REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Génie Logiciel

Par :

BOUREBAI Ismail
BOURAS Mohamed Elamine

Sur le thème

La recherche de commentaires pertinents

Soutenu le 15/ 11 /2020 à Tiaret devant le jury composé de :

M.KHARROUBI Sahraoui	M.C.B	Université IBN-KHALDOUN Tiaret	Président
Mme. LAKHDARI Aicha	M.A.A	Université IBN-KHALDOUN Tiaret	Encadrante
M. CHENINE Abdelkader	M.A.A	Université IBN-KHALDOUN Tiaret	Examinateur

Année Universitaire : 2019 - 2020

DÉDICACE

Du plus profond de nos cœurs, nous dédions ce travail à:

Nos familles, BOURAS et BOURBIE, en particulier nos parents. Ce memoire n'aurait pas vu le jour sans leurs encouragements au fil des ans.

À tous nos frères et sœurs.

À nos cousins et cousines, en particulier GARMITTE Soumia et ZEGGO Ihab, sans oublier nos collègues CHARFOUI Younes, BOUKELIKHA Djalil, BENCHOHRA Anoir et BETTAHER Freiha qui nous ont aidés tout au long de nos études.

À tous nos amis et aux personnes qui ont toujours cru en nous.

REMERCIEMENTS

Tout d'abord, nous remercions Dieu Tout-Puissant de nous avoir accordé le courage, la patience et la force morale et physique pour pouvoir accomplir ce travail.

Notre gratitude s'adresse à Mme LAKHDARI Aicha pour son encadrement, son orientation, ses conseils et la disponibilité qu'elle nous a témoignée pour nous permettre de mener à bien ce travail.

Nous tenons à exprimer nos vifs remerciements à MrKHARROUBI Sahraoui qui a accepté de présider le jury de soutenance, pour tout ce qu'il a pu nous apprendre ; qu'il trouve ici l'expression de notre profonde et sincère reconnaissance. Mr CHENINE Abdelkader pour nous avoir fait l'honneur d'accepter d'examiner ce travail.

Nous addressons nos remerciements à tous les professeurs, pour leurs conseils et leurs critiques qui ont guidé nos reflexions durant nos recherches, et nous remercions également tous nos collègues et amis du département d'informatique de l'université Ibn Khaldoun à Tiaret.

Nous souhaitons exprimer nos profondes gratitudesà nos parents qui nous ont soutenustout au long de notre projet, ainsi que toute la famille, les amis pour leur soutien indefictible.

Enfin, on remercie tous ceux qui nous ont aidés de près ou de loin dans l'élaboration de ce travail.

ABSTRACT

Our work lies at the crossroads of information research (IR) and social networks, which is Social Information Research (SIR). More specifically, we are interested in the search for relevant social textual signals such as (comments, status, tweets, etc.) in Twitter, Facebook and YouTube. To do so, we start with a syntactic search approach and then enhance it with a semantic search. The latter is based on two measures of conceptual similarity: Wu-Palmer [1] and Path-Similarity [2]. Then, we integrate, in a separate and combined way, all non-textual user interactions (e.g. Like, dislike, share, reactions, etc.) left on their textual contents within the search process as an additional source of information to improve the relevance of the results. The combination of non-textual signals can thus be a criterion of social relevance.

Our IR model is based on a vector representation of the data. It performs matches between the query entered and the documents already prepared in order to return the ranking of the latter according to their overall relevance.

Our experiments have shown an increase in the better satisfaction rate (Relevance) compared to traditional information retrieval systems.

Key words : Social Information Search, Information Retrieval System, Relevance, Social Networks, Social Signals.

RÉSUMÉ

Notre travail se situe au carrefour de la recherche d'information (RI) et des réseaux sociaux qui est la Recherche d'Information Sociale (RIS). Nous nous intéressons, plus précisément, à la recherche de signaux sociaux textuels pertinents comme (les commentaires, les statuts, les tweets, etc.) dans Twitter, Facebook et YouTube. Pour le faire, nous commençons d'abord par une approche de recherche syntaxique bonifiée à la suite par une recherche sémantique. Cette dernière repose sur deux mesures de similarité conceptuelle celles de Wu-Palmer [1] et Path-Similarity [2]. Ensuite, nous intégrons, d'une manière séparée et combinée, toutes les interactions des utilisateurs non textuelles (ex : Like, dislike, share, reactions, ...etc.) laissées sur leurs contenus textuels au sein du processus de recherche en tant que source d'information additionnelle pour améliorer la pertinence des résultats. La combinaison des signaux non textuels peut être ainsi un critère de pertinence sociale.

Notre modèle de RI se base sur une représentation vectorielle des données. Il effectue un appariement entre la requête saisie et les documents déjà préparés pour retourner le classement (ranking) de ces derniers selon leur pertinence globale.

Nos expériences menées ont montré une augmentation du taux de satisfaction meilleur (Pertinence) par rapport aux systèmes de recherche d'information classiques.

Mots clés : Recherche d'information sociale, Système de recherche d'information, Pertinence, Réseaux sociaux, Signaux sociaux.

TABLE DES MATIÈRE

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

API	Application Programming Interface
BSD	Berkeley Software Distribution
CACM	Communications of the Association for Computing Machinery
CSV	Comma Separated Values
EER	Extended Entity Relationship
FTP	File Transfer Protocol
GPS	Geographic Positioning System
IA	Intelligence Artificiel
IDE	Integrated Development Environment
IDF	Inverse Document Frequency
LCS	Least Common Subsumer
MESH	MEdical Subject Headings
NLTK	Natural Language Toolkit
PHP	Hypertext Preprocessor
RI	Recherche d'Information
RIS	Recherche d'Information Sociale
RSV	Retrieval Status Value
SRI	Systèmes de Recherche d'Information
TF	Term Frequency
TIC	Technologies de l'Information et de la Communication
UGC	User Generated Content
URL	Uniform Resource Locator
VSC	Visual Studio Code
WSD	Word Sense Disambiguation

INTRODUCTION GENERALE

Contexte

Aujourd'hui, l'information joue un rôle primordial dans le quotidien des individus et dans l'essor des entreprises. Cependant, le développement des Technologies de l'Information et de la Communication (TIC) et notamment la mise en place des technologies du Web 2.0 ont permis de faire émerger de nouveaux médias sociaux et ainsi de nouvelles sources d'informations : les blogs, les wikis, les podcasts, les plates-formes de partage de fichiers et les sites de réseaux sociaux. En effet, la quantité d'information disponible, particulièrement à travers le web, se mesure en milliards de pages. Il est par conséquent, de plus en plus difficile de localiser précisément ce que l'on recherche dans cette masse d'information. La recherche d'information (RI) est le domaine par excellence qui s'intéresse à répondre à ce type d'attente. En effet, l'objectif principal de la RI est de fournir des modèles, des techniques et des outils pour stocker et organiser des masses d'informations et localiser celles qui seraient pertinentes relativement à un besoin en information d'un utilisateur, souvent, exprimé à travers une requête. Ces outils sont appelés des Systèmes de Recherche d'Information (SRI). De manière générale, le fonctionnement d'un SRI consiste à construire une représentation des documents et de la requête et d'établir une comparaison entre ces deux représentations (requête, documents) pour retourner les documents pertinents. Cette comparaison est réalisée au moyen d'un modèle de recherche. Afin d'obtenir un SRI performant, il est nécessaire de construire une bonne représentation du document et de la requête et de développer un modèle de RI qui supporte ces représentations. [3]

Avec les réseaux sociaux comme par exemple Twitter, Facebook, LinkedIn ou encore pour les plus récents Foursquare, Gowalla, les internautes passent d'un état passif où ils étaient de simples consommateurs à un état actif où ils deviennent producteurs d'informations. Ce qui a conduit à l'émergence d'une nouvelle branche de recherche d'informations : c'est la Recherche d'Information Sociale (RIS). [3]

Motivation

La motivation derrière l'exploitation des signaux sociaux (ex. commentaire, j'aime, etc.) sur la performance des SRI est d'essayer de tirer profit de ces activités sociales provenant des interactions des utilisateurs de différents réseaux sociaux (Twitter, YouTube, Facebook, etc.) pour améliorer la RI par rapport à un besoin en information. Donc, la pertinence thématique des ressources Web (pertinence textuelle ou bien Baseline model) sera bonifiée par l'importance sociale de ces ressources (pertinence sociale) (figure 1).

Ces interactions peuvent être exploitées à différents niveaux, à savoir au niveau de l'utilisateur (profilage) pour mieux comprendre ses besoins, ou bien du côté de la ressource pour mieux la décrire et mesurer une certaine pertinence. Notre travail se situe dans la seconde classe, l'exploitation des signaux sociaux pour améliorer la RI. [2]

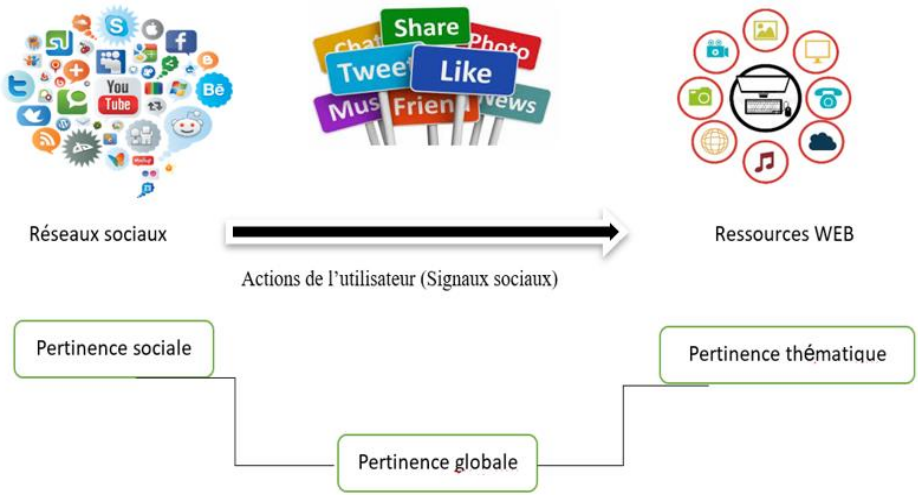


Figure 1: Utilisation de signaux sociaux pour améliorer la RI

Défis et enjeux

La problématique principale de la recherche d'information (RI) derrière les données générées par les utilisateurs porte sur la manière de transformer ces contenus hétérogènes en critères permettant de les intégrer dans des modèles d'évaluation de pertinence pour l'exploitation efficace dans des tâches de RI. De nombreux défis découlent de cette problématique tels que :

- La gestion à grande échelle de ces quantités massives d'informations sociales générées par l'utilisateur pour les traiter et les rendre utilisables et exploitables.

- La diversité des structures de réseaux sociaux qui les différencient de ses concurrents apporte des difficultés supplémentaires concernant l'analyse et l'exploitation.
  de ce fait, de nouvelles problématiques se posent quant aux méthodes qu'il faut utiliser pour parvenir à des résultats satisfaisants.

Questions de recherche

Les questions de recherche auxquelles nous avons répondu durant notre projet de fin d'études (PFE) porte sur la définition de la pertinence via une exploitation séparée et groupée des signaux sociaux non textuels (ex. like, dislike, share, reactions, ... etc.) sont les suivantes :

1- Quelle est l'approche la plus efficace afin de restituer les signaux textuels pertinents ?

2- Quelles sont les signaux non textuels qui peuvent être des critères de pertinence ?

3- Comment combiner ces signaux non textuels pour améliorer la RI ?

4- Quel est l'impact des signaux sociaux textuels et non textuels sur les performances d'un SRI ?

Objectifs

Nous nous intéressons plus particulièrement à la définition d'une approche basée sur un modèle de SRI, permettant de construire une représentation vectorielle des documents et de la requête et d'établir une correspondance syntaxique, sémantique et sociale entre ces deux représentations pour restituer les contenus textuels pertinents.

Nous prenons en compte l'exploitation séparée et groupée des signaux non textuels issus de chaque réseau social (Twitter, YouTube ou Facebook), qui sont sous forme d'actions relevant d'activités sociales telles que le nombre de *j'aime*, de *partage*... . La combinaison des signaux peut être aussi un critère de pertinence sociale.

Organisation du mémoire

Afin de parvenir à notre but, le présent document est divisé en deux volets. Un volet théorique ainsi qu'un volet pratique. Le volet théorique (partie 1) dénommé « *Etat de l'art* » comprend les chapitres 1 et 2. Le chapitre 1 présente la RI classique. Premièrement, la RI sera définie historiquement, en plus de son domaine d'application. Deuxièmement, les principales notions de base de traitement de l'information et les différentes étapes de la RI seront présentées, dans le but de fournir une compréhension sur son processus. Troisièmement, nous décrivons les différents modèles de la RI en particulier le modèle booléen, le modèle vectoriel

et le modèle probabiliste. Le chapitre 2 est consacré à la RI sociale. Nous donnons certaines définitions concernant la RIS et les différents réseaux sociaux. Puis nous expliquons l'exploitation des contenus sociaux afin d'améliorer la RI. Enfin, nous présentons les approches basées sur les signaux sociaux.

Le volet pratique (partie 2) intitulé «*Conception et réalisation de l'application* ». Il comprend un seul chapitre qui représente la conception de notre application, et les résultats obtenus grâce à sa mise en œuvre.

Enfin, nous concluons notre mémoire par une conclusion générale, où nous présentons les perspectives de nos propositions.

Partie I

I. ETAT DE L’ART

Chapitre 1

I.1 RECHERCHE D'INFORMATION CLASSIQUE

I.1.1 Introduction

La Recherche d'Information (RI) n'est pas un domaine récent, il date des années 40. Ce domaine peut être défini comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur. [4]

Le processus de la RI consiste à mettre en correspondance et à calculer le degré d'appariement des représentations internes des documents et de la requête. Les documents retournés à l'utilisateur (documents dits pertinents), sont ordonnés dans une liste par ordre décroissant de degré de pertinence. Afin d'améliorer les résultats de la recherche, le système de recherche d'information (SRI) crée une interface entre la base documentaire, ou collection de documents, et les utilisateurs qui recherchent des informations contenues dans cette base.

I.1.2 Bref Historique de la RI

Le domaine de recherche d'information remonte au début des années 1940, peu après l'invention des ordinateurs, nous allons retracer brièvement son évolution à travers le temps :

- 1940 : Apparition des SRI, focalisation de la RI sur les applications des bibliothèques.
- 1950 : Apparition du modèle booléen et l'élaboration des petites expérimentations sur des petites collection de documents.
- 1960 et 1970 : Apparition du système SMART, développement d'une méthodologie d'évaluation de système et conception du corpus de test (CACM).
- 1980 : Développement de l'intelligence artificiel, ainsi on tentait d'intégrer les techniques de l'IA en RI.
- 1990 et 1995 : Apparition d'internet, la RI a été modifié et sa problématique plus élargie comprenant la recherche d'information sur les réseaux sociaux. [5]

I.1.3 Définition et domaine d'application

- **Recherche d'Information (RI) :** *Salton* a définit la recherche d'information en 1968 comme « *un domaine qui traite de la représentation, du stockage, de l'organisation et de l'accès à l'information* ». [6]

  D'après Van *Rijsbergen* en 1980:*« la RI consiste à restituer les documents qui peuvent être pertinents par rapport au besoin d'information exprimé dans la requête »*. [7]

  Une autre définition que « *la RI est l'ensemble des méthodes, procédures, techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes* ». [8]

La recherche d'information est un domaine très large, et qui peut être utilisée dans plusieurs types d'application afin de répondre à un besoin en information. Comme le montre le tableau suivant :

Catégorie	Description	Exemple de la requête
Ad hoc recherche	Retrouver les documents pertinents dans une collection fixe.	Find documents which tell me about investment strategies.
Question\Réponse	Extraire les réponses dans les documents récupérés.	Who is the prime minister of Australia?
Annuaire	Navigation dans une Web page spécifique.	Where is the ELSNET home page?
Diffusion sélective d'information	Contrôlez un flot de documents correspondant à un profil.	Send me any new information on high tech companies.
Classification de documents	Regroupement automatique de documents.	Find the natural grouping in this set of scientific publications.
Catégorisation de documents	Affecter un document à une catégorie prédéfinie.	Classify incoming books according to their Dewey decimal category.
Synthèse de documents	Extraire l'information à partir des documents retrouvés.	Construct a personalized travel guide for my visit to Athens in July 2000.
Recherche dans la base de données	Extraire des enregistrements à partir d'une base de données structurée.	Find books where author = Salton and year =2001

Tableau 1: Domaines d'application de la RI *[9]*

Compte tenu de ces définitions, la RI est un domaine dont la finalité est de répondre à un besoin d'information dans une grande base de documents, à l'aide d'un ensemble des méthodes et outils qui facilitent la recherche appelés systèmes de recherche d'information.

- **Systèmes de Recherche d'Information (SRI) :** de nombreuses définitions sur les systèmes de recherche d'information expriment le même principe :

  Selon *Tebri*, « *un SRI est un ensemble de programmes informatiques qui a pour but de sélectionner des informations pertinentes qui répondent aux besoins d'un utilisateur, exprimés sous forme de requêtes* ». [10]

  Alors que *Chein* décrit un SRI comme étant « *un système informatique dont le but est d'aider un utilisateur à trouver des documents contenant des informations pertinentes pour un besoin d'information exprimé par une requête au système* ». [11]

  Une autre définition plus simple donnée par *Mathias* :« *un SRI est un système qui facilite l'accès à un ensemble de documents, pour permettre de retrouver ceux dont le contenu correspond le mieux à un besoin d'information d'un utilisateur* ». [12]

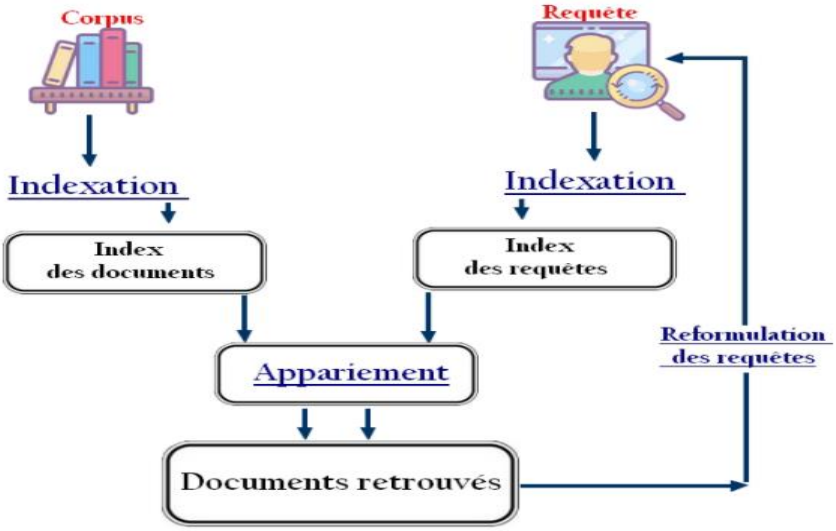


Figure 2:Architecture générale d'un Système de Recherche d'Information *[13]*

D'après l'architecture générale d'un SRI illustrée par la figure ci-dessus (figure 2), nous avons besoins donc de faire un aperçu sur les principales notions concernant le SRI qui nous permettent de comprendre son fonctionnement et de détailler son processus.

I.1.4 Notions de base

- **Document :** l'une des définitions possibles du terme document est de le considérer comme un support physique de l'information, qui peut être du texte, une page web, une image, une séquence vidéo, etc. [14]. Le document constitue l'information élémentaire d'une collection de document [15].

- **Corpus :** représente l'ensemble de documents exploitables et accessibles par l'utilisateur dont lesquels il cherche une information, on peut le définit aussi comme un ensemble des documents manipulés par un SRI qui se nomme collection de documents (ou base documentaire).
- **Besoin d'information :** c'est une représentation mentale de ce que l'utilisateur souhaite chercher. Ce besoin est représenté sous forme d'une requête. [16]
- **Requête :** c'est une expression du besoin de l'utilisateur, elle représente l'interface entre le système de recherche d'information et l'utilisateur.
- **Pertinence :** la définition la plus simple de cette notion qui représente le critère le plus important dans la RI et l'objet de tout SRI est : « *la pertinence est la correspondance entre un document et une requête, ou encore un degré de relation du document à la requête et une mesure d'utilité du document pour l'utilisateur* » [17]. Essentiellement, deux types de pertinence sont définis :
  1- **La pertinence Système :** est souvent présentée par un score attribué par le SRI afin d'évaluer l'adéquation du contenu des documents vis-à-vis de celui de la requête. [18]
  2- **La pertinence utilisateur :** pour l'utilisateur la pertinence correspond à la satisfaction d'ensemble de documents restitués par le SRI. Deux utilisateurs peuvent juger différemment un même document renvoyé pour une même requête. [19]

I.1.5 Processus de la RI

Ce processus est composé de trois étapes essentielles : l'indexation, l'appariement document-requête et la reformulation de requête, comme illustré dans la figure 2. [20]

I.1.5.1 Indexation

Une étape très importante consiste à analyser les documents et la requête d'utilisateur. Cette opération doit s'effectuer avant l'étape de recherche effective de l'information afin de construire des index qui faciliteront le processus de recherche. Un des objectifs de l'indexation est donc de permettre de retrouver rapidement les documents contenant les termes (mots-clés) de la requête. Plusieurs manières étaient donc proposées par les développeurs des SRI afin de la procéder. Les principales manières sont :

- **Indexation manuelle :** dans ce cas, chaque document est analysé par un spécialiste du domaine ou par un documentaliste. Après la lecture des documents ce spécialiste détermine, selon ses connaissances, les mots-clés qui lui semblent les plus adéquats pour représenter le contenu du document. Ce mode d'indexation est fondé sur le jugement humain, il se caractérise par sa profondeur, cohérence et sa qualité [17].

  L'indexation manuelle a l'avantage d'assurer une meilleure correspondance entre les documents et les termes choisis par les indexeurs pour les représenter. Cependant, elle présente un effort trop coûteux en temps et en besoin humain. De plus, l'inconvénient majeur de cette méthode d'indexation est que des termes différents peuvent être sélectionnés par des indexeurs différents. Il peut même arriver qu'une personne, à des moments différents, index différemment le même document. [15]
- **Indexation automatique :**ce type rend le processus d'indexation complètement automatisé, car il détecte automatiquement les termes les plus représentatifs du contenu du document, il comprend un ensemble de traitement sur le document : l'analyse de texte du document mot à mot, extraire les mots vides qui ne jouent qu'un rôle syntaxique, éliminer les mots qui n'ont aucun intérêt, pondérer les termes et finalement la création de l'index. Actuellement, l'indexation automatique est la plus répandu, basée essentiellement sur une approche statistique, est adoptée par la majorité des systèmes de RI en raison de son coût réduit par rapport à l'indexation manuelle. [21]
- **Indexation semi-automatique :** est un rapport qui combine les deux types d'indexation manuelle et automatique. Les premiers éléments de l'indexation sont effectués par l'indexation automatique, puis elle fait appel à une intervention humaine (par un expert) pour corriger manuellement les informations sélectionnées par l'indexation automatique. [22]

Généralement, l'indexation comprend une série de traitements automatisés appliqués sur les documents et aussi sur les requêtes comme le montre la figure ci-dessous :

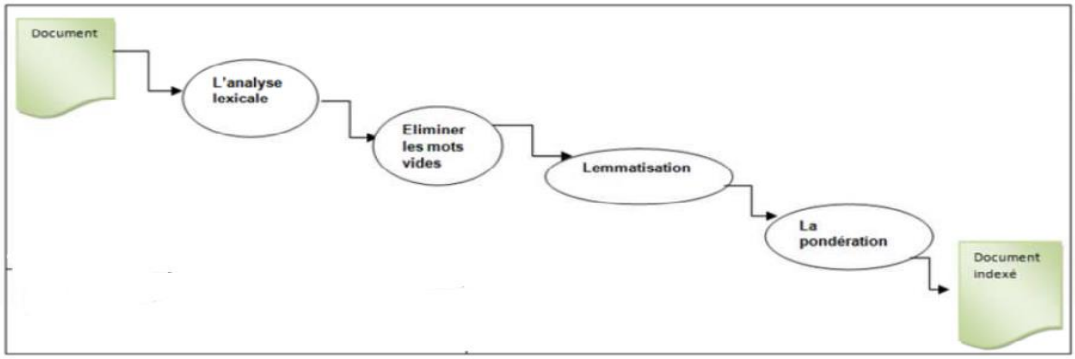


Figure 3: Processus d'indexation

I.1.5.1.1 Analyse lexicale (Segmentation)

C'est un processus qui convertit le texte d'un document en un ensemble de termes où un terme est un radicale ou une unité lexicale. Cette analyse permet de reconnaitre les espaces de séparation, des chiffres, des mots et des ponctuations. [23]

I.1.5.1.2 Elimination des mots vides

L'élimination des mots vides est l'une des étapes de processus d'indexation permettant d'améliorer la fiabilité d'un SRI au sens de qualité logiciel (temps d'exécution) et de performance. [24]

Elle consiste à éviter les mots vides (prénom personnel, prépositions, articles, mots mathématiques, etc.). Et choisir seulement les termes significatifs qui représentent au mieux un document donné. Afin d'éliminer ces mots de force, on utilise une liste, appelée Stop-List (ou anti-dictionnaire) qui contient tous les mots qu'on ne veut pas garder. L'élimination de ces mots permet de réduire l'index, on gagne alors en espace mémoire, mais aussi le non-traitement des mots vides permet de réduire le temps d'exécution.

I.1.5.1.3 Radicalisation (Normalisation)

L'utilisation de la lemmatisation, permet de pouvoir indexer les différentes formes d'un mot désignant le même sens dans un texte par un seul mot qui porte un concept commun. Elle consiste à éliminer les terminaisons des mots et garder seulement la racine. Grâce à ce processus, les documents contenant différentes formes d'un même mot auront les mêmes chances d'être restitués [25]. Par conséquent, elle réduit la taille de l'index et améliore le rappel, mais elle peut réduire la précision car elle supprime dans certains cas la sémantique des termes originaux. Plusieurs méthodes sont utilisées pour retrouver la racine lexicale d'un mot tel que : « *Tree-tagger* » [26], « *Algorithme de Porter* » [27], « *la troncature à 7 caractères* » et la méthode des n-grammes [28].

I.1.5.1.4 Pondération

La pondération est une phase primordiale vient après l'identification des termes et leur normalisation puisqu'elle traduit l'importance de ces termes en indices qui reflètent le poids relatif des mots dans les documents. On distingue deux types de pondération :

1- **Pondération locale :** indique l'importance locale du terme dans un document, elle utilise la fonction suivante :

TF **(Term Frequency)** : ce facteur prend en compte le nombre d'occurrence d'un terme dans un document. Soit le document $\boldsymbol{d_j}$et le terme $\boldsymbol{t_i}$, alors la fréquence $\boldsymbol{TF_{ij}}$ du terme dans le document peut être donnée selon l'une des formulations suivantes :

$$\mathbf{TF_{ij}= 1 + \log (td_{ij})} \quad \text{Ou bien :} \quad \mathbf{TF_{ij}} = \frac{\mathbf{td_{ij}}}{\mathbf{\Sigma_k td_{kj}}} \qquad (1)$$

Avec td_{ij} représente le nombre d'occurrences du terme t_i dans le document d_j. Le dénominateur est la somme des occurrences de tous les termes dans le document d_j. [25]

2- **Pondération globale :** elle utilise la fonction suivante :

IDF **(Inverse Document Frequency) :** ce facteur mesure la fréquence d'un terme dans toute la collection. En effet, un terme fréquent dans la collection, a moins d'importance qu'un terme moins fréquent. Cette mesure est exprimée selon l'une des déclinaisons suivantes :

$$\boldsymbol{IDF_i} = \mathbf{\log_n}\left(\frac{\mathbf{N}}{\mathbf{n_i}}\right) \qquad \text{ou bien :} \quad \boldsymbol{IDF_i} = \mathbf{\log_n}\left(\frac{\mathbf{N}}{\mathbf{n_{i+1}}}\right) \qquad (2)$$

Avec N représente le nombre de documents de la collection et n_i est le nombre de documents dans lesquels le terme $t_i$apparaît. [25]

De manière générale, la méthode de pondération la plus utilisée est construite par la combinaison de ces deux facteurs (TFIDF) :

$$\boldsymbol{TFIDF_{td} = TF_{t,d} * IDF_t} \qquad (3)$$

I.1.5.2 L'appariement document-requête

La relation d'appariement donne la possibilité de rechercher parmi les documents transformés, ceux qui répondent le mieux à une requête d'utilisateur. Le SRI procède à la mesure de pertinence de chaque document vis-à-vis du besoin d'information (requête) selon une fonction de correspondance relative au modèle de recherche, et à renvoyer ensuite à l'utilisateur une liste de résultats après avoir calculer un score de correspondance. Ce score est

calculé au moyen d'une fonction de similarité appelée RSV (D, Q) **(Retrieval Status Value)**, où *Q* représente une requête et *D* un document.

I.1.5.3 La reformulation de requête

Le processus de la reformulation des requêtes permet de générer une nouvelle plus adéquate en rajoutant de nouveaux termes et/ou supprimant des termes inutiles, afin de coordonner le langage de recherche utilisé par l'utilisateur (requête) et le langage d'indexation des documents. Deux approches principales sont utilisées dans la reformulation des requêtes : [29]

1- **Expansion automatique des requêtes :** cette approche consiste à rajouter à la requête initiale des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections. On peut alors utiliser des ontologies linguistiques (exemple. Word Net). [30]
2- **Reformulation interactive :** à l'inverse de la reformulation automatique, le système et l'utilisateur sont ensemble responsables de la détermination et du choix des termes candidats à la reformulation. Le système joue un rôle dans la suggestion des termes, le calcul des pois des termes, l'affichage de la liste ordonnée des termes à l'écran. Tandis que l'utilisateur examine cette liste et prend la décision ultime dans la sélection des termes à ajouter dans la requête. [31]

Un SRI peut être basé sur plusieurs algorithmes afin de mesurer l'adéquation entre une requête et un ensemble de mots-clés issus de l'indexation d'un document. Dans la suite, nous présentons les principaux modèles de l'état de l'art.

I.1.6 Modélisation du système de recherche d'information

Un modèle de RI est une représentation des documents et requêtes qui définit une stratégie d'ordonnancement des documents retournés vis-à-vis de la requête. Il a pour rôle de fournir un cadre théorique pour la modélisation de mesure de pertinence [15]. Plusieurs modèles de RI ont été proposés dans la littérature, ils s'appuient sur des cadres théoriques différents, théorie des ensembles, algèbre, probabilités, etc. Globalement, on distingue trois principales catégories de modèles : modèles booléens, modèles vectoriels et modèles probabilistes. [32]

I.1.6.1 Modèle booléen ou ensembliste

Les premiers SRI développés sont basés sur le modèle booléen, même aujourd'hui beaucoup de systèmes commerciaux (moteurs de recherche) utilisent le modèle booléen, il est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, les documents et les requêtes sont représentés sous forme d'un ensemble de termes. D'une part un document (d) est représenté par une conjonction de mots-clés (exemple. $d = t_1 \wedge t_2 \wedge t_3 \ldots \wedge t_n$), d'autre part une requête (q) est représentée par une expression logique composée de mots connectés par des opérateurs booléens : [8]

- **La Conjonction (∧) :** les termes soient présents simultanément dans la description d'un document.
- **La Disjonction (**V**) :** au moins un des termes soit présent dans la description d'un document à retourner.
- **La Négation (¬) :** utilisée pour écarter les documents qui contiennent un terme.

Par exemple : $q = (\mathrm{t}_1 V\ \mathrm{t}_2) \wedge \neg(\mathrm{t}_3\ V\ \mathrm{t}_4)$

L'appariement (RSV) entre une requête et un document est un appariement exact, autrement dit si un document implique au sens logique la requête alors le document est pertinent. Sinon, il est considéré non pertinent. La correspondance entre document et requête est déterminée comme suit :

$$\boldsymbol{RSV(d,\ q)}=\begin{cases} \mathbf{1} & \boldsymbol{si\ d\ appartient\ \text{à}\ l'ensemble\ décrit\ par\ q} \\ \mathbf{0} & \boldsymbol{sinon} \end{cases} \quad (4)$$

Ce modèle est très largement utilisé grâce à la simplicité et à la rapidité de sa mise en œuvre, mais il présente un certain nombre de faiblesses, en particulier la complexité de la formulation de requête car il est difficile pour les utilisateurs de formuler de bonnes requêtes. Par conséquent, l'ensemble des documents trouvés est souvent trop grand, pour les requêtes courtes, ou complètement vide dans le cas de requêtes longues. Ainsi que l'impossibilité d'ordonner les documents retournés.

I.1.6.2 Modèle vectoriel

Le modèle vectoriel a été popularisé en 1971 par *Salton* [21], il se base sur une formalisation géométrique. En effet, les documents et les requêtes sont représentés sous forme de vecteurs dans un espace de N-dimensions engendré par les termes d'indexation. Chaque document est représenté par un vecteur : $\mathrm{d}_j = (\mathrm{w}_{1j}, \mathrm{w}_{2j}, \ldots, \mathrm{w}_{Mj})$. De même chaque requête

est représentée par un vecteur : q_i= (w_{1i}, w_{2i},, ..., w_{Mi}). Avec : w correspond au poids d'un terme dans le document $d_j$ou dans la requête q_i. L'appariement document-requête dans le modèle vectoriel, consiste à trouver les vecteurs documents qui s'approchent le plus de vecteur de la requête. Cet appariement est obtenu par l'évaluation de la distance entre les deux vecteurs, elle estimée selon les propriétés et les mesures populaires issus de la théorie des espaces vectoriels [24]. Le tableau 2 montre les fonctions les plus utilisées pour mesurer le degré de similarité entre ces deux vecteurs (document, requête) [7].

Mesures	Formules
Le produit scalaire	$RSV(q_i, d_j) = \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)$
Lamesure de consinus	$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right) + \sum_{k=1}^{M} \left(w_{k_j}{}^2\right) - \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}$
La mesure de Dice	$RSV(q_i, d_j) = \frac{2 * \sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sqrt{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right) + \sum_{k=1}^{M} \left(w_{k_j}{}^2\right)}}$
La mesure de Jaccard	$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} \left(w_{k_i} * w_{k_j}\right)}{\sqrt{\sum_{k=1}^{M} \left(w_{k_i}{}^2\right)} * \sqrt{\sum_{k=1}^{M} \left(w_{k_j}{}^2\right)}}$

Tableau 2: Mesures de similarité dans le modèle vectoriel *[7]*

La figure 4 illustre une représentation vectorielle dans un espace composé de deux termes (T1,T2), avec deux documents (d1,d2), et une requête (q), plus l'angle formé par le document et la requete est petit, plus le consinus de cet angle est grand et par la suite ce document est consideré comme étant le plus pertinent à la requête, comme le document (d1) avec la requête (q).

Le modèle vectoriel reste le plus modèle simple à utiliser et facile à implémenter avec l'algèbre linéaire. Il offre une meilleure qualité des résultats, et permet une correspondance partielle ou approximative entre les requêtes et les documents qui sont triés selon leur degré de similarité. Cependant, l'un des principaux inconvénients de ce modèle réside dans l'obligation pour un texte de contenir au moins un des mots de la requête pour pouvoir être retrouvé car les dimensions de l'espace vectoriel étant orthogonales, les termes sont donc considérés comme indépendants.

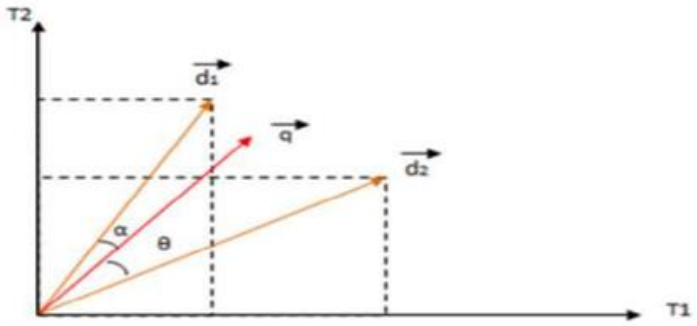


Figure 4: Représentation vectorielle de deux documents et une requête [33]

I.1.6.3 Modèle probabiliste

Le modèle probabiliste propose de modéliser le processus de sélection des documents dans un SRI en basant sur la théorie des probabilités [34]. Il trie ces documents selon leur probabilité de pertinence vis-à-vis d'une requête [17]. En utilisant deux probabilités conditionnelles pour estimer la probabilité que le document appartient à la classe des documents pertinents (ou non-pertinents) :

1- la probabilité qu'il soit pertinent à la requête, notée P(R/d).
2- La probabilité qu'il soit non pertinent à la requête, notée P(NR/d).

Le score d'appariement entre le document d et la requête Q, noté RSV (d, Q) est donné par [35] :

$$RSV\ (d, Q) = \frac{P(R/d)}{P(NR/d)} \qquad (5)$$

Un document est alors sélectionné si la probabilité qu'il soit pertinent à Q, est supérieure à la probabilité qu'il soit non pertinent à Q.

I.1.7 Au-delà des mots simples

Les modèles développés en RI sont souvent basés sur l'utilisation des mots simples comme unités de représentation des documents et des requêtes (représentation en sac de mots). Ces approches posent des problèmes liés à l'ambiguïté des mots et leur disparité.

L'ambiguïté des mots, se rapporte à des mots lexicalement identiques et portant des sens différents, ce qui conduit à avoir des documents non pertinents en réponse à une requête contenant des mots ambigus (ex : documents sur le « *langage java* » peuvent être renvoyés en réponse à la requête « *aéroport de java* »). Alors que, la disparité des mots se réfère à des mots lexicalement différents mais portant un même sens. Ce problème implique que des

documents pertinents ne sont pas retrouvés en réponse à une requête, car ils utilisent des mots différents que ceux de la requête pour exprimer le même concept (ex : documents contenant le terme « *tablette tactile* » peuvent ne pas être retrouvés en réponse à une requête « *IPad* ») [36]. Nous présentons ci-dessous trois types d'approches les plus utilisées pour remédier à ces problèmes, l'indexation par des mots composés, l'indexation sémantique et l'indexation conceptuelle.

I.1.7.1 L'indexation par des mots composés

L'indexation par des mots composés est une technique qui permet l'utilisation des mots composés comme unités d'indexation, car ils sont moins ambigus et plus précis que les mots simples (ex : le terme « *java* » est ambigu, en revanche les mots composés « *aéroport de java* » et « *langage java* » sont non ambigus). Ceci a pour objectif une représentation plus précise du contenu sémantique des documents et des requêtes. On trouvera plus de détails sur les paramètres qui sont considérés dans l'exploitation des mots composés comme unités d'indexation dans [37] [38] [39].

I.1.7.2 L'indexation sémantique

L'indexation sémantique se base sur la technique de désambiguïsation sémantique (Word Sense Disambiguation WSD), qui a pour but la sélection du sens approprié pour un terme dans un contexte donné [40]. Ce type d'indexation consiste à représenter les documents et les requêtes par les sens des termes qu'ils contiennent plutôt que par les termes eux-mêmes.

Plusieurs études ont été menées sur cette approche, nous passons en revue celles réalisées par « *Voorhees* ». Il a proposé une méthode de désambiguïsation basée sur Wordnet, qui est une structure conceptuelle organisée autour de la notion de synset. Un synset regroupe des termes (simples ou composés) ayant un même sens dans un contexte donné. Pour déterminer le sens d'un mot ambigu, les synsets (sens) de ce mot sont classés en utilisant la valeur de cooccurrence calculée entre le contexte de ce mot et un voisinage contenant les mots du synset dans la hiérarchie de WordNet. Le synset le mieux classé est alors choisi comme le sens approprié du mot ambigu analysé. [41]

I.1.7.3 L'indexation conceptuelle

L'indexation conceptuelle consiste à représenter un document par un ensemble de concepts qui sont tirés de structures conceptuelles (ex : WordNet pour la langue anglaise, MESH pour le domaine médical). Ces structures incluent les taxonomies de concepts, les ontologies, les réseaux sémantiques, les dictionnaires, les thésaurus, etc. Plusieurs travaux en

RI ont utilisé ce type d'indexation, comme « *Baziz et al* » [15] qui ont défini une méthode d'indexation conceptuelle basée sur l'utilisation de l'ontologie linguistique WordNet. Chaque document est représenté sous forme d'un réseau sémantique particulier (appelé noyau sémantique), dans lequel les nœuds représentent les Concepts et les arcs (bidirectionnels) représentatifs de la distance sémantique entre concepts liés.

I.1.8 L'évaluation d'un SRI

L'évaluation des SRI est abordée selon deux angles différents. L'un est dit « *paradigme système* », qui vise à évaluer les performances du système qui doit être capable de trouver tous les documents pertinents et rejeter tous les documents non pertinents afin de répondre de façon satisfaisante aux besoins d'information de l'utilisateur en termes de qualité des résultats retournés, de rapidité du système ainsi que la facilité d'utilisation du système qui représentent les principaux facteurs à évaluer pour un SRI. L'autre est dit *« paradigme usager* », qui est centré sur la satisfaction de l'utilisateur, et non sur les performances intrinsèques du système, en modélisant le comportement des utilisateurs en situation de recherche. [42]

I.1.9 Conclusion

Dans ce chapitre nous avons passé en revue les principaux concepts de la RI. Nous avons, particulièrement, introduit des notions de base, telles que le besoin en information, la requête, le document et la pertinence. Nous avons aussi décrit les processus de base de la RI, à savoir l'indexation, l'appariement requête-document et la reformulation de la requête. Ensuite, nous avons étudié les différents modèles de la RI. Enfin, l'évaluation des systèmes de recherche d'information est traitée.

Avec l'avènement du web et surtout l'émergence des technologies web 2.0, les utilisateurs sont devenus des producteurs de l'information, et donc le document présente de nouvelles dimensions (sources d'information) autre que le contenu textuel.

Ces nouvelles sources d'information sur le document doivent être intégrées dans les modèles de RI, afin d'améliorer les performances de la recherche d'information. Dans le chapitre suivant, nous abordons la recherche d'information sociale.

Chapitre 2

I.2 RECHERCHE D'INFORMATION SOCIALE

I.2.1 Introduction

Dans le contexte du web 2.0 et avec l'émergence des blogs, des wikis et des réseaux sociaux, l'utilisateur ne passe sans laisser sa trace. Il est producteur d'information et non plus consommateur uniquement. L'information produite est dite « *information sociale* », qui est donc toute information fournie par l'usage du web 2.0 : les tags utilisés pour l'annotation, les traces de l'utilisateur (navigation sur le web, visualisation des pages web et documents...), les relations entre les utilisateurs et les profils des utilisateurs. [3]

L'information sociale est utilisée pour prédire les intérêts et les intentions des utilisateurs. Elle est incorporée dans le processus de la RI pour améliorer la recherche et rendre à l'utilisateur la réponse la plus pertinente à son besoin.

Ce chapitre présente la recherche d'information sociale. Dans la première section, nous présentons les généralités et les bases théoriques concernant la définition des médias sociaux et ses différents types y compris les réseaux sociaux. Dans la seconde section nous définissons la notion de la RI sociale, et les distinguons de la RI classique. Nous décrivons dans la troisième section les différentes approches liées à l'exploitation des informations sociales dans le processus de la RI, dans le but d'améliorer la pertinence de la recherche d'information.

I.2.2 Distinction entre Web social, Médias Sociaux, Réseaux sociaux :

I.2.2.1 Web 2.0

La notion de Web 2.0 marque une évolution du Web vers plus de simplicité et d'interactivité (chaque utilisateur peut contribuer). Le Web 2.0 désigne l'ensemble des techniques, des fonctionnalités et des usages permettant aux internautes ayant peu de connaissances techniques de s'approprier les nouvelles fonctionnalités du Web pour échanger de l'information, interagir, partager… créant ainsi le Web social. [43]

I.2.2.2 Web social

Le web social fait partie du Web 2.0 et se concentre sur les structures sociales et les interactions sur Internet. Le terme "web social" comprend des applications basées sur le Web qui soutiennent l'échange d'informations, la création de relations et la communication dans un contexte social. Il présente un espace de socialisation sur Internet, dont l'objectif est de regrouper l'ensemble des applications qui permettent la production et le partage des contenus et aussi l'accès à des données diversifiées [44]. Le web social est le terme générique qui englobe les médias et les réseaux sociaux.

I.2.2.3 Médias sociaux

Les médias sociaux est l'appellation des services internet et mobiles qui autorisent aux utilisateurs de communiquer en ligne, de partager des contenus, en s'ouvrant sur d'autres communautés électroniques. D'après *Dupin*en 2010 :*« Les médias sociaux peuvent se définir comme l'ensemble des plateformes en ligne créant une interaction sociale entre différents utilisateurs autour de contenus numériques (photos, textes, vidéos) et selon divers degrés d'affinités»*. [45]

Figure 5: Panorama des médias sociaux 2019 [46]

Comme le montre la figure ci-dessus, les médias sociaux regroupent plusieurs services certains sont des réseaux sociaux. Ces services peuvent être classés en deux types :

Le premier, concerne les médias sociaux généralistes qui sont tous les outils mettant en jeux des technologies, de la création de contenu et des interactions entre les hommes. Ils peuvent être exploités sur des ordinateurs et des téléphones afin de créer des différents contenus. Alors que le seconde, représente les médias sociaux spécialisés qui ont pour fonction ludique et professionnelle. Le tableau suivant montre des exemples de chacun de ces types :

Médias sociaux généralistes	**Médias sociaux spécialisés**
L'e-mail (e-mailing):Gmail, Yahoo mail, Hotmail, Outlook…	Les médias sociaux de multimédia : partage de vidéos, jeux en ligne…
La messagerie instantanée : MSN, Yahoo Messenger, Google Talk…	Les médias sociaux de localisation et géolocalisation : GPS…
Les réseaux sociaux :Facebook, Twitter, LinkedIn…	Les médias sociaux d'Achats/Ventes : shopping en ligne.
La gestion d'événement :Google, Agenda, Facebook…	Les médias sociaux de recherche et de veille :trouver des personnes, blog…
Les questions /réponses :Yahoo Answers, LinkedIn, Viadeo…	Les médias sociaux dédiés à l'internet : sauvegarde des favoris sur internet.
Les jobs/recrutements: Youjob, CarriereOnLine.com, LinkedIn…	Les médias sociaux professionnels : diffusion et partage de contenu sur internet.
Les blog/pages :Blogger, OverBlog, WordPress…	Les médias sociaux pour la création des communautés : création des contenus.

Tableau 3: Types de médias sociaux *[47]*

En regardant le tableau ci-dessus (tableau 3), il devient clair qu'il existe une différence entre les médias sociaux et les réseaux sociaux. *Dubin* voit que les réseaux sociaux comme des sites reposant sur un lien social et les médias sociaux comme l'ensemble des sites proposant une interaction sociale. Dans le premier cas, c'est l'individu qui est au centre des échanges, alors que pour le second c'est l'ensemble des objets présents qui favorisent l'interaction [48].

Dans cette logique, les réseaux sociaux sont une partie des médias sociaux. Ils sont la plus pure représentation du terme "social" qui connote la relation entre différents individus et dont l'expression se centralise par un profil utilisateur.

I.2.2.4 Réseaux sociaux

Le terme « *Réseaux Social* » est apparu pour la première fois en 1954, employé par le sociologue anglais *Barnes*. Depuis, ce terme qui a été largement repris, a bien évolué. Dorénavant, nous ne définirions plus les réseaux sociaux uniquement d'un point de vue sociologique, mais aussi d'un point de vue technologique. [48]

D'un point de vue sociologique et selon Lazega, un réseau social peut être défini comme : *«un ensemble de relations spécifiques (par exemple collaboration, soutien, conseil, contrôle ou influence) entre un ensemble fini d'acteurs* » [49]. D'un point de vue technologique, *Kaplan et Haenlein* définissent les réseaux sociaux comme étant : « *des outils permettant aux individus de se connecter en créant des profils contenant des données personnelles, en invitant des amis et collègues dans le but d'avoir accès à ces profils. Ces données personnelles peuvent contenir n'importe quel type d'information comme des photos, vidéos, fichiers audios, et blogs* ». [50]

Ainsi, selon *Esther Dyson*, le réseau social sur Internet peut-être définit en reprenant les aspects sociaux et technologiques, de la façon suivante : *« les réseaux sociaux fournissent des outils qui facilitent le processus de mise en relation autour d'un centre d'intérêt commun et permettent la prise de contact en ligne* ». [51]

Figure 6: Example d'un réseau social *[52]*

D'après ce que représente la figure 6, le réseau social est donc un ensemble d'acteurs qui partagent plusieurs relations avec d'autres. Les acteurs réfèrent principalement à des personnes, mais représentent, dans un autre contexte, les institutions, les communautés, les éléments d'information, etc.

I.2.2.4.1 Diversité des réseaux sociaux

Aujourd'hui, Les réseaux sociaux sont incontournables dans notre vie quotidienne, notamment dans le domaine de la communication et de l'actualité. Nous listons ci-dessous quelques exemples de réseaux sociaux :

- **Les réseaux personnels (Facebook) :** axés sur les centres d'intérêt. D'après Dupin, « *les réseaux sociaux personnels sont créateurs d'un lien social autour de thématiques individuelles* ». [48]

  « *Facebook* [1]» est un réseau social en ligne qui permet à ses utilisateurs de publier des images, des photos, des vidéos, des fichiers et documents, d'échanger des messages, joindre et créer des groupes et d'utiliser une variété d'application. Facebook est fondé en 2004, il compte aujourd'hui, selon Mark Zuckerberg son fondateur, plus d'un milliard d'utilisateurs actifs. [53]
- **Les réseaux de médias (YouTube) :** axés sur la diffusion de contenu vidéos, photos, musique. [54]

  « *YouTube*[2] » est un site web d'hébergement de vidéos et un média social sur lequel les utilisateurs peuvent envoyer, regarder, commenter, évaluer et partager des vidéos. Il a été créé en février 2005 par Steve Chen, Chad Hurley et Jawed Karim, trois anciens employés de PayPal, et racheté par Google en octobre 2006 pour 1,65 milliard de dollars. Le service est situé à San Bruno, en Californie. [55]
- **Les réseaux d'actualité (Twitter) :** axés sur la diffusion d'informations. D'après *Fanneli-isla,* ce sont des sites où peuvent se mélanger professionnels et internautes pour diffuser, relayer et commenter l'information mondiale. [54]

  « *Twitter* [3]» est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet. Twitter a été créé le 21 mars 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass. Le service en ligne est rapidement devenu populaire, son siège social se situe aux États-Unis à San Francisco [56].
- **Les réseaux professionnels (LinkedIn) :** axés sur le carnet d'adresses et les échanges professionnels.

[1] https://www.facebook.com/
[2] https://www.youtube.com/
[3] https://twitter.com/

« *LinkedIn*[4] » est un réseau social à utiliser dans un contexte d'affaire. Les pages des utilisateurs exposent leurs carrières professionnelles et leur permettent de préciser leurs intérêts en matière de débouchés professionnels, d'emplois, loisirs et autre vie sociale, et cela en partageant des liens, textes, vidéos, etc. LinkedIn est créé en mai 2003 par *Reid Hoffman* et *Allen Blue*. Selon les dernières statistiques plus de 259 millions de professionnels dans le monde sont inscrits, ainsi que plus de 150 secteurs d'activité dans 200 pays sont présents sur LinkedIn. [57]

En se référant au graphique ci-dessous on peut établir un lien direct avec l'impact que la connexion peut avoir sur ces réseaux pour une entreprise sociale. [58]

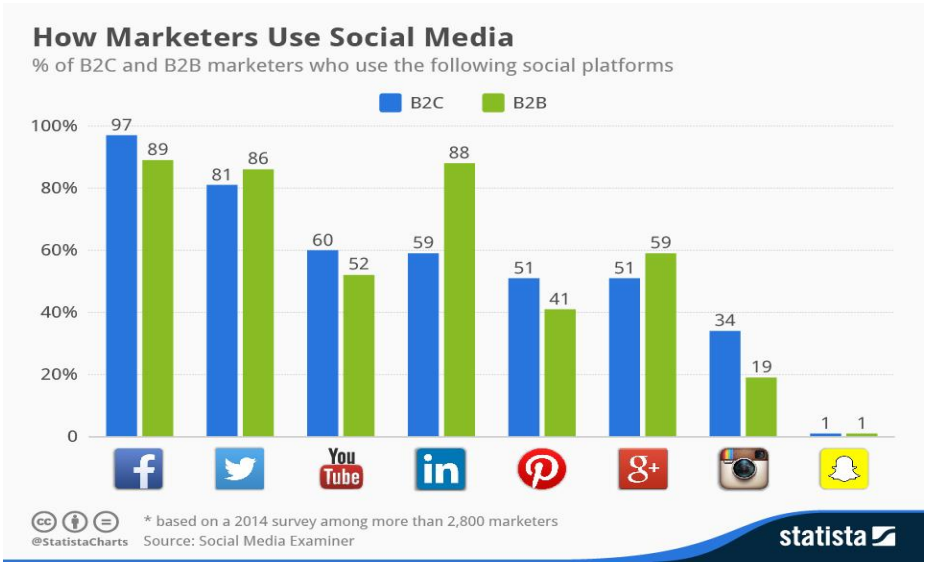


***Figure 7: Marketing des médias sociaux** [58]*

Où les barres bleues représentent les médias sociaux les plus utilisés par le grand public (B2C), tandis que les barres colorées en vert représentent les médias sociaux employés par les professionnels (B2B).

En ce qui concerne la science de l'information, une évolution importante apportée par ces médias a été la prolifération des contenus générés par l'utilisateur.

I.2.2.4.2 Contenus générés par les utilisateurs

Le terme UGC (Les contenus générés par l'utilisateur) est cité plusieurs fois et avec différentes définitions selon sa nature informationnelle (tag, j'aime, commentaire, vidéo, etc.). Selon *Baeza-Yates* l'UGC est défini comme suit :

[4] https://www.linkedin.com/

"User Generated Content is one of the main current trends in the Web. This trend has allowed all people that can access the Internet to publish content in different media, such as text (e.g. blogs), photos or video." [59]. Le contenu généré par l'utilisateur n'est pas uniquement un document, une image ou une vidéo partagée ou créée par l'utilisateur. Il existe plusieurs types de contenu y compris les signaux sociaux.

- **Signaux sociaux :** un signal social est donc une mesure de l'activité des médias sociaux. C'est une interaction sociale d'une personne réelle avec une ressource sur le Web à travers des fonctionnalités offertes par les réseaux sociaux. Le tableau 4 résume les signaux sociaux les plus populaires sur les réseaux sociaux :

Type	Exemple	Réseaux Sociaux
Commentaire	Commentaire, Répondre	Facebook, Google+, LinkedIn, Twitter
Message	Tweet, Publication	Facebook, Google+, LinkedIn, Twitter
Partage	Partage, Re-tweet	Facebook, Google+, LinkedIn, Twitter
Vote	J'aime, +1	Facebook, Google, LinkedIn, StumbleUpon
Relation	Abonnés, Amis	Facebook, Twitter

Tableau 4: Types de signaux sociaux *[25]*

Dans le contexte des réseaux sociaux, ces contenus générés par les utilisateurs ont conduit à l'émergence de nouveaux problèmes en recherche d'information, qui a donné la naissance à une nouvelle thématique en RI, la Recherche d'Information Sociale (RIS).

I.2.3 Notion de la RI sociale

La recherche d'information sociale (RIS) est une nouvelle branche de recherche d'information (RI) qui bénéficie du cadre social pour avoir un reclassement meilleur des résultats de recherche. [60]

En 2006, *Kirsch* et ses collègues ont défini la recherche d'information sociale par la prise en compte des données des réseaux sociaux dans le processus de recherche d'information :

"Social information retrieval systems are distinguished from other types of information retrieval systems by the incorporation of information about social networks and relationships into the information retrieval process." [61]

Une définition générale par *Alonso* en 2011 considère les réseaux sociaux comme une source d'intelligence collective :

"Social search is a general term used to describe searches that utilize social networks or involve a collective intelligence process to help the user satisfy an information need." [62]

I.2.4 Différences entre la RI classique et la RI sur le web social

Le web social diffère sur plusieurs points avec les autres ressources documentaires rencontrées habituellement en recherche d'information, où les collections de documents sont généralement de petites tailles et les utilisateurs ont des besoins en informations bien spécifiques. Nous décrivons brièvement les facteurs distinctifs qui peuvent être considérés comme défis dans le tableau 5:

Facteur	Description
Le volume	Le stockage, l'accès et l'analyse des quantités massives d'informations sociales (Big Data) nécessitent des études rigoureuses.
La structure	La diversité des structures de réseaux sociaux apporte des difficultés supplémentaires.
La nature dynamique	La nature dynamique du web (les informations se changent fréquemment) rend la mise à jour et la maintenance des index des moteurs de recherche extrêmement difficile.
L'hétérogénéité	L'inclusion de plusieurs thèmes qui sont traités sur le web sous différents formats par diverses sources, a pour effet d'augmenter le bruit lors d'une recherche.
La disparité	L'occurrence disséminée de l'information dans de larges collections de documents a pour effet d'augmenter le silence en recherche d'information sur le web.
La fiabilité	L'information récupérée peut être une bonne information, une information non complète ou une information fausse ce qui est plus nuisible.
Le requêtage	Les courtes requêtes rendent difficile la tâche de sélection d'information désirée parmi le grand nombre de ressources qui répondent aux besoins des utilisateurs.
La pertinence	La pertinence sociale est le degré de popularité d'un document exprimée par les activités sociales y relatives dans le réseau social.

Tableau 5: Facteurs distinctifs de la RI sociale *[63]*

I.2.5 Modèles de recherche sociale

Les contenus sociaux ont conduit à l'émergence de nouvelles approches afin de satisfaire des motivations sociales derrière les besoins d'information de l'utilisateur, ainsi qu'à la réactualisation des tâches de RI pour mieux appréhender les données sociales. Nous discutons donc la RIS selon 2 portes :

Le premier, concerne la recherche d'information de nature sociale. Alors que le deuxième, porte sur l'exploitation des contenus sociaux pour améliorer la RI, dans laquelle l'information sociale est utilisée afin d'améliorer le processus de recherche d'information :

I.2.5.1 Recherche d'information dans les contenus sociaux

Les contenus sociaux qui sont générés par les utilisateurs actifs sur les différents réseaux sociaux tels que Twitter et YouTube permettent de trouver des nouvelles informations sociales qui répondent aux utilisateurs (exemple. La recherche des experts, la recherche de conversations, la recherche d'information dans les blogs, microblogs, ou encore des réponses à des questions spécifiques auprès des amis, familles, collègues, etc.). [25]

I.2.5.2 Exploitation des contenus sociaux pour améliorer la RI

L'information sociale joue un rôle très important dans l'amélioration du processus de la RI classique, qui peut être utilisée comme une nouvelle source d'évidence pour améliorer l'index, reformuler les requêtes à l'aide de connaissances supplémentaires et reclasser (*re-ranking*) les documents retournés par un SRI. Nous allons donc détailler l'exploitation des contenus sociaux dans ces trois niveaux principaux [25] **:**

I.2.5.2.1 Indexation sociale

L'information sociale peut être utile pour les documents qui contiennent quelques termes où le processus d'indexation simple ne fournit pas une bonne performance de RI, donc elle est utilisée pour améliorer la représentation du document, soit par l'ajout de métadonnées sociales au contenu de ce document pour l'enrichir afin de permettre aux auteurs d'indexer le document à la fois avec son contenu textuel et ses contenus sociaux associés (tags et commentaires), ou bien en personnalisant la représentation des documents par des indexes personnalisés, en supposant que chaque utilisateur a sa propre vision sur un document donné (exemple. Décrire, commenter, annoter, etc.).

I.2.5.2.2 Reformulation de la requête

L'utilisation de L'information sociale dans la reformulation de la requête permet d'améliorer ce processus, soit par réduire la requête de telle sorte que l'information inutile soit éliminée (Raffinement), ou bien par rajouter de nouvelles informations à la requête initiale pour la rendre moins ambiguë et élargir son champ de recherche (Expansion). Koolen et ses collègues [64] proposent une approche d'expansion de requêtes utilisant Wikipédia comme collection externe. En outre, Bao et ses collègues [65] suggèrent l'utilisation d'un graphe bipartite entre les annotations sociales et les pages Web avec des arêtes indiquant le nombre d'utilisateurs, en utilisant des algorithmes « SocialSimRank et SocialPageRank » (voir la figure 8).

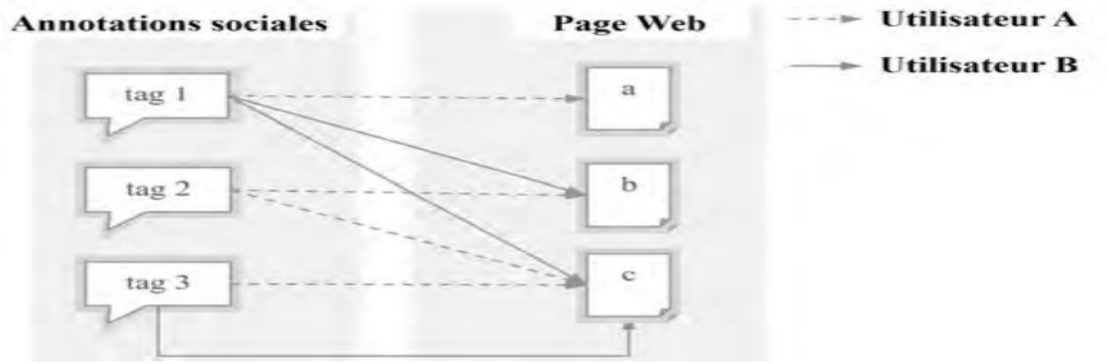


Figure 8: Graphe SocialSimRank de Bao [65]

Le SocialSimRank est utilisé comme une forme d'expansion de requête, où des tags similaires sont inclus dans le calcul de similarité entre la requête et le document.

I.2.5.2.3 Reclassement de résultats

L'idée générale du reclassement de résultats est de quantifier les similarités entre les documents et les requêtes en utilisant l'information sociale. Cette dernière peut être utilisée par deux classes différentes pour améliorer le classement des documents retournés vis-à-vis d'une requête, soit en se basant sur la pertinence sociale qui se réfère à des facteurs sociaux qui caractérisent un document en termes (d'intérêt social, sa popularité, sa réputation, etc.), en utilisant différents algorithmes par exemple, l'algorithme que nous avons mentionné précédemment « *SocialPage Rank* » [65] qui calcule la qualité de la page (popularité) par le nombre d'annotations sociales. Ou bien en se basant sur la fonction personnalisée pour améliorer les résultats de recherche, qui peut être utilisée pour trier les documents différemment selon chaque utilisateur afin de satisfaire leurs intérêts, profils et habitudes différentes.

I.2.6 Approches basées sur les signaux sociaux

La majorité d'approches se situent dans le contexte de l'annotation sociale, d'autres critères sociaux, en particulier les signaux sociaux, sont exploités pour mesurer la pertinence d'une ressource afin d'améliorer la recherche. Ces approches peuvent être classées selon deux classes :

I.2.6.1 Signaux sociaux indépendants du temps

Ces dernières années, de nouvelles approches se concentrent sur la façon d'améliorer la recherche d'information en exploitant les actions des utilisateurs, cependant ces approches ne prennent pas en compte le moment où l'action s'est produite et le moment où la ressource a été publiée.

De nombreux travaux s'intéressent à exploiter les caractéristiques sociales pour améliorer la recherche d'information sur le Web. Les chercheurs ont étudié l'impact des signaux sociaux (*aime*, *n'aime pas*, *commentaire,* etc.) sur l'efficacité de la recherche sur YouTube en utilisant des techniques de sélection d'attribut ainsi que des fonctions d'apprentissage d'ordonnancement, ces critères sociaux sont utiles dans l'amélioration du classement des résultats pour la majorité des requêtes. [66]

D'autre études proposent une approche sociale appelée « *Social Score Method »* basée sur plusieurs signaux sociaux issus de différents réseaux sociaux, afin de déterminer quelles ressources devraient être retournées en premier. Le score social est estimé avec un simple comptage des signaux (*partage, bookmark, tweet*), et est combiné avec le TF-IDF. [67]

En générale, les signaux sociaux sont utiles en RI, car l'utilisateur peut bénéficier de ces actions sociales de diverses façons, y compris la découverte de recommandations sélectionnées socialement, la recherche personnalisée, estimer la popularité des pages Web, etc. [68]

I.2.6.2 Signaux sociaux dépendants du temps

Contrairement aux travaux dont nous avons discuté précédemment, il existe d'autres approches qui sont basées sur la temporalité des signaux sociaux c.-à-d. prennent en compte le moment où l'action s'est produite et le moment où la ressource a été publiée.

En 2010, les chercheurs proposent une méthode appelée « *ClickBuzz* » comme une mesure pour déterminer si une page Web reçoit un niveau inhabituel d'intérêt des utilisateurs par rapport au passé, En basant sur le nombre de clics sur le document au cours d'un intervalle de temps donné pour améliorer la qualité des résultats de recherche en favorisant les URL qui ont un intérêt récent pour les utilisateurs. [69]

En 2012, d'autres travaux sont intéressés juste à l'évaluation des intérêts des usagers dans le temps, en examinant comment les utilisateurs produisent et consomment la grande masse des contenus générés par eux dans les réseaux sociaux au fil du temps, puis ces intérêts sociaux des utilisateurs sont classés en cinq classes ("recent", "ongoing", "seasonal", "past" et "random"), enfin les données de Twitter ou bien Facebook sont analysées sur les activités sociales des usagers. [70]

Dans le domaine de la recherche d'information sociale (RSI), nous pouvons conclure que la gestion et l'exploitation des contenus sociaux (exemple. Des *annotations* sociales, des *clics*, des *tweets*, des *commentaires*, *des relations sociales*, des *actions* tels que le *j'aime*, le *partage*, le *+1*, etc.) générés par les utilisateurs actifs dans ces réseaux sociaux ; conduit à une émergence de nouvelles informations qui répondent à un besoin spécifique, ainsi qu'une amélioration de la pertinence des résultats pour mieux décrire une certaine importance du côté de la ressource.

I.2.7 Positionnement et Conclusion

Dans cet état de l'art, nous avons passé en revue de nombreuses approches existantes selon 3 axes : 1) le premier axe concerne la recherche d'information de nature sociale. Il s'agit de trouver des informations sociales qui répondent à l'utilisateur. On distingue par exemple la recherche d'information dans les blogs, microblogs et la recherche de conversations ; 2) le deuxième porte sur l'exploitation des contenus sociaux pour améliorer la RI, dans laquelle l'information sociale est utilisée afin d'améliorer le processus de recherche d'information, par exemple, les tags pour améliorer la recherche Web et la recherche personnalisée , le reclassement (re-ranking) des résultats de recherche et 3) le troisième paradigme concerne la recherche d'information effectuée par plusieurs personnes, recherche collaborative. [25]

Les travaux les plus liés à notre travail incluent le deuxième axe. Ces travaux s'intéressent à l'exploitation des caractéristiques sociales pour améliorer la RI sur le Web et sur les réseaux sociaux. Nous proposons donc une approche de RI sur YouTube, Twitter et Facebook qui permet d'allier efficacité et simplicité de mise en œuvre.

Pour le chapitre suivant, notre approche s'inspire de l'une des contributions de I. Badache [25] que nous proposons d'améliorer, nous partons donc de l'hypothèse qu'un document doit être reclassé (re-ranking) en fonction de sa pertinence thématique et sa pertinence sociale. Nous allons donc présenter le cœur de notre travail, en montrant l'expérimentation de notre système et les résultats obtenus.

Partie II

II. CONCEPTION ET RÉALISATION DE L'APPLICATION

Chapitre 3

II.1 CONCEPTION ET RÉALISATION DE L'APPLICATION

II.1.1 Introduction

Dans ce chapitre, nous décrivons notre approche pour l'exploitation des signaux sociaux laissés par les utilisateurs sur les ressources pour mesurer la pertinence (l'intérêt) d'une ressource. Cette connaissance est combinée avec la pertinence thématique dans un modèle de recherche vectoriel qui prend en compte explicitement ces sources d'évidence.

Nous évaluons la performance de notre approche sur des collections de données tirées de YouTube, Twitter et Facebook.

II.1.2 Environnement de travail

Nous présentons brièvement notre environnement de travail en montrant les principaux outils, langages et bibliothèques utilisées pour la mise en œuvre de notre SRI :

II.1.2.1 Outils de développement

- **Anaconda**[5] **:** est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets Conda. [71]
- **Spyder**[6] **:**est un environnement de développement pour Python. Libre (Licence MIT) et multiplateforme (Windows, Mac OS, GNU/Linux), il intègre de nombreuses bibliothèques d'usage scientifique : Matplotlib, NumPy, SciPy et IPython. Créé et développé par Pierre Raybaut en 2008. [72]
- **VS code**[7] **: (**Visual Studio Code) est un éditeur de code open-source, gratuit et multi-plateforme (Windows, Mac et Linux), développé par Microsoft, à ne pas confondre avec Visual Studio, l'IDE propriétaire de Microsoft. Principalement

[5] https://www.anaconda.com/
[6] https://pypi.org/project/spyder/
[7] https://code.visualstudio.com/

conçu pour le développement d'application avec JavaScript, TypeScript, PHP et Node.js, l'éditeur peut s'adapter à d'autres types de langages grâce à un système d'extension bien fourni. [73]

- **XAMPP**[8] **:**est un ensemble de logiciels permettant de mettre en place un serveur Web local, un serveur FTP et un serveur de messagerie électronique. Il s'agit d'une distribution de logiciels libres (X (cross) Apache MariaDB Perl PHP) offrant une bonne souplesse d'utilisation, réputée pour son installation simple et rapide. Il permet de configurer un serveur de test local avant la mise en œuvre d'un site internet. [74]
- **Composer** [9]**:** est un logiciel gestionnaire de dépendances libre écrit en PHP. Il permet à ses utilisateurs de déclarer et d'installer les bibliothèques dont le projet principal a besoin. Le développement a débuté en avril 2011 et a donné lieu à une première version sortie le 1er mars 2012. [75]

II.1.2.2 Langages de programmation

- **Python** [10]**:** est un langage de programmation open source créé par le programmeur Guido van Rossum en 1991. Il s'agit d'un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. [76]
- **PHP** [11]**:** (Hypertext Preprocessor) est un langage de programmation libre, principalement utilisé pour produire des pages Web dynamiques via un serveur HTTP, mais pouvant également fonctionner comme n'importe quel langage interprété de façon locale. PHP a permis de créer un grand nombre de sites web célèbres, comme Facebook et Wikipédia. Il est considéré comme une des bases de la création de sites web dits dynamiques mais également des applications web. [77]

II.1.2.3 Bibliothèques principales

- **NLTK** [12]**:** (Natural Language Toolkit) est une bibliothèque logicielle en Python permettant un traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'université de Pennsylvanie. En

[8] https://www.apachefriends.org/fr/index.html
[9] https://getcomposer.org/
[10] https://www.python.org/
[11] https://www.php.net/
[12] http://nltk.org/

plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation (API). [78]

- **Wordnet**[13] **:** est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Par rapport aux outils fournis, un développeur peut aussi accéder à la base de données à partir des interfaces disponibles pour plusieurs langages de programmation (Java, Perl, PHP, Prolog, Python...). [79]
- **Pendas**[14] **:** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Pandas est un logiciel libre sous licence BSD. [80]

II.1.3 Architecture générale de notre SRI

Notre SRI (figure 9) possède trois fonctions fondamentales qui définissent le modèle de recherche : représenter le contenu des documents, représenter le besoin de l'utilisateur et comparer ces deux représentations. La représentation des documents et de la requête se fait à l'issue d'une phase appelée indexation qui consiste à choisir les termes représentatifs des documents et à les ajouter à un index qui à chaque terme associe le document dans lequel il se trouve. En effet, la préparation des documents ainsi que les requêtes est l'étape la plus importante de notre SRI. Elle se base sur une représentation vectorielle des données. Par la suite, le modèle effectue un appariement entre la requête saisie et les documents déjà préparés pour retourner le classement de ces derniers selon leur pertinence. Cette dernière est traduite par la similarité de leurs vecteurs associés. Le mécanisme d'appariement (syntaxique-sémantique) consiste donc à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête en s'appuyant sur deux mesures de similarité de Wu-Palmer [1] et Path [2]. Ces deux mesures exploitent les termes communs ou proches (synonymes) aux documents comparés pour évaluer leur ressemblance.

[13] https://wordnet.princeton.edu/
[14] https://pandas.pydata.org/

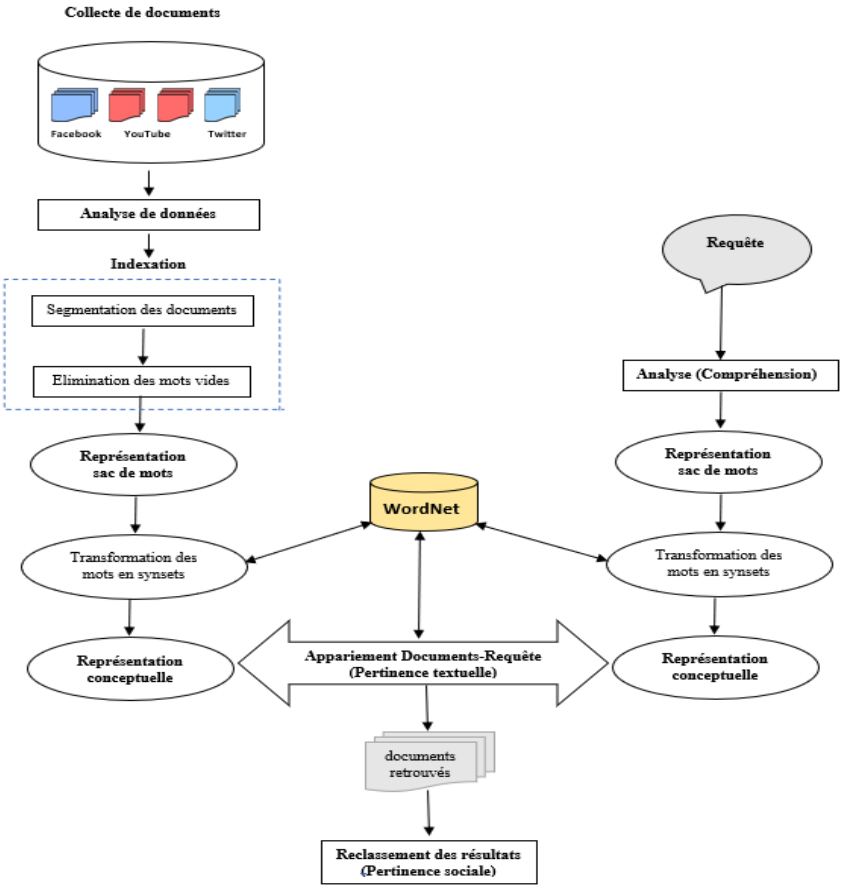


Figure 9: Architecture générale de notre SRI

La pertinence sociale se focalise sur les contenus textuels et non textuels (signaux sociaux) générés par des utilisateurs dans la recherche de ressources web (pages, vidéos, etc.). En particulier et afin d'améliorer le processus de tri des résultats de recherche, nous nous intéressons à identifier, extraire et quantifier , à partir de YouTube, Twitter et Facebook, les contenues textuels ainsi que leurs interactions sociales, telles que "Likes, Dislikes, Share, Reactions...". Ces dernières vont être intégrées dans le modèle de classement (ranking) combinant la pertinence thématique et la pertinence sociale.

II.1.3.1 Constitution du dataset

Pour évaluer notre approche, nous avons mené une série d'expérimentations sur les principales collections de chaines d'information (CNN, BBC NEWS, The New York Times, ABC News, RT, CBC News : The National, Fox News et Al Jazeera English) via Twitter, YouTube et Facebook. Le choix de ces collections se justifie par des raisons techniques. Il est donc nécessaire d'extraire ces données à l'aide du « Scraping » [81] qui est une technique d'extraction du contenu de sites Web, via un script ou un programme.

Afin de collecter les contenus textuels et leurs signaux sociaux non textuels pour chaque ressource web, il est primordial d'avoir les URLs des documents pour pouvoir les extraire à travers les APIs des réseaux sociaux qui prennent en argument ces URLs.

Afin de construire notre « Dataset » mentionnée dans la figure 19 et de collecter les informations textuelles et sociales, nous utilisons une API de réseau social « Twitter » et deux APIs « YouTube »: l'une pour les vidéos, alors que l'autre est pour les commentaires de ces vidéos. Leur algorithme de collecte des données est décrit comme suit:

Entrée : tableau de paramètres
Sortie : fichier csv
Début :
Connecter l'API
Data ◄——— API.getData (paramètres)
Ouvrir le fichier csv
Pour chaque ligne de Data **faire**
Sélectionner les signaux sociaux utiles
Ajouter la ligne au fichier csv
Finpour
Fermer le fichier csv
Fin.

Figure 10: Algorithme de collecte des actions sociales

Nous allons expliquer cet algorithme en détaillant chaque API séparément :

1. **Collecte des actions sociales « Twitter » :** cet algorithme nous a permis d'extraire des données liées au réseau Twitter en identifiant d'abord les paramètres suivants :

   Screen_name : pour déterminer la source des documents à collecter. Dans notre cas, nous choisissons les chaines d'information (CNN et BBC), car elles contiennent une variété de sujets qui aident à traiter différentes requêtes.

   Count : signifie le nombre de documents à collecter. Dans notre cas, le nombre maximal sélectionné est de 200 documents.

   Max_id : pour collecter les documents afin de profiter du grand nombre de leurs signaux en fixant la date de publication. Dans notre cas, les documents sélectionnés représentent ceux publiés avant la fin septembre.

   Trim_user : signifie la récupération des informations de la page (date de création, auteur, adresse…). Ce type d'informations est à éviter car nous ne nous intéressons qu'au contenu des pages.

Après avoir collecté les documents, nous éliminons les données inutiles telles que la date de création, URL, les mentions d'utilisateur…, et gardons uniquement les données utiles (Id, Tweet, nombre de Like et nombre de Retweet) c.-à-d. Les signaux sociaux qui peuvent être des critères de pertinence (Like, Retweet), comme le montre la figure 11.

Enfin, ces résultats sont stockés au format CSV, qui est un format texte ouvert représentant des données tabulaires sous forme de colonnes séparées par des virgules:

post_id	text	like_count	retweet_count
1300639996173582336	BTS: K-pop group reacts with 'tears' after making Billboard history htt	20605	5509
1301660270297788416	India NEET, JEE exams: 'They are playing with student lives' https://t.c	4121	2959
1301284163874631680	RT @BBCSport: Brazil's women's national players will now be paid the	0	6724
1301057826287222784	Honeybee venom 'kills some breast cancer cells', Australian scientists	2975	854
1300976569087913984	Nancy Pelosi pictured without mask in hair salon https://t.co/ZfV4q79	2969	855
1301487854619684864	National Security Agency surveillance exposed by whistleblower Edwa	2542	858
1301144045813673984	Israeli soldier condemned for putting knee on Palestinian protester's	2370	840
1301797279498752000	Daredevil David Blaine takes flight with helium-filled balloons over Ar	2558	553
1301917216238186496	'Third sighting in 100 years' of blue whale off coast of Sydney, Austral	2569	489
1301169574990221312	Gravitational "shockwave" from black hole collision half-way across u	2217	685
1301157383041880064	RT @BBCBreaking: Russian opposition politician Alexei Navalny was p	0	2812
1301071872419942400	Riot police have been filmed by the BBC attacking and forcibly arresti	1659	974
1301089692633399296	Thailand's king reinstates his consort after she was accused of "disloy	649	1943
1300725922233020416	Rare rallies in China over Mongolian language curb in schools https:/	1325	897
1301344134087741440	Dwayne 'the Rock' Johnson: Actor and family had Covid-19 https://t.c	1586	520
1302018208900280320	Coronavirus: Russian vaccine shows signs of immune response https:/	1591	499
1301099229260271616	Ugandan gorilla family has 'baby boom', with five births in six weeks f	1762	250
1301739550511955968	Hong Kong security law: UN experts voice deep concerns https://t.co/	1229	775
1301819549470294016	Britney Spears appears to endorse the Free Britney movement, as she	1585	377
1301431973597110272	Hong Kong media tycoon Jimmy Lai found not guilty in intimidation t	1128	801
1301810126643494912	RT @bbcworldservice: Covid-19 is known to affect the lungs and resp	0	1843
1301298251002503168	Coronavirus in Africa: Could poverty explain mystery of low death rat	1161	504
1301180618932391936	New Taiwan passport shrinks words 'Republic of China' https://t.co/jvl	1304	340
1301518785241849856	Facebook to freeze political ads before US presidential election https:	1336	269

Figure 11: Collection des actions sociales « Twitter »

2. **Collecte des actions sociales « YouTube » :** comme mentionné précédemment, la collecte de données du réseau YouTube est divisée en deux extractions: la première est l'extraction des interactions sociales des vidéo et la seconde est pour celles des commentaires. Nous allons expliquer la collecte de documents pour chaque collecte séparément :

- **Collecte des actions sociales « vidéo » :** nous suivons les étapes de l'algorithme précédent, en donnant d'abord les paramètres API mentionnés ci-dessous:

  Channel_Id : son rôle est de déterminer la source des données de la vidéo à collecter. Dans notre cas, nous choisissons les chaines d'information (CNN, BBC NEWS, The New York Times, ABC News, RT, CBC News : The National, Fox News et Al Jazeera English).

  Max_Results : signifie le nombre maximal de vidéos à collecter. Nous sélectionnons 30 vidéos (le maximum).

Order : représente l'ordre des vidéos extraites. Dans notre cas, les vidéos sont classées en fonction du nombre de vues (vidéos les plus vues).

Ensuite, nous supprimons les données inutiles des documents extraits (ex. tags, description, audio language, etc.). Enfin, nous obtenons les résultats finaux de cette collection (Id, titre de la vidéo, nombre de vues, nombre de réactions (J'aime, Je n'aime pas)et nombre de commentaires (S'il y a une vidéo sans commentaires, le nombre est zéro)) comme mentionné ci-dessous :

video_id	title	view_count	reaction_count	comment_count
smkyorC5qwc	The Third Presidential Debate: Hillary Clinton And Donald Trump (Full Deb	14761315	83953	28297
T9lpyl0ibiY	Egypt Opens Ancient Coffins To Find Perfectly Preserved Mummies | NBC	11224492	84710	10750
vHbLRf3C9KI	President Obama Remembers 'Biggest Disappointment' As President | NB	8051976	94017	17997
_FLo14GMYos	1980s: How Donald Trump Created Donald Trump | NBC News	7330559	57104	11999
kEQACtNXl6g	High School Football Team Too Good, Nobody Wanted To Play Them | NB	5875897	47504	4489
inqFCbCOSyE	L.A. Lakers Arrive At LAX Visibly Emotional After Kobe Bryant Death | NBC	5369448	29522	2620
KoBP74TWGSc	Boston Bomber Tsarnaev's Obscene Gesture Shocks Court | NBC Nightly	4438018	25467	2679
oejaHE5jUaA	Video Shows New Angle Of George Floyd's Arrest With Multiple Officers	4139768	16023	9714
7-WTLEMzbUw	Live: George Floyd Death Protests Around The U.S. | NBC News	3859183	42403	3959
B6K0-7oj_38	Live: SpaceX, NASA Cancels Launch Of U.S. Astronauts To International Sp	3807231	17055	1628
-HzOqZeX3Yk	Alex Jones Of 'Infowars,' Conspiracy Theories, And Trump Campaign (Full)	3139141	100869	24948
DI9ZLawGE4g	1990s: After Bankruptcies, Donald Trump Goes From Building To Branding	2947808	18925	3589
uYvTftH2rpE	NBC Nightly News Broadcast (Full) - July 13th, 2020 | NBC Nightly News	2506120	13672	4991
pnTMy0Yk8Ic	Nightly News Full Broadcast (April 5th)	2253362	14314	4948
btWcjFkn4E8	NBC News Special Report: Nationwide Protests Over Death Of George Flo	2189081	11400	5941
JVkmaBCXMx4	Meet The Press Broadcast (Full) - April 12th, 2020 | Meet The Press | NBC	2151141	9998	6347
PN3H0oFsQNg	NBC Nightly News Broadcast (Full) - June 6th, 2020 | NBC Nightly News	2147892	18510	7841
rRA1LTsyZqA	Attendee At Packed Memorial Day Gathering Tests Positive For Coronavir	2145154	12047	8038
k5clrgpXci4	NBC Nightly News Broadcast (Full) - May 27th, 2020 | NBC Nightly News	2141073	10575	3047
exsOim0Lyl4	President Donald Trump's State Of The Union Address 2018 (Full) | NBC N	2111237	41491	5279
jfiDh2yGn_g	Hasan Minhaj Calls Out Congress Over Student Loans: 'You Paid Far Less	2090119	95926	10371
WQ_VLCMdjxQ	Nightly News Broadcast (Full) - March 30th, 2020 | NBC Nightly News	2068280	12461	4480
jFO9hwRSrls	Watch Live Primary Night Coverage From NBC News NOW | NBC News (L	2054418	5097	843
I5_b3oH5MBM	NBC Nightly News Broadcast (Full) - May 21st, 2020 | NBC Nightly News	2044219	10984	2663

Figure 12: Collection de actions sociales « Vidéos_YouTube »

- **Collecte de actions sociales « commentaire »:** les paramètres utilisés pour extraire les actions sociales de commentaire dépendent des données liées aux vidéos YouTube précédemment présentées, nous allons les détailler comme suit:

  Video_Id : pour sélectionner la vidéo mentionnée précédemment comme entrée afin d'extraire les données de ses commentaires.

  Max_Results : signifie le nombre maximal de commentaires à collecter, nous récupérons le maximum (30 commentaires) pour chaque vidéo.

  Order : représente l'ordre des commentaires extraits, Dans notre cas, les commentaires sont classés en fonction de pertinence à leur vidéo (commentaires les plus pertinents afin de profiter du grand nombre de leurs signaux sociaux).

Contrairement aux signaux sociaux liés aux vidéos YouTube, leurs commentaires ne contiennent que des actions (j'aime, réponse) qu'après la suppression des données indésirables telles que le nom du commentateur et la date du commentaire.(voir la figure 13)

video_id	text	like_count	reply_count
smkyorC5qwc	Anyone watching this is September 2020?	791	61
smkyorC5qwc	Who's watching in September?	267	11
smkyorC5qwc	Who didn't care about politics much, but as 2020 is unfolding, you were forced to draw attention to it and v	80	3
smkyorC5qwc	30 YEARS of bad decisions,. That's what she brought!!!	156	3
smkyorC5qwc	Who's ready for the Biden vs Trump debates? They are gonna be halarious! 😂	2393	206
smkyorC5qwc	He should really get her where she belongs!! Prison	51	0
smkyorC5qwc	Cinton loves Guns. She sold Guns to the enemy. The Ambassador was going to blow the whistle. The day before	11	0
smkyorC5qwc	The rumor goes, that Hillary did wear a body skeleton to stand up straight. I nick name her Skillary 😈	10	1
smkyorC5qwc	Hilary should audition for the next Joker movie. That smile is creepy scary	185	11
smkyorC5qwc	Never laughed so hard. "We accept the results even when we lose."-hillary 2016. I wonder when this will happen.	118	15
smkyorC5qwc	Hillary's Mao Zedong outfit fit well with her views.	14	0
smkyorC5qwc	I love how Trump fixes his mic from time to time, like he's preparing to call her out on her bull.	29	1
smkyorC5qwc	I puke a little every time I hear Hillary's voice.	198	4
smkyorC5qwc	Does Biden know he's even running for President?	110	12
smkyorC5qwc	The look on Hillary's face lol wow whenever Trump talks	87	4
smkyorC5qwc	The look on crooked killary's face, I love it! She hates President Trump!	5	0
smkyorC5qwc	Based on the 3 debates I just watched I can say that USA is lucky to have a president like Trump. I think he deserv	66	4
smkyorC5qwc	Who is ready for Trump vs Biden debate???	1473	148
smkyorC5qwc	Trump saying he would question election results was absolutely genius. Hillary thinks she'll win and tells him to ac	10	1
smkyorC5qwc	I love listening to this now because her lies are so evident now	4	0
smkyorC5qwc	Her answers seem so scripted, almost like she had the questions before the debate....oh wait, she did. (And still lo	33	2
smkyorC5qwc	Clinton has no shame lying right to the American peoples face	122	7
smkyorC5qwc	look at how cocky she is lmao and she lost the election lololol	86	6
smkyorC5qwc	she thinks shes winning the whole time hahahah	11	1

Figure 13: Collection des actions sociales «Commentaires_YouTube »

3. **Collecte de données « Facebook » :** concernant le réseau social « Facebook », Nous réutilisons l'ensemble de données (Dataset) mentionné dans le lien ci-dessous[15]. En raison de la difficulté d'accès à ce réseau.

   Le réseau «Facebook» comprend de nombreux signaux sociaux qui sont exploités à partir des documents extraits. La figure 14 représente ces signaux avec leurs types, liens, numéros et dates de publication. Par la suite nous analysons ces données en ne laissant que les actions utiles.

[15] https://github.com/minimaxir/interactive-facebook-reactions/tree/master/data?fbclid=IwAR2jQQ7uG2kHsDV3O4utDVj0zwcO9-gwWTlJLHUTyE9mh1kBMIYYWMHplNo

status_message	link_name	status_type	status_link	status_publishe	num_reactions	num_comments	num_shares	num_likes	num_loves
Texas man arrested in the death of his infant daughter, whom he left in a	Dead Child's Da	link	http://abcn.ws/2	2016-06-22 18:4	3317	713	1374	800	6
Campaigners in England are making one last push with voters before tom	Campaigners To	link	http://abcn.ws/2	2016-06-22 18:1	89	15	14	81	3
Father shocked to discover that his two toddlers had covered their bedro	Dad Shocked by	video	https://www.fac	2016-06-22 17:5	10016	1238	3175	5627	126
Witnesses reported to police that a struggle ensued and the man pulled	Deputy Killed in	link	http://abcn.ws/2	2016-06-22 17:3	849	54	124	319	1
Two crew members stationed at a U.S. science facility at the South Pole a	2 Crew Member	link	http://abcn.ws/2	2016-06-22 17:1	1940	33	108	1742	102
Kentucky community rallies together to support a beloved, friendly neigh	Community Ralli	link	http://abcn.ws/2	2016-06-22 16:5	2218	107	273	1247	164
	ABC News Politi	video	https://www.fac	2016-06-22 16:5	1146	359	1	917	127
Family set to inherit gun stockpile worth millions plans to destroy the wea	Family to Destro	link	http://abcn.ws/2	2016-06-22 16:3	11208	2092	2211	8312	1321
Donald J. Trump painted Hillary Clinton as a "world class liar" in a speech	Fact Check: Trun	link	http://abcn.ws/2	2016-06-22 16:2	3956	1506	1049	2555	70
Colorado groom bit by a rattlesnake while taking wedding photos. "I adn	Groom Bit by Ra	link	http://abcn.ws/2	2016-06-22 16:1	1723	77	224	1277	24
Timelapse video shows intense flash flooding pour into a metro station in	Timelapse Show	video	https://www.fac	2016-06-22 15:5	10138	684	5039	6813	24
Bernie Sanders inches closer to admitting defeat in the race for the Demc	Sanders Said He	link	http://abcn.ws/2	2016-06-22 15:3	1355	409	147	813	20
Hong Kong is the world's most expensive city for expatriates, dethroning	Hong Kong Is W	link	http://abcn.ws/2	2016-06-22 15:2	168	17	17	157	2
California father spends $900 to buy full-page ad in Idaho newspaper see	Dad Buys Full-Pa	link	http://abcn.ws/2	2016-06-22 14:5	194	47	42	112	1
Celebrities and animal activists in the U.S. and China are rallying to stop a	Celebrities Squa	link	http://abcn.ws/2	2016-06-22 14:2	5351	768	1333	2711	53
UPDATE: One body found during search in the Gulf of Mexico for a father	Body Found in S	link	http://abcn.ws/2	2016-06-22 14:0	1428	77	139	552	3
Man who was stopped near the Holland Tunnel with an arsenal of guns a	Man Stopped in	link	http://abcn.ws/2	2016-06-22 13:4	320	88	64	243	14
DOGGONE EXCITED: Police dog can't control his excitement while being	Happy Police Dc	video	https://www.fac	2016-06-22 13:2	19397	351	1896	15733	3299
An estimated 1.3 million fans gather to celebrate Cleveland Cavaliers' hist	Photos from ABC	photo	https://www.fac	2016-06-22 13:0	3501	106	374	3201	178
	ABC News Politi	video	https://www.fac	2016-06-22 12:5	2718	683	0	2083	347
Former Speaker of the House Dennis Hastert reports to prison in wheelch	Former Speaker	link	http://abcn.ws/2	2016-06-22 12:5	1078	272	215	775	24
House Democrats are engaged in a sit-in on the House floor to protest G	Dems Staging Si	link	http://abcn.ws/2	2016-06-22 12:3	4365	949	521	3440	565
The name 'Hillary' rose 142% and the name 'Donald' rose 8%. But not a si	More Babies Bei	link	http://abcn.ws/2	2016-06-22 12:1	547	141	60	360	9
Cargo ship filled with junk from the International Space Station burns up	Space Station Tr	link	http://abcn.ws/2	2016-06-22 11:4	272	25	36	253	1

Figure 14: Collection initiale des actions sociales de« Facebook »

II.1.3.2 Analyse et organisation de données

Étant donné que les différentes collections de données sont biaisées et bruyante, leur nettoyage est notre première tâche. Ce prétraitement consiste à structurer et à faciliter l'utilisation des collections de données originales. Cependant, Le prétraitement des collections de données est le processus de nettoyage et de préparation des différents contenus pour l'indexation.

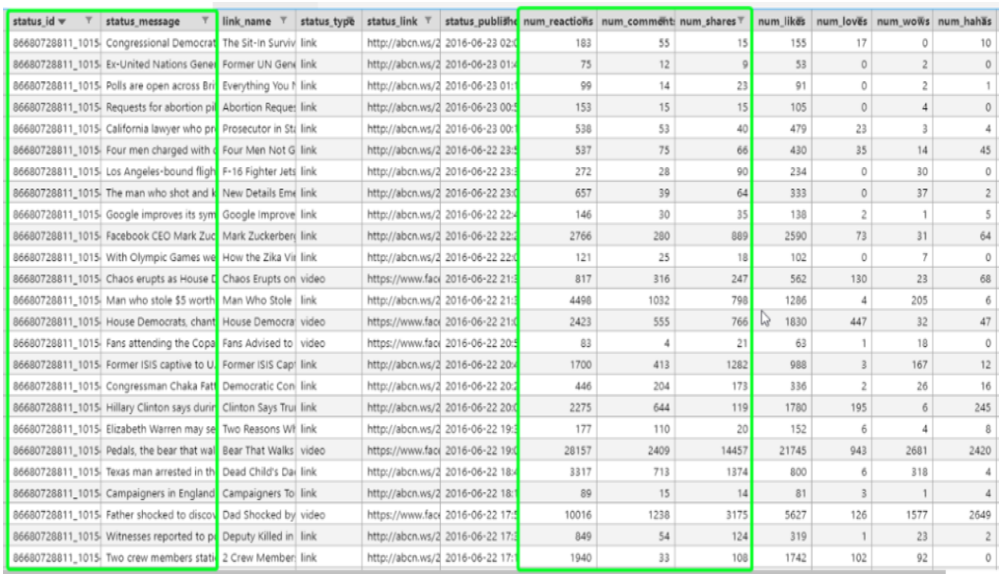
Tout d'abord, notre outil identifie les données inutiles et gênantes pour le processus de la RI. L'étape de prétraitement consiste soit à les supprimer pour réduire la complexité du système, ou bien à les mieux organiser pour donner une structure simple et lisible aux documents qui y sont exploités. Enfin, la collection finale de documents représente les données prêtes à être utilisées dans le processus de recherche. Nous présentons l'ensemble des problèmes détectés par exemple dans les données « Facebook » :

- Le problème des espaces et des nouvelles lignes, qui occupent un espace important, ce qui réduit les performances du système et augmente sa complexité.
- De plus, Le problème des lignes vides et des contenus sans texte (par exemple, une image seule, une vidéo sans statut, etc.) et donc, ces données ne seront pas des documents pertinents aux requêtes d'utilisation, ainsi que tous leurs signaux sociaux deviennent inutiles car ils ne sont pas considérés comme des critères pour décrire l'importance de la ressource. (voir la figure 15)

status_id	status_message	link_name	status_type	status_link	status_publishe	num_reactions
86680728811_10154512704773812	Campaigners in England are making one last push with voters before tom	Campaigners To	link	http://abcn.ws/2	2016-06-22 18:1	89
86680728811_10154512652233812	Father shocked to discover that his two toddlers had covered their bedro	Dad Shocked by	video	https://www.face	2016-06-22 17:5	10016
86680728811_10154512610958812	Witnesses reported to police that a struggle ensued and the man pulled a	Deputy Killed in	link	http://abcn.ws/2	2016-06-22 17:3	849
86680728811_10154512574188812	Two crew members stationed at a U.S. science facility at the South Pole ar	2 Crew Member	link	http://abcn.ws/2	2016-06-22 17:1	1940
86680728811_10154512514773812	Kentucky community rallies together to support a beloved, friendly neigh	Community Ralli	link	http://abcn.ws/2	2016-06-22 16:5	2218
86680728811_10154512503878812		ABC News Politi	video	https://www.face	2016-06-22 16:5	1146
86680728811_10154512469693812	Family set to inherit gun stockpile worth millions plans to destroy the wea	Family to Destro	link	http://abcn.ws/2	2016-06-22 16:3	11208
86680728811_10154512436248812	Donald J. Trump painted Hillary Clinton as a "world class liar" in a speech	Fact Check: Trum	link	http://abcn.ws/2	2016-06-22 16:2	3956
86680728811_10154512416983812	Colorado groom bit by a rattlesnake while taking wedding photos. "I adm	Groom Bit by Ra	link	http://abcn.ws/2	2016-06-22 16:1	1723
86680728811_10154512360043812	Timelapse video shows intense flash flooding pour into a metro station in	Timelapse Show	video	https://www.face	2016-06-22 15:5	10138
86680728811_10154512304768812	Bernie Sanders inches closer to admitting defeat in the race for the Democ	Sanders Said He	link	http://abcn.ws/2	2016-06-22 15:3	1355
86680728811_10154512280223812	Hong Kong is the world's most expensive city for expatriates, dethroning	Hong Kong Is W	link	http://abcn.ws/2	2016-06-22 15:2	168
86680728811_10154512197528812	California father spends $900 to buy full-page ad in Idaho newspaper see	Dad Buys Full-Pa	link	http://abcn.ws/2	2016-06-22 14:5	194
86680728811_10154512109978812	Celebrities and animal activists in the U.S. and China are rallying to stop a	Celebrities Squa	link	http://abcn.ws/2	2016-06-22 14:2	5351
86680728811_10154512063883812	UPDATE: One body found during search in the Gulf of Mexico for a father	Body Found in S	link	http://abcn.ws/2	2016-06-22 14:0	1428
86680728811_10154512023113812	Man who was stopped near the Holland Tunnel with an arsenal of guns a	Man Stopped in	link	http://abcn.ws/2	2016-06-22 13:4	320
86680728811_10154511971203812	DOGGONE EXCITED: Police dog can't control his excitement while being a	Happy Police Do	video	https://www.face	2016-06-22 13:2	19397
86680728811_10154511905413812	An estimated 1.3 million fans gather to celebrate Cleveland Cavaliers' hist	Photos from AB	photo	https://www.face	2016-06-22 13:0	3501
86680728811_10154511877088812		ABC News Politi	video	https://www.face	2016-06-22 12:5	2718
86680728811_10154511872203812	Former Speaker of the House Dennis Hastert reports to prison in wheelch	Former Speaker	link	http://abcn.ws/2	2016-06-22 12:5	1078
86680728811_10154511842323812	House Democrats are engaged in a sit-in on the House floor to protest G	Dems Staging Si	link	http://abcn.ws/2	2016-06-22 12:3	4365
86680728811_10154511786308812	The name 'Hillary' rose 142% and the name 'Donald' rose 8%. But not a si	More Babies Bei	link	http://abcn.ws/2	2016-06-22 12:1	547
86680728811_10154511728323812	Cargo ship filled with junk from the International Space Station burns up	Space Station Tr	link	http://abcn.ws/2	2016-06-22 11:4	272
86680728811_10154511640218812	18-year-old woman dies in Ohio after being infected by Naegleria fowler	Teen Dies After	link	http://abcn.ws/2	2016-06-22 11:2	1726

Figure 15: Détection des contenus sans texte

- Le problème des données non structurées contenues dans des documents séparés, ce qui rend l'indexation difficile et fastidieuse, en raison de l'illisibilité du corpus. De plus, le problème du trop grand nombre de champs dans l'ensemble de données Facebook, dont nous n'avons pas besoin pour mesurer l'importance des documents telles que : le type, le lien et la date du publication des statuts. Nous devons donc garder que les données utiles (id de statut, message de statut, nombre de réactions (J'aime, J'adore, Haha, Wouah, Triste et Grrr.), nombre de commentaires, nombre de partages) comme la montre la figure 16.



status_id	status_message	link_name	status_type	status_link	status_publishe	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas
86680728811_1015	Congressional Democrat	The Sit-In Survivi	link	http://abcn.ws/2	2016-06-23 02:0	183	55	15	155	17	0	10
86680728811_1015	Ex-United Nations Gener	Former UN Gene	link	http://abcn.ws/2	2016-06-23 01:4	75	12	9	58	0	2	0
86680728811_1015	Polls are open across Br	Everything You N	link	http://abcn.ws/2	2016-06-23 01:1	99	14	23	91	0	2	1
86680728811_1015	Requests for abortion pil	Abortion Reques	link	http://abcn.ws/2	2016-06-23 00:5	153	15	15	105	0	4	0
86680728811_1015	California lawyer who pr	Prosecutor in Sta	link	http://abcn.ws/2	2016-06-23 00:1	538	53	40	479	23	3	4
86680728811_1015	Four men charged with c	Four Men Not G	link	http://abcn.ws/2	2016-06-22 23:5	537	75	66	430	35	14	45
86680728811_1015	Los Angeles-bound flight	F-16 Fighter Jets	link	http://abcn.ws/2	2016-06-22 23:3	272	28	90	234	0	30	0
86680728811_1015	The man who shot and k	New Details Eme	link	http://abcn.ws/2	2016-06-22 23:0	657	39	64	333	0	37	2
86680728811_1015	Google improves its sym	Google Improve	link	http://abcn.ws/2	2016-06-22 22:4	146	30	35	138	2	1	5
86680728811_1015	Facebook CEO Mark Zuc	Mark Zuckerberg	link	http://abcn.ws/2	2016-06-22 22:2	2766	280	889	2590	73	31	64
86680728811_1015	With Olympic Games we	How the Zika Vi	link	http://abcn.ws/2	2016-06-22 22:0	121	25	18	102	0	7	0
86680728811_1015	Chaos erupts as House D	Chaos Erupts on	video	https://www.face	2016-06-22 21:3	817	316	247	562	130	23	68
86680728811_1015	Man who stole $5 worth	Man Who Stole	link	http://abcn.ws/2	2016-06-22 21:3	4498	1032	798	1286	4	205	6
86680728811_1015	House Democrats, chant	House Democra	video	https://www.face	2016-06-22 21:0	2423	555	766	1830	447	32	47
86680728811_1015	Fans attending the Copa	Fans Advised to	video	https://www.face	2016-06-22 20:5	83	4	21	63	1	18	0
86680728811_1015	Former ISIS captive to U	Former ISIS Cap	link	http://abcn.ws/2	2016-06-22 20:4	1700	413	1282	988	3	167	12
86680728811_1015	Congressman Chaka Fatt	Democratic Con	link	http://abcn.ws/2	2016-06-22 20:2	446	204	173	336	2	26	16
86680728811_1015	Hillary Clinton says durir	Clinton Says Trui	link	http://abcn.ws/2	2016-06-22 20:0	2275	644	119	1780	195	6	245
86680728811_1015	Elizabeth Warren may se	Two Reasons Wh	link	http://abcn.ws/2	2016-06-22 19:3	177	110	20	152	6	4	8
86680728811_1015	Pedals, the bear that wal	Bear That Walks	video	https://www.face	2016-06-22 19:0	28157	2409	14457	21745	943	2681	2420
86680728811_1015	Texas man arrested in th	Dead Child's Dad	link	http://abcn.ws/2	2016-06-22 18:4	3317	713	1374	800	6	318	4
86680728811_1015	Campaigners in England	Campaigners To	link	http://abcn.ws/2	2016-06-22 18:1	89	15	14	81	3	1	4
86680728811_1015	Father shocked to discov	Dad Shocked by	video	https://www.face	2016-06-22 17:5	10016	1238	3175	5627	126	1577	2649
86680728811_1015	Witnesses reported to p	Deputy Killed in	link	http://abcn.ws/2	2016-06-22 17:3	849	54	124	319	1	23	2
86680728811_1015	Two crew members stati	2 Crew Member	link	http://abcn.ws/2	2016-06-22 17:1	1940	33	108	1742	102	92	0

Figure 16: Détection des données inutiles

L'algorithme suivant nous permet de résoudre et d'éviter ces problèmes:

Entrée : corpus
Sortie : fichier csv
Variable : tableau : résultat
fichier : Data
Début :
Ouvrir le corpus
Pour chaque document de corpus **faire**
Data ⟵ lire fichier
Data ⟵ supprimer les espaces pour chaque ligne de Data
Data ⟵ supprimer les lignes vides de Data
Sélectionner les signaux sociaux utiles
Ajouter le fichier dans le tableau de résultat
Finpour
Concaténer les fichiers de résultats et les exporter vers le fichier csv
Fin.

Figure 17: Algorithme de curation de données Facebook

Les résultats obtenus à la suite du prétraitement de la collection de Facebook sont présentés dans la figure ci-dessous :

post_id	text	num_reactions	num_comments	num_shares
86680728811_10154513660173812	Congressional Democrats' staging an hour-long sit-in on the House floor to prot	183	55	15
86680728811_10154513562358812	Requests for abortion pills have increased significantly in 7 Latin American count	153	15	15
86680728811_10154500162653812	Experts say it's too soon to gauge whether a week of horrific news out of Orland	441	73	15
86680728811_10154440665588812	Leftist rebels free a Spanish correspondent and two other journalists who went n	266	16	15
86680728811_10154427431863812	The crucial black boxes from EgyptAir Flight 804 have yet to be recovered after t	229	18	15
86680728811_10154414299673812	Concern that Donald J. Trump could affect November voting further down the ba	139	78	15
86680728811_10154408439623812	Supreme Court says it won't rule in one of the most controversial legal challenge	175	28	15
86680728811_10154378970043812	While the possibility of a contested convention is essentially off the table, the Re	326	117	15
86680728811_10154301361213812	Community comes together to donate dresses to shop who gives them to peopl	117	8	15
86680728811_10154300160113812	Some people who pre-ordered Oculus Rift virtual reality headset are going to ha	184	21	15
86680728811_10154211888743812	In the biggest day of voting since Super Tuesday, Americans in five states and or	256	81	15
86680728811_10154189699638812	Mission to drill beneath surface of Mars is being postponed two years due to eq	192	30	15
86680728811_10154168911478812	The Republican presidential candidates lobbed attacks at each other in Detroit to	106	62	15
228735667216_10153694273617217	We'll soon have our second #FacebookLive today at Wembley Arena where we're	146	70	15
228735667216_10153664470817217	It's EU Referendum Questions Day on the BBC so what do you want to know? Po	628	216	15
228735667216_10153583933412217	A minister urged Francophones on Twitter to "rise up" in reaction to the official F	270	73	15
228735667216_10153664470817217	It's EU Referendum Questions Day on the BBC so what do you want to know? Po	628	216	15
25902406772_10153995033036773	Nothing like barbershop talk to get to know someone. And Anthony Davis' new	77	2	15
25902406772_10153994245036773	Don't mess with Baltimore Ravens WR Steve Smith.	193	8	15
25902406772_10153983066591773	The USMNT has been here before, but can they do it again?	288	17	15
25902406772_10153976670001773	Turns out Leo Messi jerseys are magical.	551	9	15
25902406772_10153968272816773	The Golden State Warriors have the Cavaliers on the brink.	989	27	15
25902406772_10153963392376773	Kevin Love's status for Game 4 is still uncertain.	459	21	15
25902406772_10153958678816773	The Texas Rangers are climbing up our power rankings.	186	9	15

Figure 18: Collection finale des actions sociales de « Facebook »

Après avoir analysé les données et structuré les documents, les résultats finaux de cette étape sont regroupés en quatre fichiers (Twitter.csv, YouTube Comments.csv, YouTube Videos.csv, Facebook.csv). La figure ci-dessous montre une représentation de la structure globale des données :

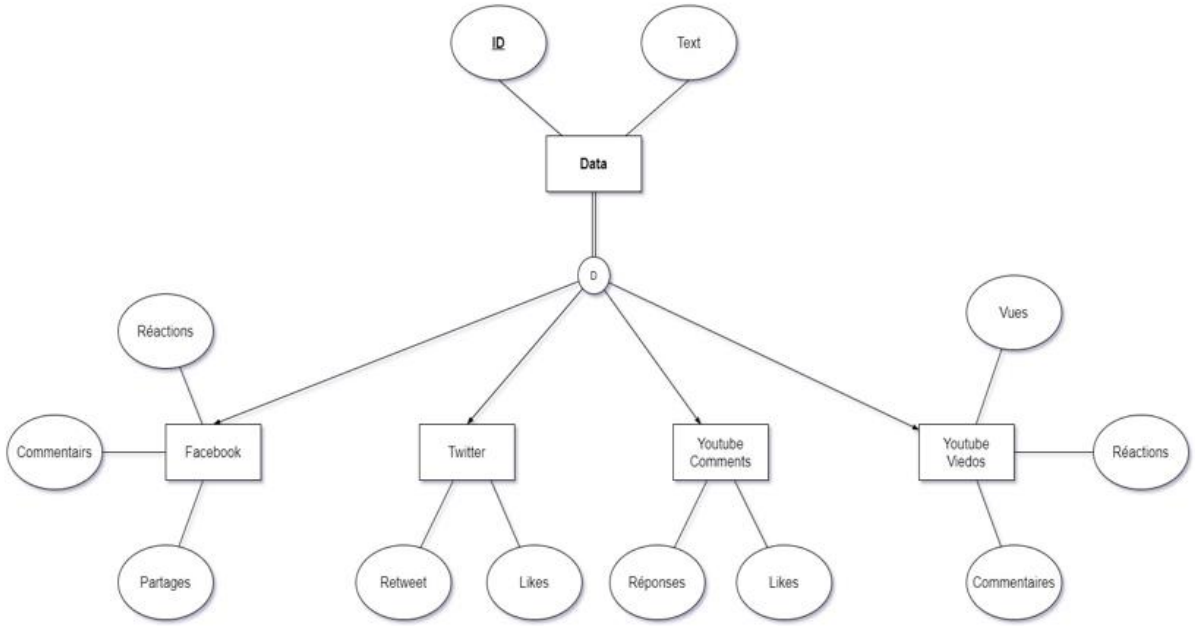


Figure 19: Organisation de données (EER Model[16])

Le schéma ci-dessus montre comment nos données sont structurées en les organisant en quatre fichiers séparés. Ces fichiers contiennent des attributs similaires telles que l'identifiant et le texte. Cependant, ils diffèrent en termes de caractéristiques liées aux signaux sociaux, car chaque réseau social a ses propres actions sociales qui sont considérées comme des critères de pertinence. Enfin, les documents de ces fichiers sont prêts à être utilisés dans notre processus de RI.

II.1.3.3 Indexation des documents

L'indexation est la première étape du processus, consiste à analyser chaque document de la collection afin de construire des index qui facilitent le processus de recherche. Nous expliquons le processus de cette étape en l'appliquant au document (contenu textuel) suivant:

Document:

Hello world! this is just an example for <testing>.

Où nous allons représenter ce document en sac de mots en suivant les étapes suivantes :

[16]https://www.tutorialride.com/dbms/enhanced-entity-relationship-model-eer-model.htm

II.1.3.3.1 Analyse lexicale

Les mots sont considérés comme une suite de caractères délimités entre deux séparateurs. Dans cette étape, il s'agit d'extraire ces mots contenus dans chaque fichier, en éliminant l'ensemble de séparateurs montré dans la figure 20:

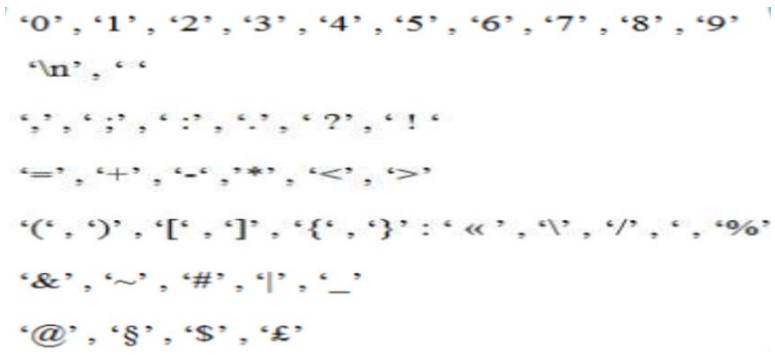


Figure 20: Ensemble de séparateurs[17]

L'algorithme suivant illustre les étapes à suivre pour l'analyse lexicale (Segmentation du corpus) :

Entrée : corpus
liste des séparateurs (délimiter)
Sortie : tableau de tokens
Variable : chaine de caractère : ligne, tableau :tab
Début :
Ouvrir le corpus
Pour chaque document du corpus **faire**
Line ←— lire ligne
Tant que line lu est différent EOF **faire**
Toknizer (line, délimiter)
Tan que il y'a des tokens **faire**
token ←— token suivant
insérer le token dans tab
Fintanque
Line ←— lire ligne
Fintanque
Document suivant
Finpour
Fin.

Figure 21: Algorithme d'analyse lexicale

[17] https://en.wikipedia.org/wiki/List_of_typographical_symbols_and_punctuation_marks

Apres l'élimination des séparateurs, les mots (tokens) sont représentés comme suit :

Hello world! this is just an example for <testing>.

Hello	**world**	**this**	**is**	**just**	**an**	**example**	**for**	**testing**

II.1.3.3.2 Elimination des mots vides

Cette étape consiste à éviter les mots vides qui n'ont aucun intérêt pour la recherche, comme indiqué dans le tableau ci-dessous, et de ne choisir que les termes significatifs qui représentent le mieux un document particulier.

Langue	Nombre de mot vides	Exemple
English	635	All, also, for, in…
Frensh	463	A, afin, notre, entre…
Germany	129	Aber, den, musst, sollen…
Spanish	351	Esta, podemos, actuelmente...
Italian	399	Deve, dice, fuori, gia…

Tableau 6: Mots vides de différentes langues [82]

Puisque nous nous intéressons aux mots en anglais, l'algorithme suivant illustre les étapes à suivre pour l'élimination de ces mots :

Entrée : tableau de tokens
liste des mot vides
Sortie : tableau de tokens

Début :
Pour chaque token du tableau **faire**
Si le token appartient à la liste des mots vides
Alors supprimer le token du tableau
Finsi
Finpour
Fin.

Figure 22: Algorithme d'élimination de mots vides

L'application de cet algorithme sur l'exemple présenté précédemment donne le résultat suivant:

Hello	world	this	is	just	an	example	for	testing

Hello	world	example	testing

Les quatre mots simples résultants sont considérés comme unités de représentation du document (représentation en sac de mots) qui peuvent être utilisées à l'étape d'appariement (approche syntaxique). Pour l'approche sémantique, nous devons indexer ce document en transformant ses mots (tokens) en synsets (indexation sémantique).

II.1.3.3.3 Indexation sémantique

Dans cette étape, nous proposons une méthode de désambiguïsation sémantique basée sur Wordnet et organisée autour de la notion de synset. Pour déterminer le sens d'un mot ambigu, les synsets (ensemble de synonymes) de ce mot sont classés en utilisant la valeur de cooccurrence calculée entre le contexte de ce mot et un voisinage contenant les mots du synset dans la hiérarchie de WordNet. Le synset le mieux classé est alors choisi comme le synset approprié du mot ambigu analysé. L'algorithme suivant nous aide à comprendre cette étape en l'appliquant aux mots de document :

Entrée : tableau de sac de mots

Sortie : tableau de synsets

Début :

Pour chaque mot du tableau **faire**

synset ⟵ Synsets (mot)

Si le synset n'est pas vide

Alors insérer le synset dans le tableau de synsets

Finpour

Fin.

Figure 23: Algorithme de Désambiguïsation du sens des mots

Les résultats de cette étape sont présentés dans le tableau suivant :

Sac de mots	Synsets du mot
Hello	Hello(N-01)
World	World(N-02), World(N-03), World(N-06), World(N-08), Worldly_Concern(N-01), Earth(N-01), Populace(N-01), Global(S-01), Universe(N-01)
Example	Example(N-01), Example(N-04), Model(N-07), Case(N-01), Exercise(N-04), Exemplar(N-01)
Testing	Test(V-01), Test(V-04), Test(V-05), Test(V-06), Test(V-07), Testing(N-01), Testing(N-02), Quiz(V-01), Screen(V-01), Examination(N-05),

Tableau 7: Indexation sémantique

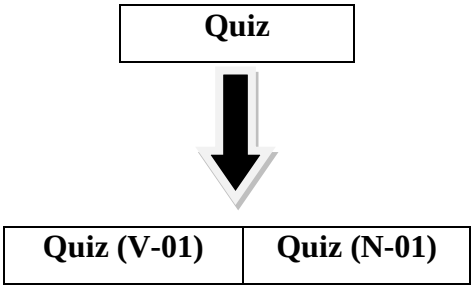
Où N, V et S représentent les types de synset : nom, verbe et adjectif satellite respectivement. D'autres types peuvent être trouvés (A : adjectif, R : adverbe). En ce qui concerne le numéro à côté d'eux, cela signifie le numéro de signification s'il y a plus d'un sens du mot.

À la fin de cette étape, une représentation vectorielle (Vecteurs de document) est prête à être utilisée au sein du modèle d'appariement (approche sémantique). Dans notre cas, les synsets de chaque mot qui sont mentionnés dans le tableau précèdent représentent un seul vecteur parmi les quatre vecteurs de notre document.

II.1.3.4 Analyse de la requête (Compréhension)

Notre principal objectif est de rechercher les documents pertinents à un sujet précis proposé par l'utilisateur. Ce sujet est représenté par une requête, soit par un seul mot (Politics, Study, Football...), soit par des mots composés (Programing_Language, Football_Stadium…). C'est pour cela, la requête est représentée directement comme un seul token dans notre système, sans passer par les étapes d'indexation (Segmentation, Elimination des mots vides). Cette représentation est prête pour l'étape d'appariement syntaxique. Par la suite, cette requête sera indexée d'une manière sémantique en transformant ce mot(token) en synsets de la même manière que le document.

Pour une meilleure compréhension, nous présentons un exemple de requête appelée « **Quiz** » qui est considérée comme un token, et montrons sa représentation en appliquant l'algorithme mentionné dans la figure 23. Le résultat est présenté ci-dessous :



À la fin de cette étape, une représentation vectorielle (Vecteur de requête mentionné ci-dessus) est prête à être utilisée à l'étape d'appariement sémantique.

II.1.3.5 Appariement document-requête

La relation d'appariement consiste à rechercher parmi les documents prétraités, ceux qui répondent le mieux à la requête c.-à-d. calculer un score de pertinence entre le vecteur requête et les vecteurs documents selon un score de correspondance entre ces deux représentations. Pour notre système, nous proposons deux approches pour calculer ce score: approche syntaxique et approche sémantique.

Le diagramme d'activité de la figure 24 montre le fonctionnement de notre système et en particulier la phase d'appariement document-requête. Tout d'abord, l'utilisateur doit sélectionner la collection de test (dataset) dans laquelle il souhaite effectuer une recherche, puis il introduit le sujet de recherche (requête). Cependant, notre système commence son processus de RI par une recherche syntaxique au sein du dataset indexé précédemment (tokens). Si la requête est trouvée dans ce document, alors son score de similarité est de 1.0 (la valeur du score est nécessaire afin de classer les documents), et donc le document est pertinent à la requête. Sinon, il faut vérifier la validité de la requête. Si elle est invalide c.-à-d. le sujet sélectionné par l'utilisateur n'existe pas dans la taxonomie de WordNet tel que les nouveaux termes (exemple Coronavirus) ou certains noms propres (comme Ismail, Bouras, Tiaret…), alors le score de pertinence sera de zéro, ce qui signifie que le document n'est pas pertinent. Sinon, la recherche orientée sémantique est lancée en exploitant deux mesures de similarité différentes (Wup [1] et Path [2]). les résultats tirés sont utilisés pour calculer le score final, si ce score est supérieur à zéro, alors le document est pertinent. Sinon, le document n'est pas pertinent à la requête.

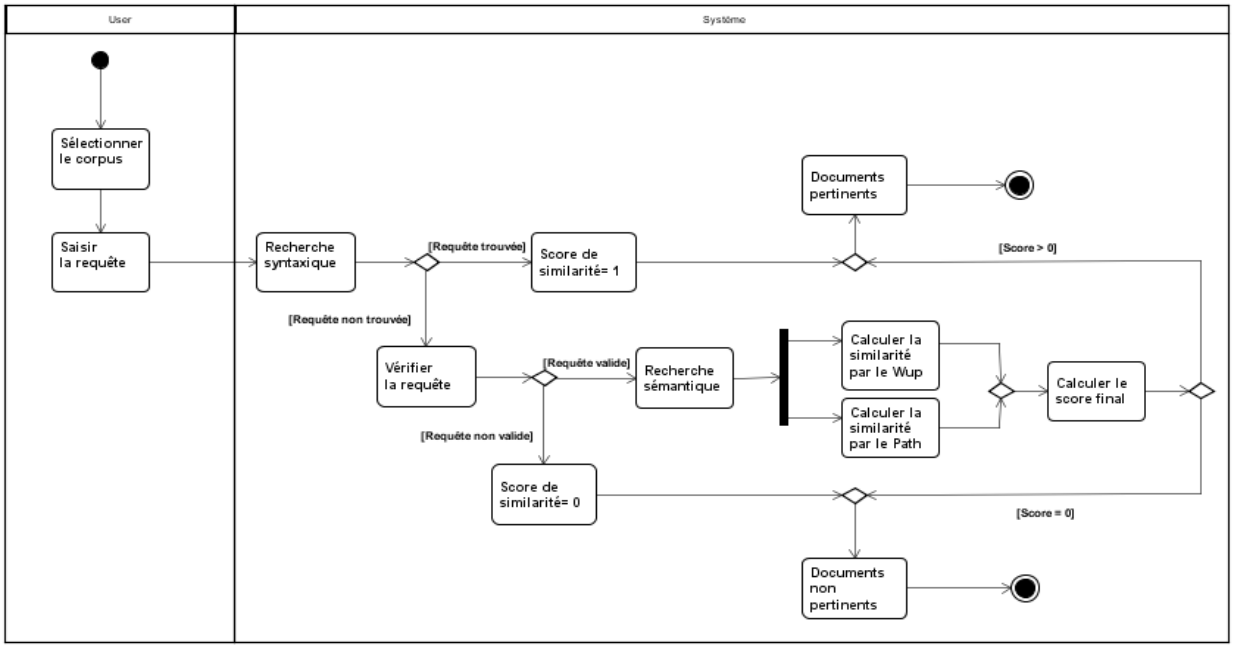


Figure 24: Processus de l'appariement (document-requête)

Afin de mieux comprendre les deux approches (Syntaxique, Sémantique) mentionnées dans ce diagramme, nous allons les expliquer séparément :

II.1.3.5.1 Approche syntaxique

L'approche syntaxique consiste à rechercher parmi les mots de document (tokens) celui qui a une syntaxe identique à la requête. Le résultat de cet algorithme (figure 25) est un score de pertinence entre le document et la requête. S'il existe un token (Représentation en sac de mots) dont la syntaxe correspond à la syntaxe de la requête, alors le score égale à 1.0 et donc le document est pertinent à cette requête. Sinon le score de pertinence sera de zéro ce qui signifie que le document n'est pas pertinent. Cette approche est utilisée afin de réduire la complexité du système en évitant l'approche sémantique en cas de résultat positif (score égale à 1.0).

Entrée : tableau de tokens du document
chaine de caractères : requête
Sortie : entier : score de similarité
Début :
Si le tableau de tokens contient la requête
Alors score de similarité ⟵ 1.0
Sinon score de similarité ⟵ 0.0
Fin.

Figure 25: Algorithme de similarité syntaxique

II.1.3.5.2 Approche sémantique

L'approche sémantique permet de calculer le degré de pertinence en faisant correspondre les mots, qui sont conceptuellement similaires mais pas nécessairement syntaxiquement similaires. Cependant, il existe plusieurs défis pour calculer la similarité sémantique entre les mots tels que la complexité des langages naturels, l'ambiguïté des mots, etc.

Notre approche définie une méthode d'indexation conceptuelle basée sur l'utilisation de l'ontologie linguistique WordNet. Chaque document est représenté sous forme d'un réseau sémantique particulier (appelé noyau sémantique), dans lequel les nœuds représentent les Synsets et les arcs (bidirectionnels) représentatifs de la distance sémantique entre les concepts liés. Nous allons expliquer l'algorithme utilisé dans cette approche en le divisant en trois séquences :

1- **La première partie** consiste à trouver le score de chaque mot (token) du document en calculant le degré de pertinence du vecteur des synsets de mot vis-à-vis le vecteur des synsets de la requête en suivant l'algorithme ci-dessous :

Entrée : tableau des synset de mot du document
tableau des synsets de requête
Sortie : score du mot
Début :
Pour chaque synset de requête **faire**
Pour chaque synset de mot du document **faire**
Score du synset ⟵ Mesure_Similarité (synset de mot, synset de requête)
Insérer le score du synset dans un tableau de scores des synsets
Finpour
Finpour
Score du mot ⟵ Max (scores des synsets)
Fin.

Figure 26: Algorithme de calcul des scores du mot

Où le score du synset est calculé au moyen d'une fonction de similarité en utilisant différentes mesures de similarité sémantique. Les mesures de similarité sémantique existantes peuvent être classées en deux groupes : basées sur le chemin et basées sur le contenu d'information. Les mesures basées sur le chemin reposent uniquement sur les informations du plus court chemin telles que « Path [2]», « WuPalmer [1]» et « LeacockChodorow [83] »,

tandis que les mesures basées sur le contenu d'information incorporent la probabilité que le concept se produise dans un corpus de texte telles que. « Resnik [84] », « JiangConrath [85]», « Lin [86]». Nous nous intéressons aux mesures basées sur le chemin, en particulier les deux mesures (Path et WuPalmer) parce qu'elles traitent l'appariement des synsets (requête-document) dont les types sont différents (V-N, V-S, N-S.…) comme le montre le tableau 8:

- **Mesure de similarité « Path »** [2]**:** la mesure basée sur le chemin (Path-based similarity) introduit la mesure de distance conceptuelle, qui est calculée comme la longueur du chemin le plus court entre deux concepts (Synsets) qui relie les concepts via leur parent le moins commun (LCS). Le LCS est l'ancêtre le plus spécifique partagé par deux concepts. La longueur est calculée en comptant le nombre de nœuds entre les deux concepts. La mesure Path est une modification de ceci et est calculée comme l'inverse de la longueur du chemin le plus court (Min_Path), comme indiqué pour les synsets S1 et S2 dans l'équation suivante :

$$Sim_{path}(S1,S2) = \frac{1}{Min\ Path\ (S1,S2)} \qquad (6)$$

Nous donnons un exemple d'un fragment de la hiérarchie des hyperonymes WordNet, montrant les longueurs de chemin (nombre d'arêtes plus 1) :

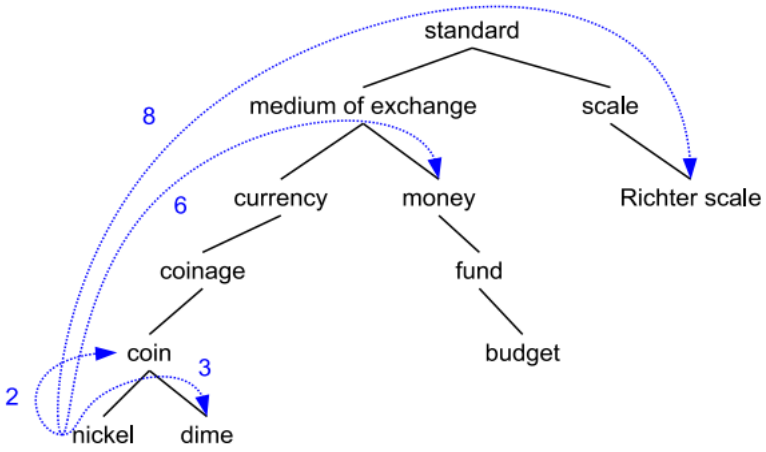


Figure 27: Exemple d'ontologie de Path [87]

La figure ci-dessus montre les longueurs de chemin du nickel à coin (2), dime (3), money (6), et Richter scale (8). Nous remarquons donc que plus le chemin entre deux concepts est long, plus le taux de similitude est faible.

- **Mesure de similarité « WuPalmer »** [1]**:** le WuPalmer calcule l'appariement en considérant les profondeurs (depths) des deux synsets (S1,S2) dans les taxonomies Wordnet, ainsi que la profondeur de LCS (Least Common Subsumer), en utilisant l'équation suivante :

$$Sim_{wup}(S1,S2) = 2 * \frac{depth\,(LCS\,(S1,S2))}{depth\,(S1) + depth\,(S2)} \qquad (7)$$

Cela signifiait que 0 <Score <= 1. Le score ne peut jamais être nul car la profondeur du LCS n'est jamais nulle (la profondeur de la racine de la taxonomie (R) en est une), le score est de un si les deux concepts d'entrée sont identiques.

Il calcule la similarité en fonction de la similitude des sens des mots et de l'emplacement des Synsets les uns par rapport aux autres dans l'arbre des hyperonymes (est-un). (voir la figure 28)

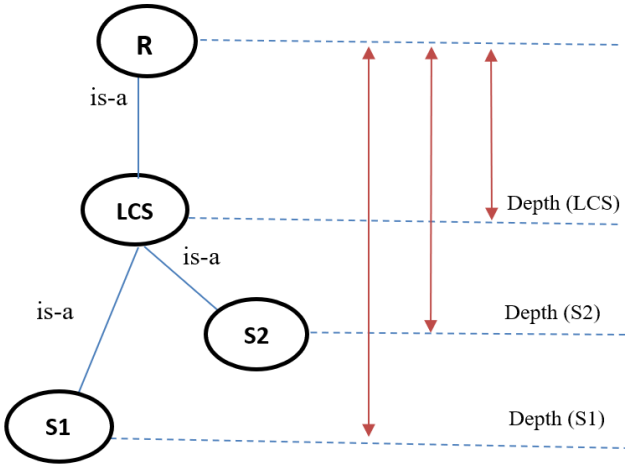


Figure 28: Exemple d'ontologie de Wu Palmer

Où :

- *LCS* est le parent commun des Synsets S1 et S2 avec une distance de nœud minimale.
- *Depth (S1)* représente le nombre de nœuds de S1 au nœud racine (R).
- *Depth (S2)* représente le nombre de nœuds de S2 au nœud racine (R).
- *Depth (LCS)* représente le nombre de nœuds de LCS(S1,S2) au nœud racine (R).

Le score de chaque mot est le score maximum de leurs synsets :

$$Score_{mot} = Max\left(Scores_{synsets}\,(R,D)\right) \qquad (8)$$

Où : R et D représentent les synsets de la requête et du mot de document respectivement.

2- **La deuxième partie** concerne le calcul du score moyen de tous les scores des mots, ainsi que leur score parfait, afin de les utiliser dans la partie suivante :

Entrée : tableau de scores des mots

Sortie : score moyen

score parfait

Début :

Score moyen ⟵ Somme (scores des mots) / Taille (scores des mots)

Score parfait ⟵ Max (scores des mots)

Fin.

Figure 29: Algorithme de calcul de scores(moyen et parfait) des mots

Où le score moyen est calculé à partir de la somme de scores des mots devisée sur la taille (n) des mots (c.-à-d. nombre de mots dans le document), sachant que cette taille est réduite s'il y'a un mot invalide (il n'a pas de synsets), comme le montre la formule suivante :

$$Score_{moyen} = \frac{\sum_{k=1}^{n} Scores_{mots}}{n} \quad (9)$$

Pour le score parfait, nous prenons le score le plus élevé parmi les scores des mots :

$$Score_{parfait} = Max\,(Scores_{mots}) \quad (10)$$

3- **La dernière partie** présente le résultat final de cette approche, en calculant le degré final de similitude entre la requête et le document. L'algorithme suivant explique comment atteindre un résultat efficace en utilisant des conditions basées sur nos tests :

Entrée : score parfait

score moyen

Sortie : score de similarité

Début :

Si (score parfait >= 0.9 **et** score moyen >= 0.4) **ou** (score parfait = 1)

Alors score de similarité ⟵ (score parfait + score moyen) / 2

Fin.

Figure 30: Algorithme de calcul de score final de la similarité

Où le score de similarité est le score final qui permet de classer les documents pertinents. Afin d'obtenir un classement efficace des résultats, nous utilisons une combinaison de scores moyen et parfait pour améliorer le score final :

$$Score_{similarité} = \frac{Score_{parfait} + Score_{moyen}}{2} \quad (11)$$

Le résultat final de cette étape d'appariement est un tableau des scores de pertinence entre la requête et les documents. Par la suite nous allons présenter les résultats finaux obtenus à partir de nos expérimentations appliquées à notre exemple.

II.1.4 Résultats et discussions

Mots du document	Synsets du mot	Appariement document-requête (WUP vs PATH)						
		Type et n⁰ de synset (requête)	Type et n⁰ de synset (document)	Score du synset	Score du mot	Score parfait	Score moyen	Score final
Hello	Hello	N-01	N-01	0.266666 0.083333	0.266666 0.083333	1.0	0.559523	0.779761
		V-01	N-01	0.117647 0.0625				
World	World	N-01	N-02	0.258714 0.090909	0.4 0.111111			
		N-01	N-03	0.4 0.1				
		N-01	N-06	0.153846 0.083333				
		N-01	N-08	0.333333 0.111111				
		V-01	N-02	0.125 0.066666				
		V-01	N-03	0.117647 0.0625				
		V-01	N-06	0.133333 0.071428				
		V-01	N-08	0.083333 0.043478				
	Universe	N-01	N-01	0.142857 0.076923				
		V-01	N-01	0.125 0.066666				
	Earth	N-01	N-01	0.117647 0.0625		1.0	0.334325	0.667162
		V-01	N-01	0.105263 0.055555				
	Populace	N-01	N-01	0.307692 0.1				
		V-01	N-01	0.133333 0.071428				
	Global	N-01	S-01	------ ------				
		V-01	S-01	0.181818 0.1				
	Worldly_Concern	N-01	N-01	0.210526 0.0625				
		V-01	N-01	0.095238 0.05				

Example	Example	N-01	N-01	0.428571 0.111111	0.571428 0.142857			
		N-01	N-04	0.25 0.076923				
		V-01	N-01	0.125 0.066666				
		V-01	N-04	0.111111 0.058823				
	Model	N-01	N-07	0.4 0.1				
		V-01	N-07	0.117647 0.0625				
	Case	N-01	N-01	0.571428 0.142857				
		V-01	N-01	0.125 0.066666				
	Exemplar	N-01	N-01	0.375 0.090909				
		V-01	N-01	0.111111 0.058823				
	Exercise	N-01	N-04	0.5 0.090909				
		V-01	N-04	0.090909 0.047619				
Testing	Test	N-01	V-01	------ ------	1.0 1.0			
		N-01	V-04	------ ------				
		N-01	V-05	------ ------				
		N-01	V-06	------ ------				
		N-01	V-07	------ ------				
		V-01	V-01	0.153846 0.083333				
		V-01	V-04	0.166666 0.090909				
		V-01	V-05	0.142857 0.076923				
		V-01	V-06	0.166666 0.090909				
		V-01	V-07	0.142857 0.076923				
	Examination	N-01	N-05	0.588235 0.125				
		V-01	N-05	0.105263 0.055555				
	Quiz	N-01	V-01	------ ------				
		V-01	V-01	1.0 1.0				
	Screen	N-01	V-01	------ ------				
		V-01	V-01	0.153846 0.083333				
	Testing	N-01	N-01	0.5 0.090909				
		N-01	N-02	0.555555 0.111111				
		V-01	N-01	0.090909 0.047619				
		V-01	N-02	0.1 0.052631				

Tableau 8: Résultats de similarité (Wup vs Path)

Le tableau ci-dessus présente des statistiques sur les résultats de l'étape d'appariement de notre document vis-à-vis la requête (Quiz) en utilisant l'approche sémantique. Nous allons discuter ces résultats en trois parties :

Premièrement : nous calculons le score du synset de chaque mot par rapport au synset de la requête comme mentionné dans la figure 26. Ce score est calculé en utilisant deux mesures différentes de similarité sémantique : le **Wu Palmer** dont les résultats sont affichés en bleu et le **Path** comme indiqué en rouge. Pour mieux comprendre le fonctionnement de ces mesures, nous allons expliquer comment arriver aux résultats du score des synsets en donnant un exemple de l'arbre des hyperonymes pour les deux synsets « Quiz.N.01 » et « Testing.N.01 », en utilisant une fonction spéciale appelée « Hypernym Paths » :

1- WuPalmer :

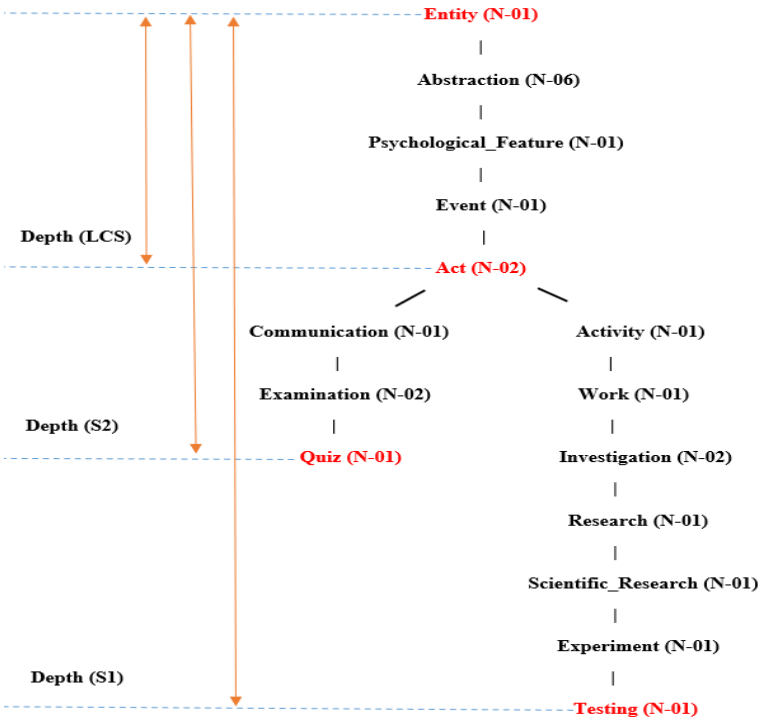


Figure 31: Représentation conceptuelle de « WuPalmer »

Avec :

- Depth (S1) représente le nombre de nœuds de S1 (Testing (N-01)) au nœud racine (Entity (N-01)), qui est égal à 12.
- Depth(S2)représente le nombre de nœuds de S2 (Quiz (N-01)) au nœud racine (Entity (N-01)), qui est égal à 8.
- Depth (LCS) représente le nombre de nœuds de LCS (Act (N-02)) au nœud racine (Entity (N-01)), qui est égal à 5.

Afin de calculer le score de similarité pour ces deux synsets, nous devons appliquer l'équation 8 mentionnée précédemment à notre exemple :

$$\boldsymbol{Score_{Wup} = 2 * \frac{\text{depth}\,(\text{Act}\,(\text{N02}))}{\text{depth}\,(\text{Testing}\,(\text{N01}))+\text{depth}\,(\text{Quiz}\,(\text{N01}))} = 2 * \frac{5}{12+8} = 0.5} \qquad (12)$$

2- Path:

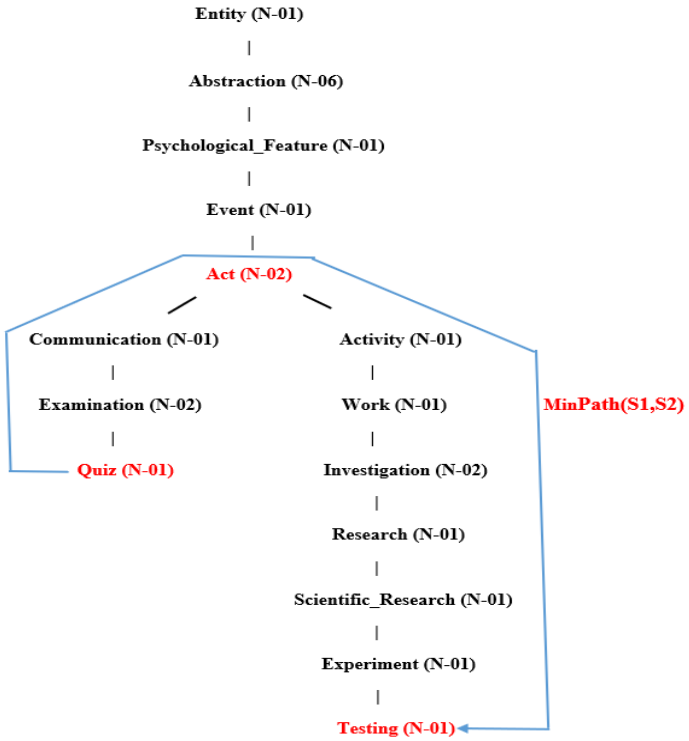


Figure 32: Représentation conceptuelle de « Path »

Où le MinPath (S1, S2) représente le chemin le plus court entre les deux synsets c.-à-d. Le nombre de nœuds de S1 (Quiz (N-01)) à S2 (Testing (N-01)), qui est égal à 11.

L'application de la formule de mesure de Path précédemment définie à cet exemple donne le résultat suivant :

$$\boldsymbol{Score_{Path} = \frac{1}{MinPath\left((Quiz\,(N01)),(Testing\,(N01))\right)} = \frac{1}{11} = 0.090909} \qquad (13)$$

Avec cela, nous avons atteint les résultats montrés dans le tableau 8. C'est ainsi que nous traitons tous les autres synsets de notre système. Cependant, il existe des synsets qui n'ont pas de score, parce qu'il n'y a pas de chemin qui les relie aux synsets de requête dans l'arbre conceptuel, ce qui rend impossible le calcul de similarité.

Deuxièmement : nous calculons les scores des mots comme indiqué dans la figure 26. Les résultats de chaque mesure sont visualisés dans le diagramme suivant :

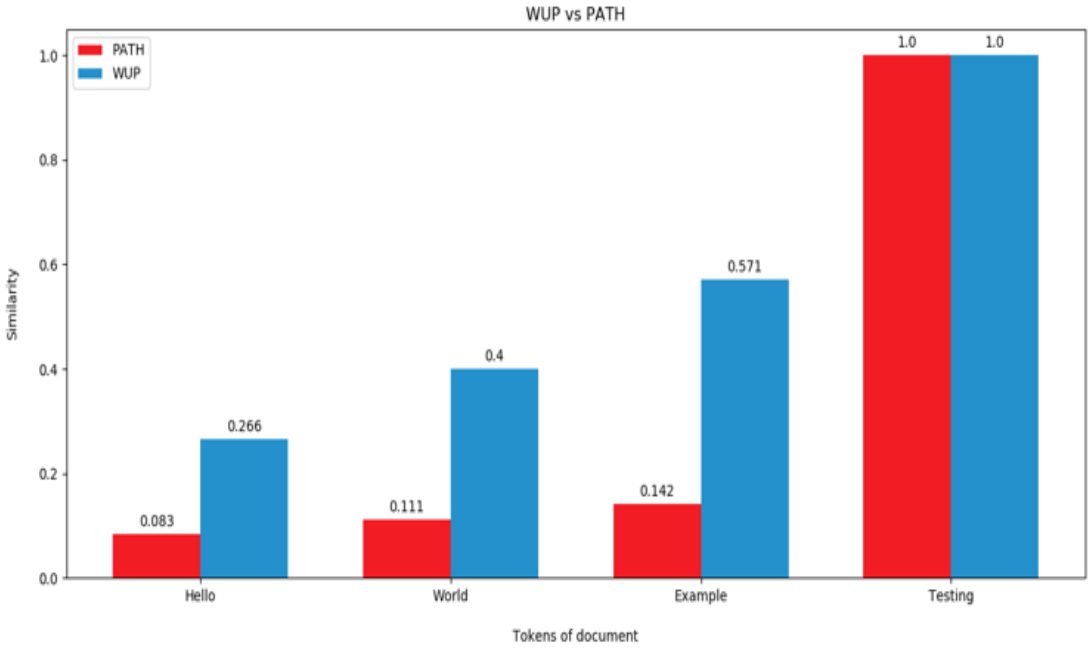


Figure 33: Scores des mots (Wup vs Path)

Ce diagramme montre les statistiques de chaque mot du document en fonction du degré de similarité sémantique avec la requête en appliquant les deux mesures (Wup et Path).

Nous remarquons que les scores des trois mots (Hello, World et Exemple) sont plus élevés dans le cas du Wup que dans le cas du Path. Alors que le score du mot (Testing) est équivalent dans les deux cas, qui est considéré comme un score parfait car il représente le degré maximum parmi tous les scores. Cette différence entre les deux mesures peut affecter le score moyen, et donc le résultat final de similarité (score final) devient inégal.

Troisièmement : nous concluons que notre document est pertinent vis-à-vis la requête (Quiz), grâce au résultat final obtenu qui représente le degré de similarité entre le document et la requête (Wup :0.779761, Path : 0.667162) en utilisant les deux scores (parfait et moyen) comme nous avons indiqué dans la figure 29. Ce score final peut être considéré comme un critère pour retrouver tous les documents pertinents dans un corpus, comme le montre la figure 34 :

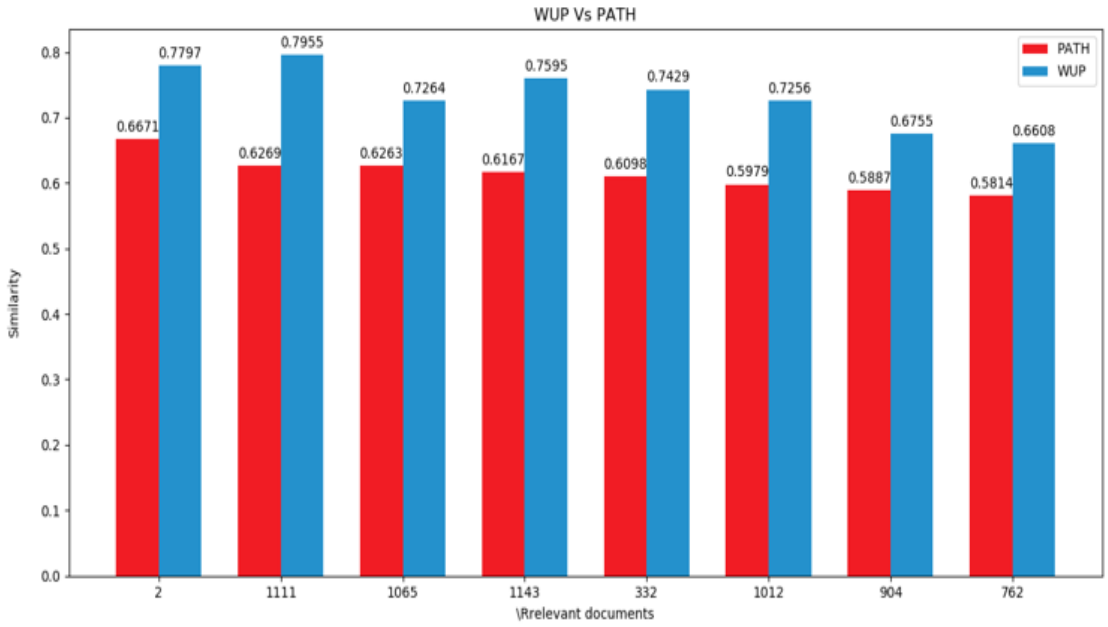


Figure 34: Scores des documents pertinents (Wup vs Path)

Le diagramme ci-dessus présente les documents pertinents (leurs index) à la requête (Quiz) y compris notre document (index : 2) en fonction du degré de similarité de chaque document. Ces documents (Tweets) qui étaient stockés dans la collection (Twitter) dont nous avons parlé précédemment, sont restitués en appliquant notre approche de similarité sémantique à travers les deux mesures (Wup et Path). Nous remarquons que les scores de tous les documents pertinents sont plus élevés dans le cas du Wup que dans le cas du Path. Par conséquent, ces deux mesures peuvent retrouver les mêmes documents pertinents mais pas les mêmes degrés de pertinence.

Par la suite, nous allons discuter des résultats obtenus en appliquant les deux approches ensemble (syntaxique, sémantique) c.-à-d. le fonctionnement de tout le système. Nous allons d'abord incorporer notre exemple (document) dans les quatre fichiers mentionnés précédemment, qui contiennent des collections de documents extraites des réseaux sociaux (YouTube, Twitter, Facebook). Ensuite, pour chaque collection, nous allons restituer tous les documents (contenus textuels) pertinents vis à vis de chacune des requêtes suivantes :

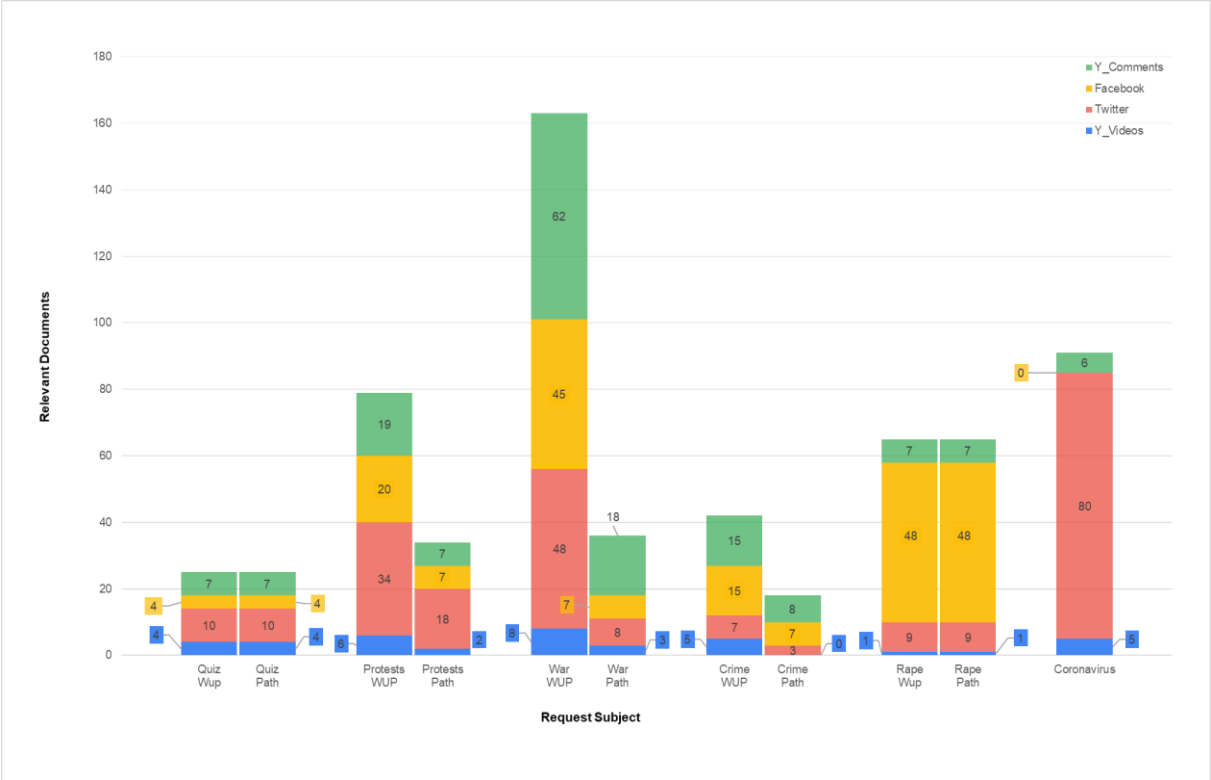


Figure 35: Documents pertinents aux différentes requêtes

Le diagramme ci-dessus montre le nombre de documents pertinents à chacune des requêtes suivantes: Quiz, Protests, War, Crime, Rape et Coronavirus. Ces documents sont retrouvés en appliquant notre algorithme de similarité à chaque collection de documents c.-à-d. aux contenus de chaque fichier que nous avons mentionné précédemment (Facebook, Twitter, Vidéos YouTube et Commentaires YouTube). Nous remarquons que les documents trouvés par le Wup sont nombreux par rapport à ceux trouvés par le Path pour certains sujets (Requêtes) telles que Protests, War et crime. Alors que dans certains cas (ex. Quiz, Rape), les deux mesures renvoient les mêmes documents pertinents. Tandis qu'il existe des documents qui ne sont renvoyés que par une approche syntaxique comme le cas de la requête (Coronavirus). Enfin, il peut y avoir aucun document pertinent à une requête dans tout le corpus comme nous constatons dans les deux cas : la collection de « Facebook » et la collection de « vidéos YouTube » concernant les requêtes « Coronavirus » et « Crime » respectivement.

II.1.4.1 Classement de résultats (Pertinence textuelle)

Dans le but de satisfaire la requête de l'utilisateur, il est nécessaire de trier le tableau des scores finaux par ordre décroissant selon le degré de similarité afin de pouvoir restituer les documents les plus pertinents.

Pour mieux comprendre, nous sélectionnons les documents pertinents vis-à-vis de la requête (Quiz) présentés dans le diagramme précèdent, qui ont été extraits syntaxiquement ou sémantiquement à travers le Wup uniquement. Ensuite, pour chaque fichier, nous classons ces documents selon le degré de similarité. Les résultats sont présentés ci-dessous :

Classement de documents	Index	ID	Texte	Degré de similarité
1	19	866807288118298812_1015451290	Hello world! this is just an example for <testing>.	**0.779761**
2	30	866807288116983812_1015451241	Colorado groom bit by a rattlesnake while taking wedding photos. "I admire Laura and Johnny so much for staying calm and holding each other\'s hand through the first great test of their brand-new marriage."	**0.682905**
3	339	866807288115653812_1015449384	Records show the Orlando shooting suspect caused trouble at an early age and, as an adult, became a proficient marksman with consistently high scores in tests.	**0.681501**
4	945	866807288114198812_1015446709	Judge allows Mississippi teen to serenade strangers instead of going to jail for failing drug test while on probation.	**0.674108**

Tableau 9: Classement des statuts pertinents sur Facebook (Pertinence textuelle)

Le tableau ci-dessus montre le classement de documents pertinents en fonction du degré de similarité. Ces documents représentent les quatre statuts pertinents sur le réseau social « Facebook » vis-à-vis de la requête (Quiz). Nous remarquons que notre document (index :19) se classe premier avec le score le plus élevé.

Classement de documents	Index	ID	Texte	Degré de similarité
1	178	akiXEfWW-V0	Coronavirus tests: how they work and what they show	**0.875816**
2	187	ldbQQo8XzCo	IELTS SPEAKING TEST - BAND 9 - MUST WATCH BEFORE YOUR EXAM	**0.780112**
3	2	QxJPKeqfn7c	Hello world! this is just an example for <testing>.	**0.779761**
4	206	14mRmD8zHOk	Watch what it's like to get tested for COVID-19	**0.763888**

Tableau 10: Classement des vidéos pertinentes sur YouTube (Pertinence textuelle)

Le tableau 10 présente les quatre documents pertinents par rapport à notre requête, qui sont les titres de vidéos YouTube. Dans ce classement, notre document (index :2) se classe troisième, ce qui signifie qu'il est moins pertinent que les deux premiers documents (Index: 178, 187).

Classement de documents	Index	ID	Texte	Degré de similarité
1	2	QxJPKeqfn7c	Hello world! this is just an example for <testing>.	**0.779761**
2	1526	7XsviBxbLvI	Poor experimental design...one test subject, one test run, not blind, not placebo controlled.	**0.749817**
3	1506	7XsviBxbLvI	Couldn't the 10% increase be due to the practice effect of repeating the same test twice? What they should've done was take 3 trials: 1 without Modafinil, 1 with Modafinil, and a placebo.	**0.705680**
4	1565	VFWD9fRtmMk	Imagine writing tests here. Who won world war 2? Kim, who discovered flamingos? Kim, how does Newton's laws work? I don't know but Kim made it.	**0.699115**
5	504	yL9UJVtgPZY	I took this test for one of my classes at school and it literally drives you crazy. It's hard to concentrate on tasks and interact with others. It's scary that too many people have to go through this in real life.	**0.687424**
6	1503	7XsviBxbLvI	Guy buys unknown pills online, doesn’t test them in a lab, doesn’t eat or drink properly and thinks he has taken modafinil. Worst way to present smart drugs.	**0.680213**
7	1094	k51L0MkRO8E	Everyone in Beirut needs to get a “I survived 3 Kiloton explosions” T-Shirt. Trinity test only used 100 tons of tnt for comparison test for nuke blast. That was a 0.1 kiloton explosion of tnt!! Yea I’d be very pissed at my gov if it happened in my city.	**0.677844**

Tableau 11: Classement des commentaires pertinents sur YouTube (Pertinence textuelle)

Les résultats présentés ci-dessus (tableau 11) montrent le classement des commentaires YouTube pertinents en fonction du degré de similitude. Nous notons que notre document (index: 2) est le plus pertinent parmi les sept documents retrouvés.

Classement de documents	Index	ID	Texte	Degré de similarité
1	1123	1306737948369522690	Quiz of the week: Who scored the Premier League's opening goal?	**1.0**
1	441	1309268147619483648	Quiz of the week: Who made history at the 'pand-Emmys'?	**1.0**
2	1111	1306873029100699653	Although these numbers reflect more comprehensive testing, it also shows alarming rates of transmission «The WHO…	**0.795516**

3	2	1302018208900280325	Hello world! this is just an example for <testing>.	**0.779761**
4	1143	1306637306489647105	BBC News: Is it a cold, flu or coronavirus? Here's some advice which could help you work out if you need to get tested for Covid-19 h…	**0.759523**
5	332	1310651881405009927	New Covid-19 test will give results 'in minutes' and is set to roll out in 133 nations.	**0.742900**
6	1065	1307069469500608514	US health chiefs reverse advice on Covid-19 testing that said people without symptoms should not get tested.	**0.726493**
7	1012	1307688166179119104	World Health Organization agrees rules for testing of African herbal remedies to fight Covid-19.	**0.725661**
8	904	1309802677224828929	A giant robot based on a character from a classic anime series has undergone testing in the Japanese city of Yokohama… pissed at my gov if it happened in my city.	**0.675515**
9	762	1310373173888774146	A giant robot resembling the 1970s anime figure Gundam has been tested in Yokohama, Japan.	**0.660883**

Tableau 12: Classement des tweets pertinents sur Twitter (Pertinence textuelle)

Nous présentons les scores de similarité correspondants aux documents pertinents restitués. Ces documents qui sont des Tweets sur le réseau « Twitter », sont classés par ordre décroissant selon leurs scores. Nous constatons que deux documents sont classés en premier (index : 1123, 441) et sont les plus pertinents de tous les documents, grâce à leurs scores élevés (1,0). Les scores signifient que ces documents sont pertinents à 100% à la requête (leurs syntaxes sont identiques). Notre document (index : 2) est classé en troisième place comme le montre le tableau 12 et la figure 36 :

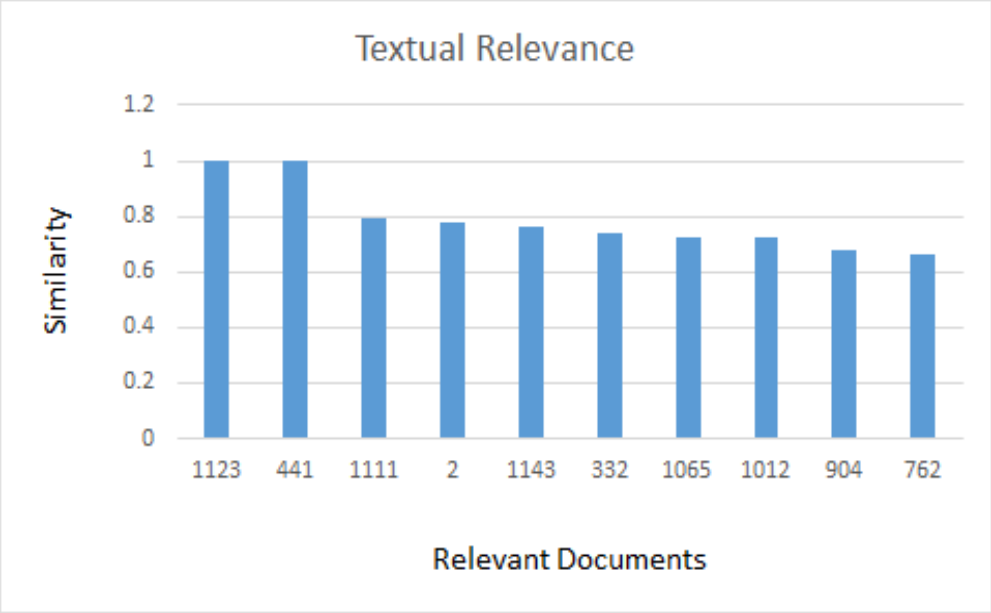


Figure 36: Classement de documents pertinents (Twitter)

II.1.4.2 Reclassement de résultats (Pertinence sociale)

Afin d'améliorer le classement des documents retournés vis-à-vis d'une requête en utilisant l'information sociale. Notre approche est basée sur la pertinence sociale qui se réfère à des facteurs sociaux (signaux sociaux), qui caractérisent un document en termes d'importance.

L'algorithme suivant illustre le reclassement de documents pertinents (signaux textuels) en exploitant leurs actions sociales (signaux non textuels) :

Entrée : tableau de signaux sociaux
score de similarité
Sortie : tableau des taux de classification
Début :
somme de signaux sociaux ⟵ Somme (signaux sociaux)
somme normalisée ⟵ Normalisation (somme de signaux sociaux)
taux ⟵ score de similarité * somme normalisée
Insérer le taux de chaque ligne correspondante dans le tableau des taux de classification
Trier le tableau des taux de classification par ordre décroissant et afficher les documents triés
Fin.

Figure 37: Algorithme de reclassement des résultats

Où nous utilisons une méthode de normalisation appelée « *min-max normalization* » (ou *min-max scaling*), qui consiste à redimensionner la plage de valeurs pour mettre à l'échelle la plage en [0, 1]. La formule générale pour un min-max de [0, 1] est donnée par [88]:

$$x' = \frac{x - min\,(x)}{max\,(x) - min(x)} \qquad (14)$$

Où : x est une valeur d'origine, x' est la valeur normalisée.

Cette normalisation est appliquée à la somme des signaux sociaux de chaque document afin de la combiner avec le score de similarité thématique calculé précédemment. Cette combinaison permet de trouver le résultat final de notre recherche qui représente le taux de reclassement de documents pertinents en utilisant la pertinence sociale, à travers la formule suivante :

$$Taux_{classement} = Score_{similarité} * Normalisation\,(\sum_{i=1}^{n} A_i(x)) \qquad (15)$$

Où :

- x, représente le réseau social (fichier) : Twitter, YouTube Videos, YouTube comment ou Facebook.
- n, représente les actions sociales (signaux non textuels) dans le fichier (x) : Twitter (j'aime, retweet), YouTube Videos (vues, réactions, commentaires), YouTube comment (j'aime, réponse), Facebook (réactions, commentaires, partages).
- $A_i(x)$, représente le nombre de chaque action sociale dans le fichier (x).

Le tableau 13 présente des statistiques sur le nombre de signaux sociaux sur « Facebook ». Ces signaux qui représentent les commentaires, les partages et les réactions (J'aime, J'adore, Haha, Wouah, Triste et Grrr.) ont reclassé les documents, où nous constatons un échange dans l'ordre entre les deux documents (index :945,339). Alors que notre cas d'étude a maintenu son classement (**N⁰** 1) avec le taux le plus élevé (0.009465), car il contient le plus grand nombre des actions sociales.

Classement de documents	Index	ID	Nombre de réactions	Nombre de commentaires	Nombre de partages	Somme des actions (normalisées)	Degré de similarité	Taux de classement
1	19	866807288 11_1015451290 8298812	2275	644	119	0.012138	0.779761	**0.009465**
2	30	866807288 11_1015451241 6983812	1723	77	224	0.007963	0.682905	**0.005438**
3	945	866807288 11_1015446709 4198812	1475	51	195	0.006715	0.674108	**0.004527**
4	339	866807288 11_1015449384 5653812	295	230	47	0.001984	0.681501	**0.001352**

Tableau 13: Classement des statuts pertinents sur Facebook (Pertinence sociale)

Le tableau 14 montre le reclassement de documents pertinents retrouvés précédemment en fonction du taux de classement, qui combine le degré de similarité et la somme des actions sociales des vidéos sur YouTube (Vues, Réactions, Commentaires). Nous remarquons que notre exemple (index :2) occupe désormais la première place après qu'il était deuxième, grâce au plus grand nombre de vues.

Classement de documents	Index	ID	Nombre de vues	Nombre de réactions	Nombre de commentaires	Somme des actions (normalisées)	Degré de similarité	Taux de classement
1	2	QxJPKeqfn7c	21945510	43418	2251	0.537380	0.779761	**0.419028**
2	187	ldbQQo8XzCo	7181206	72143	16861	0.164805	0.780112	**0.128566**
3	178	akiXEfWW-V0	5243953	70833	27229	0.116004	0.875816	**0.101598**
4	206	14mRmD8zHOk	1658971	53617	16616	0.024567	0.763888	**0.018766**

Tableau 14: Classement des vidéos pertinentes sur YouTube (Pertinence sociale)

Contrairement aux résultats mentionnés dans le tableau précédent, les résultats du tableau 15 qui représente le reclassement des commentaires pertinents sur YouTube montrent que notre document (index :2) est devenu en deuxième position, après qu'il était en haut du classement, en raison du petit nombre de likes (734) et de réponses (12) par rapport au premier document du classement (index :1565) dont le taux dépasse 0.016023.

Classement de documents	Index	ID	Nombre de (J'aime)	Nombre de réponses	Somme des actions (normalisées)	Degré de similarité	Taux de classement
1	1565	VFWD9fRtmMk	4599	40	0.022919	0.699115	**0.016023**
2	2	QxJPKeqfn7c	734	12	0.003685	0.779761	**0.002873**
3	1506	7XsviBxbLvI	552	21	0.002830	0.705680	**0.001997**
4	1503	7XsviBxbLvI	411	10	0.002079	0.680213	**0.001414**
5	1526	7XsviBxbLvI	333	8	0.001684	0.749817	**0.001263**
6	504	yL9UJVtgPZY	27	0	0.000133	0.687424	**0.000091**
7	1094	k51L0MkRO8E	21	1	0.000108	0.677844	**0.000073**

Tableau 15: Classement des commentaires pertinents sur YouTube (Pertinence sociale)

Le tableau 16 et la figure 38 présentent des statistiques sur le nouveau classement des documents précédemment trouvés sur Twitter :

Classement de documents	Index	ID	Nombre de (J'aime)	Nombre de (Retweet)	Somme des actions (normalisées)	Degré de similarité	Taux de classement
1	762	1310373173888774146	1620	589	0.026565	0.660883	**0.017556**
2	332	1310651881405009927	990	281	0.015192	0.742900	**0.011286**
3	904	1309802677224828929	816	208	0.012197	0.675515	**0.008239**
4	1012	1307688166179119104	447	162	0.007165	0.725661	**0.005199**
5	2	1302018208900280325	249	87	0.003855	0.779761	**0.003006**
6	1065	1307069469500608514	244	70	0.003588	0.726493	**0.002607**
7	1123	1306737948369522690	106	27	0.001394	1.0	**0.001394**
8	1143	1306637306489647105	0	162	0.001745	0.759523	**0.001326**
9	441	1309268147619483648	66	12	0.000727	1.0	**0.000727**
10	1111	1306873029100699653	22	15	0.000230	0.795516	**0.000183**

Tableau 16: Classement des tweets pertinents sur Twitter (Pertinence sociale)

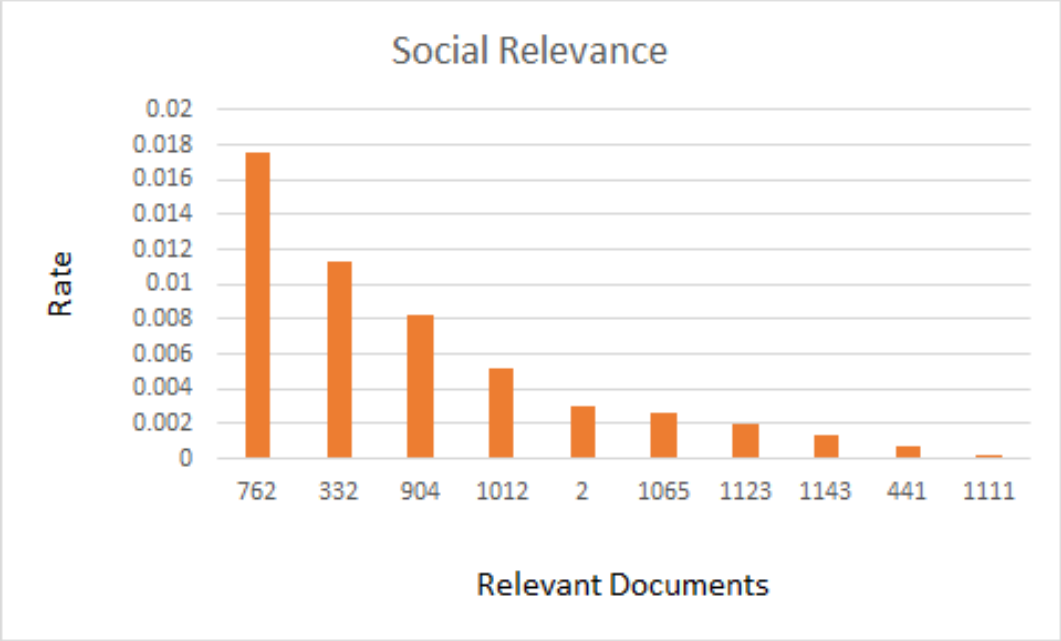


Figure 38: Reclassement de documents pertinents (Twitter)

La figure 38 représente le reclassement des documents de twitter de la figure 36 en fonction du score social global. Ceci prouve l'impact des actions sociales (likes, retweets) sur le classement de ces documents, comme le montre la figure 39 :

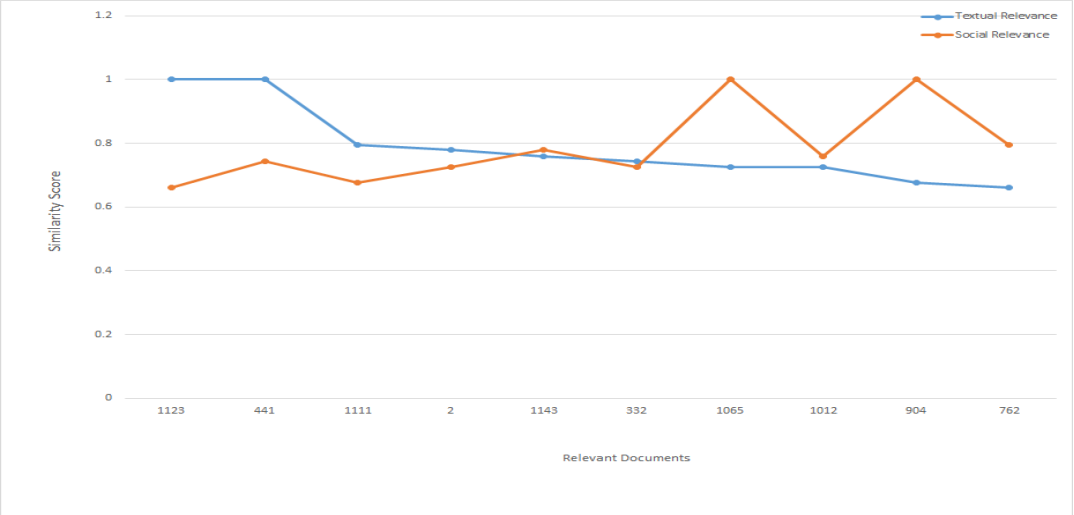


Figure 39: Impact des signaux sociaux

le graphique ci-dessus présente les deux courbes de classement des documents pertinents sur Twitter en fonction de leurs degrés de similarité. la première (en bleu) représente le classement des documents selon le score de pertinence textuelle, tandis que la seconde (en orange) représente le nouveau classement de ces documents selon la pertinence sociale.

II.1.5 Conclusion

Le tableau suivant récapitule et positionne notre SRI par rapport aux travaux de recherche antérieurs :

Source	Modèle	Auteur
Modèle de RI sociale pour l'accès aux ressources bibliographiques basée sur la pertinence thématique et de la pertinence sociale	Combinaison linéaire	LB Jabeur, L Tamine - 2010 [60]
Impact des signaux sur l'efficacité de la recherche sur YouTube	Combinaison linéaire et technique d'apprentissage	Chelaru et al. – 2012 [66]
Social Score Method basée sur plusieurs signaux et le TF-IDF	Combinaison linéaire	Marco Buijs, Marco Spruit. - 2014 [67]
Exploitation des Signaux Sociaux pour Améliorer la Recherche d'Information basée sur RSV	Combinaison linéaire Probabilité a priori (modèle de langue)	Ismail BADACHE. - 2016,2017a,2017b,2018 [25]
La recherche de commentaires pertinents dans Facebook, Twitter, YouTube basée sur WuPalmer & Path-similarity	Combinaison linéaire	Notre travail

Tableau 17: Limites et positionnement

Notre approche s’inspire donc de celle de Ismail Badache [25] qu’il propose d’améliorer, nous nous sommes partis donc de l’hypothèse qu’une ressource web doit être reclassé en fonction d’une combinaison linéaire de sa pertinence thématique et sociale.

Le rang d'une ressource dans le classement global des résultats est donc déterminé par son score global, qui est calculé par la combinaison des valeurs des différents scores, notre but est de fournir des résultats aussi pertinents que possible et aussi frais que possible.

CONCLUSION GÉNÉRALE

Le Web 2.0 a conduit à l'émergence des contenus sociaux générés par les utilisateurs (UGC) dans les services sociaux sur Internet. Ces UGC sont généralement évolutifs et de nature différente : des annotations sociales, des clics, des tweets, des commentaires, des relations sociales, des actions relevant d'activités sociales telles que le j'aime, le partage, le +1, le rating, etc. Les utilisateurs interagissent de plus en plus entre eux et/ou avec les ressources. Ces interactions associées aux ressources peuvent être considérées comme une des sources que l'on peut également exploiter pour améliorer la RI.

Pour ce mémoire de Master2, nous avons proposé un modèle de recherche d'information vectoriel basé sur les signaux sociaux textuels et non textuels. Ces signaux, pris séparément et groupée, sont considérés comme une information additionnelle permettant de mesurer la pertinence de la ressource à laquelle ils sont associés.

Cependant, notre travail présente quelques limites. D'abord, nous avons considéré que les signaux sont tous de même importance. Ils ne se différencient que par leur nombre vis-à-vis de la ressource correspondante. Selon nos résultats , il semblerait que certains soient plus important que d'autres pour la recherche d'information. Ensuite, nous n'avons pas pu évaluer l'impact d'autres signaux et les auteurs des différents signaux, sur le processus de RI. La récupération de ces informations n'est pas accessible via les APIs des réseaux sociaux actuels.

Une autre limite de notre travail réside dans la non prise en compte des facteurs temporels (temporalité des signaux, date de publication de la ressource, date du signal) ou des facteurs imagerie (images, émojis, ...). Nous pensons qu'un comptage simple de la quantité des signaux associés à une ressource privilégieront les ressources anciennes.

Nous traitons plus finement ces aspects dans le futur Inchaâllah où nous envisagerons inclure le domaine de la «*Learning Machine* » dans la recherche d'information sociale.

BIBLIOGRAPHIE

[1] «NLP | WuPalmer – WordNet Similarity,» 28 01 2019. [En ligne]. Available: https://www.geeksforgeeks.org/nlp-wupalmer-wordnet-similarity/. [Accès le 02 10 2020].

[2] R. Rada , H. Mili , E. Bicknell et M. Blettner , Development and application of a metric on semantic nets, vol. 19, IEEE Transactions on Systems, Man, and Cybernetics, 1989, pp. 17-30.

[3] Dridi, Amna, Haddad et Hatem, Recherche d'Information Sociale Reclassement des résultats de recherche à base de pertinence sociale, 2012.

[4] A. Bouramoul, «Recherche d'information contextuelle et sémantique,» Thèse de doctorat en informatique,Université Mentouri, Constantine, 2011.

[5] N. Jian-Yun, «Le domaine de recherche d'information – Un survol d'une longue histoire,» Département d'informatique et recherche opérationnelle,Université de Montréal.

[6] M. J. McGill et G. Salton, «introduction to modern information retrieval,» McGraw Hill Publishing Company, New York, 1983.

[7] V. Rijsbergen, C. J. Information retrieval, London: Butterworth, 1979.

[8] N. Ismail, «contribution à l'analyse et à la recherche d'information en texte intégral : Application de la transformée en ondelettes pour la recherche et l'analyse de texte,» 2010.

[9] s. Karbasi, «Modéle de ponderation basé sur le rang des termes dans les documents,» These de doctorat en informatique, Paul Sbatier, 2007.

[10] H.TEBRI, «Formalisation et specification d'un system de filtrage incrémental d'information,» Thèse de doctoorat de l'université Paul Sbatir, Toulous, 2004.

[11] M. Chein, «Introduction à la recherche d'information,» Master d'informatique, Montepellier, 2005.

[12] G. Mathias, «Indexation et interrogation de chemins de lecture en contexte pour la recherche d'information structurée sur le Web,» Thèse pour obtenir le grade de Docteur de l'université Joseph Fourier, 2002.

[13] «Indexation Non Supervisée de Documents Textuels,» Application aux documents Biomédicaux - Scientific Figure on ResearchGate, [En ligne]. Available: https://www.researchgate.net/figure/Processus-de-recherche-dinformation-15-Modeles-de-recherche-dinformation-Des-exemples_fig1_327395681. [Accès le 2020 11 09].

[14] N. Fuhr, Information Retrieval - From Information Access to Contextual Retrieval, Verlagsgesellschaft: In M. Eibl, C. Wolf, and C. Womser-Hacker, editors, Designing Information Systems. Festschrift für Jürgen Krause, 2005, pp. 47-57.

[15] M. Baziz, «Indexation conceptuelle guidée par ontologie pour la recherche d'information,» Thèse de doctorat , Université Paul Sbatier, 2005.

[16] H. Brini, «Un model de recherche d'information basé sur les réseaux posibilistes,» Thèse de doctorat en informatique, Université Paul Sbatier, 2005.

[17] N. Abbas, «vers une éxtension semantique d'analyse formelle de concept : Application à la recherche d'information,» Memoire de Magister , 2014.

[18] Cleverdon, «C. Progress in documentation. Evaluation of information retrieval systems,» *Journal of Documentation*, n° %126, pp. 55-67, 1970.

[19] S. Harter, «Psychological relevance and information science,» *Journal of the American Society for Information Science (JASIS)*, n° %143, pp. 602-615, 1992.

[20] G. R. E. Stephane , «Prise en compte du profil de l'utilisateur pour l'adaptation des systèmes d'information : Cas de TECHNIPEDIA,» Thèse de doctorat en informatiqu, University of Yaounde I, 2015.

[21] G. Salton, Automatic information organization and retrieval, 1968.

[22] . C. Jacquemin, B. Daille, J. Royanté et Polanc, In vitro evaluation of a program for machine-aided indexing, vol. 38, Inf. Process. Manage, 2002, pp. 765-792.

[23] R. El charif, «Analyse des paramettres de ponderation dans le cadre collections volumineuses,» DEA d'informatique , 2006.

[24] M. Ben Aouicha, «Une approche algebrique pour la recherche d'information structurée,» These de doctorat d'université Paul Sbatier, 2009.

[25] I.Badache, «exploitation des signaux sociaux pour améliorer la recherche d'information,» Thèse de doctorat en informatique,Université Paul Sabatier,Toulouse III, 2016.

[26] S. Helmut , Probabilistic part-of-speech tagging using decision trees, vol. 12, Citeseer: In Proceedings of the international conference on new methods in language processing, 1994, p. 44–49.

[27] M. F. Porter, *An algorithm for suffix stripping,* Program, 1980.

[28] E. M. Melvin et J. L Kuhns, «On relevance, probabilistic indexing and information retrieval,» *Journal of the ACM (JACM)*, n° %17(3):216–244,, 1960.

[29] N. Zimerli, «modele d'acces personnalisé a l'information bassé sur les diagrmmes d'influence integrant un profil d'utilisateur evolutif,» Thése de doctorat, l'université Paul Sbatier , Toulouse, 2009.

[30] Miller, «G.A,Miller.Word Net: A lexical data base for english,» In HLT, 1994.

[31] H. Alliane, «Un systeme de reformulation de requétes pour la recherche d'information,» Centre de recherche sur l'information scientifique et technique, alger, 2004.

[32] S. Dominich, Mathematical Foundations of Information Retrieval, London: Kluwer Academic Publishers, 2001.

[33] «File:Vector space model.jpg,» 20 10 2020. [En ligne]. Available: https://commons.wikimedia.org/wiki/File:Vector_space_model.jpg. [Accès le 09 11 2020].

[34] S. Robertson, «The Probability Ranking Principle in IR,» *Journal of Documentation,* n° %133, pp. 294-304, 1977.

[35] F. Boubkeur, «Contribution à la definition de modeles de recherche d'information flexibles basés sur les CP-NET,» These de doctorat en informatique, Paul Sbatier, 2008.

[36] . F. Boubekeur, «Contribution à la dé¿nition de modèles Àexibles de recherche d'information basés sur les CPNET,» Thèse de doctorat, Université Paul Sabatier, 2008.

[37] J. L. Fagan, «Experiments in Automatic Phrase Indexing For Document Retrieval:A Comparison of Syntactic and Non-Syntactic Methods,» Ph.D. thesis, Department of Computer Science, Cornell University, NY, 1987.

[38] W. .. R. Martin, B. P. AI et van Strenkenbu, On the Processing of Test Corpus: FroniTextual Data to Lexicographical Information. In Lexicography. Principles and Practice, R. R. K. I-Tartmann, London: Academic Press, 1983, pp. 77-88.

[39] E. Mittendort, B. Mateev et B. Schauble, Using the Co-occurrence of Words for Retrieval Weighting, 2000, pp. 243-251.

[40] R. Navigli, Word sense disambiguation: A survey, vol. 41, ACM Comput. Surv, 2009, pp. 1-69.

[41] . E. M. Voorhees, «Using WordNet to disambiguate word senses for text retrieval,» chez *International Conference on Research and Development in Information Retrieval*, 1993.

[42] C. Cleverdon, . J. Mills. et . M. Keen, Factors determining the performance of indexing systems, Cranfield: College of Aeronautics, 1966.

[43] B. GREGORY , le marketing digital, Dunod, 2016, pp. 12-13.

[44] . M. Florence, P. Serge et R. Julien , Web social. Mutation de la communication, Québec: Presses de l'Université du Québec, 2010.

[45] A. DUPIN, Communiquer sur les réseaux sociaux : les méthodes et les outils indispensables pour vos stratégies de communication sur les médias sociaux, FYP, 2010, p. 14.

[46] «Panorama des médias sociaux 2019,» 12 05 2019. [En ligne]. Available:

https://fredcavazza.net/2019/05/12/panorama-des-medias-sociaux-2019/. [Accès le 09 11 2020].

[47] R. ROMAIN , les réseaux sociaux (Facebook ,Twitter,LinkedIn,Viadeo, Google+..)Comprendre et maitriser ses nouveaux outils de communication, 2éme éditions, éni-éditions, octobre 2011.

[48] A. Dupin , Communiquer sur les réseaux sociaux, France: Editions Fyp, 2010, p. 93.

[49] E. Lazega, «Réseaux sociaux et structures relationnelles,» 1998.

[50] A. Kaplan et . M. Haenlein, Users of the world, unite! The challenges and opportunities, vol. 53, Business Horizons, 2010, p. 60.

[51] E. Dyson, «Enjeux et perspectives des réseaux sociaux,» 2006.

[52] «IEMS5720 Social Networking Sharing,» 15 03 2012. [En ligne]. Available: http://www.orbitalrpm.com/wp-content/uploads/2007/10/sna-graphic-7.jpg. [Accès le 09 11 2020].

[53] «Facebook--Wikipédia,» 28 6 2017. [En ligne]. Available: https://fr.wikipedia.org/wiki/Facebook. [Accès le 15 06 2020].

[54] Fanelli-Isla, Guide pratique des réseaux sociaux,, France: Dunod, 2010.

[55] «YouTube--Wikipédia,» 21 1 2019. [En ligne]. Available: https://fr.wikipedia.org/wiki/YouTube. [Accès le 15 06 2020].

[56] «Twitter--Wikipédia,» 18 10 2019. [En ligne]. Available: https://fr.wikipedia.org/wiki/Twitter. [Accès le 15 06 2020].

[57] «LinkedIn-Wikipedia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/LinkedIn. [Accès le 19 09 2020].

[58] F. Richter, «How Marketers Use Social Media,» 23 5 2014. [En ligne]. Available: https://www.statista.com/chart/2289/how-marketers-use-social-media/.

[59] R. B. Yates, User generated content: how good is it? In Proceedings of the 3rd workshop on Information credibility on the web, ACM, 2009, pp. 1-2.

[60] M. Boughanem, L. Tamine et L. Benjabeur, «Un modèle de recherche d'information sociale pour l'accès aux ressources bibliographiques vers un réseau social pondère,» Atelier de Recherche et de Recommandation d'Information dans les Réseaux sociaux, 2010.

[61] S. M. Kirsch, M. Gnasa et A. B. Cremers, Beyond the web:Retrieval in social information spaces. In Advances in Information Retrieval., 2006, pp. 84-95.

[62] O. Alonson, M. Gamon, K. Haas et P. pantel, Diversity and relevance in social search, 2012.

[63] A. Hammache , «Recherche d'Information : un modèle de langue combinant mots simples et mots composés,» Tizi-Ouzou, 2013.

[64] M. Koolen, G. Kazai et a. N. Craswell, Wikipedia pages as entry points for book search, In Proceedings of the Second ACM International Conference on Web Search and Data Mining, 2009, pp. 44-53.

[65] S. Bao, . G. Xue, . X. Wu et . Y. Y. B, Optimizing web search using social annotations, In Proceedings of the 16th international conference on World Wide Web, 2007, pp. 501-510.

[66] S. V. Chelaru, C. Orellana-Rodriguez et I. S, Can social features help learning to rank youtube videos?, Berlin: WISE, 2012, pp. 552-566.

[67] M. Buijs et M. Spruit, The social score - determining the relative importance of webpages based on online social signals, Rome-Italy: Proccedings of the International Conference on Knowledge Discovery and Information Retrieval, 2014, pp. 71-77.

[68] P. Pantel, . M. Gamon, . O. Alonso et K. Haas, Social annotations: Utility and prediction modeling, New York: In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 285-294.

[69] Y. Inagaki, . N. Sadagopan, G. Du, A. Dong, C. Liao, Y. Chang et Z. Zheng, Session based click features for recency ranking, vol. 10, In AAAI, 2010, p. 1334–1339.

[70] A. Khodaei et O. Alonso., Temporally-aware signals for social search, In SIGIR Workshop on Time-aware Information Access, 2012.

[71] «Anaconda (Python distribution),» 12 05 2019. [En ligne]. Available: https://fr.wikipedia.org/wiki/Anaconda_(Python_distribution). [Accès le 19 09 2020].

[72] «Spyder (logiciel),» 20 06 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/Spyder_(logiciel). [Accès le 19 09 2020].

[73] «Visual studio code,» 14 09 2020. [En ligne]. Available: https://edutechwiki.unige.ch/fr/Visual_studio_code#Installation. [Accès le 19 09 2020].

[74] «XAMPP,» 07 05 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/XAMPP. [Accès le 19 09 2020].

[75] «Composer (logiciel),» 19 06 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/Composer_(logiciel). [Accès le 20 09 2020].

[76] «LeBigData,» [En ligne]. Available: https://www.lebigdata.fr/python-langage-definition. [Accès le 20 09 2020].

[77] «PHP-Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/PHP. [Accès le 20 09 2020].

[78] «Natural Language Toolkit,» 28 03 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/Natural_Language_Toolkit. [Accès le 21 09 2020].

[79] «WordNet,» 04 06 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/WordNet. [Accès le 21 09 2020].

[80] «Pandas,» 20 06 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/Pandas. [Accès le 21 09 2020].

[81] «Web scraping,» 28 09 2020. [En ligne]. Available: https://fr.wikipedia.org/wiki/Web_scraping. [Accès le 27 10 2020].

[82] «On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-words Lists of International Languages,» Scientific Figure on ResearchGate, [En ligne]. Available: https://www.researchgate.net/figure/Snap-shot-of-First-25-Stop-words-of-First-19-Languages_fig1_306361861. [Accès le 09 11 2020].

[83] C. Leacock et M. Chodorow, Combining local context and WordNet similarity for word sense identification, vol. 49, WordNet: An electronic lexical database, 1998, p. 265–283.

[84] P. Resnik , Using information content to evaluate semantic similarity in a taxonomy, Montreal, Canada.: Proc of the 14th Intl Joint Conference on Artificial Intelligence, 1995, p. 448–453.

[85] J. Jiang et D. Conrath , Semantic similarity based on corpus statistics and lexical taxonomy, 1997, p. 19–33.

[86] D. Lin , An information-theoretic definition of similarity, Intl Conf ML Proc, 1998, p. 296–304.

[87] J. Daniel et H. M. James , «Speech and Language Processing,» 19 09 2018. [En ligne].

[88] «Feature scaling,» 16 08 2020. [En ligne]. Available: https://en.wikipedia.org/wiki/Feature_scaling?fbclid=IwAR2eGJ_7e4LhV_okpVGo5tm_Hebt51iTcH4u-EJoC2DueQoqPd7V5ONFDac#Rescaling_(min-max_normalization). [Accès le 03 10 2020].