

Université Mouloud MAMMERI de Tizi-Ouzou
Faculté De Génie Électrique et Informatique
Département Informatique.



Mémoire de fin d'étude:

En vue d'obtention du Diplôme de MASTER 2 en Informatique

Option: Réseaux, Mobilité et Systèmes Embarqués

Thème:

***Prédiction avancée et multidimensionnelle dans le
contexte de données d'archives de faible qualité***

Présenté par :

- ❖ BENAMROUCHE Nadjib
- ❖ DJENNOUNE Salah Yanis

Jury :

DAOUI Mehammed	Professeur, UMMTO	Président
AHMED-OUAMER Rachid	Professeur, UMMTO	Encadrant
HADJALI Allel	Professeur, LIAS/ENSMA Poitiers (France)	Co-Encadrant
IDMHAND Fatiha	Professeur, Univ. Poitiers (France)	Co-Encadrante
BELKADI Malika	Professeur, UMMTO	Examinatrice

Remerciements

Nous adressons nos remerciements tout particulièrement :

*Au bon dieu pour toute la volonté et le courage qu'il nous a accordés pour
l'achèvement de ce travail.*

*À notre promoteur Mr Rachid AHMED-OUAMER Professeur à l'UMMTO, pour son
soutien et ses précieux conseils qui ont contribué à alimenter notre réflexion et qui
nous ont vraiment été utiles dans ce projet.*

*Du fond du cœur M. Allel HADJALI, Mme Fatiha IDMHAND et Mme Chourouk
BELHEOUANE, nos Co-Encadrants, qui nous ont énormément épaulé sincèrement et
sans la moindre hésitation le long de notre stage de fin d'études. Nous tenons
également à remercier tous les membres du LIAS et CRLA-Archivos sans oublier ceux
de notre université de Tizi-Ouzou.*

Aux membres du jury qui ont accepté d'évaluer notre travail.

*À nos très chers frères, sœurs, ami (es) et à toutes les personnes qui ont contribué de
loin ou de près à la réussite de ce travail.*

*À nos chers parents qui ont œuvré pour notre réussite, par leur amour, leur soutien,
tous les sacrifices consentis et leurs précieux conseils, pour toute leur assistance et
leur présence dans notre vie.*

Et enfin à tous ceux et celles qui me sont chers. «Que Dieu bénisse ce travail».

Dédicaces

A mes chers parents, pour tous leurs sacrifices, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A mes chères sœurs Lamia et Amina pour leurs encouragements permanents, et leur soutien moral,

A mes beaux-frères, Said et Sid Ali pour leurs appuis et leurs encouragements,

A tous mes chères amis plus particulièrement Amine, Hacene, Younes, Karim, Hichem et Adel pour leur soutien et pour les bons moments et les diverses aventures qu'on a eu ensemble durant mon parcours à l'université,

A mon binôme Yanis qui a toujours cru en moi, et m'a soutenu, poussé à me dépasser,

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible,

Merci d'être toujours là pour moi.

En témoignage de l'attachement, de l'amour et de l'affection que je porte pour vous.

Nadjib.B

Dédicaces

A mes chers parents, pour tous leurs sacrifices, leur tendresse, leur soutien tout au long de mes études,

A mon oncle Said et ma tante Houria pour m'avoir aidé, encouragé et accueilli,

A mon oncle Hamid et ma tante Fatima pour leur support et leur épaullement

A mes chères sœurs Aziza, Dalila et Djouher pour leurs encouragements permanents, et leur soutien moral,

A mes frères, Samir, Azedine et Amayas pour leurs appuis et leurs encouragements,

A tous mes chères amis plus particulièrement Amine, Hacene, Younes, Krimo, Hichem, Adel, Yacine, Lyes et Nassim pour leurs soutiens et pour les bons moments et les diverses aventures qu'on a eu ensemble durant mon parcours à l'université, A mon binôme Nadjib qui a toujours cru en moi, et m'a soutenu, poussé à me dépasser,

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible,

Merci d'être toujours là pour moi.

En témoignage de l'attachement, de l'amour et de l'affection que je porte pour vous.

Salah.D.Yanis

Résumé:

Contexte et problématique

Les sciences humaines produisent actuellement des masses de données très variées mais peinent à proposer de nouvelles observations et/ou connaissances à partir de celles-ci. Ces données sont soit de qualité faible car collectées à l'aide de processus semi-automatisés, à partir du terrain (entretiens, fouilles, photos, etc.), soit de qualité insuffisante, du point de vue de l'informatique, car réalisées à partir de différentes formes d'éditorialisation des sources. Ainsi, elles peuvent se révéler peu ou pas exploitables et requièrent une phase de « curation ». Les chercheurs dans ce domaine structurent leurs informations/données d'archives sous formes de bases de données, ils désirent faire ressortir un ensemble de métadonnées et explorer les approches/technologies modernes pour exploiter ces données. Le stage a pour but de développer des outils d'extraction de connaissances, de fouilles de données ou de clustering adaptés au contexte des données d'archives littéraires avec toute leur imperfection et leur variété, il s'agit de permettre aux chercheurs SHS de découvrir et d'aller vers de nouveaux questionnements de leurs données.

Objectifs

Un premier travail sur la prédiction de liens et de relations a été réalisé en 2019. Ce travail a conduit au développement d'un outil Link&Pred qui permet de calculer les relations possibles entre des objets à partir de leurs métadonnées (propriétés/attributs). L'objectif essentiel de ce stage est de faire évoluer cet outil vers une prédiction plus pertinente fondée sur un nombre d'attributs plus important. Dans un second temps, il s'agit de développer une interface plus « user-friendly » de l'outil afin de faciliter son exploitation par les experts des sciences humaines. Enfin, il faudra enrichir l'outil d'une fonctionnalité de visualisation des prédictions identifiées. Il sera aussi demandé au candidat une réflexion sur la manière de sécuriser l'outil développé ainsi qu'une étude en rapport avec la confidentialité et la sécurité des données manipulées.

Mots clés

Masses de données, Qualité de données, Données d'archives, Fouille de données, Prédiction de liens, Décision.

Abstract

Context and issues

The human sciences currently produce masses of very varied data but are struggling to propose new observations and/or knowledge from them. These data are either of low quality because they are collected using semi-automated processes, from the field (interviews, excavations, photos, etc.), or of insufficient quality, from a computer point of view, because they are produced using different forms of editorialization of the sources. Thus, they may prove to be of little or no use and require a "curation" phase. Researchers in this field structure their information/archival data in the form of databases, they want to highlight a set of metadata and explore modern approaches/technologies to exploit these data. The internship aims to develop tools for knowledge extraction, data mining or clustering adapted to the context of literary archival data with all their imperfection and variety, it is to allow SHS researchers to discover and go towards new questioning of their data.

Objectives

A first work on linkage and relationship prediction was carried out in 2019. This work led to the development of the Link&Pred tool that allows to calculate the possible relationships between objects from their metadata (properties/attributes). The main objective of this internship is to make this tool evolve towards a more relevant prediction based on a larger number of attributes. In a second step, the aim is to develop a more "user-friendly" interface for the tool in order to facilitate its use by experts in the human sciences. Finally, it will be necessary to provide the tool with a visualization functionality for the predicted results. The applicant will also be asked to consider how to secure the developed tool as well as a study about the confidentiality and security of the data handled.

Key words

Big Data, Data Quality, Archive Data, Data Mining, Link Prediction, Decision.

Table des matières

Introduction générale	13
-----------------------------	----

Chapitre I : Etat de l'art sur l'extraction de connaissances

Introduction	16
1. Besoin de l'Extraction de Connaissances à partir des Données	17
2. Extraction de Connaissances à partir de Données ECD.....	17
2.1. Etapes de l'ECD	18
2.1.1. Sélection et prétraitement des données	18
2.1.2. Fouille de données.....	19
2.1.3. Interprétation.....	20
3. ETL Extract-Transform-Load.....	21
3.1. Etapes de l'ETL	21
3.1.1. Extract.....	21
3.1.2. Transform.....	22
3.1.3. Load	22
3.2. ETL vs ELT	23
3.3. ETL Talend.....	24
Conclusion.....	25

Chapitre II : Contexte de l'étude et Analyse.

Introduction	27
1. Contexte du projet	28
1.1. Partenariat « LIAS – Equipe CRLA-Archivos ».....	28
1.1.1. Laboratoire LIAS	28
1.1.2. Equipe Archivos de l'Institut des textes et manuscrits modernes	28
1.2. Problématique.....	29
1.3. Objectifs détaillés du projet.....	30
2. Etude de l'existant	31
2.1. Données des sciences humaines	31
2.2. Normalisation des données	31

2.2.1.	Evaluation de la qualité des données	31
2.2.2.	Traitement des données	31
2.3.	Outil Link&Pred	32
2.3.1.	Présentation de l'outil.....	32
2.3.2.	Insuffisances et limitations de l'outil	34
2.3.3.	Solution proposée.....	34
3.	Analyse	35
3.1.	Intelligence Artificielle	35
3.2.	Machine Learning.....	35
3.2.1.	Approches du Machine Learning	36
3.2.1.1.	Approche supervisée	36
3.2.1.2.	Approche non-supervisée.....	37
3.2.1.3.	Différence entre les deux approches.....	37
3.3.	Représentation graphique.....	38
3.3.1.	Graphe.....	38
3.3.2.	Réseaux.....	38
3.4.	Réseau social	40
3.5.	Prédiction de liens.....	42
3.5.1.	Intérêt de la prédiction de liens dans les réseaux sociaux	43
3.5.2.	Domaine d'application de la prédiction de liens	43
3.5.3.	Approches de la prédiction de liens	44
3.5.3.1.	Approche basée sur les nœuds.....	44
3.5.3.2.	Approche probabiliste	44
3.5.3.3.	Approche topologique.....	45
3.5.4.	Algorithmes de la prédiction de liens.....	45
3.5.4.1.	Algorithme de Adamic Adar	45
3.5.4.2.	Algorithme Preferential Attachment.....	46
3.5.4.3.	Algorithme Resource Allocation	46
3.5.4.4.	Algorithme SimRank	47

3.5.4.5. Algorithme Jaccard Similarity	47
3.5.4.6. Algorithme Common Neighbors	48
3.5.5. Comparatif des différents algorithmes de la prédiction de liens ..	49
3.5.6. Variante de Common Neighbors.....	49
Conclusion.....	50

Chapitre III : Conception, Implémentation et Réalisation.

Introduction	52
1. Algorithme de la prédiction de liens utilisé CNGF	52
1.1. Notion de capacité de guidage des nœuds	52
1.2. Algorithme CNGF	54
2. Comparaison des Algorithmes de la prédiction de liens	57
2.1. Comparaison entre Common Neighbors et CNGF.....	58
3. Aspect sécurité dans notre plateforme.....	64
3.1. Importance de la sécurité dans le web	64
3.2. Protocole HTTPS.....	64
3.3. Authentification	65
3.4. Fonction de Hachage.....	66
3.5. Algorithme Bcrypt	66
3.6. Protection contre Cross-Site Request Forgery	67
4. Conception de la plateforme web Future Links	68
4.1. Langage UML	68
4.2. Diagramme des cas d'utilisation de notre cas d'étude	68
4.3. Diagramme de classes de notre cas d'étude.....	69
4.4. Diagramme de séquences de notre cas d'étude.....	70

5. Outils, langages et bibliothèques utilisés.....	71
6. Présentation de notre application Future Links	76
Conclusion.....	79
Conclusion générale.....	80
Bibliographie	82

Liste des figures :

Figure 1 : Les étapes du processus ECD	20
Figure 2 : Les étapes du processus ETL	23
Figure 3 : Comparaison des étapes des processus ETL et ELT	24
Figure 4 : Liste d'adjacence entre les auteurs	32
Figure 5 : L'outil Gephi	32
Figure 6 : L'interface de l'outil Link&Pred	33
Figure 7 : Résultats de l'outil Link&Pred	33
Figure 8 : Représentation d'un réseau à l'aide d'un graphe	39
Figure 9 : Représentation d'un graphe orienté et d'un graphe non orienté	40
Figure 10 : Structure d'un réseau social	41
Figure 11 : Aperçu du processus de prédiction de liens	42
Figure 12 : Graphe pour comprendre Common Neighbors	48
Figure 13 : Les graphes représentant l'exemple 1	53
Figure 14 : Les sous-graphes représentant l'exemple 1	54
Figure 15 : Les graphes représentant l'exemple 2	56
Figure 16 : Les sous-graphes représentant l'exemple 2	56
Figure 17 : Le graphe de la courbe ROC.	60
Figure 18 : Résultats du test 1 avec Link&Pred	61
Figure 19 : Résultats du test 1 avec Future Links	61
Figure 20 : Résultats du test 2 avec Link&Pred	62
Figure 21 : Résultats du test 2 avec Future Links.	63

Figure 22 : Diagramme de cas d'utilisation .	69
Figure 23 : Diagramme de classes .	70
Figure 24 : Diagramme de séquences pour Upload un fichier (.net)	70
Figure 25 : Diagramme de séquences pour Upload un fichier csv	71
Figure 26 : Logo Python	71
Figure 27 : Logo Flask	72
Figure 28 : Logo Heroku	72
Figure 29 : Logo Visual studio code	73
Figure 30 : Logo Git	74
Figure 31 : Logo MySQL	74
Figure 32 : Interface d'inscription	76
Figure 33 : Interface de connexion	77
Figure 34 : Interface d'ajout de fichier	77
Figure 35 : Interface de prédiction de liens et affichage de graphe	78
Figure 36 : Fonctionnalité de recherche	78
Figure 37 : Interface historique	79

Liste des tableaux :

Tableau 1 : Tableau comparatif des différents algorithmes de prédiction de liens.	49
Tableau 2 : Mesure des performances pour les petits ensembles de données.	57
Tableau 3 : Mesure des performances pour les grands ensembles de données.	58
Tableau 4 : Différence de similarité entre Common Neighbors et CNGF.	59

Introduction générale

Avec l'explosion de la quantité d'informations à stocker il y a eu une évolution des modèles et des techniques de stockage de données. Cependant la gestion de données relatives aux sciences humaines rencontre des problèmes liés aux masses de données produites ainsi que leurs manipulations. Ces masses d'informations rendent difficiles l'analyse, le résumé et l'extraction manuelle des potentielles connaissances qui peuvent être contenues dans ces données. Cette extraction de données offre de nombreux avantages, notamment une meilleure analyse, une prise de décision améliorée, une accessibilité accrue des données et une productivité améliorée. Au carrefour des méthodes et domaines variés tels que les statistiques, l'intelligence artificielle, l'interaction homme-machine ou encore les bases de données, la recherche scientifique a consacré ces 30 dernières années de réels efforts pour développer des solutions aux problèmes d'aide à la décision, de conception et de développement d'outils permettant d'extraire automatiquement, ou du moins plus facilement, de la connaissance à partir de ces données volumineuses.

Un outil nommé Link&Pred a été développé au LIAS pour faciliter aux chercheurs de l'équipe Archivos-CRLA l'exploration des données collectées à partir de leurs champs d'études. Cet outil a été développé dans le cadre du projet initié en 2019 par l'équipe Archivos-CRLA de l'Institut des Textes et Manuscrits Modernes (UMR 8132) en collaboration avec le laboratoire LIAS (Laboratoire d'Informatique et d'Automatique pour les Systèmes) de l'ISAE/ENSMA (Institut Supérieure de l'Aéronautique et de l'Espace/ École Nationale Supérieure de Mécanique et d'Aéronautique) de Poitiers (France). Il s'agit d'un outil d'exploration et d'analyse des données d'archives de très faible qualité.

Link&Pred est un outil de prédiction de liens entre auteurs, qui utilise un algorithme de prédiction de liens prédéfini. Une interface d'utilisation pour les archivistes concernés est également développée [1].

Le but de notre travail est dans un premier temps, l'amélioration et l'évolution de cet outil vers une prédiction plus pertinente fondée sur un ensemble d'attributs. Dans un second temps, développer une plateforme qui offre de nouvelles fonctionnalités tout en améliorant l'interface pour qu'elle soit plus « user-friendly » afin de faciliter son fonctionnement et son exploitation.

Afin de concrétiser ce projet, notre travail sera organisé de la façon suivante :

- Dans un 1^{er} chapitre intitulé «Etat de l'art sur l'extraction de connaissances», nous expliquerons les différentes méthodes d'extraction des connaissances à partir des données, notamment ECD et ETL.
- Dans un 2^{ème} chapitre intitulé «Contexte de l'étude et Analyse», nous présenterons le contexte de notre projet de master, et l'étude de l'existant. Ensuite nous enchaînerons avec l'analyse en expliquant certaines notions telles que, le Machine Learning, les graphes ainsi que le concept de réseau social. Enfin, nous introduirons les algorithmes de prédiction de liens et proposerons une amélioration.
- Dans un 3^{ème} chapitre intitulé «Conception, Implémentation et Réalisation», nous donnerons une explication plus détaillée de la technique de prédiction utilisée ainsi qu'une amélioration de la version actuelle, et nous la comparerons avec la version initiale de l'outil en testant et en vérifiant les résultats obtenus par notre application. Pour finir, nous listerons les outils utilisés et nous montrerons les différentes fonctionnalités de la plateforme développée.
- Enfin, nous terminerons par une conclusion générale qui reprendra toutes les idées contenues dans notre travail.

Chapitre I :

Etat de l'art sur l'extraction de connaissances

Introduction

L'informatique moderne a permis la production et l'archivage d'énormes masses de données numériques lors de ces deux dernières décennies. Ces données sont généralement collectées pour rendre un service spécifique ou répondre à une problématique précise. Parallèlement le problème d'accès à ces connaissances dépasse les capacités humaines d'analyse, tant ces éléments sont disséminés dans une quantité importante de données souvent complexes.

Au début des années 90, une nouvelle discipline scientifique appelée communément Extraction de Connaissances à partir de Données (ECD) a vu le jour, pour répondre à ce besoin d'exploiter pleinement les énormes quantités de données récoltées.

L'ECD est une discipline, à l'intersection des domaines des bases de données, des interfaces homme / machine (IHM), de l'intelligence artificielle et des statistiques [2]. Elle propose des connaissances nouvelles qui enrichissent les interprétations du champ d'application, à partir de données collectées par des experts tout en fournissant des méthodes automatiques qui exploitent ces informations.

En effet l'ECD repose sur des algorithmes de fouilles de données (Data Mining). En général de tels algorithmes travaillent sur des données ayant un format bien particulier et adaptées au type de connaissance que nous cherchons à extraire. Nous appelons une telle représentation des données « contexte d'extraction ». Suite à l'extraction, l'algorithme renvoie un ou plusieurs motifs (ou modèles) construits à partir du jeu de données. Un motif est une représentation de tout ou partie du contexte d'extraction initial. Deux types de motifs se distinguent : les motifs globaux et les motifs locaux. Les motifs globaux ont pour but de modéliser le jeu de données dans son ensemble. Parmi les motifs globaux, on trouve par exemple les arbres de décision ou les réseaux de neurones. Les motifs locaux quant à eux ont pour but de décrire des propriétés locales des données, par exemple des corrélations entre certains enregistrements.

En recherchant des motifs locaux, nous cherchons à décrire le jeu de données par un ensemble de caractéristiques. Parmi les motifs locaux, on trouve les ensembles d'items, les règles d'association et les motifs séquentiels [3].

Nous avons réparti ce chapitre en trois sections, la première introduit la genèse du besoin d'extraction de connaissances à partir des données. La deuxième section est dédiée à la définition du concept d'ECD ainsi que des différentes étapes constituant son processus. Dans la troisième et dernière section de ce chapitre nous présentons l'ETL Extract-Transform-Load.

1 Besoin de l'Extraction de connaissances à partir des données

L'homme a toujours mémorisé sur des supports différents des informations qui lui ont permis d'inférer des lois. L'exploitation de données pour en extraire des connaissances est une préoccupation constante de l'être humain car c'est l'une des conditions essentielles à son évolution [4]. La physique, la chimie ou la sociologie et biens d'autres disciplines, font usage de l'approche empirique pour faire ressortir des éléments structurants dans des populations et découvrir des lois. Devenue une science, la statistique a pour objet de donner un cadre rigoureux à la démarche empirique. Le Data Mining a puisé une large partie de ses outils depuis la statistique et du domaine des bases de données. L'explosion de la quantité d'information à stocker a donc provoqué une évolution des modèles et des techniques de stockage de données. Ainsi dans de nombreux domaines, les besoins ont rapidement évolué et la nécessité d'automatiser l'acquisition des données a été ressentie et a fait naître le processus de l'ECD qui est l'ensemble des opérations qui permettent d'exploiter facilement et rapidement des données stockées.

2 Extraction de connaissances à partir des données ECD

L'extraction de connaissances à partir de données (ECD) se définit comme un processus de découverte d'informations inconnues, implicites et potentiellement utiles à partir de données. Ce processus se déroule en plusieurs étapes : préparation des données, fouille des données, validation et interprétation du résultat et enfin intégration des connaissances apprises [5]. Cependant, il se trouve que dans beaucoup de domaines, les données représentées sont incomplètes et imprécises ce qui rend leur

exploitation très difficile ou impossible d'autant plus que la tâche d'extraction de connaissances à partir des masses de données se complique à cause de la dimensionnalité élevée des bases de données.

Le but de l'ECD était initialement d'extraire de la connaissance dans des bases de données existantes. Cela a un double impact : d'une part, sur les algorithmes, qui doivent prendre en compte certaines caractéristiques propres aux bases de données (accès aux données, standardisation, volumes à traiter). D'autre part, une adaptation du modèle des bases de données est souhaitable afin de mieux prendre en compte certaines propriétés des algorithmes d'ECD.

2.1 Etapes de l'ECD

Il y a trois principales étapes dans le processus ECD : sélection et prétraitement de données, fouille de données et l'interprétation.

2.1.1 Sélection et prétraitement des données

C'est la transformation (nettoyage, intégration des sources multiples, agrégation, normalisation) d'une masse d'informations dans le but de construire une base de données. Le prétraitement porte sur l'accès aux données en vue de construire des Data Marts, des corpus de données spécifiques. Le prétraitement concerne la mise en forme des données entrées selon leur type (son, numérique, symbolique, texte, image), ainsi que la sélection d'attributs ou la sélection d'instances, le nettoyage des données et le traitement des données manquantes [6].

La nécessité de cette première phase est due au choix des descripteurs et de la connaissance précise du domaine et va dépendre de la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillons d'apprentissage peut faire échouer l'opération.

2.1.2 Fouille de données (Data Mining)

Une confusion subsiste encore entre Data Mining, que nous appelons en français «fouille de données» et Knowledge-Discovery in Databases, que nous appelons en français «Extraction de Connaissances à partir de Données (ECD)» [7]. Le Data Mining est l'une des étapes de la chaîne de traitement pour la découverte des connaissances à partir des données. Nous pouvons dire, sous un autre angle, que l'ECD est un véhicule dont le Data Mining est le moteur. L'ECD, par le biais du Data Mining, est alors vue comme une ingénierie pour extraire des connaissances à partir de données [8].

Le traitement de données par un algorithme d'extraction est un processus interactif d'analyse dans le but d'extraire des connaissances exploitables. L'émergence du Data Mining n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques.

Nous utilisons le Data Mining pour découvrir et interpréter les relations, les tendances ou les tendances cachées dans les grandes sources de données. Avec la surcharge d'informations actuelle, il est devenu de plus en plus difficile d'analyser l'énorme quantité de données et de générer des décisions d'extraction appropriées. Le Data Mining est une technique automatisée ou semi-automatisée qui comprend un ensemble de stratégies et de pratiques utilisées pour déterminer, créer, représenter, distribuer, qui peut constituer une véritable méthodologie [9]. Bien que le Data Mining soit méthodiquement similaire à l'extraction d'informations et à l'entreposage de données, la différence principale est que le résultat de l'extraction va au-delà de la création d'informations structurées ou de la transformation en un schéma relationnel. Elle exige soit la réutilisation des connaissances formelles existantes, soit la production d'un schéma basé sur les données sources.

L'objectif de la mise en œuvre des techniques de Data Mining est d'aboutir à des connaissances opérationnelles. Ces connaissances sont exprimées sous forme de modèles plus ou moins complexes : une série de coefficients pour un modèle de prévision numérique, des règles logiques du type « si Condition alors Conclusion » ou

des instances. Ces modèles peuvent être prédictifs, dans ce cas, les motifs trouvés ont pour but de prédire des comportements futurs, ou bien ils peuvent être descriptifs, les motifs extraits ont alors pour but de décrire les données de façon compréhensible et intelligible pour l'utilisateur. Selon les objectifs que nous souhaitons remplir, les algorithmes de fouille employés pour analyser les données seront différents.

2.1.3 Interprétation

L'interprétation est faite pour évaluer les découvertes réalisées lors de la fouille de données. Une fois cette interprétation terminée, l'utilisateur peut enfin savoir quels schémas intéressants se cachent dans la quantité de données qu'il a stockées. Pour cela, des mesures quantitatives et qualitatives sont définies, afin d'évaluer les motifs extraits et d'aider à leur utilisation [10].

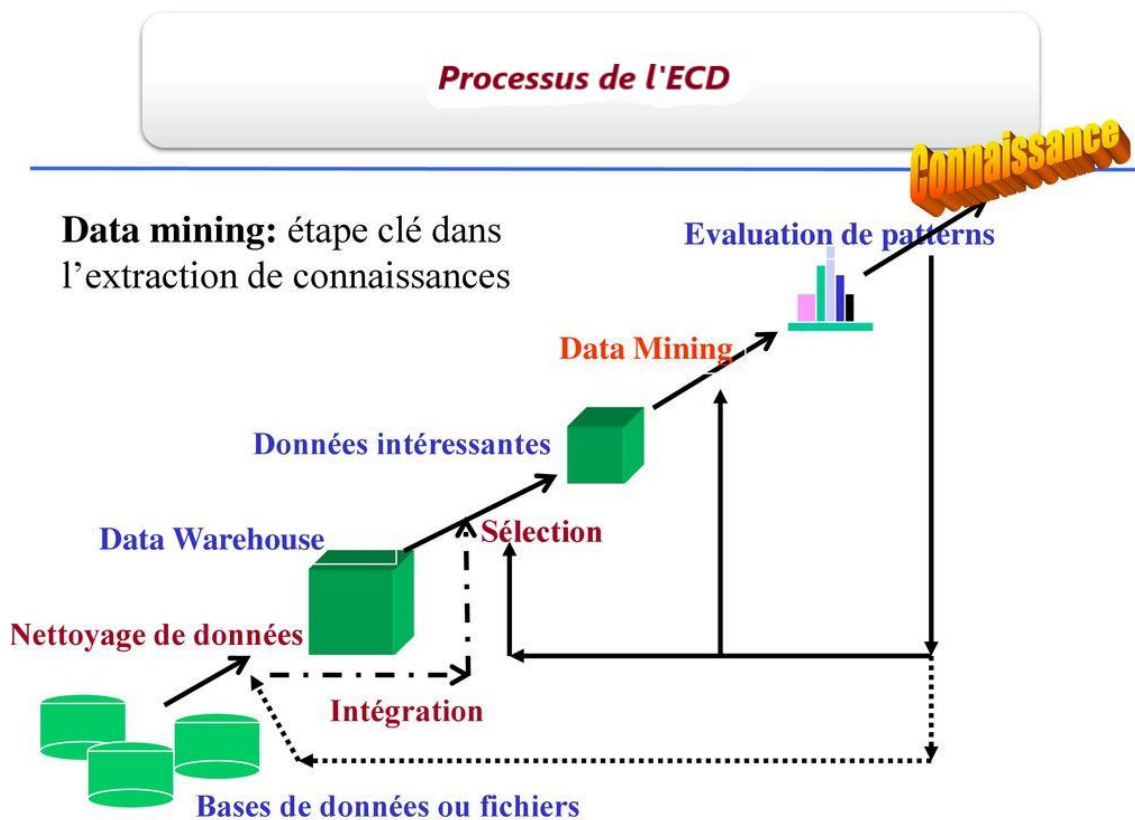


Figure 1 : Les étapes du processus ECD

3 ETL Extract-Transform-Load

Extract, Transform, Load est défini comme un mécanisme pour acquérir des données de divers systèmes sources (Extract), les normaliser (Transform) et ensuite remplir les données transformées dans l'entrepôt de données cible (Load). L'ETL est une approche d'intégration qui recueille des informations auprès de sources distantes, les transforme en formats et styles définis, puis les charge dans des bases de données [11].

ETL a été introduit dans les années 1970 comme un processus d'intégration et de chargement des données dans les ordinateurs centraux ou les supercalculateurs pour le calcul et l'analyse. De la fin des années 1980 au milieu des années 2000, c'était le principal processus de création d'entrepôts de données qui soutiennent les applications de veille stratégique (Business Intelligence). ETL est maintenant recommandée plus souvent pour créer des dépôts de données cibles plus petits qui nécessitent une mise à jour moins fréquente tandis que d'autres méthodes d'intégration des données, y compris l'ELT (extraction, chargement, transformation) [12], et la virtualisation des données sont utilisées pour intégrer des volumes de plus en plus importants de données en constante évolution ou de flux de données en temps réel.

Le fonctionnement d'ETL est un processus à trois étapes.

3.1 Etapes de l'ETL

Il y a trois principales étapes dans le processus ETL : Extract, Transform, Load.

3.1.1 Extract

Lors de l'étape de l'extraction, les données sont copiées ou exportées des emplacements sources vers une zone de transit. Les données peuvent provenir de n'importe quelle source structurée ou non structurée : serveurs SQL ou NoSQL, systèmes CRM (Customer Relationship Management) et ERP (Enterprise Resource Planning), LOB (Large Object), fichiers textes et documents, courriels, pages web et plus encore [13].

3.1.2 Transform

Dans la zone de transit, les données brutes sont transformées pour être utiles à l'analyse et pour s'adapter au schéma de l'entrepôt de données cible éventuel, qui est généralement alimenté par un traitement analytique en ligne structuré OLAP (Online Analytical Processing) ou une base de données relationnelle [14]. Il peut s'agir de ce qui suit :

- Filtrage, nettoyage, désencombrement, validation et authentification des données.
- Effectuer des calculs, des traductions ou des résumés à partir des données brutes. Il peut s'agir de modifier les en-têtes de lignes et de colonnes pour assurer la cohérence, de convertir des devises ou des unités de mesure, de modifier des chaînes de texte, de faire des sommes ou de calculer la moyenne des valeurs, tout ce qui est nécessaire pour répondre aux besoins de l'organisation en matière d'analyse.
- Supprimer, chiffrer, cacher ou protéger autrement les données régies par les règlements du gouvernement ou de l'industrie.
- Formatage des données en tables ou tables jointes pour correspondre au schéma de l'entrepôt de données cible.

L'exécution de ces transformations dans une zone de transit, plutôt que dans les systèmes sources eux-mêmes, limite l'incidence du rendement sur les systèmes sources et réduit la probabilité de corruption des données.

3.1.3 Load

Dans cette dernière étape, les données transformées sont déplacées de la zone de transit vers un entrepôt de données cible. Généralement, cela implique un chargement initial de toutes les données, suivi d'un chargement périodique des changements de données supplémentaires et, moins souvent, des actualisations complètes pour effacer et remplacer les données dans l'entrepôt [15].

Pour la plupart des organisations qui utilisent l'ETL, le processus est automatisé, bien défini, continu et axé sur les lots exécuté en dehors des heures de travail lorsque le trafic sur les systèmes sources et l'entrepôt de données est à son plus bas.

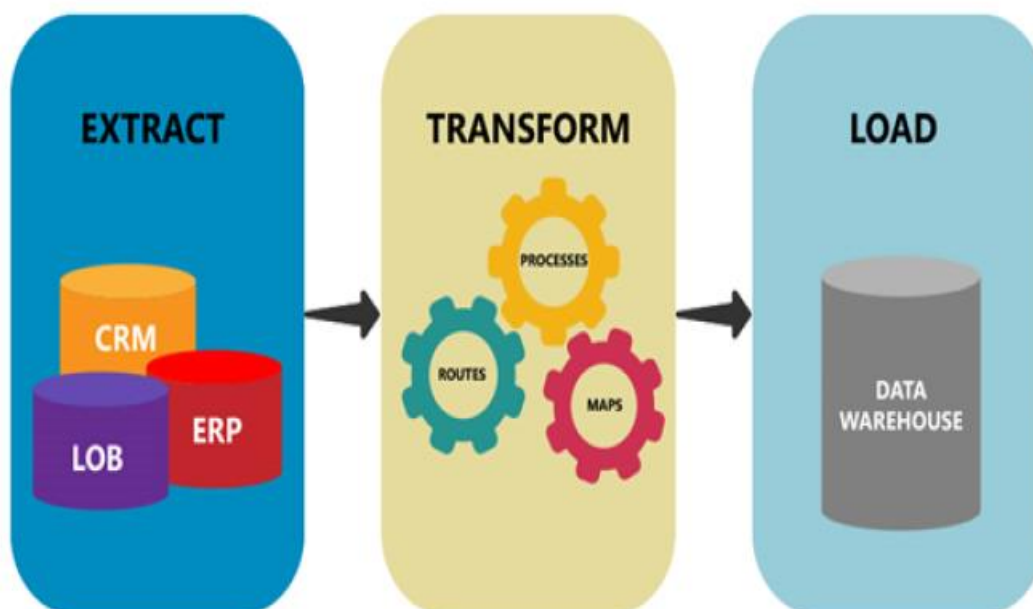


Figure 2 : Les étapes du processus ETL

3.2 ETL vs ELT

ELT (extraction, chargement, transformation) inverse les deuxième et troisième étapes du processus ETL. Il copie ou exporte les données à partir des emplacements sources, mais au lieu de les déplacer vers une zone de transit pour transformation, il charge les données brutes directement vers le magasin de données cible, où elles peuvent être transformées au besoin [16].

L'ordre des étapes n'est pas la seule différence. Dans ELT, le magasin de données cible peut être un entrepôt de données, mais le plus souvent c'est un lac de données, qui est un grand magasin central conçu pour contenir des données structurées et non structurées à grande échelle. Les lacs de données sont gérés à l'aide d'une plateforme Big Data (comme Apache Hadoop) ou d'un système de gestion de données NoSQL

distribué. Le plus souvent, ils sont créés pour soutenir l'intelligence artificielle, l'apprentissage automatique, l'analyse prédictive et les applications pilotées par des données en temps réel et des flux d'événements [17].

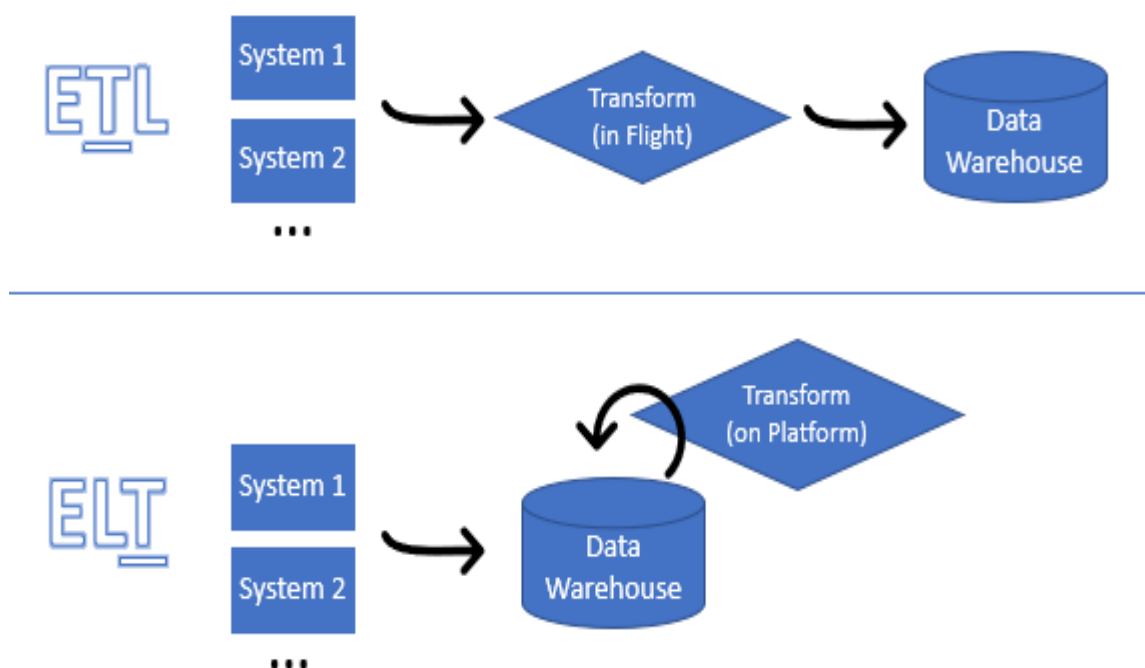


Figure 3 : Comparaison des étapes du processus ETL et du processus ELT

3.3 ETL Talend

Une large panoplie d'outils ETL existe que nous pouvons décomposer en deux principales catégories : des outils propriétaires comme IBM InfoSphere DataStage et Microsoft SSIS, mais aussi des outils Open sources comme Pentaho, Cloudera et Talend.

Talend est outil de l'ETL (Extract Transform and Load) qui permet d'extraire des données d'une source, de modifier ces données, puis de les recharger vers une destination. La source et la destination des données peuvent être une base de données, un service web, un fichier csv et bien d'autres.

Talend peut donc être utilisé dans n'importe quel contexte où des données sont véhiculées. L'IDE de développement est basé sur Eclipse et des connecteurs standards existent pour la majorité des sources de données, tels que les ERP, les BDD ou encore des solutions de commerce en ligne [18].

Talend est capable de transformer les données vers presque tous les formats existants et si une transformation ou un connecteur est manquant, il suffit de le développer très facilement en Java [19].

Parmi les outils que peut utiliser Talend nous citons :

- Hadoop base de données qui permet de gérer des bases de données de tailles colossales [20].
- Cloud Computing c'est l'utilisation et l'accès à différents services depuis internet à partir d'un fournisseur quelconque [21]. Le Cloud est caractérisé par sa disponibilité mondiale libre-service.
- Data Warehouse est une base de données utilisée pour manipuler plusieurs informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise [22].

Conclusion

Dans ce chapitre, nous avons commencé par introduire l'Extraction de Connaissances de Données et le besoin de l'utiliser puis nous sommes passés à la définition de l'ECD toute en parcourant les différentes étapes constituant son processus. Nous avons ensuite défini l'ETL et ses étapes. Enfin, nous avons présenté la différence entre l'ETL et l'ELT.

Dans le prochain chapitre, nous présenterons le contexte de notre projet et l'étude de l'existant. Ensuite, nous introduirons les notions, l'Intelligence Artificielle, le Machine Learning, la théorie des graphes, la théorie des réseaux et les différents algorithmes de prédiction de liens. Enfin nous expliquerons brièvement notre solution pour l'amélioration de la prédiction.

Chapitre II :

Contexte de l'étude et Analyse.

Introduction

L'an dernier, le Pr. Fatiha Idmhand a fait appel au laboratoire LIAS pour développer un outil permettant l'analyse et l'exploitation des données d'archives de l'équipe Archivos de l'institut des Textes et Manuscrits modernes (UMR8132) sous formes de textes et d'images. L'outil est également destiné à soutenir des systèmes de recommandations qui aident à la décision dans le contexte de la recherche scientifique.

Dans le cadre d'un stage de Master 2 et en collaboration avec le Pr. Hadj Ali du laboratoire LIAS, une première version d'un outil d'exploitation de données d'archives a été développée. Cette année, le laboratoire LIAS a proposé dans la continuité de ce travail un projet de fin d'études, le premier objectif du projet consiste à améliorer l'outil déjà développé, le deuxième objectif concerne l'amélioration de l'interface graphique de l'outil et le développement d'une plateforme web sécurisée.

Dans ce chapitre, nous avons divisé notre travail en trois sections :

- Première section intitulée «Contexte du projet» : Nous allons commencer par présenter les deux partenaires, à savoir le laboratoire LIAS, et l'équipe Archivos, puis nous allons discuter de la problématique du projet, ainsi que des objectifs que nous voulons atteindre à la fin de ce projet.
- Deuxième section intitulée «Etude de l'existant»: Nous allons présenter l'étude de l'existant et l'outil Link&Pred déjà développé ainsi que ses limites et les différentes solutions à proposer.
- Troisième section intitulée «Analyse» : Après avoir introduit les notions d'Intelligence Artificielle, de Machine Learning, des graphes et leurs utilisations et la théorie des réseaux, nous allons définir la notion de prédiction de liens, les différentes approches et les algorithmes utilisés. Enfin, nous en proposons une amélioration.

1 Contexte du projet

1.1 Partenariat «LIAS –Equipe CRLA-Archivos»

1.1.1 Laboratoire LIAS

Le LIAS (Laboratoire d'Informatique et d'Automatique pour les Systèmes) comporte des enseignants chercheurs dans les disciplines de l'Automatique, du Génie électrique et de l'Informatique. Il a été créé le 1er janvier 2012, suite à la fusion des laboratoires du LAII (Laboratoire d'Automatique et d'Informatique Industrielle) et du LISI (Laboratoire d'Informatique Scientifique et Industrielle). Le laboratoire LIAS est composé de trois équipes de recherche : l'équipe Ingénierie des Données et des Modèles, l'équipe Systèmes embarqués Temps Réel et l'équipe Automatique & Système. Notre projet de fin d'études est proposé par le Pr. Allel Hadj Ali de l'équipe Ingénierie des données et des Modèles (IDD). L'équipe IDD s'intéresse aux différentes problématiques liées à la construction de systèmes de gestion de données permettant la collecte, l'intégration, la persistance des données et des modèles et l'exploitation des données d'une manière efficace, flexible et intelligente [23].

1.1.2 Equipe Archivos de l'Institut des textes et manuscrits modernes

Le CRLA-Archivos (Centre de Recherches Latino-Américaines-Archivos) est une équipe de l'Institut des Textes et Manuscrits Modernes –ITEM (UMR8132). Située à Poitiers, elle rassemble des chercheurs hispanistes et lusistes qui se consacrent à des travaux d'analyse critique des œuvres littéraires et des manuscrits des écrivains.

L'équipe réalise des travaux d'analyse critique de documents historiques, des études sur la genèse des œuvres, elle coordonne la préparation scientifique et éditoriale des éditions critiques et génétiques de la collection Archivos et organise également la conservation, la diffusion numérique et l'exploitation de fonds d'archives et de manuscrits. Ces travaux permettent de construire des bases de données [24].

1.2 Problématique

Les sciences humaines produisent actuellement des masses de données très variées mais peinent à proposer de nouvelles observations et connaissances à partir de celles-ci. Ces données sont souvent de qualité faible car collectées à l'aide de processus semi-automatisés et peuvent donc se révéler peu ou pas exploitables et requièrent une phase de « curation ». Le problème crucial se manifeste dans le stockage, la gestion et l'exploitation de ces données.

A vrai dire, pour étudier les archives et données produites par les experts, les chercheurs optent pour la construction de métadonnées, soit le recueil des propriétés des archives décrites : côte, sujet, auteurs, dates, notes, etc. A cette échelle, cette pratique, qui prend énormément de temps, se fait manuellement à l'aide de tableurs Microsoft Excel ou Ods: ceux-ci génèrent des problèmes et produisent des données bruitées, mal structurées et incertaines. De plus, d'autres bruits inhérents aux pratiques du domaine existent également : des dates de création des documents incertaines, des formats de dates différents, des champs vides, etc. Ces défaillances font forcément partie des problèmes à régler lors du nettoyage des données. Les chercheurs du domaine désirent donc organiser leurs informations et données d'archives sous formes de bases de données pour les explorer avec les méthodes des technologies modernes.

Plus spécifiquement, il s'agit de développer des outils d'extraction de connaissances, de fouilles de données ou de clustering adaptés au contexte des données d'archives avec toute leur imperfection et leur variété, d'une part, et permettant aux chercheurs scientifiques de découvrir et d'aller vers de nouveaux questionnements de leurs données, d'autre part.

L'an dernier, les experts du projet ont ressenti le besoin de repenser leurs méthodes en vue de l'exploitation de leurs corpus : ils ont fait appel aux services du laboratoire LIAS pour explorer ces données et fournir un outil d'analyse de données général qui permette d'exploiter ces données afin de pouvoir en extraire des connaissances.

Le prototype d'outil de recherche et d'exploitation des données de ces corpus devait permettre aux chercheurs des sciences humaines d'extraire des relations entre les

données de différents auteurs. Ce qui suit est un petit exemple pour comprendre la relation entre les auteurs : Si l'auteur A partage les mêmes mots-clés avec un autre auteur B dans une période de temps définie, sachant que l'auteur A n'a jamais collaboré avec l'auteur B, quelles possibilités de relations basées sur des critères prédéfinis peuvent être réalisables ?

L'outil Link&Pred permet de calculer les relations possibles entre des objets à partir de leurs métadonnées (propriétés/attributs). L'objectif essentiel de ce stage est de faire évoluer cet outil vers une prédiction plus pertinente fondée sur un ensemble d'attributs.

Dans un second temps, développer une plateforme qui offre de nouvelles fonctionnalités tout en améliorant l'interface pour qu'elle soit plus «user-friendly» afin de faciliter son fonctionnement et son exploitation.

1.3 Objectifs du projet

Après avoir présenté la problématique du projet nous allons énumérer les différents objectifs à réaliser pour ce projet :

- Parcourir les jeux de données.
- Intégrer une nouvelle méthode de prédiction de liens suivant les besoins.
- Amélioration de la précision des résultats.
- Développer une plateforme web munie d'une interface fonctionnelle et intuitive pour les non-informaticiens.
- Offrir de nouvelles fonctionnalités sur la plateforme web :
 - Upload et Download des fichiers.
 - Génération du graphe avec les nouvelles prédictions et amélioration de la visualisation des résultats.
 - Rechercher un auteur, visualiser ses liens et consulter l'historique des prédictions faites au préalable.
- Développer un système d'authentification pour sécuriser la plateforme et offrir des privilèges aux utilisateurs.

2 Etude de l'existant

2.1 Données des sciences humaines

En premier, l'archiviste et le chercheur collectent les différents types d'informations concernant les documents d'archives à partir des ressources et des archives éditées en ligne ou se trouvant dans les locaux de la Maison des Sciences de l'Homme et de la Société (MSHS). Ensuite, les informations sont saisies dans des fichiers CSV avec une structuration prédéfinie, un schéma de métadonnées, défini par le chercheur en lien avec ses domaines scientifiques. En dernier, les données collectées sont stockées dans des fichiers CSV.

2.2 Normalisation des données

Pour pouvoir travailler sur une masse d'informations importante, la normalisation des archives est une étape cruciale afin de pouvoir effectuer des traitements sur ces données.

Ce travail a été mis en œuvre l'année passée dans le cadre d'un projet de master et nous n'avons pas trouvé nécessaire de refaire cette tâche. Nous utiliserons des fichiers CSV normalisés. Nous allons brièvement décrire le travail déjà accompli dans les deux points suivants :

2.2.1 Evaluation de la qualité des données

Les données reçues ont été analysées selon plusieurs indicateurs : nombre de lignes, nombre de valeurs nulles, nombre d'uniques, compte des doublons, fréquence de la valeur, fréquence basse de valeur (Value low frequency), modèle date (Patterns).

2.2.2 Traitement des données

Les données ont été traitées selon les étapes suivantes : concaténation des fichiers, nettoyage de données (saut de ligne, format de dates, caractères spéciaux), normalisation de données (déterminer un seul référentiel de donnée et prendre un format de date précis), filtrage des données (trier les fichiers selon des critères et supprimer les vides).

2.3 Outil Link&Pred

2.3.1 Présentation de l'outil

Link&Pred est un outil d'analyse et d'exploitation des données d'archives qui a été développé en 2019, dans le but d'aider les archivistes dans leur travail. En effet, il sert également à soutenir des systèmes de recommandations qui aident à la décision dans le contexte de la recherche scientifique.

En premier, l'archiviste sépare les auteurs en relation dans des colonnes.

	A	B	C	D	E
1	Auteur 1	Auteur 2	Auteur 3	Auteur 4	Auteur 5
2	Goodpaster, Edwin W.				
3	Aínsa, Isabel				
4	Osorio Tejeda, Nelson				
5	Obligado, Alberto				
6	Gamarra, Antonio de				
7	Gamarra, Antonio de				
8	Anderson Imbert, Enrique				
9	Lyon, Ted				
10	Silver, Philip W.				
11	Fayard, Daniel				
12	Aínsa, Fernando				
13	Andreón, Roberto	Aínsa, Fernando			
14	Amigues Graham, M. L.				
15	Salom, Rodolfo	Aínsa, Fernando			
16	Kalenberg, Ángel	Aínsa, Fernando			
17	García Capurro, Federico				
18	Ricci, Julio				
19	Sanguinetti, Julio María				
20	Aratanha, Mario de				
21	Aínsa, Fernando				
22	Aínsa, Fernando				
23	Aínsa, Fernando				
24	Aínsa, Fernando				
25	Aínsa, Fernando				
26	Aínsa, Fernando				

Figure 4 : Liste d'adjacence entre les auteurs

Puis convertit le fichier CSV en fichier graphe (.net) grâce à l'outil Gephi.

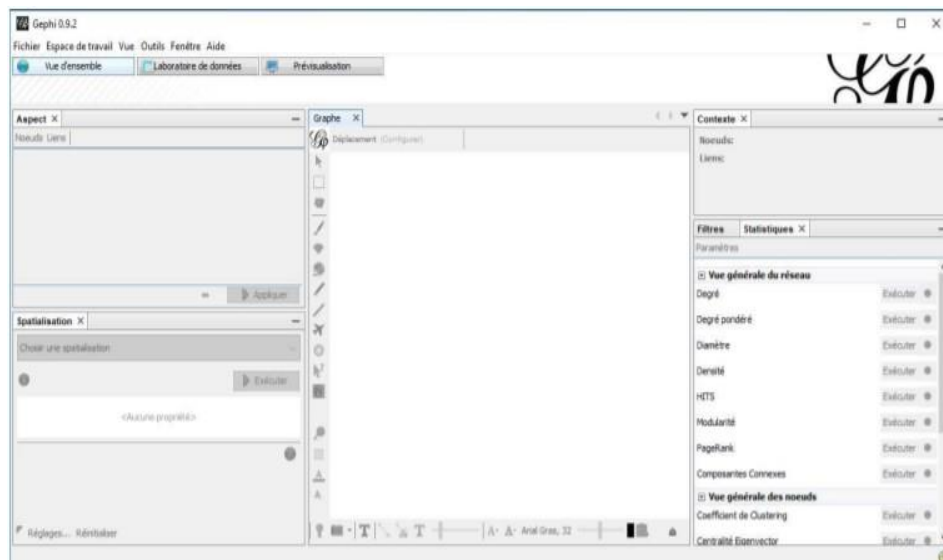


Figure 5 : L'outil Gephi

Ensuite, les quatre fonctionnalités que propose l'outil peuvent être utilisées :

- +Add : Ajouter un fichier graphe .net.
- -Del : Supprimer un fichier graphe .net.
- Self-loop Remover : Supprimer une boucle de lien.
- Predict : Prédire de futurs nouveaux liens entre les auteurs.

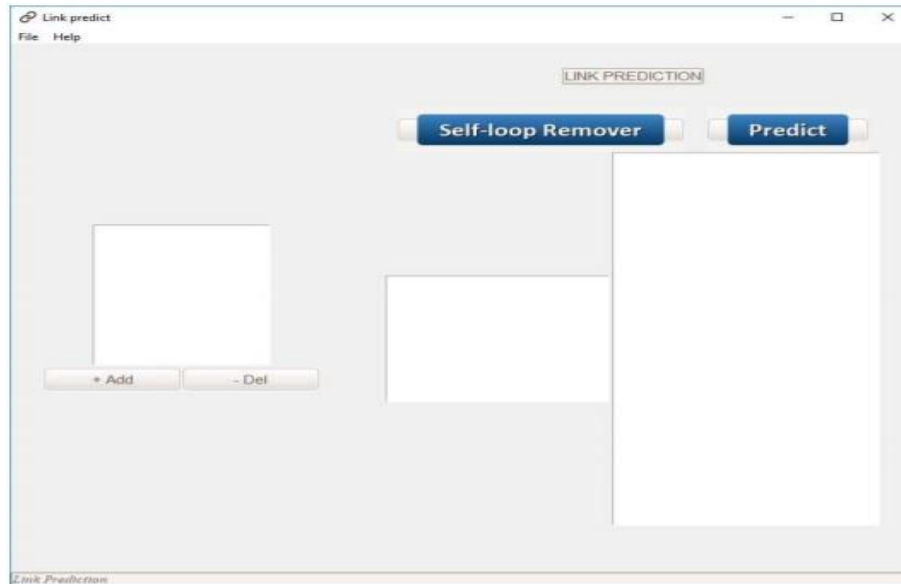


Figure 6 : L'interface de l'outil Link&Pred

Enfin, les résultats de la prédiction de liens s'afficheront sous le bouton Predict.

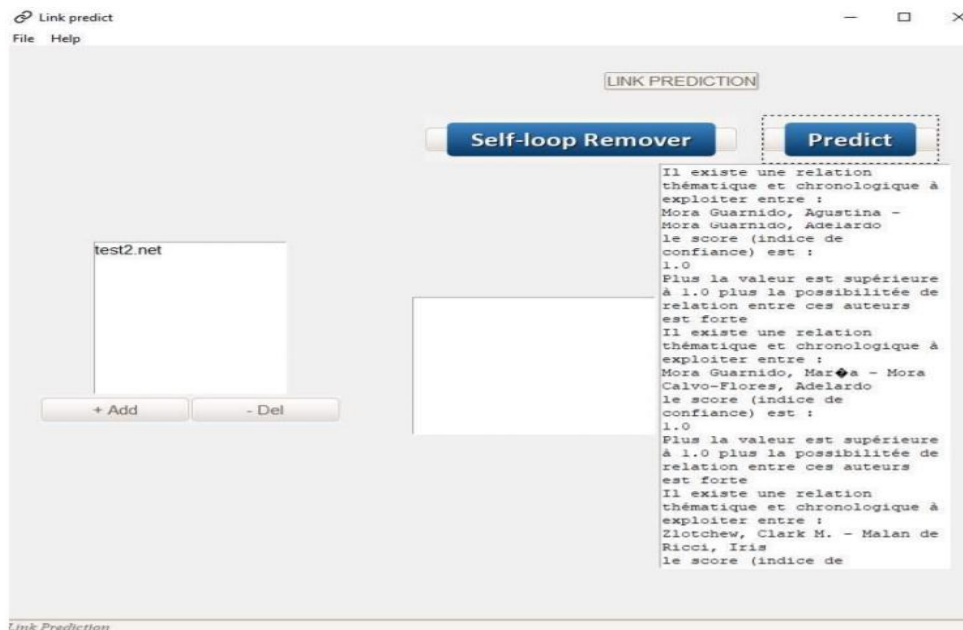


Figure 7 : Résultats de l'outil Link&Pred

L'application consiste à chercher des liens entre des auteurs (des données d'archives), ces données sont stockées dans des fichiers CSV de façon non-triées. Pour pouvoir arriver à ces résultats il faut transformer les CSV en liste d'adjacence puis générer un graphe (.net) à l'aide de l'outil gephi et c'est ce graphe qui est donné comme entrée à l'application, pour plus ou moins faciliter les études de ces archives aux chercheurs des sciences humaines l'outil a été conçu en utilisant des méthodes/techniques de prédictions liées au Machine Learning.

2.3.2 Insuffisances et limitations de l'outil

- La précision de l'algorithme utilisé pour la prédiction de liens peut être revue en vue d'un meilleur taux de précision.
- L'outil est limité à un certains nombres de résultats.
- Difficulté pour comprendre les différentes fonctionnalités de l'outil pour un utilisateur lambda et surtout l'enchaînement des étapes n'est pas intuitif.
- Absence des contrôles sur les types de fichiers introduits et absence de message lorsqu'il n'y a pas de nouvelle prédiction.
- Si on ajoute un fichier on doit toujours parcourir de nouveau tout le chemin vers le dossier où se trouvent les fichiers CSV.
- L'utilisateur doit supprimer les boucles avant de prédire les liens, donc la non-pertinence de la fonctionnalité Self loop remover.
- L'obligation de transformer les CSV en liste d'adjacence puis d'utiliser l'outil Gephi pour les transformer en graphe (.net).
- Résultat mal formulé et ambiguë.
- L'interface graphique n'est pas dynamique et ne respecte pas les règles d'Interfaces-Homme-Machine (IHM).

2.3.3 Solution proposée

Mise en place d'une plateforme web qui permet aux utilisateurs de générer des prédictions facilement interprétables et plus précises, visualiser les graphes, consulter l'historique des prédictions précédemment générés et télécharger les résultats obtenus, sécuriser la plateforme et gérer les droits d'accès.

3 Analyse

3.1 Intelligence Artificielle

L'intelligence artificielle (IA ou AI en anglais) est l'ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables d'imiter une forme d'intelligence réelle qui utilise des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critiquée.

Avec l'intelligence artificielle, l'homme côtoie un de ses rêves prométhéens les plus ambitieux : fabriquer des machines dotées d'un « esprit » semblable au sien. Nous utilisons l'IA énormément dans notre vie quotidienne comme avec les assistants vocaux (Siri, Cortana, Google Assistant) [25]. Les plus grandes entreprises de l'informatique l'utilise pour leurs algorithmes comme Google, YouTube ou Netflix. L'IA propose plusieurs avantages, entre autre, la limitation des erreurs de calcul, le remplacement de l'homme dans les tâches pénibles ou dangereuses et l'optimisation du travail qui aurait pu être réalisé par les hommes [26].

Parmi les nombreux champs d'études que propose l'intelligence artificielle se trouve le Machine Learning. En effet, le Machine Learning est une branche de l'intelligence artificielle qui concerne la conception, l'analyse, le développement et l'implémentation de modèles, permettant à une machine d'apprendre à partir des données par un processus systématique afin de remplir une tâche.

3.2 Machine Learning

Le Machine Learning est l'idée qu'il y a des algorithmes génériques qui peuvent vous dire quelque chose d'intéressant sur un ensemble de données sans que vous n'ayez à écrire n'importe quel code personnalisé spécifique au problème. Au lieu d'écrire du code, vous donnez des données à l'algorithme générique et il construit sa propre logique basée sur les données [27].

Par exemple, un type d'algorithme qui est un algorithme de classification peut placer des données dans différents groupes. Le même algorithme de classification utilisé pour reconnaître les chiffres manuscrits pourrait également être utilisé pour classer les courriels en pourriels et non pourriels sans changer de ligne de code. C'est le même

algorithmes, mais il alimente des données d'entraînement différentes, ce qui donne une logique de classification différente [28].

Voici un aperçu rapide de ce que l'apprentissage automatique (Machine Learning) est capable de faire :

- Santé : Prédire les diagnostics des patients pour les médecins.
- Réseau social : Prédire certaines préférences de match sur un site de rencontres pour une meilleure compatibilité.
- Finances : Prédire une activité frauduleuse sur une carte de crédit.
- Commerce électronique : Prédire le roulement des clients.
- Biologie : Trouver des modèles de mutations génétiques qui pourraient représenter le cancer.

De façon plus explicite, les machines « apprennent » en trouvant des modèles dans des données similaires. En effet, plus nous donnons de données à une machine, plus elle est intelligente. Cependant toutes les données ne sont pas les mêmes, plus l'information et les données obtenues sont bonnes, plus l'incertitude est réduite, et vice versa. Il est donc important de garder à l'esprit le type de données qu'on donne à la machine pour qu'elle apprenne.

3.2.1 Approches du Machine Learning

3.2.1.1 Approche supervisée

Dans l'apprentissage machine et l'intelligence artificielle, l'approche supervisée fait référence à une classe de systèmes et d'algorithmes qui déterminent un modèle prédictif en utilisant des points de données dont les résultats sont connus. Le modèle est appris par l'entraînement au moyen d'un algorithme d'apprentissage approprié qui fonctionne généralement par le biais d'une routine d'optimisation pour minimiser une fonction de perte ou d'erreur [29].

En d'autres termes, l'approche supervisée est le processus qui consiste à enseigner un modèle en lui fournissant des données d'entrée ainsi que des données de sortie correctes. Cette paire entrée/sortie est généralement appelée "données étiquetées". Elle

peut être comparée à l'apprentissage qui se déroule en présence d'un superviseur ou d'un enseignant.

Cette approche nécessite l'intervention d'une équipe d'experts et d'informaticiens. De plus, les scientifiques des données doivent eux-mêmes reconstruire les modèles pour s'assurer du bon déroulement de la phase d'apprentissage [30].

3.2.1.2 Approche non-supervisée

L'approche non-supervisée est une sorte d'apprentissage machine où un modèle doit chercher des motifs dans un ensemble de données non étiquetée et avec une supervision humaine minimale. Cela s'oppose aux techniques d'approches supervisées, où un modèle reçoit une série d'entrées et une série d'observations, et doit apprendre à établir une correspondance entre les entrées et les observations. Dans l'approche non-supervisée, seules les entrées sont disponibles, et un modèle doit rechercher des motifs dans les données. [31].

Les algorithmes d'approches non-supervisées permettent d'effectuer des tâches de traitement plus complexes que l'approche supervisée. Bien que l'approche non-supervisée puisse être plus imprévisible que d'autres méthodes naturelles d'apprentissage, les méthodes non supervisées aident à trouver des résultats sans avoir besoin de passer par une étape d'entraînement qui est généralement coûteuse en ressources et en temps d'exécution [32].

C'est cette approche que nous avons choisi pour réaliser nos différents objectifs.

3.2.1.3 Différence entre les deux approches

La principale différence entre les deux approches réside dans le fait que l'approche supervisée se fait sur la base d'une vérité fondamentale. En d'autres termes, nous avons une connaissance préalable de ce que devraient être les valeurs de sortie de nos échantillons.

Par conséquent, l'objectif de l'approche supervisée est d'apprendre une fonction qui, à partir d'un échantillon de données et des résultats souhaités, se rapproche le mieux de la relation entre entrée et sortie observable dans les données. En revanche, l'approche non-supervisée ne possède pas d'étape d'apprentissage. Son objectif est donc de déduire la structure naturelle présente dans un ensemble de points de données.

L'approche non-supervisée est beaucoup plus complexe puisqu'ici le système va devoir détecter les similarités dans les données qu'il reçoit et les organiser en fonction de ces dernières. Cette façon de travailler présente un avantage indéniable en ce sens que la phase de catégorisation de l'approche supervisée est un processus gourmand en ressources humaines et machines. Son élimination, ou tout du moins sa réduction retire un frein à l'implémentation de la technologie [33].

3.3 Représentation graphique

3.3.1 Graphe

La théorie des graphes est la discipline mathématique et informatique qui étudie les graphes, lesquels sont des modèles abstraits de dessins de réseaux reliant des objets.

Ces modèles sont constitués par la donnée de sommets (aussi appelés nœuds ou points) et d'arêtes (aussi appelées liens ou lignes). Les graphes constituent donc une méthode de pensée qui permet de modéliser une grande variété de problèmes en se ramenant à l'étude de sommets et d'arcs. Les derniers travaux en théorie des graphes sont souvent effectués par des informaticiens, du fait de l'importance qu'y revêt l'aspect algorithmique [34].

Les algorithmes élaborés pour résoudre des problèmes concernant les objets de cette théorie ont de nombreuses applications dans tous les domaines liés à la notion de réseau (réseau social, réseau informatique, télécommunications, etc.).

3.3.2 Réseaux

La théorie des réseaux est une représentation graphique des relations. Beaucoup de relations et de connexions dans le monde peuvent être modélisées ou décrites comme un réseau. La connexion d'amitié dans un réseau social, la connexion de vol globale, notre réseau neuronal de cerveau, et ainsi de suite. Les réseaux sont tout simplement très répandus, ils peuvent être trouvés dans de nombreux domaines, à savoir social, biologique, informationnel, et même logistique [35].

Elle peut être utilisée pour étudier l'influence d'un nœud particulier dans le réseau, par exemple, il peut aider à identifier qui est la personne la plus influente dans le

monde en analysant les informations de connexion dans l'application de réseautage social comme Twitter ou Facebook [36].

Le regroupement est une autre façon d'utiliser la théorie des réseaux. Nous pouvons regrouper les nœuds en fonction de leur interconnexion. Il peut aider à identifier le lien potentiel entre les nœuds, reliant les entités avec des attributs similaires. Il peut être utilisé pour prédire la relation future entre deux nœuds non liés, par exemple, A et B sont amis, B et C sont amis, A et C sont très susceptibles de devenir amis dans le futur basé sur la connexion réseau de voisins communs. Cette application est appelée prédiction de liens (Link Prediction en anglais), qui est utilisé dans les systèmes de recommandation de beaucoup de sociétés bien connues comme Amazon, Netflix et Spotify.

La théorie des réseaux est une partie essentielle de la théorie des graphes, elle est définie comme un graphe dans lequel les nœuds et/ou les arêtes ont des attributs (par ex. les noms) [37].

- Sommet (nœud/vertex) : représenté par un point dans le graphe
- Arête (edge) : ligne reliant deux sommets, montrant la relation/connexion entre deux nœuds

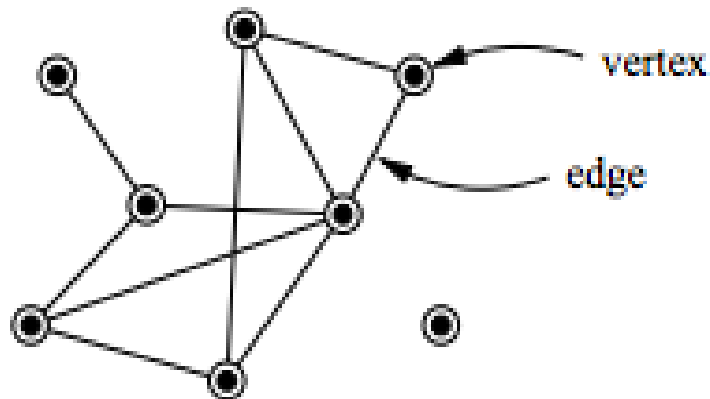


Figure 8 : Représentation d'un réseau à l'aide d'un graphe

Dans la théorie des graphes, un graphe orienté est un couple formé de un ensemble de nœuds et un ensemble d'arcs, chaque arc étant associé à un couple de nœuds selon une direction représentée par une flèche.

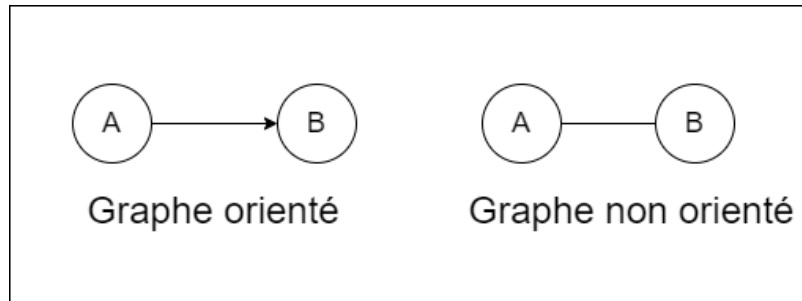


Figure 9 : Représentation d'un graphe orienté et d'un graphe non orienté

3.4 Réseau social

Le réseau social apparaît comme une structure sociale composée de différents nœuds du réseau. Chaque « nœud » signifie une personne ou une organisation [38].

D'une manière générale, un réseau social est une carte de nœuds et de connexions. Chaque nœud représente une entité unique. Ils peuvent être soit une personne ou un groupe. De nombreuses connexions / liens relient les nœuds entre eux. Ces liens peuvent être des relations familiales, des amis ainsi que des collègues.

De l'étude des graphes et de ses processus d'évolution dynamique, nous pouvons généralement trouver des informations précieuses qui peuvent nous aider à résoudre des problèmes pratiques dans le monde réel. Un réseau social est défini comme un réseau d'interactions ou de relations, dans lequel les nœuds sont constitués d'acteurs, et les arêtes représentent des relations ou des interactions entre les acteurs [39]. La généralisation de l'espace des réseaux sociaux est celle des réseaux d'information, dans lesquels les nœuds peuvent comprendre des acteurs ou des entités, et les contours indiquent les relations qui les unissent. De toute évidence, le concept de réseau social ne se limite pas au cas particulier d'un réseau social basé sur Internet tel que Facebook.

Il existe deux catégories de réseaux sociaux. La première porte sur les réseaux sociaux humains et la deuxième catégorie porte sur les réseaux sociaux en ligne. Le problème des réseaux sociaux a souvent été étudié dans le domaine de la sociologie

sous l'angle des interactions génériques entre tous les groupes d'acteurs. De telles interactions peuvent être dans n'importe quelle forme conventionnelle, que ce soit des interactions entre contacts différents, des interactions de télécommunication, des interactions de courrier électronique ou des interactions de courrier postal.

Dans la vie réelle, les individus ne sont pas indépendants, ils sont mutuellement contactés. Si nous prêtons attention aux attributs individuels, tout en ignorant les relations entre individus, nous risquons d'affecter l'exactitude et l'exhaustivité de l'analyse. L'analyse des réseaux sociaux peut expliquer les motifs cachés et les effets de ces relations. Elle est basée sur une hypothèse, à savoir, les individus dans les groupes sociaux sont interdépendants, et non autonomes. Le réseau social comprend un ensemble d'objets et de relations entre eux [40]. Ces relations peuvent être de tout type de relations sociales, comme l'amitié, etc. Les réseaux sociaux peuvent être représentés par un graphe.

Exemple dans un graphe G qui contient des nœuds V représentant les acteurs et des arêtes E représentant les liens entre les acteurs. Nous pouvons utiliser $G = (V, E)$ pour représenter le graphe G .



Figure 10 : Structure d'un réseau social

3.5 Prédiction de liens

La prédiction et la recommandation de liens constituent un point fondamental dans l'analyse des réseaux sociaux. Le principal défi de la prédiction de liens vient de la rareté des réseaux en raison de la forte disproportion des liens. La prédiction de liens consiste à utiliser des états de réseau observés précédemment pour trouver des connexions cachées ou prévoir des liens qui sont les plus susceptibles d'apparaître à l'avenir [41].

Différentes études ont montré qu'il est possible de prédire de nouvelles relations entre les éléments présents dans la topologie d'un réseau. Cette thématique qui consiste à chercher de nouvelles relations dans les réseaux est appelée prédiction de liens. Elle vise à prédire le comportement de lien, c'est-à-dire si une relation entre deux éléments dans un réseau peut être créée ou si une relation entre eux est manquante en se basant sur les relations actuellement observées. Beaucoup d'études et de recherches se sont orientées vers ce domaine compte tenu de son champ d'application. Pour cette raison plusieurs méthodes ont été conçues et appliquées pour rechercher et prédire des liens dans différents types de réseaux [42].

Il y a plusieurs approches pour traiter ce problème. Les plus populaires sont basées sur les caractéristiques des nœuds (contenu et/ou sémantique), les modèles de probabilités (apprentissage relationnel) et les approches topologiques/structurales.

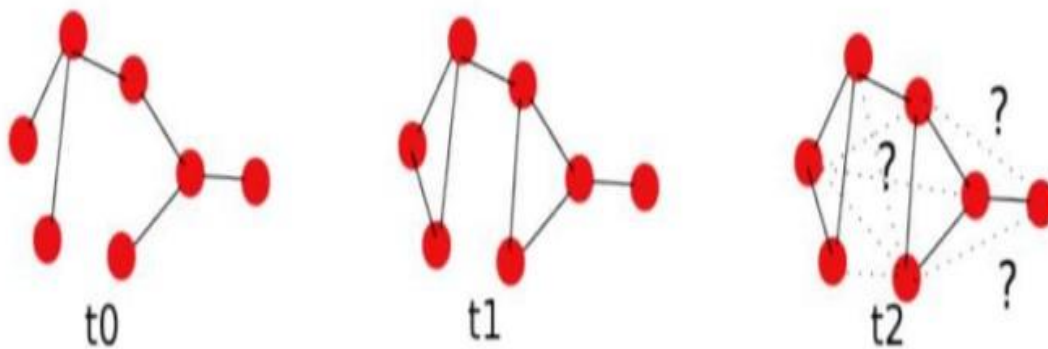


Figure 11 : Aperçu du processus de prévision de liens

3.5.1 Intérêt de la prédiction de liens dans les réseaux sociaux

L'analyse des réseaux sociaux est devenue un sujet de recherche populaire en informatique. Il est bien connu que la prédiction sur les réseaux sociaux est un domaine complexe et difficile, car les réseaux sociaux, particulièrement, ceux en ligne se composent d'un grand nombre d'utilisateurs comptant des millions de nœuds ou plus, voire des milliards d'arêtes. De plus, les données des réseaux sociaux en ligne sont très dynamiques. Les activités sociales des utilisateurs dans ce type de réseau sont imprévisibles. Le regroupement ou la sortie des utilisateurs, ainsi que l'émergence ou l'élimination des contours, peuvent survenir à tout moment [43].

Les relations dans les réseaux sociaux présentent une grande diversité. Les différents types de systèmes ont différents types de relations, leur degré d'importance, leur sens de l'orientation, etc.

Si nous pouvons prédire avec précision les limites qui seront créées entre deux nœuds du réseau pendant un intervalle de temps allant de t à un temps futur donné t' ($t' > t$), nous pouvons comprendre comment un réseau social évolue et quelle est la dynamique qui est derrière [44].

Plus important encore, étant donné que les liens du réseau représentent leur maintien et leur qualité reflétant les comportements sociaux d'individus et de communautés, la recherche de prédiction de liens peut donc être très utile pour l'évaluation quantitative et qualitative des relations humaines en cette ère d'informations où davantage de personnes participent à des communautés d'un réseau social en ligne ou humain par exemple les clubs sportifs. Ainsi, la prédiction de liens est une tâche importante dans l'analyse de réseau social.

3.5.2 Domaine d'application de la prédiction de liens

Les techniques de prédiction de liens ont trouvé un grand nombre d'applications dans des domaines très différents. Tout domaine dans lequel les entités interagissent de manière structurée peut potentiellement bénéficier de la prédiction de liens. Ces techniques sont utilisées pour améliorer la sélection des utilisateurs similaires dans les systèmes de recommandation qui adoptent une approche collaborative, ce qui permet d'obtenir de meilleurs résultats de recommandation.

Une application similaire est liée aux réseaux sociaux, qui sont devenus extrêmement populaires dans la société moderne. Les utilisateurs de ces systèmes s'attendent à disposer de mécanismes simples et efficaces leur permettant de se familiariser avec l'énorme quantité d'utilisateurs enregistrés. La plupart des réseaux sociaux utilisent des techniques de prédiction de liens pour suggérer automatiquement des connaissances avec un haut degré de précision [45].

Dans le domaine de la biologie, des techniques de prédiction de liens sont appliquées pour trouver des interactions possibles entre des paires de protéines dans un réseau d'interaction protéine-protéine (réseau PPI) [46].

Une autre application se trouve dans la prédiction de collaboration dans les réseaux de co-auteurs scientifiques. Les données de collaboration sont facilement accessibles, car certains sites d'indexation de journaux rendent publiques leurs collections. Les méthodes de prédiction de liens sont devenues un outil permettant de mieux comprendre les domaines de recherche des réseaux de prédiction de groupes d'auteurs ou de groupes de co-auteurs ou de collaboration potentielle à l'avenir.

3.5.3 Approches de la prédiction de liens

Les approches pour prédire des liens sont les suivantes :

3.5.3.1 Approche basée sur les nœuds

Dans l'approche basée sur les nœuds, les mesures de similitude sont adoptées pour associer les nœuds en fonction de leur contenu ou sémantique [47]. Les nœuds sont représentés comme un vecteur de caractéristiques qu'ils possèdent, et les mesures de similarité sont ensuite appliquées à des paires de nœuds pour déterminer leur proximité.

3.5.3.2 Approche probabiliste

L'approche probabiliste tente de trouver un modèle qui représente le mieux le réseau. L'idée est de construire un modèle probabiliste défini par un ensemble de paramètres θ , estimés en utilisant le réseau social observé. Ensuite, l'existence d'une

connexion entre une paire donnée de nœuds est déterminée par la probabilité conditionnelle $P(e^{<x,y>}|\theta)$ [48].

Des exemples de modélisations dans cette approche sont les réseaux de Markov relationnels, les réseaux relationnels et les réseaux de dépendance relationnelle.

3.5.3.3 Approche topologique

L'approche basée sur les modèles topologiques du réseau consiste à extraire des scores à partir de nœuds non connectés du réseau au moyen de métriques topologiques.

Ces mesures offrent un degré de similitude entre deux nœuds en explorant les structures du réseau en analyse. Ces scores sont ensuite utilisés comme base pour construire des modèles qui peuvent effectuer la prédiction.

L'approche basée sur la topologie est la plus répandue. Elle présente également de bonnes performances et est facile à mettre en œuvre [49].

3.5.4 Algorithmes de la prédiction de liens

3.5.4.1 Algorithme de Adamic Adar

L'algorithme consiste à calculer la proximité des nœuds en fonction de leurs voisins partagés, en effet, il s'intéresse plus particulièrement aux caractéristiques les plus rares pour prédire de futurs liens [50].

Cette mesure construit les voisins communs, mais plutôt que de compter ces voisins, il calcule la somme du log inverse du degré de chacun des voisins. Le degré d'un nœud est le nombre de voisins qu'il a, et l'intuition derrière cet algorithme est que quand il s'agit de triangles de fermeture, les nœuds de faible degré sont susceptibles d'être plus influents. Le score est calculé à l'aide de cette formule :

$$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

$N(u)$ est l'ensemble des nœuds adjacents à u . $N(x)$ est l'ensemble des nœuds adjacents à x , et $N(y)$ est l'ensemble des nœuds adjacents à y . Plus le score est élevé plus les nœuds sont proches.

3.5.4.2 *Algorithme Preferential Attachment*

L'algorithme est utilisé dans le but de calculer la proximité des nœuds, en fonction de leurs voisins partagés. L'intuition est que les nœuds avec beaucoup de relations gagneront plus de relations donc plus deux nœuds ont des voisins plus il y a de chances qu'un lien entre ces deux nœuds se crée [51]. Cette mesure est l'une des plus faciles à calculer, une telle possibilité est corrélée au produit du nombre de nœuds qu'ils ont.

$$PA(x, y) = |N(x)| * |N(y)|$$

$N(x)$ est l'ensemble des nœuds adjacents à x , et $N(y)$ est l'ensemble des nœuds adjacents à y .

3.5.4.3 *Algorithme Resource Allocation*

Cet indice de similitude est inspiré par les idées de ressources réseau complexes allouées dynamiquement. Dans les paires de nœuds (x, y) qui n'ont pas de lien direct, le nœud x peut allouer certaines ressources au nœud y par l'intermédiaire de leur voisin commun. Leurs voisins communs assument le rôle de passants. Dans le cas le plus simple, nous supposons que chaque passeur dispose d'une unité de ressource, il attribue ces ressources à ses voisins de façon égale. Par conséquent, la similarité du nœud x et du nœud y peut être définie comme le nombre de ressources que le nœud x obtient du nœud y [52]. Le calcul se fait à l'aide de la formule suivante :

$$RA(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{|N(u)|}$$

$N(u)$ est l'ensemble des nœuds adjacents à u . $N(x)$ est l'ensemble des nœuds adjacents à x , et $N(y)$ est l'ensemble des nœuds adjacents à y .

C'est une mesure utilisée pour calculer la proximité des nœuds en fonction de leurs voisins partagés. Si le RA est élevé alors les nœuds sont plus proches.

3.5.4.4 Algorithme SimRank

SimRank est une mesure de similarité largement adoptée pour les objets modélisés comme des nœuds dans un graphe, basée sur l'intuition que deux objets sont similaires s'ils sont référencés par des objets similaires [53].

En quelques sortes les objets liés aux objets similaires sont eux même similaires. Deux nœuds sont similaires si leurs voisins sont similaires. La formule est la suivante :

$$W(u, v) = \frac{c}{|N(u)| \cdot |N(v)|} \sum_{p \in N(u)} \sum_{q \in N(v)} W(p, q)$$

Point de démarrage : un nœud est totalement similaire à lui-même $W(u, v) = 1$

$N(u)$ et $N(v)$ sont les ensembles des nœuds adjacents respectivement à u et v .

c est un facteur d'importance compris entre 0 et 1

3.5.4.5 Algorithme Jaccard Similarity

C'est le calcul de coefficient des amis en commun entre deux nœuds sur la totalité de leurs amis. Ce qui veut dire qu'il mesure la possibilité que deux nœuds soient liés en calculant le rapport entre le nombre de voisins qu'ils partagent et le nombre total de voisins distincts qu'ils ont. Cette métrique résout le problème où deux nœuds peuvent avoir plusieurs voisins communs car ils ont beaucoup de voisins et non pas parce qu'ils sont fortement liés [54]. La formule est la suivante :

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$|A \cap B|$ est l'ensemble des voisins en commun de A et B, $|A \cup B|$ est l'ensemble de tous des voisins de A et B.

La fonction Jaccard Similarity calcule la similitude de deux listes de nombres. Nous pouvons l'utiliser pour calculer la similitude de deux listes codées.

3.5.4.6 Algorithme Common Neighbors

Comme son nom l'indique, cette mesure calcule le nombre de voisins communs qu'une paire de nœuds partage. S'il y a les mêmes nœuds voisins entre deux nœuds alors ces deux nœuds ont plus de chance d'avoir un lien entre eux [55]. Il est décrit comme suit: Le prédicteur commun-voisins capture la notion que deux étrangers qui ont un ami commun peuvent être introduits par cet ami.

Formellement la métrique est définie comme suit :

$$C(x, y) = |N(x) \cap N(y)|$$

$N(x)$ est l'ensemble des nœuds adjacents à x , et $N(y)$ est l'ensemble des nœuds adjacents à y .

- L'outil Link&Pred se base sur cet algorithme pour la prédiction de liens.

Dans le graphe ci-dessous, les nœuds A et D ont 2 voisins communs (nœuds B et C), alors que les nœuds A et E n'ont qu'un seul voisin commun (nœud B). Par conséquent, les nœuds A et D seraient considérés comme plus proches et plus susceptibles d'être reliés par un lien à l'avenir.

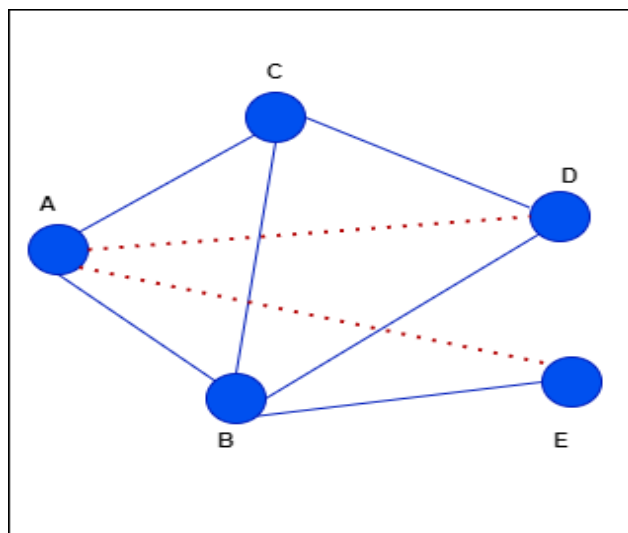


Figure 12 : Graphe pour comprendre Common Neighbors

3.5.5 Comparatif des différents algorithmes de la prédiction de liens

Tableau 1 : Tableau comparatif des différents algorithmes de prédiction de liens

Algorithmes de Prédiction	Performance		Input	Output	Approches	
	Small Dataset	Large Dataset			Machine Learning	Link Prediction
Common Neighbors	Performant	Performant	-Nœud 1 -Nœud 2 -Relation entre les nœuds 1 et 2 -La direction du lien entre les nœuds 1 et 2	Score de similarité	Non-supervisée	Basée sur le nœud
Adamic Adar	Performant	Pas performant				
Resource Allocation	Moyenne performance	Moyenne performance				
Preferential Attachment	Pas performant	Pas performant				
Jaccard Similarity	Pas performant	Moyenne performance				
SimRank	Performant	Moyenne performance	-Nœud 1 -Nœud 2 -La longueur de chemin entre les nœuds 1 et 2 -La direction du lien entre les nœuds 1 et 2	Score de similarité	Non-supervisée	Basée sur le chemin

3.5.6 Variante de Common Neighbors

Cette variante est une amélioration de Common Neighbors et elle se nomme CNGF, cette variante se repose sur le principe des nœuds voisins en communs ajoutant à cela le principe de la topologie du réseau.

En effet, pour calculer la similarité entre deux nœuds « A » et « B », il faut ajouter ce principe qui consiste entre autres à diviser réseau en un sous graphe qui contient les nœuds « A » et « B » et cela dans le but de prendre en considération les degrés des

voisins en communs. Cette différence met l'accent sur les différents chemins ignorés par le Common Neighbors classique et apporte une plus grande précision sur les scores calculés. C'est cet algorithme que nous avons retenu pour faire la prédiction de liens dans notre application. Nous présenterons cette variante en détail dans le chapitre suivant.

Conclusion

Dans ce chapitre, nous avons commencé par présenter notre projet et préciser l'étude de l'existant ainsi que certaines notions comme le Machine Learning. Ensuite, nous avons souligné les problèmes et les limites de l'outil Link&Pred et avons envisagé certaines améliorations. Enfin, nous avons expliqué la prédiction de liens en abordant les différents algorithmes existants.

Dans le prochain chapitre nous allons développer les aspects de conception, implémentation et réalisation de notre application.

Chapitre III

Conception, Implémentation et Réalisation

Introduction

La conception et la réalisation de notre solution est le résultat de la mise en œuvre de l'analyse présenté dans le chapitre précédent. C'est l'aboutissement final de notre projet et la raison d'être du projet lui-même.

Pour pouvoir mener à bien cette étape, il est nécessaire de choisir des technologies adaptées à sa réalisation.

Nous présentons dans ce chapitre la conception, l'implémentation et réalisation de notre application. Pour ce faire, nous avons structuré notre chapitre comme suit :

- Introduction de la notion de «Node Guidance Capability» et explication de façon plus approfondie l'algorithme utilisé, à savoir CNFG.
- Comparaison des différents algorithmes.
- L'aspect sécurité dans notre plateforme web.
- Conception de notre plateforme web dénommée Future Links
- Présentation de l'environnement de développement (outils, langages) utilisés durant cette phase.
- Présentation des fonctionnalités de notre plateforme web à travers ses interfaces utilisateurs.

1 Algorithme de la prédiction de liens utilisé « CNGF »

1.1 Notion de capacité de guidage des nœuds « Node Guidance Capability »

Cette notion est basée sur la densité du sous-graphe des voisins communs. Pour réaliser cette méthode, il faut extraire le sous-graphe contenant les nœuds « A » et « B » et leurs voisins communs.

Plus le sous-graphe des voisins communs est plus dense, plus la possibilité qu'un lien entre les nœuds « A » et « B » se forme est plus grande. Ce qui équivaut à dire que les nœuds du sous-graphe apportent une immense contribution dans la formation de liens.

Pour mieux expliquer cette notion, nous attribuons la densité du sous-graphe à chaque nœud. Si le voisin commun occupe plus de proportion dans le voisinage du nœud, alors il y a une plus grande capacité pour former un nouveau lien entre les nœuds « A » et « B » [56].

Nous définissons la formule de la Guidance d'un nœud comme suit :

$$\text{Guidance}(\mathbf{z}) = \frac{|\Phi(\mathbf{z})|}{\log(d_{\mathbf{z}})}$$

- \mathbf{z} est un nœud commun entre « A » et « B ».
- $d_{\mathbf{z}}$ est le degré du nœud \mathbf{z} dans le graphe original.
- $|\Phi(\mathbf{z})|$ est le degré de nœud \mathbf{z} dans le sous-graphe extrait.

Exemple 1 :

Nous allons expliquer cette notion introduite par l'exemple suivant :

La figure 13 présente deux graphes différents. Dans le premier graphe se trouvent les nœuds « X » et « Y » et dans le second graphe les nœuds « A » et « B » :

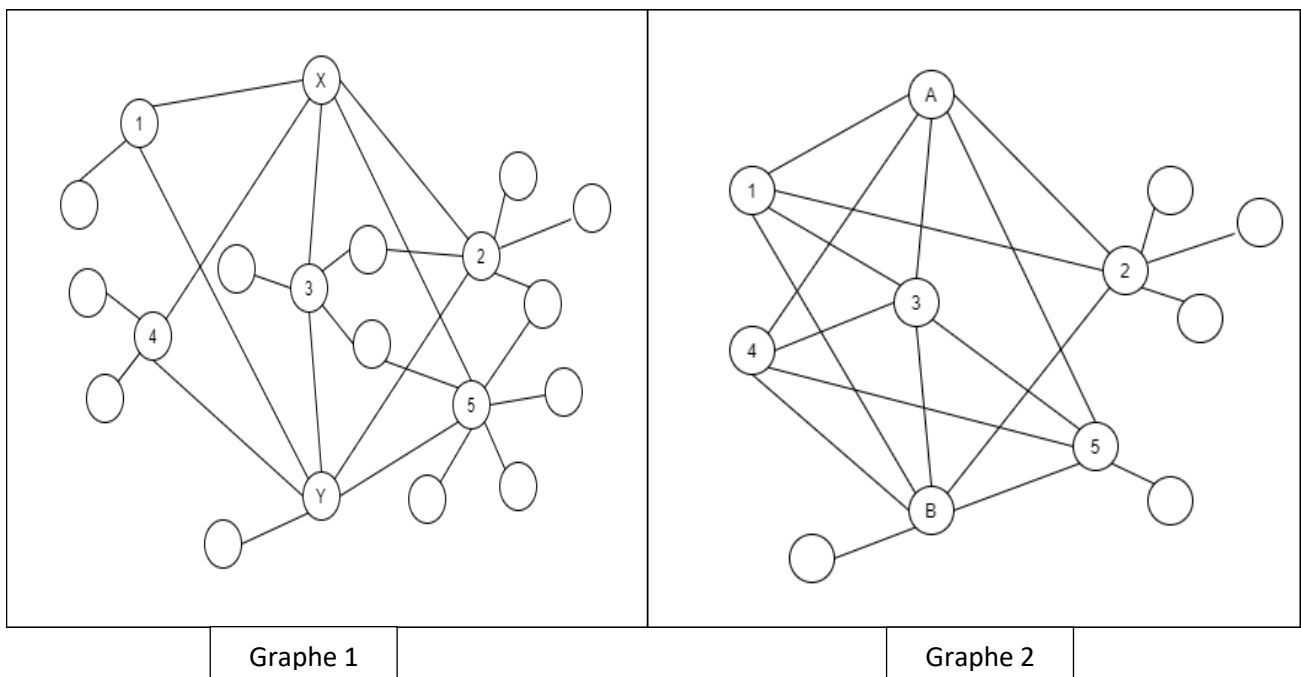


Figure 23 : Les graphes représentant l'exemple 1

La figure 14 présente les sous-graphes extraits à partir de la figure 13 :

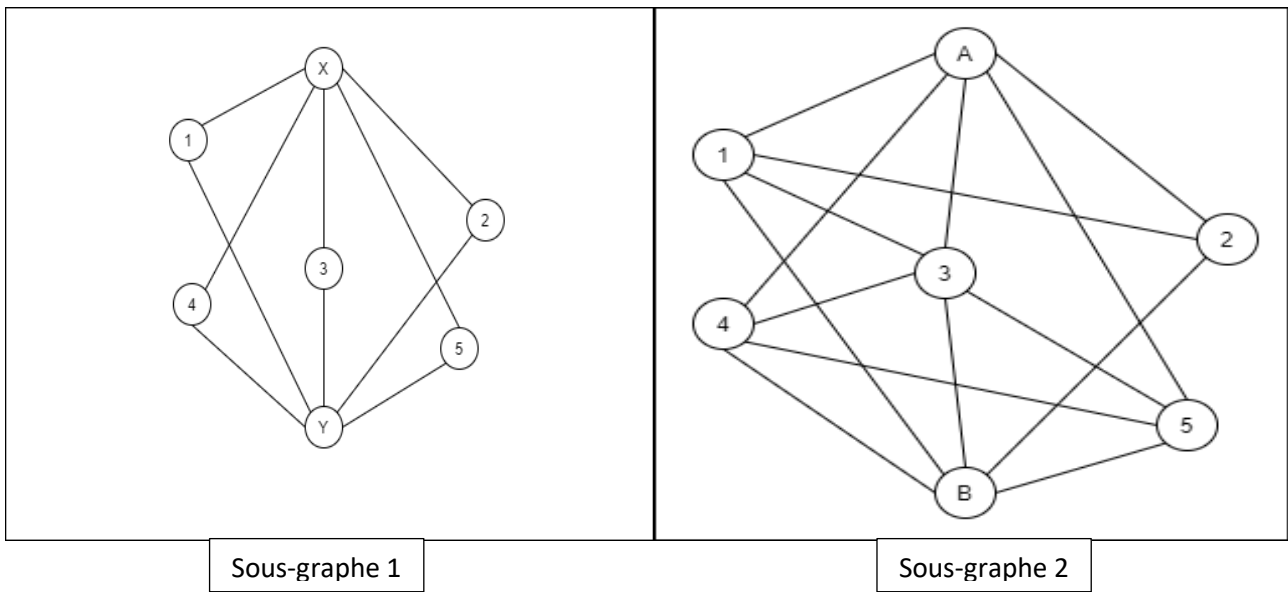


Figure 14 : Les sous-graphes représentant l'exemple 1

Observation : Pour la plupart des algorithmes de prédiction de liens, la similarité calculée entre le nœud « A » et le nœud « B » est la même que celle entre le nœud « X » et le nœud « Y ». Cependant, lorsque nous extrayons le sous-graphe qui contient le nœud « A », le nœud « B » et leurs voisins communs, ainsi que le sous-graphe qui contient le nœud « X », le nœud « Y » et leurs voisins communs, comme le montre la figure 14, nous pouvons voir qu'il y a beaucoup plus de chemins entre le nœud « A » et le nœud « B » que de chemins entre le nœud « X » et le nœud « Y ». Selon la notion de Node Guidance Capability la similarité entre le nœud « A » et le nœud « B » est plus élevée que la similarité entre le nœud « X » et le nœud « Y ».

1.2 Algorithme CNGF

En introduisant la notion de la capacité de guidage des nœuds (NGC), nous savons que les capacités de guidage des différents nœuds sont différentes. La similarité entre les deux nœuds peut être représentée par la somme des capacités de guidage de chaque nœud du voisinage commun entre « A » et « B ». Plus la capacité de guidage des voisins communs est grande, plus la possibilité d'un nouveau lien entre les deux

nœuds est plus grande. Sur cette base, nous définissons la formule du degré de similarité entre deux nœuds :

$$Similarity^{CNGF}(A, B) = \sum_{z \in N(A) \cap N(B)} \frac{|\Phi(z)|}{\log(d_z)}$$

- z est un nœud commun entre « A » et « B ».
- $N(A) \cap N(B)$ représente tous les nœuds en communs entre « A » et « B ».
- d_z est le degré du nœud z dans le graphe original.
- $|\Phi(z)|$ est le degré de nœud z dans le sous-graphe extrait.

Avec la formule précédente, nous pouvons donner le pseudo-code de l'algorithme CNGF basé sur la capacité de guidage des nœuds (voir Algorithme 1) :

Algorithme 1 :

Input (entrée) : graphe d'un réseau social $G = (V, E)$. Nœud x , nœud y .

Output (sortie) : la similarité entre le nœud x et le nœud y .

Description de l'algorithme :

- (1) Trouver l'ensemble des voisins communs de la paire de nœuds x et y .
- (2) Extraire le sous-graphe qui contient la paire de nœuds testée et leurs voisins communs.
- (3) **While** (l'ensemble des voisins en communs *is not null*) {
 - (4) Calculer le degré du nœud v , et *get v.degree* (récupération du degré de v), le nœud v est un nœud de l'ensemble des voisins communs.
 - (5) Calculer le degré du nœud v dans le sous graphe extrait dans l'étape 2, *get v.common_degree*.
 - (6) Calculer la capacité de guidage du nœud v ,
 $Guidance(v) = v.common_degree / \log(v.degree)$.
 - (7) La similarité entre les nœuds x et y : $Similarity.xy += Guidance(v)$.

Exemple 2 :

Calcul de la similarité entre les nœuds « A » et « B » ainsi que la similarité entre les nœuds « C » et « D » :

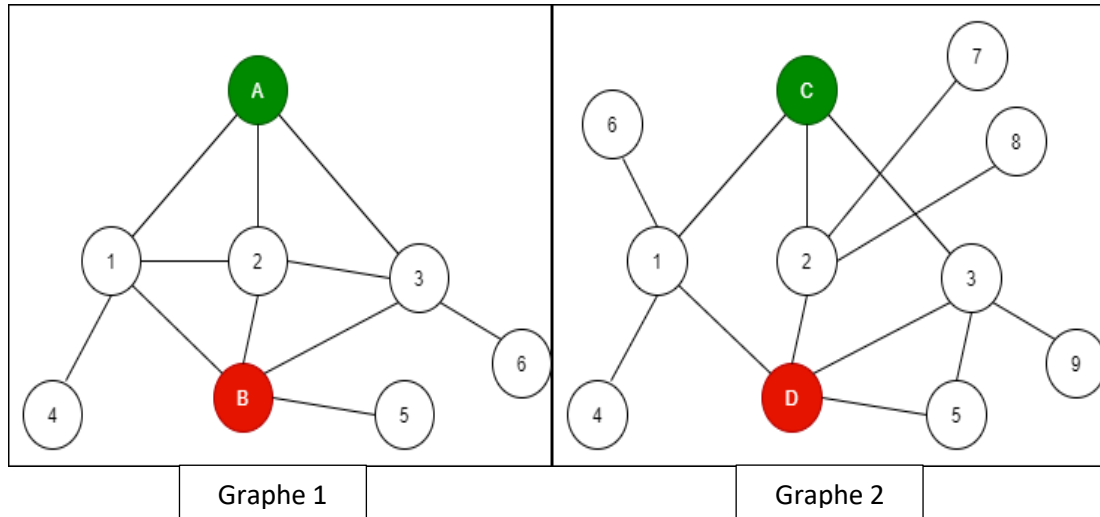


Figure 15 : Les graphes représentant l'exemple 2

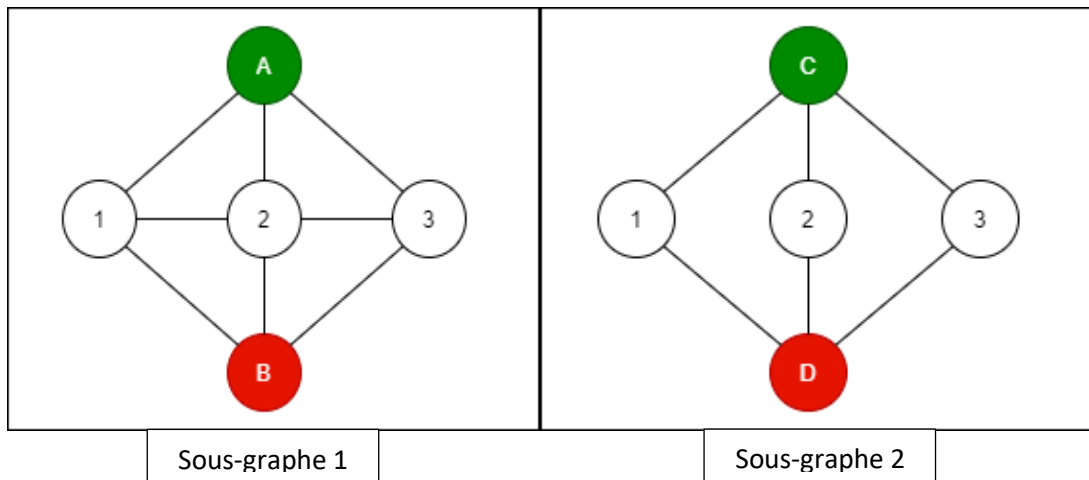


Figure 16 : Les sous-graphes représentant l'exemple 2

Calcul de la similarité du graphe 1 : $Score_{AB} = \frac{3}{\log 4} + \frac{4}{\log 4} + \frac{3}{\log 4} = 16.60$

Calcul de la similarité du graphe 2 : $Score_{CD} = \frac{2}{\log 4} + \frac{2}{\log 4} + \frac{2}{\log 4} = 13.28$

Nous constatons que la similarité entre les nœuds « A » et « B » est plus grande que celle entre « C » et « D ».

2 Comparaison des Algorithmes de la prédiction de liens

A partir d'une étude préalable publiée par Sahil Gupta et ses collègues dans l'International Journal of Computer Applications (0975-8887) [57], des résultats ont été obtenus à partir des tests de comparaison sur les performances des différents algorithmes de la prédiction de liens. Nous mentionnons ces résultats dans les deux tableaux ci-dessous :

- True Positive (TP): liens prévus correctement qui sont en fait des liens.
- True Negative (TN): nombre de non-liens correctement prédits.
- False Positive (FP): nombre de liens prévus qui ne sont pas des liens.
- False Negative (FN): nombre de liens non prédits qui sont en fait des liens.

$$\text{Calcul de l'Accuracy (ACC)} \longrightarrow \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Calcul de True Positive Rate (TPR)} \longrightarrow \text{TPR} = \frac{TP}{TP+FN}$$

$$\text{Calcul de Precision (PRE)} \longrightarrow \text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Calcul de True Negative Rate (TNR)} \longrightarrow \text{TNR} = \frac{TN}{TN+FP}$$

Tableau 2 : Mesure des performances pour les petits ensembles de données.

	TPR	TNR	PRE	ACC
Common Neighbors	0.588	0.942	0.910	0.765
Jaccard's Coefficient	0.803	0.488	0.611	0.646
Adamic Adar	0.614	0.932	0.900	0.773
Preferential Attachment	0.628	0.716	0.688	0.672
Resource Allocation	0.613	0.932	0.900	0.773
CNGF	0.713	0.963	0.951	0.838

Tableau 3 : Mesure des performances pour les grands ensembles de données.

	TPR	TNR	PRE	ACC
Common Neighbors	0.729	0.963	0.952	0.846
Jaccard's Coefficient	0.862	0.888	0.885	0.875
Adamic Adar	0.862	0.888	0.885	0.875
Preferential Attachment	0.762	0.660	0.691	0.711
Resource Allocation	0.824	0.930	0.922	0.877
CNGF	0.812	0.975	0.971	0.893

Nous déduisons de cette étude que l'algorithme Common Neighbors et l'algorithme CNGF donnent les meilleurs résultats par rapport aux autres algorithmes avec des précisions de 95% pour le Common Neighbors et 97% pour le CNGF.

2.1 Comparaison entre Common Neighbors et CNGF

Comme vu précédemment lors de la comparaison des algorithmes de prédiction de liens, deux algorithmes sortent du lot à savoir l'algorithme Common Neighbors et l'algorithme CNGF.

Dans cette partie nous allons comparer et mettre en évidence les différences entre ces deux algorithmes. Nous allons répartir cette étude en trois points : le premier point concerne un exemple de calcul de similarité, le deuxième présente le graphe de comparaison et le dernier se rapporte à la comparaison des résultats obtenus par la prédiction de liens de l'outil Link&Pred basée sur Common Neighbors et ceux obtenus dans notre plateforme basée sur CNGF.

- ***Exemple de calcul de similarité :***

Dans la continuité de l'exemple 2, figures 15 et 16 ci-dessus, nous allons calculer les scores de similarité pour les paires de nœuds « A, B » et « C, D » avec les deux algorithmes :

Tableau 4 : Différence de similarité entre Common Neighbors et CNGF.

	Algorithme Common Neighbors	Algorithme CNGF
Score de similarité entre A et B	3	16.60
Score de similarité entre C et D	3	13.28

Observation : Dans le cas de l'algorithme Common Neighbors, nous avons le $scoreCN_{AB} = scoreCN_{CD} = 3$ (le nombre des voisins en communs entre les deux paires de nœud), puisqu'il y a une égalité des scores, nous avons une possibilité équivalente de prédiction de liens entre la paire (A, B) et la paire (C, D).

Par contre dans le cas de CNGF, nous avons le $scoreCNGF_{AB} > scoreCNGF_{CD}$ car cette algorithme prend en considération la notion de « Node Guidance Capability », ce qui équivaut à dire qu'il y a une plus grande possibilité de prédire un lien entre A et B qu'un lien entre C et D. Nous concluons donc sur cet exemple que le score de CNGF est bien plus précis que le score de Common Neighbors.

- **Le graphe de comparaison entre Common Neighbors et CNGF :**

La courbe Receiver Operating Characteristic (ROC) est la variation du True Positive Rate (TPR) par rapport au False Positive Rate (FPR) à différents seuils. L'axe des x représente le FPR et l'axe des y représente le TPR. En fixant des seuils différents, nous calculons le True Positive Rate (TPR) et le False Positive Rate (FPR) respectivement. Ensuite, nous déduisons le taux ROC correspondant à la courbe.

Tout d'abord, nous donnons la courbe ROC de l'algorithme local Common Neighbors et la courbe ROC de l'algorithme CNGF, comme le montre la figure 17. La courbe ROC montre l'impact global des différents seuils pour la prédiction de liens. Nous pouvons constater dans la figure 17 [58] que les performances des deux algorithmes sont très similaires au début ; cependant, avec l'augmentation du False Positive Rate, le

True Positive Rate de CNGF est supérieur à celui de CN. Cela montre que la capacité de prédiction de l'algorithme de CNGF est meilleure.

Calcul de True Positive Rate (TPR) $\rightarrow TPR = \frac{TP}{TP+FN}$

Calcul de False Positive Rate (FPR) $\rightarrow FPR = \frac{FP}{TN+FP}$

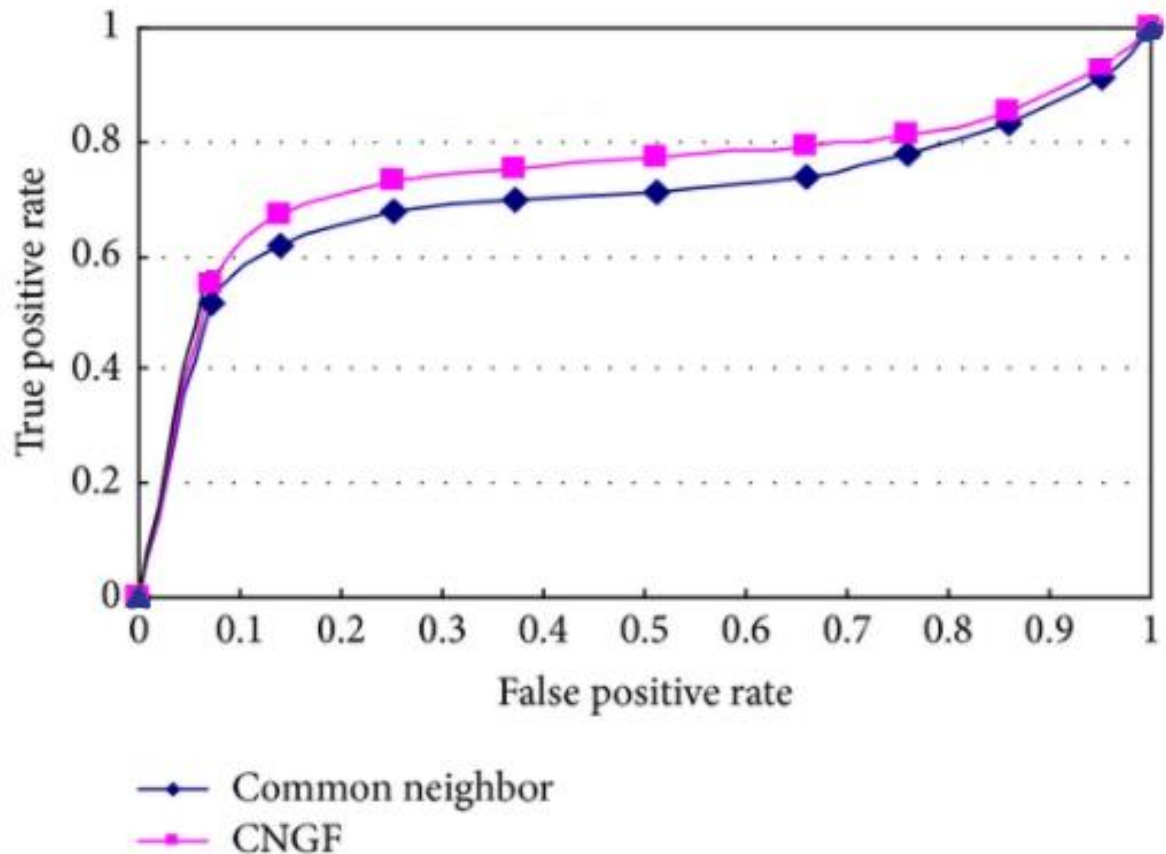


Figure 17 : Le graphe de la courbe ROC.

- *Comparaison des résultats de la prédiction de liens de l'outil Link&Pred et ceux de la prédiction de liens de notre plateforme Future Links :*

Nous montrons dans ce point la différence entre les résultats de la prédiction de liens obtenus par l'application Link&Pred qui utilise l'algorithme Common Neighbors et les résultats obtenus par la prédiction de liens de notre plateforme qui utilise l'algorithme CNGF sur les données d'archives fournies et sur des fichiers similaires trouvés sur le net.

Test 1 :

- **Link&Pred**

over **Predict**

Il existe une relation thématique et chronologique à exploiter entre :
Mora Guarnido, Agustina - Mora Guarnido, Adelardo
le score (indice de confiance) est : 1.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre :
Mora Guarnido, Mar◆a - Mora Calvo-Flores, Adelardo
le score (indice de confiance) est : 1.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre :
Zlotchew, Clark M. - Malan de Ricci, Iris
le score (indice de confiance) est : 1.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre :
Mora Guarnido, Rafael - Mora Guarnido, Adelardo
le score (indice de confiance) est : 2.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Figure 18 : Résultats du test 1 avec Link&Pred

- **Future Links**

Noeud Source	Noeud Destination	Score	Proportion
Mora Guarnido, Rafael	Mora Guarnido, Adelardo	11.5275	100.00 %
Malan de Ricci, Iris	Zlotchew, Clark M.	4.9829	43.23 %
Mora Guarnido, Agustina	Mora Guarnido, Adelardo	4.1918	36.36 %
Mora Guarnido, Mar ◆ a	Mora Calvo-Flores, Adelardo	3.1439	27.27 %

Figure 19 : Résultats du test 1 avec Future Links

Observation : La différence qui subsiste dans ce test est dans les trois dernières prédictions (Figure 19). Avec l'outil Link&Pred les liens ont la même possibilité d'apparaître mais sur notre plateforme Future Links, les liens n'ont pas la même possibilité d'apparaître. La prédiction est donc plus précise.

Test 2 :

- *Link&Pred*

ver **Predict**

Il existe une relation thématique et chronologique à exploiter entre : Fischer, TC - Dinh, QT
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Glenisson, P - Debackere, K
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Kusma, B - Chung, KF
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure

le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Eowles, CA - Bowles, CA
le score (indice de confiance) est : 8.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Eowles, CA - Briggs, MB
le score (indice de confiance) est : 8.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Kusma, B - Dinh, QT
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Leta, J - Debackere, K
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre :

exploiter entre : Thijs, B - De Moor, B
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Wooding, S - Frame, I
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Il existe une relation thématique et chronologique à exploiter entre : Young, T - Wooding, S
le score (indice de confiance) est : 4.0
Plus la valeur est supérieure à 1.0 plus la possibilité de relation entre ces auteurs est forte

Figure 20 : Résultats du test 2 avec Link&Pred

- ***Future Links***

Prédictions :			
Noeud Source	Noeud Destination	Score	Proportion
Bowles, CA	Eowles, CA	68.1612	100.00 %
Briggs, MB	Eowles, CA	68.1612	100.00 %
Donnadieu, S	Caillieux, N	24.2765	35.62 %
Kusma, B	Dinh, QT	23.6659	34.72 %
Fischer, TC	Dinh, QT	23.6659	34.72 %
Chung, KF	Kusma, B	23.6659	34.72 %
Fischer, TC	Chung, KF	23.6659	34.72 %
Wooding, S	Young, T	22.5399	33.07 %
Wooding, S	Frame, I	22.5399	33.07 %
Covert-vail, L	Onghena, P	19.9316	29.24 %
Debackere, K	Leta, J	19.0302	27.92 %
Thijs, B	De Moor, B	17.539	25.73 %
Mustar, P	Mercier, S	16.5858	24.33 %

Figure 21 : Résultats du test 2 avec Future Links

Observation : La différence qui subsiste dans ce test est dans le classement et le nombre de nouveaux liens prédits : avec l’outil Link&Pred il y a des liens qui s’affichent où les possibilités sont faibles (ces liens ont le même score de similarité avec le Common Neighbors), par contre dans notre plateforme Future Links le classement est respecté et tous les nouveaux liens sont prédits.

Déduction générale : Après avoir constaté les différences entre Common Neighbors et CNGF sur plusieurs points, exemples, graphe, et tests notamment les deux exemples présentés dans ce mémoire nous concluons que l’algorithme CNGF est bien plus précis et affiche plus de nouveaux liens que l’algorithme Common Neighbors.

3 Aspect sécurité dans notre plateforme

3.1 Importance de la sécurité dans le web

L'évolution des technologies de l'information et de la communication, notamment avec le développement d'Internet, a fait que les réseaux et les systèmes d'informations jouent désormais un rôle crucial dans notre société, [59]. Les cyber-attaques sont de plus en plus fréquentes, il est donc nécessaire de savoir comment nous pouvons protéger nos informations confidentielles. C'est pourquoi il est important de s'informer sur la cyber-sécurité et la sécurité informatique pour empêcher les hackers et les cyber-voleurs d'accéder aux informations des utilisateurs, et de voler des données sensibles. Sans une stratégie de sécurité proactive, les entreprises risquent la propagation et l'escalade des logiciels malveillants, des attaques sur d'autres sites web, réseaux et autres infrastructures informatiques.

Il existe de nombreuses façons de sécuriser une application web, notamment en mettant en place le protocole HTTPS sur notre site.

3.2 Protocole HTTPS

Le protocole de transfert hypertexte sécurisé (HTTPS) est la version sécurisée de HTTP, qui est le protocole principal utilisé pour envoyer des données entre un navigateur web et un serveur web. HTTPS est chiffré afin d'augmenter la sécurité du transfert de données. Tous les sites web, en particulier ceux qui nécessitent des informations de connexion, doivent utiliser HTTPS [60].

HTTPS permet au visiteur de vérifier l'identité du site web auquel il accède, grâce à un certificat d'authentification émis par une autorité tierce, réputée fiable (et faisant généralement partie de la liste blanche des navigateurs internet). Il garantit théoriquement la confidentialité et l'intégrité des données envoyées par l'utilisateur et reçues par le serveur. Il peut permettre de valider l'identité du visiteur, si celui-ci utilise également un certificat d'authentification client.

HTTPS est généralement utilisé pour les transactions financières en ligne : commerce électronique, banque en ligne, courtage en ligne, etc. Il est aussi utilisé pour la consultation de données privées, comme les courriers électroniques, par exemple.

Exemple :

- Avant le chiffrement :

Il s'agit d'une chaîne de texte parfaitement lisible

- Après le chiffrement :

ITM0IRyiEhVpa6VnKyExMiEgNveroyWBPlgGyfkflYjDaaFf/Kn3bo3OfghBPDWo6
AfSHINtL8N7ITEwIXc1gU5X73xMsJormzzXlwOyrCs+9XCPk63Y+z0=

3.3 Authentification

L'authentification pour un système informatique est un processus permettant au système d'assurer la légitimité de la demande d'accès faite par une entité (être humain ou un autre système, etc.) afin d'autoriser l'accès de cette entité à des ressources du système, conformément au paramétrage du contrôle d'accès. L'authentification permet donc, pour le système, de valider la légitimité de l'accès de l'entité, ensuite le système attribue à cette entité les données d'identité pour cette session (ces attributs sont détenus par le système ou peuvent être fournis par l'entité lors du processus d'authentification). C'est à partir des éléments issus de ces deux processus que l'accès aux ressources du système pourra être paramétré (contrôle d'accès) [61].

La majorité des sites web utilise l'authentification par mot de passe. En effet, dans les réseaux informatiques privés et publics, notamment Internet, l'authentification implique couramment l'utilisation d'un identifiant de connexion (nom d'utilisateur) et d'un mot de passe. L'utilisateur qui connaît les informations de connexion est réputé authentique.

Chaque utilisateur commence par s'inscrire (ou est inscrit par une autre personne, comme un administrateur système) avec un mot de passe attribué ou choisi. A chaque

utilisation suivante, l'utilisateur doit fournir le mot de passe précédemment déclaré. Pour assurer la sécurité du système il faut bien conserver les mots de passe.

Dans notre plateforme les mots de passe sont stockés après être hachés par l'algorithme BCrypt. Avant d'introduire BCrypt nous allons expliquer la fonction de hachage.

3.4 Fonction de Hachage

On nomme fonction de hachage, de l'anglais hash function (hash : pagaille, désordre, recouper et mélanger), c'est une fonction particulière qui, à partir d'une donnée fournie en entrée, calcule une empreinte numérique servant à identifier rapidement la donnée initiale, au même titre qu'une signature pour identifier une personne. Les fonctions de hachage sont utilisées en informatique et en cryptographie notamment pour reconnaître rapidement des fichiers ou des mots de passe.

L'ordinateur ne va pas envoyer le mot de passe au serveur, mais une signature du mot de passe. Le serveur ne va pas enregistrer le mot de passe mais enregistrera cette signature. Lorsque l'utilisateur se connectera, le serveur ne va pas vérifier si le mot de passe est identique, mais il va vérifier que la signature du mot de passe saisi est bien la même que la signature du mot de passe enregistrée.

Ce qui est intéressant dans le hachage, ce n'est pas les données elles-mêmes mais leurs signatures. Puisque chaque donnée a sa propre signature, on peut se dire que si les signatures sont identiques alors les données sont identiques, et à l'inverse, si les signatures sont différentes alors les données sont forcément différentes. Le hachage est donc utilisé pour comparer les données (en comparant les signatures) [62].

3.5 Algorithme BCrypt

L'algorithme de hachage BCrypt est une fonction de hash issu à la base de l'algorithme blowfish (type de cryptage). Cette fonction de hachage a plusieurs avantages. Tout d'abord elle utilise nativement un salt (un salt est une séquence qui est

rajoutée à un mot de passe pour en améliorer la sécurité), le salt est généré aléatoirement, ce qui empêche la création des look-up tables, qui sont considérés comme des failles dans les fonctions de hachage plus anciennes (md5, sha1, etc). En fait, une look-up table est possible mais elle demanderait un espace de stockage phénoménal, ainsi qu'une puissance de calcul immense pour sa création, puisqu'il faudrait stocker tous les salts possibles. La consultation d'une telle table demanderait aussi des ressources considérables [63]. L'autre avantage du BCrypt est que l'on peut choisir le nombre d'itérations pour rendre le résultat du hachage plus long et donc plus difficile à « brute-forcer ». Avec l'avancée des capacités hardware, on peut imaginer un jour une puissance de calcul suffisante pour rendre le BCrypt moins sécurisé, mais pour le moment, cela reste une des fonctions de hachage les plus difficiles à « casser » [64].

3.6 Protection contre Cross-Site Request Forgery

En sécurité des systèmes d'information, le cross-site request forgery abrégé CSRF est un type de vulnérabilité des services d'authentification web [65]. L'objet de cette attaque est de transmettre à un utilisateur authentifié une requête HTTP falsifiée qui pointe sur une action interne au site, afin qu'il l'exécute sans en avoir conscience et en utilisant ses propres droits. L'utilisateur devient donc complice d'une attaque sans même s'en rendre compte. L'attaque étant actionnée par l'utilisateur, un grand nombre de systèmes d'authentification sont contournés.

Pour se protéger du CSRF nous pouvons utiliser des jetons de validité (ou Token) dans les formulaires par la création d'un Token lors de l'envoi des formulaires à l'utilisateur, une fois le formulaire rempli et envoyé vers le serveur, ce dernier doit alors vérifier la correspondance du jeton envoyé et du jeton reçue.

4 Conception de la plateforme web Future Links

4.1 Langage UML

Le langage UML (Unified Modeling Language, ou langage de modélisation unifié) a été pensé pour être un langage de modélisation visuelle commun, et riche sémantiquement et syntaxiquement. Il est destiné à l'architecture, la conception et la mise en œuvre de systèmes logiciels complexes par leur structure aussi bien que leur comportement. L'UML a des applications qui vont au-delà du développement logiciel, notamment pour les flux de processus dans l'industrie.

UML nous fournit donc des schémas appelés diagrammes pour représenter le logiciel à développer : son fonctionnement, sa mise en route, sa limite, sa structure, les actions susceptibles d'être effectuées par le logiciel, etc. [66].

Initialement, les buts des concepteurs d'UML étaient les suivants :

- ❖ Représenter des systèmes entiers (pas uniquement logiciels) par des concepts objets.
- ❖ Lier explicitement des concepts et le code qui les implantent.
- ❖ Pouvoir modéliser des systèmes à différents niveaux de granularité, (pour permettre d'appréhender des systèmes complexes).

4.2 Diagramme des cas d'utilisation de notre cas d'étude

Le diagramme des cas d'utilisation est un diagramme UML utilisé pour donner une vision globale du comportement fonctionnel d'un système logiciel. Un cas d'utilisation représente une unité discrète d'interaction entre un utilisateur (Humain ou Machine) et un système. Il est une entité significative de travail. Dans un diagramme de cas d'utilisation, il existe des acteurs qui interagissent avec des cas d'utilisation. Les cas d'utilisation permettent de structurer les besoins des utilisateurs et les objectifs du système. Une fois identifiés et structurés, ces besoins permettent d'identifier les fonctionnalités principales ou critiques du système [67].

Présentation du diagramme des cas d'utilisation dans notre cas d'étude :

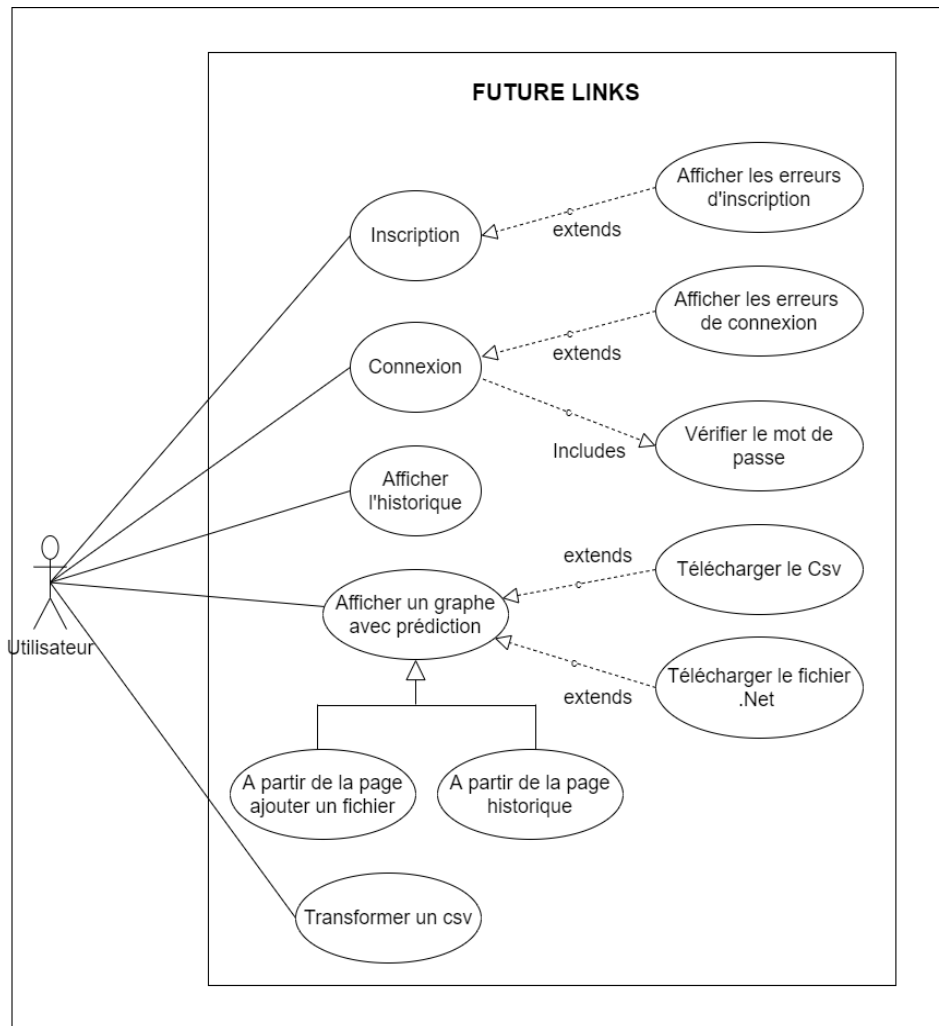


Figure 22 : Diagramme des cas d'utilisation.

4.3 Diagrammes de classes dans notre cas d'étude

Le diagramme des classes est un diagramme structurel (statique) qui montre la structure interne du système. Il permet de représenter les classes (attributs + méthodes) et les associations (relations) entre ces classes. Il est nécessaire lors de la modélisation objet d'un système.

Une classe est une représentation abstraite d'un ensemble d'objets, elle contient les informations nécessaires à la construction de l'objet (c'est-à-dire la définition des attributs et des méthodes). La classe peut donc être considérée comme le modèle, le moule ou la notice qui va permettre la construction d'un objet. On dit également qu'un objet est l'instance d'une classe (la concrétisation d'une classe) [68].

Nous présentons ci-dessous le diagramme de classes de notre cas d'étude :

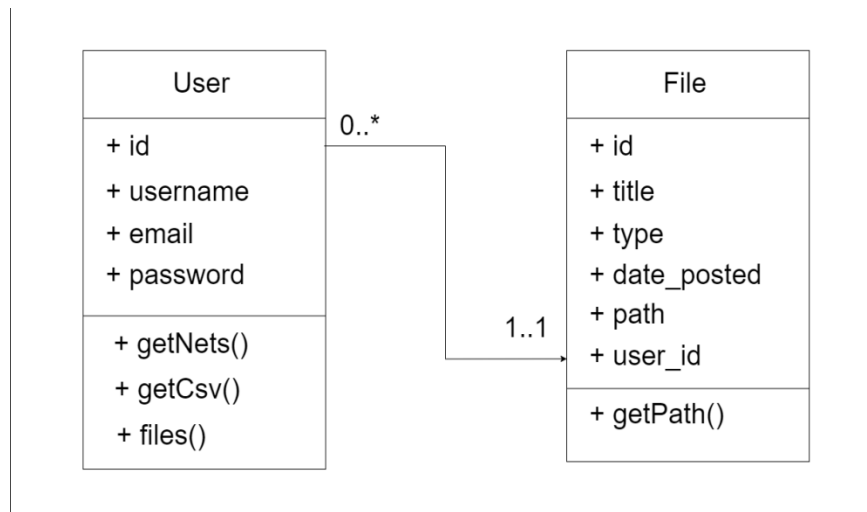


Figure 23 : Diagramme de classes.

4.4 Diagrammes de séquences dans notre cas d'étude

Les diagrammes de séquences permettent de décrire l'aspect dynamique du système qui représente comment les éléments du système interagissent entre eux et avec les acteurs. Les objets au cœur d'un système interagissent en s'échangeant des messages. Les acteurs interagissent avec le système au moyen des IHM (Interfaces Homme-Machine) [69]. Nous présentons ci-dessous les diagrammes de séquences de notre cas d'étude liés au cas d'utilisation « Upload de fichier (.net) et afficher les résultats » et « Upload de fichier csv et télécharger le fichier (.net) correspondant ».

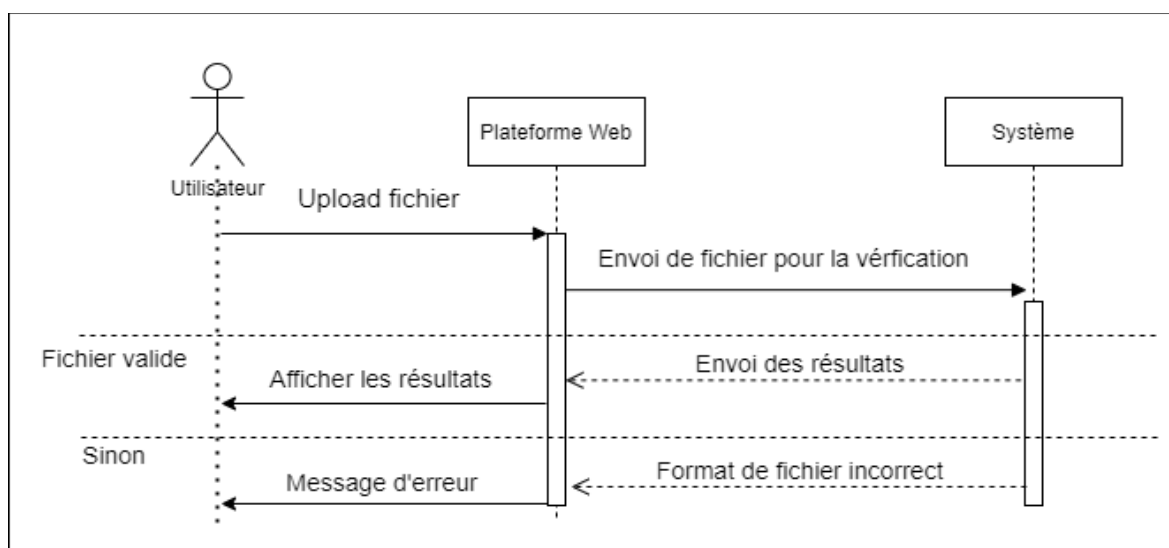


Figure 24 : Diagramme de séquences pour Upload un fichier (.net).

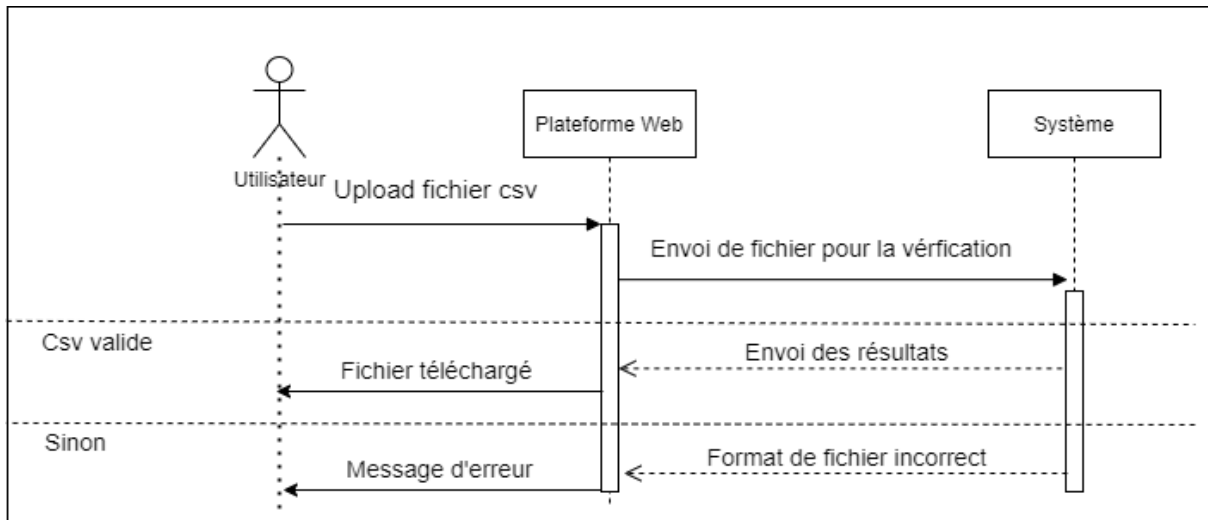


Figure 25 : Diagramme de séquences pour Upload un fichier csv.

5 Outils, langages et bibliothèques utilisés

- **Python :**

Python est un langage de programmation de haut niveau. Il fournit des outils qui permettent de construire des programmes et applications à petite et à grande échelle. Python dispose d'un système de typage dynamique et d'une gestion automatique de la mémoire. Il permet une programmation multi-paradigmes comme par exemple l'orienté objet et la programmation fonctionnelle, et dispose d'une bibliothèque standard vaste et complète ainsi que de nombreuses autres librairies conçues par la communauté à travers le gestionnaire de dépendance PIP. Les interpréteurs de Python sont disponibles pour de nombreux systèmes d'exploitation.

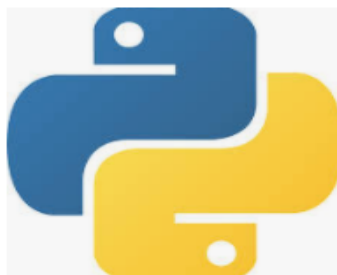


Figure 26 : Logo Python

- **Flask :**

Flask est un micro framework open-source de développement web en Python. Il est classé comme microframework car il est très léger. Flask a pour objectif de garder un noyau simple mais extensible. Il n'intègre pas de système d'authentification, pas de couche d'abstraction de base de données, ni d'outil de validation de formulaires. Cependant, de nombreuses extensions permettent d'ajouter facilement des fonctionnalités.



Figure 27 : Logo Flask

- **Heroku :**

Heroku est une plateforme en tant que service, qui permet d'héberger des applications web sans avoir à configurer les serveurs et les ressources de ce dernier, et permet aussi le déploiement continu (qui est une stratégie de développement logiciel où toute validation de code qui réussit le cycle de test automatisé est automatiquement transférée dans l'environnement de production, propulsant ainsi les modifications vers les utilisateurs du logiciel.). La plateforme Heroku prend en charge le développement dans Ruby on Rails, Php, Java, Node.js, Python, Go, Scala et Clojure.



Figure 28 : Logo HEROKU

- **Visual studio code :**

Visual Studio Code est un éditeur de code open-source et extensible développé par Microsoft pour Windows, Linux et macOS. Il supporte un très grand nombre de langages grâce à des extensions. Il supporte aussi l'autocomplétion, la coloration syntaxique, le débogage, et les commandes Git. Il est aussi bien performant et agréable à utiliser.



Figure 29 : Logo Visual studio code

- **Git :**

Git est un logiciel de gestion de versions décentralisé. C'est un logiciel libre créé par Linus Torvalds, auteur du noyau Linux. Git est un système de contrôle de version distribué et open source :

- ❖ **Système de contrôle:** cela signifie essentiellement que Git est un outil de suivi de contenu. Git peut donc être utilisé pour stocker du contenu. Il est principalement utilisé pour stocker du code en raison des autres fonctionnalités qu'il fournit.
- ❖ **Système de contrôle de version:** Le code qui est stocké dans Git continue de changer au fur et à mesure que l'on en ajoute. De nombreux développeurs peuvent également ajouter du code en parallèle. Ainsi, le système de contrôle de version aide à gérer cela en conservant un historique des changements qui se sont produits.
- ❖ **Système de contrôle de version distribué:** Git possède un référentiel distant qui est stocké sur un serveur et un référentiel local qui est stocké dans l'ordinateur de chaque développeur. Cela signifie que le code n'est pas seulement stocké sur

un serveur central, mais la copie complète du code est présente sur tous les ordinateurs des développeurs.



Figure 30 : Logo Git

- **MySQL :**

MySQL est un système de gestion de bases de données relationnelles (SGBDR) à code source libre qui utilise le langage SQL (Structured Query Language). SQL est le langage le plus populaire pour ajouter, accéder et gérer le contenu d'une base de données. Il est surtout connu pour sa rapidité de traitement, sa fiabilité éprouvée, sa facilité et sa flexibilité d'utilisation.



Figure 31 : Logo MySQL

- **SQLite :**

SQLite est une bibliothèque qui met en œuvre un moteur de base de données SQL transactionnel, autonome, sans serveur et sans configuration. Le code de SQLite est dans le domaine public et peut donc être utilisé librement à toutes fins, commerciales ou privées. SQLite est la base de données la plus largement déployée dans le monde, avec plus d'applications que nous ne pouvons en compter, y compris plusieurs projets de grande envergure.

- ***Vis.js :***

Vis.js est une bibliothèque JavaScript de visualisation dynamique, basée sur le navigateur. La bibliothèque est conçue pour être facile à utiliser, traiter de grandes quantités de données dynamiques et permettre la manipulation des données.

- ***AutoComplete.js :***

AutoComplete.js est une bibliothèque Javascript, conçue progressivement pour la vitesse, la grande polyvalence et l'intégration transparente avec un large éventail de projets et de systèmes, à l'intention des utilisateurs et des développeurs.

- ***NetworkX :***

NetworkX est une librairie Python pour la création, la manipulation et l'étude de la structure, de la dynamique et des fonctions de réseaux complexes.

- ***LinkPred :***

LinkPred est une librairie Python pour la prédiction de liens basé sur NetworkX, étant donné un réseau, LinkPred fournit un certain nombre d'heuristiques (connues sous le nom de prédicteurs) qui évaluent la possibilité de liens potentiels dans un futur instantané du réseau.

- ***Flask-Login :***

Flask-Login permet la gestion des sessions utilisateur pour Flask. Il gère les tâches courantes de connexion, de déconnexion et de mémorisation des sessions de vos utilisateurs sur de longues périodes.

Il permet de :

- ❖ Stocker l'identifiant de l'utilisateur actif dans la session, et de vous permettre de vous connecter et de vous déconnecter facilement.

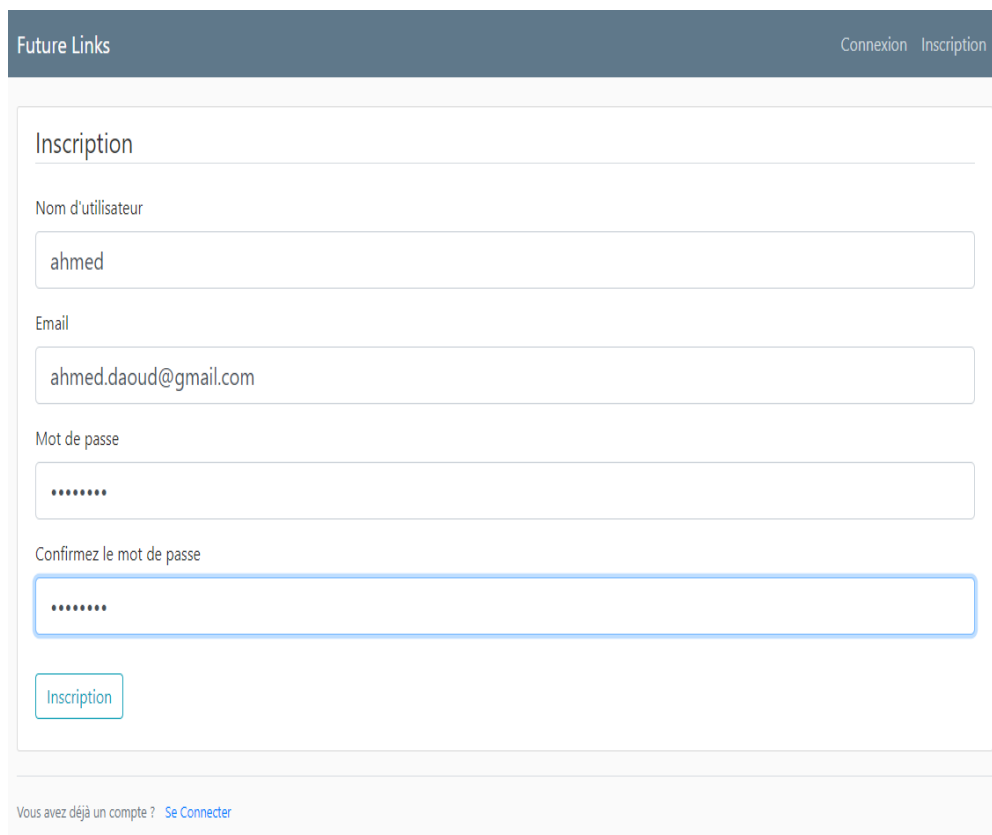
- ❖ Limiter les vues aux utilisateurs connectés (ou déconnectés).
- ❖ Gérer la fonction normalement difficile "se souvenir de moi".
- ❖ Protéger les sessions de vos utilisateurs contre les vols de cookies.
- ❖ Fournir la possibilité d'intégration avec Flask-Principal ou d'autres extensions d'autorisation par la suite.

- ***Flask-SQLAlchemy :***

Flask-SQLAlchemy est une extension pour Flask qui ajoute un support pour SQLAlchemy à votre application. Il vise à simplifier l'utilisation de SQLAlchemy avec Flask en fournissant des valeurs par défaut utiles et des aides supplémentaires qui facilitent l'exécution de tâches communes.

6 Présentation de notre application Future Links

1) Interface d'inscription



The screenshot shows a web interface for 'Future Links'. At the top, there is a dark blue header with the text 'Future Links' on the left and 'Connexion' and 'Inscription' on the right. Below the header is a light gray box containing the registration form. The form is titled 'Inscription' and has four input fields: 'Nom d'utilisateur' (containing 'ahmed'), 'Email' (containing 'ahmed.daoud@gmail.com'), 'Mot de passe' (containing seven dots), and 'Confirmez le mot de passe' (containing seven dots). Below these fields is a blue button labeled 'Inscription'. At the bottom of the form, there is a link that says 'Vous avez déjà un compte ? [Se Connecter](#)'.

Figure 32 : Interface d'inscription.

2) Interface d'authentification

Future Links Connexion Inscription

Connexion

Email

ahmed.daoud@gmail.com

Mot de passe

.....

☐ Se souvenir de moi

Connexion

[Mot de passe oublié ?](#)

Besoin d'un compte ? [Inscrivez Vous Maintenant !](#)

Figure 33 : Interface d'authentification.

3) Interface d'ajout de fichier (graphe et csv)

Future Links Ajouter un Fichier Historique Déconnexion

Ajouter un graphe

Choisir un fichier Aucun fichier choisi Confirmer

Transformer un csv en graphe

Choisir un fichier Aucun fichier choisi Confirmer

Figure 34 : Interface d'ajout de fichier.

4) Interface de prédiction de liens et affichage de graphe

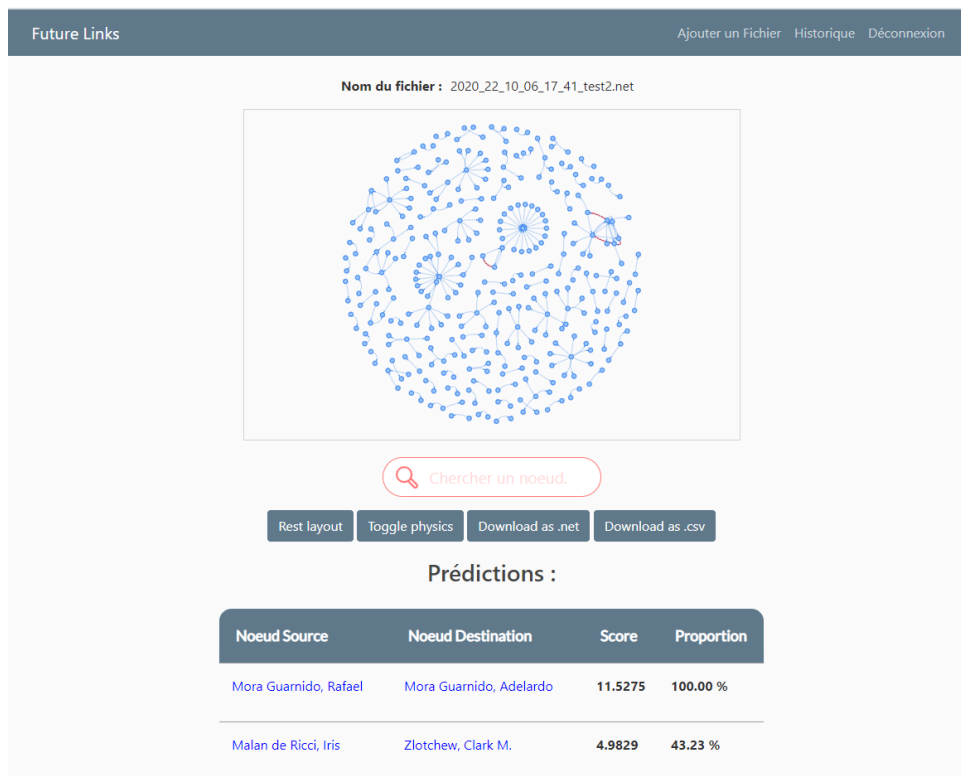


Figure 35 : Interface de prédiction de liens et affichage de graphe

5) Fonctionnalité de recherche

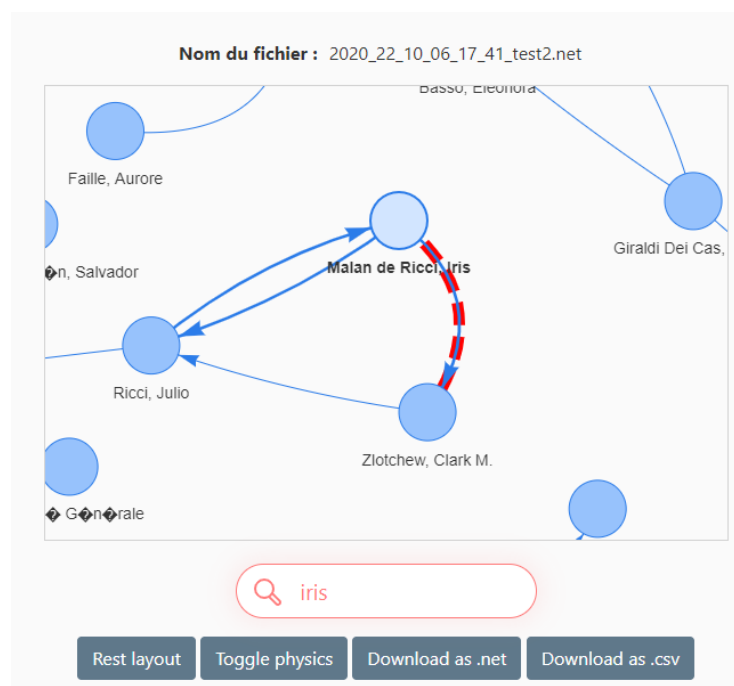


Figure 36 : Fonctionnalité de recherche

6)Interface Historique

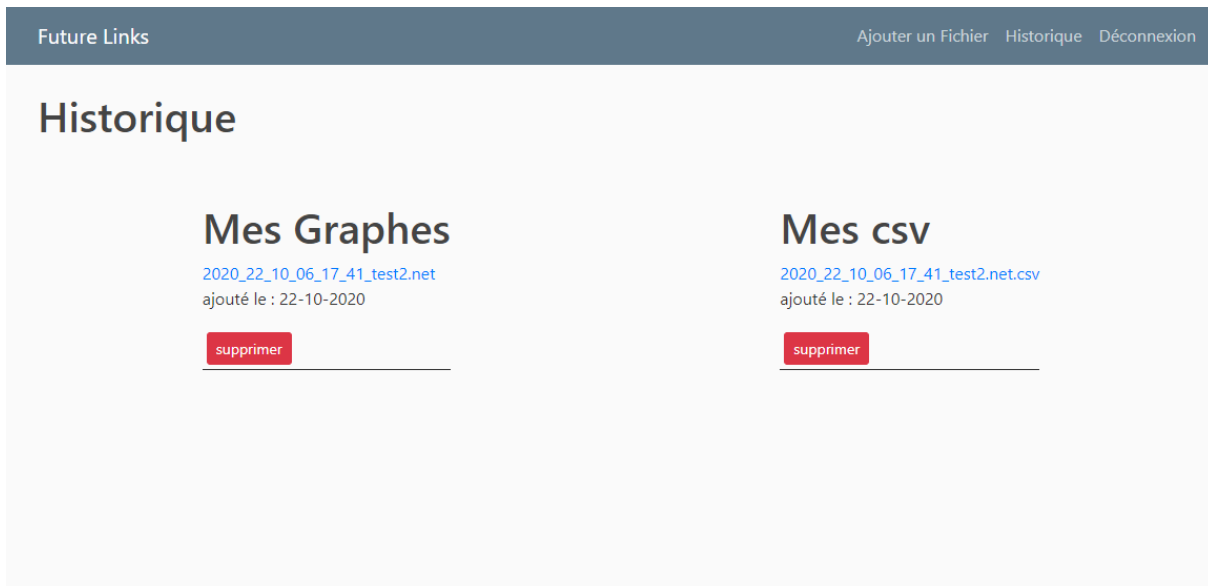


Figure 37 : Interface historique

Conclusion

Dans ce dernier chapitre, nous avons présenté la conception, l'implémentation et la réalisation de notre application. Nous avons expliqué en détail l'algorithme implémenté dans notre plateforme web (dénommée Future Links). Il s'agit de l'algorithme CNGF qui est basé sur la notion de la « Node Guidance Capability ».

Ensuite nous avons comparé les différents algorithmes de prédiction de liens en mettant en évidence les améliorations apportées par CNGF par rapport à Common Neighbors, puis nous avons évoqué l'aspect sécurité dans notre plateforme web.

Enfin nous avons listé les outils, langages et bibliothèques utilisés avant de présenter les principales fonctionnalités de notre application Future Links.

Conclusion générale

Une première version d'un outil d'exploitation de données d'archives nommé Link&Pred a été développée dans le cadre d'un stage de Master 2 au laboratoire du LIAS à Poitiers en France. Dans la continuité de ce travail, l'objectif principal de notre projet de Master 2 est l'amélioration de cet outil en termes de performances et précisions ainsi que de fournir de nouvelles fonctionnalités et le munir d'une interface utilisateur fonctionnelle.

Nous avons commencé notre travail par un état de l'art sur l'extraction de connaissances (ECD). Nous nous sommes documentés sur les différentes étapes de l'ECD et plus particulièrement sur la phase fouille de données (Data Mining).

Ensuite nous nous sommes penchés sur l'étude de l'existant et l'analyse de ce projet, ainsi que les différentes approches et méthodes de prédiction de liens que l'on peut trouver dans la littérature.

Nous avons alors expliqué la notion de « Node Guidance Capability » sur laquelle se base l'algorithme CNGF que nous utilisons dans notre étude, et l'avons comparé aux autres algorithmes sur différents jeux de données. La méthode que nous proposons est testée particulièrement sur un ensemble de fichiers fournis par l'équipe Archivos-CRLA (Poitiers, France). Les expérimentations réalisées ont donné des résultats satisfaisants tout particulièrement en termes de précision.

Plusieurs technologies nous ont été nécessaires pour l'implémentation de l'algorithme CNGF dans notre application.

La solution que nous proposons s'insère dans le cadre du développement d'une application web sécurisée dénommée Future Links qui a pour but la prédiction de nouveaux liens dans un réseau social. L'application propose diverses fonctionnalités comme l'authentification ainsi que l'upload d'un fichier graphe (.net) ou bien d'un fichier CSV, une visualisation du graphe et des résultats obtenus, une section de

recherche et d'historique, et aussi le téléchargement des résultats de prédiction de liens obtenus.

Notre application ne s'arrête pas juste au traitement des données d'archives mais pour prédire et orienter la recherche au future. Les projets de recherche actuels tendent à être plus coopératifs en raison de la multidisciplinarité des domaines abordés (Le programme ERASMUS par exemple nécessite la collaboration de plusieurs équipes de recherche de pays différents, voire de divers continents).

Bien que de nombreux algorithmes de prédiction de liens aient été développés récemment dans divers domaines, la prédiction de liens reste un domaine qui est encore dans une phase d'exploration et pour lequel il faudra attendre quelques années avant d'arriver à un stade de maturation. Ceci dit, il est vrai que la solution que nous proposons présente certaines limites dont, la nécessité d'une liste d'adjacences pour la transformer en graphe (.net), cette limite nécessite le développement d'une plateforme plus complète qui entre autre disposerait d'un ensemble d'outils permettant aux utilisateurs de visualiser et manipuler leurs fichiers csv de tel sorte à générer des graphes à partir de ces derniers et éliminer le besoin de passer par d'autres outils tiers ou d'avoir recours à l'intervention d'un informaticien.

Ce projet qui jumelle les thèmes de recherche et de la méthodologie pratique nous a permis d'acquérir de nouvelles connaissances couvrant le cycle de vie de conception d'un entrepôt de données (la phase logique et ETL), les techniques de fouilles de données, le Machine Learning, la théorie des graphes et la théorie des réseaux. Nous avons également pu apprendre de nouveaux langages de programmation, outils, techniques et concepts de développements d'applications web, qui nous ont aidés à développer notre solution.

Finalement, ce projet nous a permis de développer nos compétences de communication et d'écoute.

Bibliographie :

- [1] Abdelghani Laifa, Vers un outil intelligent d'exploitation et d'analyse de données d'archives de faible qualité, Mémoire de Stage M2, LIAS/ENSMA & CRLA-ITEM/UP, Poitiers, 2019.
- [2] Frédéric Pennerath, Méthodes d'extraction de connaissances à partir de données, Application à des problèmes de synthèse organique Informatique, Université Henri Poincaré – Nancy, France, 2009.
- [3] Dominique Crié, De l'extraction des connaissances au Knowledge Management Dans Revue française de gestion (no 146), pages 59 à 79, 2003.
- [4] Frédéric Pennerath, Méthodes d'Extraction De Connaissances à Partir De Données Modélisables Par Des Graphes, Application à Des Problèmes De Synthèse Organique, Thèse de doctorat, INRIA Lorraine, LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications, 2009.
- [5] HELA LTIFI, Mounir Ben Ayed, Christophe Kolski, Démarche Centrée Utilisateur Pour La Conception De SIAD Basés Sur Un Processus d'ECD, Application Dans Le Domaine De La Santé, Journal d'Interaction Personne-Système, Vol. 1, Num. 1, Art. 1, Septembre 2010.
- [6] BRAHIMI Belgacem, Extraction De Connaissances à Partir De Données Incomplètes et Imprécises, Magistère, Blida, 2011.
- [7] JEROME AZE, Extraction De Connaissance à Partir de Données Numériques et Textuelles, Thèse de doctorat, Université Paris Sud - Paris XI, 2003.
- [8] Catarina Dudas, Amos Ng, Henrik Boström, Knowledge Extraction in Manufacturing using Data Mining Techniques, Article Universitaire Centre for Intelligent Automation, University of Skövde, Sweden, 2009.
- [9] Tipawan Silwattananusarn, Kulthida Tuamsuk, Data Mining and Its Applications for Knowledge Management, International Journal of Data Mining & Knowledge Management Process (IJDMP) Vol.2, No.5, pp. 13-24, 2012.
- [10] Thomas Burger, Extraction de connaissances à partir de données de protéomique de découverte haut-débit, Bio-informatique [q-bio.QM]. UGA - Université Grenoble Alpes, 2017.
- [11] Labio WJ, JL Wiener, H Garcia-Molina, 2000, Efficient resumption of interrupted warehouse loads. In Proc. ACM SIGMOD Int. Conf. on Management of Data, pp. 46–57, 2002.

- [12] Ítalo Cunha, distributed on-demand ETL intelligence Journal of Internet Services and Applications volume 10, Article number: 21, 2019.
- [13] Margaret Rouse, «<https://www.lemagit.fr/definition/ETL-et-ELT>», New York, États-Unis, 2015.
- [14] Simitsis A, P Vassiliadis, T Sellis, 2005, Optimizing ETL processes in data warehouses, In Proc. 21st Int. Conf. on Data Engineering, pp. 564–575, 2005.
- [15] Trujillo J et al, UML based approach for modeling ETL processes in data warehouses In Proc, 22nd Int Conf. on Conceptual Modeling, pp. 307–320, 2003.
- [16] C. Desrosiers, Intégration des données et ETL, «https://cours.etsmtl.ca/mti820/public_docs/acetates/MTI820-Acetates-ETL_1pp.pdf», Cours du département de génie logiciel et des TI MTI820, 2011.
- [17] Panos Vassiliadis, A Survey of Extract-Transform-Load Technology, International Journal of Data Warehousing and Mining 5:1-27, 2009.
- [18] Le site talend «<https://www.talend.com/fr/resources/elt-vs-etl>».
- [19] Pierre Liseron, «<https://blog.axopen.com/2016/11/talend-etl-definition>», Lyon, Villeurbanne – France, 2016.
- [20] Mathieu, «<https://www.saagie.com/fr/blog/hadoop-et-big-data>», Angers, France, 2017.
- [21] George Thomas, «<https://www.futura-sciences.com/tech/definitions/informatique-cloud-computing-11573>», Flickr, CC by-sa 2.0, 2019.
- [22] Analyse web, Procedia Computer Science Volume 124, 2017, Pages 93-99, Leo WillyantoSantoso, Surabaya, Indonesia, 2017.
- [23] Site Lias, «<https://www.lias-lab.fr>».
- [24] L'équipe Archivos, «<http://crla-archivos.labo.univ-poitiers.fr/le-centre-de-recherches-latino-americaines--archivos-crla--archivos/presentation>».
- [25] Gordon Scott, «<https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp>», Review from investopedia, 2018.
- [26] L. Breiman, Random forests, Machine Learning, 45(1):p5, 2001.
- [27] Peter Harrington, 2012, Machine Learning in Action, Manning Publications Co.3 Lewis Street Greenwich, CT United States ISBN:978-1-61729-018-3, 2012.
- [28] Steven Lemm, NeuroImage, Volume 56, Issue 2, 15, Pages 387-399, 2011.

- [29] M. Al Hasan, V Chaoji, S Salem, M Zaki, 2006, Link prediction using supervised learning. In Workshop on Link Discovery: Issues, Approaches and Apps, 2005.
- [30] Antonio Criminisi et al, Foundations and Trends in Computer Graphics and Vision February, 2012.
- [31] Constantinos S.Hilas, Knowledge-Based Systems Volume 21, Issue 7, Pages 721-726, Grèce, 2008.
- [32] Jason J. Yu, AW Harley, KG Derpanis, Back to Basics: Unsupervised Learning, Conférence en ligne, 2016.
- [33] David L, 2017, ActiIA, Le Magazine de l'intelligence artificielle, France, 2018.
- [34] Olivier Cogis et Claudine Schwartz, Théorie des graphes, Cassini, 2018.
- [35] L. Barabási, H Jeong, Z Néda, E Ravasz, Evolution of the social network of scientific collaboration. Physica A, 311(3-4):590 614, 2002.
- [36] Tang, F. et al, Notice of Retraction The implementation of information service based on social network systems, Second International Workshop on Education Technology and Computer Science, 2010.
- [37] Emmanuel Pannier, L'analyse des réseaux sociaux : théories, concepts et méthodologies, Sociological Review of Vietnam, 4 (104), Hanoi : 100-114, 2008.
- [38] Lindsay A. Thompson et al, The Intersection of Online Social Networking with Medical Professionalism, Journal of General Internal Medicine volume 23, pages 954–957, 2008.
- [39] Christine A. Halverson, Social Networks and Social Networking, IEEE Internet Computing Volume : 9, Issue: 5, Page(s): 14 – 15, 2005.
- [40] Peng Wang, BW Xu, YR Wu, XY Zhou, «Link prediction in social networks: the state-of-the-art» Article universitaire, Paris, France & Shanghai, China, 2014.
- [41] Mohammad Al Hasan et al, A Survey of Link Prediction in Social Networks, Social Network Data Analytics pp 243-275, 2011.
- [42] Shiping Huang, MJ Zaki, Link Prediction Based on Timevaried Weight, Conference on Computer Supported Cooperative Work in Design (CSCWD), 2014.
- [43] Niladri Sett, Neurocomputing, Volume 172, 8, Pages 71-83, 2016.
- [44] J Ben Schafe, D Frankowski, S Sen, Collaborative filtering recommender systems, In The adaptive web, Department of Computer Science University of Northern Iowa, 2007.

- [45] Zhou T, L Lü, Link prediction in complex networks: A survey Physica A: statistical mechanics and its applications, Physica A: Statistical Mechanics and its Applications, Volume 390, Issue 6, 15, Pages 1150-1170, 2011.
- [46] Pavlov M, R Ichise, Finding experts by link prediction in co-authorship networks University of Waterloo, Waterloo ON N2L 3G1, Canada, 2007.
- [47] Zan Huang et al, Link prediction approach to collaborative filtering, Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL05), 2005.
- [48] David Liben-Nowell, The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology, 2007.
- [49] NV Chawla. Lichtenwalter et al, KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 243–252, 2010.
- [50] Adamic and E. Adar, Friends and neighbors on the web, Social Networks, 25:211--230, 2001.
- [51] Ryan N. Lichtenwalter, RN Lichtenwalter, JT Lussier, «New perspectives and methods in link prediction», 2010.
- [52] Zne-JungLee et al, Information Sciences volume 173, Issues 1–3, 16, Pages 155-167, 2005.
- [53] European Journal of Operational Research volume 195, Issue 3, Pages 803-809, 2009.
- [54] LingxiaDu et al, Information Sciences volume 295, 20, Pages 521-535, February 2015.
- [55] Wang, P., Xu, B., Wu, Y., & Zhou, X. Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 58(1), 1-38, 2015.
- [56] Shafiur Rahmanet, LR Dey, S Haider, Link Prediction by Correlation on Social Network, 20th International Conference of Computer and Information Technology (ICCIT), 2017.
- [57] Sahil Gupta, Shalini Pandey, K.K.Shukla, Comparison Analysis of Link Prediction Algorithms in Social Network, International Journal of Computer Applications (0975 – 8887) Volume 111 – No 16, February 2015.

- [58] Liyan Dong, Y Li, H Yin, H Le, M Rui, The Algorithm of Link Prediction on Social Network, Article de recherche College of Computer Science and Technology, Jilin University, Changchun, China, 2013.
- [59] Yves Barlette, Une étude des comportements liés à la sécurité des systèmes d'information, Dans Systèmes d'information & management (Volume 13), pages 7 à 30, 2008.
- [60] Brian Harnish, «<https://www.semrush.com/blog/what-is-https/>», 2018.
- [61] Jean L, Authentification De L'accès Web Pour Les Apps, 2019.
- [62] Jean-René Reinhard, Etude de primitives cryptographiques symétriques : chiffrements par flot et fonction de hachage, Thèse de doctorat en Informatique, Versailles-St Quentin en Yvelines, 2011.
- [63] Daniel Boterhoven, «<https://medium.com/@danboterhoven/why-you-should-use-bcrypt-to-hash-passwords-af330100b861>», 2016
- [64] Katja Malvoni, Are Your Passwords Safe: Energy-Efficient Bcrypt Cracking with Low-Cost Parallel Hardware, University of Zagreb, 2014
- [65] Laurent P, «<https://docs.djangoproject.com/fr/1.8/ref/csrf>», 2014.
- [66] Stéphane Jorge, Qu'est-ce que le langage UML (langage de modélisation unifié)? : « <https://www.lucidchart.com/pages/fr/langage-uml> », 2018.
- [67] Pierre Gérard IUT de Villetaneuse DUT informatique, S2 : «<https://lipn.univ-paris13.fr/~gerard/uml-s2/uml-cours04.html>», 2013.
- [68] Rémy Manu “UML : Langage de modélisation objet unifié Cours n°3 : Diagramme des classes : « <http://remy-manu.no-ip.biz/UML/Cours/coursUML3.pdf> », 2014.
- [69] Pierre Gérard IUT de Villetaneuse DUT informatique, S2 : “UML Cours 5 : Diagramme de séquences”: « <https://lipn.univ-paris13.fr/~gerard/uml-s2/uml-cours05.html> », 2013.