REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
NISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID - TLEMCEN

THÈSE

Présentée à :

FACULTE DES SCIENCES – DEPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

DOCTORAT EN SCIENCES

Spécialité : Informatique

Par :

Mme Hamza-Cherif épouse Rahmoun Souaad

Sur le thème

Conceptualisation des bonnes pratiques au sein d'une communauté de pratique

---

Soutenue publiquement le : 28 Juin 2022 à Tlemcen devant le jury composé de :

Mr CHIKH Med Amine	Professeur	Université de Tlemcen	Président
Mr CHIKH Azeddine	Professeur	Université de Tlemcen	Directeur de thèse
Mr ABDERRAHIM Med El Amine	Professeur	Université de Tlemcen	Examinateur 1
Mr BELALEM Ghalem	Professeur	Université d'Oran1	Examinateur 2
Mr SOUIER Mehdi	Professeur	Ecole Supérieure de Management Tlemcen	Examinateur 3
Mr BENOMAR Med Lamine	MCA	Université de Temouchent	Examinateur 4

Laboratoire de Recherche en Informatique de Tlemcen (LRIT)
BP 119, 13000 Tlemcen - Algérie

Je dédie ce travail à toutes les personnes chères à mon cœur qui m'ont aidé à arriver là où j'en suis : ma chère maman que dieu me la garde, mon père lah yarahmou, mes enfants qui sont ma raison d'être, mon mari, ma sœur, mes frères, ma sœur de cœur et ceux que je ne cite pas vous êtes nombreux. Aussi je dédie ce travail à toutes les personnes surtout les femmes, les mamans qui malgré toutes les difficultés de la vie réussissent à entreprendre un projet et le finaliser aussi modeste soit-il

Remerciements

Louange à Allah qui m'a donné la force, la patience et le courage pour finaliser ce modeste travail de thèse malgré toutes les difficultés rencontrées.

Je voudrais tout d'abord exprimer mes plus profonds remerciements à Mr CHIKH Azeddine, mon directeur de thèse pour son aide, ses encouragements, ses conseils, sa disponibilité et sa sympathie qui m'ont permis de mener à bien cette thèse.

J'aimerais également remercier les membres du laboratoire d'Image Signaux et Systèmes Intelligents LISSI au niveau de l'Université Paris Est Créteil "UPEC", notamment Monsieur CHIBANI Abdelghani qui m'a co-encadré durant une certaine pèriode, Monsieur AMIRAT Yacine le directeur du laboratoire pour son accueil, et surtout Madame AYARI Nouel pour son aide précieuse, ses conseils, et sa sympathie.

Je remercie très vivement mes collègues et amis du département de Génie Biomédicale pour leurs aides, leurs soutiens et leurs encouragements qui m'ont été précieux durant cette thèse.

J'adresse mes sincères remerciements à Monsieur CHIKH Amine, Professeur à l'université de Tlemcen, qui m'a fait l'honneur de présider le jury de cette thèse.

J'exprime ma profonde reconnaissance à Monsieur ABDERRAHIM Mohammed El Amine Professeur à l'université de Tlemcen, Monsieur GHALEM Belalem Professeur à l'université USTO1 d'Oran, Monsieur SOUIER Mehdi Professeur de l'école supérieure de management à Tlemcen, et Monsieur BENOMAR Mohammed Lamine Maitre de conférence à l'université de Temouchent, pour l'intérêt qu'ils ont bien voulu porter à ce travail en acceptant de l'examiner et d'en être rapporteurs.

Mes remerciements vont à toutes les personnes (amis, familles, voisins) que j'ai rencontrées au cours de mon séjour de 18 mois à Paris, pour leur convivialité, leur aide et leur soutien surtout durant la période de COVID qui a été très difficile .

Je conclurais en remerciant de tout cœur ma famille pour son soutien, son écoute et ses encouragements tout au long de cette thèse. Mes parents, mon époux, mes sœurs et belles soeurs, mes frères et beau frère.

Lieu, le 28 juin 2022.

ملخص

منذ ظهور الويب الاجتماعي والدلالي ، في السنوات الأخيرة ، ظهرت أدوات جديدة ومشاركة مواقع مثل ميتا و طويتير و ويكي هاو ، إلخ ، مما جعل الويب مجموعة عالمية من المعرفة ، حيث يشكل المستخدمون جغرافيًا مجتمعات الممارسة عبر الإنترنت ، التي تعد في الأصل مفهومًا لعلم الاجتماع ولكنها تجد تطورها الكامل في الويب الحالي حيث يشارك المستخدمون ويتبادلون معرفتهم في مجالات مختلفة في شكل معرفة إجرائية تسمى الممارسات الجيدة.

يتم تحديد هذه الممارسات الجيدة من خلال مجموعة من الخطوات المتتالية المتخذة لتحقيق الهدف. أصبح تصور هذه المعرفةيتم تحديد هذه الممارسات الجيدة من خلال مجموعة من الخطوات المتتالية المتخذة لتحقيق الهدف. أصبح تصور هذه المعرفة الإجرائية تحديًا كبيرًا في العديد من المجالات (استرجاع المعلومات ، والتطبيقات الذكية ، والروبوتات ...) ، واستخراج المعرفة من قاعدة البيانات هو المجال الذي يتطور لتقديم الحلول، حيث يجمع بين طرق مختلفة للتعلم وتمثيل المعرفة من أجل إيجاد حلول لاستكشاف البيانات غير المهيكلة من أجل تسهيل استغلالها وفي هذا السياق ركز العديد من الأعمال على استكشاف المعرفة الإجرائية في أغراض مختلفة ، أحيانًا لإنشاء قاعدة معرفية أو تحديد التعليمات من المعرفة الإجرائية. معظم هذا العمل في مجال معالجة اللغة الطبيعية ، والهدف الذي نسعى إليه هو هدف آخر ، في هذه الأطروحة نقدم نهجًا جديدًا لاستخراج وتصور الممارسات الجيدة من الويب ، واستخراج أفضل الممارسات لاستعلام معين. يتم تنفيذ النهج المقترح على مرحلتين : في المرحلة الأولى ، يتم استخراج الممارسات الجيدة من الويب باستخدام طريقة تخريد

يتم تنفيذ النهج المقترح على مرحلتين : في المرحلة الأولى ، يتم استخراج الممارسات الجيدة من الويب باستخدام طريقة تخريد الويب ، بعد أن نقوم بتمثيلها من خلال الرسوم البيانية للبيانات الموجهة. في المرحلة الثانية ، نستخرج أفضل الممارسات لاستعلام معين من خلال تطبيق تقنيات التعلم الآلي وتلخيص النص على الرسوم البيانية. تحدث هذه المرحلة في ثلاث خطوات : (١) البحث عن ممارسات مشابهة لاستعلام المستخدم ، وهنا نستخدم كلمة نموذج التضمين لتحديد الجمل المشابهة للهدف الذي يسعى إليه المستخدم ؛ (٢) تجميع ودمج الخطوات المتشابهة ، حيث نستخدم تقنيات التعلم غير الخاضع للإشراف وتلخيص النص لتجميع العقد القريبة المعنوية التي ندمجها في نفس الخطوة ؛ (٣) استخراج أفضل الممارسات التي يتم تحديدها من خلال مسار الرسم البياني الذي يمر عبر أهم الخطوات للوصول إلى الهدف ، ويتم حساب هذه الأهمية من خلال مقاييس مركزية الرسوم البيانية التي تحدد أهمية العقد في الرسم البياني الموجه بواسطة عدد الْقواس الواردة والصادرة. أظهرت النتائج التي تم الحصول عليها تفوق نهجنا في : (١) التقاط ممارسات مماثلة للهدف الذي يسعى إليه المستخدم ، وذلك من خلال تحسين وقت التنفيذ ، (٢) استخراج أفضل الممارسات للاستعلامات مقارنة بمحرك بحث من واقع حقيقي. مجموعة
:الكلمات المفتاحية مجتمع الممارسة ، الممارسة الجيدة ، استخراج المعرفة الإجرائية ، الرسم البياني للمعرفة ، تركيب النص ، التعلم الآلي.

Résumé

Depuis l'avènement du web social et sémantique, il ne cesse d'émerger ces dernières années de nouveaux outils et sites de partage tels que Meta, Twitter, WikiHow,... faisant du web un recueil universel de connaissances, où les utilisateurs répartis géographiquement forment des communautés de pratique (CdP) en ligne, ces CdP sont à l'origine un concept de sociologie mais trouvent tout leur essor dans le web actuel où des individus partagent et échangent leur savoir faire dans différents domaines sous forme de connaissances procédurales (CP) appelées bonnes pratiques.

Ces bonnes pratiques sont définies par un ensemble d'étapes successives acheminées pour atteindre un objectif. Conceptualiser ces connaissances procédurales est devenu un enjeu majeur dans plusieurs domaines (recherche d'information, applications intelligentes, Robotique...), l'extraction des connaissances à partir de données (ECD) est le domaine qui évolue pour offrir des solutions. L'ECD combine différentes méthode d'apprentissage et de représentation des connaissances afin de trouver des solutions pour explorer les données non structurées dans le but de faciliter leur exploitation et dans ce contexte plusieurs travaux se sont penchés sur l'exploration des connaissance procédurales dans des buts différents, parfois pour créer une base de connaissance ou encore pour identifier les instructions à partir des connaissances procédurales. La plupart de ces travaux relèvent du domaine de traitement du langage naturel, le but que nous poursuivons est autre, dans cette thèse nous présentons une nouvelle approche pour extraire et conceptualiser les bonnes pratiques du web, et extraire la meilleure pratique pour une requête donnée.

L'approche proposée se déroule en deux phases : durant la première on extrait les bonnes pratiques du web par une méthode de web scrapping, qu'on représente par des graphes orientés de données. Dans la seconde phase on procède à l'extraction de la meilleure pratique pour une requête donnée ceci en appliquant les techniques d'apprentissage artificiel et de résumé de texte sur les graphes. Cette phase de déroule en trois étapes : (1) recherche des pratiques similaires à la requête de l'utilisateur, on utilise ici le modèle de prolongement lexicale de mots pour identifier les phrases similaires au but recherché par l'utilisateur ; (2) regroupement et fusion des étapes similaires, où nous faisons appel aux techniques d'apprentissage non supervisé (DBScan) et de résumé de texte (PageRank) afin de regrouper les noeuds sémantiquement proches que nous fusionnons dans une même étape ; (3) extraction de la meilleure pratique qu'on identifie par le chemin du graphe parcourant les étapes les plus importantes pour atteindre l'objectif, cette importance est calculée grâce aux mesures de centralitè des graphes qui quantifient l'importance des noeuds dans un graphe orienté par ses le nombre de leur arc entrant et sortant.

Les résultats obtenus ont démontré la supériorité de notre approche pour : (1) capturer les pratiques similaires au but recherché par l'utilisateur, et ceci en optimisant le temps d'exécution, (2) extraire les meilleure pratiques pour des requêtes par rapport à un moteur de recherche à partir d'un jeu de données réel.

Mots clés : Communauté de pratique, Bonne pratique, Extraction des connaissance procédurale, Graphe de connaissance, Synthèse de texte, Apprentissage artificiel.

Abstract

Since the advent of the social and semantic web, in recent years new tools and sharing sites such as Meta, Twitter, WikiHow, etc. have emerged, making the web a universal collection of knowledge, where users geographically form communities of practice (CoP) online, these CoPs are originally a concept of sociology but find their full development in the current web where users share and exchange their know-how in different fields in the form of procedural knowledge (PK) called good practices.

These good practices are defined by a set of successive steps taken to achieve an objective. Conceptualizing this procedural knowledge has become a major challenge in several fields (information retrieval, intelligent applications, robotics...), knowledge extraction from data base (KDD) is the field that is evolving to offer solutions. KDD combines different methods of learning and knowledge representation in order to find solutions to explore unstructured data in order to facilitate their exploitation and in this context several works have focused on the exploration of procedural knowledge in different purposes, sometimes to create a knowledge base or to identify instructions from procedural knowledge. Most of this work is in the field of natural language processing, the goal we pursue is another, in this thesis we present a new approach to extract and conceptualize good practices from the web, and extract the best practice for a given query.

The proposed approach takes place in two phases: in the first one extracts good practices from the web using a web scrapping method, after we represent them by oriented data graphs. In the second phase, we extract the best practice for a given query by applying the techniques of machine learning and text summarization on graphs. This phase takes place in three steps: (1) search for practices similar to the user's query, here we use the word embedding model to identify sentences similar to the goal sought by the user; (2) Grouping and fusion of similar steps, where we use unsupervised learning (DBScan) and text summarization (PageRank) techniques to group semantically close nodes that we merge in the same step; (3) Extraction of the best practice that is identified by the path of the graph traversing the most important steps to reach the objective, this importance is calculated by measures of centrality of the graphs which quantify the importance of the nodes in a graph oriented by the number of their incoming and outgoing arc.

The results obtained demonstrated the superiority of our approach for: (1) capturing practices similar to the goal sought by the user, and this by optimizing the execution time, (2) extracting the best practices for queries compared to a search engine from a real data set.

Keywords: Community of practice, Good practice, Procedural knowledge extraction, Knowledge graph, Text synthesis, Machine learning.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

1

Sommaire

1.1 Contexte

Depuis toujours l'esprit communautaire est au cœur du rapprochement des gens soit par la religion, la géo-localisation, la langue, le domaine d'intérêt, les liens du sang etc. La communauté rassemble en elle-même les individus partageant un intérêt commun quel qu'il soit, en sociologie Jean Lave et Etienne Wenger Lave et Wenger [1991] définissent un genre particulier de communautés, il s'agit de communauté de pratiques (CdP), où l'on favorise le partage de savoir faire dans un domaine particulier sous forme de bonnes pratique appelée aussi connaissance procédurale décrite par un ensemble d'étapes successives pour atteindre un objectif particulier, ainsi l'apprentissage collaboratif est inhérent à ces communauté en ce que leurs membres apprennent les uns des autres en rendant leurs connaissances et pratiques explicites, en les partageant avec d'autre membre de la même communauté et en raisonnant à leur sujet.

Les CdP sont différentes des organisations formelles et des situations d'apprentissage. En fait, ce sont des groupes de personnes qui partagent une préoccupation, des problèmes ou une passion pour un sujet (domaine), approfondissent leurs connaissances pratiques et leur expertise dans le domaine considéré (pratique) et interagissent de manière continue (communauté). D'habitude, les CdP émergent souvent du contexte d'organisations ou d'associations professionnelles existantes physiquement, où les gens sont déjà impliqués dans des pratiques professionnelles communes, mais avec l'ascension du web social et sémantique ces dernières années, on a vu émerger de multiples sites de parages, réseaux sociaux, et de plate forme collaborative tels que Facebook, Twitter, WikiHow, etc. ce qui a donner un nouvel essor aux CdP en ligne en les affranchissant des limitations géographiques et el leur facilitant via leur outils le partage et l'échange instantanées et sans restrictions de connaissances.

Dans ce contexte notre travail vise à développer une approche pour conceptualiser les bonnes pratiques au sein d'une CdP de santé en ligne afin d'accroître l'apprentissage individuel et en groupe dans les communautés de pratique et de favoriser la réutilisation des connaissances existantes du référentiel dans des cas similaires notamment celles qui traitent du domaine de la santé, l'objectif est d'extraire les bonnes pratiques dans une CdP et de chercher à identifier la meilleure pour une requête donnée en s'appuyant sur les techniques d'apprentissage artificielles sur les graphes.

1.2 Problématique et motivations

Notre travail vise d'une part à conceptualiser les bonnes pratiques au sein d'une Communauté de pratique (CdP) en ligne et d'autre part à extraire la meilleure pratique pour une requête donnée, ce qui revient à extraire et formaliser des connaissances procédurales dans une base de données et de les explorer afin d'identifier la meilleure pratique qui répond au mieux à la requête recherchée. Plusieurs travaux de recherche se sont penchés sur le problème d'acquisition de connaissances procédurales qui est devenue primordiale pour les applications avancées actuelles telles que SIRI l'assistant intelligent d'Apple, ou ALEXA l'assistant intelligent développé par Amazon Lab qui effectuent des tâches rapidement et simplement sur des appareils dédiés ou encore dans le domaine de la robotique à titre d'exemple l'application Tenorth et al. [2011] qui utilise les connaissances procédurales extraites de site de partage du web pour réaliser une crêpe étape par étape par un robot. Un autre exemple d'application de telles connaissances procédurales est la recherche d'information où les moteurs de recherches actuels exploitent les graphes de connaissances pour répondre aux requêtes des nouveaux utilisateurs qui recherchent davantage des connaissances explicites sur la façon de faire les choses plutôt que sur des informations basiques (date, heure, etc.).

Dans cette perspective le web offre l'avantage d'être un recueil de connaissance mondiale caractérisé par une abondance d'informations, mais exploiter de tel type d'information reste

très difficile de par le fait que ces connaissances sont semi ou généralement non structurées. L'extraction des connaissances à partir de base de données « ECD » dit Knowledge Discovery in Data Base «KDD» en anglais est le domaine qui évolue pour offrir des solutions à ce problème. L'ECD propose un processus pour extraire et analyser des données à partir de différente sources d'informations, dans notre cas il s'agit d'analyser et d'extraire les connaissances procédurales à partir de donnée web partagées par des membres de CdP en ligne et de les représenter par un formalise adéquat dans le but d'extraire la meilleure pratique pour une requête donnée et assister l'utilisateur dans son processus de recherche. Le problème d'exploration des données à fait l'objet de plusieurs travaux de recherches comme nous l'expliciterons dans le chapitre 3, ce processus s'appuient sur différentes techniques d'analyse de réduction, et de machine learning, etc. pour arriver à un résultat exploitable.

1.3 Contribution

Ce travail a pour objectif de conceptualiser les meilleures pratiques au sein d'une CdP, de ce fait notre contribution est la proposition d'une approche pour extraire et formaliser les connaissances procédurales d'une part et d'autre part identifier la meilleure pratique pour une requête donnée.

Il est judicieux de constater que cette thématique que nous traitons touche des concepts propres aux domaines de sociologie, marketing, etc. tel que définit par Jean Lave et Etienne Wenger Lave et Wenger [1991] et il faut préciser aussi que dans la littérature à notre connaissance on ne trouve pas de critère de classification de meilleures pratiques, de ce fait dans le chapitre 2 nous proposerons une définition formelle des bonnes pratiques, nous allons aussi suggérer une définition de la meilleure pratique et proposer des moyens de classifications pouvant aboutir à l'identification des meilleures pratiques. Dans ce document, nous présenterons aussi un état de l'art sur les deux principaux domaines autour des quels s'axent notre problématique à savoir de l'extraction et la représentation des connaissances : dans le chapitre 3 nous énumérons les principaux processus d'extraction existants et expliquerons ainsi les différentes méthodes d'explorations de données et dans le chapitre 4 nous présenterons les différents formalismes de représentation des connaissances, on poussera l'état de l'art au chapitre 5 ou nous relaterons les approches récentes connexes à notre problématique ainsi qu'une synthèse afin de comparer ces travaux.

Notre contribution majeure est la proposition d'une mèthode pour extraire les meilleures pratiques au sein d'une communauté de pratique en ligne basée sur les techniques d'apprentissage artificiel sur les graphes présentée dans le chapitre 6 de ce mémoire. Cette contribution recouvre deux parties : la première concerne la conceptualisation des bonnes pratiques du web dans une base de données graphique par une technique de web scraping qui recueille les données textuelles du web. Ensuite les connaissances procédurales extraites sont formalisé en un graphe orienté. La seconde partie de l'approche proposée porte sur l'extraction de la meilleure pratique pour répondre à une requête donnée, et ceci en se basant sur l'hypothèse qu'une meilleure pratique est la pratique ayant en son sein les étapes les plus utilisées par les autres pratiques visant à atteindre un objectif commun. La première étape de cette phase est la recherche dans notre base de connaissances des pratiques similaires à l'objectif recherché par l'utilisateur, dans ce cas la on propose une méthode mathématique intuitive basée sur le modèle de prolongement lexical de mots dit Word embedding en anglais, plus précisément on utilise le réseau neuronal à deux couche Word2Vec Mikolov et al. [2013], qui effectue la représentation vectorielles des mots grâce à leur contexte et on calcule ainsi la distance sémantique entre les phrases représentant la requête et les titres des bonnes méthodes existantes on obtient ainsi un graphe orienté reliant un nœud initial qui représente la requête de l'utilisateur avec les nœuds qui représente les bonnes pratiques qui lui sont similaires ainsi que toutes les étapes qui les décrivent. Afin de remédier à la redondance de certaines étapes qui peuvent être

sémantiquement identiques et unifier ainsi notre représentation graphique, nous procédons dans la seconde phase au regroupement des nœuds similaires en utilisant la classification non supervisée par l'algorithme DBSCAN Ester et al. [1996], et à partir de chaque cluster de nœuds obtenu on fusionne les étapes similaires par une technique de résumé de texte appelée classement qui est inspirée du célèbre algorithme de Google PageRank Brin et Page [1998]. Enfin dans la dernière étape on procède à l'extraction de la meilleure pratique répondant à la requête de l'utilisateur qui se traduit par l'identification du chemin ayant les étapes les plus utilisées par les autres méthodes, et pour se faire on se base sur la théorie des graphe, notamment sur la mesure d'importance dans un graphe exprimée en terme de centralité de degré qui quantifie la popularité des nœuds dans les graphes par le nombre de leur voisins (nombre d'arcs entrants et sortant d'un nœud), cette idée de centralité n'est pas nouvelle, elle est utilisé dans plusieurs domaines surtout dans l'analyse des réseaux sociaux pour identifier les personnes influentes et les communautés, etc. et en recherche d'information pour mesurer l'importance d'un document. Dans notre cas on se basera sur le nombre d'arcs entrant des nœuds parcourus par chaque chemin afin d'atteindre l'objectif recherché pour quantifier la valeur d'importance de tous les chemins existants dans le graphe et identifier ainsi le chemin ayant les étapes les plus utilisée représentant la meilleure pratique pour la requête de l'utilisateur.

1.4 Organisation du document

Cette thèse est constituée de deux parties principales : la première intitulée «État de l'art», introduit le contexte dans lequel se place cette thèse et montre un état de l'art de notre problématique. La seconde partie du document intitulée « Contributions », présente l'approche proposée pour extraire et formaliser les bonne pratique d'une CdP en ligne d'une part et d'autre part identifier la meilleure pratique pour une requête donnée ainsi que l'ensemble des expérimentations effectuées. L'état de l'art quant à lui inclut quatre chapitres :

- **Chapitre 2 : Communauté de pratique**
  Dans ce chapitre on introduit les fondements et les définitions de base des communautés de pratiques, et des communautés en ligne en particulier, nous proposons aussi de définir les meilleures pratiques et plus formellement les bonnes pratiques, et enfin nous présentons les différents systèmes de classification des connaissances procédurales.
- **Chapitre 3 : Extraction des connaissances**
  Ce chapitre est consacré à une étude bibliographique des différentes méthodes d'extraction de connaissances à partir de base de données ECD, nous présenterons les différentes définitions de l'ECD, et nous ferons une distinction avec la fouille de donnée, nous verrons les principaux types de données à explorer et la classification des méthodes de fouilles existantes.
- **Chapitre 4 : Représentation des connaissances**
  Ce chapitre est consacré à l'autre domaine de la problématique qui est la représentation des connaissances, nous présentons ainsi les différentes définitions du domaine ainsi que les principaux formalismes de représentation, et les différents problèmes qu'on peut retrouver lors du processus de modélisation.
- **Chapitre 5 : Travaux connexes : approches d'extraction et de représentation des connaissances procédurales**
  Ce chapitre présente les principaux travaux récents traitant de notre domaine ainsi qu'une étude comparative de ces derniers, nous présentons ainsi les différentes méthodes d'ECD utilisées par les travaux connexes pour extraire et formaliser des connaissances procédurales afin de favoriser leur réutilisation et leur exploitation dans différents domaines.

La seconde partie du document regroupe trois chapitres :

— **Chapitre 6 : Extraction des meilleures pratiques au sein d'une communauté de pratique par l'apprentissage artificiel sur les graphes**
Ce chapitre est consacré à notre proposition, nous y exposerons notre contribution d'une approche visant à conceptualiser les bonnes pratiques au sein d'une communauté de pratique en ligne et identifier la meilleure pratique pour une requête donnée. Nous déroulerons aussi un exemple d'application à petite échelle afin de mieux expliciter notre méthode.

— **Chapitre 7 : Implémentation et expérimentations**
Ce chapitre est consacré à l'expérimentation et à l'évaluation de notre approche Nous y présentons les résultats d'une expérimentation visant à valider l'approche adoptée et à tester les algorithmes conçus.

— **Chapitre 8 : Conclusion et Perspectives**
Ce chapitre est une synthèse des nos principales contributions et du travail que nous présentons dans ce document, nous y débattons aussi des perspectives possibles de ce travail.

Communautés de pratiques 2

Sommaire

2.1 Introduction

On parle souvent de communauté : communauté musulmane communauté de médecin, communauté de jeunes, etc. Ce concept englobe en fait un ensemble d'individus ayants soit des caractéristiques communes ou partageant des objectifs, des croyances ou encore des savoirs faires similaires. Cette notion prend une identité particulière quand l'activité principale de la communauté s'axe autour de partage de pratique bien définit, on parle alors de communauté de pratique. La pratique en elle-même peu importe le domaine qu'elle recouvre, est vue comme un modèle à suivre afin de réaliser une tâche bien définie. Elle peut être caractérisée comme étant une bonne pratique si celle-ci à fait ses preuve au sein d'un membre de la communauté.

Dans ce chapitre nous allons présenter selon la littérature les communautés de pratique leurs objectifs et caractéristiques, nous étendrons le concept aux communautés du web. Nous spécifierons par la suite la différenciation entre le concept de bonne et meilleure pratique

2.2 Communauté de pratique (Cdps / Cops : Communities of parctices)

2.2.1 Concept et définition

La notion de communauté de pratique a été développée par Jean Lave et Etienne Wenger Lave et Wenger [1991], Wenger [2000] comme base d'une théorie sociale de l'apprentissage. Une communauté de pratique est une collection de personnes qui s'engagent de façon continue dans une entreprise commune.

La valeur de la notion de communauté de pratique réside dans le fait qu'elle identifie un groupement social non en raison de caractéristiques abstraites partagées par exemple la classe sociale, le genre ou le rapprochement géographique par exemple le quartier, le lieu de travail, mais en vertu de la pratique partagée. Dans le cadre d'une activité conjointe régulière, une communauté de pratique développe des façons de faire des choses, des points de vue, des valeurs, des relations de pouvoir, des moyens de parler. Au sein du groupe, les individus créent des connaissances et les échangent entre eux.

Afin de conclure à une communauté de pratique, un sentiment d'appartenance au groupe est nécessaire, et doit être constaté par les membres eux-mêmes. Ce sentiment dépend entre autres du niveau d'implication de chacun au sein du groupe, et de l'impression de partager une identité commune. De manière globale, trois éléments sont essentiels à la constitution d'une communauté de pratique :

- **Le domaine :** une communauté de pratique a une identité définie par un domaine d'intérêt partagé. L'adhésion implique donc un engagement envers le domaine.
- **La communauté :** dans la poursuite de but particulier, les membres du groupe s'engagent dans des activités et des discussions conjointes, s'entraident et partagent des informations. Ils établissent des relations qui leur permettent d'apprendre les uns des autres.
- **La pratique :**une communauté de pratique n'est pas seulement une communauté d'intérêt. Les membres d'une communauté de pratique sont des praticiens. Ils développent un répertoire de ressources partagé : les expériences, les histoires, les outils, les moyens d'aborder les problèmes récurrents, bref une pratique partagée.

2.2.2 Objectifs d'une communauté de pratique

Les communautés de pratique peuvent avoir différents objectifs Wenger et Snyder [2000] tels que piloter une stratégie, donner naissance à une nouvelle activité, résoudre un problème, promouvoir la diffusion des bonnes pratiques, développer les compétences professionnelles

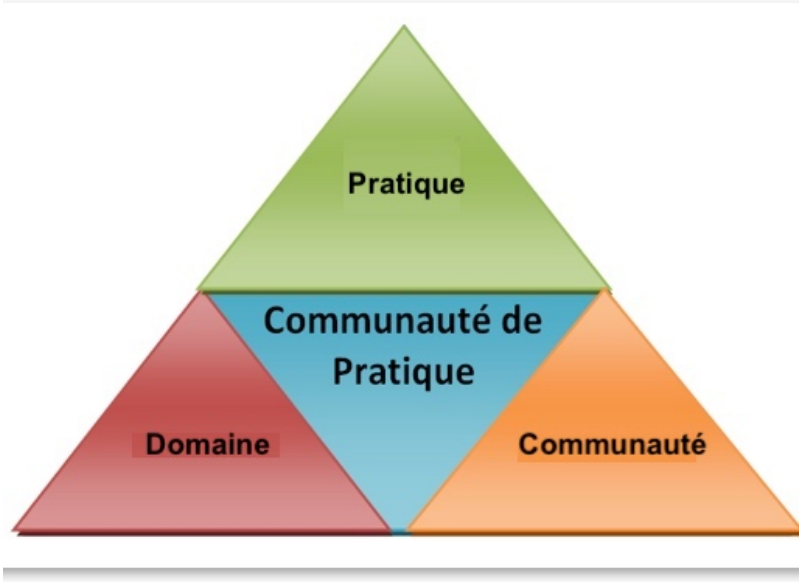

FIGURE 2.1 – *Les composants principaux d'une communauté de pratique*

des individus, aider les entreprises à embaucher et à retenir les meilleurs talents. Au sein d'une communauté de pratique, les gens partagent des connaissances qui sont habituellement tacites et c'est au moyen du dialogue que ces connaissances s'expriment et peuvent alors être repérées, identifiées, échangées. Les membres d'une communauté échangent des idées sur leur pratique de travail et expérimentent des nouvelles méthodes et idées. Ils s'engagent dans des discussions qui contribuent parfois à remettre en question les théories en usage. Ils innovent en développant de nouvelles routines de résolution de problèmes. Autrement dit, les participants d'une communauté de pratique s'engagent développent et raffinent leurs structures cognitives. De plus, ces communautés participent à la formation et à la transformation des pratiques, des idées et de la culture organisationnelle.

2.2.3 Communauté de pratique et apprentissage

L'origine et l'utilisation principale du concept de communauté de pratique ont été dans la théorie de l'apprentissage Wenger [1998]. Les gens pensent généralement à l'apprentissage comme une relation entre un étudiant et un enseignant, mais les études d'apprentissage révèlent un ensemble de relations sociales plus complexes par lesquelles l'apprentissage a lieu principalement avec des compagnons et des apprentis plus avancés. Le terme «communauté de pratique» a été inventé pour se référer à la communauté qui agit comme un programme d'études vivant pour l'apprenti.

L'apprentissage implique la participation à une communauté de pratique. Et cette participation se réfère non seulement aux événements locaux d'engagement dans certaines activités avec certaines personnes, mais à un processus plus englobant d'être des participants actifs dans les pratiques des communautés sociales et de construire des identités par rapport à ces communautés. L'apprentissage est un ensemble de relations évolutif et continuellement renouvelé. Au début, les gens doivent rejoindre les communautés et apprendre à la périphérie, a mesure qu'ils deviennent plus compétents, ils se déplacent plus vers le «centre» de la communauté en particulier. L'apprentissage n'est donc pas considéré comme l'acquisition de connaissances par les individus tant qu'un processus de participation sociale.

2.2.4 Les communautés de pratiques du web social

Avec l'avènement du web social et sémantique, nous avons vu l'apparition de nombreux sites de partages tels que Facebook, Youtube, Wikipedia, Wikihow, etc. cet essor a crée ce qu'on appelle les communautés virtuelles, où des groupes de personne à travers internet interagissent entre eux, afin de partager et d'échanger des données et ainsi créer des liens, qui forment ce qu'on appelle aujourd'hui les réseaux sociaux.

À travers les années, ces communautés se sont de plus en plus spécifié et affinée grâce aux différents outils et plates formes du web selon l'objectif et les intérêts des participants, créant ainsi par exemple des blogs spécialisés en cuisine, en jeux vidéo, en TIC, etc. Ou encore des groupes de partages de praticiens ou d'experts en santé ou en enseignement, etc. Par analogie au concept communautés de pratiques définit par Wenger, les communautés en lignes ont développé des identités spécifique à des domaines d'intérêt, basé sur l'échange et le partage d'information, savoir faire, on encore de pratique. Donc une communauté virtuelle partageant les pratiques peut être communément appelée communauté de pratique en ligne, néanmoins le processus de réification dans ce cas reste difficile du fait que les connaissances partagé ne sont pas structurées ou sont semi structurées.

2.3 Les bonnes pratiques et les meilleures pratiques

2.3.1 Définitions

Selon le dictionnaire Sensagent [2016], le terme «bonnes pratiques» désigne, dans un milieu professionnel donné surtout en entreprise, un ensemble de comportements qui font consensus et qui sont considérés comme indispensables par la plupart des professionnels du domaine, qu'on peut trouver sous forme de guides de bonnes pratiques(GBP). Ces guides sont conçus par les filières ou par les autorités. Ils peuvent se limiter aux obligations légales, ou les dépasser. Comme les chartes, ils ne sont opposables que s'ils ont été rendus publics. Ils sont souvent établis dans le cadre d'une démarche qualité par les filières.

Selon ONUAA [2014], Une bonne pratique n'est pas seulement une pratique qui est bonne, mais une pratique ayant fait ses preuves et permis d'obtenir de bons résultats. C'est une expérience réussie, testée et validée à travers différentes phases de reproductibilité et qui, dès lors, peut être recommandée comme modèle et mérite d'être partagée, de sorte qu'un plus grand nombre de personnes peuvent l'adopter.. Elle doit aussi être évolutive et adaptative afin de s'adapter à de nouveaux défis et s'améliorer à mesure que des améliorations sont découvertes.

Plus formellement, nous définissons une bonne pratique comme une méthode ou un processus utilisé pour atteindre un objectif particulier ce processus n'est pas figé dans le temps il peut à tout moment évoluer, ou changer selon le contexte. En outre ce processus représente un savoir-faire sous forme de procédure constituée d'un ensemble d'étapes successives, chaque processus peut en amont nécessiter l'utilisation d'objet ou de contrainte spatio-temporelle spécifique au contexte de la connaissance véhiculée et de l'objectif à atteindre, par exemple si la communauté partagent des pratiques culinaires, elles peuvent avoir recours à des ustensiles de cuisines ou encore à des ingrédients spécifiques, et en même temps être assujetties à des contraintes temporelles et environnementales de temps et de degrés de cuisson nécessaires à la réussite de la pratique. La figure 2.2 représente une bonne pratique comme un processus tel qu'il est définit dans cette section.

Par ailleurs si on passe maintenant au concept de meilleure pratique nous trouverons des définitions spécifiques au domaine telles qu'en économie et en gestion, où l'on définit une meilleure pratique comme une méthode, ou un ensemble de méthodes de travail, qui est officiellement reconnue comme la meilleure à utiliser dans une entreprise ou une indus-

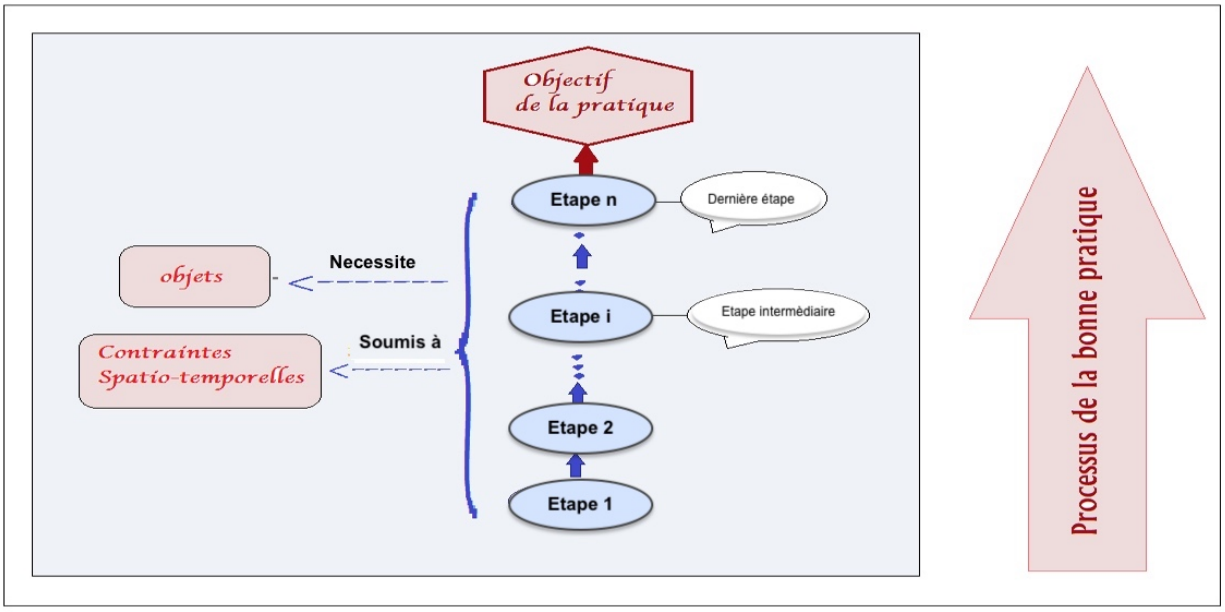

FIGURE 2.2 – *Représentation de la bonne pratique*

trie particulière Dictionary [2020]. Elle est considérée comme le moyen le plus efficace pour atteindre un résultat particulier Techopedia [2012].

Faire la différence entre une bonne et une meilleure pratique est une chose assez subjective, une bonne pratique est certainement une méthode qui fonctionne et qui a fait ses preuves pour atteindre un certain objectif, une meilleure pratique serait donc une bonne pratique triée et sélectionnée parmi d'autres bonnes des méthodes dans un contexte bien défini selon des critères de comparaison ou de votes qui ont démontré la supériorité de cette pratique par rapport aux autres.

2.3.2 Comparaison des bonnes pratiques

Afin d'identifier la meilleure pratique, il est nécessaire de mettre au point un système d'évaluation. Dans cette section nous proposons des critères de comparaison des bonnes pratiques notamment sur le web.

— **Système de notation et d'évaluation**
Selon le dictionnaire CNRTL [2012] la notation est l'action de traduire l'appréciation d'un travail par une note en chiffres ou en lettres accompagné généralement d'un commentaire. L'évaluation, consiste à établir une comparaison entre un produit, une réponse, une idée, etc. donné et une norme de référence. Ces deux concepts nous suivent depuis notre scolarisation, on les considère comme des repères dans tout système de comparaison afin de se positionner, et ils ont pris plus d'ampleur depuis l'apparition du e-commerce, où toute transaction, tout produit en vente est soumis à un système d'évaluation, et de notation, si un client veut acheter un article, louer un hébergement, consulter un médecin, etc. il n'a qu'a se référer au système de notation mis en ligne grâce aux algorithmes du web, le nombre d'étoile lui donnera un avis sur l'appréciation des gens, leurs commentaires et avis peut renforcer ou dissuader le client dans sa démarche. De même un membre d'une communauté en ligne dans sa quête de la meilleure pratique pour réaliser un but précis peut se référer aux systèmes de notation afin de choisir la méthode pour achever son objectif, à titre d'exemple quelqu'un cherchant à faire des crêpes trouvera un grand nombre de façon de faire qui différent, la notation et les avis peuvent l'aider à choisir la bonne pratique. Néanmoins ces systèmes

d'évaluation restent subjectifs à l'avis de chacun et à la vision personnelle de l'individu selon son contexte qu'une façon de faire est plus avantageuse qu'une autre.

— **Système basé sur un critère d'optimisation :**
Dans certain cas, un membre d'une communauté de pratique peut chercher à réaliser un but tout en optimisant un critère particulier ou même en combinant plusieurs, dans ce cas là le facteur d'optimisation peut être utilisé pour comparer les bonnes pratiques et identifier la meilleure. L'objet ou le facteur à optimiser peut différer selon la situation : le temps peut être un facteur important, on peut chercher la meilleure façon d'accomplir une tâche dans un temps précis, ou encore dans un espace limité en utilisant une liste fini d'objet, par exemple quelqu'un peut vouloir faire du pain avec un nombre minimum d'ingrédient et dans un laps de temps. Dans ce cas le système d'optimisation peut se baser sur les méthodes de recherches opérationnelles afin de comparer les bonnes pratiques. Ces situations là restent à exploiter dans un contexte bien précis relatif aux choix et conviction de l'utilisateur et peuvent ne pas refléter de manière fiable et globale la supériorité d'une pratique par rapport à une autre.

— **Système de comparaison par étapes :**
Si l'on se réfère à la définition formelle d'une bonne pratique citée dans les paragraphes précédents, ce serait un ensemble d'étapes successives suivies pour achever un but précis, comparer donc les bonnes pratiques entre elle revient à comparer leur complétude par rapport aux étapes qu'elles empruntent, c.à.d. qu'une meilleure pratique serait une procédure utilisant en son sein un ensemble d'étapes communes à d'autres bonnes pratiques pour atteindre un objectif commun. En d'autres termes, la meilleure pratique est une bonne pratique regroupant l'ensemble des étapes les plus utilisées par d'autres bonnes pratiques pour atteindre le même objectif. L'idée d'un tel système de comparaison existe déjà dans plusieurs domaines par exemple dans l'analyse des réseaux sociaux on peut identifier les personnes influentes par l'importance de leurs liens d'amitié, ou encore dans la recherche d'information pour le classement des documents web (Ranking) notamment si l'on fait référence au fameux algorithme «PageRank» Brin et Page [1998] utilisé par les moteurs de recherche de Google qui dit qu'une page est importante si elle est pointée par un nombre élevé d'autres pages, donc les sites les plus importants sont classés en fonction de leur popularité est qui vaut au nombre de fois qu'il seront cités par d'autre sites. Par analogie nous pouvons appliquer le fondement du Page Rank pour avoir un système de comparaison des bonnes pratiques par rapport à la popularité des étapes qu'elles contiennent, ainsi la pratique qui empruntera les étapes les plus utilisées par les autres pratiques pour atteindre le même objectif sera considérée comme supérieure aux autres.

Dans cette section, nous avons proposé trois systèmes de comparaison afin d'évaluer les bonnes pratiques car il n'existe pas dans la littérature concrètement de fondement pour le faire. Il est possible toutefois de combiner deux ou les trois systèmes de comparaison proposés selon le contexte de la requête du membre de la communauté afin d'identifier la meilleure pratique, par exemple si un utilisateur veut savoir quelle est la meilleure méthode pour perdre du poids dans une durée de temps limitée, le système de comparaison peut s'appuyer sur la popularité étapes des bonnes pratiques ou sur le système de notation tout en prenant en compte le facteur de temps comme variable à optimiser.

2.4 Conceptualisation des bonnes pratiques

La conceptualisation des bonnes pratiques au sein d'une communauté est le fait de rendre sous forme de concept les connaissances partagées en les formalisant, l'enjeu principal est l'identification et le partage des connaissances procédurales afin de favoriser le processus

d'apprentissage au sein de la communauté de partage. Il est donc nécessaire de développer une approche pour extraire les connaissances dans une communauté, et de choisir un model de représentation afin de rendre ces pratiques réutilisables par les autres membres, toutefois il est laborieux de faire un tel processus manuellement, il est donc nécessaire d'automatiser l'extraction et la représentation des connaissances, de plus il reste difficile de trouver une base de bonnes pratiques exploitable. Dans ce contexte le web fournit une source de connaissance idéale puisqu'il héberge un grand nombre de site de partage et de communauté en ligne, le problème reste que le formalisme de données en ligne est généralement semi ou pas structuré du tout ce qui rend leur exploitation difficile. Actuellement, plusieurs travaux de recherches tentent d'extraire et de formaliser les connaissances procédurales du web, car de tels données sont devenues la base des application modernes tel que SIRI ou Alexa, leurs utilisation est tout aussi importante pour les moteurs de recherches qui ne se contentent plus de retourner des informations élémentaires tel que l'heure ou la météo, mais de répondre au mieux aux requêtes complexes des utilisateurs cherchant plus les méthodes de faire tel que «comment soigner le COVID» ou encore «comment faire des plantations en lasagne». Donc la problématique de conceptualisation des bonnes pratiques au sein des communautés de pratiques traitée dans ce mémoire est relative au domaine d'extraction et de représentation des connaissances du web.

2.5 Conclusion

Dans ce chapitre nous avons vu le concept de communautés de pratique définit par Wenger. Conceptualiser les bonnes pratiques partagées dans de ces communautés notamment celles du web est une entreprise très intéressante à exploiter surtout dans le domaine de la recherche d'information avancée et dans les technologies d'informations et de communications actuelles. Ceci reviendrait à traiter les méthodes d'extraction et de représentation des connaissances procédurales sur le web. Dans ce contexte dans le chapitre suivant nous présenterons les principales approches d'extraction de connaissances existantes dans la littérature.

Extractions des connaissances 3

Sommaire

3.1 Introduction

De nos jours, la quantité d'informations partagée sur le web dépasse de loin notre capacité à réduire et à analyser ces données sans l'utilisation de techniques d'analyse automatisées, surtout avec l'avènement du web social et sémantique et des différentes technologies en ligne. Parmi les données partagées, on retrouve de plus en plus les connaissances sous forme procédurales dîtes bonnes pratiques qui priment dans les communautés en ligne du fait que l'internaute d'aujourd'hui recherche davantage des informations sur la façon de faire les choses plutôt que des informations basiques telles que l'heure et le climat. L'enjeu majeur dans ce cas est d'extraire ces connaissances procédurales afin de les automatiser et faciliter leur réutilisation.

La découverte de connaissances dans les bases de données (ECD) est le domaine qui évolue pour offrir des solutions d'analyse automatisés. Ce sont de nouvelles approches méthodologiques qui tendent à extraire et à explorer les connaissances à partir de différente source. Le concept d'Extraction de Connaissance à partir de Données (ECD) ou Knowledge Discovery in Data Base (KDD) a été inventé lors du premier atelier KDD en 1989, pour souligner que la connaissance est le produit final d'une découverte piloté par les données. Il a été popularisé dans l'intelligence artificielle et dans le domaine d'apprentissage automatique Piatetsky-Shapiro [1991].

L'ECD est souvent confondu dans la littérature avec le concept de fouille de donnée (datamining en anglais, appelé également exploitation stratégique de données) et sont parfois considérés comme synonymes Feldman et al. [1999]. Cependant, la définition la plus partagée du concept de datamining le voit comme la phase essentielle de recherche de connaissances intervenant dans le processus plus général de découverte de connaissances dans les données Fayyad et al. [1996].

Dans ce contexte nous présenterons dans ce chapitre l'extraction des connaissances à partir de données, nous parcourons les principaux modèles d'ECD présents dans la littérature, et nous verrons par la suite les méthodes de fouille de données.

3.2 Extraction des connaissances

La notion de découverte de connaissances revêt plusieurs termes comme celui d'Extraction de Connaissances à partir de Données (ECD) et sa traduction en anglais «Knowledge Discovery in Databases (KDD) », aussi comme celui de constitution de modèles (patterns) à partir des données, ... Quelque soit la terminologie utilisée, l'essence même de l'ECD est la découverte non triviale, à partir de données, d'une information implicite, précédemment inconnue et potentiellement intéressante. Une telle « information » extraite d'une base de données devient alors une « connaissance » Frawley et al. [1992] et Kodratoff [1995].

En 1996 Fayyad et Shapiro définissent le processus KDD Fayyad et al. [1996].comme un processus non trivial, permettant l'identification, au sein des données, de nouveaux patterns valides (pour de nouvelles données avec un bon degré de certitude), potentiellement utiles (i.e. devraient conduire vers des décisions utiles), et les plus compréhensibles possible (par des humains).. . ., par des moyens automatiques ou semi-automatiques, de grandes quantités de données en vue d'extraire des motifs intéressants.

3.3 Les modèles du processus de l'ECD

L'extraction des connaissances à partir des données s'effectuent selon un processus bien précis, qui englobe par un ensemble d'étapes successives. Chaque étape commence après que les précédentes soient bien finalisées, parce qu'elle utilise les sorties des étapes précédentes et

Le processus est aussi itératif, ce qui signifie que parfois il peut être nécessaire de refaire les pas précédents.

Plusieurs modèle d'ECD ont été proposé dans littérature, le premier processus KDD de base a été proposé en 1996 par Oussama Fayyad Fayyad et al. [1996], qui a été amélioré ultérieurement. Le modèle de Fayyad comprend 9 étapes successives comme montré dans la figure 3.1 :

- **Première étape :**
  C'est le pas initial de ce processus, où on cherche à développer et comprendre le domaine de l'application et à identifier l'objectif du processus de KDD du point de vue du client.
- **Deuxième étape :**
  A ce niveau on se concentre sur la création d'un groupe de données cible sur lequel va être appliqué le processus d'exploration.
- **Troisième étape :**
  Dans cette étape, on procède au nettoyage des données grâce à des opérations de base comprennent la suppression du bruit ou les valeurs aberrantes.
- **Quatrième étape :**
  Cette étape concerne la transformation des données, elle est très importante pour la réussite du projet et doit être adaptée en fonction de chaque base de données et des objectifs du projet. A ce stade il faut chercher les méthodes correctes de représentation des connaissances permettant de réduire le nombre effectif de variables à étudier. Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie de datamining.
- **Cinquième étape :**
  A ce niveau on doit sélectionner la tâche d'exploration de données appropriée ; afin de réaliser l'objectif du processus de l'ECD. Plusieurs méthodes peuvent être utilisées : la classification, la régression, le clustering, etc.
- **Sixième étape :**
  Cette étape repose sur le choix de l'algorithme d'exploration de données (datamining). On doit sélectionner la méthode (s) à utiliser pour la recherche de tendances ou modèles dans les données.
- **Septième étape :**
  C'est ici qu'on effectue l'exploration de données au sens propre du terme à travers l'implémentation de l'algorithme de Data Mining . Comme l'utilisateur final peut être plus intéressé à comprendre un modèle particulier. Cette étape propose de faire une recherche de motifs d'intérêt dans une forme particulière de représentation ou d'un ensemble de représentations telles que les règles de classification ou d'arbres, de régression, ou le clustering, etc.
- **Huitième étape :**
  Cette étape inclue l'évaluation et l'interprétation des motifs découverts. Cette étape donne la possibilité de retourner à une des étapes précédentes, mais aussi d'avoir une représentation visuelle des motifs, d'enlever les motifs redondants ou non-représentatifs et de les transformer dans des termes compréhensibles pour l'utilisateur.
- **Neuvième étape :**
  C'est ici qu'on effectue la consolidation des connaissances découvertes, à travers leur intégration dans un autre système.

Une autre approche plus moderne de l'extraction des connaissances à partir des données est celle de Zigheb et Rakotomalala [2002], qui comme le montre le schéma de la figure 3.2, se veut un peu différente, on distingue ainsi deux niveaux dans le processus d'ECD : «niveau opérationnel et décisionnel» et le «niveau analyse». Le niveau opérationnel ou décisionnel s'appelle le front office. Le front office exploite les connaissances qui lui sont fournies par

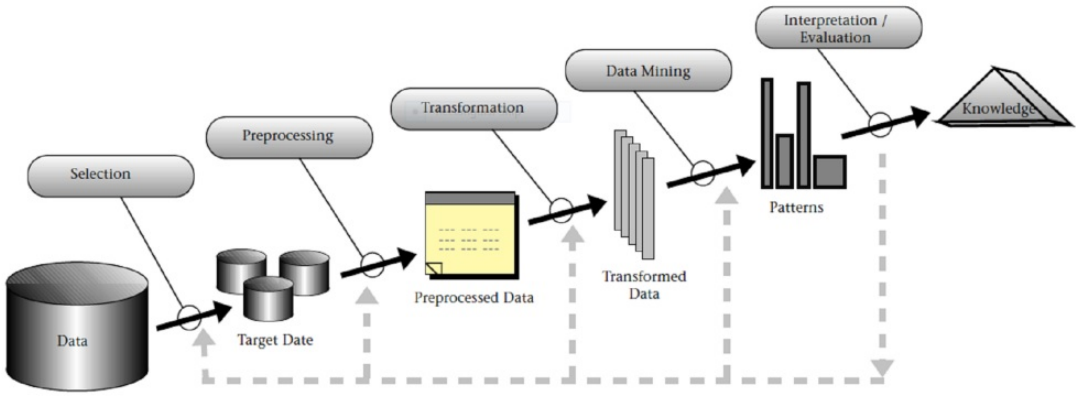

FIGURE 3.1 – *Le processus d'ECD Fayyad et al. [1996]*

les services études en vue de gérer au mieux la relation avec le client, par exemple dans les sites de vente en ligne, dés qu'un nouveau client se connecte, les connaissances sur ses préférences et son profil sont extraites afin de lui proposer les services et produits qui lui sont le plus adaptés. Les services études sont généralement situés en back office c'est le niveau analyse, qui est au centre des opérations d'extraction des connaissances à partir des données. Les données issues des bases de données de production, en service en front office, alimentent les entrepôts de données qui seront utilisées en ECD. Généralement, le processus d'ECD, se déroule en quatre phases : acquisition des données, prétraitement et mise en forme, fouille de et analyse, validation et mise en forme des connaissances (figure 3.2) :

- **Phase 1 Acquisition des données :**
  Dans cette phase on cible, même de façon grossière, l'espace des données qui va être exploré. Il est généralement question ici aussi de nettoyer les données qui sont visées. A l'issue de la phase, l'analyste est, a priori, en possession d'un stock de données contenant potentiellement l'information ou la connaissance recherchée.
- **Phase 2 Prétraitement des données :**
  Les données issues de la phase d'acquisition ne sont pas nécessairement toutes exploitables par des techniques de fouille de données. On peut y trouver des textes de longueurs variables, des images, des enregistrements quantitatifs ou des séquences vidéo. La préparation consiste à homogénéiser ces données. Le pré-traitement des données est donc l'acte de modélisation des connaissances dans la quelle l'expert devra choisir un canevas pour représenter les données et éventuellement effectuer une série de transformations afin de les adapter aux méthodes d'exploitation.
- **Phase 3 de fouille de données :**
  La fouille de données concerne le datamining et est au cœur du processus de l'ECD. Cette phase fait appel à de multiples méthodes issues de la statistique, de l'apprentissage automatique, de la reconnaissance de formes ou de la visualisation. Les méthodes de datamining permettent de découvrir ce que contiennent les données comme informations ou modèles utiles.
- **Phase 4 de validation et de mise en forme :**
  Les modèles extraits, notamment par les méthodes d'apprentissage supervisées, ne peuvent être utilisées directement en toute fiabilité. Nous devons les évaluer, c'est-à dire les soumettre à l'épreuve de la réalité et apprécier leur justesse. Le procédé habituel consiste à estimer au mieux le taux d'erreur du modèle. Ainsi, l'utilisateur décidera d'appliquer ou non le modèle de prédiction en connaissance des risques qu'il prend.

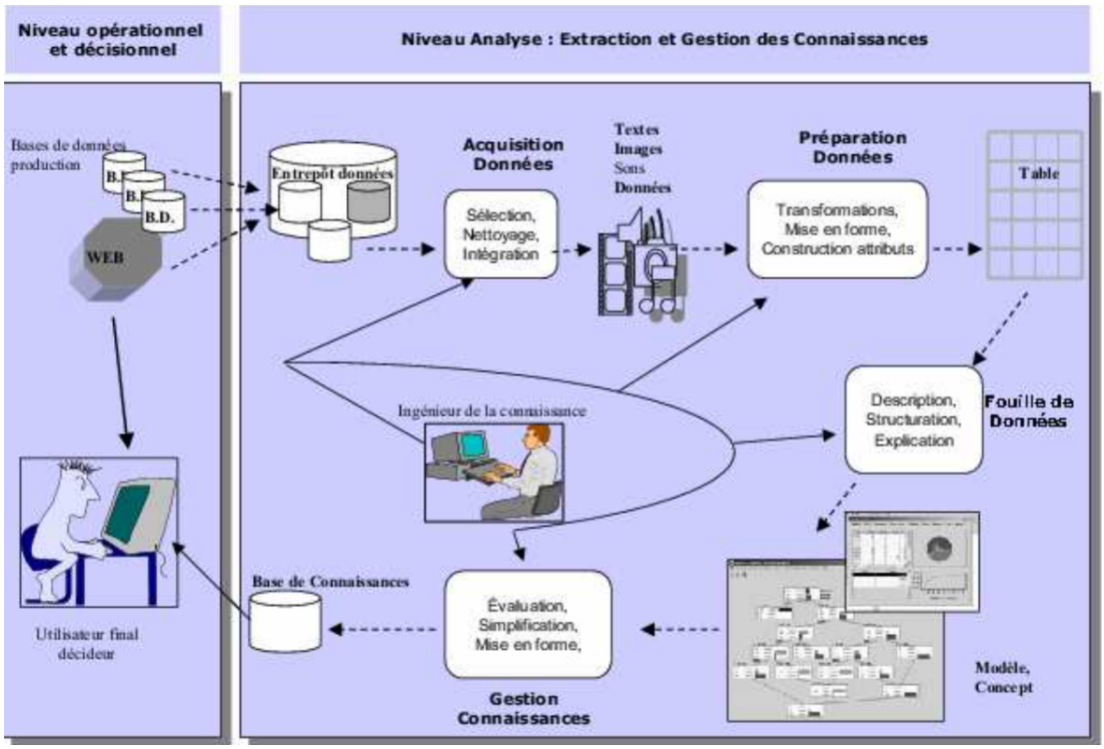

FIGURE 3.2 – *Processus général d'ECD Zigheb et Rakotomalala [2002]*

En partant du modèle de Fayyad Fayyad et al. [1996] dans les dernières années, les efforts ont été orientés pour trouver d'autres processus ou méthodologies d'ECD. Il existe plusieurs autres modèles qu'on ne peut pas citer de manière exhaustives par exemple on retrouve dans Kurgan et al. [2001] une classification des modèles d'ECD orientés vers la recherche ou l'industrie. Parmi les modèles orientés vers la recherche on cite celui d'Anand et Buchner Anand et al. [1998a], Anand et al. [1998b] qui ont développé une méthodologie hybride en 8 étapes pour résoudre les problèmes de ventes croisées et pour analyser les données de marketing sur Internet. Ces huit étapes comportent : ***l'identification des ressources humaines*** (identifie les ressources humaines et leur rôle), ***la spécification du problème*** (divise le projet en plusieurs tâches et chaque tâche sera résolue par une méthode particulière de fouille de données), ***la prospection de données*** (analyse l'accessibilité et la disponibilité des données) ***l'licitation (extraction) des connaissances du domaine, la spécification de la méthodologie de data mining*** pour résoudre le problème.,***le pré-traitement des données***(suppression des valeurs aberrantes, des données bruitées, transformation et codage, etc.), ***la fouille des données*** : découvre des motifs dans les données pré-traitées. et enfin ***la validation et visualisation de connaissances découvertes***. L'autre modèle orienté aussi vers la recherche, est celui de Kurgan et al. [2001] qui a été proposé pour répondre aux besoins académiques. Ce modèle est constitué de six étapes, on commence en premier lieu par comprendre et déterminer les objectifs du domaine et du data mining, et ensuite par comprendre les mécanismes pour collecter, explorer et vérifier les données, les étapes suivantes concernent respectivement : la préparation des donnée, l'implémentation de différents algorithmes de fouille de donnée, l'interprétation des résultats et la recherche des améliorations possibles pour les algorithmes et en dernier lieu la création d'un plan pour superviser l'implémentation des connaissances découvertes, la documentation du projet, l'extension de l'application dans d'autres domaines.

Dans la catégorie des modèles orientés vers l'industrie on retrouve le modèle de l'institut SAS SAS [1998] qui divise la fouille de données en cinq étapes représentées pas l'acronyme SEMMA « Sample, Explore, Modify, Model, Asses ». SEMMA est intégré comme outil en enterprise, ce qui diffère de l'ECD qui est un processus ouvert pouvant être appliqué dans

plusieurs environnements. Cette méthodologie extrait des échantillons d'un vaste ensemble de données, ensuite explore ces données en recherchant les tendances et les anomalies imprévues afin de mieux les cerner. En troisième point les données sont modifiées en transformant les variables afin de s'axer sur le processus de sélection de modèles. On effectue alors la modélisation automatique des connaissances grâce à un logiciel qui permet de rechercher automatiquement une combinaison des données qui prédit de façon fiable le résultat souhaité. Et enfin on évalue l'utilité et la fiabilité des résultats du processus de Data Mining. La figure 3.3 illustre les tâches d'ECD par la méthodologie SEMMA.

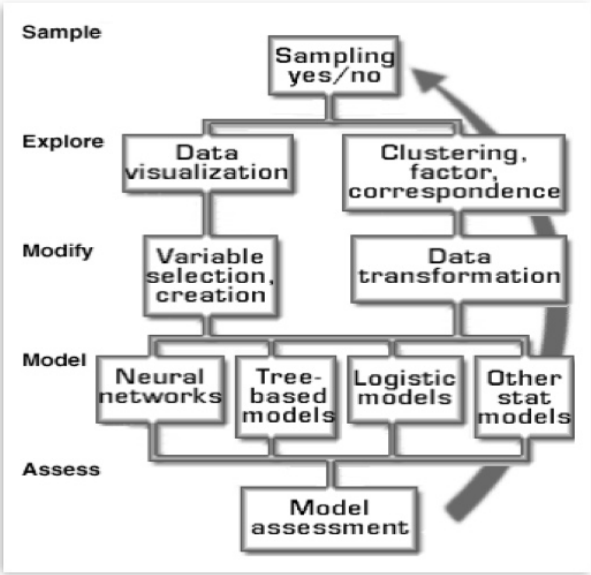

FIGURE 3.3 – *Les étapes dans la méthodologie SEMMA SAS [1998]*

Le modèle industriel le plus connu reste le CRISP-DM développé pour répondre aux besoins des projets industriels de data mining, il a été utilisé dans de domaines comme : l'ingénierie, la médecine, les ventes, le marketing, etc. CRISP-DM est décrit comme un processus hiérarchique constitué par plusieurs tâches, avec quatre niveaux d'abstraction : la phase, la tâche générique, la tâche spécialisée et l'instance du processus Shearer [2000] . CRISP-DM est l'acronyme de CRoss-Industry Standard for Data Mining et contient un cycle de six étapes : ***Business understanding (La compréhension du business)*** : cette phase initiale porte sur la compréhension des objectifs et des exigences du projet, ***Data understanding (La compréhension des données)*** : cette phase sert a se familiariser et identifier les données exploitables ; ***Data preparation (La préparation des données)*** : cette phase concerne le pré-traitement des données ; ***Modeling (La modélisation)*** : durant cette étape on sélectionne l'ensemble des méthodes pour modéliser les données ; ***Evolution (L'évolution)*** : à ce niveau le(s) modèle(s) sont évalué(s) et les étapes suivies pour la construction du modèle sont réévaluées pour s'assurer que le projet respecte les objectifs du business, définis au début du projet ; et ***Deployment (Déploiement)*** : cette phase concerne l'organisation et la présentation des connaissances extraites d'une manière utilisable par le client. Cette séquence de phases n'est pas obligatoire. On peut aller entre les phases, comme suggéré dans la figure 3.4 par la flèche qui indique les plus importantes et fréquentes dépendances entre les phases. CRISP-DM ne guide pas l'utilisateur sur comment les tâches doivent être réalisées, mais le modèle est facile à comprendre et très bien documenté.

On retrouve dans tous ces modèles les mêmes étapes importantes qui se répètent des fois sous diverse appellations, et qui font le processus d'extraction des connaissances, à savoir de manière successive : la sélection et le pré-traitement des données, la modélisation ou la transformation des données, la fouille des données, et dans certain cas l'évaluation ou le

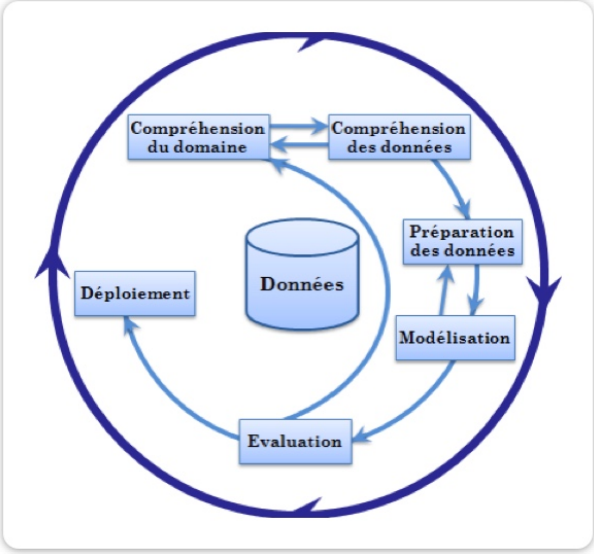

FIGURE 3.4 – *Le cycle de vie du CRISP-DM Shearer [2000]*

déploiement des données. Le choix de la méthodologie reste au bon vouloir de l'analyste et dépend du type de donner à extraire et des objectifs de l'ECD.

3.4 Fouille de données

La fouille de données est connue sous différentes appellations dans les différentes communautés, on retrouve les termes comme : la recherche de patterns dans les données ou encore l'exploration de données, l'extraction de la connaissance, la découverte d'informations, la récolte d'information, l'archéologie des données, et le traitement de modèle de données et le datamining Fayyad et al. [1996]. Cette discipline est apparue au milieu des années 90 avec le développement des datawarehouse, vu l'expansion de la masse de données il est devenu nécessaire de trouver le moyen d'analyser et d'exploiter toutes ces informations stockées. La fouille de données représente la phase de recherche de connaissances dans le processus plus général de découverte de connaissances dans les données (KDD : knowledge discovery in data) Fayyad et al. [1996], qui consiste à utiliser un ensemble de techniques et d'algorithmes afin d'extraire des connaissances implicites et potentiellement utiles, pouvant servir de support au processus de décision. Son objectif principal est d'être soit prédictive, soit descriptive. Prédictive dans le sens de prédire la valeur future des variables étudiées et descriptive dans le sens de la production de modèle expliquant les données sous étude Kantardzic [2003].

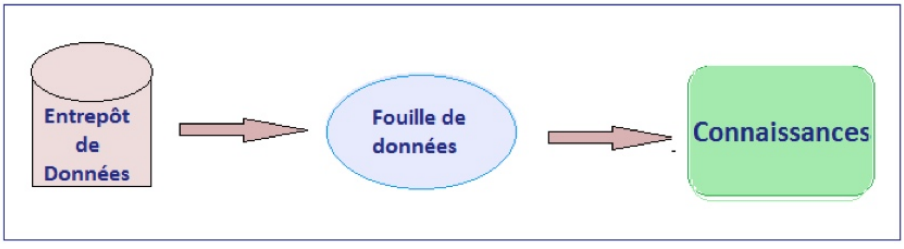

FIGURE 3.5 – *La fouille de données dans le processus d'ECD*

Le data mining diffère selon le type de données à analyser (texte, image, multimédia, web), on retrouve alors :

Texte Mining :

La fouille de données textuelles ou texte mining vise à définir des stratégies pour exploiter les textes en format libre Zigheb et Rakotomalala [2002]. Le texte à explorer se retrouve souvent sous forme de rapports, de courriers, de publications de manuel, etc.

L'image mining :

A l'instar des données textuelles, les données sous formes d'images peuvent également être traitées par des les techniques de data mining en vue d'extraire des connaissances. Celles-ci permettraient d'identifier, de reconnaitre ou de classer automatiquement des bases volumineuses d'images utilisant des fichiers d'index. A chaque image sera associée une série d'index qui donnent des indications sur son contenu. Le plus souvent cette indexation est effectuée manuellement. Pour être exploitées par des méthodes de data mining, les images doivent également subir une série de prétraitement en vue d'obtenir des tableaux numériques.

Le multimédia mining :

Le multimédia mining obéit aux mêmes principes que ceux établis pour le texte et les images. Dans le multimédia mining nous avons au moins deux objets à coder : les images et le son.

Le web mining :

Le web constitue une formidable source d'information de par son large volume, sa richesse, et sa densité. Le web mining est le processus d'exploration de données permettant de découvrir et d'extraire des informations et données pertinentes à partir des documents web. L'intérêt de fouiller dans ces données sont multiples et variés par exemple cela permet d'améliorer la puissance du moteur de recherche web en classant les documents web et en identifiant les pages web ou encore pour prédire le comportement des utilisateurs. Cependant le processus de fouille de données web se heurte à des problèmes majeurs de par le fait que la majorité des données sont non structurées. En effet, l'abondance de ces ressources, leur évolution perpétuelle et l'aspect polymorphe de leur contenu (format, média, type d'utilisation) sont autant d'obstacles qu'il faut contourner.

L'exploration web peut être divisée en quatre types de techniques d'extraction Srivastava et al. [2000] : l'exploration de contenu web, l'exploration de structures web, l'exploration de l'utilisation web et l'exploration relatives au profil de l'utilisateur. L'exploration de contenu web est l'application qui consiste à extraire des informations utiles du contenu des documents web (texte, image, audio, vidéo, etc..). Et selon le type de données extrait cette exploration peut être assimilé à du texte, image ou multimédia mining. L'exploration de structures web est l'application de découverte d'informations de structure à partir du web.La structure du graphique web se compose de pages web en tant que nœuds et d'hyperliens en tant que bords reliant les pages associées.

L'exploration de structure montre essentiellement le résumé structuré d'un site web particulier.Il identifie la relation entre les pages Web liées par des informations ou une connexion par lien direct, il est souvent assimilé au web scraping qui se base sur différentes programmes informatiques qui exploite la structure arborescente du web pour extraire les informations pertinentes, par exemple pour déterminer la connexion entre deux sites Web commerciaux, l'exploration de structures Web peut être très utile. L'exploration de l'utilisation Web est l'application qui permet d'identifier ou de découvrir des modèles d'utilisation intéressants à partir de grands ensembles de données afin de comprendre les comportements des utilisateurs et

d'anticiper leur choix et leurs besoins. L'exploration relative aux profils de l'utilisateur exploite les données fournissant des informations démographiques sur les utilisateurs du site Web, il est souvent utilisé pour identifier les communautés en ligne.

3.5 Approches de fouilles de données

Les méthodes utilisées dans la fouille de données peuvent être classifiée selon 3 critères : l'objectif de la tâche d'exploration (classification, prédiction, association, et segmentation) et le type de modèle obtenu (prédictif ou descriptif) ou encore le type d'apprentissage utilisé (supervisé et non supervisé) Atif [2015].Dans ce qui suit nous allons décrie chaque classe de méthodes.

3.5.1 Typologie des méthodes de fouille de données selon l'objectif de l'exploration

Les méthodes de fouille de données peuvent être classifier selon l'objectif de la tâche d'exploration des données, à savoir : la classification, la prédiction, l'association et la segmentation.

La classification :

C'est le fait d'examiner les caractéristiques d'un objet et lui attribuer une classe. Elle est souvent confondue avec le regroupement (dit clustering en Anglais), et est de loin l'une des tâches de fouille de données la plus utilisée car intervenant dans plusieurs domaines d'activité (Banque, Médecine,...) Larose et Larose [2014]. La classification est définit comme le processus d'apprentissage d'une fonction (ou d'un modèle) qui permet d'affecter un individu à une classe donnée parmi un ensemble de classes prédéfinies. La classification suppose qu'il existe un ensemble de données d'apprentissage (catégories, les classes), puis à générer des critères d'attribution de chaque élément présent dans les données à une des classes. Cette tâche peut être vue comme un problème d'apprentissage supervisé dans lequel le système apprend à bien classer les éléments.) Hand [1981], Weiss et Kulikowski [1991]. Les modèles de classification sont le plus souvent utilisés comme des modèles prédictifs. Parmi les algorithmes de classification on note : les arbres de décision, les réseaux de neurones, les réseaux de Bayés.

La prédiction

Appelée aussi apprentissage « supervisé ». La prédiction se déroule souvent comme la classification, car une fois la structure permettant la classification générée à partir des données, la prédiction revient à déterminer à quelle classe appartient tout nouvel élément. Ces méthodes requièrent généralement de l'utilisateur la définition d'une variable cible dont on veut par exemple prédire la valeur. Les algorithmes désignés comme supervisés fonctionnent généralement sur la base de trois(3) jeux de données Larose et Larose [2014] : ***jeu de données d'essai ou training set*** : contient l'ensemble des attributs y compris les valeurs de la variable à prédire : ces valeurs aident à la supervision du processus en mettant à nus les erreurs quand l'algorithme utilise le modèle pour prédire les résultats Hornick et al. [2006], ***jeu de données test ou test set*** : contient les différentes variables exceptées les valeurs de l'attribut à prédire. Ce jeu de données contient après le lancement de l'algorithme, les valeurs prédites de la variable cible. ***Jeu de données de validation ou validation set*** : est semblable au test set avec toutefois les vraies valeurs de l'attribut cible. Ce jeu de données sert à faire une confrontation avec les valeurs prédites de la variable cible afin d'estimer le pourcentage d'efficacité de l'algorithme utilisé.

L'association

Appelée aussi extraction de corrélations et d'associations, ou Analyse de dépendances. C'est aussi une tâche couramment répandue dans la fouille de donnée, utile pour la modélisation prédictive. Introduite en 1993 par des chercheurs en base de données d'IBM, ayant pour objectif de rechercher des conjonctions significatives d'évènement, cette approche consiste à extraire des données de dépendances, par exemple des règles d'association du type : si un individu à la caractéristique A alors il a également la caractéristique B. Les dépendances peuvent être strictes ou probabilistes Agrawal et al. [1993]. Elles décrivent les affinités des éléments de données (les éléments de données ou événements qui se produisent fréquemment ensemble).

La segmentation

La segmentation, ou clustering, est une tâche descriptive commune où on cherche à identifier un ensemble fini de catégories ou clusters pour décrire les données Jain et Dubes [1988]. Les membres de chaque cluster partagent certaines caractéristiques significatives. Les catégories peuvent être mutuellement exclusives et exhaustives, ou bien elles consistent en une représentation plus riche comme les catégories hiérarchiques ou en chevauchement. La segmentation peut être à elle seule la tâche d'un processus d'ECD. Ainsi, la détection des clusters serait l'objectif principal de l'exploration des données. Cependant, la segmentation est souvent une étape intermédiaire pour la réalisation d'autres tâches d'ECD. Dans ce cas, l'objectif de la segmentation peut être de garder une taille raisonnable des données ou de détecter des sous-ensembles de données homogènes qui sont plus simples à analyser.

3.5.2 Typologie des méthodes de fouille de données selon le modèle obtenu

Construire des modèles a toujours été une activité des statisticiens. Un modèle est un résumé global des relations entre variables, permettant de comprendre des phénomènes, et d'émettre des prévisions. « Tous les modèles sont faux, certains sont utiles » Box et Draper [1987]. L'objectif de l'exploration de données est de construire un modèle qui soit prédictif ou bien descriptif.

Modèles prédictifs

La modélisation prédictive, regroupe un ensemble de méthodes permettant de collecter et d'analyser des données définies, de manière à les interpréter pour en déduire des pronostics concernant des tendances futures, des événements à venir. Elle se base sur des données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données. Elle donne ainsi lieu à des pronostics devant toutefois être considérés comme des probabilités et non comme des prédictions certaines. La probabilité est alors envisageable en fonction de la taille de l'ensemble des données étudiée, de tel sorte que plus le nombre de données analysées est important, plus les résultats des modèles de pronostics peuvent être considérés comme des résultats envisageables et précis. Les techniques qui entrent dans cette catégorie sont la classification, la régression et l'analyse des séries chronologiques Coheris [2020].

Modèles descriptifs

Contrairement aux méthodes prédictives, les méthodes descriptives n'utilisent pas de cible. Elles fonctionnent plutôt sur la base de recherche de structures intrinsèques, des relations, ou affinités dans le jeu de données fourni en entrée. En d'autres termes, il s'agit de trouver des tendances et corrélations qui résument les relations entre données Larose et Larose [2014], Tan

et al. [2006]. Les techniques les plus connues dans les modèles descriptifs sont le clustering, l'association, segmentation.

3.5.3 Typologie des méthodes de fouille de données selon le modèle type d'apprentissage utilisé

L'apprentissage (machine learning) est la discipline qui a pour but de construire des règles d'inférence et de décision pour le traitement automatique des données Garivier [2013]. Il existe cinq type d'apprentissage automatique Ah-Pine [2020] : supervisé, non supervisé, semi supervisé, par renforcement, et actif. Nous présentons ci-dessous chaque type d'apprentissage.

Apprentissage supervisé

L'objectif de la classification supervisée est principalement de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets Labatte [2013]. Plus formellement est qu'à partir d'un ensemble d'observations $x1, \ldots =, xn$ X et de mesures $y1, \ldots, yn \in Y$, on cherche à estimer les dépendances entre l'ensemble X et Y.

On peut classifier l'apprentissage supervisé en deux familles Cornuéjols et al. [2010] : l'apprentissage supervisé symbolique, et l'apprentissage supervisé numérique. La première famille d'apprentissage sont des méthodes inspirées de l'intelligence artificielle et dont les fondements reposent beaucoup sur des modèles de logique, une représentation binaire des données (vrai/faux), et sur les méthodes de représentation des connaissances. L'apprentissage supervisé numérique regroupe les méthodes inspirées de la statistique, les données sont en général des vecteurs de réels, et les méthodes font intervenir des outils provenant des probabilités, de l'algèbre linéaire et de l'optimisation.

Il existe deux types de sous-problèmes en apprentissage supervisé numérique qui sont la régression : lorsque la valeur cible à prédire est continu, et le classement, appelée aussi classification ou catégorisation : lorsque la valeur cible à prédire est discrète. Il existe bon nombre d'algorithmes supervisés on peut citer quelques exemples comme les réseaux de neurones artificiels qui sont comme on peut le voir dans la figure 3.6 des imitations simples des fonctions d'un neurone dans le cerveau humain qui s'appuient sur des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit, et ceci dans le but de résoudre des problématiques d'apprentissage de la machine (Machine Learning Belhaouci [2020]. Le neurone reçoit donc un ensemble d'entrées : x1,..xi,..., xn. Le potentiel d'activation du neurone p est défini comme la somme pondérée (les poids sont les coefficients synaptiques wi) des entrées. La sortie o est alors calculée en fonction du seuil θ Touzet [1992].

Les arbres de décision sont un autre exemple d'algorithme supervisé, ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille Crucianu et al. [2020].

Il reste beaucoup d'autres algorithmes de classification supervisé qu'on ne peut citer de manière exhaustive tels que : K Nearest Neighbours, SVC linéaire (classificateur de vecteur de support), régression logistique, Naive Bayes, la régression linéaire, la régression vectorielle de support (SVR), les arbres de régression, etc.

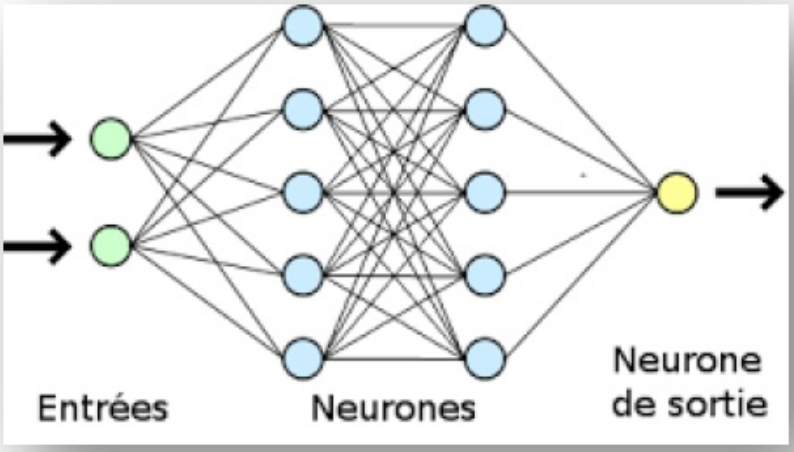

FIGURE 3.6 – *Réseaux de neurones artificiels*

Apprentissage non supervisé

Dans ce type d'apprentissage nous ne disposons pas d'un ensemble d'objets étiquetés pour prédire les nouvelles entrées ; il faut apprendre un modèle capable d'extraire les régularités présentes au sein des objets pour mieux visualiser ou appréhender la structure de l'ensemble des données. Comme seules les observations $x1, .., xn \in X$ sont disponibles, l'objectif est de décrire comment les données sont organisées et d'en extraire des sous-ensembles homogènes. L'apprentissage non supervisé comprend deux catégorie d'algorithme : regroupement appelé aussi clustering, ou segmentation, qui consiste à grouper des points de données en fonction de leurs similitudes, et l'association qui consiste à découvrir des relations entre les attributs de ces points de données (Vu dans les paragraphes précédents) Ah-Pine [2020]. Parmi les exemples d'algorithmes non supervisés on retrouve la classification hiérarchique ascendante appelée aussi clustering agglomératif dans le quel on considère tout d'abord que chaque point est un cluster. Il y a donc autant de clusters que de points. Ensuite, on cherche les deux clusters les plus proches, et on les agglomère en un seul cluster. On répète cette étape jusqu'à ce que tous les points soient regroupés en un seul grand cluster. Une autre approche descendante, appelée clustering divisif fait l'inverse puisqu'on part ici d'un grand cluster contenant tous les points, puis on le divise successivement jusqu'à obtenir autant de clusters que de points. Le K-means est aussi très l'un des algorithmes de clustering les plus répandus. Il permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données "similaires" en groupes (ou clusters) DataScientist [2019].

Apprentissage semi-supervisé

Dans ce cas on dispose d'un petit ensemble d'objets avec pour chacun une valeur cible associée et d'un plus grand ensemble d'objets sans valeur cible ; il faut tirer profit à la fois des données avec et sans valeurs cibles pour résoudre des tâches d'apprentissage supervisé ou non-supervisé. En d'autres termes parmi les observations $x1, .., xn \in X$, seulement un petit nombre d'entre elles ont un label yi. L'objectif est le même que pour l'apprentissage supervisé mais on aimerait tirer profit des observations non labellisées. Parmi les algorithmes semi supervisé on retrouve l'auto-apprentissage (self-training) Zhu [2005] qui consiste à entrainer un classifieur avec les données étiquetées (DL), ensuite le classifieur utilisé pour étiqueter les don-

nées incomplètes. Co-apprentissage est aussi une méthode semi supervisée basée sur l'idée que s'il existe 2 projections indépendantes d'un même espace de données, deux classifieurs entrainés selon ces 2 projections, doivent étiqueter de manière identique la même donnée. Il existe bien entendu beaucoup d'autres algorithmes (auto-encodeur, forêts isolées, etc.).

Apprentissage par renforcement

On dispose dans ce cas d'un ensemble de séquences de décisions (politiques ou stratégiques) dans un environnement dynamique, et pour chaque action de chaque séquence une valeur de récompense (la valeur de récompense de la séquence est alors la somme des valeurs des récompenses des actions qu'elle met en œuvre) ; il faut apprendre un modèle capable de prédire la meilleure décision à prendre étant donné un état de l'environnement Ah-Pine [2020]. Ce type d'apprentissage est inspiré des travaux de psychologie expérimentale de Thorndike (le comportement animal au niveau biologique) Thorndike [1911], ainsi lorsqu'une action (décision) prise par le réseau engendre un indice de satisfaction positif, alors la tendance du réseau à prendre cette action doit être renforcée. Autrement, la tendance à prendre cette action doit être diminuée Adam [2015].

L'apprentissage par renforcement est utilisé dans plusieurs applications : robotique, gestion de ressources, vol d'hélicoptères, etc. Parmi les algorithmes utilisés on retrouve le Q-learning (Q signifie qualité) : l'idée ici est donc d'effectuer plusieurs cycles de recherche de récompenses, de l'état initial vers un état but et de renforcer à chaque passage l'utilité/la qualité de l'action qui mène à des récompenses ou qui mène a des états menant à des récompenses Watkins et Dayan [1992].

Il existe aussi d'autre exemple d'algorithme tels que SARSA signifie État-action-récompense-État-Action (Stat Action Reward State Action) qui est une variante de l'algorithme Q-learning, ou encore l'algorithme d'apprentissage par différence temporelle qui utilise un mécanisme d'estimation temporelle pour la prédiction du temps d'arrivée d'une récompense Kung et Huang [2018].

Apprentissage actif

L'apprentissage actif est une extension de l'apprentissage semi-supervise. Plutôt qu'exploiter les données non annotées l'idée est d'enrichir la qualité des données utilisées en autorisant les interactions entre l'apprentissage et l'utilisateur qui effectue l'étiquetage des données afin qu'il donne la valeur cible d'un nouvel objet dans le but de mieux apprendre le modèle de prédiction. Ainsi Le choix des instances à étiqueter pour l'apprentissage, peut influencer considérablement la qualité du classifieur apprit.

Parmi les méthodes d'apprentissage actif on retrouve celles basées sur l'incertitude : la stratégie est de demander au classifieur l'étiquetage des instances pour lesquelles il est le moins certain (ou le plus incertain) de leur classe. Il existe d'autres méthodes basées sur la réduction de l'erreur cette stratégie vise à sélectionner les éléments qui, une fois ajoutés à la base d'apprentissage, minimisent l'erreur de généralisation Zhu [2005], Revel [2020].

3.6 Conclusion

Dans ce chapitre nous avons présenté le processus d'extraction de connaissances, nous avons vu qu'il existe différents modèles d'ECD qui passent globalement par les mêmes étapes (sélection, pré-traitement et transformation des données, data mining et évaluation), par la suite nous avons décrit une étape importante du processus ECD à savoir la fouille de données et nous avons présentés les différentes typologies d'approches existantes dans la littérature, le

choix d'une approche de fouille de données reste au bon vouloir de l'expert, des données en sa possession ou l'objectif de la tâche d'explorations.

Ce chapitre est essentiel dans l'état de l'art de ce mémoire car le processus d'extraction des connaissances est le domaine recherche sur le quel se situe principalement le travail de cette thèse où l'on vise à extraire les connaissances procédurales du web et identifier la meilleure pratique pour une requête.

L'autre facette de ce travail est la conceptualisation des meilleures pratiques, donc il s'agit ici de choisir un bon formalisme de représentation des connaissances pour faciliter leur réutilisation, et c'est en ce sens que dans le chapitre suivant nous présentons le domaine de représentation des connaissances.

Représentation des connaissances 4

Sommaire

4.1 Introduction

Pour échanger, transmettre ou traiter des connaissances, il est nécessaire de pouvoir les représenter sur un support externe à notre cerveau, sous une forme transmissible ou traitable. Pareillement échanger et partager les bonnes pratiques au sein d'une communauté de pratique revient à les conceptualiser et donc à trouver un bon formalisme de représentation des connaissances procédurales afin de permettre la réutilisation de celles-ci et favoriser l'apprentissage au sein de ces communautés.

Les systèmes de représentation des connaissances sont dans l'essence même de l'histoire de l'humanité, à commencer par les langages oraux, écrits ou picturaux permettant de représenter des objets et des idées par des mots ou des symboles et d'établir des relations entre ces objets et ces idées, autre exemple de systèmes de représentation d'usage, les représentations géométriques et les diagrammes cartésiens que nous étudions courant dans le domaine des mathématiques et des sciences.

Dans ce chapitre nous aborderons le domaine de représentation des connaissances, nous présenterons : en premier lieu les définitions de base du domaine, par la suite nous définirions les principaux formalismes de représentations de connaissances et nous aborderons les difficultés que l'on peut rencontrer lors de l'élaboration de modèles de représentation de connaissances

4.2 Définition de la connaissance

En premier lieu, il importe d'établir une distinction claire entre les concepts de données d'information et de connaissance. La donnée est une notion abstraite typée, elle peut être numériques, symboliques, textuelles, logiques, ... La donnée ne porte pas de sens en elle-même, c'est un élément brut, qui n'a pas encore été interprétée, mis en contexte. L'information est aussi une notion abstraite, mais d'un niveau d'abstraction supérieur à celui de la donnée. On définit l'information comme une donnée qui un sens, ou qui a été interprété. Tout comme la donnée et l'information, la connaissance est aussi une notion abstraite, mais d'un niveau d'abstraction supérieur à celui de l'information Malle [2017]. Par définition la connaissance est le fait de comprendre, de connaître les propriétés, les caractéristiques, les traits spécifiques de quelque chose Larousse [2020]. Elle peut être considérée comme une croyance assurée, un fait connu, ou tout simplement comme ce qu'on a appris par l'étude ou par la pratique. Dans Paquette [2002] la connaissance est considérée comme le résultat de toute construction mentale effectuée par un individu à partir d'informations ou d'autre stimuli. L'apprentissage par un individu consiste à transformer les informations qu'il obtient par différents moyens en connaissances.

Afin de sélectionner un formalisme adéquat de représentations des connaissances, il est intéressant de cerner les différents types de connaissances que l'on sera amené à représenter et de délimiter les problèmes à résoudre pour avoir une représentation adaptée. Les experts du knowledge management Nonaka et al. [1995] différencient deux formes de connaissance : la connaissance tacite et la connaissance explicite. La connaissance tacite : c'est la connaissance que possèdent les individus. Elle n'est pas formalisée et difficilement transmissible. Ce sont les compétences, les expériences, l'intuition, les secrets de métiers, les tours de main qu'un individu a acquis et échangés lors d'échanges internes et externes à l'entreprise. La connaissance explicite est la connaissance formalisée et transmissible sous forme de documents formalisés et normalisés qu'on peut réutiliser.

Dans Laurière [1988] on classifie les connaissances en éléments granulaires selon leur types, on peut citer : les perceptions immédiates du domaine à représenter tels que les objets du monde réel, les assertions et définitions sur les objets de base, les concepts ou abstractions qui permettent le regroupement ou la généralisation d'objets du domaine étudié,

les relations dîtes propriétés élémentaires des éléments de base ou des relations de cause à effet entre concepts, les théorèmes et règles, les algorithmes de résolution, les stratégies et heuristiques (connaissances empiriques reflétant les stratégies de résolution acquises par les experts humains par expérience), les procédures, les méta-connaissances ou connaissances sur la connaissance, etc.

4.3 Représentation des connaissances (Knowledge representation)

La modélisation et la représentation peuvent être assimilées ou confondues suivant chaque école de pensées : selon le dictionnaire OQLF [2002] la modélisation est une description dans un langage compréhensible par l'ordinateur de la forme, du mouvement et des caractéristiques d'un objet ou d'un ensemble d'objets qui crée un modèle et la représentation des données est une manière de structurer une donnée, selon la catégorie à laquelle elle appartient (analogique, numérique, logique, etc.) et selon une convention établie.

En résumé, la modélisation permet de décrire les connaissances, et la représentation permet de les structurer afin de les intégrer dans les systèmes informatiques. Dans ce travail, nous considérons qu'il n'existe pas une barrière stricte entre modélisation et représentation de la connaissance. Certains modèles peuvent structurer les données et certaines représentations les décrire.

Les auteurs de Barr et Feigenbaum [1981] définissent la représentation des connaissances comme une combinaison de structures de données et de procédures interprétatives qui, utilisées correctement dans un programme, conduiront à un comportement intelligent. L'objectif d'une représentation de connaissance est d'exprimer donc la connaissance sous une forme informatisée, afin qu'elle puisse être utilisée Bullinaria [2005]. Une langue de représentation des connaissances est définie par deux aspects la syntaxe d'un langage qui définit quelles configurations des composantes du langage constituent des phrases valides et la sémantique qui définit a quels faits dans le monde les phrases se rapportent.

4.4 Les principaux formalismes de représentation de connaissances

Le choix d'une bonne représentation des connaissances n'est pas aisé puisqu'aucun formalisme universel n'existe ; au contraire, plusieurs sont disponibles tels que : la logique, les systèmes de production, les ontologies, les graphes de données, etc. Dans cette section nous présenterons de façon succincte un état de l'art dur les principaux formalismes de représentations des connaissances.

4.4.1 Le langage naturel

Le langage naturel quelle que soit la langue utilisée (arabe, français, anglais..) est le moyen qui permet aux êtres humains de communiquer et représenter leurs connaissances. C'est un formalisme de représentation très expressive puisqu'on peut exprimer pratiquement tout en langage naturel mais qui reste néanmoins très difficile à utiliser par une machine tant l'ambigüité qu'il engendre, ainsi que la complexité de la syntaxe et la sémantique qui le caractérise.

4.4.2 La logique

La logique est utilisée en informatique pour modéliser de manière formelle des objets afin que l'informaticien puisse raisonner sur ces objets grâce aux modèles obtenus. La logique a été le premier formalisme de représentation des connaissances en intelligence artificielle à l'exception des langages de programmations Israel [1983]. La logique classique ou propositionnelle sert à exprimer des énoncés auxquels on attribue une valeur dite de vérité : un énoncé est soit

vrai soit faux et il n'y a pas d'autre valeur possible Paulin-Mohring [2020]. Le vocabulaire du langage de la logique propositionnelle est composé :

— D'un ensemble $V = p, q, r, ...$ dénombrable de lettres appelées variables propositionnelles. Il s'agit des propositions atomiques telles que par exemple « 6 est divisible par 2 ».
— les constantes vrai et faux.
— Un ensemble (fini) de connecteurs logiques : et (noté $\wedge$), ou (noté $\vee$), non (noté $\neg$), implique (noté $\implies$), équivalent (noté $\iff$) ; les parenthèses : $(,)$.

Le principe du raisonnement logique est la relation de conséquence logique entre les énoncés, l'idée est qu'un énoncé découle logiquement d'un autre énoncé au quel est associée une valeur de vérité vraie ou fausse. Le processus de résolution effectue en premier lieu le filtrage des données afin d'obtenir un sous-ensemble de formules méritant d'être comparée. La deuxième opération consiste à unifier les formules ainsi obtenues avec le but à résoudre par comparaison de leurs termes. Ainsi toute nouvelle formule obtenue par inférence est vraie puisqu'elle est prouvée à partir de prémisses déjà connues comme étant vraies.

La force de ce formalisme se caractérise par les bases théoriques solides, dont dispose la logique qui lui permettent ainsi de s'étendre à de nouvelles utilisations. De plus le naturel avec lequel les éléments de connaissance peuvent être exprimés sous forme de formules logiques ; permet d'exprimer un fait sans se soucier de ses manipulations Kayser [1987].

D'autres intérêts de la logique est que la cohérence et la complétude sont garanties mais cette caractéristique peut être, dans certains cas, un inconvénient car la logique est une représentation très formelle et mathématique. Elle ne permet pas, la manipulation des informations incertaines ou incomplètes.

4.4.3 Les systèmes de production

Un système de production est un programme informatique issu de l'intelligence artificielle, qui étant donné un ensemble de règles sur le comportement, infère suivant les différents états de son environnement. Il est particulièrement appropriée quand la connaissance est décomposable en une série d'actions comme c'est le cas dans les systèmes experts Richard [2004], Jouve [1992]. Dans la plupart des systèmes experts, dont le plus connu est MYCIN Shortliffe [1976], spécialisé dans le diagnostic et la prescription des infections bactériennes du sang, les connaissances sont représentées par des règles de production, et le Modus Ponens est utilisé comme mécanisme de raisonnement. Ainsi un système de production est composé de règles de production dont la syntaxe est :

SI <Condition> **ALORS** <Action>.

Par exemple : SI le feu est rouge ALORS vous devez vous arrêter. La partie gauche d'une règle de production exprime les caractéristiques d'une situation pour lesquelles il est approprié d'activer la partie droite de la règle. La condition d'une règle de production peut contenir une ou plusieurs conditions appelées les prémisses ou les antécédents de la règle. De même, la conclusion peut être multiple, c'est à dire que la règle peut avoir une ou plusieurs conséquences. Une règle de production décrit donc les actions à réaliser si l'ensemble des conditions est vérifié. Chaque règle est indépendante des autres parce qu'elle est la description d'une "réalité" élémentaire plutôt qu'une suite d'actions Jouve [1992].

Un système de production raisonne selon trois modes possibles : chainage avant, chainage arrière ou en chainage mixte qui combine les deux modes de chainages avant et arrière. En chaînage avant, si les prémisses d'une règle sont satisfaites, la règle est déclenchée engendrant ainsi de nouvelles connaissances. Ce processus est itéré jusqu'à ce qu'aucune inférence ne satisfasse un but ou qu'il y ait épuisement des inférences exécutables. Ce mécanisme permet d'arriver à une conclusion inconnue a priori alors qu'en chaînage arrière un but est fixé et le

système va tenter de le résoudre, le système va alors examiner les règles concluant sur ce but et vérifier si elles sont satisfaites. Des sous-buts peuvent apparaître lorsqu'une règle peut être unifiée avec la prémisse d'une règle concluant sur le but initial ou un autre sous-but.

Les systèmes de production présentent un avantage qui est la modularité dans la représentation des connaissances Jouve [1992] ce qui facilite la modification des éléments de connaissance. De plus avec ce formalisme, il est possible de gérer des informations incertaines, comme dans MYCIN qui permet la pondération de la valeur de vérité de chaque règle par un coefficient de certitude. Néanmoins l'utilisation de tels systèmes reste difficile du fait que la représentation sous forme de règles de production ne s'adapte pas à tous les domaines d'applications. De plus si le nombre de règles est grand, l'exécution du système devient inefficace.

4.4.4 Les réseaux sémantiques

A l'origine, les réseaux sémantiques ont été développés par Quillian Quillian [1968] à partir de travaux faits sur la modélisation en psychologie de la mémorisation associative des êtres humains. Dans son modèle, les concepts sont représentés par des noeuds et les relations entre ces concepts par des arêtes. Un réseau sémantique comprend :

- Un ensemble de nœuds représentant des objets,
- Les liens orientés ou arcs ou chaque arc représente une relation entre deux objets
- Des étiquettes : les arcs sont étiquetés en fonction des relations qu'ils représentent.
- L'interprétation sémantique du réseau dépend de chaque application concernée. Les réseaux sémantiques utilisent généralement deux relations très particulières concernant des objets de l'un des types individu ou classe :
  - **Est_un** : relation entre un individu et une classe exprimant l'appartenance ;
  - **Sorte_de** : relation entre deux classes exprimant l'inclusion.

Par exemple :

- **Est_un** (Tommy, chat) représente le fait : Tommy est un chat.
- **Sorte_de** (chat, félin) représente le fait : chat est un félin.

L'un des avantages de cette représentation est qu'un réseau sémantique autorise des déductions à l'aide d'inférences par héritage que l'on peut retrouver sans l'exemple précédent ou l'on déduit que Tommy est un félin. Ce type d'héritage permet de représenter de façon naturelle des domaines ayant des taxonomies compliquées ou une organisation complexe. Néanmoins du fait de la simplicité structurelle des nœuds, un réseau sémantique devient complexe dès qu'il contient beaucoup d'informations. Sa manipulation devient délicate et difficile à étendre ou à modifier Jouve [1992].

4.4.5 Les ontologies

Une ontologie est définit comme étant une spécification formelle d'une conceptualisation partagée Borst [1997]. **Formelle** : l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel. **Explicite** : la définition explicite des concepts utilisés et des contraintes de leurs utilisations. **Conceptualisation** : le modèle abstrait d'un phénomène du monde réel par identification des concepts clefs de ce phénomène. **Partagée** : l'ontologie n'est pas la propriété d'un individu, mais elle représente un consensus accepté par une communauté d'utilisateurs. L'ontologie est relative à un domaine, et est constituée de concepts et de relations les reliant les uns aux autres. Les connaissances sont formalisées dans les ontologies en utilisant les cinq types de composants Gruber [1993] : concepts (ou classes), relations (ou propriétés), fonctions, axiomes (ou règles) et instances (ou individus) :

- **Les concepts :** appelés aussi classe, représentent une collection ou un type d'objet.
- **Les relations :** représentent les relations entre concepts, elles peuvent être binaires, tertiaire (connectés à, sous classe de, etc.).

— **Les fonctions :** ce sont des cas particuliers de relations dans lesquelles un élément de la relation est défini à partir des autres éléments.
— **Les axiomes :** les axiomes désignent des vérités indémontrables qui doivent être admises. Ce sont des affirmations considérées comme évidentes sans preuve. Ils permettent de contraindre les valeurs de classes ou d'instances.
— **Les instances :** les instances représentent les éléments des concepts et des relations dans un domaine donné.

Il existe différentes typologies pour distinguer les ontologies existantes : selon l'objet de conceptualisation, en fonction de la granularité de l'ontologie ou encore du niveau de formalisation. Dans Gómez-Pérez et al. [2004] on distingue selon l'objet de conceptualisation six types d'ontologie :

— **Ontologie de représentation de connaissances :** elle regroupe les primitives utilisées pour formaliser les connaissances, et fournit ainsi les primitives nécessaires pour décrire les concepts des autres types d'ontologies
— **Ontologie supérieure ou de Haut niveau :** ce sont des ontologies qui décrivent des concepts générales de haute abstraction tels que : les entités, les événements, les états, les processus, les actions, le temps, etc. qui sont indépendants d'un domaine particulier.
— **Ontologie Générique :** appelée aussi méta-ontologies, véhicule des connaissances génériques moins abstraites que celles véhiculées par l'ontologie de haut niveau, mais assez générales néanmoins pour être réutilisées à travers différents domaines.
— **Ontologie du Domaine :** contient de la connaissance se rapportant à un domaine, elle décrit les concepts et leurs relations, elle se doit d'être réutilisable.
— **Ontologie de tâches :** ce type d'ontologie décrit le vocabulaire relatif à une tâche ou une activité générique d'un domaine particulier (faire un diagnostic, tâche de planification, etc.)
— **Ontologie d'application :** elle contient les connaissances requises pour une application particulière et par conséquent décrit des concepts qui dépendent à la fois d'un domaine particulier et d'une tâche particulière. Ce type d'ontologie ne peut pas être réutilisé pour d'autres applications.

D'autres auteurs dans Valéry et al. [2003] classifient les ontologies en deux catégories selon le niveau de détail utilisé lors de la conceptualisation de cette dernière.

— **Granularité fine :** elle correspond aux ontologies ayant un niveau de détails très poussé avec un vocabulaire riche pouvant décrire de manière pertinente les concepts d'un domaine ou d'une tâche ;
— **Granularité large :** concerne surtout les ontologies génériques décrivant des concepts plus larges n'ayant pas besoin d'un niveau de détails importants ;

Une autre typologie a été proposée dans Uschold et Gruninger [1996] selon le niveau de formalisation utilisé dans les ontologies :

— **Hautement informelle :** ce qui signifie qu'elle est exprimée en langue naturelle.
— **Semi-informelle :** elle est exprimée sous une forme structurée du langage naturel.
— **Rigoureusement formelle :** l'ontologie est exprimée dans un langage contenant une sémantique formelle, des théorèmes, qui permettent de vérifier les propriétés telles que la validité et la complétude.

La représentation des connaissances par les ontologies a plusieurs avantages puisqu'elles favorisent la réutilisation et le partage de données, et elles permet aussi l'inférence sur ces connaissances. Toutefois lors l'élaboration d'une ontologie il est toujours difficiles d'avoir accès aux connaissances, à la constitution du corpus, les principes d'organisation sont souvent abstraits et font appel à des notions philosophiques. De plus lors de l'élaboration de grande ontologie il est difficile de les mettre jour.

4.4.6 Les graphes de données

Les graphes sont largement utilisés dans différents domaines pour représenter les connexions de structures complexes en exprimant les relations entre ses éléments grâce aux sommets et aux arcs, nous pouvons citer divers exemple tels que le réseau de communication, réseaux routiers, réseaux sociaux, etc.

Un graphe simple ou non orienté G est définit comme une paire de deux ensembles finis G= (V,E) , tel que V =v1, v2,..., vn représentent les sommets du graphe (Vertices en anglais), et E =e1, e2,..., em représentent les arêtes (Edges en anglais) Didier [2012]. Une arête e de l'ensemble E est définie par une paire non ordonnée de sommets. On appelle ordre d'un graphe le nombre de sommets n de ce graphe. Les graphes peuvent être simple ou multi-graphe, un graphe est dit simple si toute ses arêtes relie au plus deux sommets et s'il n'y a pas de boucle sur un sommet. Contrairement aux multi-graphes qui contiennent au moins une arête qui relie un sommet à lui-même (une boucle), ou plusieurs arêtes reliant deux mêmes sommets.

Il existe plusieurs autre type de graphe non orienté tel que les graphes connexes, complets, biparti. Un graphe est connexe s'il est possible, à partir de n'importe quel sommet, de rejoindre tous les autres en suivant les arêtes. Un graphe est complet si chaque sommet du graphe est relié directement à tous les autres sommets. Un graphe est biparti si ses sommets peuvent être divisés en deux ensembles X et Y, de sorte que toutes les arêtes du graphe relient un sommet dans X à un sommet dans Y. En donnant un sens aux arêtes d'un graphe, on obtient un digraphe (ou graphe orienté). Le mot « digraphe » est la contraction de l'expression anglaise « directed graph ». Un digraphe fini G = (V,E) est défini par l'ensemble fini V = v1, v2,..., vn dont les éléments sont appelés sommets, et par l'ensemble fini E = e1, e2,..., em dont les éléments sont appelés arcs. Un arc e de l'ensemble E est défini par une paire ordonnée de sommets. Lorsque e = (u, v), on dit que l'arc e va de u à v. On dit aussi que u est l'extrémité initiale et v l'extrémité finale de e.

Soit v un sommet d'un graphe orienté. On note d+(v) le degré extérieur du sommet v, c'est-à-dire le nombre d'arcs ayant v comme extrémité initiale. On note d-(v) le degré intérieur du sommet v, c'est-à-dire le nombre d'arcs ayant v comme extrémité finale. On définit le degré : d(v) = d+(v) +d-(v).

L'utilisation des graphes de données pour représenter les connaissances est largement utilisé dans différent domaine vu ses avantages indéniables, tel que la flexibilité dans l'ajout de nouvelles relations et de nouveaux nœud, la facilité d'adaptation des structures de graphes aux environnements dynamiques à titre d'exemple nous pouvons citer les réseaux sociaux tel que Twitter ou Facebook qui s'appuient sur les graphes, ajouter à cela la facilité de calcul et d'analyse sur les graphes, par exemple on peut identifier des communauté en ligne facilement par les graphes ou encore calculer un chemin optimal pour atteindre un but précis, tout cela font les forces de représentation par les graphes de données.

Côté contrainte, les utilisateurs de bases de données graphiques volumineuses rencontrent parfois des problèmes, surtout dans les environnements distribués, où l'on doit parfois positionner les données sur plusieurs serveurs, tout en sélectionnant les informations fréquemment utilisées ensemble sur le même serveur. Ce qui peut poser des problèmes de de partitionnement et de densité. Autre limite, les bases de données graphique ne sont pas performantes pour calculer de grandes agrégations de données car dans ce cas le temps de calcul peut devenir très important Serries [2018].

4.5 Les problèmes liés à la représentation des connaissances

La représentation des connaissances doit faire face à divers problèmes et qui doivent être traités par le formalisme choisi. Il est intéressant de voir quelles sont les difficultés que l'on

peut rencontrer lors du processus de représentation. Dans cette section nous présentons un ensemble de problèmes relayés dans Jouve [1992] aux quels on peut avoir à faire lors du processus de représentation des connaissances :

4.5.1 Le traitement des exceptions

En traitant certain type de connaissances on peut faire face à des exceptions qui contredisent une loi générale, par exemple «les oiseaux peuvent voler» est une loi générale caractérisant l'élément de base oiseau, mais il existe des exceptions comme les autruches qui ne volent pas. Par conséquent, il est toujours difficile, si ce n'est impossible, de généraliser le comportement d'objets du monde réel appartenant à une même classe.

4.5.2 Évolutivité constante des connaissances

Bon nombre de connaissances sont en constante évolution, l'exemple le plus concret est le domaine des TIC technologies d'information et de communication : de nouveaux composants de réseaux apparaissent continuellement sur le marché de même que de nouvelles techniques d'implantation de réseaux. Un autre exemple qu'on peut citer : les données sur des virus ou maladie tels que la COVID que nous vivons en ces temps ci et qui ne cessent de changer, ce qui se pose le problème de la mise à jour des connaissances. Donc il est nécessaire dans ces cas de figures de choisir un formalisme qui prenne en charge cette évolutivité d'un côté en déterminant quelles sont les connaissances obsolètes, et s'il est nécessaire de les détruire donc de les perdre ou de les garder car bien qu'obsolètes elle n'en reste pas moins vraies dans un certain contexte, et d'un autre côté d'accepter de nouvelles relations et mise à jour sans pour autant refaire toute la représentation.

4.5.3 Le traitement des ambigüités

Certaines connaissances dans leur définition peuvent entraîner une certaine ambigüité comme par exemple le fait de dire «il fait trop chaud aujourd'hui» introduit une certaine ambiguïté au niveau de la signification de «trop chaud», car tout dépend du contexte dans lequel se situe la règle. Ainsi «trop chaud» peut signifier que la température supérieur à vingt cinq degrés pour les habitants d'une ville dans le climat est doux et de quarante degré pour les habitants du Sahara, ou encore de mille degrés pour un réacteur alors qu'elle est équivalente à supérieur à zéro degré pour un canon à neige. Une bonne représentation doit permettre au système de poursuivre son raisonnement en traitant ces ambiguïtés par la définition de règles ou le recours à des traitements sémantique.

4.5.4 Connaissances incomplètes, incertaines ou implicite

Certain type de connaissances peuvent être incertaine ou imprécise par exemple dans un système d'aide au diagnostique si on prend un patient qui a des symptômes comme la fièvre, la toux on peut dire que celui-ci peut être atteint du COVID mais ce n'est pas un diagnostic sur pour autant. Dons Le fait d'avoir des connaissances incertaines implique d'introduire cette incertitude quant à leur véracité dans le formalisme de représentation (généralement par l'affectation de coefficient à chaque granule de connaissance reflétant leur degré de véracité). Ainsi le système doit être capable d'inférer sur ces connaissances incertaines en fournissant en sortie des résultats incertains, pondérés par des estimations de l'incertitude.

4.5.5 Connaissances contextuelles

Certaines connaissances pour être vraies doivent être prises dans leur contexte. Généralement quand on parle de contexte on cible la variable du temps, la géo localisation, les

contraintes d'objets et sur les objets, etc. Le contexte temporel peut être vu comme une contrainte sur la durée d'un événement par exemple le temps de cuisson d'un gâteau ou encore le temps de prise de médicaments. La géo-localisation est aussi un autre facteur contextuelle à prendre en considération à titre d'exemple pour être connecté à un réseau local LAN (local area network) d'une entreprise il faut être géo-localisé dans les locaux de cette même entreprise. Par contraintes d'objets on vise l'ensemble des objets indispensables pour la réalisation d'une tâche par exemple dans réalisation du gâteau il est nécessaire d'avoir un ensemble d'ustensiles de cuisine et d'ingrédient mais il faut aussi par exemple que le four soit à une certaine température ce qui représente une contrainte sur l'objet four. Dans ce cas la lors du choix du formalisme de représentation il est nécessaire d'introduire ces variables contextuelles.

4.5.6 Contrainte de précédence entre les connaissances

Une autre contrainte à prendre en considération lors du choix du formalisme de représentation de connaissances est les relations de précédence entre certaine connaissances à formaliser. Si nous prenons comme exemple les connaissances procédurales comme lors d'une opération chirurgicale il y'a un certain séquencement dans les étapes qu'il faudra suivre afin que le processus se fasse.

4.6 Conclusion

La représentation des connaissances est tout comme le processus d'extraction de ces dernières un volet important du travail de cette thèse, ainsi nous avons tout au long de ce chapitre présenté les définitions du domaine et les principaux formalismes de représentation, nous avons par la suite explicité les différents problèmes que l'on peut rencontrer lors d'une représentation des connaissances, c'est pourquoi il est nécessaire de choisir le formalisme adéquat au type de connaissances qu'on veut modéliser, il n'existe pas de méthodologie universelle, ce choix reste subjectif aux choix de l'expert, par exemple dans notre cas on s'intéresse aux bonne pratiques qui sont considérées comme des connaissances procédurales.

L'extraction et la représentation de ce type de connaissances a fait l'objet de plusieurs travaux dans la littérature, dans le chapitre suivant nous présenterons les travaux connexes de ce domaine et nous comparons les différentes méthodologies du domaine existant.

Travaux connexes : approches d'extraction et de représentation des connaissances procédurales

5

Sommaire

5.1 Introduction

Les bonnes pratiques sont des méthodologies de travail qui ont fait leur preuve pour réaliser un objectif. Elles représentent des connaissances procédurales formées par un ensemble successives d'étapes pour atteindre l'objectif souhaité (voir chapitre 2). Extraire et formaliser ces connaissances est devenu un enjeu majeur, surtout avec l'avènement du web sémantique et social et la popularisation de sites et outils de partage. Aujourd'hui Internet est devenu une source indéniable de connaissances notamment procédurales : les internautes ne cherchent plus des informations élémentaires sur la météo mais plutôt des connaissances sur les façons de faire par exemple comment soigner une grippe, comment réparer son ordinateur, ou encore comment faire des crêpes.

Vu les avantages indéniables d'exploiter ce type de connaissance par les moteurs de recherches tel que Google ou dans les nouvelles applications modernes comme SIRI ou ALEXA, actuellement beaucoup de travaux de recherches tentent d'extraire ces connaissances et de les formaliser dans des bases dédiés pour faciliter leur exploitation par des machines, car bien que le web soit une source abondante de savoir faire, son contenu n'en reste pas moins non structuré ce qui rend la tâche d'exploration ardue.

Dans ce chapitre nous présentons un ensemble de travaux de recherches tirés de la littérature traitant du domaine de l'extraction des connaissances procédurales nous fournirons par la suite une étude comparative de ces travaux.

5.2 Travaux connexes

5.2.1 Construction d'une base de connaissance de savoir faire pour alimenter les graphes de connaissances

L'approche de Chu et al. [2017] vise à extraire des connaissances procédurales à partir de sites de partage tel que wikihow afin de créer une base de connaissances «HowtoKB» sous forme de taxonomie hiérarchique de tâche non ambiguë dans le but d'alimenter les graphes de connaissances. Ici l'extraction des connaissances a pour objectif d'identifier à partir du texte de site web les éléments tels que : le nom de la tâche, et les informations contextuelles sur cette tâche (objet participant, agent participant, lieu et heure. Par exemple pour la phrase «corriger fissure dans un mur avec du mastic» l'extraction revient à identifier la tâche :«corriger fissure» lieu :«mur» et objet participant «mastic».

Pour ce faire l'approche proposée ici passe par un pipeline d'étapes : tout d'abord le logiciel OpenIE (Open Information Extraction Systems) Etzioni et al. [2011], est utilisé pour fournir les n-uplets «sujet, prédicat, objet» qui forment une phrase lexicale. Par la suite les phrases obtenues sont normalisées et nettoyées (supprimant des mots vides). la désambigüisation des tâches synonymes est effectuée grâce à un algorithme de clustering hybride qui s'effectue en 2 phases et qui se base sur 3 différentes mesures de similarité : (1) la mesure de Wu-Palmer Wu et Palmer [1994] pour calculer la similarité entre 2 catégorie d'article du site de partage ; (2) la mesure Word2Vec Mikolov et al. [2013] pour calculer la similarité lexicale entre 2 nom de tâches ; (3) la similarité vectorielle Singhal [2001] qui calcule la similarité entre deux vecteurs de chaines qui représentent l'emplacement, l'heure, l'agent et l'objet participant. Durant la première phase du clustering (la phase ascendante) les auteurs utilisent la mesure Word2Vec pour regrouper les tâches sémantiquement proche, s'ensuit la phase descendante qui fait appel à une heuristique simple pour scinder les clusters, afin d'identifier les tâches dissemblables dans le but de remédier aux cas de faux positifs et les faux négatifs résultant de la phase1.

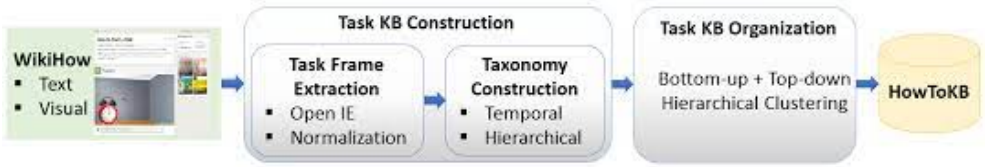

FIGURE 5.1 – *Approche de construction d'une base de connaissance HowToKB Chu et al. [2017]*

5.2.2 Extraction des relations d'un texte procédurale par une architecture de réseaux de neurone

Ce travail Park et al. [2018] propose une approche d'extraction de connaissances procédurales à partir du texte de site web de partage en utilisant l'architecture de réseaux de neurones de bout en bout afin d'apprendre sélectivement d'importantes relations spécifiques à une procédure et propose de modéliser ces connaissances sous forme de graphe où le nœud principale est le but ou l'objectif composé de méthodes, tâches et sous tâches, chaque objectif, méthode, tâche et sous-tâche peut avoir des contextes tels que l'heure, le lieu et l'acteur (voir figure 5.2). Cette architecture comprend deux modèles : encodeur hiérarchique d'attention mots : HAE (Hiearchical attention Encoder), et le classifieur de relations enrichis Memory-net : MARC (Memory-net Augmented Relation Classifier). Le modèle HAE utilise les réseaux de neurone LSTM (long short term memory) Mei et al. [2016], pour modéliser l'état des phrases d'un texte en mémorisant les informations importantes et en ignorant celles qui sont superflues, ceci passe par trois étapes importantes : la première étape consiste à encoder les mots : à ce niveau une couche LSTM bidirectionnelle est utilisée pour représenter les mots d'une phrase dans un vecteur : le LSTM avant lit la phrase du premier au dernier mot et le LSTM arrière effectue la lecture des mots dans le sens inverse, les vecteurs avant et arrière de mots obtenus sont concaténés pour résumer le sens de la phrase. La deuxième étape consiste à mettre en place un mécanisme d'attention de mot qui capture les informations sémantiques les plus importantes véhiculées dans une phrase, et les entités de mots sont fusionnées en vecteurs d'entités de phrase à chaque étape. En dernier lieu à partir du texte de site web de partage l'encodeur de phrase est utilisé pour apprendre la représentation d'une relation en utilisant les vecteurs de phrases de manière similaires.

L'autre modèle MemoryNet proposé désigne un nouveau type de représentation en mémoire pour modéliser et stocker un ensemble de phrase séquentiellement liées afin d'identifier les relations de niveaux supérieurs telles que les informations de contextes et les relations d'hiérarchie entre les taches d'une procédure. Le module d'entrée de MemoryNet calcule les vecteurs de phrases indépendamment et le stocke ou le télécharge si besoin est.

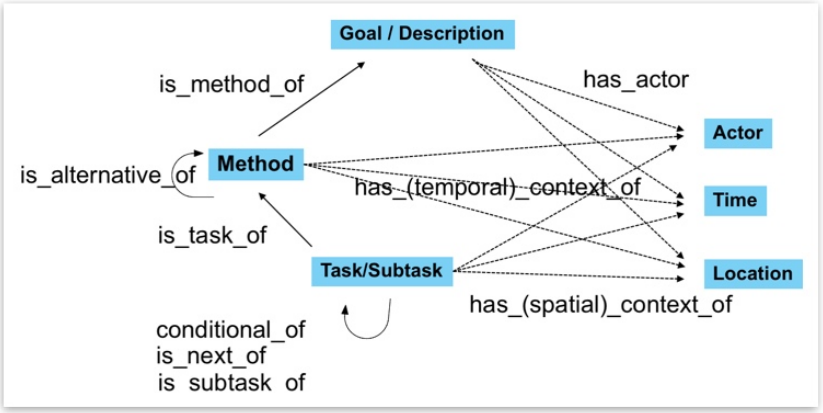

FIGURE 5.2 – *Méta modèle des connaissances procédurales Park et al. [2018]*

5.2.3 Extractions des séquences d'actions à partir du texte procédural par l'apprentissage par renforcement

Ce travail propose une approche nommée «EASDRL» (Extracting Action Sequences from text based on Deep Reinfocement Learning) permettant d'extraire à partir de texte procédurale des séquences d'actions en se basant sur l'apprentissage artificiel par renforcement Feng et al. [2018]. Plus formellement, étant donné un jeu d'entrainement le but est de chercher à apprendre deux modèles afin de prédire les noms d'actions et leurs arguments en utilisant l'apprentissage par renforcement. Pour ce faire, les auteurs considèrent les mots dans un texte procédurale associés aux opérations comme des états représentés par des vecteurs aux quels est associée une séquence d'opération (sélectionner : si le mot représente un nom d'action, éliminer : le mot n'est pas une action, null : le mot n'est pas encore traité). Ensuite, l'apprentissage en profondeur est utilisé pour extraire d'abord les noms d'actions à partir des quels on extrait les arguments en se basant un système de récompense. Enfin le modèle d'entrainement est mis en place afin de traiter de nouveaux textes pour l'extraction des noms d'actions et leurs arguments.

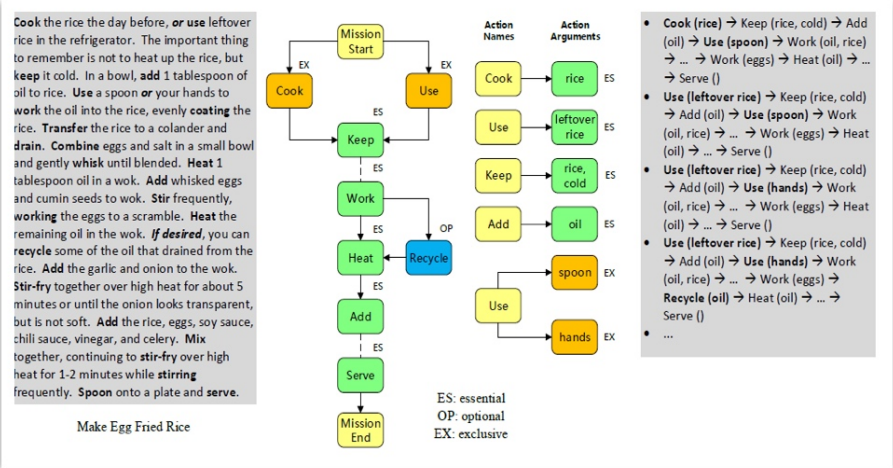

FIGURE 5.3 – *Processus d'extraction des séquences d'actions à partir du texte procédural Feng et al. [2018]*

5.2.4 Exploration de connaissances à partir de manuel de support web

Ce travail Gupta et al. [2018] vise à identifier les procédures à partir de page web représentant des manuels de procédures d'aide au diagnostique en ligne sous forme html, Les auteurs se basent sur l'idée intuitive que les procédures à extraire sont représentées par un ensemble d'instruction énumérés par des liste dans un texte. Donc le problème se résume ici à un problème de classification de liste sur les pages web de support technique. Les auteurs ont implémenté une ligne de base qui utilise une règle de classification naïve en se basant sur Slot-Grammaire McCord et al. [1992], un framework qui décrit les règles d'analyse de phrases pour de nombreuses langues (dans ce cas les puces dans le texte, les mots impératifs).

L'autre contribution de ce travail est l'extraction des blocs de décision grâce aussi à l'implémentation d'une simple ligne de base de classification où l'on identifie d'abord les points de décisions : Ce sont les points dans une procédure où les instructions bifurquent (si alors, sinon). Ainsi l'ensemble d'instructions qui suit immédiatement le point de décision est extrait comme bloc de décision. Les connaissances extraites sont alors représentées dans un graphe de données.

5.2.5 Exploration des instructions procédurales à partir du web pour la construction d'ontologie de situation

Cet article Jung et al. [2010] propose une approche pour construire automatiquement une ontologie de situation à grande échelle en exploitant des ressources du web contenant des ensembles d'instructions de savoir faire. L'approche proposée est constituée de 2 étapes principales : la première étape est l'exploration des connaissances qui extrait les paires verbes, ingrédient (emplacement, date, heure. . ..) sous forme de phrase verbale impérative, la seconde étape est la représentation des connaissances extraites par la normalisation et l'intégration des cas de situation pour former l'ontologie de situation.

L'extraction des connaissances concerne des données textuelles de sites web représentant des procédures, elle commence par le pré-traitement des données qui consiste à appliquer un ensemble de tâches afin de détecter les phrases dans les actions parsées, les auteurs font appel à l'outil Stanford paser5 Stanford [2020] pour parcourir les articles du web, tout d'abord un sujet est ajouté artificiellement aux phrases impératives pour éviter les erreurs, ensuite l'arbre d'analyse généré est converti à des dépendances typées correspondantes, et enfin les phrases et déterminants sont regroupées dans le même cluster d'arbre d'analyse. L'exploration est traitée ici par une méthode d'apprentissage non supervisé basée sur un modèle syntaxique, il s'agit d'une simple heuristique qu'un verbe d'action et ses ingrédients se présente comme un prédicat sous forme verbale. Bien que cette approche atteigne un haut niveau de précision sa couverture est limitée aux modèles construits manuellement en fonction des phrases qui ont été analysées. Dans ce cas les auteurs appliquent l'extraction d'information basée sur les Champs Aléatoires Conditionnels «CRF» Wallach [2004] pour traiter les phrases non couvertes par les règles d'association.

Pour représenter les connaissances extraites (action sous forme de verbe et informations contextuelles dites ingrédients) les auteurs utilisent une ontologie de situations qui s'appuient sur les variables de contexte afin de modéliser la nature dynamique de la vie quotidienne. L'ontologie construite ici est définit avec 6 classes : (sujet, objectif, action, objet, heure, emplacement), et 6 types de relations sémantique : has_time, has_objectif, has action, has_object, has_heure, has_emplacement). Il est nécessaire toutefois de faire un ensemble traitements de normalisation afin de regrouper les actions et objectifs similaires. Pour la normalisation d'action une méthode de clustering agglomerative gourmande «le Clustering de Resolver» Yates et Etzioni [2007] est utlisée. Ainsi les similitudes entre les actions sont calculées, et l'action la plus dominante dans le cluster par la fréquence est choisie. Ensuite les auteurs procèdent à la normalisation des objectifs car que les articles extraits peuvent avoir le même objectif et des étapes différentes, ainsi pour une requête lancée le titre d'un article est utilisé pour trouver les objectifs de recherche d'autres articles similaire.

5.2.6 Approche d'extraction et représentation des connaissances techniques pour l'amélioration de l'efficacité de réponses aux questions technique

Ce travail Yang et al. [2017] tente d'améliorer la pertinence des réponses automatiques aux questions techniques en exploitant les graphes de connaissances. Contrairement aux méthodes traditionnelles de recherche d'informations qui se concentrent uniquement sur les mots clés apparaissant dans une question, ici le but est de comprendre les intentions des requêtes des utilisateurs. Pour cela les auteurs extraient des connaissances sur les savoirs faire techniques et les représentent dans un graphe de connaissance. Le graphe de connaissances construit est basé sur un corpus technique sous forme d'hiérarchie à quatre niveaux correspondants au niveau d'une question technique : catégorie (groupe de produits ayant des fonctions similaires), produit (produit et attribut de produit), composant (erreur appartenant généralement à un composant d'un produit) et événement (phénomènes d'erreur».). L'extraction des catégories, des produits et des informations sur les produits se fait manuellement à partir de support

techniques. Par contre les nœuds au niveau des composants sont extraits automatiquement à partir du corpus technique en utilisant la méthode d'étiquetage séquentiel Lafferty et al. [2001], qui associe chaque mot d'une phrase à une étiquette grammaticale (ou tag). L'extraction des relations entre les nœuds des produits et des catégories (Component_Of) ainsi qu'au niveau des événements (EventWordOf et RelatedTo) est basée sur la cooccurrence de composants et de produits dans le corpus technique (respectivement la cooccurrence d'événement et de composants ou de produits), une mesure probabiliste courante de la force d'association entre deux termes nommée information mutuelle ponctuelle (PonctuelMutuel Information PMI) Manning et Schütze [1999] est alors utilisé pour identifier les relations entre les nœuds du graphe.

5.2.7 Extraction automatique des connaissances pour l'élaboration d'applications web

Ce travail Noura et al. [2019] propose une approche baptisée KE4WoT (Knowledge Extraction for the Web of Things) qui s'appui sur les techniques d'apprentissages automatiques non supervisées dans le but d'extraire des connaissances à partir de données déjà structurées afin d'identifier automatiquement les sujets (concepts et propriétés) les plus importants à partir d'ontologies existantes dans les applications Internet des objets (IdO).

La méthodologie d'extraction utilisée suit le processus d'ECD vu dans le chapitre 3 et englobe un pipeline d'étapes dont : la sélection d'ontologies sur les quelles s'effectuera la tâche d'exploration, le prétraitement impliquant le nettoyage et la normalisation des concepts des ontologies, l'extraction de vocabulaire : la chaîne d'outils d'analyse interroge les termes d'ontologie tels que les classes, les sous-classes, les propriétés, les étiquettes et les concepts, etc. s'en résulte est une liste de vocabulaires uniques, l'étape suivante concerne l'extraction de termes : qui revient à tokeniser le texte en mot unique, et à supprimer des caractères spéciaux et les mots vides, etc. Par la suite on calcule la fréquence de chaque terme dans toutes les ontologies afin de l'utiliser comme métrique pour identifier les sujets les plus importants. Les auteurs utilisent le modèle d'apprentissage Word2vec Mikolov et al. [2013] qui identifie l'association d'un mot avec d'autres mots, et ont formé ainsi un corpus à partir de publications scientifiques, enfin l'algorithme de clustering K-means MacQueen [1967] est appliqué en exploitant word2vec afin d'identifier les sujets populaire dans les clusters.

5.2.8 Interprétation non supervisée d'instructions pédagogiques

Ce travail Kiddon et al. [2015] présente une approche non supervisée basée sur les modèles probabilistes pour extraire des connaissances à partir d'un texte procédurale (recette de cuisine) dans le but de les organiser sous forme d'un graphe d'action, définissant quelles actions doivent être effectuées sur quels objets et dans quel ordre. Plus précisément, étant donné un texte procédural l'objectif est d'identifier les segments de texte qui décrivent des actions individuelles afin de construire un graphe d'instructions orienté où les sommets représentent les verbes ou les arguments identifiés dans les phrases du texte et les arêtes sont les connections de chaque argument à un verbe.

Donc l'extraction de connaissances ici revient à identifier en premier lieu les paires verbes et arguments dans le texte à travers la segmentation de ce dernier et ensuite à définir les connections entres les arguments et les verbes identifiés. Pour ce faire les auteurs appliquent une méthode d'apprentissage non supervisé sur deux modèles conçu pour apprendre les aspects des connaissances procédurales : un modèle de segmentation pour extraire les verbes et les arguments du texte, qui parmi toutes les segmentations possibles choisit celle avec la probabilité la plus élevée, et un modèle de connexion probabiliste qui définit une distribution sur les connexions entre les actions extraites, cette probabilité s'appui sur une probabilité a priori sur les connexions antérieures C nommée P (C) et la probabilité de voir une recette

segmentée R étant donné un ensemble de connexions C nommée P (R | C). L'ensemble de connexions le plus probable maximisera la probabilité conjointe : P (R | C) P (C).

5.2.9 Extraction et représentation de la connaissance dans les scripts

Ce travail Regneri et al. [2010] tend à extraire à partir d'un scénario les phrases pouvant décrire le même événement dans un script, et de modéliser les contraintes sur l'ordre temporel dans lequel ces événements se produisent. Pour se faire les auteurs utilisent un algorithme de la bio-informatique qui calcule des alignements de séquences multiples ASM utilisé généralement pour trouver des éléments correspondants dans des protéines ou de l'ADN Durbin et al. [1998], et l'adaptent à des descriptions en langage naturel des séquences d'événements (DSE) spécifiques au script recueillies auprès de bénévoles afin d'identifier les phrases de différentes DSE qui décrivent le même événement.

De cet alignement, on extrait le graphe orienté de script temporel où les nœuds représentent les événements d'un scénario s, et l'ensemble d'arêtes (ei, ek) indiquent que l'événement ei se produit généralement avant ek dans s. Ce graphe initial représente exactement les mêmes informations que le ASM, dans une notation différente. Dans une seconde étape le graphe est affiné pour éliminer le bruit à l'origine des erreurs ASM en identifiant les nœuds parasites qui ne contiennent qu'une seule description d'événement et en fusionnant les nœuds dont les descriptions d'événements seraient suffisamment cohérentes selon le calcul de mesure de similarité basée sur une simple heuristique de style dépendance peu profondes qui effectue le balisage des mots dans les phrases décrivant les séquences d'événements grâce à WordNet : le premier verbe potentiel de la phrase est identifié comme prédicat, le nom précédent comme sujet et tous les noms potentiels suivants comme objets, la mesure de similarité est donc la somme des valeurs de similarité pour les prédicats, les sujets et les objets respectivement.

5.3 Synthèse

Afin de comparer les différents travaux énumérés nous présentons le tableau ci-dessous qui reprend les principaux points relatifs à chaque approche à savoir :

- **Le but d l'extraction** : ou de la tâche d'exploration qui résume la problématique traitée dans l'article
- **Recueil, type et pré-traitement des données** : ce point spécifie la méthode de recueil de données, typologie des données traitées (connaissances procédurales quelles soient textuelles, multimédia, etc.) et l'ensemble des tâches de nettoyage qui ont été utilisées dans le cas ou cela est spécifié.
- **Algorithme de Data Mining** : ce point est l'un des plus important puisqu'il illustre la méthodologie d'exploration de données utilisée.
- **Formalisme de Représentation des connaissances** : ce point présente le type de formalisme utilisé pour représenter les connaissances extraites.

Approche étudiée	But de l'extraction	Recueil, type et Prétraitement des données	Algorithme de Data Mining	Formalisme de Représentation des connaissance
Chu et al. [2017]	Création d'une base de connaissances pour alimenter les graphes de connaissances	Données textuelles tirées de sites web par un outil automatique OpenIE : Open Information Extraction Systems pour obtenir les n-uplets « sujet, prédicat, objet » dans une phrase d'un texte donnée. Le prétraitement des données comprend la normalisation et l'élimination des mots vides	Clustering hybride en 2 phases basé sur 3 différentes mesures de similarité : (1) la mesure de Wu-Palmer ; (2) la mesure Word2Vec ; (3) la similarité vectorielle, pour regrouper les tâches synonymes et les désambigüiser	Taxonomie hiérarchique de tâche de 5 éléments : tâche, objet participant, agent participant, lieu et heure
Park et al. [2018]	Modélisation des connaissances procédurales sous forme de graphe	Données textuelle du web disponible en ligne à partir du site github	Les réseaux de neurones de bout en bout basés sur 2 modèles :le modèle HAE s'appuient sur les réseaux de neurone LSTM (long short term memory), pour modéliser l'état des phrases d'un texte, et ensuite met en place un mécanisme d'attention de mot qui capture les informations sémantiques les plus importantes véhiculées dans une phrase et le modèle MemoryNet calcule en entrée les vecteurs de phrases indépendamment et le stocke ou le télécharge si besoin est	Méta-modèle décrivant une procédure sous forme de graphe orienté ou l'objectif à atteindre est le nœud principale, les nœuds secondaires : méthode, tache et sous tache, acteur, temps, localisation

Approche étudiée	But de l'extraction	Recueil, type et Prétraitement des données	Algorithme de Data Mining	Formalisme de Représentation des connaissance
Feng et al. [2018]	Identification des séquences d'actions à partir de texte procédurale afin de remédier au manquement des connaissances procédurale pour les applications connectées	Données textuelles de sites de partage disponibles en ligne	Apprentissage artificiel par renforcement : les états (mots associés aux actions) sont représentés par des vecteurs aux quels est associée une séquence d'opération (sélectionner : si le mot représente un nom d'action, éliminer : le mot n'est pas une action, null : le mot n'est pas encore traité). on utilise l'apprentissage en profondeur pour extraire d'abord les noms d'actions à partir du quel on peut procéder à l'extraction d'argument. on met en place le modèle d'entrainement pour traiter de nouveaux textes afin d'extraire les noms d'actions et leurs arguments	Non abordé
Gupta et al. [2018]	Identification des procédures d'aide au diagnostic	Page web disponibles en ligne représentant des manuels de procédures d'aide au diagnostique en ligne sous forme html	Implémentation d'une ligne de base qui utilise une règle de classification naïve basée sur Slot Grammaire pour extraire les instructions et les points de décisions	Graphe de données
Jung et al. [2010]	Construction automatique d'une ontologie de situation à grande échelle	Données textuelles de sites web recueillies par une méthode automatique basée sur l'outil Stanford parser	Méthode d'apprentissage non supervisé basée sur un modèle syntaxique, et les Champs Aléatoires Conditionnels «CRF» pour extraire les actions sous forme de verbe	Ontologie de situation

Approche étudiée	But de l'extraction	Recueil, type et Prétraitement des données	Algorithme de Data Mining	Formalisme de Représentation des connaissance
Yang et al. [2017]	améliorer la pertinence des réponses automatique aux questions techniques	Texte procédurale tirés à partir de support technique manuellement	Une partie de l'extraction des connaissances modélisées dans le graphe se fait manuellement et l'autre automatiquement par la méthode d'étiquetage séquentiel qui associe chaque mot d'une phrase à un tag L'extraction des relations entre les nœuds est basée sur une mesure probabiliste courante de la force d'association entre deux termes nommée information mutuelle ponctuelle	Graphe de connaissances sous forme d'hiérarchie à quatre niveaux
Noura et al. [2019]	Identification automatiquement les sujets les plus importants à partir d'ontologies existantes dans les applications Internet des objets (IdO).	Ontologie disponible sur le web sur les quelles on effectue la normalisation des concepts des ontologies, tokenisation des termes	Algorithme de clustering non supervisé K-means basé sur la mesure de similarité word2vec afin d'identifier les sujets populaires dans les clusters	Non abordé

Approche étudiée	But de l'extraction	Recueil, type et Prétraitement des données	Algorithme de Data Mining	Formalisme de Représentation des connaissance
Kiddon et al. [2015]	Mise en place d'un graphe d'action	Texte procédurale représentant des recettes de cuisine disponible en ligne	Algorithme d'apprentissage non supervisé basé sur un modèle de segmentation pour extraire les verbes et les arguments du texte, qui parmi toutes les segmentations possibles choisit celle avec la probabilité la plus élevée, et un modèle de connexion probabiliste qui définit une distribution sur les connexions entre les actions extraites	Graphe orienté
Regneri et al. [2010]	Identification de phrase dans un script décrivant le même événement	Texte sous forme de scénario dans un script recueilli auprès de bénévoles	Alignements de séquences multiples ASM pour extraire les événements dans les nœuds du graphe Mesure de similarité basée du Wordnet pour fusionner les nœuds similaires	Graphe orienté

TABLE 5.1 – *Récapitulatif des approches étudiées*

Nous pouvons déduire à partir du tableau ci-dessus que la majorité des travaux étudiés traitent des donnés sous format procédurale c.à.d. qui décrivent une façon de faire ou une méthodologie que ce soit une recette de cuisine un scénario rassemblant un ensemble d'étape ou des séquences d'actions, etc, de plus ces données sont principalement du texte et en majorité tirées du web ce qui confirme que premièrement le web est la source d'information la plus importante mais néanmoins son exploitation nécessite des processus d'extraction pour faciliter sa réutilisation et aussi l'importance de l'exploration de données procédurales dans divers domaine. Nous observons aussi que seuls les travaux de Chu et al. [2017] et Jung et al. [2010] ont procédé au recueil des données de manière automatique grâce à des outils disponible qui permettent de le faire, sans pour autant avoir recours à la programmation.

Une autre remarque est que la plupart des travaux n'abordent pas la phase de pré-traitement des données qui est une étape importante du processus d'ECD car c'est elle qui définit la qualité des connaissances extraites, mise à part dans Chu et al. [2017] et dans Noura et al. [2019] où l'on aborde certaines étapes du processus de nettoyage des données tel que la suppression des mots vide la tokenisation des termes,.etc. Donc cette phase de l'ECD reste au bon vouloir des concepteurs et tout dépend de la source, du type et du volume des données et l'étape d'exploration qui suit.

Dans la phase de Data mining qui est la plus importante, diverse approche sont utilisés tel que l'apprentissage (supervisé, non supervisé) les méthodes probabilistes, l'alignements de séquences multiples etc. afin d'atteindre l'objectif de l'exploration. Dans les travaux de Chu et al. [2017], Park et al. [2018], et Yang et al. [2017], Jung et al. [2010] on cherche à construire respectivement des taxonomies ou encore des métas-modèles, des graphes de connaissances et des ontologies pour formaliser les connaissances procédurales, d'autres tel que Noura et al. [2019], Feng et al. [2018]. Gupta et al. [2018] vise à identifier les séquences d'actions ou les étapes dans une procédure ou même les phrases décrivant un événement (Regneri et al. [2010]).

L'autre facette de l'exploration des données est que dans la majorité des cas on se retrouve avec des données similaires ou ambigües donc la plupart des tâches d'exploration utilisent différentes techniques pour remédier à ce problème tel que Regneri et al. [2010] qui utilise la mesure de similarité basée du Wordnet pour fusionner les nœuds similaires, ou encore dans Chu et al. [2017] qui tentent de fusionner les tâches et les catégories similaires en se basant sur différentes mesures de similarité.

Il est à conclure dans cette analyse que la phase de l'exploration est l'aboutissement de chacun des travaux étudiés elle peut avoir comme but final la représentation de connaissances dans un formalisme facilitant sa réutilisation ou simplement la création de technique pouvant identifier des points important d'une méthode ou pratique. Aussi notons que le formalisme pour modéliser les données utilisé par la plupart des approches (mise à part dans Jung et al. [2010] où on tente de modéliser l'aspect dynamique des connaissances par une ontologie situation).

5.4 Conclusion

Ce chapitre marque la fin de la partie théorique de ce mémoire, donc il nous a semblé intéressant de voir en dernier point de l'état de l'art les différents travaux connexes à notre problématique à savoir l'extraction et la représentation des connaissances procédurales qui ne sont autre que des bonnes pratiques appelées différemment mais qui véhiculent un même principe i.e. un processus regroupant un ensemble d'étapes pour atteindre un objectif.

A la fin de ce chapitre nous avons présenté une synthèse des travaux étudiés afin d'établir une étude comparatives de ces derniers en faisant ressortir les point les plus important des deux domaines qu'ils impliquent à savoir : l'extraction et la représentation des connaissances. Nous avons pu voir que tous les approches étudiées suivent les principales étapes du processus d'ECD à noter : le recueil des données, leur pré-traitement, la fouille de don-

nées. Nous n'avons pas spécifié l'étape de validation et de vérification pour les approches étudiés car ceci reste un choix particulier pour chaque tâche d'exploration. Pour ce qui est du côté représentation des connaissances nous avons constaté que c'est un aspect primordial du processus d'ECD car il peut véhiculer en lui-même l'objectif à atteindre pour chaque travail d'exploration et c'est pour cette raison que nous lui avons consacré tout un chapitre dans ce mémoire.

Le chapitre suivant entamera la seconde partie de notre travail à savoir la partie contribution où nous proposerons une approche pour conceptualiser les bonnes pratiques au sein des communautés de pratique et extraire meilleure pratique pour une requête donnée.

6 Extraction des meilleures pratiques au sein d'une communauté de pratique par l'apprentissage artificiel sur les graphes

Sommaire

6.1 Introduction

Le web est devenu ces dernières années une source constante et permanente de connaissances évolutives où des utilisateurs répartis géographiquement ne cessent de partager leur savoir faire, leur expérience acquise sous forme de connaissances procédurales dans différents domaines d'expertise. Sans le savoir les internautes ont créé des communautés de pratique sur la toile ou les uns deviennent apprenants des savoir faire des autres dans des domaines précis, ceci grâce a différente plate forme collaborative de partage tél que Facebook, Youtube, Wikihown, etc.

L'avantage principal du processus d'apprentissage via internet est qu'il est gratuit et ne nécessite pas une proximité. Néanmoins, la réutilisation de ces données et l'inférence dessus par des machines reste un défi majeur vu la diversité des formats de représentation. Dans ce contexte nous proposons dans ce chapitre une nouvelle approche pour découvrir les connaissances procédurales cachées à partir d'un data set de données sur le web, et les représenter dans une base de connaissances. L'extraction de données est un processus qui implique la récupération de données provenant de différentes sources. Souvent, les entreprises extraient des données afin de les traiter plus avant de les migrer vers un référentiel de données (tel qu'un entrepôt de données) ou de les analyser plus en profondeur, dans notre cas l'approche que nous proposons pousse l'analyse des données extraites à identifier la meilleure pratique parmi toutes les bonnes pratiques pour une requête spécifique.

Ce chapitre introduit en premier lieu le contexte et la problématique traitée dans ce mémoire, par la suite on passera les points majeurs de notre modeste contribution, on présentera en troisième point l'approche que nous proposons et nous l'expliciterons en dernier lieu par un exemple applicatif.

6.2 Contexte et Problématique

La problématique abordée dans ce mémoire concerne l'extraction et la représentation formelle du savoir-faire humain à partir de CdP en ligne pour faciliter sa réutilisation et l'extraction par la même occasion de la meilleure pratique pour une requête lancée afin d'assister l'utilisateur dans son processus de recherche, ceci présente un défi majeur : car d'une part la connaissance en elle-même peut être vague, ambigüe et incomplète ; les données véhiculées sont fournies par des humains non expert en information, aussi elles ne sont pas structurées ce qui rend difficile leur représentation et l'inférence dessus. D'autre part devant l'abondance d'information dans les sites web de CdP en ligne, l'utilisateur ne s'y retrouve plus, par exemple si on cherche sur le web une simple recette de pain on se retrouve face à un dilemme qui est de choisir parmi la multitude de résultat retourné la meilleure recette. De plus comme nous l'avions spécifié dans le chapitre 2 il n'existe pas dans la littérature concrètement de fondement pour évaluer les bonnes pratiques. On peut toutefois se baser sur l'un des 3 systèmes que nous avions proposé (chapitre 2) afin de comparer les bonnes pratiques à savoir en premier lieu ***le système de notation et d'évaluation des sites web :*** dans ce cas on se réfère au système d'évaluation qu'offre certain site web qui se traduit par exemple par le nombre d'étoiles obtenues ou on peut aussi lire les commentaires des utilisateurs mais leurs avis restent subjectifs a leur contexte et il faudra un temps précieux pour analyser les différents avis, le second ***système de comparaison de bonnes pratiques est basé sur un critère d'optimisation :*** par exemple on peut chercher la meilleure façon d'accomplir une tâche dans un temps précis, ou encore dans un espace limité en utilisant une liste finie d'objet, dans ce cas le système d'optimisation peut se baser sur les méthodes de recherche opérationnelle afin de comparer les bonnes pratiques. Ces situations là restent à exploiter dans un contexte bien précis relatif aux choix et conviction de l'utilisateur et peuvent ne pas refléter de manière fiable et globale la supériorité d'une pratique par rapport à une autre. Le troisième

système de comparaison est par étapes : dans ce cas là comparer donc les bonnes pratiques entre elle revient à comparer leur complétude par rapport aux étapes qu'elles empruntent, c.à.d. qu'une meilleure pratique serait une procédure utilisant en son sein un ensemble d'étapes communes à d'autres bonnes pratiques pour atteindre un objectif commun. En d'autres termes, la meilleure pratique est une bonne pratique regroupant l'ensemble des étapes les plus utilisées par d'autres bonnes pratiques pour atteindre le même objectif. Et c'est sur la base de cette hypothèse que nous proposons une approche spécifique qui permet la conceptualisation des bonnes pratiques (savoir-faire) du web et l'extraction des meilleures pratiques pour une requête donnée.

6.3 Etude Comparative

L'approche que nous proposons dans ce mémoire a pour objectif la conceptualisation des meilleures pratiques au sein d'une communauté de pratique, pour notre part nous visons les communautés de pratique en ligne car comme souligné dans les travaux étudiés précédemment le web offre une base de connaissance très riche à exploiter, et d'autre part les bonnes pratiques nommées aussi connaissances procédurales sont devenues l'essence même des applications modernes, robotique et des moteurs de recherches actuels. L'exploration de ces connaissances peut avoir différentes raisons, notre objectif à nous diffère des approches relatées dans le chapitre 5 car nous poussons la tâche d'exploration plus loin que la création d'une base de connaissances comme dans Chu et al. [2017] ou la modélisation des connaissances par un formalisme de représentation comme dans Park et al. [2018], dans notre cas outre l'extraction et le recueil des bonnes pratiques dans une base de connaissance graphiques, nous proposons une approche basée sur les techniques d'apprentissage artificiel pour traiter et analyser, automatiquement les pratiques disponibles et assister un utilisateur dans son processus de recherche de la meilleure pratique. Une autre distinction à faire concerne le type des données traitées, dans Kiddon et al. [2015] ces données prennent la forme de recette de cuisine, dans Yang et al. [2017] on parle de manuels de procédure, etc. le point commun entre toutes ces données c'est qu'elles représentent des connaissances procédurales qui regroupent un ensemble d'étapes successives pour atteindre un objectif quel qu'il soit, pour notre part nous définissons les connaissances procédurales partagées dans des sites du web comme des bonnes pratiques issues de communautés de pratiques en ligne dans le but de favoriser l'apprentissage entre les membre de la communauté.

Le recueil des données est le premier pas dans tout processus d'ECD, la plupart des travaux comme Park et al. [2018], Feng et al. [2018], Gupta et al. [2018], Kiddon et al. [2015] utilisent des données libres et disponible sur des sites ou recueillis manuellement. D'autre mèthodes comme dans Jung et al. [2010] ou encore dans Chu et al. [2017] choisissent des outils automatiques pour extraire des données tels que respectivement : l'outil Stanford parser Stanford [2020] qui parcoure les pages web et génère des arbres de dépendances et l'outil OpenIE Etzioni et al. [2011] : Open Information Extraction Systems qui permet obtenir les n-uplets « sujet, prédicat, objet » dans une phrase d'un texte donnée. Ces outils facilitent les concepteurs dans leur recueil d'informations, mais trouvent rapidement leur limite, les développeurs font souvent face à de fortes limitations en terme d'usage ou même en terme de fonctionnalités. Pour notre part nous proposons d'alimenter un algorithme de web scraping pour extraire les bonnes pratiques à partir de site web, ces lignes de codes implémentés parcourent l'arborescence des pages web et extraient les informations qui nous sont nécessaires afin des les représenter dans un graphe de bonnes pratiques. Dans Yang et al. [2017] l'extraction des connaissances se fait manuellement et automatiquement par la méthode d'étiquetage séquentiel qui associe chaque mot d'une phrase à un tag modélisées dans le graphe, l'extraction des relations entre les nœuds est basée sur une mesure probabiliste courante de la force d'association entre deux termes nommée information mutuelle ponctuelle, Dans notre cas il

serait laborieux d'extraire une partie des connaissances manuellement, les nœuds de notre graphes représentent les bonnes pratiques ainsi que les étapes qui les forment et les relations entre les nœuds du graphe témoignent des relations de précédences entre les étapes, ceci dans l'ordre temporel des choses et dans le quelle elle sont citées dans le texte par exemple en lisant une recette on voit que l'on ne peut mettre un gâteau au four avant d'avoir mélanger les ingrédients.

Le formalisme de représentation aussi diffère d'une approche à une autre, dans notre approche nous avons opté pour les graphes de données car leurs avantages sont indéniables en terme de flexibilité, facilité de mise à jour, facilité de calcul, etc. Contrairement aux autres travaux tel que Jung et al. [2010] qui ont créer une ontologie de situation afin de modéliser l'aspect dynamique des connaissances procédurales mais l'inférence sur ce type d'ontologie semble coûteuse en temps et en espace mémoire, un autre exemple l'approche proposée dans Chu et al. [2017] vise à créer une taxonomie de 5 éléments : tâche, objet participant, agent participant, lieu et heure, ou encore l'approche de Park et al. [2018] qui élabore un méta-modèle décrivant une procédure sous forme de graphe orienté ou l'objectif à atteindre est le nœud principale, et les nœuds secondaires représentent : méthode, tache et sous tache, acteur, temps, localisation, ce qui semble se rapprocher de notre représentation, sauf que le modèle que nous proposons concerne uniquement les bonnes pratiques et les étapes qu'elles englobent indépendamment des contraintes temporelles ou d'objets car notre objectif majeur n'est pas seulement la création d'une base de connaissances mais l'extraction de la meilleure pratique en se basant uniquement sur les étapes qu'elle parcoure.

Différente approches sont utilisées dans la tâche d'exploration selon l'objectif de l'ECD, il y'a des approches qui cherchent à extraire ou identifier des entités dans du texte comme Feng et al. [2018] qui s'appuie l'apprentissage artificiel par renforcement pour extraire les noms d'actions et leurs arguments, ou encore Gupta et al. [2018] qui implémente une ligne de base qui utilise une règle de classification naïve basée sur Slot-Grammaire pour extraire les instructions et les points de décisions. Un autre exemple dans Jung et al. [2010] on propose une méthode d'apprentissage non supervisé basée sur un modèle syntaxique, et les Champs Aléatoires Conditionnels «CRF» pour extraire les actions sous forme de verbe, ainsi que dans Kiddon et al. [2015] où les auteurs utilisent un algorithme d'apprentissage non supervisé basé sur un modèle de segmentation pour extraire les verbes et les arguments du texte. L'approche proposée dans Park et al. [2018] utilise les réseaux de neurones de bout en bout basés sur 2 modèles : le modèle HAE s'appuient sur les réseaux de neurone LSTM (long short term memory), pour modéliser l'état des phrases d'un texte, et ensuite met en place un mécanisme d'attention de mot qui capture les informations sémantiques les plus importantes véhiculées dans une phrase et le modèle MemoryNet calcule en entrée les vecteurs de phrases indépendamment et le stocke ou le télécharge si besoin est. Toutes ces méthodes relèvent du domaine traitement du langage naturel (TAL) et leur finalité est l'analyse syntaxique du texte afin d'identifier les actions dans les bonnes pratiques, le but que nous poursuivons est autre puisque nous extrayons et modélisons les bonnes pratiques directement grâces aux données semi structurées du web en explorant l'arborescence des pages web.

Noura et al. [2019] utilise l'algorithme de clustering non supervisé K-means basé sur la mesure de similarité Word2Vec Mikolov et al. [2013] afin d'identifier les sujets populaires dans les clusters, la tâche d'exploration que nous suivons ne se concentre pas sur la popularité d'un sujet mais plutôt de la supériorité des pratiques, nous faisons appel aussi aux méthodes d'apprentissage non supervisé en se basant sur la mesure de similarité sémantique Word2Vec de Mikolov mais dans le but d'assister l'utilisateur dans sa tâche de recherche de la meilleure pratique afin de regrouper les méthodes similaires à l'objectif recherché par l'utilisateur. On effectue ensuite le regroupement des étapes similaires, ceci est le cas aussi dans l'approche de Chu et al. [2017] qui utilise un clustering hybride en 2 phases basé sur 3 différentes mesures de similarité : la mesure de Wu-Palmer ; la mesure Word2Vec ; et la similarité vectorielle, pour

regrouper les tâches synonymes et les désambigüiser, notre technique ne nécessite pas de phase ascendante et descendante lors du clustering des tâches et se fait en une seule phase par l'algorithme non supervisé DBScan Ester et al. [1996] en se basant sur Word2vec ce qui est mois coûteux temps, de plus dans Chu et al. [2017] le regroupement se fait lors de la conception de la base connaissances, dans notre tâche d'exploration on ne regroupe pas les pratiques similaires entre elles dans la base de connaissances car quelque soit le type de clustering utilisé il engendrera une perte d'information, et c'est pourquoi le regroupement est fait lors du processus de recherche de la meilleure pratique.

Dans Regneri et al. [2010] les alignements de séquences multiples ASM sont utilisés pour extraire les événements dans les nœuds du graphe ensuite la mesure de similarité basée du Wordnet est appliquée pour fusionner les nœuds similaires, sauf que fusionner 2 ou plusieurs nœuds revient à choisir parmi toutes les étapes regroupées celle qui nommera le nouveau nœud dans le graphe et là encore notre approche diffère puisque nous utilisons les techniques de résumé de texte dit sumurization of text en anglais où on fait appel à l'algorithme de classement de PageRank Brin et Page [1998] afin de classer les étapes similaires et choisir ainsi celle qui aura le score le plus élevée afin de représenter le nouveau nœud dans le graphe des bonnes pratiques.

6.4 Approche proposée

Dans cette section nous proposons une nouvelle méthode pour conceptualiser les connaissances procédurales du web et extraire les meilleures pratiques au sein d'une communauté de pratique en ligne par l'apprentissage artificiel sur les graphes. L'approche comme illustrée dans la figure ci-dessous se déroule en deux phases la première concerne la conceptualisation des bonnes pratiques et la seconde comprend l'extraction des meilleures pratiques pour une requête donnée. Les deux phases parcourent les principales étapes du processus de l'ECD, durant la première phase on procédera au recueil des données, que nous modéliserons par un graphe orienté, dans le but de construire une base graphiques de bonnes pratiques, la seconde phase est celle qui assistera l'utilisateur lors de son processus de recherche de la meilleure pratique grâce aux techniques d'apprentissage artificiel sur les graphes et aux techniques de résumé de texte. Dans ce qui suit nous présentons de manière détaillée chaque étape de notre approche :

6.4.1 Conceptualisation des bonnes pratiques

Cette phase à pour objectif la conceptualisation des bonnes pratiques, comme nous l'avions déjà souligné les bonnes pratiques sont des pratiques qui ont fait leur preuve pour réaliser un objectif ; nous les définissons formellement comme un ensemble d'étapes successive dont le parcoure amène à un but bien précis, conceptualiser ces bonnes pratiques revient à les formaliser dans une base de connaissances afin de faciliter leur réutilisation et favoriser ainsi l'apprentissage au sein d'une communauté de pratique, ou bien pour d'autres utilités aussi importantes comme leur exploitation dans les moteurs de recherches, applications intelligentes, etc. Cette phase comprend deux étapes majeures : le recueil des données et la modélisation des bonnes pratiques que nous présentons dans les sections suivantes :

Recueil des données

De nombreux projets d'analyse, et d'apprentissage automatique requièrent le raclage de sites Web pour recueillir les données à analyser surtout dans le cadre de l'exploration de texte (texte mining). L'extraction de données d'un site web peut être réalisée de plusieurs façons, spécifiquement par le biais d'APIs (interface de programmation d'application traduit

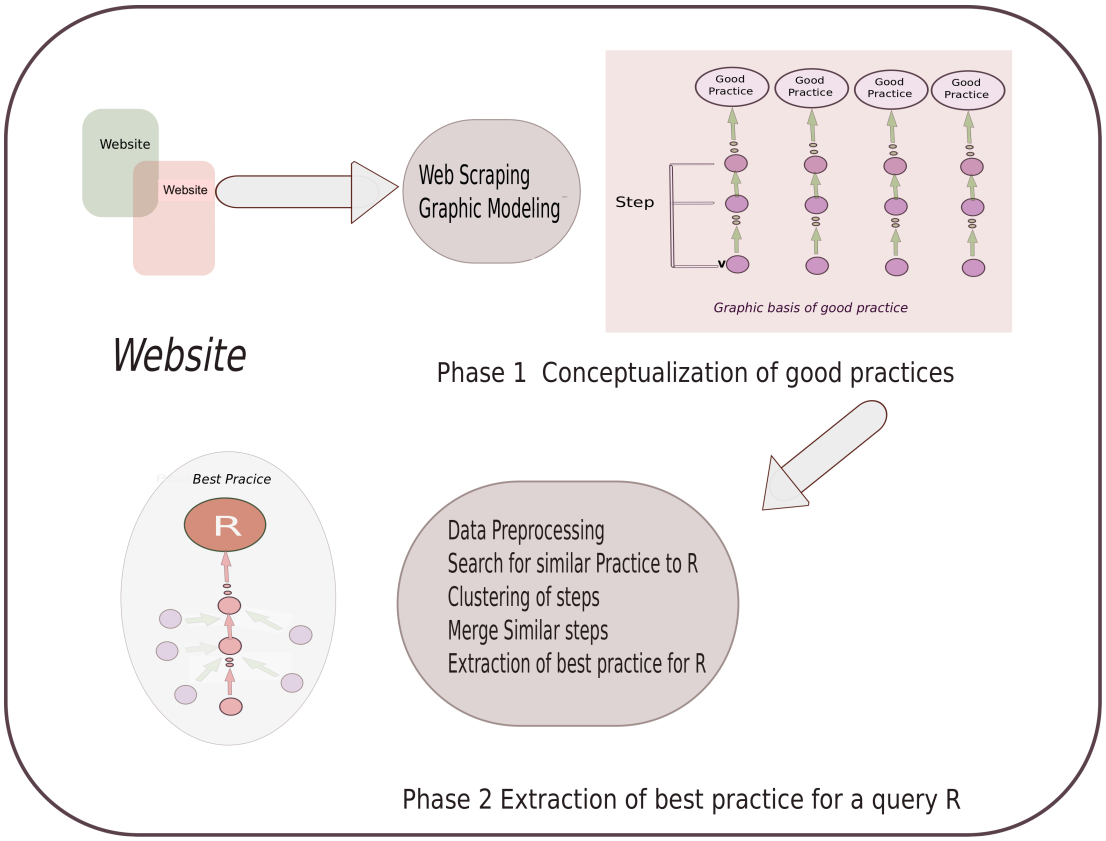

FIGURE 6.1 – *Processus de conceptualisation et d'extraction des meilleures pratiques*

de l'anglais Application Programming Interface) permettant l'utilisation d'un service web sans passer par l'interface utilisateur. Toutefois, tous les sites Web ne fournissent pas une API, soit pour une question de sécurité ou soit par manque de connaissances techniques. Une autre alternative est d'utiliser les robots d'exploration web qui grâce à un programme ou un script automatisé parcourt le web afin de pratiquer l'extraction d'une ou plusieurs parties d'un site web.

Dans l'approche que nous proposons on utilise le web scraping (également appelé récolte du web ou extraction de données Web) une version plus récente des robots d'exploration web. Le web scraping est une technique d'extraction du contenu d'un site web de manière automatisée à l'aide d'un programme informatique Glez-Peña et al. [2013], utilisée surtout dans le référencement. Le web scraping se concentre sur la transformation de données non structurées sur le web, généralement en Format HTML, en données structurées qui peuvent être stockées et analysées dans une base de données locale. Pour se faire il est nécessaire de lancer une requête http afin de parser les contenus des pages web à extraire. Ainsi on analyse des documents HTML ou XML dans une arborescence facilitant la recherche et l'extraction des données. Il faut savoir que le web scraping peut être contraire aux conditions d'utilisation de certains sites Web comme tout autre API d'extraction de contenu web. Dans notre cas on ne s'intéresse qu'aux enjeux scientifiques liés à l'adoption du web scraping.

Modélisation des bonnes pratiques

Pour rappel dans le chapitre 2 nous avions définit les bonnes pratiques comme un processus utilisé pour atteindre un objectif, chaque pratique représente un savoir faire sous forma de connaissance procédurale constituée d'un ensemble d'étapes successives. Pour modéliser les bonnes pratiques nous utilisons les graphes de données orientés. Les avantages d'une telle

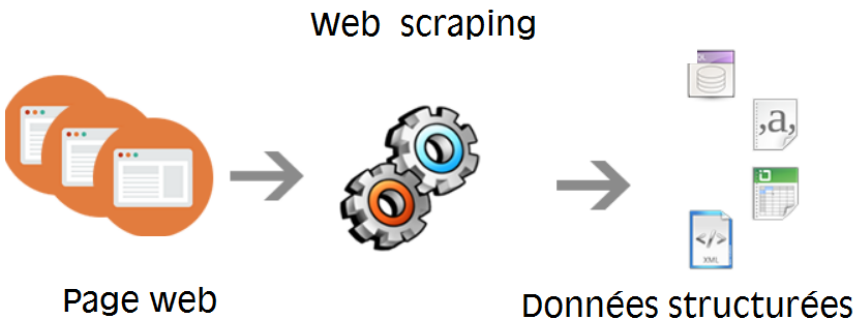

FIGURE 6.2 – *Processus du web scraping*

représentation sont indéniables : elle permet de représenter n'importe quel objet en nœud et n'importe quelle propriété en arc de plus elle est très flexible, on peut augmenter un graphe sans pour autant atténuer le graphe existant. Un graphe orienté G est définit par une paire d'ensemble V et E, noté G= (V, E), tel que :

— V est l'ensemble des nœuds représentant les étapes de la bonne pratique
— E est l'ensemble des arcs qui relient les étapes

On considère alors chaque phrase verbale (exemple «être en bonne santé») désignant le nom d'une bonne pratique comme des nœuds parents c.à.d. des nœuds sans successeur mais qui peuvent avoir des prédécesseurs. S'en suit l'ensemble des étapes qui composent la bonne pratique : la première étape de chaque pratique est représentée par un nœud n'ayant aucun prédécesseur liée par un arc à l'étape suivante jusqu'à arriver au but à atteindre représenté dans le nœud parent. La figure 6.3 présente un exemple de modélisation de la bonne pratique "Faire un gâteau" (Make a cake en Anglais), cette pratique regroupe 5 étapes successives à savoir : préparer les ingrédients, préchauffer le four, mélanger les ingrédients, cuire le gâteau dans le four , démouler le gâteau.

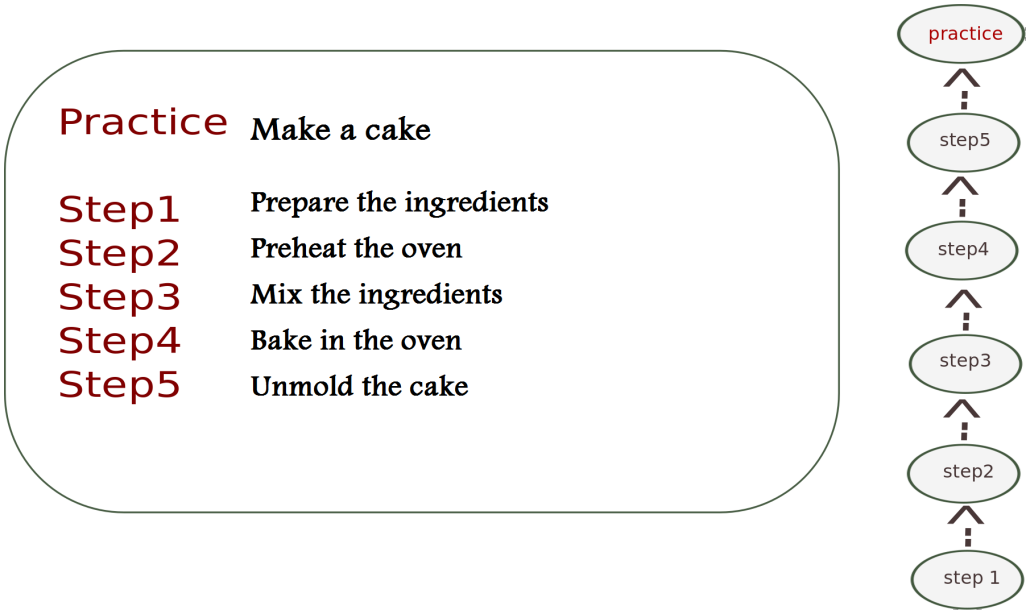

FIGURE 6.3 – *Exemple de modélisation d'une bonne pratique*

6.4.2 Extraction des meilleures pratiques

Cette seconde phase consiste à extraire la meilleure pratique parmi celles existantes dans notre base de connaissance pour une requête donnée afin d'assister l'utilisateur dans son processus de recherche. Compte tenu du nombre croissant de connaissances procédurales partagées dans la communauté web, il est difficile de se retrouver parmi tout ce qui est proposé, par exemple si nous lançons une requête pour trouver une recette pour un gâteau nous sommes confrontés au dilemme de choisir parmi tous les résultats retournés, donc identifier la meilleure façon de faire un gâteau devient laborieux, on peut certainement se référer aux systèmes d'évaluation en ligne mais encore une fois cela nécessite l'analyse et l'évaluation de toutes les notes et de tous les commentaires pour se forger une opinion, et comme nous l'avions déjà souligné dans le chapitre 2 différencier entre une bonne et une meilleure pratique reste subjective, car une bonne pratique aura certes démontrer ses preuves pour achever un but mais une meilleure pratique sera sélectionnée parmi toute les bonnes pratiques selon un critère, un contexte pouvant démontré sa supériorités par rapport aux autres.

Dans la recherche d'informations on considère qu'un site est important s'il est populaire, le célèbre algorithme de Google «PageRank» Brin et Page [1998] utilise la théorie des graphes pour évaluer la popularité d'un site grâce aux liens pointant vers un site représenté par des arcs entrants dans le graphique Web. L'approche que nous proposons est également basée sur la représentation des connaissances par des graphiques de données où les nœuds sont les étapes et les arcs orientés modélisent la successivité entre les étapes, pour trouver la meilleure pratique nous nous fondons alors sur l'hypothèse qu'une meilleure pratique sera la pratique ayant les étapes les plus utilisées par toutes les bonnes méthodes cherchant à atteindre le même but, dans notre graphe de connaissances cela se traduit par la méthodes passant par les étapes ou les nœuds les plus importants ayant le plus de liens ou d'arcs entrants.

Cette phase de l'approche se déroule en trois étapes : la première concerne la recherche des pratiques similaires à la requête de l'utilisateur, dans la seconde étape on procède au regroupement et à la fusion des noeuds similaires, et la dernière étape est consacrée à l'extraction de la meilleure pratique. Etant donné une requête "r" d'un utilisateur, on cherche d'abord toutes les pratiques ayant le même objectif que "r" : on utilise une méthode intuitive mathématique s'appuyant sur le prolongement de morts lexicales dit en anglais «word embedding» pour regrouper toutes les pratiques sémantiquement proches de "r". Le word embedding est un ensemble de techniques d'apprentissage artificiel qui visent à représenter les mots ou les phrases d'un texte par des vecteurs de nombres réels, ceci dans le but de capturer le contexte, et la similarité sémantique et syntaxique d'un mot en réduisant sa dimension. Historiquement le prolongement de mots repose sur l'hypothèse distributionnelle (distributional hypothesis) fondée par Zelling Harris qui établit que les mots sont caractérisés par leur contexte. Ainsi, des mots similaires partageront des contextes similaires. Comme l'a écrit John Rupert Firth en 1954, «Vous connaîtrez un mot par ses fréquentations.»Harris [1954]. Il existe plusieurs approches de word embedding. Les premières remontent aux années 1960 et reposent sur des méthodes de réduction de dimension. Plus récemment, de nouvelles techniques basées sur des modèles probabilistes et des réseaux de neurones, comme Word2Vec, ont permis d'obtenir de meilleures performances, dans notre cas nous utiliserons le modèle Word2Vec que nous présenterons de manière plus détaillée dans la section suivante.

A partir de là, on définit alors un nouveau graphe G constitué d'un nœud parent représentant l'objectif (la requête) "r" de l'utilisateur, auquel on relie par des arcs l'ensemble des bonnes pratiques sélectionnées sur la base de la distance sémantiques entre l'objectif et les noms des pratiques existantes. Trouver la meilleure pratique revient alors à évaluer tous les chemins possibles dans le graphe pour atteindre l'objectif "r", dans ce cas les différents chemins trouvés peuvent emprunter les mêmes étapes ou des étapes similaires, dans l'exemple comment faire un gâteau les étapes comme préparer ou mélanger les ingrédients, préchauffer le four à 180 ° ou allumer le four à 180 ° peut être commun à toutes les pratiques, ce qui si-

gnifient qu'elles sont importantes pour toute les méthodes. Dans ce cas dans la seconde étape de cette phase de l'approche on procède à la fusion des nœuds similaires qui représentent des étapes identiques, on utilise alors dans un premier temps un algorithme de clustering non supervisé DBSCAN pour regrouper les nœuds sémantiquement proches, par la suite on effectue la fusions des nœuds similaires par une technique de synthèse de texte.

Dans la dernière partie de l'approche, on identifie la meilleure méthode répondant à la requête r de l'utilisateur à partir du nouveau graphe G. Dans ce cas on comparera l'importance des chemins parcourus dans le graphe représentants les bonnes pratiques pour atteindre l'objectif r recherché par l'utilisateur, cette mesure d'importance se base sur les chemins traversant les nœuds ayant le plus grand nombre d'arcs entrants pour atteindre r. La figure 6.4 présente la phase d'extraction de la meilleure pratique pour une requête. Nous expliquons de manière plus détaillée ci-dessous chacune des étapes parcourues lors de cette phase de l'approche :

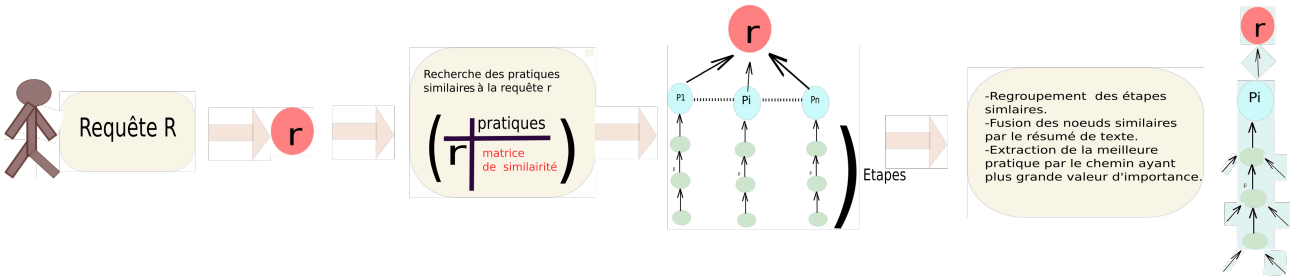

FIGURE 6.4 – *Phase d'extraction de la meilleure pratique*

Recherche des pratiques similaires

La recherche d'information classique s'appuie généralement sur différentes méthodes de classification supervisée (exemple les réseaux de neurones) et non supervisées (exemple l'algorithme Kmeans) ou des mesures de similarité tels que le word embedding afin de répondre aux requêtes des utilisateurs dans le but de retourner les documents sémantiquement proches de la requête émise. Ces méthodes dans le cas classique tentent de trouver parmi le corpus de document existant les documents dont l'index est le plus proche de la requête ceci dit en passant par un pipeline d'étapes. Dans notre cas, l'approche que nous proposons s'inspire de la méthode «Word mover's distance » WMD traduit littéralement en Français par distance par déplacement de mots Kusner et al. [2015]. Rappelons que le WDM, exploite les techniques d'intégration avancées telles que le word embedding pour calculer la distance sémantique entre des documents texte, pour cela on suggère que la distance entre deux documents texte A et B est calculée par la distance cumulative minimale que les mots du document texte A doivent parcourir pour atteindre les mots incorporés du document texte B. Dans notre cas, étant donnée une requête de l'utilisateur, nous considérons chaque titre de pratique en parallèle avec les mots constituant la requête et après une série de pré-traitent et de nettoyage que nous expliciterons ci-dessous nous transformons ces mots en vecteurs en utilisant le modèle de prolongement de mots Word2Vec présenté plus bas et nous sélectionnerons ainsi les pratiques ayant des représentations vectorielles les plus proches de la requête, pour cela nous considérons une matrice de similarité sémantique sur la quelle on reportera les distances sémantiques entres les éléments (phrase représentant la requête de l'utilisateur et les phrases représentants les bonnes pratiques existantes) à comparer. Nous présentons dans ce qui suit chaque partie de cette étape.

— **Prétraitement des données**
Les données réelles sont souvent incomplètes : valeurs manquantes, données simplifiées bruitées, incohérentes. C'est pourquoi il est rare que le texte brut se prête directement à l'analyse, surtout que Les résultats de la fouille dépendent de la qualité des

données. Le prétraitement des données est l'étape de base de tout processus d'extraction et d'exploration des connaissances. Cette étape commence par la sélection et la détermination de l'ensemble des données sur lequel s'effectuera le processus d'ECD, ensuite on procède à différentes opérations de nettoyage, de réduction et de normalisation de données. Les principales opérations que nous suivons dans notre approche sont les suivantes :

- La suppression de mots vides
  appelée en anglais les stopwords. Ce sont les mots très courants dans la langue étudiée ("et", "à", "le"... en français) qui n'apportent aucune valeur informative pour la compréhension du "sens" d'un document et corpus, qu'on isole et qu'on supprime.
- La lemmatisation
  La lemmatisation consiste à remplacer chaque mot par sa forme canonique par exemple le mot connaissons se rapporter à sa forme canonique connaitre. Cette étape est utile pour la classification thématique de textes car elle permet de traiter comme un mot unique les différentes variantes issues d'une même forme canonique ou racine.
- La tokenisation
  la tokenisation est le fait de parser du texte en token en d'autre terme on segmente le texte en unités linguistiques comme les mots, la ponctuation, les nombres, les données alphanumériques...Chaque élément correspondant à un token qui sera utile à l'analyse.

— **Mesure de similarité Word2Vec :**
Le regroupement des pratiques et des étapes du graphe de données nécessite l'utilisation de mesure de similarité ; par conséquent, nous utilisons le modèle d'intégration Word2Vec de Tomas Mikolov Mikolov et al. [2013]. Ce modèle a l'avantage d'être facilement accessible par un langage de programmation comme Python de plus il est rapide à entraîner car il est basé sur un réseau de neurones à deux couches formé pour prédire la représentation vectorielle des mots en contexte, plus simplement Word2Vec prend comme entrée un corpus de texte et donne en sortie un ensemble de vecteurs pour les mots de ce corpus. Son but est de regrouper les vecteurs de mots similaires dans un espace vectoriel, c'est-à-dire qu'il détecte les similitudes mathématiquement pour cela Word2vec crée des vecteurs qui sont des représentations numériques distribuées de caractéristiques de mot, telles que le contexte de mots individuels. Ainsi avec suffisamment de données, d'utilisation et de contextes, Word2vec peut faire des suppositions très précises sur la signification d'un mot en fonction des apparences passées. Ces suppositions peuvent être utilisées pour établir l'association d'un mot avec d'autres mots (par exemple, «homme» signifie «garçon» ce que «femme» signifie «fille»), ou regrouper des documents et les classer par sujet.
Word2Vec a deux architectures neuronales, appelées CBOW et Skip-Gram. CBOW prédit la probabilité d'apparition d'un mot en fonction de son contexte. Skip-Gram que nous utiliserons dans l'expérimentation fait exactement le contraire : il prend un mot comme entrée et essaie de prédire son contexte. Cet apprentissage automatique nécessite une base de formation, pour notre part, nous utiliserons le modèle d'apprentissage de Google Google [2013] disponible sur le web, formé de 300 dimensions vectorielles pour 3 millions de mots et de phrases.

— **Matrice de similarité**
Dans notre approche nous définissons une matrice dite de similarité sémantique que nous nommerons M (m,n) tels que m et n sont le nombre de phrases que nous souhaitons comparer et pour les quelles on utilise le modèle Word2Vec afin de trouver leur représentation vectorielle. Ensuite, nous définissons une fonction F (x, y), cette fonction comprise entre [0,1] renvoie le taux de similitude entre deux phrase x et y sur le base

de la proximité entre leur représentation vectorielle grâce au modèle Word2vec. On définit un seuil de similarité s = 0.7, au-delà duquel deux mots sont considérés comme sémantiquement proches, c'est-à-dire :
— Si F (x, y)> = s : M (x, y) = M (y, x) = F (x, y)
— Sinon M (x, y) = M (y, x) = 0
Dans cette partie de l'approche nous construisons une matrice de similarité entre toutes les pratiques de notre corpus et la requête de l'utilisateur, et nous prendrons en considération toutes les correspondances qui seront supérieures au seuil s que nous avons définit afin de sélectionner parmi notre base de connaissances les bonnes pratiques répondant à l'objectif recherché par l'utilisateur et construire par conséquent le nouveau graphe de connaissance G comprenant la requête de l'utilisateur et les pratiques qui lui sont similaires
Cette matrice de similarité sera utilisée tout au long des étapes suivantes : on calculera une matrice de similarité entre les étapes sélectionnées lors du processus de classification des nœuds similaires, et ensuite on construira une autre matrice lors du processus de synthèse de texte.

Fusion des nœuds identiques

Après avoir sélectionné les pratiques similaires au but recherché par l'utilisateur nous nous retrouvons avec un nouveau graphe pouvant contenir des nœuds similaires c'est-à-dire des étapes syntaxiquement différentes mais sémantiquement identiques, il serait judicieux donc de les fusionner. Cette partie de l'approche nécessite de faire appel à la classification non supervisée en premier lieu pour regrouper les étapes similaires dans des clusters distincts. Ensuite il faudra déterminer pour chaque groupe d'étapes dans chaque cluster quelle phrase résumera l'ensemble des étapes à fusionner dans un nœud final et pour cela nous utiliserons une technique de résumé de texte que nous présenterons de façon plus détaillée ci-dessous :

— **Algorithme de classification non supervisée :**
Pour fusionner les nœuds similaires il est nécessaire de les regrouper selon leur proximité sémantique en d'autre terme il s'agit de classifier les nœuds représentant les étapes sémantiquement proches, et c'est pour cela que nous faisons appel aux techniques d'apprentissage non supervisé. Ne connaissant pas a priori le nombre de clusters, nous avons choisi d'utiliser l'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise), qui présente l'avantage d'être efficace en temps de calcul sans nécessiter de prédéfinir le nombre de clusters. Cet algorithme est utilisé pour regrouper les points de données en haute densité et ne tient pas compte des valeurs aberrantes dans les régions de faible densité Ester et al. [1996]. Il a deux paramètres principaux : "eps" qui détermine le seuil au-dessus duquel 2 points sont considérés comme voisins et "min_point" est le nombre minimum de points pour former une région dense.
L'algorithme DBSCAN prend en entrée la matrice de similarité obtenue en appliquant le modèle d'apprentissage Word2Vec (que nous avions présenté dans la section précédente) entre les nœuds du graphe qui représentent les étapes de méthodes sélectionnées lors du processus de recherche des pratiques similaires dans l'étape précédente et générera ainsi un groupe de clusters avec des éléments fortement liés dans le but d'obtenir des ensembles distincts contenants chacun un groupe de nœuds similaires.
— **La synthèse de texte :**
La synthèse de texte (text summarization en anglais) est une tâche importante de l'apprentissage automatique et le traitement du langage naturel (Processing Natural Language PNL), elle fait référence à la technique de raccourcissement des textes volumineux dans le but de créer un résumé cohérent et fluide contenant les principaux points soulignés dans le document. Les modèles d'apprentissage automatique sont générale-

ment formés pour comprendre les documents et distiller les informations utiles avant de produire les résumés requis.
Il existe deux types principaux de synthèse de texte en PNL : l'extraction et l'abstraction. Les méthodes d'extraction sélectionnent un sous-ensemble de mots, d'expressions ou de phrases existants dans le texte d'origine pour former un résumé. En revanche, les méthodes abstractives construisent d'abord une représentation sémantique interne, puis utilisent des techniques de génération de langage naturel pour créer un résumé. Un tel résumé peut contenir des mots qui ne sont pas explicitement présents dans le document original. La plupart des systèmes de synthèse de texte sont basés sur une forme de synthèse extractive Michael, Gudivada et Rao [2018].
Dans notre approche, nous utilisons la synthèse de texte pour fusionner les nœuds représentant les étapes similaires, nous recherchons donc parmi chaque ensemble d'étapes regroupées la représentation lexicale la plus proche des autres. Pour cela, nous utilisons une méthode d'extraction de similarité qui identifie les points les plus importants d'un ensemble de phrases. Pour chaque groupe d'étapes obtenu lors du processus de clustering, nous construisons d'abord une matrice de similitude entre les phrases représentant chaque nœud (étape) du cluster en se basant sur le modèle d'apprentissage Word2Vec, cette matrice renvoie le taux de similarité entre les étapes regroupées dans chaque cluster. A partir de cette matrice on génère un graphe de données où les nœuds représentent les phrases du corpus et les arcs les liens de similitude entre ces phrases, nous appliquons ensuite l'algorithme de classification PageRank Brin et Page [1998] afin de classer les phrases les plus importantes. Rappelons que cet algorithme identifie un nœud comme étant important s'il est pointé par d'autres nœuds importants. De là, nous pouvons identifier la phrase la plus importante dans chaque ensemble d'étapes c'est celle qui aura le plus grand nombre d'arcs entrants et ainsi nous obtenons la synthèse lexicale des étapes regroupées ensemble. La figure 6.5 schématise un exemple de processus de synthèse de texte pour des nœuds similaires.
Enfin nous porterons ces nouvelles modifications sur notre graphe de connaissance G par la fusion des nœuds de chaque clusters par un nœud final qui sera représenté par la phrase désignée dans le processus de synthèse de texte au niveau de chaque cluster, chaque nœud final héritera des arcs entrants et sortants des autres étapes qui lui sont similaires.

Extraction de la meilleure pratique

Dans cette dernière étape de l'approche proposée, on fait appel aux théories des graphes de données pour extraire la meilleure pratique pour une requête donnée. Plus précisément on a recours aux indicateurs qui mesurent la notion d'importance dans un graphe en identifiant les sommets populaires. Cette notion d'importance représente un enjeu majeur dans plusieurs domaines et notamment dans l'analyse des réseaux sociaux, où l'on cherche à identifier les communautés en lignes et les personnes influentes grâce a différentes mesures de centralité des graphes Wasserman et Faust [1994], en recherche d'information le classement des résultats est aussi influencé par le degré d'importance d'un document, le meilleure exemple reste le célèbre algorithme PageRank de Google Brin et Page [1998] qui dit qu'une page est importante si elle est taguée par d'autre page importante.

Dans notre approche aussi nous exploitons cette notion d'importance des sommets dans les graphes pour évaluer les bonnes pratiques, ainsi la meilleure pratique pour une requête donnée sera celle qui empruntera le chemin regroupant les sommets les plus importants, et afin de quantifier cette importance de chaque chemin menant à l'objectif recherché, nous nous basons sur la mesure de centralité de degré appelée aussi mesure de prestige Freeman [1978]. Cette mesure est la forme la plus simple de la notion de centralité, elle est fondée sur l'idée qu'un sommet important dépend du nombre de ses sommets voisins ce qui revient à calculer

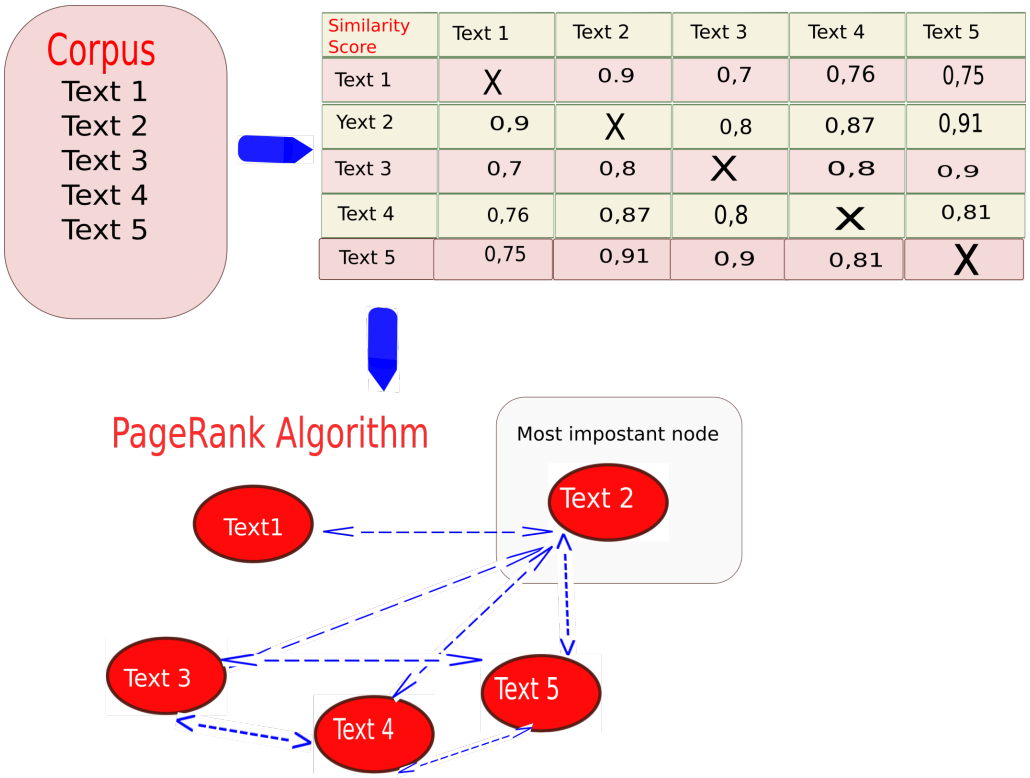

FIGURE 6.5 – *Processus de synthèse de texte des nœuds représentants les étapes similaires*

le nombre de ses liens incidents. En théorie des graphes ce nombre est appelé degré entrant ou sortant du nœud i, qui se mesure soit par le nombre de ses arcs entrant on parle alors de degré de centralité entrant ou par le de ses arcs sortants on parle alors de degré de centralité sortant.

Dans notre approche nous allons comparer la valeur d'importance de chaque chemin afin d'identifier la meilleure pratique, pour quantifier cette importance nous calculons le rapport entre la somme des arcs entrants des nœuds que parcoure chaque chemin avec le nombre total des nœuds parcourus, plus formellement on définit l'importance de chaque chemin par la formule suivante :

$$Valeur\_Importance(Cj) = \frac{\sum_{i=1}^{n} de(i)}{n} \tag{6.1}$$

Avec :

— Cj : le chemin j qui mènent au but recherché par la requête de l'utilisateur,
— n : le nombre de nœuds total que parcoure le chemin Cj pour atteindre le but recherché
— de(i) : le degré entrant de nœud i contenu dans Cj qui se traduit par le nombre de ses arcs entrants.

Donc au terme de cette étape grâce à notre algorithme nous extrayons le chemin ayant les étapes les plus importantes menant à l'objectif recherché par l'utilisateur qui représente la meilleure pratique.

6.4.3 Algorithme d'extraction de la meilleure pratique

Dans cette section nous présentons notre algorithme pour extraire la meilleure pratique, pour une requête donnée. Cet algorithme retrace toutes les étapes de la phase 2 de l'approche proposée afin d'identifier et d'assister l'utilisateur dans son processus de recherche de la meilleure pratique.

Algorithm 1 Extraction of best practice

Initialisation
s=0,7
G= graph of good practice
R User Request
P=Practice in G
G'= New graph with node R
Selection of similar practice to request R :
for P in good practice of graph G **do**
Cleaning and stemming (P)
if $Semantic\_similarity(R, P, Word2vec_model)= S\_S > s$ **then**
Similarity_Matrix(R, P) = S_S
Add practice P and her steps to G' with arcs entering towards node R.
else
Similarity_Matrix(R, practice) =0
end if
end for
Fusion of step node :
Clustering Steps
for S1 in Step of G' **do**
for S2 in Step of G' **do**
Cleaning and stemming (S1,S2)
if $Semantic\_similarity(S1, S2, Word2vec_model)=S\_S > s$ **then**
Similarity_Matrix(S1,S2)=S_S
else
Similarity_Matrix(S1,S2)= 0
end if
end for
end for
Clustering_DBSCAN (Steps of G', Similarity_Matrix)
Summarizing of text
for *steps*in each *clusters_step* **do**
Caclsulate Similarity Matrix (steps, Word2vec_model)
end for
Ranked node of GS
Fusion node of each cluster by node in G' with maximum score ranked
Extraction of best practice
for each path j in G' from node R to final node **do**
Calculate Importance_Value of path *j*
end for
Select path with maximum Importance_Value as best practice to achieve R

6.5 Exemple d'application

Dans cette section nous présentons un exemple applicatif afin d'expliquer au mieux l'approche que nous proposons. Pour se faire, nous définissons une base de connaissances regroupant quelques bonnes pratiques portant sur le domaine de la santé extraites du site de partage WikiHow wikiHow [2019], notamment les pratiques à suivre pour être en bonne santé chacune de ces bonnes pratiques regroupent différentes étapes pour les réaliser. Le tableau 6.1 ci-dessous, présente les étapes et pratiques de l'exemple.

Pratique	**Etapes**
P1 : Having a Healthy Diet	S1 : Drink more water S2 : Eat breakfast. S3 : Eat well throughout the day S4 : Eat meatless at least a few days a week. S5 : Eat at the right times S6 : Consider going meatless at least a few days a week. S7 : Limit simple sugars in your diet S8 : Read food labels to make the healthiest choices. S9 : Talk to your doctor about incorporating supplements and sugar in your diet
P2 : Having a Healthy Exercise Plan	S10 : Get in shape S11 : Maintain a healthy weight S12 : Cross train S13 : Exercise often S14 : Take advantage of opportunities to be active
P3 : Being Emotionally Healthy	S15 :Think positively S16 : Be satisfied. S17 : Think small. S18 : Manage stress. S19 : Choose your friends wisely. S20 : Be productive. S21 : Take a break S22 : Find emotional balance. S23 : Include the arts in your life, such as music, theater, and visual arts. S24 : Travel as much as you can
P4 : Having a Healthy Routine	S25 : Create a daily routine. S26 : Stop engaging in risky behavior S27 : Exercise several times. S28 : Get a good night's rest S29 : Learn how to cook. S30 : Maintain your personal hygiene. S31 : Bolster your immune system
P5 : Creating the Right Mindset	S32 : Focus on the positive. S33 : Don't compare yourself to others S34 : Manage stress in your life S35 : Find time to relax. S36 : Establish healthy, manageable goals. S37 : Express gratitude for the good things in your life. S38 : See a mental health professional if you feel depressed, anxious, or suicidal
P6 : Eating for Health and Mood	S39 : Practice mindful eating to increase satisfaction S40 : Consume 5-6 servings of fruit and vegetables a day. S41 : Choose foods high in fiber S42 : Find sources of omega-3 fatty acids. S43 : Avoid processed foods and fast food. S44 : Substitute unhealthy ingredients with healthier choices
P7 :Practicing Beneficial Habits	S45 : Get enough sleep. S46 : Exercise for at least 30 minutes a day S47 : Get 12-15 minutes of sun exposure a day S48 : Meditate once a day
P8 : Maintaining a Healthy Social Life	S49 : Establish lasting bonds with family and friends. S50 : Adopt a pet if you can care for it. S51 : Help others S52 : Distance yourself from toxic or needy personalities

TABLE 6.1 – *Exemple de bonnes pratiques dans le domaine de la santé*

A partir des informations du tableau 6.1 nous pouvons conceptualiser les bonnes pratiques en les modélisant dans un graphe de connaissances afin de faciliter leur réutilisation. La figure 6.6 montre la base de graphe des bonnes pratiques de l'exemple, nous y constatons un ensemble de graphes orientés éparpillés qui représentent les bonnes pratiques décrit dans le tableau 6.1 ainsi que les étapes qui les décrivent (pour plus de lisibilité la partie de droite de la figure représente le même graphe labellisé de la partie gauche).

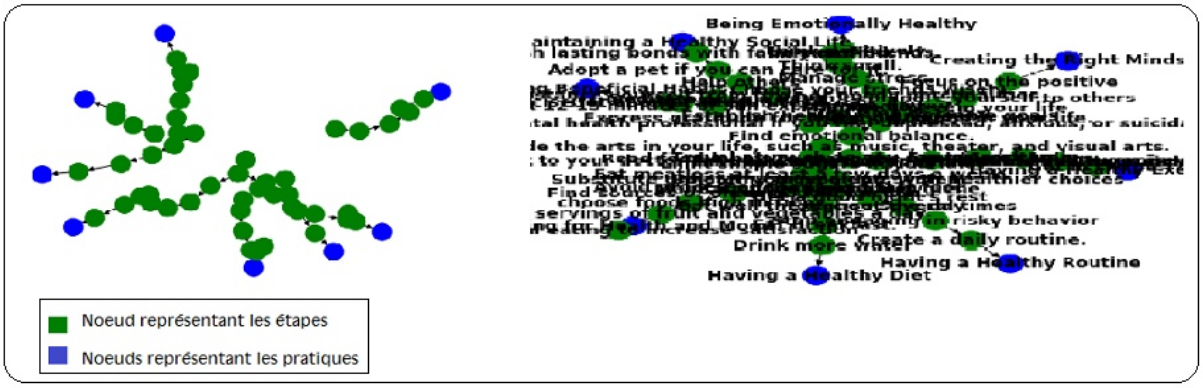

FIGURE 6.6 – *Base graphique des bonnes pratiques*

Soit R la requête d'un utilisateur portant sur la question «how to be healthy». La seconde phase de l'approche proposée tente d'extraire la meilleure pratique pour répondre à la requête de l'utilisateur. La première étape dans ce cas consiste à regrouper les bonnes pratiques similaires à la requête R, on utilise alors le modèle de prolongement lexicale Word2Vec pour calculer la distance sémantique entre les phrases grâce à leur représentation vectorielle, on définit le seuil de similarité s =0.7 à partir du quel deux phrases sont considérées sémantiquement porches. On obtient ainsi les taux de similarité présentés dans le tableau 6.2 :

Pratique	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
Score de similarité	0.7769	0.7001	0.5898	0.7431	0	0	0	0.6437

TABLE 6.2 – *Taux de similarité sémantique entre la requête et les bonnes pratiques*

Nous modélisons ainsi le nouveau graphe G' qui initialement contient un seul nœud racine représentant la requête R de l'utilisateur, auquel on relie les bonnes pratiques qui lui sont similaires, extraites à partir du graphe de bonnes pratique G par des arcs entrants. La figure 6.7 schématise le nouveau graphe G', l'image de droite représente le même graphe que celui de gauche labellisé

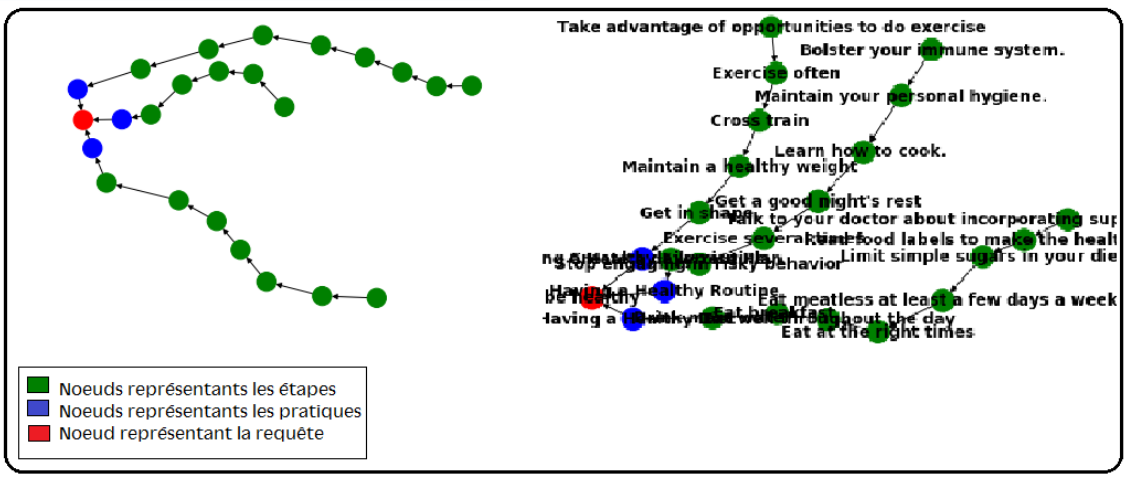

FIGURE 6.7 – *Graphe de connaissances G' représentant les bonnes pratiques reliées à la requête de l'utilisateur*

Nous passons ensuite à l'étape suivante qui a pour objectif de fusionner les nœuds similaires, pour se faire nous regroupons en premier lieu les étapes similaire par l'algorithme DBSCAN basé sur la similarité sémantique Word2vec, nous obtenons quatre clusters d'étapes similaires qui sont illustrés dans le tableau 6.3 :

Cluster	**Etapes**
Cluster1	Eat breakfast Eat well throughout the day Eat at the right times Eat meatless at least a few days a week.
Cluster2	Limit simple sugars in your diet Talk to your doctor about incorporating supplements and sugar in your diet
Cluster3	Get in shape Get a good night's rest
Cluster4	Exercise often Take advantage of opportunities to do exercise Exercise several times

TABLE 6.3 – *Tableau représentant les clusters regroupant les étapes similaires de G'*

Afin de fusionner les nœuds similaires, nous procédons à la synthèse des étapes regroupées dans chaque cluster, nous calculons d'abord la matrice de similarité sémantique entre les nœuds de chaque groupe, et on applique l'algorithme PageRank pour classer ces derniers et les synthétiser par une seule proposition qui résumera les autres étapes afin de les fusionner. La figure 6.8 montre les résultats obtenus dans cette étape.

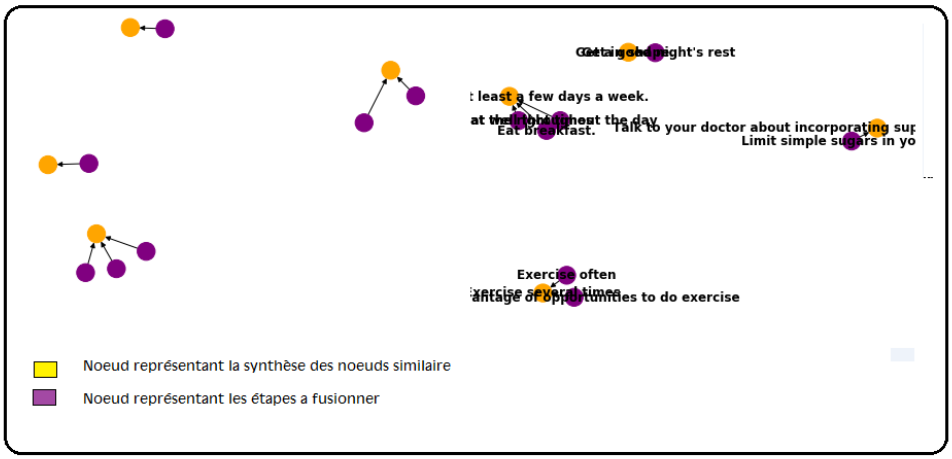

FIGURE 6.8 – *Fusion des nœuds similaires*

En dernier lieu après fusion des étapes similaires, nous obtenons un graphe plus dense comme illustré dans la figure 6.9 à partir duquel nous pouvons identifier les différents chemins possibles à partir des nœuds finaux (c.à.d.des nœuds n'ayant pas de successeurs) menant à l'objectif r (c.à.d. le nœud racine R qui représente la requête de l'utilisateur),

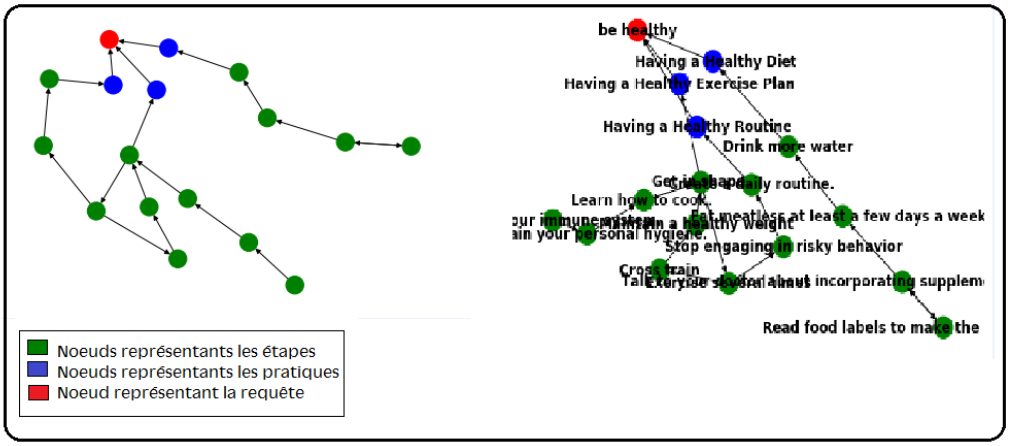

FIGURE 6.9 – *Graphe G' après fusion des nœuds similaires*

Pour finir, nous calculons la valeur d'importance des différents chemins menant à la requête de l'utilisateur et nous comparons ainsi les valeurs obtenues afin d'extraire la meilleure pratique qui représente le chemin ayant la plus grande valeur d'importance en d'autre terme le chemin qui emprunte les étapes les plus utilisées par les autres pratiques. Le tableau ci-dessous montre le résultat final de l'exemple.

Chemin identifié	**Valeur d'importance**
Bolster your immune system-Maintain your personal hygiene-Learn how to cook-Get in shape-Having a Healthy Exercise Plan-be healthy	1,6
Bolster your immune system-Maintain your personal hygiene-Learn how to cook-Get in shape-Exercise several times-Stop engaging in risky behavior-Create a daily routine-Having a Healthy Routine-be healthy	1,625
Read food labels to make the healthiest choices-Talk to your doctor about incorporating supplements and sugar in your diet-Eat meatless at least a few days a week-Drink more water-Having a Healthy Diet -be healthy	2,8 Best Practice

TABLE 6.4 – *Identification de la meilleure pratique*

6.6 CONCLUSION

Dans ce chapitre nous avons présenté une nouvelle méthode pour conceptualiser les connaissances procédurales du web sous forme de graphe orienté et extraire la meilleure pratique pour une requête lancée afin d'assister l'utilisateur dans son processus de recherche. L'utilisation des différentes technologies comme le web scraping pour récupérer le contenu web et des systèmes d'apprentissage et de raisonnement de l'intelligence artificielle ainsi que les techniques de traitements du langage naturels notamment le résumé de texte et les principes de théories de graphes ont été très utiles pour lever le défi de la problématique traitée par l'approche proposée.

Dans la dernière partie de ce chapitre, nous avons déroulé un exemple applicatif dans le but de mieux cerner notre proposition. Et afin de valider l'approche proposée nous présenterons dans le chapitre suivant l'expérimentation que nous avons menée.

Expérimentation et résultats 7

Sommaire

7.1 Introduction

Ce dernier chapitre est consacré à l'expérimentation que nous avons menée afin de tester la faisabilité de notre approche à savoir la conceptualisation des meilleures pratiques au sein d'une communauté de pratique et l'extraction de la meilleure pratique pour une requête donnée. Et dans cette perspective nous nous somme penchés sur les communautés en ligne afin de mener notre expérience car comme nous l'avons déjà souligné le web est devenu ces dernières années une source constante et permanente de connaissances évolutives où des utilisateurs répartis géographiquement ne cessent de partager leur savoir faire, leur expérience acquise sous forme de connaissances procédurales dans différents domaines d'expertise. Sans le savoir les internautes ont créé des communautés de pratique sur la toile ou les uns deviennent apprenants des savoir faire des autres dans des domaines précis, ceci grâce a différente plate forme collaborative de partage tél que le site de partage que nous utilisons pour mener nos tests Wikihow qui pour un sujet bien définit présente un ensemble de solutions sous forme de méthodes ou pratiques avec des étapes à suivre pour chaque pratique.

Dans ce qui suit nous présentons en premier lieu le langage de programmation et- l'environnement de développement que nous avons utilisé pour mettre en œuvre notre approche ensuite nous parlerons du jeu de données de la plate forme WikiHow sur lequel s'effectuera l'expérimentation et enfin nous discuterons des résultats obtenus.

7.2 Mise en œuvre : langage de programmation, environnement de développement

Afin de mettre en œuvre notre expérimentation, nous avons implémenté des lignes de code en langages de programmation Python sur l'environnement de développement intégré (IDE) Spyder. Dans ce qui suit nous présentons la description et les avantages de Python et de l'IDE Spyder qui nous ont orientés dans notre choix.

7.2.1 Langage de programmation

Python est un langage de programmation open source créé par le programmeur Guido van Rossum en 1991. Le langage de programmation Python est apparu à l'époque comme une façon d'automatiser les éléments les plus ennuyeux de l'écriture de scripts ou de réaliser rapidement des prototypes d'applications. Toutefois depuis quelques années, Python est devenu le langage de programmation le plus utilisé dans le domaine du Machine Learning, du Big Data, de la robotique, etc. lebigdata [2020].

Parmi ses avantages est que Python est un langage de programmation interprété, donc il ne nécessite pas d'être compilé pour fonctionner. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci le rend plus lent que d'autres langages. Un autre avantage est que Python est facile à apprendre et à utiliser, ses caractéristiques sont peu nombreuses, sa syntaxe est conçue pour être lisible et directe, de plus c'est un langage multiplateforme qui fonctionne sur tous les principaux systèmes d'exploitation et plateformes informatiques.

Python dispose de nombreuses bibliothèques, telles que TensorFlow, Pandas et Numpy qui permettent d'effectuer une large variété de tâches et qui ont été très utiles pour notre expérimentation. Les bibliothèques comme Lxml et BeautifulSoup permettent d'extraire des données depuis internet pour le scraping Web, tandis que Seaborn et Matplotlib aident à la visualisation des données. De leur côté, Tensorflow, Keras et Theano permettent le développement de modèles de Deep Learning, et Scikit-Learn aide au développement d'algorithmes de Machine Learning. La bibliothèque Pandas que nous avons utilisé est aussi connue pour

ses nombreuses fonctionnalités telles que de lire des données en provenance de nombreuses sources, de créer de larges data frames à partir de ces sources.

7.2.2 Environnement de développement

Spyder dont l'acronyme signifie «Scientific PYthon Development EnviRonment» est un environnement scientifique gratuit et open source écrit en Python, pour Python, et conçu par et pour des scientifiques, des ingénieurs et des analystes de données.

Cet environnement de développement intégré (IDE) a été créé et développé par Pierre Raybaut en 2008, Spyder est maintenu depuis 2012 par une communauté de développeurs appartenant à la communauté Python. Il présente une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de déboggage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les superbes capacités de visualisation d'un package scientifique Spyder [2020].

7.3 Jeu de données

Notre expérimentation a été menée sur un jeu de données provenant de la plate forme collaborative du net Wikihow wikiHow [2019]. Ce projet d'écriture collaboratif basé sur la technologie wiki a comme objectif de construire le plus grand manuel d'instruction de qualité au monde afin de faciliter l'échange de données et d'informations. Ce lieu de recueil d'information appelée aussi Howto est universel et multilingue. Il a été fondé en 2005 par Jack Herrick, le site Web vise à créer des instructions sur presque tous les sujets imaginables et permettre à n'importe qui d'apprendre à faire quoi que ce soit. Un exemple tronqué d'un article de WikiHow et de la façon dont les paires de données sont construites est présenté dans la figure 7.1.

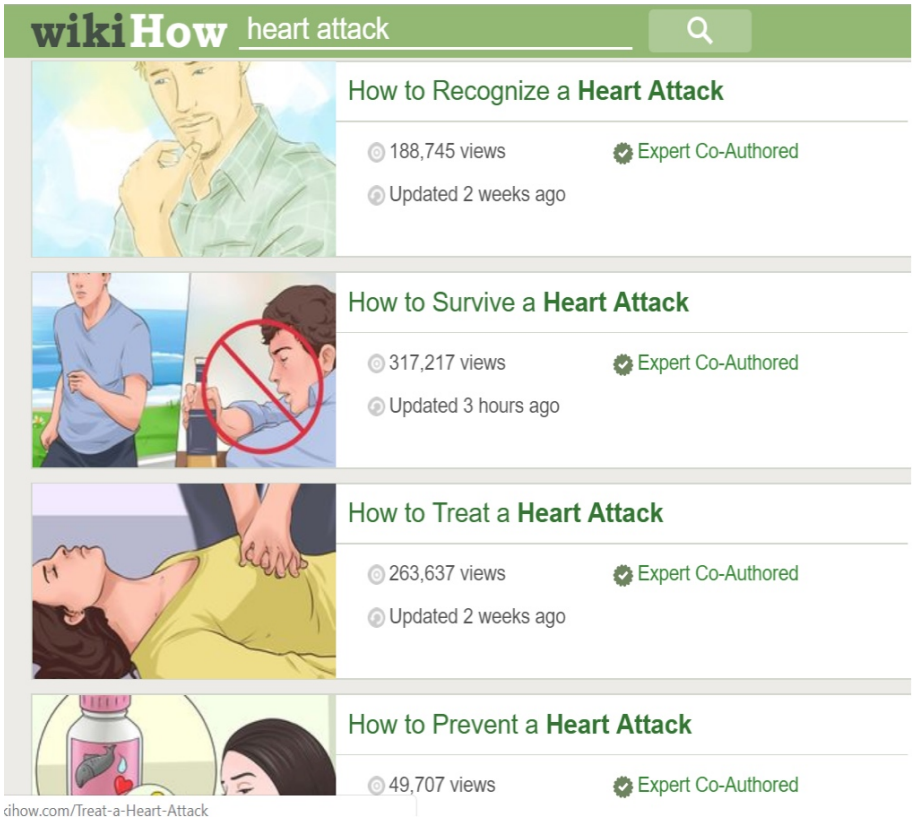

Figure 7.1 – *Exemple des résultats d'une requête pour un sujet particulier*

Le principe de fonctionnement de cette plate forme est similaire à celle d'une communauté de pratique car les individus répartis géographiquement peuvent partager un savoir faire, une pratique ou une méthodologie de travail, ils ne créent pas une communauté proprement dite mais ils partagent un savoir faire en forma de page web ou ils expliquent les solutions à un problème de manière procédurale avec un ensemble d'étape successives, en énumérant les méthodologies de résolution. Nous pouvons schématiser le processus de participation à la plateforme d'apprentissage comme suit (figure 7.2) :

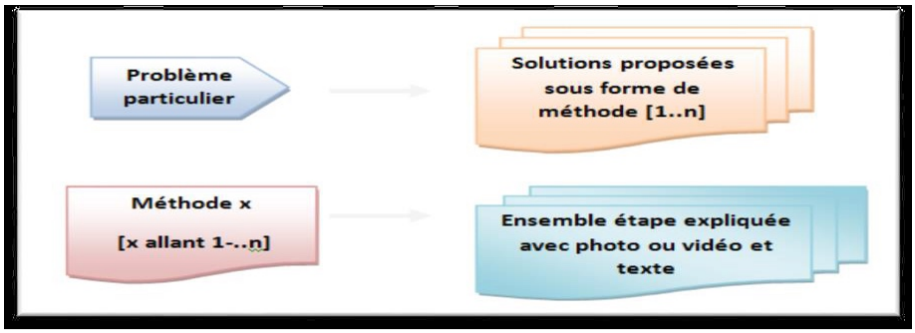

FIGURE 7.2 – *Processus participation à WikiHow*

En consultant différentes pages du site WikiHow, nous relevons quelques observations, notamment que les données de wikihow représentent un savoir faire sous forme procédurale, le problème à résoudre y est défini sous forme d'interrogation «how to do anything /comment faire n'importe quoi». Pour chaque problème un ensemble de solutions sont présentées sous forme de «méthodes ». Chaque méthode est structurée par un ensemble de «tâches successives» et un ensemble de contraintes d'objets exemple «ingrédient, ustensiles,..». On relève aussi des contraintes temporelles entre les étapes d'une méthode : chaque étape doit suivre un ordre de successivité bien précis, et ne doit se déclencher qu'après que l'étape précédente soit finie. On remarque aussi qu'une même méthode peut représenter d'un point de vue sémantique une étape dans une autre méthode ou encore une pratique à rechercher. La figure 7.3 illustre un exemple du modèle d'une page Wikihow.

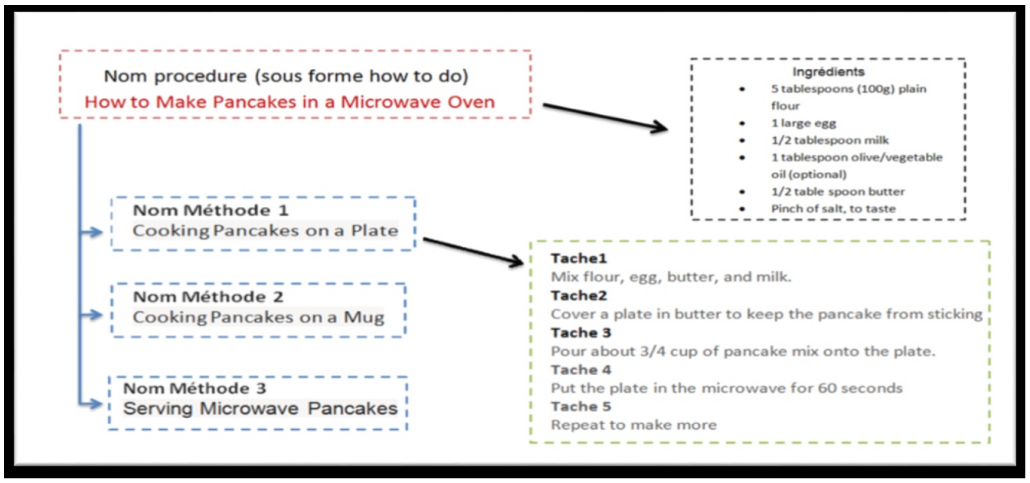

FIGURE 7.3 – *Modèle de page Wikihow*

7.4 Expérimentation

L'expérimentation que nous avons menée est réalisée en deux phases selon l'approche proposée, la première concerne le recueil et la modélisation des bonnes pratiques et la seconde concerne l'extraction de la meilleure pratique pour une requête donnée. Dans ce qui suit nous présentons l'expérimentation que nous avons mené et nous discuterons des résultats que nous avons obtenus.

7.4.1 Recueil et modélisation des bonnes méthodes

Afin de sélectionner les bonnes pratiques du web, notamment du site de partage WikiHow nous avons réalisé une implémentation de ligne de codes en python en faisant appel à la bibliothèque BeautifulSoup Richardson [2015] et Lxml Richter [2021] qui rappelons le aident à l'extraction de données web en se plaçant comme parseur xml et html afin de parcourir l'arborescence du site Wikihow telle que schématisée dans la figure 7.4 et extraire les données nécessaires.

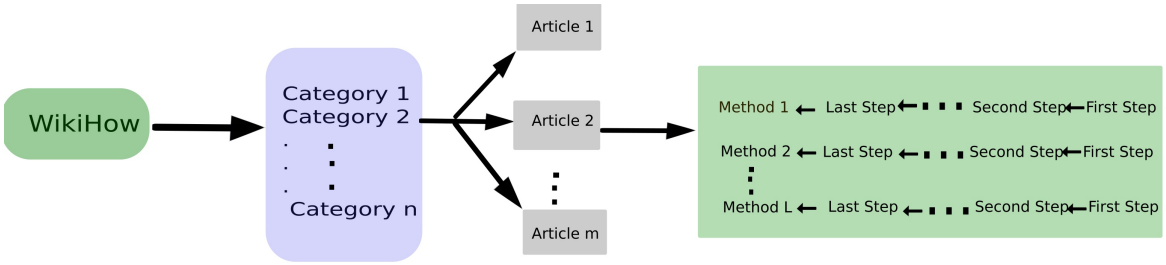

Figure 7.4 – *Structure du site web Wikihow*

Comme nous l'avons déjà spécifié dans la section précédente WikiHow est constitué d'un ensemble de catégories, qui elles mêmes regroupent un ensemble d'articles, et chaque article regroupe un ensemble de méthodes et enfin chaque méthode est formée par un ensemble d'étapes. Pour notre part on ne s'intéresse qu'aux articles touchant au domaine de la santé. Ainsi notre algorithme de scraping recherche les bonnes pratiques dans les articles regroupés dans la catégorie santé et des catégories liées qui sont connexes au même sujet. Ainsi nous avons extrait les articles contenus dans plus de 400 catégories et pour chaque article nous avons scrappé toutes les méthodes et les étapes qu'il décrit. Le résultat de l'extraction est reporté dans le tableau 7.1.

Nombre de catégories	407
Nombre d'articles	9659
Nombre de méthodes	27266
Nombre d'étapes	348634

Table 7.1 – *Résultat de l'extraction du site Wikihow*

L'ensemble des méthodes extraites appelées communément bonnes pratiques et les étapes nécessaires pour les réaliser sont modélisés dans un graphe orienté, où chaque bonne pratique représente un nœud parent précédé par un ensemble de nœuds successives correspondant aux étapes nécessaires pour achever chaque pratique. La figure 7.5 illustre cette partie de l'expérimentation.

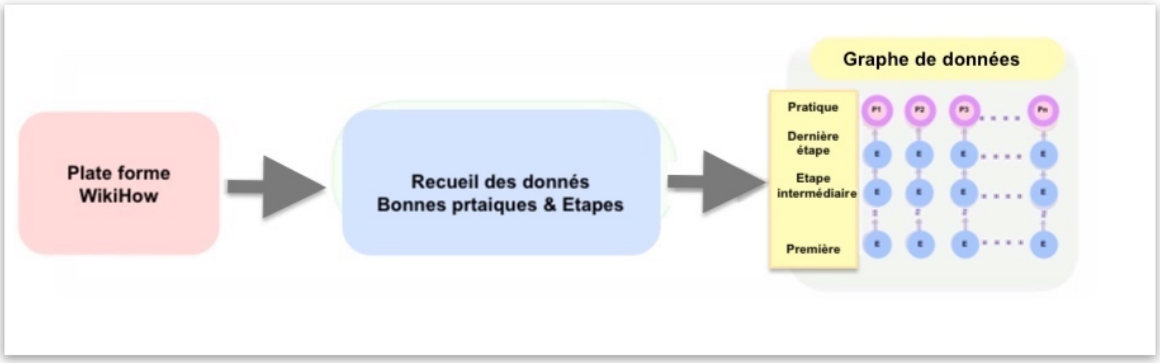

FIGURE 7.5 – *Extraction et modélisation des bonnes pratiques de Wikihow*

7.4.2 Extraction des meilleures pratiques

Dans cette section, nous analysons les performances de la seconde phase de notre approche qui concerne l'extraction de la meilleure pratique pour une requête donnée. Nous avons mené ainsi une série d'expérimentations à chaque niveau de cette phase (recherche des pratiques similaires, fusion des nœuds similaires, et identification de la meilleure pratique) afin de répondre aux questions suivantes :

- QR1 : Quelle est la performance de notre méthode de recherche d'information en termes de qualité des résultats retrouvés et de temps d'exécution ?
- QR2 : Etudier l'influence des paramètres utilisés dans DBSCAN par rapport au nombre de cluster et de points traités ?
- QR3 : Etudier l'impact de la métrique choisie lors de la classification des étapes similaires ainsi que le type de regroupement en termes de qualité de clustering ?
- QR4 : Vérifier la cohérence des résultats retrouvés en terme meilleure pratique ?

Les expérimentations ont été réalisées sur une machine sous Windows 7 dont les performances sont : 2.20 GHz Intel i5 core CPU, et 4GB de RAM.

Expérimentation pour QR1

Pour répondre à la première question de recherche, il nous faut évaluer la performance de notre système de recherche d'information en fonction de sa capacité à retourner des documents pertinents et de l'optimalité du temps d'exécution. En recherche d'information classique les deux principaux facteurs d'évaluation des SRI (systèmes de recherche d'information) sont le rappel qui signifie la capacité d'un système à sélectionner tous les documents pertinents de la collection et la précision qui est la capacité d'un système à sélectionner que des documents pertinents, mais ceci nécessite d'avoir un modèle préétabli pour effectuer la comparaison, ce qui n'est pas notre cas, de ce fait notre évaluation repose sur une comparaison entre approches du domaine notamment le WMD Kusner et al. [2015] que nous avons présenté dans le chapitre 6, qui pour rappel exploite les techniques d'intégration avancées telles que le word embedding pour calculer la distance sémantique entre des documents textuels, ainsi la distance entre deux documents texte A et B est calculée par la distance cumulative minimale que les mots du document texte A doivent parcourir pour atteindre les mots incorporés du document texte B.

L'autre technique de comparaison n'est autre la mesure TF*IDF Salton et McGill [1983] l'acronyme de «Term Frequency - Inverse Document Frequency», traduit en Français par «Fréquence de terme - Fréquence inverse de document», cette technique statistique est utilisée surtout en recherche d'information et en exploration de données pour quantifier des mots dans un ensemble de documents. Cette technique se base sur la fréquence de mot dans un texte qui est donnée par la loi Zipf Zipf [1949] : TF mesure l'importance d'un terme dans

un document, par contre IDF mesure si le terme est discriminant (ou non-uniformément distribué). Ainsi, un terme qui a une valeur de TF*IDF élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents. En règle générale, pour mesurer finement la similarité entre des séquences de texte, les vecteurs sont construits d'après un calcul de type TF IDF pour obtenir deux vecteurs à valeur réelle, et on calcule par la suite la similarité Cosinus Singhal [2001] qui est issue de l'algèbre linéaire pour quantifier la similitude entre deux vecteurs de dimensions n en déterminant l'angle entre eux. Soit deux vecteurs A et B a n dimensions, la similarité cosinus entre deux vecteurs A et B est calculée selon la formule suivante :

$$Similarite\_Cosinus(A,B) = A.B\frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} Ai.Bi}{\sqrt{\sum_{i=1}^{n} Ai}.\sqrt{\sum_{i=1}^{n} Bi}} \tag{7.1}$$

En gardant en tête que le but principal de notre approche est l'extraction de la meilleure pratique pour une requête donnée, il est nécessaire que notre système de recherche retourne les pratiques les plus proches sémantiquement de l'objectif recherché par l'utilisateur. Etant donnée ceci nous avons implémenté des algorithmes de recherche de pratiques similaires utilisant les 3 variantes de mesure de similarité à savoir : notre approche basée sur Word2Vec qui calcule le taux de similarité entre la requête de l'utilisateur et les bonnes pratiques existantes par leur distance sémantique, WMD et TF-IDF. Dans notre méthodologie d'évaluation nous lançons au hasard 3 requêtes qui portent sur le domaine de la santé sur les 3 SRI : la première s'intitule : «be healthy», la seconde « treat flu » et la dernière «manage stress» traduit respectivement par : «comment être en bonne santé», «comment traiter la grippe» et «comme gérer le stress», Pour chacune des trois requêtes, nous sélectionnons les 10 meilleurs résultats retournés par chaque méthode implémentée avec le taux de similarité entre la pratique retournée et la requête lancée. Les 10 premiers résultats obtenus sont présentés dans les tableaux suivants :

Approche utilisée	Taux de Similarité	Pratique retournée
W2V	0,83764699	Eating Healthy
	0,82523564	Eat Healthy at McDonalds
	0,82045749	Maintaining Healthy Habits and a Healthy Diet
	0,81725908	Being Healthy
	0,80938269	Getting Healthy
	0,80877639	Looking Healthy
	0,80531424	Keeping healthy
	0,79907151	Encouraging a Healthy Lifestyle
	0,79509763	Staying Healthy
WMD	0,71959318	Keeping healthy
	0,70518442	Treating the health problem
	0,69483534	How to Chewable Jewelry
	0,69010762	Recharge Mentally
	0,6887936	Hold the needle by its flat side using the needle holder
	0,68666016	Dealing with health challenges
	0,68608443	Eating Healthy for Eye Health
	0,68381789	Eat Well and Stay Healthy the Mediterranean Way
	0,6814472	Eat Healthy in College
	0,68033145	Meeting the Challenges of the Career
TF-IDF	0,73309878	Be Emotionally Healthy
	0,63373524	Be Yourself
	0,60620182	Be Healthy Without Dieting
	0,52221059	Be Calm
	0,45208903	Getting Healthy
	0,4457038	Be Hypnotized
	0,43886204	Eating Healthy for a Healthy Mind
	0,42509441	Be a Yoga Teacher
	0,41530484	Maintaining Healthy Habits and a Healthy Diet
	0,40844829	Eating Healthy

Table 7.2 – *Les 10 meilleurs résultats obtenus pour la première requête R1*

Approche utilisée	Taux de Similarité	Pratique retournée
W2V	0,93879732	Treating the Flu
	0,88598146	Treating the Flu with Supplements
	0,84002069	Treating a Cold or Flu at Home
	0,80490138	Diagnosing the Flu
	0,78100522	Getting the Flu Vaccine
	0,77854632	Understanding the Flu
	0,7711956	Identifying the Flu
	0,7579924	Preventing the Flu
	0,753766	Treating Flu Symptoms with Medicinal Remedies
	0,74978356	Deciding When to Get the Flu Vaccine
WMD	0,73948178	Meditate for Health
	0,73494954	Be a Health Nut
	0,722894593	Do a Butterfly Stretch
	0,71997842	Dealing with the Aftereffects
	0,71978031	Insertin²g the Catheter
	0,71931725	Treating the health problem
	0,71546335	Treating the Cut
	0,71481449	Start a Healthy Diet
	0,71299623	Treating the Flu
	0,7128484	Create the Charts
TF-IDF	0,705501486	Treating the Flu
	0,686500064	Understanding the Flu
	0,683075903	Preventing the Flu
	0,669423643	Identifying the Flu
	0,640187568	Diagnosing the Flu
	0,551265696	Treating the Flu with Supplements
	0,53290678	Recognizing Flu Symptoms
	0,525375984	Getting the Flu Vaccine
	0,47400423	Fighting the Flu with Food
	0,463912506	Preventing the Common Cold and Flu

TABLE 7.3 – *Les 10 meilleurs résultats obtenus pour la deuxième requête R2*

Approche utilisée	Taux de Similarité	Pratique retournée
W2V	0,90251938	How to manage stress
	0,86817032	Managing Stress
	0,86817032	Managing Anticipatory Stress
	0,82919288	Cooking to Help Manage Stress
	0,8218777	Managing Anxiety and Stress
	0,82086623	Managing Stress Naturally
	0,78440743	Managing Life Stress
	0,78203427	Managing Daily Stress
	0,77555298	Manage Stress Under Time Constraints
	0,77340661	Managing Your Stress
WMD	0,81355212	How to massage stance
	0,76852747	Mental Assessment
	0,75885291	Long-term strategies
	0,75872011	Assessing the Emergency
	0,75401504	Writing the Assessment
	0,75249832	How to manage stress
	0,73712977	Undergoing Assessments and Tests
	0,73594397	Inserting the Pessary
	0,735883	Using Water Treatments
	0,73484042	Assessing the Results
TF-IDF	0,778133643	How to manage stress
	0,691120393	Cooking to Help Manage Stress
	0,55938813	Manage Stress Under Time Constraints
	0,549763593	Use Relaxation Techniques to Manage Your Relationship Stress
	0,495079705	Manage Orthorexia
	0,466191111	Manage a Broken Arm
	0,45808205	Manage Osteoarthritis Pain
	0,453832431	Manage a Painful Injection
	0,44417183	Understanding Stress
	0,440813743	Preventing Stress

TABLE 7.4 – *Les 10 meilleurs résultats obtenus pour la troisième requête R3*

Afin de visualiser les résultats obtenus nous les avons schématisés dans les graphes suivants :

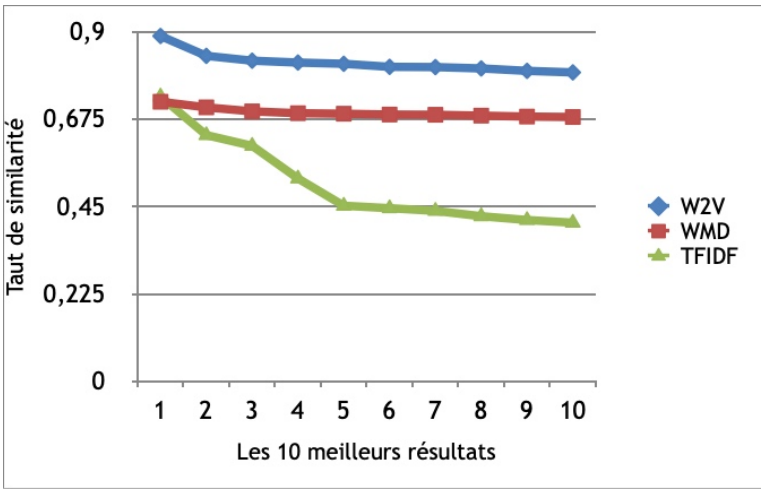

FIGURE 7.6 – *Courbes représentants les 10 meilleurs taux de similarité obtenus pour les 3 approches pour la première requête*

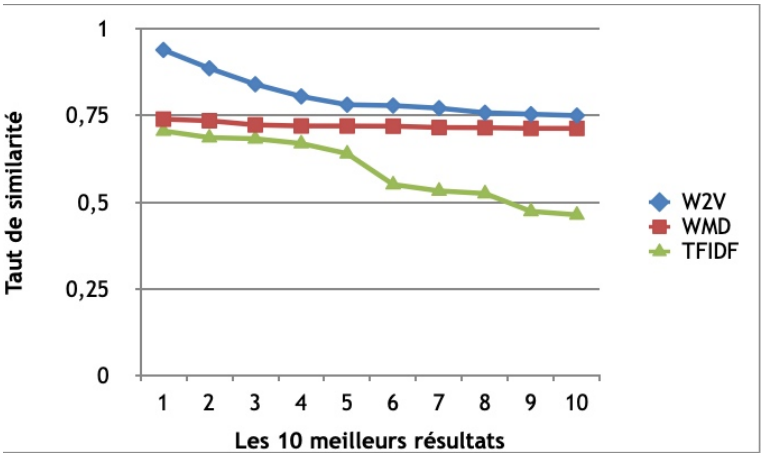

FIGURE 7.7 – *Courbes représentants les 10 meilleurs taux de similarité obtenus pour les 3 approches pour la seconde requête*

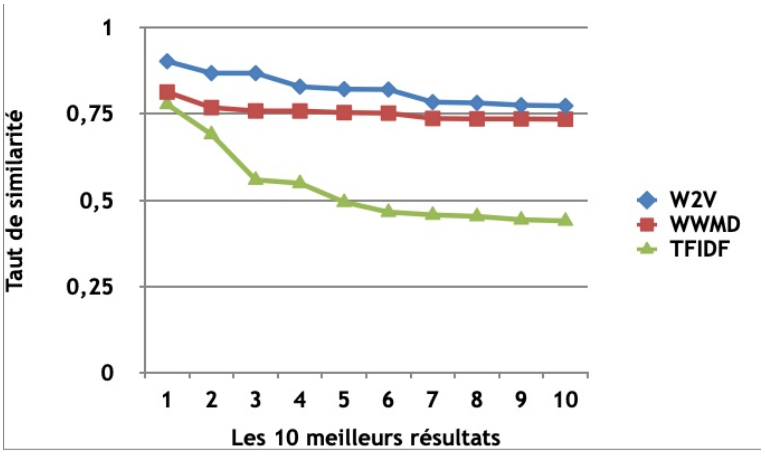

FIGURE 7.8 – *Courbes représentants les 10 meilleurs taux de similarité obtenus pour les 3 approches pour la troisième requête*

A partir des résultats obtenus nous observons que les taux de similarité des dix premiers résultats retournés par notre approche sont largement supérieurs à ceux obtenus avec TF-

IDF ou WMD, nous remarquons aussi à partir des tableaux 7.2, 7.3, et 7.4 que les pratiques retournées par notre approche sont plus pertinentes pour les requêtes lancées en comparaison avec WMD et TF-IDF. Prenons comme exemple la requête « treat the flu», l'idéal aurait été de retourner des pratiques portant sur comment traiter la grippe dans les premiers résultats sauf qu'on retrouve parmi les résultats retournés par WMD des éléments comme «Meditate for Health», «Be a Health Nut» ou encore «Do a Butterfly Stretch» qui ne sont pas en corrélation avec le but recherché, nous remarquons aussi dans le cas de WMD que la méthode la plus similaire à la requête «treating the flu» n'est classé qu'a la 9ème position, ce qui assez aberrant comme résultat. Par contre TF-IDF a tendance a retourner de meilleurs résultats que WMD mais avec un taux de similarité inférieur, on retrouve dans les deux dernières requêtes comme «treat the flu» et «manage stress» que les premiers résultats retournés par notre approche et TF-IDF sont identiques, mais à partir de là dans l'approche utilisant TF-IDF, les pratiques suivantes dans le classements commence a perdre de leur pertinence comme dans la second requête on retrouve respectivement en 2ème et 3ème position du classement les pratiques telles que «Preventing the flu», ou encore «identifing the flu» qui sont moins pertinentes pour la requête, par contre dans notre approche on retrouve les pratiques comme «Treating the Flu with Supplements» et «Treating a Cold or Flu at Home» en 2ème et 3ème position du classement qui sont plus pertinentes pour la requête lancée.
Un autre point que nous soulignons concerne la première requête que nous avons lancé qui est une recherche assez vaste c.à.d. qui n'est pas ciblé comme les deux requêtes suivantes mais là encore l'approche que nous proposons retourne des résultats en corrélation avec le but recherché on retrouve des pratiques comme «Eating Healthy» ou encore «Maintaining Healthy Habits and a Healthy Diet», «Being Healthy», «Encouraging a Healthy Life style», dans les premiers résultats, mais dans l'approche utilisant TF-IDF nous retrouvons dans les premiers classement des pratiques non similaires au but recherché comme «Be Hypnotized» «Be a Yoga Teacher» et de même dans WMD, on retrouve des résultats assez aberrants comme «How to Chewable Jewelry», « Meeting the Challenges of the Career».
A partir de ces observations nous pouvons affirmer la supériorité de notre approche dans la recherche des pratiques similaires, surtout pour capturer et retourner les résultats pertinents avec des taux de similarité supérieurs aux autres approches, car dans notre cas nous cherchons à extraire la meilleure pratique pour une requête donnée, donc nous cherchons a filtrer que les résultats pertinents comme nous l'avions spécifié nous sélectionnons que les résultats avec un taux de similarité supérieur à un seuil s=0.70.
Dans les tableaux suivants nous énumérons pour chaque requête le nombre d'éléments retrouvés par chaque approche par rapport au taux de similarité obtenu : qu'il soit supérieur à 70%, compris entre 0 et 70%, et égale a 0.

Approche	0 < Score de similarité <0.70	Score de similarité=0	Score de similarité=0.70
W2V	15088	207	85
WMD	0	15378	2
Tf-IDF	15037	343	1

Table 7.5 – *Nombre de résultats pour chaque seuil de score de similarité pour la première requête*

Approche	0 < Score de similarité <0.70	Score de similarité=0	Score de similarité=0.70
W2V	15191	160	25
WMD	0	15355	25
TF-IDF	13059	2319	2

Table 7.6 – *Nombre de résultats pour chaque seuil de score de similarité pour la seconde requête*

Approche	0 < Score de similarité <0.70	Score de similarité=0	Score de similarité=0.70
W2V	15211	138	29
WMD	0	15303	77
TF-IDF	15212	166	2

TABLE 7.7 – *Nombre de résultats pour chaque seuil de score de similarité pour la troisième requête*

Comme nous pouvons le constater quelque soit la requête WMD trouve des résultats avec un score de similarité supérieur à 0, ce qui nous confirme le peu de fiabilité dans la pertinence des résultats. Une autre constatation est que notre approche dans les deux premières requêtes retourne un corpus supérieur aux autres approches avec un score de similarité supérieur à 70%, sauf pour la troisième requête ou WMD la surpasse avec 77 éléments contre 29, mais la encore si nous jetons un coup d'œil au tableau 7.4 parmi les meilleurs résultats retournés par l'approche WMD, nous retrouvons beaucoup de pratique non pertinente pour la recherche effectuée comme «How to massage stance», «Writing the Assessment » ou encore «Assessing the Emergency» qui sont assez aberrante en comparaison avec la pertinence des pratiques sélectionnées par notre approche, donc cette dernière variation des résultat ne discrédite en aucun cas la supériorité de notre approche dans la recherche des pratiques similaires en terme de pertinence.
Pour finaliser cette première évaluation, nous rapportons dans le tableau suivant les temps d'exécution qui ont été nécessaires pour les trois approches pour répondre aux trois requêtes..

Approche	Temps d'exécution de R1	Temps d'exécution de R2	Temps d'exécution de R3
W2V	1,466084003 secondes	1,717098236 secondes	1,456083059 secondes
WMD	98,96766067 secondes	102,3228524 secondes	100,2217326 secondes
TF-IDF	1,899108648 secondes	1,992113829 secondes	2,011115074 secondes

TABLE 7.8 – *Temps d'exécution moyen des trois méthodes pour R1, R2 et R3*

Les graphiques représentés dans la figure 7.9 schématise les résultats obtenus dans le tableau 7.8, où l'on peut voir le temps d'exécution moyen pour chacune des méthodes nécessaires pour répondre à chaque requête.

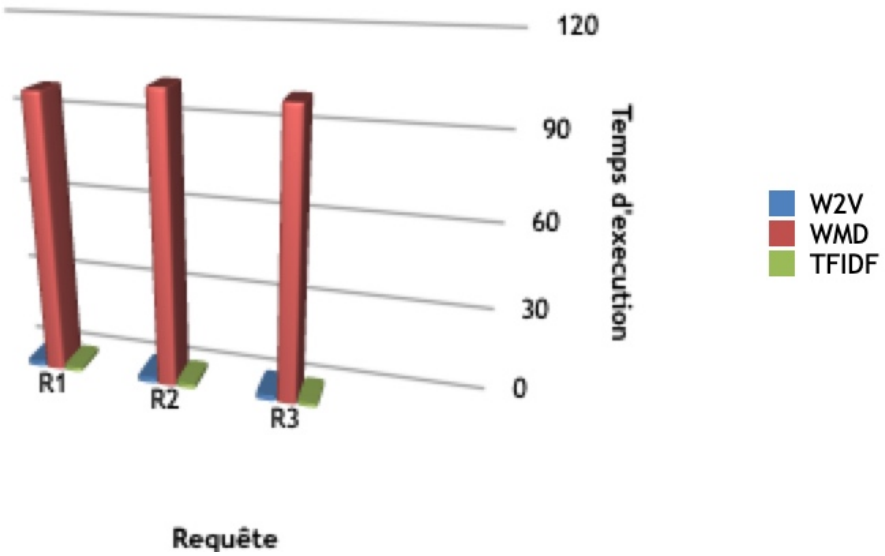

FIGURE 7.9 – *Graphiques présentants les temps d'exécution moyen pour les trois approches pour R1, R2 et R3*

Nous constatons clairement que le temps d'exécution nécessaire à l'approche WMD est beaucoup plus long que celui nécessaire à notre approche et l'approche utilisant TF-ID, il diffère en moyenne de 98,656 secondes de notre approche et de 98,537 de l'approche TF-IDF, donc même en comparant le temps d'exécution, notre approche et celle utilisant TF-IDF gagnent en optimalité quant au temps d'exécution des requêtes. Il est à noter aussi que TF-IDF est plus rapide de 0.421 secondes par rapport à notre approche ce qui est clairement négligeable en comparant l'écart calculé avec WMD et vu aussi la performance de notre approche quant à la pertinence des résultats.

Experimentation pour QR2

L'algorithme de clustering DBSCAN se base sur le principe assez simple que si un point a suffisamment de voisins récupérés arbitrairement il sera considéré comme appartenant à un cluster, autrement le point sera considéré comme du bruit. DBSCAN requiert donc 2 paramètres en entrée : le rayon de recherche du voisinage des points (Eps) ainsi que le nombre minimum de points qu'il doit y avoir dans ce voisinage pour que ceux-ci soient considérés comme un cluster (MinPts). Ces paramètres d'entrées correspondent donc à une estimation de la densité de points des clusters. Notre scénario d'évaluation vise à détecter l'influence des points traités, par les paramètres de DBSCAN. Les facteurs à modifier pour cette phase d'évaluation sont le nombre minimum de points pour constituer un cluster (MinPts) et le rayon de recherche du voisinage (Eps). Pour les trois requêtes R1, R2, R3 nous avons en premier lieu fixé MinPts à 5 point et nous avons fait varier epsilon. Les résultats sont présentés dans les tableaux 7.9, 7.10 et 7.11 :

Eps	Nombre de point traités	Nombre de Clusters
0.2	535	5
0.25	413	4
0.3	284	4
0.35	284	4

Table 7.9 – *Variation des paramètres Eps pour Min_Pts=5 pour R1*

Eps	Nombre de point traités	Nombre de Clusters
0.2	0	0
0.25	213	2
0.3	172	2
0.35	172	2

Table 7.10 – *Variation des paramètres Eps pour Min_Pts=5 pour R2*

Eps	Nombre de point traités	Nombre de Clusters
0.2	0	0
0.25	213	2
0.3	172	2
0.35	172	2

Table 7.11 – *Variation des paramètres Eps pour Min_pts=5 pour R3*

En second lieu nous avons inversé en fixant Eps à 0,25 pour R1 et R2 et R3, et en variant Min_pts. Les résultats sont présentés dans les tableaux ci-dessous :

MinPts	Nombre de point traités	Nombre de Clusters
2	337	13
5	413	4
10	468	4
15	568	2

TABLE 7.12 – *Variation des paramètres Min_pts pour Eps=0.2 pour R1*

MinPts	Nombre de point traités	Nombre de Clusters
2	137	7
5	156	1
10	159	1
15	0	0

TABLE 7.13 – *Variation des paramètres Min_pts pour Eps=0.2 pour R2*

MinPts	Nombre de point traités	Nombre de Clusters
2	181	8
5	213	2
10	0	0
15	0	0

TABLE 7.14 – *Variation des paramètres Min_pts pour Eps=0.2 pour R3*

A partir de ces tableaux nous avons tracer des courbes montrant le rapport entre les variations de des paramètres Eps et Min_Pts de DBSCAN sur le nombre de points traités et les nombre de cluster retourné :

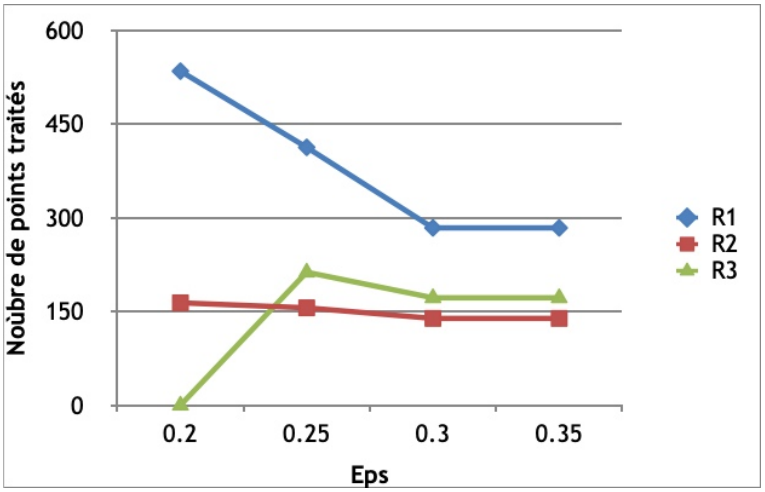

FIGURE 7.10 – *Influence d'Eps sur les nombre de points traités pour R1, R2, et R3*

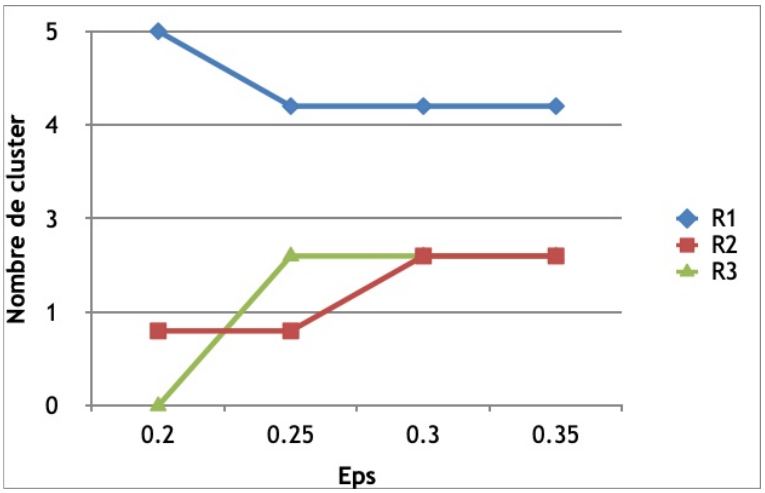

FIGURE 7.11 – *Influence d'Eps sur les nombre de clusters trouvés pour R1, R2, et R3*

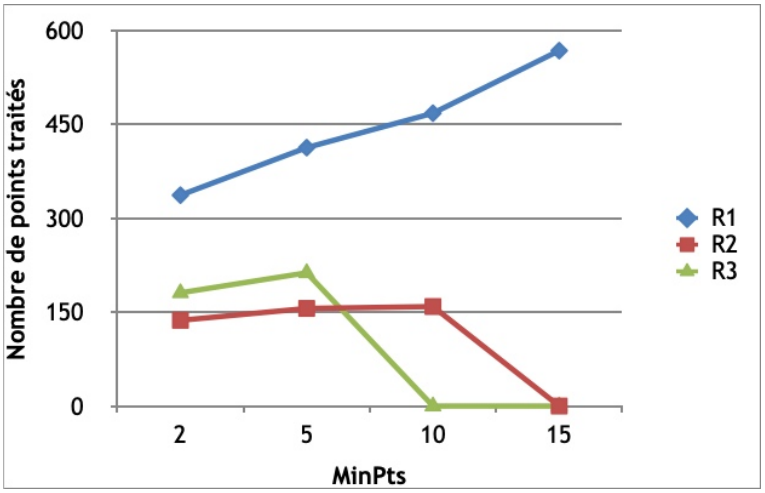

FIGURE 7.12 – *Influence de MinPts sur les nombre de points traités pour R1, R2, et R3*

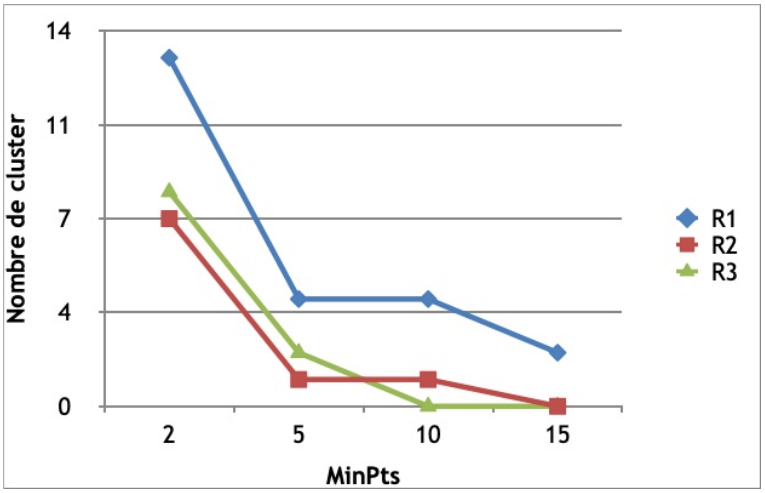

FIGURE 7.13 – *Influence de MinPts sur les nombre de clusters trouvés pour R1, R2, et R3*

Avant de commencer l'analyse des résultats, nous tenons à rappeler que le clustering avec DBSCAN sert a regrouper les étapes similaires sélectionnées auprès des bonnes pratiques retournées lors de l'étape précédente, donc pour les requêtes R1,R2,R3 le nombre de points

(étapes) à traiter diffère : 751 points pour R1, 175 points pour R2, et 223 points pour R3. Et c'est en ce sens que nous présenterons nos constatations au cas par cas : pour R2 et R3 le nombre de points traités et le nombre de clusters trouvés augmentent à chaque variation d'Eps par contre pour R1 on voit une diminution du nombre de clusters avec l'augmentation d'Eps. Nous constatons aussi que R1, R2 et R3 convergent au bout d'Eps=0,3, où l'on observe une stagnation des valeurs de points traités et du nombre de cluster. En faisant varier MinPts on remarque que pour R2 et R3 le nombre de points traités diminuent avec l'augmentation de MinPts, par contre dans R1 la courbe est exponentielle ce qui est dû au fait que le nombre de points à traiter pour R1 soit largement supérieur aux autres requêtes.
Une dernière remarque est que le nombre de clusters diminuent avec l'augmentation de MinPts pour les 3 requêtes ce qui est tout a fait normal.
Ces résultats nous montrent que DBSCAN est très influencé par le changement de ses paramètres. Le nombre de points à classifier influe aussi sur la valeur de Min_Pts, et plus la valeur de cette dernière augmente plus le nombre de clusters trouvés diminue ce qui est tout à fait logique. La variation du rayon de recherche (Eps) agit énormément sur le nombre de clusters et points isolés formés. Ces derniers se stabilisent et stagnent avec l'augmentation d'Eps, ce qui est cohérent.

Experimentation pour QR3

Dans cette partie de l'évaluation, nous cherchons à mesurer la qualité de notre partitionnement, pour ce faire nous avons fait appel au coefficient de silhouette appelé aussi le score silhouette Rousseeuw [1987]. Cette métrique calcule la qualité d'une partition d'un ensemble de données dans le domaine de la classification automatique. Elle se base sur le Silhouette Score (SS) de chaque point individuel de données, et permet de voir si un échantillon appartient effectivement au cluster qui lui a été assigné, par rapport aux autres clusters. Le score silhouette varie entre -1 et 1 :

- 1 : Signifie que les groupes sont bien séparés les uns des autres et clairement distingués.
- 0 : signifie que les clusters sont indifférents, ou on peut dire que la distance entre les clusters n'est pas significative.
- -1 : signifie que les clusters sont attribués dans le mauvais sens.

Dans le contexte de l'expérimentation, cette mesure permet de comparer DBSCAN avec une autre ligne de code et d'étudier l'impact de la métrique de similarité sur notre algorithme de partitionnement.
Tout algorithme de clustering est sensible au choix de la métrique de similarité utilisée : cette dernière est exprimée en termes d'une fonction de distance D(a ;b) tel que D(a,b)<D(a,c) si les objets a et b sont considérés plus similairement proches que les objets a et c. Dans le cas de DBSCAN la métrique utilisé par default est la distance euclidien aucun paramètrage n'est nécessaire alors.
Pour notre expérimentation nous avons implémenté DBSCAN avec deux métrique différentes pour voir l'impact qu'il peut y avoir sur la qualité de clustering. En premier lieu nous avons paramètré DBSCAN avec la métrique pré-calculée «precomputed» et qui prend comme entrée la matrice carrée de distance ou de dissimilarité qui n'est autre que la matrice inverse de la matrice de similarité entre les étapes sélectionnées pour chaque requête que nous calculons avec le modèle Word2Vec. Comme nous l'avions précisé dans le chapitre 6 cette matrice nous retourne les scores de similarité entre les étapes, de sorte que nous ne considérions que les valeurs supérieures à 0.70. En second lieu nous paramètrons DBSCAN avec le métrique cosinus «Cosine» qui calcule la distance entre deux points vectorisées grâce à la mesure TF-IDF Singhal [2001].
Nous avons par la suite comparé DBSCAN à un autre algorithme de classification non supervisé Kmeans MacQueen [1967] appelé aussi K-moyenne. Cet algorithme nécessite de choisir à l'avance le nombre de clusters K pour faire le partitionnement des données. Ensuite, k points

sont choisis semi aléatoirement comme centre des clusters. Toutes les instances sont assignées au centre le plus proche d'eux.
Afin de répondre à l'évaluation QR3, nous avons calculé le score silhouette de chaque algorithme pour les 3 requêtes. Les résultats sont présentés dans le tableau 7.15 :

Clustering	SS(R1)	SS(R2)	SS(R3)
DBSCAN (metric precomputed)	0,649	0,554	0,522
DBSCAN (metric cosine)	0,351	0,243	0,241
Kmeans	0,565	0,149	0,329

TABLE 7.15 – *Le Score silhouette pour chaque méthode de clustering pour les requêtes R1, R2 et R3*

Les résultats obtenus sont reporté dans la figure suivante :

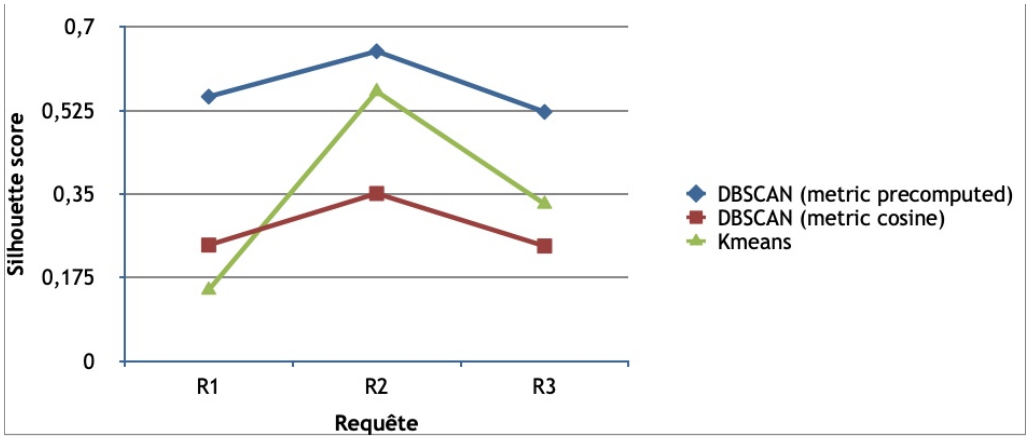

FIGURE 7.14 – *Variation des scores silhouette pour chaque méthode de clustering pour les requêtes R1, R2 et R3*

La première constatation que nous pouvons faire est que notre approche surpasse les autres en termes de qualité de clustering. En effet, on arrive à des scores de 0.649 pour la deuxième requête ce qui est assez bon, il est vrai que dans la première requête et la troisième on est à un score avoisinant les 50% mais ça reste largement supérieur aux autres, surtout que la qualité du clustering dans le cas des données textuelles est un peu spécifique puisque les données ne sont pas décrites par des variables quantitatifs mais plutôt qualitatifs Une autre constatation est que l'algorithme Kmeans dans les 2 dernières requêtes surpasse celui de DBSCAN utilisant la métrique Cosinus. Nous pouvons confirmer ainsi que la métrique utilisée impacte fortement la qualité du partitionnement DBSCAN, et dans cette expérimentation DBSCAN avec la métrique pré-calculée a surpassé Kmeans. Par ailleurs on remarque aussi que les scores silhouettes pour R1 et R3 sont au plus bas pour les trois partitionnements et sont au plus haut pour R2, et ceci est dû peut être au fait que dans ces requêtes il y'est un chevauchement entre les clusters, ainsi les classes obtenues sont très proches les une des autres et que les étapes présentent une confusion au niveau de leur classification.

Experimentation pour QR4

Cette dernière partie de l'évaluation porte sur l'identification de la meilleure pratique, et comme nous l'avions déjà précisé, il n'y a pas de manière précise de déterminer la supériorité d'une façon de faire par rapport à d'autres, de plus nous ne possédons aucun modèle d'évaluation prédéfinie, vérifier alors la cohérence des résultats est assez délicat.
Dans notre scénario d'évaluation, nous allons comparer les résultats des extractions des meilleures pratiques obtenues pour les 3 requêtes R1, R2 et R3 avec notre approche, ceci en parallèle avec les résultats obtenus en lançant ces même requêtes sur la plateforme WikiHow.

Dans les figures ci-dessous on retrouve l'évolution du graphe de sélection des bonnes pratiques pour R1, R2 et R3, au moment du lancement de chaque requête, ensuite lors de la sélection des pratique similaires, et enfin après la fusion des étapes similaires.

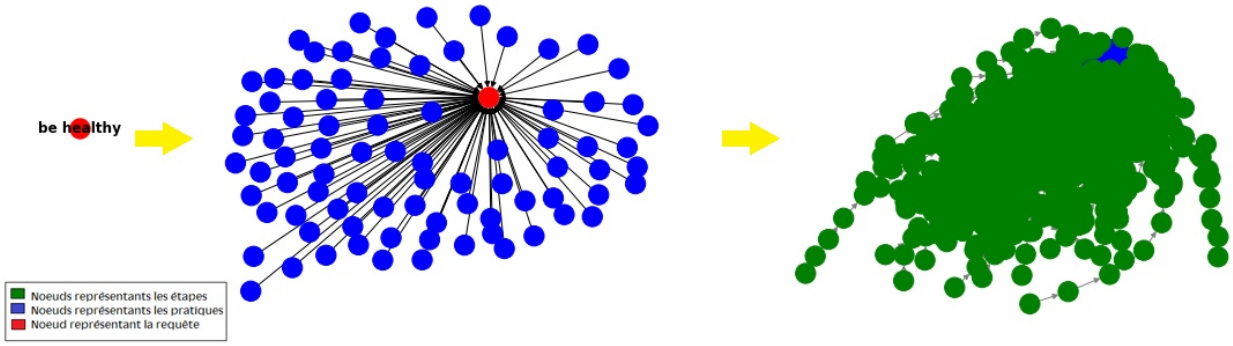

FIGURE 7.15 – *Evolution du graphe des bonnes pratiques pour R1*

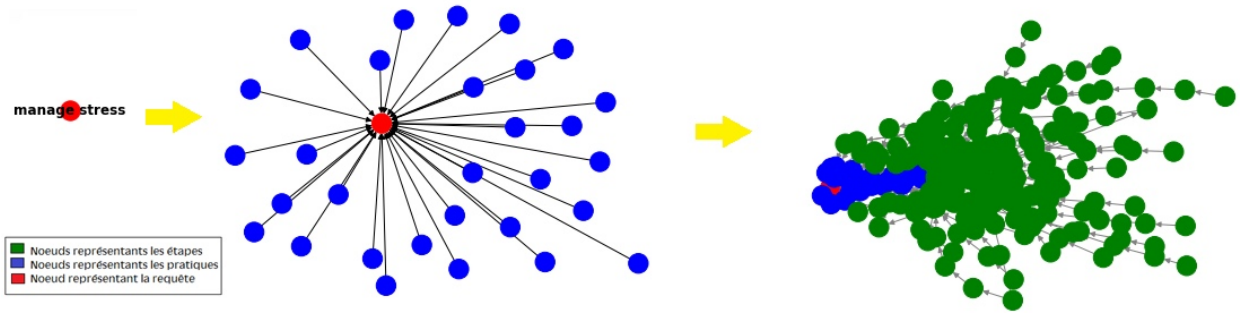

FIGURE 7.16 – *Evolution du graphe des bonnes pratiques pour R2*

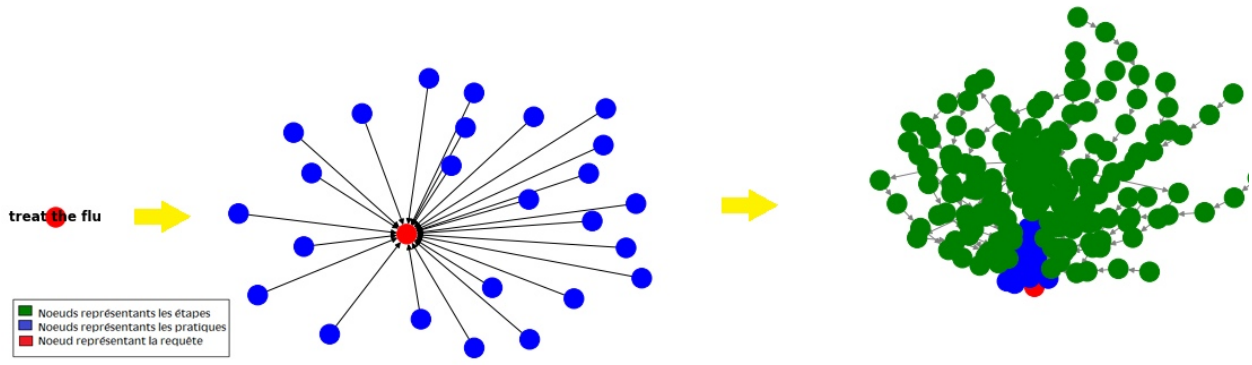

FIGURE 7.17 – *Evolution du graphe des bonnes pratiques pour R3*

Tout d'abord, nous constatons que l'approche proposée fonctionne correctement quant à la conceptualisation des bonnes pratiques sous forma de graphe de données Pour chaque requête on voit le graphe se modéliser autour du nœud initial représentant la requête qui correspond à l'objectif recherché (nœud en rouge sur les figures), d'abord les nœuds représentants les bonnes méthodes similaires à la requête sont connectés à l'objectif (nœuds en bleu sur les figures 7.15, 7.16, et 7.17), ensuite après la fusion des étapes similaires les nœuds représentant les étapes sont reliés par des arcs entrants (nœuds en vert sur les figures 7.15, 7.16, et 7.17).
A partir de ces représentations notre approche extrait la meilleure pratique pour chaque requête, en identifiant parmi tous les chemins existants dans le graphe et qui mènent à chaque

objectif recherché ceux qui empruntent les étapes les plus utilisées, comme nous l'avions mentionné ces nœuds sont caractérisés par une centralité de degré entrante élevée qui traduit leur importance dans le graphe. Afin d'identifier les meilleures pratiques, nous avons donc calculé grâce à l'équation 6.1 la valeur de l'importance de chaque chemin qui est égale au rapport entre la somme des degrés entrants des nœuds et le nombre total de ces mêmes nœuds dans chaque chemin identifié. Les meilleures pratiques obtenues sont reportées dans les tableaux 7.16, 7.17 et 7.18 ci-dessous.
Par la suite pour parfaire notre évaluation nous avons lancé les trois requêtes sur le site Wiki-How, les figures ci-dessous montrent les résultats des articles qui sont apparus au lancement de R1, R2 et R3 :

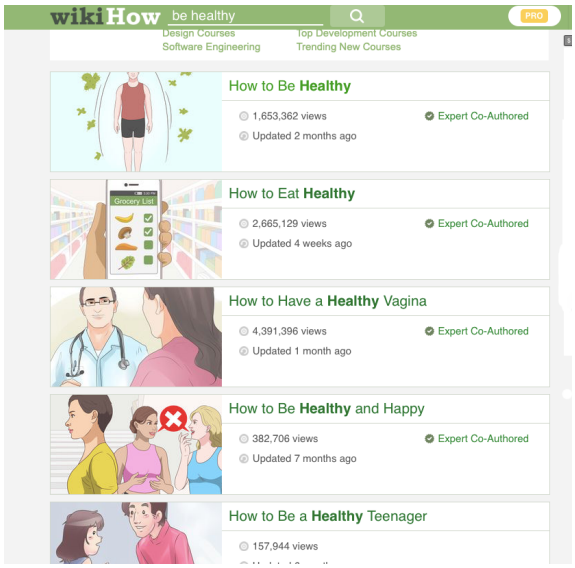

FIGURE 7.18 – *Résultats de la requête R1 sur le site Wikihow*

la requête "how treat the flu"

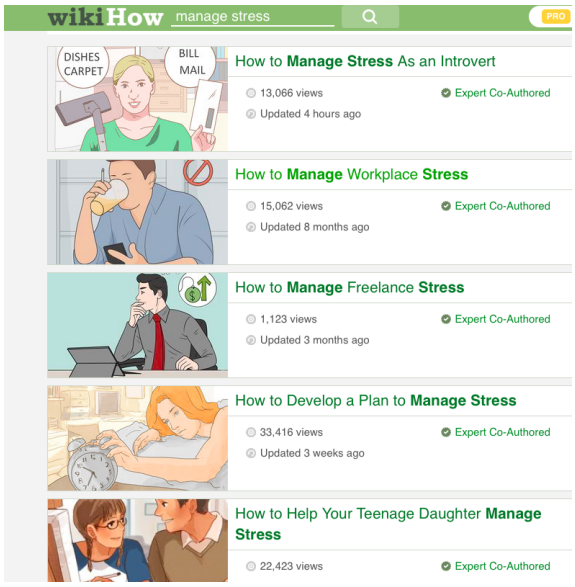

FIGURE 7.19 – *Résultats de la requête R3 sur le site Wikihow*

Comme nous l'avions spécifié WikiHow est regroupement d'un ensemble d'article chaque article est constitué d'un ensemble de méthodes, en lançant des requêtes sur le site les articles les plus proches de nos requêtes sont apparus en premier, pour notre évaluation nous allons sélectionner les articles en tête de liste, et par la même occasion l'ensemble des bonnes pratiques qui y figurent, ce qui fera office pour nous du meilleur résultat retourné par WikiHow pour les requêtes R1, R2, et R3. les résultats obtenus sont présentés dans les tableaux 7.16, 7.17 et 7.18 ci-dessous. :

Approche	Résultats
Mèthode proposée	**Best Practice : Staying Healthy** Talk to people who are older than you. Exercise regularly. Eat a healthy diet. Quit smoking. Visit your doctor. Stay on top of personal hygiene.
WikiHow	**Best Practice 1 : Having a Healthy Diet** Drink more water Eat breakfast Eat well Eat at the right times Consider going meatless at least a few days a week Limit or completely eliminate simple sugars in your diet. Eat food when they're in season Read food labels to make the healthiest choices Talk to your doctor about incorporating supplements into your diet Use intermittent fasting to control calories and boost endurance
	Best Practice 2 : Having a Healthy Exercise Plan Get in shape Maintain a healthy weight. Cross train Exercise wisely Take advantage of opportunities to be active.
	Best Practice 3 : Being Emotionally Healthy Think positively Be satisfied and practice gratitude daily Think small. Manage stress Get outside Choose your friends wisely Be productive Take a break Find emotional balance. Include the arts in your life, such as music, theater, and visual arts. Travel as much as you can.
	Best Practice 4 : Having a Healthy Routine Create a daily routine Stop engaging in risky behavior. Exercise several times a week Get a good night'rest Learn how to cook Maintain your personal hygiene Bolster your immune system.

TABLE 7.16 – *Meilleures pratiques retournées par Wikihow et notre approche pour R1*

Approche	Résultats
Mèthode proposée	**Best practice : Treating the Flu** Rest. Stay hydrated and avoids cigarettes and alcohol. Fight a low fever with OTC medication Take cold medication for other symptoms. Identify a dangerous fever based on age Watch for warning signs See a doctor early if you are at risk of complications.
WikiHow	**Best practice 1 : Identifying the Flu** Recognize the symptoms of the flu. Distinguish between the flu and a cold. Distinguish between the flu and a stomach bug. Know when to seek emergency medical treatment.
	Best practice 2 : Treating Flu Symptoms with Natural Remedies Get some rest. Stay hydrated. Take a vitamin C supplement Clear mucus from your nose often Use a heating pad Relieve fever symptoms with a cool cloth. Gargle with saltwater Try an herbal remedy to relieve your symptoms. Try a eucalyptus steam treatment.
	Best practice 3 : Taking Medication to Treat Your Symptoms Buy over-the-counter medicine to treat symptoms. Give children the correct dosage Take prescription medication as directed. Understand that antibiotics will not treat the flu.
	Best practice 4 : Preventing the Flu Get vaccinated before flu season Talk to your doctor before getting the vaccine if you have certain conditions.‘ Choose between the flu shot and the nasal spray vaccine. Practice good hygiene. Keep your body in good general health. Take the flu seriously.

Table 7.17 – *Meilleures pratiques retournées par Wikihow et notre approche pour R2*

Approche	Résultats
Mèthode proposée	**Best Practice : : Reducing Stress Say no sometimes** Take care of yourself Try meditation Keep a journal
WikiHow	**Best Practice 1 : Getting Personal Space** Take time to process things internally. Be assertive Locate a place that makes you feel safe and at peace. Unplug from your phone on the internet Take a walk in nature Go on a solo vacation
	Best Practice 2 : Doing Relaxing Activities Practice deep breathing Problem-solve with journaling exercises Listen to soothing music Practice meditation Get some rest
	Best Practice 3 : Reducing Stressors Set and enforce personal boundaries with others. Prepare in advance for activities that deplete your energy. Practice batching work tasks and personal chores

TABLE 7.18 – *Meilleures pratiques retournées par WikiHow et notre approche pour R3*

A partir de ces résultats, nous pouvons constater en premier lieu que les meilleures pratiques extraites par notre approche sont tout à fait cohérentes quant aux requêtes émises ce qui déjà un bon résultat.
Nous constatons par ailleurs que les bonnes pratiques contenues dans l'article en tête de liste du site Wikihow diffèrent les une des autres, chacune traite d'un aspect de la requête, contrairement aux meilleures pratiques que nous avons extraite qui traitent le sujet de la requête dans sa globalité, comme par exemple en réponse à R1 notre approche extrait la pratique «Staying Healthy» et Wikihow revoie les pratiques par ordre de classement «Having a Healthy Diet», «Having a Healthy Exercise Plan», «Being Emotionally Healthy», «Having a Healthy Routine» qui sont différentes et répondent chacune à un aspect de la requête R1.
Certaines meilleures pratiques renvoyées par Wikihow contiennent des étapes parfois redondantes comme dans le premier résultat de la requête R1 dans la bonne pratique «Having a healthy diet» on retrouve des étapes telles que «eat well» , «eat at the right times» ou encore «eat food when they're in season» qui peuvent être regroupées ensemble. Par contre la meilleure pour R1 pratique extraite par notre approche est plus optimale car elle englobe plusieurs facettes du sujet de la requête sans se répéter pour autant. En comparant les résultats obtenus pour la requête R2 la meilleure pratique extraite par notre approche «Treating the Flu» répond précisément à la question posée «how treat the flu» même les étapes qu'elles contient sont cohèrente par rapport au sujet traité par contre les meilleurs résultats retournés dans Wikihow certes ont une relation avec la requête recherchée mais dérivent là encore un peu du fondement de cette dernière, on retrouve alors les réponses par ordre de classement comme «Identifying the Flu», «Treating Flu Symptoms with Natural Remedies», et «Taking Medication to Treat Your Symptoms Preventing the Flu».
La meilleure pratique «Reducing stress» extraite en réponse à R3 reste aussi tout a fait correcte et traite aussi la requête dans sa globalité, par contre là aussi les meilleures pratiques

retournées par Wikihow «Getting Personal Space», «Doing Relaxing Activities» et «Reducing Stressors» représentent de bons résultats mais ne couvrent pas l'aspect principal de la requête. Nous pouvons ainsi conclure suite a ces constatations que notre approche arrive a extraire pour chaque requête la meilleure pratique de manière pertinente par rapport au sujet recherché et regroupe aussi les étapes les plus importantes que doit suivre chaque processus pour atteindre son objectif sans être redondant et sans dériver du but recherché pour autant.

7.5 Conclusion

Dans ce chapitre, nous avons décrit la mise en œuvre de l'approche proposée à savoir la conceptualisation des bonnes pratiques et l'extraction de la meilleure pratique pour une requête donnée. Nous avons présenté les outils, les langages et la base de données que nous avons utilisée lors de notre l'expérimentation.
Cette expérimentation que nous avons menée a englobé chaque partie de l'approche proposée : du scrapping des données web, à la recherche des pratiques similaires à des requêtes posées, jusqu'au regroupement des nœuds similaires et l'extraction des meilleures pratiques. Les résultats obtenus lors de chaque évaluation ont montré la supériorité de notre approche à chaque niveau, par exemple lors de la comparaison de notre approche de recherche par rapport aux approches utilisant les mesures de similarités WMD ou TF-IDF ou encore dans le score silhouette obtenu par notre algorithme de clustering DBSCAN, nous avons pu confirmer l'importance du choix des mesure de similarité telle que Word2Vec que nous avons utilisé influe largement sur la pertinence des algorithmes de recherche d'information aussi nous avons constaté que la qualité du clustering dépend largement du choix de sa métrique de calcul de similarité et de ses paramètres (Eps et MinPoints). La dernière évaluation menée dans ce chapitre montre que notre approche réussit à extraire des pratiques pertinentes pour des requêtes émises et qui englobent des étapes importantes quant au but recherché par rapport aux meilleures pratiques retournées par Wikihow.
Nous pouvons conclure au terme de ce chapitre que nous avons réussit à mettre en œuvre l'approche que nous avons proposée, nous avons ainsi confirmé sa validité par notre expérimentation. Arrivant à la fin de ce mémoire nous présenterons dans le chapitre suivant les conclusions générales déduites, et les perspectives possibles quant à notre travail

Conclusion générale

«Il est facile de manquer le but et difficile de l'atteindre »
Aristote

Communauté de pratiques, bonnes pratiques, meilleure pratique : tant de concepts émergeants de la sociologie et difficilement interprétables en logique et informatique, sauf que l'essor du web actuel, le développement de ses outils et de ses plates formes ont fait que ces concepts ont trouvé tout leur essor dans l'ère actuelle, où l'on voit des groupes de personnes se réunissant autour d'un sujet, d'un axe de recherche quel qu'il soit (culinaire, artistique, scientifique, etc.) afin de favoriser le partage d'un savoir faire et l'apprentissage entre les membres de ces communautés.
En commençant cette thèse beaucoup ont suggéré que la problématique été trop vaste, l'abstrait y été plus que le concret, mais en analysant les choses de plus prêt on ses retrouve face à un problème d'actualité, qui est de conceptualiser les meilleures pratiques au sein d'une communauté de pratique. Ces bonnes pratiques sont définies comme des connaissances procédurales qui présentent de manière structurée sous forme d'étapes successives un savoir faire pour réaliser une tâche précise. En effet, il est constaté que l'utilisateur du web actuel cherche au delà des informations élémentaires davantage des connaissances d'un niveau cognitive plus élevé : à titre d'exemple un étudiant en informatique cherchera des savoirs faire explicites pour apprendre à programmer, un amateur en cuisine cherchera la meilleure façon de faire de bon plats ou gourmandises, etc. Le web est ainsi devenu un recueil mondial de connaissances procédurales, néanmoins explorer ces dernières est une tâche ardue du fait que leur format est généralement non structurées.
Dans cette perspective, nous avons présenté dans cette thèse une nouvelle approche pour conceptualiser les bonnes pratiques et extraire la meilleure pratique pour atteindre un but recherché.

Synthèse

Dans la première partie de cette thèse nous avons présenté un état de l'art en partant de la problématique, nous avons d'abord abordé le domaine des communautés de pratiques qui est au coeur de notre thématique de recherche.
Nous avons par la suite passer en revue les principes d'Extraction des Connaissances à partir des bases de Données (ECD), en effet l'ECD est le domaine qui évolue pour offrir des solutions afin d'explorer les connaissances procédurales issues des communautés de pratiques.
Nous avons aussi ouvert un chapitre sur un autre domaine offrant une solution à notre problématique à savoir la représentation des connaissances, car le choix d'un bon formalisme de représentation est primordiale dans la tâche d'exploration des données et dans notre cas l'utilisation des principes de théorie de graphe à été très avantageux dans notre recherche de la meilleure pratique.
A la fin de la l'état de l'art nous avons passé en revue les travaux de recherche actuels connexes à notre thématique et nous avons présenté une étude comparative entre ces derniers.
Dans la deuxième partie de cette thèse, nous avons présenté une nouvelle approche pour conceptualiser les bonne pratiques et extraire la meilleure méthode pour atteindre un but re-

cherché. Notre approche se déroule en deux phase, durant la première on extrait les connaissances procédurales du web grâce aux techniques du web scraping et on les formalise par des graphes de données, où les étapes sont représentés par des noeuds reliés par des arcs orientés jusqu'a atteindre un noeud final qui est le nom de pratique.
Dans la seconde phase on extrait la meilleure pratique pour répondre à une requête d'un utilisateur. Cette phase repose sur un pipeline d'étapes et utilise différentes technologies : tout d'abord on commence par rechercher les bonnes pratiques similaires à la requête grâce au modèle de prolongement lexicaux de mots "Word2Vec" Mikolov et al. [2013]. Ensuite nous avons procède au regroupement des étapes similaires grâce à l'algorithme d'apprentissage non supervisé DBScan Ester et al. [1996]. Après nous avons fait appel aux techniques de synthèse de texte notamment au célèbre algorithme de ranking Brin et Page [1998] afin de fusionner les étapes similaires dans chaque cluster obtenu. Et pour finir nous avons fait appel aux notions de centralité dans la théorie des graphes pour identifier la meilleure pratique.
Pour finir nous avons démontré la faisabilité de notre approche sur des connaissances réelles extraites du site de partage WikiHow. L'utilisation des graphes pour modéliser les connaissances procédurales s'est avéré très avantageuse en terme de flexibilité mais aussi en vue des possibilités qu'offre un tel formalisme pour identifier l'importance des chemins représentants les meilleures pratiques grâce aux différentes mesures de centralité des graphes. Les résultats de l'expérimentation obtenus ont démontré la supériorité du modèle Word2Vec pour la recherche des pratiques similaires à une requête, nous avons constaté que cette métrique arrive a capturer les scores de similarité élèves par rapport aux mesures de similarité WMD, et TF-IDF et aussi que les résultats retournés par notre approche étaient plus pertinents. Nous avons aussi constaté que DBSCAN est sensible au choix de la métrique utilisée, les scores silhouettes obtenus pour les différentes lignes de codes ont démontré que le clustering utilisé est meilleur que les autres.

Perspectives

Dans la continuité directe de notre travail de thèse, nous pensons que cela ouvre des perspectives à plusieurs voix de recherches en vue de :

- enrichir les bases des connaissances procédurales et diversifier leurs domaines(informatique, cuisine, etc) car telles bases sont très utiles notamment en recherche d'information pour l'exploration des graphes du web et aussi dans les applications automatiques actuelles tels que SIRI ou ALEXA.
- étendre la recherche des bonnes pratiques pour une requête non pas seulement aux phrases décrivants les méthodes mais aussi aux étapes qu'elles englobent afin de relever les connaissances véhiculées dans la pratique dans leur globalité en vue de capter des savoir faire encore plus pertinents.
- Explorer d'autres techniques d'apprentissage pour valider l'approche tels que le shot learning.

Mes contributions scienfifiques

— Hamza-Cherif, Souaad., Chikh, Azzedine. Procedural knowledge mining - A new method for extracting best practices by applying machine learning on data graph. Revue d'Intelligence Artificielle RIA,Vol. 36, No. 2, April, 2022, pp. 297-304 : -. https ://doi.org/10.18280/ria.360214

Bibliographie

Emmanuel Adam. Intelligence artificielle apprentissage par renforcement. Universit e de Valenciennes et du Hainaut-Cambr esis.http://emmanuel.adam.free.fr/site/IMG/pdf/iacollective_apprentissage.pdf. Page consultée le 6 Janvier 2022, 2015.

Rakesh Agrawal, Tomasz Imieliński, et Arun Swami. Mining association rules between sets of items in large databases. Dans *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, page 207–216, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915925. URL https://doi.org/10.1145/170035.170072.

Julien Ah-Pine. Méthodes avancées en apprentissage supervisé et non supervisé. Universit e Lyon 2. http://eric.univ-lyon2.fr/~jahpine/cours/m2_sise-dm/cm.pdf. Page consultée le 6 Janvier 2022, 2020.

S. S. Anand, A. R. Patrick, J. G. Hughes, et D. A. Bell. A data mining methodology for cross-sales. 10(7) :449–461, may 1998a. ISSN 0950-7051. URL https://doi.org/10.1016/S0950-7051(98)00035-5.

S.S. Anand, A.G. Büchner, et Financial Times Management. *Decision Support Using Data Mining*. Financial Times management briefings : Information technology. Financial Times Management, 1998b. ISBN 9780273632696. URL https://books.google.dz/books?id=U12YtAEACAAJ.

Jamal Atif. Analyse et fouille de données. Université Paris-Dauphine. https://www.lamsade.dauphine.fr/~atif/lib/exe/fetch.php?media=teaching:coursafd_ch1.pdf. Page consultée le 6 Janvier 2022, 2015.

Avron. Barr et Edward A. Feigenbaum. Dans *The Handbook of artificial intelligence*, volume 1. Stanford, Calif. : HeurisTech Press, 1981.

Djamel Belhaouci. Démystifier le machine learning, partie 2 : les réseaux de neurones artificiels. Juri'Predis. https://www.juripredis.com/fr/blog/id-19-demystifier-le-machine-learning-partie-2-les-reseaux-de-neurones-artific Page consultée le 6 février 2021, 2020.

Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Twente, Netherlands, Septembre 1997.

George E P. Box et Norman Richard. Draper. *Empirical model-building and response surfaces / George E. P. Box, Norman R. Draper*. Wiley series in probability and mathematical statistics Applied probability and statistics. J. Wiley Sons, New York Chichester Brisbane [etc, 1987. ISBN 0-471-81033-9.

Sergey Brin et Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1) :107–117, 1998. ISSN 0169-7552. Proceedings of the Seventh International World Wide Web Conference.

John A. Bullinaria. Iai : Knowledge representation. School of Computer Science University of Birminghaml. https://www.cs.bham.ac.uk/~jxb/IAI/w5.pdf. Page consultée le 6 Janvier 2022, 2005.

Cuong Xuan Chu, Niket Tandon, et Gerhard Weikum. Distilling task knowledge from how-to communities. Dans *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 805–814, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. URL https://doi.org/10.1145/3038912.3052715.

CNRTL. *Analyse de systèmes distribué : Projets*. Centre National de Ressources Textuelles et Lexicales. https://www.cnrtl.fr/. Page consultée le 6 février 2021, 2012.

Coheris. Qu'est ce que la modélisation prédictive. Coheris . https://ia-data-analytics.fr/modelisation-predictive. Page consultée le 6 Janvier 2022, 2020.

Antoine Cornuéjols, Miclet Laurent, et Jean-Paul Haton. *Apprentissage artificiel*. Algorithmes. Eyrolles, Paris, 2e édition édition, 2010. ISBN 978-2-212-12471-2.

Michel Crucianu, Marin Ferecatu, Nicolas Thome, et Nicolas Audebert. Cours - arbres de décision. Cnam-UE RCP209. http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html. Page consultée le 2 Janvier 2021, 2020.

DataScientist. Apprentissage supervisé vs. non supervisé. DataScientist. https://le-datascientist.fr/apprentissage-supervise-vs-non-supervise. Page consultée le 10 octobre 2020, 2019.

Dictionary. *Analyse de systèmes distribué : Projets*. Press Cambridge Unversity https://dictionary.cambridge.org/fr/dictionnaire/anglais/best-practice. Page consultée le 6 février 2020, 2020.

Müller Didier. *Introduction à la théorie des graphes*, volume 6. Cahiers de la Commission Romande de Mathématique CRM, 2012.

Richard Durbin, Sean R. Eddy, Anders Krogh, et Graeme Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, et Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Dans *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, et Mausam Mausam. Open information extraction : The second generation. Dans *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, page 3–10. AAAI Press, 2011. ISBN 9781577355137.

Usama Fayyad, Gregory Piatetsky-Shapiro, et Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37, Mar. 1996. URL https://doi.org/10.1609/aimag.v17i3.1230.

Ronen Feldman, Yonatan Aumann, Moshe Fresko, Orly Liphstat, Binyamin Rosenfeld, et Yonatan Schler. Text mining via information extraction. Dans Jan M. Zytkow et Jan Rauch, éditeurs, *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99, Prague, Czech Republic, September 15-18, 1999, Proceedings*, volume 1704 de *Lecture Notes*

in Computer Science, pages 165–173. Springer, 1999. URL https://doi.org/10.1007/978-3-540-48247-5_18.

Wenfeng Feng, Hankz Hankui Zhuo, et Subbarao Kambhampati. Extracting action sequences from texts based on deep reinforcement learning. Dans *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4064–4070. International Joint Conferences on Artificial Intelligence Organization, 7 2018. URL https://doi.org/10.24963/ijcai.2018/565.

William J. Frawley, Gregory Piatetsky-shapiro, et Christopher J. Matheus. Knowledge discovery in databases : an overview, 1992.

Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1 (3) :215–239, 1978. ISSN 0378-8733. URL https://www.sciencedirect.com/science/article/pii/0378873378900217.

Aurélien Garivier. *Analyse de systèmes distribué : Projets*. Université de Toulouse. https://www.math.univ-toulouse.fr/~agarivie/mydocs/apprentissageSupervise.pdf. Page consultée le 6 février 2021, 2013.

Daniel Glez-Peña, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato, et Florentino Fdez-Riverola. Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15 (5) :788–797, 04 2013. ISSN 1467-5463. URL https://doi.org/10.1093/bib/bbt026.

Asunción Gómez-Pérez, Mariano Fernández-López, et Óscar Corcho. Ontological engineering : With examples from the areas of knowledge management, e-commerce and the semantic web. Dans *Advanced Information and Knowledge Processing*, 2004.

Google. word2vec. .google code. https://code.google.com/archive/p/word2vec. Page consultée le 6 Janvrier 2021, Montréal 2013.

Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2) :199–220, 1993. ISSN 1042-8143. URL https://www.sciencedirect.com/science/article/pii/S1042814383710083.

Venkat Gudivada et C.R Rao. *Computational Analysis and Understanding of Natural Languages : Principles, Methods and Applications*, volume 38 de *Handbook of Statistics*. Elsevier, 2018.

Abhirut Gupta, Abhay Khosla, Gautam Singh, et Gargi Dasgupta. Mining procedures from technical support documents. *ArXiv*, abs/1805.09780, 2018.

D. J. (David J.) Hand. *Discrimination and classification*. Wiley series in probability and mathematical statistics. Wiley, Chichester, 1981. ISBN 0471280488.

Zellig S. Harris. Distributional structure. WORD, 10(2-3) :146–162, 1954. URL https://doi.org/10.1080/00437956.1954.11659520.

Mark F. Hornick, Erik Marcadé, et Sunil Venkayala. *Java Data Mining : Strategy, Standard, and Practice : A Practical Guide for Architecture, Design, and Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006. ISBN 0123704529.

David . Israel. The role of logic in knowledge representation. *Computer*, 16(10) :37–41, oct 1983. ISSN 1558-0814.

Anil K. Jain et Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., USA, 1988. ISBN 013022278X.

Christine Jouve. *Représentation des connaissances pour les problèmes de conception. Application à un système à base de connaissances pour la conception de réseaux informatiques : NEST. (Knowledge representation for design problems. Application to a knowledge based system for network design : nest)*. PhD thesis, École nationale supérieure des mines de Saint-Étienne, France, 1992. URL https://tel.archives-ouvertes.fr/tel-00832243.

Yuchul Jung, Jihee Ryu, Kyung-min Kim, et Sung-Hyon Myaeng. Automatic construction of a large-scale situation ontology by mining how-to instructions from the web. *Web Semantics : Science, Services and Agents on the World Wide Web*, 8 :110–124, 07 2010.

Mehmed Kantardzic. Datamining concepts. Dans *Data Mining : Concepts, Models, Methods, and Algorithms*, pages 1–18. 2003.

Daniel Kayser. Raisonnement : logique et informatique. *Logique naturelle et argumentation*, pages 81–103, 1987.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, et Yejin Choi. Mise en place : Unsupervised interpretation of instructional recipes. Dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Lisbon, Portugal, Septembre 2015. Association for Computational Linguistics. URL https://aclanthology.org/D15-1114.

Yves Kodratoff. Technical and scientific issues of kdd (or : Is kdd a science ?). Dans *Proceedings of the 6th International Conference on Algorithmic Learning Theory*, ALT '95, page 261–265, Berlin, Heidelberg, 1995. Springer-Verlag. ISBN 3540604545.

Hsiang Kung et Steeve Huang. Introduction to various reinforcement learning algorithms. part i (q-learning, sarsa, dqn, ddpg. Towards Data Science. http://www.crim.ca/fr/r-d/systemes_distrib/projetssd.html. Page consultée le 6 Janvrier 2022, 2018.

Lukasz Kurgan, Krzysztof Cios, Ryszard Tadeusiewicz, Marek Ogiela, et Lucy Goodenday. Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial intelligence in medicine*, 23 :149–69, 11 2001.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, et Kilian Q. Weinberger. From word embeddings to document distances. Dans *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org, 2015.

Jean-Marc Labatte. Classification supervisée. Université d'Anger département de mathèmatiquel. https://math.univ-angers.fr/~labatte/enseignement%20UFR/M1MIM.html. Page consultée le 6 février 2021, 2013.

John D. Lafferty, Andrew McCallum, et Fernando C. N. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

D.T. Larose et C.D. Larose. *Discovering Knowledge in Data : An Introduction to Data Mining*. Wiley Series on Methods and Applications in Data Mining. Wiley, 2014. ISBN 9781118873571. URL https://books.google.co.id/books?id=UGu8AwAAQBAJ.

Larousse. Définitions. Larousse Dictionnaire Français. https://www.larousse.fr/dictionnaires/francais/connaissance/18273. Page consultée le 4 février 2020, 2020.

Jean-Louis Laurière. *Intelligence artificielle Tome 2 Représentation des connaissances*. #o, Eyrolles, Paris, 1988. ISBN 2-212-08190-1. URL http://www.sudoc.fr/001261266.

Jean. Lave et Etienne Wenger. *Situated Learning : Legitimate Peripheral Participation*. Learning in Doing : Social, Cognitive and Computational Perspectives. Cambridge University Press, 1991. ISBN 9780521423748. URL https://books.google.dz/books?id=CAVIOrW3vYAC.

lebigdata. Python : tout savoir sur le principal langage big data et machine learning. openclassrooms. https://www.lebigdata.fr/python-langage-definition. Page consultée le 6 février 2021, 2020.

J. MacQueen. Some methods for classification and analysis of multivariate observations. page 281–297, 1967.

Jean Pierre Malle. donnee-information-connaissance. Clever cognitive automation solutions Cleverm8l. hhttps://cleverm8.com/donnee-information-connaissance/. Page consultée le 2 Janvier 2022, 2017.

Christopher D. Manning et Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262133601.

Michael C. McCord, Arendse Bernth, Shalom Lappin, et Wlodek Zadrozny. Natural language processing within a slot grammar framework. *Int. J. Artif. Intell. Tools*, 1 :229–278, 1992.

Hongyuan Mei, Mohit Bansal, et Matthew R. Walter. Listen, attend, and walk : Neural mapping of navigational instructions to action sequences. Dans *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2772–2778. AAAI Press, 2016.

Garbade. Michael. A quick introduction to text summarization in machine learning.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, et Jeff Dean. Distributed representations of words and phrases and their compositionality. Dans C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, et K. Q. Weinberger, éditeurs, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

I. Nonaka, I.N.H. Takeuchi, I. Nonaka, N. Ikujiro, T. Hirotaka, , H. Takeuchi, P.K.I. Nonaka, et B.P.M.I.B.R.H. Takeuchi. *The Knowledge-creating Company : How Japanese Companies Create the Dynamics of Innovation*. Everyman's library. Oxford University Press, 1995. ISBN 9780195092691. URL https://books.google.dz/books?id=B-qxrPaU1-MC.

Mahda Noura, Amelie Gyrard, Sebastian Heil, et Martin Gaedke. Automatic knowledge extraction to build semantic web of things applications. *IEEE Internet of Things Journal*, 6(5) : 8447–8454, 2019.

ONUAA. Bonnes pratiques et résilience. KORE - Plateforme de partage des connaissances sur la résilience.Organisation des Nations Unies pour l'alimentation l'agriculture. https://www.fao.org/in-action/kore/good-practices/fr/?page=14&ipp=5&tx_dynalist_pi1%5Bpar%5D=YToxOntzOjE6IkwiO3M6MToiMCI7fQ%3D%3D. Page consultée le 2 Janvier 2022, 2014.

OQLF. Fiche terminologique. Office québécois de la langue française. http://gdt.oqlf.gouv.qc.ca/Resultat.aspx. Page consultée le 16 février 2021, 2002.

Gilbert Paquette. *Modélisation des connaissances et des compétences - pour concevoir et apprendre*. 05 2002. ISBN ISBN 2-7605-1163-4.

Hogun Park, Motahari Nezhad, et Hamid Reza. Learning procedures from text : Codifying how-to procedures in deep neural networks. Dans *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 351–358, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. URL https://doi.org/10.1145/3184558.3186347.

Christine Paulin-Mohring. Logique propositionnelle classique. Institut universitaire Informatique et réseaux industriels IRI Lille. https://www.lri.fr/~paulin/Logique/html/cours004.html. Page consultée le 6 Janvier 2022, 2020.

Gregory Piatetsky-Shapiro. Knowledge discovery in real databases : A report on the ijcai-89 workshop. *AI Mag.*, 11 :68–70, 1991.

M. Ross Quillian. Semantic networks. Dans Marvin L. Minsky, éditeur, *Semantic Information Processing*. MIT Press, 1968.

Michaela Regneri, Alexander Koller, et Manfred Pinkal. Learning script knowledge with web experiments. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988, Uppsala, Sweden, Juillet 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1100.

Arnaud Revel. Apprentissage semi-supervisé et apprentissage transductif. Université de la Rochellel. https://pageperso.univ-lr.fr/arnaud.revel/MesPolys/SemiSupervise.pdf. Page consultée le 12 Janvier 2022, Rochelle 2020.

Raphaël Richard. Système de production. 24PM Academy. https://www.24pm.com/117-definitions/466-systeme-de-production. Page consultée le 6 février 2021, 2004.

Leonard Richardson. Beautiful soup documentation. https://beautiful-soup-4.readthedocs.io/en/latest/. Page consultée le 2 Janvier 2022, 2015.

Stephan Richter. lxml - xml and html with python. https://lxml.de. Page consultée le 2 Janvier 2022, 2021.

Peter J. Rousseeuw. Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65, 1987. ISSN 0377-0427. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

G. Salton et M.J. McGill. *Introduction to Modern Information Retrieval*. International student edition. McGraw-Hill, 1983. ISBN 9780070544840. URL https://books.google.dz/books?id=7f5TAAAAMAAJ.

Institute SAS. *Data Mining and the Case for Sampling*. A SAS Institute, 1998. URL https://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf.

Sensagent. les bonnes pratiques. Dictionnaire.http://dictionnaire.sensagent.leparisien.fr/BONNE%20PRATIQUE/fr-fr/. Page consultée le 16 Janvierr 2022, l 2016.

Guillaume Serries. 5 points pour mieux comprendre les bases de données graph. zdnet. https://www.zdnet.fr/pratique/comment-comprendre-les-bases-de-donnees-graph-39839720.htm. Page consultée le 6 février 2020, 2018.

Colin Shearer. The crisp-dm model : the new blueprint for data mining. *Journal of data warehousing*, 5(4) :13–22, 2000. URL https://www.bibsonomy.org/bibtex/24e676fa2d25f47a2c4937c781a1b0106/becker.

Edward H. Shortliffe. *Computer-based medical consultations, MYCIN / Edward Hance Shortliffe*. Elsevier New York, 1976. ISBN 0444001794.

Amit Singhal. Modern information retrieval : a brief overview. *BULLETIN OF THE IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING*, 24 :2001, 2001.

Spyder. *Analyse de systèmes distribué : Projets*. Spyder IDE. https://www.spyder-ide.org/. Page consultée le 6 Janvrier 2021, 2020.

Jaideep Srivastava, Robert Cooley, Mukund Deshpande, et Pang-Ning Tan. Web usage mining : discovery and applications of usage patterns from web data. *SIGKDD Explor.*, 1 :12–23, 2000.

NLP Group Stanford. The stanford natural language processing group. Stanford, NLP Group. https://nlp.stanford.edu/software/lex-parser.shtml. Page consultée le 6 Novembre 2021, 2020.

P.N. Tan, M. Steinbach, et V. Kumar. *Introduction to Data Mining*. Pearson International Edition. Pearson Addison Wesley, 2006. ISBN 9780321420527. URL https://books.google.dz/books?id=_XdrQgAACAAJ.

Techopedia. Best practice. Dictionary. IT Business Alignment. https://www.techopedia.com/definition/14269/best-practice. Page consultée le 6 Novembre 2021, 2012.

Moritz Tenorth, Ulrich Klank, Dejan Pangercic, et Michael Beetz. Web-enabled robots. *IEEE Robotics Automation Magazine*, 18(2) :58–68, 2011.

Edward L. Thorndike. A scale for measuring the merit of english writing. *Science*, 33(859) :935–938, 1911. ISSN 00368075, 10959203. URL http://www.jstor.org/stable/1638715.

Claude Touzet. Les réseaux de neurones artificiels, introduction au connexionnisme. Université Aix Marseille LNIA - Laboratoire de Neurosciences intégratives et adaptatives. ttps://hal-amu.archives-ouvertes.fr/hal-01338010/file/. Page consultée le 12 December 2020, Montréal 1992.

Mike Uschold et Michael Gruninger. Ontologies : principles, methods and applications. *The Knowledge Engineering Review*, 11(2) :93–136, 1996.

Psyché. Valéry, Mendes. Olavo, et Bourdeau. Jacqueline. Apport de l'ingénierie ontologique aux environnements de formation à distance. volume 10, pages 89–126. 2003.

Hanna M. Wallach. Conditional random fields : An introduction, 2004.

Stanley Wasserman et Katherine Faust. *Social Network Analysis : Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

Christopher J. C. H. Watkins et Peter Dayan. Q-learning. Dans *Machine Learning*, pages 279–292, 1992.

Sholom M. Weiss et Casimir A. Kulikowski. *Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991. ISBN 1558600655.

Etienne Wenger. *Communities of Practice : Learning, Meaning, and Identity*. Learning in Doing : Social, Cognitive and Computational Perspectives. Cambridge University Press, 1998.

Etienne Wenger. Communities of practice and social learning systems. *Organization*, 7(2) : 225–246, 2000. URL https://doi.org/10.1177/135050840072002.

Etienne C Wenger et William M Snyder. Communities of practice : The organizational frontier. *Harvard business review*, 78(1) :139–146, 2000.

wikiHow. About wikihow. About wikiHow. https://www.wikihow.com/wikiHow:About-wikiHow. Page consultée le 6 Janvier 2022, 2019.

Zhibiao Wu et Martha Palmer. Verbs semantics and lexical selection. Dans *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 133–138, USA, 1994. Association for Computational Linguistics. URL https://doi.org/10.3115/981732.981751.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, et Ji-Rong Wen. Efficiently answering technical questions — a knowledge graph approach. Dans *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3111–3118, 2017.

Alexander Yates et Oren Etzioni. Unsupervised resolution of objects and relations on the web. Dans *HLT-NAACL*, pages 121–130, 2007. URL http://www.aclweb.org/anthology/N07-1016.

Xiaojin Zhu. Semi-supervised learning literature survey. Rapport Technique 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

Abdelkader Djamel Zigheb et Ricco Rakotomalala. Extraction de connaissances à partir de données (ecd). *Techniques de l'ingénieur Technologies logicielles Architectures des systèmes*, base documentaire : TIP402WEB(ref. article : h3744), 2002.

G.K. Zipf. *Human Behavior and the Principle of Least Effort : An Introduction to Human Ecology*. Addison-Wesley Press, 1949. URL https://books.google.dz/books?id=1tx9AAAAIAAJ.

ملخص:
منذ ظهور الويب الاجتماعي والدلالي ، :ظهرت أدوات جديدة ومواقع مشاركة مثل ميطا, تويتر, ويكي هاو وما إلى ذلك ، مما جعل الويب مجموعة عالمية من المعرفة ، حيث يتم توزيع المستخدمين جغرافيًا من المجتمعات عبر الإنترنت ، ويتبادلون ويشاركون معرفتهم في مجالات مختلفة في شكل معرفة إجرائية تسمى(م ج)
يتم تحديد هذه الممارسات الجيدة من خلال مجموعة من الخطوات المتتالية المتخذة لتحقيق الهدف. لقد أصبح وضع تصورات (م ج) هذه تحديًا كبيرًا في العديد من المجالات (استرجاع المعلومات ، والتطبيقات الذكية ، والروبوتات ، وما إلى ذلك) ، وفي هذا السياق نقدم في هذه الأطروحة نهجًا جديدًا لاستخراج وتصور الممارسات الجيدة من الويب، واستخراج أفضل ممارسة لاستعلام معين من خلال تطبيق تقنيات التعلم الآلي وتلخيص النص على الرسوم البياني

الكلمات الرئيسية: مجتمع الممارسة ، الممارسة الجيدة ، استخراج المعرفة الإجرائية ، الرسم البياني للمعرفة ، تركيب النص ، التعلم الآلي

Résumé
Depuis l'avènement du web social et sémantique, il ne cesse d'émerger ces dernières années de nouveaux outils et sites de partage tels que Meta, Twitter, WikiHow,... faisant du web un recueil universel de connaissances, où les utilisateurs répartis géographiquement forment des communautés de pratique (CdP) en ligne, ces CPs sont à l'origine un concept de sociologie mais trouvent tout leur essor dans le web actuel où des individus partagent et échangent leur savoir faire dans différents domaines sous forme de connaissances procédurales (CPs) appelées bonnes pratiques.
Ces bonnes pratiques sont définies par un ensemble d'étapes successives acheminées pour atteindre un objectif. Conceptualiser ces CPs est devenu un enjeu majeur dans plusieurs domaines (recherche d'information, applications intelligentes, Robotique…). Et c'est dans ce contexte que dans cette thèse nous présentons une nouvelle approche pour extraire et conceptualiser les bonnes pratiques du web, et extraire la meilleure pratique pour une requête donnée, ceci en appliquant les techniques d'apprentissage artificiel et de résumé de texte sur les graphes .

Mots clés: Communauté de pratique, Bonne pratique, Extraction des connaissance procédurale, Graphe de connaissance, Synthèse de texte, Apprentissage artificiel.

Abstract
Since the advent of the social and semantic web, new tools and sharing sites such as Meta, Twitter, WikiHow, etc. have emerged, making the web a universal collection of knowledge, where users geographically distributed form online communities, sharing and exchanging their know-how in different fields in the form of procedural knowledge (PK) called good practices.
These good practices are defined by a set of successive steps taken to achieve an objective. Conceptualizing these PK has become a major challenge in several fields (information retrieval, intelligent applications, robotics, etc.), and it is in this context that in this thesis we present a new approach to extract and conceptualize good practices from web, and extract the best practice for a given query by applying machine learning and text summarization techniques on graphs.

Keywords: Community of practice, Good practice, Procedural knowledge extraction, Knowledge graph, Text synthesis, Machine learning.