaginer un système permettant aux hommes de voler à leur tour...etc).

La vie de l'humain a connu durant l'histoire beaucoup de guerres et le développement des stratégies des guerres donne l'avantage au camp qui détient la dernière mise à jour. Après les deux guerres mondiales et la guerre froide, aujourd'hui le développement de la science donne naissance à une guerre électronique, on prédit même que la troisième guerre mondiale sera purement électronique. Dans l'histoire, le support de donnée faisait l'objet d'attaque afin d'intercepter, modifier ou détruire l'information. De nos jours tout et surtout dans les pays développés ou tout est informatisé et stockée dans des serveurs, comme :

— les informations personnelles de l'être depuis sa naissance : nom, prénom, adresse, poids, taille, RDV de jour, CV, état de santé etc,

— la fortune d'une personne est devenue juste que des fichiers de propriété ou des tuplets dans une base de données à la mairie ou des chiffres à la banque, et ça ne s'arrête pas là,

— l'arrivé des réseaux sociaux à informatisé la vie personnelle et quotidienne, les opinions et les sentiments.

Nous les informaticiens, nous visons à reproduire le monde en virtuelle. Ce qui fait que les serveurs et les systèmes informatiques sont devenu cible d'attaque et de criminalité. La criminalité électronique est une mode et un challenge entre des jeunes hackers, mais il ne faut pas négliger que cela devienne un nerf de la concurrence entre les entreprises et les services secrets des pays, En effet de plus en plus d'entreprises subissent des attaques qui peuvent entraîner des pertes conséquentes.

Le besoin des entreprises en sécurité informatique est de plus en plus important. La mise en œuvre d'une politique de sécurité globale est assez difficile, essentiellement par la diversité des aspects à considérer. Une politique de sécurité peut se définir par un certain nombre de caractéristiques : les niveaux où elle intervient, les objectifs de cette politique et enfin les outils utilisés pour assurer cette sécurité.

Pour assurer une bonne protection des données d'une entreprise, différents outils sont disponibles. Ils sont en général utilisés ensembles, de façon à sécuriser les différentes failles existantes dans un système. La pièce maitresse d'un système de sécurité est l'IDS : Système de détection d'intrusion ; c'est le seul outil qui assure la permanence, il est responsable du déclenchement ou l'arrêt des stratégies et des outils de réponse en cas d'attaque.

Un IDS est un équipement permettant de surveiller l'activité d'un réseau ou d'un hôte donné, afin de détecter toute tentative d'intrusion et éventuellement de réagir à cette tentative. Il existe différents sorte d'IDS dans la littérature, ça diffère dans le domaine de surveillance, mode de fonctionnement ou mode de réponse.

L'approche de la sécurité des systèmes d'information qui prédomine encore aujourd'hui est trop passive. On attend de détecter une attaque tout en faisant (aveuglément) confiance aux multiples outils de protection qu'on a mis en place et qui ne sont pas infaillibles.

Il est nécessaire de faire évoluer nos postulats et nos modèles en matière de sécurité des systèmes d'information. Pour cela, une nouvelle approche proactive est indispensable. La réponse active appelée aussi défense offensive est légale vis-à-vis de la loi de ce faite qu'en 2011 le Ministère de la Défense des Etats Unis publie sa stratégie pour opérer dans le cyberespace. Il y annonce de mettre en place des capacités de Défense Active pour bloquer les intrusions visant ses réseaux et systèmes informatiques.

Si vous remontez vers la fin du deuxième paragraphe de cette introduction, vous allez trouver une phrases frappante : « Nous les informaticiens, nous visons à reproduire le monde en virtuelle », donc en tant qu'informaticiens nous adhérons et contribuons à ce but en cherchant dans la nature un système de sécurité solide ayant une défense offensive. Nous avons été attirés par une citation d'Albert Einstein

qui dit « si l'abeille disparaît, l'humanité en a pour quatre ans à vivre », comment les abeilles sociales peuvent elles se protéger et protéger leur miel face à la loi du plus fort?

Personne d'entre nous n'osera s'approcher d'une ruche sans protection, car non seulement il ne le pourra pas, mais il sera poursuivie et attaquée à des centaines de mètres par les habitants de cette ruche qui se sacrifient pour la sécurité de leur ruche. Cela nous a inspiré à modéliser un système de détection d'intrusion basé sur une méta-heuristique en l'occurrence le système de protection des abeilles sociales.

4.1 Les systèmes multi-agents

Les agents sont des entités partiellement autonomes disposant d'un nombre restreint de fonctionnalités, capables de communiquer et d'évoluer dans un environnement qu'ils peuvent le percevoir et modifier. On distingue deux catégories d'agents : les agents cognitifs et les agents réactifs. Les agents réactifs se contentent de réagir à des stimuli de l'environnement. Les agents cognitifs héritent de ce comportement mais possèdent en plus une couche d'apprentissage ainsi qu'une couche décisionnelle leur permettant d'évoluer. [15].

Les systèmes multi-agents désignent les systèmes où un ensemble d'agents exécutent des tâches locales dans un environnement afin de résoudre une tâche globale. La puissance des systèmes multi-agents vient en partie de la capacité des agents à interagir entre eux. Ces interactions peuvent être de deux formes : directes ou indirectes.

Les interactions directes sont en général caractéristiques d'agents cognitifs qui, œuvrant dans un but précis, sont capables de communiquer intentionnellement. Ce type d'interactions se rapproche d'un acte de langage. Les interactions indirectes sont au contraire caractéristiques d'agents réactifs qui en réponse à des stimuli de l'environnement déposent des informations dans celui-ci [43].

En 1959, P.P. Grassé définit ce type d'interactions indirectes dans les sociétés (en particulier les insectes) sous le nom de stigmergie : « *La coordination des tâches, la régulation des constructions ne dépendent pas directement des ouvriers, mais des constructions elles-mêmes. L'ouvrier ne dirige pas son travail, il est guidé par ce dernier. C'est à cette stimulation d'un type particulier que nous donnons le nom de stigmergie* » [49].

Vincent Chevrier consacre une partie de son mémoire à la description de son travail sur les interactions entre agents. Ce travail met en avant l'importance des interactions entre agents et le besoin de modéliser ces interactions [22].

Les buts des interactions entre agents peuvent être classés en deux catégories : coopération et compétition. Un agent peut en effet collaborer en vue de résoudre un but commun ou au contraire chercher à être le plus performant pour résoudre une tâche [105].

4.2 Les abeilles sociales

Quoi de plus simple qu'une abeille, quoi de plus complexe qu'une colonie d'abeille? et en particulier de l'abeille à miel? L'abeille noire (Apis mellifera mellifera) (En particulier l'abeille à miel ou l'abeille noire (Apis mellifera)), qui est une abeille à miel indigène de plus en plus domestique dans les ruches, de moins en moins sauvage en pleine nature, elle traverse avec difficultés les changements qui lui sont imposés par les bouleversements de son environnement :l'agriculture intensive et les mauvaises habitudes domestiques [1].

Le cerveau d'une abeille mesure moins de $1mm^3$, pèse autour d'un gramme et contient environ 900.000 neurones. Ce n'est pas bien lourd comparé au dm3 du cerveau humain, et surtout à ses 100 milliards de neurones, les performances de l'abeille à miel ne résident pas dans son individualité mais par sa vie sociale dans la ruche [50].

Les abeilles possèdent des propriétés assez différentes de celles des autres espèces d'insectes. Elles vivent en colonies, en construisant leurs nids dans des troncs d'arbre ou d'autres espaces clos similaires. Les abeilles sont des insectes sociaux qui forment des colonies permanentes, constituant ainsi un «super organisme» soudé par des relations complexes de travail entre frères, sœurs et mère. A priori, la vie de la colonie est infinie alors que la vie de l'individu est éphémère [50].

Une colonie d'abeilles est composée d'une reine, de quelques centaines de mâles et de 10.000 à 80.000 ouvrières. De ces trois genres d'abeilles à l'aspect très différent, deux sont des femelles, à savoir la reine et les ouvrières [1].

La reine est la seule femelle fertile pour chaque colonie et sa taille est beaucoup plus grande que celle des autres abeilles. Sa principale tâche est de pondre des œufs. En plus de ça, elle sécrète aussi des substances communicatives importantes qui maintiennent l'unité de la colonie et le bon fonctionnement des différents systèmes à l'intérieur de celle-ci, elle sécrète aussi des substances odorantes (phéromones) qui assurent la cohésion de la colonie; elle est la mère de tous les membres de la colonie, et influence donc directement les performances de cette dernière [53].

Les faux-bourdons sont plus gros que les femelles ouvrières, mais ils n'ont pas d'aiguillon et d'organes nécessaires à récolter leur propre nourriture. Leur unique fonction est de féconder la reine.

Les abeilles ouvrières qui constitue 95% soit 30 000 à 60 000 individus de la colonie toutes non fécondes, elle exécutent toutes les autres tâches que vous pouvez imaginer : nettoyer la ruche, nourrir la reine et les faux-bourdons, faire le miel, construire et entretenir les rayons, ventiler la ruche, récolter et emmagasiner des substances comme le nectar, le pollen, l'eau et la résine, et Assurer la sécurité de la ruche [53].

L'ordre à l'intérieur de la ruche, avec ses dizaines de milliers d'abeilles, est assuré par chacun des individus exécutant parfaitement ses tâches. Le scientifique allemand Gustav Rosch a cherché le secret de cet ordre, il a conclu que les tâches exécutées par les ouvrières dans la ruche dépendent de leur âge. Selon ces résultats, les abeilles ouvrières remplissent des rôles complètement différents durant les premières trois semaines de leur vie [50].

Mais l'âge n'est pas le seul facteur impliqué dans la détermination des tâches d'une abeille. Bien que chaque abeille ait ses responsabilités spécifiques, en cas d'urgence, les abeilles peuvent aussi modifier leurs fonctions instantanément. Cela est un énorme avantage dans une société aussi peuplée que la ruche. Si la distribution du travail entre les abeilles était contrainte par des règles fixes alors, dans le cas d'un événement imprévu, la colonie pourrait faire face à de graves difficultés. Par exemple, dans le cas d'une attaque d'une importance majeure, si seules les abeilles sentinelles participaient au combat et que les autres continuaient à exécuter leurs propres tâches, cela représenterait un sérieux danger pour la ruche. Cependant, ce qui se produit en réalité est qu'une grande partie de la colonie sauf celles qui montent la garde sur les autres entrées prend part à la défense et la sécurité devient une priorité immédiate [1].

La coordination repose sur la diffusion et l'échange d'informations entre les individus; les messages sont quasi essentiellement chimiques (phéromones), mais aussi tactiles et comportementaux (danse) [53].

Les phéromones émises par la glande de Nasanov (située à l'extrémité de l'abdomen des ou-vrières) permettent aux abeilles de battre le rappel et de se regrouper.

Chaque colonie possède son odeur propre et chaque abeille peut être identifiée par ses sœurs; cette odeur circule, au sein de la colonie par trophallaxie. Trophallaxie : Transferts "buccaux" réguliers des liquides sucrés entre les abeilles [1].

La prospection se fait plus ou moins au hasard mais aussi guidée par des signaux visuels (couleurs, formes) et chimiques (odeurs). Après découverte d'une source de nectar et/ou de pollen, la qualité, la quantité et différents paramètres de cette source sont mémorisés et transmis à d'autres butineuses après le retour à la ruche. Selon la richesse de la zone et les conditions de « miellée »; le butinage intéresse un rayon de quelques mètres à 2 ou 3 kilomètres (parfois plus).

La danse fournie aux membres de la colonie des informations sur la distance et la direction des sources de nourriture. L'information de direction est donnée par l'angle que forme la partie rectiligne de la danse avec la verticale : cet angle est identique à celui observé, à la sortie de la ruche, entre la position du soleil et celle de la source. L'information de distance est fournie par le frétillement de l'abdomen : plus le rythme est lent, plus la distance est grande.

4.3 Le cycle de vie des abeilles sociales

— Durant les trois premiers jours, elle joue le rôle de nettoyeuse et veille à la propreté des cellules. Sa deuxième mission est celle de nourricière, elle distribue la gelée royale à toutes les larves, qui donneront naissance aux jeunes abeilles, et aux reines et ce jusqu'au environ du dixième jour suivant sa naissance [50].

— Du 11ième au 20ième jours, les abeilles exécutentdes travaux de nettoyage, débarrassent la ruche des détritus, des cadavres de leurs sœurs. Elles vont aussi à la rencontre des butineuses rentrantes pour les décharger du nectar récolté en le disposant dans les alvéoles, et de s'occuper également du pollen ramené par leurs compagnes [50].

— Pendant la troisième phase les ouvrières magasinières procèdent encore à la construction des cellules de miel de réserve et celles des nymphes [50].

— Du 18ième au 21ième jour, elles deviennent les gardiennes en prenant part à la défense de la ruche et montent la garde au trou de vol à l'affût des pillards comme des bourdons, guêpes ou abeilles de ruches voisines. Elles communiquent grâce à ses antennes avec les abeilles qui entrent dans la ruche. Celles qui ne font pas partie de la colonie sont repoussées. Les voleuses de miel sont chassées à coup de dard [50].

La collecte est la dernière et plus longue tâche d'une ouvrière s'étalant du 21ème jour jusqu'à sa mort ou Elle part récolter le pollen et le nectar des fleurs pour la production de miel [50].

On peut se demander comment une abeille, noyée dans la foule des autres, parvient à connaître la tâche qui lui incombe. Journalièrement, la population s'accroît de quelque 2000 individus, et pourtant il y règne une entente et un ordre parfaits, les fonctions y sont remplies avec rigueur et précision, sans que le moindre frottement, la moindre rivalité se fassent ressentir. Tout est réglé, comme si nous nous trouvions en présence d'une sorte de cerveau collectif. Tel est l'instinct, connaissance innée, héréditaire, qui se manifeste ici par un comportement plus ou moins commun à tous les individus d'une société, mais non immuable, susceptible au contraire de réflexion et d'adaptation. Les abeilles ne sont pas des automates [50].

Comme toute organisation, la société des abeilles repose sur deux principes : la différentiation ou distribution du travail entre ses différents membres, et la coordination ou direction de toutes les facultés individuelles. Ainsi, dans une colonie d'abeilles, tous les individus sont tributaires les uns des autres, et sont incapables de subsister par eux-mêmes[50].

4.4 Le modèle de comportement de sécurité

Durant la quatrième phase de leurs vies, les abeilles ouvrières servent de gardes à l'entrée de la ruche en interdisant l'accès aux intrus. A ce stade, un changement physique se produit au niveau du corps de l'abeille ou leurs glandes à venin se développent et commencent à produire du venin [51].

Signalons que les gardiennes ne quittent jamais leurs postes, même dans le cas d'une défense offensive qui se réalise sur une autre entrée, les gardiennes des entrées ne rejoindront pas l'essaim car elles ne peuvent pas laisser une entrée sans surveillance ce qui sera une vulnérabilité et pourra être détectée et utilisée par l'intrus [51].

Toutes les abeilles se ressemblent énormément, pourtant les abeilles étrangères qui entrent dans la ruche sont immédiatement identifiées. Les scientifiques qui ont étudié la question à savoir comment les abeilles accomplissent cela en sont arrivés à de surprenantes conclusions :

L'odeur de la ruche est le plus important facteur qui permet aux abeilles de se reconnaître et de se distinguer les unes des autres. Celles qui n'ont pas l'odeur distinctive de la ruche représentent donc un danger et sont immédiatement expulsées ou tuées par les abeilles sentinelles. (Voir Figure 4.1).

FIGURE 4.1 – Défense offensive par piqûre –sacrifice (abeilles gardiennes entrain de tuer une abeille intrus)

les sentinelles font preuve d'une réaction énergique, en utilisant leurs aiguillons contre toute créature perçue comme n'appartenant pas à la ruche. Tout de suite après l'intervention initiale des gardes, les autres abeilles de la ruche se joignent généralement à l'attaque. Le signal qui amorce une attaque générale par les abeilles ouvrières de la ruche est une substance chimique (la phéromone) émise par les aiguillons des sentinelles attaquant l'intrus. Dans certains cas, en plus de l'émission des phéromones qui amorcent l'attaque, la posture caractéristique et le comportement agité des abeilles représentent également un signal d'alarme pour les autres abeilles de la ruche. Suite à la diffusion de la phéromone, des centaines d'abeilles se groupent à l'entrée de la ruche pour la protéger. Plus la phéromone émise par les gardes est forte, plus considérables sont l'excitation et l'agressivité des autres abeilles [51].

La contre-attaque menait par les abeilles s'arrêtera dés que l'intrus est mort ou loin du périmètre de la ruche, ce périmètre s'élève à quelque centaine de mètres .

Durant la période où ces abeilles ouvrières servent de gardes, elles mettent leur propre vie en danger en piquant l'agresseur car elles seront incapables de retirer leur aiguillon de l'agresseur ce qui endommage leurs organes internes et provoque leur mort en très peu de temps (Voir Figure 4.2) [51].

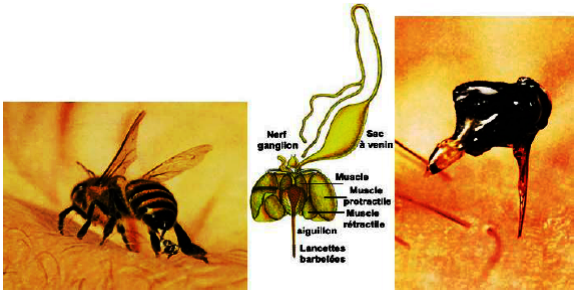


FIGURE 4.2 – Piqûre d'abeille

La défense de la ruche est une responsabilité majeure qui concerne la colonie en entier, surtout les abeilles gardiennes qui doivent l'assumer pleinement même au dépends de leurs vie.

Le comportement du sacrifice des abeilles réfute l'affirmation des évolutionnistes qu'il y a une "lutte pour la survie" dans la nature et que tous les êtres vivants cherchent à protéger seulement leur propre descendance. Ce sacrifice est une forme de comportement qui ne peut pas être expliquée par la thèse de la "lutte pour la survie" de la théorie de l'évolution. Les affirmations des évolutionnistes prennent la position que les êtres vivants luttent pour se protéger et survivre. Mais le fait est qu'il est inexact de dire que la nature consiste seulement en des individus guerriers, puisque les êtres vivants font considérablement preuve de comportements tels la coopération mutuelle et l'esprit de sacrifice. Comme réponse, certains évolutionnistes affirment que les êtres vivants se sacrifient pour assurer la continuité de leur progéniture, en d'autres mots que ceci représente un avantage pour eux. Cependant, cette affirmation contient un certain nombre d'inconsistances. Par exemple, les abeilles gardiennes attaquent et combattent des insectes comme les frelons, lesquels sont beaucoup plus gros qu'elles, sans un moment d'hésitation. L'affirmation que les abeilles le font représente un avantage pour la colonie [51].

Les abeilles gardiennes adoptent deux stratégies en réponse d'une intrusion, ses deux stratégies reposent sur le principe de la défense offensive :

— La première stratégie repose sur la poursuite et l'attaque de pique jusqu'à ce que l'intrus sorte du

périmètre et donc les abeilles ne le considère plus comme une attaques, cette stratégie est mortelle pour les abeilles puisque le mécanisme de piqûre blesse mortellement les abeilles.

— La deuxième stratégie, que les abeilles utilise pour ce défendre, est l'utilisation de la chaleur pour détruire l'ennemi ayant une vulnérabilité envers la chaleur : Les frelons japonais sont littéralement un cauchemar pour les abeilles européennes introduites au Japon. Une colonie de 30.000 abeilles européennes peut être tuée en approximativement en trois heures par environ 30 frelons qui occupent la ruche. Cette stratégie n'est pas mortelle aux abeilles en elles mêmes mais le fait de s'approcher de l'intrus peut causer des abeilles victimes surtout dans le début de processus d'attaque par la chaleur (Voir Figure 4.3) [51].

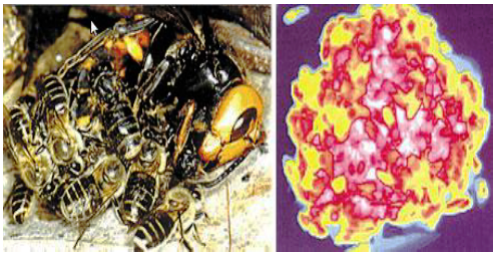


FIGURE 4.3 – Défense offensive utilisant la vulnérabilité de l'ennemi : chaleur

Dans le cas ou les abeilles gardienne ne réussissent pas à défendre la ruche et que l'intrus les bat, les abeilles qui n'ont pas pris part à la défense offensive délocalisent le maximum de nourriture et du contenu de ruche vers d'autre endroit. Elles optent même dans les pires cas à délocaliser la reine et les faux-bourdons en assurant l'escorte durant la délocalisation. Ceci dans un but d'assurer une continuité de vie de ruche et fondre une autre vie dans un emplacement plus sécurisé [51].

4.5 Le Modèle Informatique

Avant de détailler l'approche de notre contribution, nous devons d'abord décrire le modèle naturel du fonctionnement du système anti-intrusion à défense active des abeilles sociales et mettre la lumière sur les aspects qui nous ont orientés à choisir cette méta-heuristique. Cette superposition donne une idée préalable sur la faisabilité de cette modélisation(Voir Tableau 4.1).

	Abeille Sociale	Détection intrusion
Vise à sécurisé	La Ruche	Réseau
Elément important à sécurisé dans le pire cas	A. La reine B. Les faux-bourdons. C. La nourriture.	D. Les informations (fichiers, BDD…) E. Vie privé …
Stratégie de détection d'intrusion	• Reconnaissance d'odeur (identificateur des habitants de ruche produit par la reine). • Reconnaissance de comportement (abeille qui ne reçoit pas la substance royale)	• Par scénario • Par comportement
Stratégie de réponse à l'intrusion	Alerter les autres abeilles (demande de soutiens) F. Attaque et poursuite G. Utilisation vulnérabilité de l'ennemi : chaleur	Alerter + Réaction Passive ou Active
Type de processus	Permanant	Permanant
Tolérance aux pannes	Ne tolère pas la panne, dans le cas ou une gardienne quitte son poste, elle doit être remplacé immédiatement	Ne tolère pas la panne, si IDS se bug, on doit appliquer des directive selon la politique de sécurité

TABLEAU 4.1 – superposition entre le système de protection des abeilles sociales et IDS

Pour modéliser le comportement de sécurité chez les abeilles, nous avons imaginé une approche générale que nous avons scindé en trois sous approches, (Voir Tableau 4.2) :

1. La première sous approche, qui fait l'objet de notre contribution, modélise la détection d'intrusion par scénario, basée sur la reconnaissance d'odeur pour identifier les intrus à l'entrée de la ruche.
2. la deuxième sous approche modélise la détection d'intrusion par comportement à l'intérieur de la ruche, basée sur le comportement des abeilles et qui fera très prochainement l'objet d'un article.
3. La troisième sous approche est réservée à la réactivité offensive du système et qui finalisera l'approche générale en donnant de la robustesse au système.

		Comportement des abeilles sociales	Sécurité informatique
SCENARIO	**Etape 1** Authentification	La reine dépose une odeur sous forme d'une substance chimique sur toutes les abeilles de sa ruche : odeur de la ruche	Crée un Training set et le mettre à jour
	Etape 2 Détection d'intrusion (venant hors ruche/ venant de l'extérieur de réseau)	Les abeilles gardiennes montent une garde sur toutes les entrées de la ruche et ne laisse passer que les abeilles qui portent l'odeur de la ruche	Mettre un agent filtre par port de connexion.
COMPORTEMENT	**Etape 3** Détection d'intrusion (à l'intérieur de la ruche / à l'intérieur du réseau)	Les abeilles gardienne chasse et tue toute abeille qui a un comportement qui nuit au bon fonctionnement de la ruche	Mettre un agent de surveillance dans le réseau pour détecter les comportements malveillants
REACTION	**Etape 4** Réponse à l'intrusion (active 1)	Quand une abeille gardienne détecte une intrusion, elle attaque l'intrus, après sa mort une substance chimique est lancé pour incité les autre abeilles à attaquer l'intrus, la défense offensive continue jusqu'a ce qu'il ya plus de substance (pas de mort d'abeille donc intrus neutralisé)	L'agent qui détecte une intrusion lance une attaque DDoS contre l'intrus, et il la stoppe après le succès de cette opération.
	Etape 4 Réponse à l'intrusion (active 2)	L'abeille essaye de détecter les points faibles de l'intrus et l'utilise pour se défendre.	Les agents essayent plusieurs attaques afin de neutraliser l'intrus.
	Etape 5 Réponse à l'intrusion (fuite)	Les abeilles gardiennes délocalisent la reine, les males et la nourriture (gelée royale).	Les agents détruisent les informations importantes et exécute le plan d'évacuation.

TABLEAU 4.2 – Modélisation du comportement de sécurité chez les abeilles sociales

A fin d'étudier et d'implémenter chacune à part, ou la première représentant notre contribution et les deux autres feront l'objet de travaux futures. Une fois les trois sous approches achevées, nous les regrouperons en une seule approche pour constituer l'approche générale de notre IDSbees robuste. Une fois cet IDSbees installé sur le réseau, il constituera le blindage adéquat de notre « réseau-ruche » contre tout intrus et crée ainsi une phobie chez les hackers et les cybers criminels, et par conséquent personne n'osera s'aventurer à attaquer ce « réseau-ruche ».

Pour ce qui concerne notre travail, nous nous somme focalisé sur la première sous approche, c'est à dire la Reconnaissance d'odeur des abeilles (modèle naturel) représentant la détection d'intrusion par scénario (le modèle artificiel) mais selon une autre vision contraire à l'approche conventionnelle ou l'implémentation des systèmes de détection d'intrusion par scénario (mono ou multi agents) responsabilise l'agent à détecter toutes les intrusions sur sur l'ensemble des ports du réseau, chose qui est très difficile à réalisé d'une manière efficace, car à un moment donné, un ou plusieurs ports peuvent se trouver sans protection et feront donc l'objet d'une vulnérabilité. Donc selon notre vision, chacune des abeilles est responsable de la sécurité de son port et exécutent leurs taches parallèlement.

4.5.1 Tableau de modélisation : le passage du modèle naturel au modèle artificielle (Voir Tableau 4.3)

Le passage du modèle naturel au modèle artificiel est résumé par le tableau de modélisation décrit ci-dessous, montrant les correspondances entre les deux systèmes (modèle naturel vers le modèle artificiel) (Voir Tableau 4.3).

Système de détection d'intrusion basé sur les abeille social ; IDSBees		Modèle Naturelle	Modèle Artificielle
		Ruche	Réseau
	E	Entrée de ruche	port
	A	Reine	Donnée à caractère sensible
	B	faux-bourdons	Applications
	C	Ouvrière	Services système
	D	Ouvrière Gardien	Agent de garde : service de sécurité
	F	Remplacement direct des abeilles gardiennes qui quittent leurs postes	Définition de nombre d'agent gardien inactif
	G	Reconnaissance d'odeur (reconnu, non-reconnu)	Training Set (intrusion, non intrusion)
	H	Reconnaissance de comportement (normal, anormal)	Training set (normal, anormal)
	I	Nourriture : Miel, Gelée Royale et pollen	Ressources (CPU, Mémoire…)
	J	Poursuite et attaque avec aiguillon avec venin	Défense offensif : Attaque DDoS
	K	Le sacrifice de soi	Dead to thread pour libérer les ressources
	L	Elimination d'intrus : mort ou hors périmètre danger	Fin attaque DDoS : pas de réponse de ping
	M	Augmentation / diminution d'Agressivité	Utilisation plus/moins de serveur ping (DDoS)
	N	Gestion stock miel et tue des faux-bourdons à l'hiver	Priorité au système de sécurité par rapport les applications dans la consommation des ressources dans les cas critique
	O	Attaque selon vulnérabilité d'intrus	Différent Attaque selon les points faibles d'intrus
	P	Délocalisation de Reine, faux-bourdons, ouvrière, ouvrière gardienne et nourriture : quitter la ruche	l'interruption d'une connexion
	Q	Maintenir l'ordre dans la ruche	Surveillance de comportement des utilisateurs

TABLEAU 4.3 – Passage du modèle naturelle au modèle artificielle

Dans les approches conventionnelles, la construction du modèle d'intrusion se faisait à partir d'un Training set généralisé englobant toutes les connexions (intrusion et non-intrusion) sur tous les ports du réseau. Générant ainsi un seul modèle d'apprentissage d'intrusion qui peut contenir un déséquilibre entre les ports, réduisant ainsi la sécurité et la fiabilité du système de détection d'intrusion puisqu'il ne prend pas en considération toutes les intrusions possibles sur tous les ports du réseau.

Selon l'approche que nous proposons, le passage du modèle naturel au modèle artificiel est décrit dans le tableau et la figure ci-dessous (Voir Tableau 4.4) (Voir Figure 4.4).

	Modèle Naturel	Modèle Artificiel
	Ruche	Réseau
E	Entrée de ruche	port
A	Reine	Les informations sensibles
B	Mâles	Les applications
D	Abeilles gardiennes	Agents gardien: service de sécurité
F	Reconnaissance d'odeur	Training Set (intrusion, no-intrusion)
G	Responsabilités limitées pour chaque abeille gardienne	Division of Training set (par port)

TABLEAU 4.4 – Modèle naturel vs Modèle Artificiel

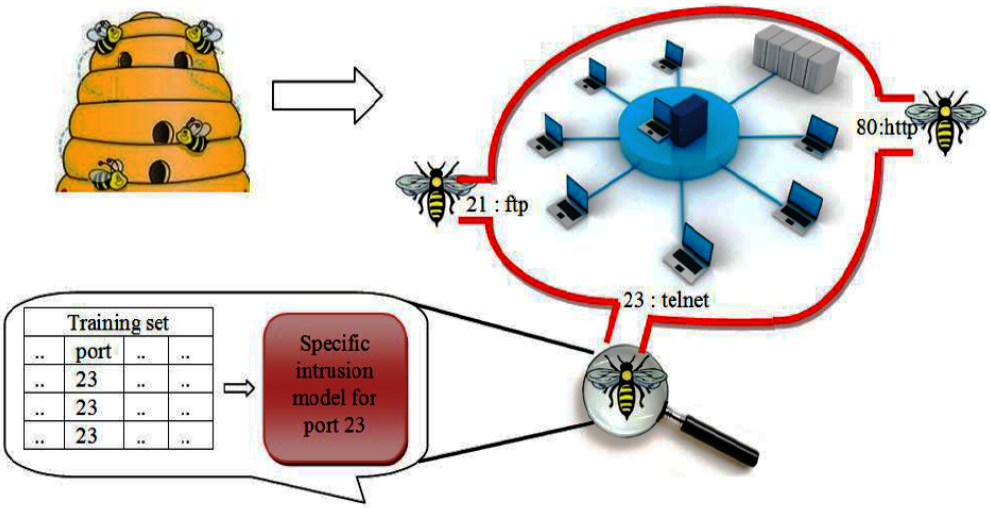


FIGURE 4.4 – Modèle artificiel d'IDS par scénario inspiré du système de protection des abeilles sociales

Par opposition à l'approche conventionnelle, notre modèle attribut à chaque agent une seule entrée (port réseau) qu'il doit protéger et ne le déserter à aucun moment et doit attendre l'appui des autres abeilles, ceci pour assurer la garde de l'entrée d'une manière permanente. De cette façon, nous garantissons la continuité du service sur chaque entrée et minimiser le risque d'être piégés par une désorientation de notre système de détection d'intrusion.

Cette modélisation va être expliqué en détail dans l'élaboration du modèle artificiel dans la section suivante :

4.5.2 État initial

Selon le tableau ci-dessus notre système de détection d'intrusion IDSbees sera implémenté après avoir défini clairement ;

— (A) Les données caractères sensibles qui représenteront la reine,

— (B) Les applications à priorité élevée

— (D) et (E) Les ports à surveiller ; dans chaque port un agent gardien montra la garde. Le nombre des agents gardiens actifs est égal au nombre du port en plus des nombres des gardiens qui maintient l'ordre l'intérieur de la ruche.

— (F) Le nombre des agents gardiens inactifs ne doit pas être inférieur à deux fois le nombre des agents gardiens actifs.

— (G) Démarrer avec une base d'apprentissage initiale (Benchmark ou des fichiers logs du réseau), sachant qu'un des attributs de la base est nommé 'Service' représentant le numéro de port de la

connexion. Nous lancerons deux simples algorithmes de parcours sur la base d'apprentissage initiale : le premier calcule la cardinalité de l'attribut'Service' qui représente le nombre de ports N du réseau et le second détermine la liste représentant toutes les valeurs possibles de l'attribut 'Service' ($V_1, V_2, ..., V_N$) représentants les ports à surveiller. Ensuite nous construisons les N sous bases d'apprentissages à partir de la base d'apprentissage initiale que nous nommons dans le reste du chapitre base d'apprentissage spécialisée, par une sélection faite sur l'attribut '$Service' = V_1, V_2, ..., V_N$. C'est-à-dire, qu'on aura à la fin N bases d'apprentissages spécialisés ayant la même configuration d'attributs que la base d'apprentissage initiale où la valeur de l'attribut 'Service' est la même dans chaque base d'apprentissage spécialisée en question. Chacune de ces sous bases est affectée à l'abeille responsable du port en question (valeur de l'attribut 'Service').

4.5.3 État d'activité

Étape 1 : Construction du modèle d'apprentissage

Au début de l'utilisation du système de détection d'intrusion IDSBees, chaque agent gardien posté sur un port, construit un modèle d'apprentissage d'intrusion en appliquant plusieurs algorithmes de Data Mining, ayant une performance élevée et reconnu pour les classifications bi-classe, sur la base d'apprentissage qui leur a été attribué lors de l'état initial.

Les algorithmes choisis sont :

— Le Naïve Bayes : a cause de ses performances qui sont fortement et négativement corrélées avec le nombre de classes (tant que le nombre de classes diminue les performances de classifieur augmente et vice-versa). Aussi, des avantages probabilistes qu'il offre.

— Les arbres de décision : a cause du gain d'information et les splits d'information qu'ils nous offrent, en plus de l'aspect visuel du modèle de classification.

— Les SVM : a cause de la souplesse qu'ils nous offrent pour la détection d'intrusion par le biais de la marge, et la notion de similarité qu'ils utilisent lors de la génération du modèle d'apprentissage.

Après l'application de ces trois algorithmes classiques de Data mining sur chaque base d'apprentissage spécialisée. Nous auront donc trois modèles de classification différents pour chaque agent. L'attribution d'une classe à une nouvelle connexion entrante par l'agent responsable du port qu'elle utilise sera faite par la méthode de Baging, qui est un vote à majorité simple, appliquée sur ses trois modèles générés au par avant.

Après évaluation, pour entamer une future modélisation et afin d'éviter un sur apprentissage (overfitting), l'agent gardien ayant effectué cette affectation doit calculer la similarité de ce nouveau cas avec les cas existant dans la base d'apprentissage qu'il détient. Si aucune ressemblance n'est trouvée bien entendu selon un seuil de similarité qui doit être suffisamment grand pour éviter des mise à jours fréquentes de la base d'apprentissage qui causent d'énormes calculs suite à la génération du nouveau modèle d'apprentissage. L'agent doit remplacer ce cas par un des éléments (les plus redondant) de la base d'apprentissage du port en question. Cet élément est choisi à partir d'un calcul de similarité appliqué à tous les éléments de cette base d'apprentissage de même classe (intrus ou non intrus) que la connexion entrante, ceci pour éviter un déséquilibre de la base d'apprentissage. Ensuite régénérer ses trois modèles de classification en appliquant les trois algorithmes de classification citée auparavant sur la nouvelle base d'apprentissage.

Le schéma général de l'algorithme de notre contribution (système de détection d'intrusion par scénario) est le suivant :

4.5.4 Algorithme général de l'approche

1. Calcule de nombre de port N de la détermination de la liste de ports à surveiller.
2. Création des N Training set.
3. Mettre une abeille (agent filtre) dans chaque entrée de la ruche (port de connexion) en lui attribuant le training set spécialisé de port qu'elle surveille.

4. Génération de modèle de bagging basé sur trois modèle d'apprentissage : Naive Bayes, SVM et arbre de décision (C4.5) pour chaque abeille (agent gardien posté sur un port)
5. Tester et évaluer le modèle par une base de test généralisé
6. Calcul de la matrice de confusion et les mesures d'évaluation pour chaque port.
7. Remplacement d'exemple dans la base d'apprentissage et régénération des modèle d'apprentissage pour éviter le sur-apprentissage

Algorithme 1 : système de détection d'intrusion IDSbees par scénario

```
Entier Nombre.abeille.gardien= Nombre de ports à sécurisé.
Entier seuil. for Chaque Port i do
    Générer la base d'apprentissage spécialisé au Port i.
    sous-base-d'apprentissage(Port − i) = Sélection(base-d'apprentissage-générale, Port i).
    //Générer les modèle d'apprentissage.
    M1(Port i)=Générer(sous-base-d'apprentissage(Port i), NaiveBayes).
    M2(Port i)=Générer (sous-base-d'apprentissage(Port i), K_Means(2)).
    M3(Port i)=Générer (sous-base-d'apprentissage(Port i), C4.5).
    //Modèle générale d'apprentissage du Port i.
    Modèle-générale (Port i) = Baging(M1(Port i)+M2(Port i)+M3(Port i)).
end
for chaque connexion de base de test j do
    X =Déterminer le port de la connexion j. Classification-modèle-1= Classification (connexion
      j, M1(X)).
    Classification-modèle-2= Classification (connexion j, M2(X)).
    Classification-modèle-3= Classification (connexion j, M3(X)).
    Classification-final = Modèle-générale(Baging(Classification-modèle-1,Classification-modèle-
      2,Classification-modèle-3)).
    Calculer similarité (sous-base-d'apprentissage(X), nouvelle connexion,Classification-final).
    if similarité > seuil then
        exemple.sortant =Déterminer l'exemple sortant(sous-base-d'apprentissage(X)).
        Remplacer (sous-base-d'apprentissage(X), exemple.sortant).
        //Régénérer les modèle d'apprentissage.
        M1(X)=Générer (sous-base-d'apprentissage(X), NaiveBayes).
        M2(X)=Générer (sous-base-d'apprentissage(X), K_Means(2)).
        M3(X)=Générer (sous-base-d'apprentissage(X), C4.5).
        //Modèle générale d'apprentissage du port i.
        Modèle-générale (X) = Baging(M1(X)+M2(X)+M3(X)).
    end
end
Calculer la matrice de confusion et mesures d'évaluation pour chaque port.
```

4.6 Expérimentations et résultats

Pour ce faire une idée sur la démarche entreprise, pour l'élaboration de notre modèle de détection d'intrusion par scénario afin de valider notre méthode d'abeille sociale sur la base KDD formée à peu près de 5 millions d'instances de données orientées détection d'intrusion, ou chaque instances caractérise une connexion et contient 41 attributs différents, tel que la durée de la connexion (nombre de seconde) ; service réseau (destination) ex : http, Telnet ; statut de la connexion (normale ou anormale) ; etc. nous présentons les étapes suivantes (Voir Figure 4.5) :

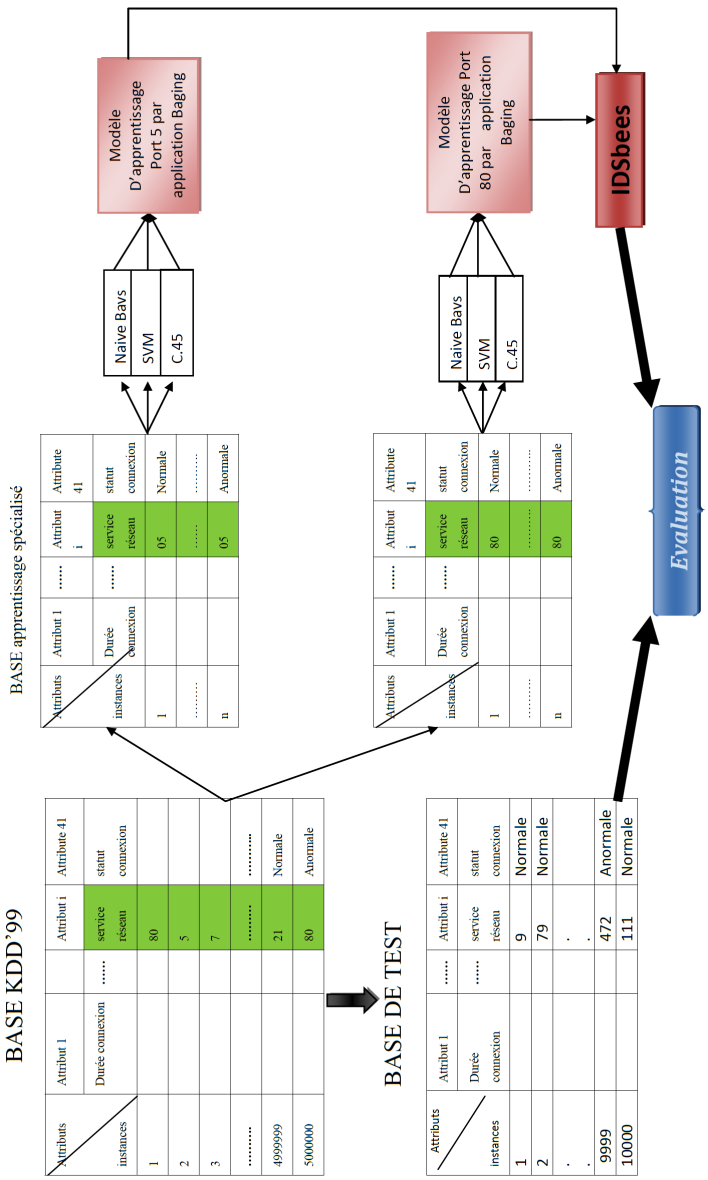


FIGURE 4.5 – Schéma général de notre approche

1. Éclater la base KDD considérée comme base d'apprentissage initiale en 51 sous bases d'apprentissage spécialisées dans notre cas ou dans chaque sous base la valeur de l'attribut 'service' est la même pour toutes les instances de la sous base.
2. Appliquer sur chaque sous base d'apprentissage spécialisée relative à un port du réseau trois algorithme de classification tel que : Naives Bayes, SVM, et arbres de décision (C4.5). Nous aurons donc trois modèles différents d'apprentissage.
3. Appliquer un Baging sur les trois modèles pour chaque sous base d'apprentissage spécialisée.
4. Former la base de test générale à partir de la base d'apprentissage initiale KDD.
5. Faire l'évaluation du modèle d'apprentissage cité au point 3 avec la base de test cite au point 4.

4.6.1 Corpus utilisé

Depuis 1999, KDD'99 est le corpus le plus utilisé pour l'évaluation des anomalies et la détection d'intrusion (Benchmark). Ce corpus a été construit par Stolfo et al [113], à partir de la base des données collectées par le programme d'évaluation des systèmes de détection d'intrusion DARPA'98. La taille de DARPA'98 est d'environ 4 gigaoctets de (binaires) relevés sur 7 semaines de trafic réseau, constitué d'environ 5 millions d'enregistrements de connexion d'environ 100 octets chacune.

Le KDD'99 est constitué d'un ensemble de données d'environ 4.900.000 vecteurs de connexion unique dont chacune contient 41 colonnes dont une est étiqueté comme normal ou une attaque, avec exactement un type d'attaque spécifique.

Les attaques peuvent être classées dans l'une des quatre catégories suivantes :

1. Attaques par déni de service (DoS).
2. Attaque U2R.
3. Attaque R2L.
4. Attaque probing.

Il est important de noter que les données de la base de test incluant des types d'attaques spécifiques qui ne figure pas dans le training set. Certains experts d'intrusion croient que la plupart des nouvelles attaques sont des variantes d'attaques connues. Le training set contient 24 types d'attaque, avec 14 types supplémentaires dans la base de test. Le nom et description détaillée des types d'attaque sont répertoriés dans le travail de chercheur Lippmann[77].

Les fonctionnalités de KDD99 peuvent être classées en trois groupes :

— Caractéristiques de base : cette catégorie englobe tous les attributs qui peuvent être extraites à partir d'une connexion TCP / IP.

— Caractéristique de trafique de réseau : cette catégorie inclut des fonctionnalités qui sont calculées par rapport à un intervalle de fenêtre et est divisé en deux groupes :

  — caractéristiques "même hôte"

  — caractéristiques "même service"

— Caractéristique de contenu.

4.6.2 Outils d'évaluation

La matrice de confusion, dans la terminologie de l'apprentissage supervisé, est un outil servant à mesurer la qualité d'un système de classification (Voir Tableau 4.5). Chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle (ou de référence). Un des intérêts de la matrice de confusion est qu'elle montre rapidement si le système parvient à classifier correctement, c'est-à-dire lorsque les VP et VN doivent être maximales alors automatiquement les FN et FP sont minimales.

TABLEAU 4.5 – Matrice de confusion (IDS)

	Classé comme intrusion	Classé comme non-intrusion
Réellement Intrusion	VP	FN
Réellement Non-Intrusion	FP	VN

De cette matrice nous calculerons toutes les métriques d'évaluation suivante : Rappel, Précision, F-mesure, Kappa statistique, Entropie et Précision dont les formules sont citées auparavant dans le chapitre III.

4.6.3 Résultats et discussions

Voici ci-dessous les tableaux des résultats (Voir du Tableau 4.6 jusqu'a Tableau 4.22).

Port n° 0 : OTHER			Port n° 5 : RJE			Port n° 7 : Echo		
1067	415	Précision = 0.687	1031	144	Précision = 0.844	1193	87	Précision =0. 937
344	674	Rappel = 0.691	248	1077	Rappel = 0.845	68	1152	Rappel = 0.938
F-Mesure= 0.689			F-Mesure= 0.844			F-Mesure= 0.938		
Entropie= 0.257		Kappa statistc=0.377	Entropie= 0.143		Kappa statistc=0.686	Entropie= 0.061		Kappa statistc=0.875
Robustesse=0.694			Robustesse= 0.843			Robustesse= 0.938		

TABLEAU 4.6 – Résultats de détection d'intrusion sur les ports 0, 5 et 7

Port n° 9 : DISCARD			Port n° 11 : SYSTAT			Port n° 13 : DAYTIME		
1009	232	Précision = 0.759	1051	144	Précision = 0.829	1101	184	Précision = 0.824
377	882	Rappel = 0.756	291	1014	Rappel = 0.828	257	958	Rappel = 0.822
F-Mesure= 0.758			F-Mesure= 0.828			F-Mesure= 0.823		
Entropie= 0.208		Kappa statistc=0.513	Entropie= 0.155		Kappa statistc=0.653	Entropie= 0.158		Kappa statistc=0.646
Robustesse=0.756			Robustesse= 0.826			Robustesse= 0.823		

TABLEAU 4.7 – Résultats de détection d'intrusion sur les ports 9, 11 et 13

Port n° 15 : NETSTAT			Port n° 21 : FTP			Port n° 22 : SSH		
995	191	Précision = 0.765	1257	64	Précision = 0.942	1248	151	Précision = 0.869
413	901	Rappel = 0.762	81	1098	Rappel = 0.941	173	928	Rappel = 0.867
F-Mesure= 0.764			F-Mesure= 0.941			F-Mesure= 0.868		
Entropie= 0.204		Kappa statistc=0.519	Entropie= 0.056		Kappa statistc=0.883	Entropie= 0.121		Kappa statistc=0.736
Robustesse=0.758			Robustesse= 0.942			Robustesse= 0.870		

TABLEAU 4.8 – Résultats de détection d'intrusion sur les ports 15, 21 et 22

Port n° 23 : TELNET			Port n° 25 : SMTP			Port n° 35 : PRINTER		
1110	101	Précision = 0.842	1178	121	Précision = 0.919	943	258	Précision = 0.737
315	974	Rappel = 0.836	79	1122	Rappel = 0.920	405	894	Rappel = 0.736
F-Mesure= 0.839			F-Mesure= 0.920			F-Mesure= 0.737		
Entropie= 0.144		Kappa statistc=0.668	Entropie= 0.076		Kappa statistc=0.839	Entropie= 0.224		Kappa statistc=0.734
Robustesse=0.833			Robustesse= 0.920			Robustesse= 0.471		

TABLEAU 4.9 – Résultats de détection d'intrusion sur les ports 23, 25 et 35

Port n° 42 : HOSTNAMES			Port n° 53 : DNS			Port n° 56 : AUTH		
1144	151	Précision = 0.882	1055	157	Précision = 0.858	1288	44	Précision = 0.968
143	1062	Rappel = 0.882	198	1090	Rappel = 0.858	35	1133	Rappel = 0.968
F-Mesure= 0.882			F-Mesure= 0.858			F-Mesure= 0.968		
Entropie= 0.110		Kappa statistc=0.764	Entropie= 0.131		Kappa statistc=0.716	Entropie= 0.031		Kappa statistc=0.936
Robustesse=0.882			Robustesse= 0.858			Robustesse= 0.967		

TABLEAU 4.10 – Résultats de détection d'intrusion sur les ports 42, 53 et 56

Port n° 63 : WHOIS			Port n° 69 : TFTP			Port n° 70 : GOPHER		
1197	81	Précision = 0.908	1219	50	Précision = 0.951	987	272	Précision=0.802
153	1069	Rappel = 0.908	72	1159	Rappel = 0.951	223	1018	Rappel = 0.802
F-Mesure= 0.908			F-Mesure= 0. 951			F-Mesure= 0.802		
Entropie= 0.087		Kappa statistc=0.812	Entropie= 0.047		Kappa statistc=0.902	Entropie= 0.176		Kappa statistc=0.604
Robustesse=0.906			Robustesse=0.951			Robustesse=0.803		

TABLEAU 4.11 – Résultats de détection d'intrusion sur les ports 63, 69 et 70

Port n° 79 : FINGER			Port n° 80 : HTTP			Port n° 84 : CTF		
1018	313	Précision = 0.783	1235	30	Précision = 0.957	927	200	Précision= 0.828
228	941	Rappel = 0.784	79	1156	Rappel = 0.956	225	1148	Rappel =0.0.829
F-Mesure= 0.784			F-Mesure= 0. 956			F-Mesure= 0.0.828		
Entropie= 0.191		Kappa statistc=0.567	Entropie= 0.041		Kappa statistc=0.912	Entropie= 0.156		Kappa statistc=0.657
Robustesse=0.783			Robustesse=0.956			Robustesse= 0.83		

TABLEAU 4.12 – Résultats de détection d'intrusion sur les ports 79, 80 et 84

Port n° 87 : PRIVATE			Port n° 95 : SUPDUP			Port n° 102 : ISO_TSAP		
1214	160	Précision=0.880	1116	271	Précision = 0.837	1061	163	Précision=0.823
134	992	Rappel = 0.882	135	978	Rappel = 0.841	286	990	Rappel = 0.821
F-Mesure= 0.881			F-Mesure= 0. 839			F-Mesure= 0.822		
Entropie= 0.111		Kappa statistc=0.762	Entropie= 0.148		Kappa statistc=0.675	Entropie=0.161		Kappa statistc=0.641
Robustesse=0.882			Robustesse=0.837			Robustesse=0.821		

TABLEAU 4.13 – Résultats de détection d'intrusion sur les ports 87, 95 et 102

Port n° 105 : CSNET_NS			Port n° 109 : POP_2			Port n° 110 : POP_3		
1012	242	Précision=0.823	1145	126	Précision=0.834	1198	136	Précision=0.867
199	1047	Rappel = 0.823	306	923	Rappel = 0.825	199	967	Rappel = 0.863
F-Mesure= 0.823			F-Mesure= 0. 830			F-Mesure= 0.865		
Entropie= 0.159		Kappa statistc=0.647	Entropie= 0.151		Kappa statistc=0.653	Entropie=0.123		Kappa statistc=0.729
Robustesse=0.823			Robustesse=0.827			Robustesse=0.866		

TABLEAU 4.14 – Résultats de détection d'intrusion sur les ports 105, 109 et 110

Port n° 111 : SUNRPC			Port n° 117 : UUCP			Port n° 119 : NNTP		
1158	95	Précision =0.865	1058	143	Précision = 0.839	1147	171	Précision=0.856
256	991	Rappel = 0.859	264	1035	Rappel = 0.838	188	994	Rappel = 0.855
F-Mesure= 0.862			F-Mesure= 0. 839			F-Mesure= 0.855		
Entropie= 0.124		Kappa statistc=0.719	Entropie= 0.146		Kappa statistc=0.837	Entropie=0.132		Kappa statistc=0.711
Robustesse=0.859			Robustesse=0.675			Robustesse=0.856		

TABLEAU 4.15 – Résultats de détection d'intrusion sur les ports 111, 117 et 119

Port n° 123 : NTP_U			Port n° 137 : NETBIOS_NS			Port n° 138 : NETBIOS_DGM		
1187	121	Précision= 0.896	1264	72	Précision = 0.934	1123	197	Précision=0.836
139	1053	Rappel = 0.895	91	1073	Rappel = 0.933	211	969	Rappel = 0.835
F-Mesure= 0.895			F-Mesure= 0. 934			F-Mesure= 0.836		
Entropie= 0.098		Kappa statistc=0.791	Entropie= 0.062		Kappa statistc=0.868	Entropie=0.149		Kappa statistc=0.672
Robustesse=0.896			Robustesse=0.934			Robustesse=0.836		

TABLEAU 4.16 – Résultats de détection d'intrusion sur les ports 123, 137 et 138

Port n° 139 : NETBIOS_SSN			Port n° 143 : IMAP4			Port n° 150 : SQL_NET		
1243	122	Précision=0.901	1122	233	Précision=0. 841	1095	153	Précision=0.836
121	1014	Rappel = 0.902	162	983	Rappel =0. 843	263	989	Rappel = 0.833
F-Mesure= 0.901			F-Mesure= 0. 842			F-Mesure= 0.834		
Entropie= 0. 093		Kappa statistc=0.803	Entropie= 0.145		Kappa statistc=0.683	Entropie=0.149		Kappa statistc=0.667
Robustesse=0.902			Robustesse=0.842			Robustesse=0.833		

TABLEAU 4.17 – Résultats de détection d'intrusion sur les ports 139, 143 et 150

Port n° 165 : COURIER			Port n° 175 : VMNET			Port n° 179 : BGP		
1202	76	Précision=0.916	1028	245	Précision=0.798	1074	136	Précision=0.869
137	1085	Rappel = 0.914	258	969	Rappel = 0.798	192	1098	Rappel = 0.869
F-Mesure= 0.915			F-Mesure= 0. 798			F-Mesure= 0.869		
Entropie= 0.081		Kappa statistc=0.829	Entropie= 0.179		Kappa statistc=0.597	Entropie=0.121		Kappa statistc=0.737
Robustesse=0.914			Robustesse=0.798			Robustesse=0.868		

TABLEAU 4.18 – Résultats de détection d'intrusion sur les ports 165, 175 et 179

Port n° 194 : IRC			Port n° 196 : RED_I			Port n° 209 : MTP		
1159	185	Précision= 0.838	1176	174	Précision= 0.861	1118	187	Précision=0.841
218	938	Rappel = 0.836	171	979	Rappel = 0.861	208	987	Rappel = 0.841
F-Mesure= 0.837			F-Mesure= 0. 861			F-Mesure= 0.841		
Entropie= 0.147		Kappa statistc=0.675	Entropie= 0.128		Kappa statistc=0.722	Entropie= 0.144		Kappa statistc=0.683
Robustesse=0.838			Robustesse=0.862			Robustesse=0.842		

TABLEAU 4.19 – Résultats de détection d'intrusion sur les ports 194, 196 et 209

Port n° 210 : Z39_50			Port n° 245 : LINK			Port n° 387 : URP_I		
1108	154	Précision=0.861	1155	180	Précision= 0.872	1079	146	Précision=0.846
194	1044	Rappel = 0.860	136	1029	Rappel = 0.874	241	1034	Rappel = 0.845
F-Mesure= 0.860			F-Mesure= 0. 873			F-Mesure= 0.846		
Entropie= 0.128		Kappa statistc=0.721	Entropie= 0.118		Kappa statistc=0.746	Entropie=0.141		Kappa statistc=0.691
Robustesse=0.861			Robustesse=0.873			Robustesse=0.845		

TABLEAU 4.20 – Résultats de détection d'intrusion sur les ports 210, 245 et 387

Port n° 389 : LDAP			Port n° 407 : TIM_I			Port n° 433 : NNSP		
1187	112	Précision=0.868	1121	157	Précision=0.855	1218	120	Précision=0.892
225	976	Rappel = 0.863	205	1017	Rappel = 0.854	149	1013	Rappel = 0.891
F-Mesure= 0.866			F-Mesure= 0. 855			F-Mesure= 0.891		
Entropie= 0.122		Kappa statistc=0.729	Entropie= 0.133		Kappa statistc=0.711	Entropie=0.101		Kappa statistc=0.783
Robustesse=0.865			Robustesse=0.855			Robustesse=0.892		

TABLEAU 4.21 – Résultats de détection d'intrusion sur les ports 389, 407 et 433

Port n° 472 : KLOGIN			Port n° 491 : LOGIN			Port n° 495 : ECO_I		
1169	71	Précision =0.904	1142	114	Précision = 0.901	1097	338	Précision=0.821
174	1086	Rappel = 0.902	133	1111	Rappel = 0.901	114	951	Rappel = 0.828
F-Mesure= 0.903			F-Mesure= 0. 901			F-Mesure= 0.825		
Entropie= 0.091		Kappa statistc=0.804	Entropie= 0.093		Kappa statistc=0.802	Entropie= 0.161		Kappa statistc=0.641
Robustesse=0.902			Robustesse=0.901			Robustesse=0.819		

TABLEAU 4.22 – Résultats de détection d'intrusion sur les ports 472, 491 et 495

— **En jaune** : la performance de l'agent est bonne, le port est sécurisé ;

— **En vert** : la performance de l'agent est très bonne, le port est très bien sécurisé ;

— **En rouge** : la performance de l'agent est mauvaise, le port n'est pas sécurisé et peut être une source d'attaque, car il est vulnérable.

Si nous supposons que la performance de notre système est égale à la moyenne de la performance de tous les agents , nous aurons le résultat global (Voir Tableau 4.23).

Globale : 127.500 instance		
TP=57.429	**FN=10.116**	**Précision = 0.856**
FP=8.190	**TN=51.765**	**Rappel = 0.855**
F-Mesure= 0.856		
Entropie= 0.132		**Kappa statistc= 0.712**
Robustesse= 0.856		

TABLEAU 4.23 – Résultats général de notre approche de IDS

Selon les résultats des tableaux 4.6 jusqu'a 4.22, nous pouvons voir clairement que chacun des agents est autonomes et indépendant des autres, il a sa propre base d'apprentissage spécialisée relative au port dont il est responsable de sa sécurité, il génère ses propres modèles d'intrusion et prend sa propre décision d'affectation des classes (intrusion ou non-intrusion) en utilsant le Baging.

Le tableau suivant (Voir Tableau 4.24), récapitule les résultats.

	N° port correspondant A la valeur de l'attribut 'service'	VP	FN	FP	VN	Précision	Rappel	F-Mesure	Entropie	Robust	Kappa statistc
1	Port n° 0 : OTHER	1067	415	344	674	0.687	0.691	0.689	0.257	0.694	0.377
2	Port n° 9 : DISCARD	1009	232	377	882	0.759	0.756	0.758	0.208	0.756	0.513
3	Port n° 15 : NETSTAT	995	191	413	901	0.765	0.762	0.764	0.204	0.758	0.519
4	Port n° 35 : PRINTER	943	258	405	894	0.737	0.736	0.737	0.224	0.471	0.734
5	Port n° 79 : FINGER	1018	313	228	941	0.783	0.784	0.784	0.191	0.783	0.567
6	Port n° 175 : VMNET	1028	245	258	969	0.798	0.798	0.798	0.179	0.798	0.597
7	Port n° 7 : ECHO	1193	87	68	1152	0. 937	0.938	0.938	0.061	0.938	0.875
8	Port n° 21 : FTP	1257	64	81	1098	0.942	0.941	0.941	0.056	0.942	0.883
9	Port n° 25 : SMTP	1178	121	79	1122	0.919	0.920	0.920	0.076	0.920	0.839
10	Port n° 56 : AUTH	1288	44	35	1133	0.968	0.968	0.968	0.031	0.967	0.936
11	Port n° 63 : WHOIS	1197	81	153	1069	0.908	0.908	0.908	0.087	0.906	0.812
12	Port n° 69 : TFTP	1219	50	72	1159	0.951	0.951	0.951	0.047	0.951	0.902
13	Port n° 80 : HTTP	1235	30	79	1156	0.957	0.956	0.956	0.041	0.956	0.912
14	Port n° 137 : NETBIOS_NS	1264	72	91	1073	0.934	0.933	0.934	0.062	0.934	0.868
15	Port n° 139 : NETBIOS_SSN	1243	122	121	1014	0.901	0.902	0.901	0. 093	0.902	0.803
16	Port n° 165 : COURIER	1202	76	137	1085	0.916	0.914	0.915	0.081	0.914	0.829
17	Port n° 472 : KLOGIN	1169	71	174	1086	0.904	0.902	0.903	0.091	0.902	0.804
18	Port n° 491 : LOGIN	1142	114	133	1111	0.901	0.901	0.901	0.093	0.901	0.802
19	Port n° 22 : SSH	1248	151	173	928	0.869	0.867	0.868	0.121	0.870	0.736
20	Port n° 53 : DNS	1055	157	198	1090	0.858	0.858	0.858	0.131	0.858	0.716
21	Port n° 87 : PRIVATE	1214	160	134	992	0.880	0.882	0.881	0.111	0.882	0.762
22	Port n° 110 : POP_3	1198	136	199	967	0.867	0.863	0.865	0.123	0.866	0.729
23	Port n° 111 : SUNRPC	1158	95	256	991	0.865	0.859	0.862	0.124	0.859	0.719
24	Port n° 119 : NNTP	1147	171	188	994	0.856	0.855	0.855	0.132	0.856	0.711
25	Port n° 123 : NTP_U	1187	121	139	1053	0.896	0.895	0.895	0.098	0.896	0.791
26	Port n° 179 : BGP	1074	136	192	1098	0.869	0.869	0.869	0.121	0.868	0.737
27	Port n° 196 : RED_I	1176	174	171	979	0.861	0.861	0.861	0.128	0.862	0.722
28	Port n° 210 : Z39_50	1108	154	194	1044	0.861	0.860	0.860	0.128	0.861	0.721
29	Port n° 245 : LINK	1155	180	136	1029	0.872	0.874	0.873	0.118	0.873	0.746
30	Port n° 389 : LDAP	1187	112	225	976	0.868	0.863	0.866	0.122	0.865	0.729
31	Port n° 433 : NNSP	1218	120	149	1013	0.892	0.891	0.891	0.101	0.892	0.783
32	Port n° 5 : RJE	1031	144	248	1077	0.844	0.845	0.844	0.143	0.843	0.686
33	Port n° 11 : SYSTAT	1051	144	291	1014	0.829	0.828	0.828	0.155	0.826	0.653
34	Port n° 13 : DAYTIME	1101	184	257	958	0.824	0.822	0.823	0.158	0.823	0.646
35	Port n° 23 : TELNET	1110	101	315	974	0.842	0.836	0.839	0.144	0.833	0.668
36	Port n° 42 : HOSTNAMES	1144	151	143	1062	0.882	0.882	0.882	0.110	0.882	0.764
37	Port n° 70 : GOPHER	987	272	223	1018	0.802	0.802	0.802	0.176	0.803	0.604
38	Port n° 84 : CTF	927	200	225	1148	0.828	0.829	0.828	0.156	0.83	0.657
39	Port n° 95 : SUPDUP	1116	271	135	978	0.837	0.841	0.839	0.148	0.837	0.675
40	Port n° 102 : ISO_TSAP	1061	163	286	990	0.823	0.821	0.822	0.161	0.821	0.641
41	Port n° 105 : CSNET_NS	1012	242	199	1047	0.823	0.823	0.823	0.159	0.823	0.647
42	Port n° 109 : POP_2	1145	126	306	923	0.834	0.825	0.834	0.151	0.827	0.653
43	Port n° 117 : UUCP	1058	143	264	1035	0.839	0.838	0.839	0.146	0.675	0.837
44	Port n° 138 : NETBIOS_DGM	1123	197	211	969	0.836	0.835	0.836	0.06	0.836	0.868
45	Port n° 143 : IMAP4	1122	233	162	983	0.843	0.843	0.842	0.145	0.842	0.683
46	Port n° 150 : SQL_NET	1095	153	263	989	0.836	0.833	0.834	0.149	0.833	0.667
47	Port n° 194 : IRC	1159	185	218	938	0.838	0.836	0.837	0.147	0.838	0.675
48	Port n° 209 : MTP	1118	187	208	987	0.841	0.841	0.841	0.144	0.842	0.683
49	Port n° 387 : URP_I	1079	146	241	1034	0.846	0.845	0.846	0.141	0.845	0.691
50	Port n° 407 : TIM_I	1121	157	205	1017	0.855	0.854	0.855	0.133	0.855	0.711
51	Port n° 495 : ECO_I	1097	338	114	951	0.821	0.828	0.825	0.161	0.819	0.641

TABLEAU 4.24 – Tableau récapitulatifs des résultats correspondants à l'attribut 'Service'

Si nous procédons à une lecture théorique du tableau 4.5 du point de vue de sécurité, nous concluons que la faiblesse du système provient des FN et FP, où nous pouvons tolérer l'erreur FP que sur FN (c'est-à-dire tolérer les non-intrusion classé comme intrus que les intrus classé comme non-intrusion).

Pour cela, le tableau 4.24 prouve clairement ce que nous venons de dire. En effet, les résultats des différentes ports dites comme importantes (coloré en vert et jaune) la majorité des valeurs FN est inférieur aux valeurs de FP (dans notre de cas cette majorité est supérieur au 3/4).

Même remarque sur les autres ports dites moins importantes (coloré en violet et rouge) où leurs valeurs FN est inférieur aux valeurs FP (dans notre de cas cette majorité est supérieur au 3/4).

Donc, dans l'ensemble de des ports, les valeurs FN sont inférieur aux valeurs de FP. Ceci dit que notre système est robuste coté sécurité mais présente perte d'information ce qui est en concordance avec notre but de cette approche fondée sur la notion de sécurité.

Concernant les valeurs de la mesure entropie, nous remarquons que les valeur des ports en vert sont

très bonne (<10%), les jaunes sont assez bonne (<13%), les violette sont ordinaire (<17%) et les rouge sont mauvaise (>20%). Ceci est dû à représentativité de la base d'apprentissage spécialisé de chaque port.

Le concept d'agent spécialisé rend l'ensemble du système plus robuste et augmente la rentabilité et minimise les dégâts en cas d'une attaque. En effet, en cas d'attaque, l'agent responsable de la sécurité du port lance une défense offensive ou au pire des cas interrompre la connexion sur ce port, sans gêner les autres agents postés sur connexions seines qui continuent à assurer la sécurité de leur port sans se préoccuper par l'attaque.

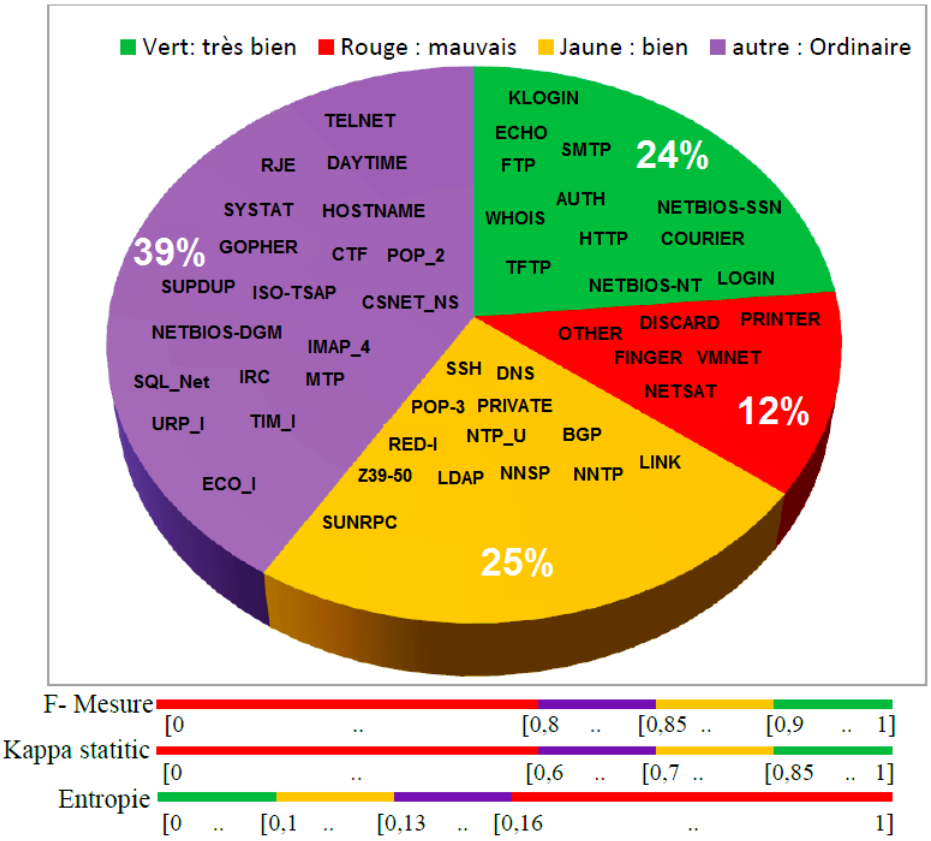


FIGURE 4.6 – Performance des agents

Dans la figure ci-dessus (Voir Figure 4.6), nous pouvons voir que le quart du ports réseau est très bien surveillé , nous constatons qu'ils sont des ports stratégiques et très actifs, tels que : ECHO, HTTP, FTP, SMTP, AUTH, WHOIS ...,de même pour la portion en jaune où les ports sont actifs. Nous constatons que 40% des ports ont une sécurité ordinaire, et qu'il ya seulement 12% des ports qui ont une mauvaise sécurisation et sont vulnérables.

Nous tenons à préciser que notre évaluation de performance des agents était très sévère, du faite que notre travail entre dans le cadre de la sécurité.

4.6.4 Comparaison

A titre de comparaison, nous avons fait une expérience avec un seul agent pour sécuriser totalement le réseau. Sachant que le plus grand défaut de kddcup'99 est sa taille énorme qui ne peut pas être utilisé dans l'ensemble, notre matériel nous permet d'aller jusqu'à 10.000 cas, voici ci-dessous le résultat (Voir Tableau 4.25) :

Technique conventionnelle: 10.000 instances max		
TP=4.784	FN=758	Précision = 0. 832
FP=901	TN=3.557	Rappel = 0.830
F-Mesure= 0.831		
Entropie= 0.152	Kappa statistc= 0.663	
Robustesse= 0.834		

TABLEAU 4.25 – Résultat de l'approche conventionnelle

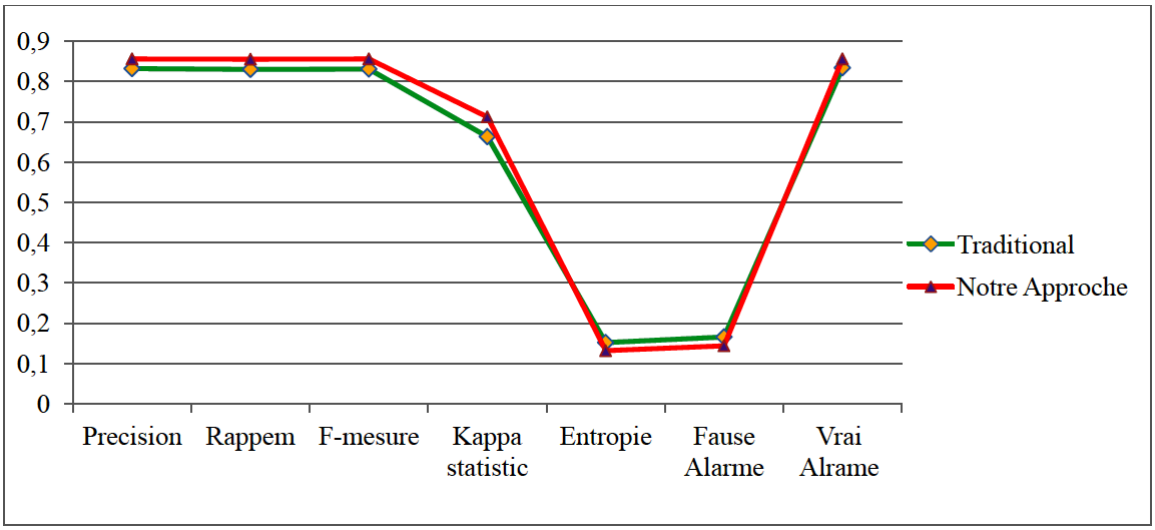


FIGURE 4.7 – Approche conventionnelle vs Notre Approche

Comme le montre le graphique (Voir Figure 4.7), notre approche est meilleure que l'approche conventionnelle sur tous les niveaux : Rappel, précision, F-mesure, Kappa statistique, les véritables alarmes sont supérieures à celle de l'approche conventionnelle; L'entropie et fausse alarme obtenu par notre approche sont réduite par rapport à celle de l'approche conventionnelle.

Pour notre approche, l'Entropie et la fausse alarme, sont réduites par rapport à celles de l'approche conventionnelle (Voir Figure 4.7).

Pour éclaircir la représentativité de notre approche par rapport à l'approche conventionnelle vis-à-vis l'utilisation des exemples du training set, nous dressons ce tableau (Voir Tableau 4.26).

	Nombre d'instance utilisé de KDD'99 pour la génération du modèle d'intrusion	Vis-à-vis 10% KDDCUP'99	Vis-à-vis KDDCUP'99
Notre approche	(2500 x 51)= 127.500	127.500/500.000 = 25%	127.500/5.000.000 = 2,5%
L'approche conventionnelle	10.000	10.000/500.000 = 2%	10.000/500.000 = 0,2%

TABLEAU 4.26 – La représentativité des approches

Nous remarquons selon le tableau 4.26 et la figure 4.7 que notre système basé sur uniquement 2.500 exemple est largement performant que l'approche conventionnelle. Nous pensons que si nous utilisons

10.000 exemples pour chaque port (capacité maximale utilisé dans l'approche conventionnelle) , nous aurons donc en totalité 510.000 exemples qui dépasse les 500.000 exemples de la base 10_percent_corrected de KDD'99, ce qui augmentera probablement la robustesse de notre système.

Conclusion

Dans ce chapitre, nous avons modélisé un modèle artificiel pour la détection d'intrusion par scénario basé sur une nouvelle méthode bio-inspirée en l'occurrence le système de protection des abeilles.

L'expérimentation de cette nouvelle approche a montré la robuste du système par rapport à l'approche conventionnelle vu les résultats atteints et la comparaison faite entre les deux. Ceci du fait que nous avons utilisés des Training set spécialisés sur chaque port, ou nous avons appliqué, pour la génération du modèle un baging qui est un vote a majorité simple, sur les trois modèles d'un même port (classifieurs bayésien, SVM et arbre de décision) qui ce comporte bien avec les classifications Bi-Classe.

Un autre point fort de notre approche est que la base d'apprentissage n'est pas totalement statique, car nous remplaçons à la fin de chaque détection d'intrusion un élément de notre base d'apprentissage par la nouvelle connexion anormal si celle ci est différente des exemples existant selon un certain seuil de similarité. Ceci afin d'assurer une grande représentativité des cas possibles. Nous pensons alors, que notre modeste contribution de l'IDSbees sera efficace, par conséquent la « ruche électronique » sera en sécurité et les criminelles n'oseront pas à l'approcher, néanmoins jusqu'à où ils fabriquent une tenue électronique qui peut résister à la défense offensive de notre IDSbee.

Conclusion et Perspectives

Conclusion générale

Vu les évolutions actuelles, les systèmes informatiques vont vraisemblablement continuer à s'immiscer d'avantage dans notre quotidien. Le développement de ces systèmes s'accompagne automatiquement de nombreux défis technologiques tels que la mobilité, l'évolutivité et l'autonomie, et la réactivité, etc. parmi ces défis, la question cruciale de la sécurité demeure un enjeu majeur en raison de la dématérialisation croissante de l'information et des risques de plus en plus importants liés à la complexité de ces systèmes.

Cette thèse traite la sécurité des systèmes informatiques. Les contributions que nous avons réalisées se focalisent principalement sur la protection de ces systèmes par la détection d'attaques informatiques, qui suscitent actuellement un intérêt croissant autant dans le monde académique que professionnel et surtout industriel en raison de la propagation de ces menaces informatiques représentées dans les intrusions, qui sont notre problématique, et qui se font aujourd'hui de plus en plus oppressantes. Donc la mise en place d'une bonne stratégie de défense contre les intrusions passe par l'amélioration des techniques de détections des IDS. Pour cela, il faut améliorer les algorithmes de détection, afin qu'ils puissent traiter les données de manière efficace.

Nous avons entamé nos contributions en proposant un nouveau modèle artificiel calqué du monde naturel qu'est le monde des abeilles (Le bio mimétisme), par l'utilisation des techniques du data mining, sur les quelles nous nous sommes basés pour identifier les intrus. Pour cerner cette problématique complexe, nous avons proposé une idée novatrice qui a dominé nos contributions en se basant sur le fait que les modèles des approches traditionnelles des IDS sont construits à partir d'un ensemble d'apprentissage générale de tout le réseau et qui génère un seul modèle d'apprentissage d'intrusion. Ce qui nous a amené proposer une nouvelle approche plus spécifique et plus décentralisée, en prenant aléatoirement le même nombre de connexions, pour les bases d'apprentissage et les bases de Test, correspondant au nombre de ports logique du réseau. Par conséquent, au lieu d'avoir un seul modèle d'apprentissage (approche traditionnelle) nous aurons ainsi comme nombre de modèles d'apprentissage différents le nombre des ports logiques du réseau, ce qui maximise la sécurité et la robustesse du réseau.

Un autre point important qu'apporte notre approche est la construction du modèle d'attaque en utilisant la base d'apprentissage qui leur a été attribué lors de l'état initial et en utilisant plusieurs algorithmes de Data Mining, tel que Naïve Bayes, le classifieur SVM et les arbres de décision, qui ont une performance élevée et reconnu pour les classifications bi-classe (intrus – non intrus) dans notre cas.

Un des points culminants de notre approche fondée sur la protection des systèmes des abeilles est l'utilisation intelligente des ressources (base d'apprentissage, base test) et matériel. Autre chose qu'apporte de notre approche est que la sécurité de la permanence est garantie, même en cas d'attaque sur un port, les ports logiques du réseau ne seront pas perturbés et continuent de fonctionner normalement.

Une des améliorations de notre approche est que l'ensemble d'entraînement (Training set) n'est pas statique, en fait nous remplaçons à la fin de chaque détection d'intrusion un élément de notre base d'apprentissage par la nouvelle connexion anormal si celle ci est différente des exemples existant selon un certain seuil de similarité. Ceci afin d'assurer une grande représentativité des cas possibles.

Donc, l'expérimentation que nous avons initiée et par le biais des ses résultats satisfaisants montre le bien fait qui nous a guidé à choisir cette méta heuristique classifiée intelligence en essaim pour notre problématique.

Perspectives

Nous prévoyons dans un avenir proche de finaliser les deux sous autres approches, en l'occurrence :

— Reconnaissance de comportement dans la ruche : détection d'intrusion par comportement (qui fait actuellement l'objet d'un article)

— Stratégie de réponse a une intrusion : la réponse active (défense offensive). Ceci pour donner plus d'ampleur à notre système de détection d'intrusion et pour qu'il soit complet et pourquoi pas opérationnel.

Comme, nous prévoyons d'apporter certaines améliorations en fonction des résultats en s'appuyant sur les points forts et essayer de corriger les faiblesses maximales de notre système basé sur l'approche par scénario. En même temps appliquer à notre modèle artificiel pour la détection d'intrusion d'autres algorithmes du data mining pour avoir une idée sur l'influence de ces derniers sur la performance de notre IDSbee.

Publications de l'auteur

A - Article

Lokbani, Ahmed Chaouki., Lehireche, A., Hamou, R.M. and Boudia, M.A., 2015. An Approach based on Social Bees for an Intrusion Detection System by Scenario. *International Journal of Organizational and Collective Intelligence (IJOCI), 5(3)*, pp.44-67.

Reda Mohamed Hamou, Abdelmalek Amine, and **Ahmed Chaouki Lokbani**. "Study of Sensitive Parameters of PSO Application to Clustering of Texts." *International Journal of Applied Evolutionary Computation (IJAEC)* 4.2 (2013) : 41-55.

Reda Mohamed Hamou, Abdelmalek Amine, Ali Rahmouni, **Ahmed Chaouki Lokbani**, Michel Simonet : " Modeling of Inclusion by Genetic Algorithms : Application to the Beta-Cyclodextrin and Triphenylphosphine". IJCCE 3(1) : 19-36 (2013)

Reda Mohamed Hamou, Abdelmalek Amine, and **Ahmed Chaouki Lokbani**. "The Social Spiders in the Clustering of Texts : Towards an Aspect of Visual Classification." *International Journal of Artificial Life Research (IJALR)* 3.3 (2012) : 1-14.

Reda Mohamed Hamou, Abdelmalek Amine, and **Ahmed Chaouki Lokbani**. "A Biomimetic Approach Based on Immune Systems for Classification of Unstructured Data." *arXiv preprint arXiv :1210.7002* (2012).

Reda Mohamed Hamou, Abdelmalek Amine, **Ahmed Chaouki Lokbani**, and Michel Simonet. "Visualization and clustering by 3D cellular automata : Application to unstructured data." *arXiv preprint arXiv :1211.5766* (2012).

Reda Mohamed Hamou, Ahmed Lehireche, **Lokbani Ahmed Chaouki**, and Rahmani Mohamed. "Text Clustering Based on the N-Grams by Bio Inspired Method (Immune Systems)." *Researchers World* 1, no. 1 (2010) : 56.

Reda Mohamed Hamou, Ahmed Lehireche, **LOKBANI Ahmed Chaouki**, and Rahmani Mohamed. "Representation of textual documents by the approach wordnet and n-grams for the unsupervised classification (clustering) with 2D cellular automata : a comparative study." *Computer and Information Science* 3, no. 3 (2010) : 240.

B - Chapitre

LOKBANI Ahmed Chaouki, et al. "Synthesis of Supervised Approaches for Intrusion Detection Systems." *Network Security Technologies : Design and Applications : Design and Applications* (2013) : 44.

C - Conférence

LOKBANI Ahmed Chaouki, Ahmed Lehireche, and Reda Mohamed Hamou. "Experimentation of Data Mining Technique for System's Security : A Comparative Study." *International Conference in Swarm Intelligence*. Springer Berlin Heidelberg, 2013.

Hamou RM, Lehireche A, **LOKBANI Ahmed Chaouki**, Rahmani M. Text clustering by 2D cellular automata based on the N-grams. InCryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE), 2010 First ACIS International Symposium on 2010 Oct 23 (pp. 271-277). IEEE.

Bibliographie

[1] ADAIF.FR, éd. *La colonie d'abeille*. Accessed: 12.01.2016. URL : http://www.adaif.fr/l-abeille/abeille-colonie.html.

[2] Ahmed AHMIM. "Système de détection d'Intrusion adaptatif et distribué". Thèse de doct. Université Badji Mokhtar Annaba, 2014.

[3] Ludovic Mé—Véronique ALANOU. *Détection d'intrusion dans un systeme informatique: méthodes et outils*. Supélec, Grande Ecole des Sciences de l'Information et de l'Energie, 1996.

[4] Massih-Reza AMINI et Eric GAUSSIER. *Recherche d'information: applications, modèles et algorithmes*. Editions Eyrolles, 2013.

[5] James P ANDERSON. *Computer security threat monitoring and surveillance*. Rapp. tech. Fort Washington, Pennsylvania, 1980.

[6] sorin ASSOCIES, éd. *Le biomimétisme*. Accessed: 25.12.2015. URL : http://www.sorin-associes.com/component/content/article/2-non-categorise/6-biomim%C3%A9tisme.html.

[7] Alain BACCINI et al. "Analyse des critères d'évaluation des systèmes de recherche d'information." In : *Technique et Science Informatiques* 29.3 (2010), p. 289–308.

[8] Ricardo BAEZA-YATES, Berthier RIBEIRO-NETO et al. *Modern information retrieval*. T. 463. ACM press New York, 1999.

[9] Nicolas BAUDOIN et Marion KARLE. *NT Réseaux: IDS et IPS*. 2000.

[10] David S BAUER et Michael E KOBLENTZ. "NIDX-an expert system for real-time network intrusion detection". In : *Computer Networking Symposium*. IEEE. 1988, p. 98–106.

[11] Michael JA BERRY. "Gordon. S. Linoff". In : *Data Mining Technique: For Marketing, Sales, and Customer Relationship Management* (1997).

[12] Clifford BESHERS et Steven FEINER. "Generating efficient virtual worlds for visualization using partial evaluation and dynamic compilation". In : *ACM SIGPLAN Notices* 32.12 (1997), p. 107–115.

[13] Philipe BIONDI. *Architecture expérimentale pour la détection d'intrusions dans un système informatique*. 2001.

[14] Laurent BLOCH et Christophe WOLFHUGEL. *Sécurité informatique: principes et méthodes*. Editions Eyrolles, 2013.

[15] Olivier BOISSIER. "Systèmes multi-agents". Thèse de doct. Ecole nationale supérieure des mines de Saint-Étienne, 2010.

[16] Lamia BOUABDELLAH et Asma BENMANSOUR. "Expansion de requête pour un système de recherche d'information par croisement de langues". Mém.de mast. Université Abou Bekr Belkaid Tlemcen, 2012.

[17] Djalila BOUGHAREB. "Recherche d'information multicritères". Thèse de doct. Université Badji Mokhtar Annaba, 2014.

[18] Abdelkrim BOURAMOUL. "RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE". Thèse de doct. Université de Constantine, 2011.

[19] S CATER et D KRAFT. “TIRS: A topological information retrieval system satisfying the requirements of the Waller-Kraft wish list”. In : *Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1987, p. 171–180.

[20] LELOU CATHÉRINE. *Moteurs d'indexation et de recherche: environnement client-serveur, Internet et Intranet*. Paris: Eyrolles, 1998.

[21] Stéphane CHAUDIRON et Madjid IHADJADENE. “Évaluer les systèmes de recherche d'information”. In : *Hermès, La Revue* 2 (2004), p. 170–178.

[22] Vincent CHEVRIER. “Etude et mise en oeuvre du paradigme multi-agents: de Atome a Gtmas”. Thèse de doct. Univsercité de Nancy, 1993.

[23] Hassan CHOUAIB. “Sélection de caractéristiques: méthodes et applications”. Thèse de doct. Université Paris Descartes, 2011.

[24] Krzysztof J CIOS, Witold PEDRYCZ et Roman W SWINIARSKI. “Data Mining and Knowledge Discovery”. In : *Data Mining Methods for Knowledge Discovery*. Springer, 1998, p. 1–26.

[25] E COLE. “Bezpieczeństwo sieci: biblia”. In : *Network Security Bible, Wiley Publishing, Inc* (2005), p. 0–7645.

[26] William S COOPER. “Getting beyond boole”. In : *Information Processing & Management* 24.3 (1988), p. 243–248.

[27] Mariam DAOUD. “Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif”. Thèse de doct. Institut de Recherche en Informatique de Toulouse, 2009.

[28] Herve DEBAR, Monique BECKER et Didier SIBONI. “A neural network component for an intrusion detection system”. In : *Research in Security and Privacy, 1992. Proceedings., 1992 IEEE Computer Society Symposium on*. IEEE. 1992, p. 240–250.

[29] Hervé DEBAR, Marc DACIER et Andreas WESPI. “A revised taxonomy for intrusion-detection systems”. In : *Annales des télécommunications*. T. 55. 7-8. Springer. 2000, p. 361–378.

[30] Dorothy E DENNING. “An intrusion-detection model”. In : *IEEE Transactions on software engineering* 2 (1987), p. 222–232.

[31] Guillaume DESGEORGE. *La sécurité des réseaux*. 2000.

[32] Yves DESWARTE. “La securite des systemes d'information”. In : *Revue de l'electricite et de l'electronique* 5 (2001), p. 21.

[33] Yves DESWARTE et Sébastien GAMBS. “Cyber-attaques et cyber-défenses: problématique et évolution”. In : *REE. Revue de l'électricité et de l'électronique* 2 (2012), p. 23–35.

[34] Georg DISTERER. *Iso/iec 27000, 27001 and 27002 for information security management*. Scientific Research Publishing, 2013.

[35] Cheri DOWELL et Paul RAMSTEDT. “The ComputerWatch data reduction tool”. In : *Proceedings of the 13th National Computer Security Conference*. T. 13. 1990, p. 99–108.

[36] Sumeet DUA et Xian DU. *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[37] Richard O DUDA, Peter E HART et al. *Pattern classification and scene analysis*. T. 3. Wiley New York, 1973.

[38] Susan T DUMAIS. “Latent semantic indexing (LSI): TREC-3 report”. In : *Nist Special Publication SP* (1995), p. 219–219.

[39] Ted DUNNING. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University, 1994.

[40] F EBEL et al. *Sécurité informatique - Ethical Hacking: Apprendre l'attaque pour mieux se défendre*. Editions ENI, 2009. ISBN : 978-2-7460-5105-8.

[41] Carl F ENDORF, Eugene SCHULTZ et Jim MELLANDER. *Intrusion detection & prevention*. McGraw-Hill Osborne Media, 2004.

[42] Wei FAN et al. “Using artificial anomalies to detect unknown and known network intrusions”. In : *Knowledge and Information Systems* 6.5 (2004), p. 507–527.

[43] Jacques FERBER. “Les systèmes multi-agents: un aperçu général”. In : *Techniques et sciences informatiques* 16.8 (1997).

[44] Stephanie FORREST, Steven A HOFMEYR et Anil SOMAYAJI. “Computer immunology”. In : *Communications of the ACM* 40.10 (1997), p. 88–96.

[45] Edward A FOX. *Extending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types*. Cornell University, 1983.

[46] Norbert FUHR. “Information Retrieval-From Information Access to Contextual Retrieval.” In : *Designing Information Systems*. 2004, p. 47–57.

[47] Patrick GALLINARI, Sylvie THIRIA et F Fogelman SOULIE. “Multilayer perceptrons and data analysis”. In : *Neural Networks, 1988., IEEE International Conference on*. IEEE. 1988, p. 391–399.

[48] K GEORGE. *Zipf. Human behavior and the principle of least effort*. 1949.

[49] Plerre-P GRASSÉ. “La reconstruction du nid et les coordinations interindividuelles chezBellicositermes natalensis etCubitermes sp. la théorie de la stigmergie: Essai d’interprétation du comportement des termites constructeurs”. In : *Insectes sociaux* 6.1 (1959), p. 41–80.

[50] Hubert GUERRIAT et Michel ITTELET. “Aperçu sur le statut du Milan noir (Milvus migrans) en Belgique”. In : *Aves* 19.3 (1982), p. 183–191.

[51] Montagne. H et J PAIN. “Analysis of trophallactic behavior of young bees (apis-mellifica l) using cinematographic recording”. In : *Comptes rendus hebdomadaires des seances de l academie des sciences serie D* 272.2 (1971), p. 297.

[52] Hadjira HACHEMI. “Moteur de recherche Sémantique.” Thèse de doct. Université de Tlemcen, 2014.

[53] Reda Mohamed HAMOU, Abdelmalek AMINE et Amine BOUDIA. “A New Meta-Heuristic Based on Social Bees for Detection and Filtering of Spam”. In : *International Journal of Applied Metaheuristic Computing (IJAMC)* 4.3 (2013), p. 15–33.

[54] Jiawei HAN et Micheline KAMBER. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[55] David J HAND, Heikki MANNILA et Padhraic SMYTH. *Principles of data mining*. MIT press, 2001.

[56] Richard HEADY et al. *The architecture of a network level intrusion detection system*. University of New Mexico. Department of Computer Science. College of Engineering, 1990.

[57] Paul HELMAN et Gunar LIEPINS. “Statistical foundations of audit trail analysis for the detection of computer misuse”. In : *IEEE Transactions on software engineering* 19.9 (1993), p. 886–901.

[58] Paul HELMAN, Gunar LIEPINS et Wynette RICHARDS. “Foundations of intrusion detection [computer security]”. In : *Computer Security Foundations Workshop V, 1992. Proceedings*. IEEE. 1992, p. 114–120.

[59] Nathalie HERNANDEZ. “Ontologies de domaine pour la modélisation du contexte en recherche d’information”. Thèse de doct. Université Paul Sabatier, 2005.

[60] Abdelkarim HERZALLAH. “Recherche d’information”. Thèse de doct. Université de Bouira, 2014.

[61] John H HOLLAND. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press, 1975.

[62] Carl J HUBERTY. *Applied discriminant analysis*. 519.535 HUB. CIMMYT. 1994.

[63] Peter INGWERSEN. “Polyrepresentation of information needs and semantic entities elements of a cognitive theory for information retrieval interaction”. In : *SIGIR’94*. Springer. 1994, p. 101–110.

[64] Russell A JACKSON. “Get the most out of audit tools: several practitioners share their approaches to maximizing the potential of automated tools. Plus, respondents to Internal Auditor’s 10th annual product and usage survey reveal their top software picks”. In : *Internal Auditor* 61.4 (2004), p. 36–45.

[65] Michel JAMBU. *Méthodes de base de l'analyse des données*. Eyrolles, 1999.

[66] Harold S JAVITZ et Alfonso VALDES. "The SRI IDES statistical anomaly detector". In : *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*. IEEE. 1991, p. 316–326.

[67] Finn V JENSEN. *An introduction to Bayesian networks*. T. 210. UCL press London, 1996.

[68] Arnold JOHNSON et al. "Guide for Security-Focused Configuration Management of Information Systems: The National Institute of Standards and Technology Special Publication 800-128". In : (2012).

[69] Auray JP, Duru G et Zighed A. *Analyse des données multidimensionnelles, volume 2 : les méthodes de structuration*. Alexandre Lacassagne, 2000. ISBN : 978-2905972217.

[70] Soheila KARBASI. "Pondération des termes en Recherche d'Information". Thèse de doct. Université Paul Sabatier, 2007.

[71] Nongdo Désiré KOMPAORÉ. "Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif". Thèse de doct. Université Paul Sabatier-Toulouse III, 2008.

[72] Sandeep KUMAR et Eugene H SPAFFORD. "A pattern matching model for misuse intrusion detection". In : (1994).

[73] Haystack LABS, éd. Accessed: 03.12.2015. URL : http://www.haystack.com/stalk.htm.

[74] Wenke LEE, Salvatore J STOLFO et al. "Data Mining Approaches for Intrusion Detection." In : *Usenix security*. 1998.

[75] René LEFÉBURE et Gilles VENTURI. *Data Mining: Gestion de la relation client personnalisation de sites web*. Eyrolles, 2001.

[76] Cédric LIORENS et al. *Tableaux de bord de la sécurité réseau*. Editions Eyrolles, 2011. ISBN : 2-212-11973-9.

[77] Richard P LIPPMANN et al. "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation". In : *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*. T. 2. IEEE. 2000, p. 12–26.

[78] M LUDOVIC. "Gassata, a genetic algorithm as an alternative tool for security audit trails analysis". In : *Proceedings of the First International Workshop on the Recent Advances in Intrusion Detection, Louvain-la-Neuve, Belgium*. 1998, p. 1–11.

[79] Hans Peter LUHN. "A statistical approach to mechanized encoding and searching of literary information". In : *IBM Journal of research and development* 1.4 (1957), p. 309–317.

[80] Oded MAIMON et Lior ROKACH. *Data mining and knowledge discovery handbook*. T. 2. Springer, 2005.

[81] Loïc MAISONNASSE. "Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision: application aux graphes pour la recherche d'information médicale." Thèse de doct. Université Joseph-Fourier-Grenoble I, 2008.

[82] D MASLENNIKOV et Y NAMESTNIKOV. *Kaspersky security bulletin 2012: The overall statistics for 2012. Kaspersky Lab*. 2012.

[83] Ludovic MÉ et Cédric MICHEL. "La détection d'intrusions: bref aperçu et derniers développements". In : *Mars* (1999).

[84] Trend MICRO. *IT Security FOR DUMMIES*. North American Small Business Edition, Wiley Publishing, Inc, 2011. ISBN : 978-1-118-08410-6.

[85] Calvin N MOOERS. "Information retrieval viewed as temporal signaling". In : *Proceedings of the international congress of mathematicians*. T. 1. 1950, p. 572–573.

[86] Klaus MÜLLER, Klaus Müller ALIAS'SOCMA et Georges TARBOURIECH. *IDS-Systèmes de Détection d'Intrusion, Partie I*. 2003.

[87] Jian-Yun NIE. *Introduction à la RI*. http://www.iro.umontreal.ca/~nie/IFT6255/Introduction.html. Accessed: 06.07.2016. 2015.

[88] Donald L PIPKIN. *Information security*. Prentice Hall PTR, 2000.

[89] Jay M PONTE et W Bruce CROFT. "A language modeling approach to information retrieval". In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, p. 275–281.

[90] Phillip A PORRAS et Alfonso VALDES. "Live Traffic Analysis of TCP/IP Gateways." In : *NDSS*. 1998.

[91] Carroll Cornelius PRATT. *The logic of modern psychology*. Macmillan, 1939.

[92] Kai RANNENBERG. "Recent Development in Information Technology Security Evaluation-The Need for Evaluation Criteria for Multilateral Security." In : *Security and control of information technology in society*. 1993, p. 113–128.

[93] Simon RÉHEL. "Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés". In : *Faculty of Science and Engineering, University LAVAL, QUEBEC* (2005).

[94] Robert RICHARDSON. "CSI : computer crime and security". In : *Computer Security Institute* 1 (2008), p. 1–30.

[95] CJ van RIJSBERGEN. "Information retrieval". In : *The Information Retrieval Group,< http://dsc. glasgow. ac. uk/preface. html* (1979).

[96] Stephen E ROBERTSON. "The probability ranking principle in IR". In : *Journal of documentation* 33.4 (1977), p. 294–304.

[97] Stephen E ROBERTSON et Steve WALKER. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". In : *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc. 1994, p. 232–241.

[98] Bernard RUYER. *Eloge de la Societe de Consommation*. philpapers, 1998.

[99] Gerard SALTON. "Experiments in Automatic Thesaurus Construction for Information Retrieval." In : *IFIP Congress (1)*. 1971, p. 115–123.

[100] Gerard SALTON. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971.

[101] Gerard SALTON et Christopher BUCKLEY. "Term-weighting approaches in automatic text retrieval". In : *Information processing & management* 24.5 (1988), p. 513–523.

[102] Gerard SALTON, Edward A FOX et Harry WU. "Extended Boolean information retrieval". In : *Communications of the ACM* 26.11 (1983), p. 1022–1036.

[103] Gerard SALTON et M. McGill. MCGRAW-HILL. *Introduction to modern information retrieval*. 1983.

[104] Warren S SARLE. *Neural networks and statistical models*. Citeseer, 1994.

[105] Guilhelm SAVIN. *Segmentation 2D et 3D par Systemes Multi-Agents*. 2008.

[106] Helmut SCHMID. "Treetagger| a language independent part-of-speech tagger". In : *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 43 (1995), p. 28.

[107] B SCHNEIER. "Secrets and Lies: digital security in a networked world. 2000". In : *New York, John Wiley & Sons. Rocco F. Grillo, CISSP Managing Director* 2.2.603 (), p. 838.

[108] Michael M SEBRING et al. "Expert systems in intrusion detection: A case study". In : *Proceedings of the 11th National Computer Security Conference*. T. 7. 1988, p. 4–81.

[109] Fethi SERRIR. "Annotation d'un corpus pour l'évaluation des systèmes de recherche d'informations". Thèse de doct. 2012.

[110] Claude Elwood SHANNON. "A mathematical theory of communication". In : *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), p. 3–55.

[111] Amit SINGHAL, Chris BUCKLEY et Mandar MITRA. “Pivoted document length normalization”. In : *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1996, p. 21–29.

[112] Stephen E SMAHA. “Haystack: An intrusion detection system”. In : *Aerospace Computer Security Applications Conference, 1988., Fourth*. IEEE. 1988, p. 37–44.

[113] Salvatore J STOLFO et al. “Cost-based modeling for fraud and intrusion detection: Results from the JAM project”. In : *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*. T. 2. IEEE. 2000, p. 130–144.

[114] Internet Security SYSTEMS, éd. Accessed: 15.11.2015. URL : http://www.iss.net/prod/rsds.html.

[115] Amel TERKIA DERDRA et Fatima Zahra BENSAFIA. “La Représentation Conceptuelle pour la Catégorisation des Textes Multilingue.” Mém.de mast. université de Tlemcen, 2014.

[116] Bhavani THURAISINGHAM. *Managing and mining multimedia databases*. CRC Press, 2001.

[117] Chih-Fong TSAI et al. “Intrusion detection by machine learning: A review”. In : *Expert Systems with Applications* 36.10 (2009), p. 11994–12000.

[118] Hank S VACCARO et Gunar E LIEPINS. “Detection of anomalous computer session activity”. In : *Security and Privacy, 1989. Proceedings., 1989 IEEE Symposium on*. IEEE. 1989, p. 280–289.

[119] Denis VALOIS et Cedric LLORENS. *Network security verification system and method*. US Patent App. 10/601,290. 2003.

[120] Sholom M WEISS et Nitin INDURKHYA. *Predictive data mining: a practical guide*. Morgan Kaufmann, 1998.

[121] Mark WOOD et Michael A ERLINGER. *Intrusion detection message exchange requirements*. Rapp. tech. 2007.

[122] Shelly Xiaonan WU et Wolfgang BANZHAF. “The use of computational intelligence in intrusion detection systems: A review”. In : *Applied Soft Computing* 10.1 (2010), p. 1–35.

[123] Lotfi A ZADEH. “Fuzzy sets”. In : *Information and control* 8.3 (1965), p. 338–353.

[124] Wahiba Nesrine ZEMIRLI. “Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif”. Thèse de doct. Université de Toulouse, Université Toulouse III-Paul Sabatier, 2008.

[125] EL Moukhtar ZEMMOURI. *Processus d'extraction de connaissances à partir de données ,Représentation et gestion des connaissances multi-points de vue*. Editions Eyrolles, 2014. ISBN : ISBN-10: 3838140486.

[126] Jacob ZIMMERMANN et Ludovic MÉ. “Les systèmes de détection d'intrusions: principes algorithmiques”. In : *disponible sur url: http://www. researchgate. net/publication/239917861_Les_systemes_de_d'etection_d'intr usions_principes_algorithmiques* (2002).