République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Yahia Fares de MEDEA
Faculté des Sciences
Département De Mathématique Et Informatique



Mémoire de Fin d'Études
Pour l'obtention du Diplôme de Master en Informatique
Spécialité : Ingénierie des Systèmes et Technologies du Web

Thème :

Conception D'une Banque De Bonnes Requêtes Comme Moyen D'aide à la Reformulation des Nouvelles Requêtes De Recherche D'information

RÉALISÉ PAR :
- **CHAIRI Rania**
- **HADRI Wissal**

ENCADRÉ PAR :
- **Dr : RAFA Tahar**

Année universitaire : 2024/2025

Remerciements

En préambule à ce travail,

Je remercie Dieu le tout-puissant de m'avoir donné le courage, la force et la santé pour réaliser ce travail.

Nous tenons à adresser nos sincères remerciements au **Dr. RAFA Tahar**. Nous les remercions pour la confiance qu'ils nous ont témoignée en nous proposant ce sujet, leur disponibilité, leur patience, leurs conseils et directives très instructifs. Nous les remercions énormément pour l'autonomie qu'ils nous ont laissée tout en nous aiguillant sur des prises de réflexions riches et porteuses.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre travail en acceptant de l'examiner et de l'enrichir par leurs propositions.

Enfin, nous remercions tous nos enseignants pour la qualité de la formation qu'ils nous ont prodiguée au cours de 5 années, pour leurs conseils, leur disponibilité, ainsi que leur patience et leur encadrement.

Dédicace

À mes chers parents qui consacrent chaque jour leur énergie à leurs enfants, vous êtes les héros invisibles qui ont façonné mon avenir avec dévouement. Votre présence et vos sacrifices silencieux sont une véritable source d'inspiration. Merci d'être les piliers de ma vie et de me guider avec amour et bienveillance.

A mon mari *Oussama* pour son amour inconditionnel, son soutien constant et sa patience infinie durant ce parcours.

À mon cher frère et mes sœurs, vous êtes bien plus qu'une simple famille, vous êtes mes complices, mes confidents et mes amis pour la vie. Que notre lien spécial continue de grandir et de s'épanouir, car vous êtes les personnes les plus précieuses à mes yeux.

urs amies, vous êtes la famille que j'ai choisie, les bons moments que nous avons passés ensemble, votre écoute bienveillante et votre présence réconfortante sont des trésors que je chéris chaque jour. Je suis très reconnaissante de vous avoir connus.

CHAIRI Rania

Dédicace

À mes parents bien-aimés,

Il n'existe pas de mots assez forts pour exprimer toute ma gratitude.
Merci pour votre amour inconditionnel, vos prières silencieuses, vos sacrifices et votre soutien constant.
Vous êtes et resterez ma plus grande source de force et de motivation.

À mes frères et sœurs,

Pour votre complicité, vos encouragements sincères et votre présence rassurante à chaque étape.
Vos mots, vos regards, vos silences m'ont souvent donné le courage d'avancer.

À mes amis les plus chers,

Pour les moments partagés, les discussions réconfortantes et l'énergie positive que vous m'avez transmise.
Merci d'avoir été là, dans les hauts comme dans les bas.
Ce mémoire est aussi le reflet de votre amitié.

HADRI Wissal

Résumé :

Ce mémoire explore les défis de la recherche d'information (RI) dans un contexte de croissance exponentielle des données. Il met en lumière les limites des approches classiques et propose une solution basée sur la recherche d'information personnalisée (RIP). Une banque de données de requêtes a été conçue pour faciliter la reformulation des requêtes en s'appuyant sur les profils utilisateurs et leurs historiques de recherche, offrant ainsi des résultats plus pertinents. Une application basée sur Django a été développée pour illustrer cette approche, avec des fonctionnalités comme la suggestion automatique et la mise à jour dynamique des profils.

Mots-clés : Recherche d'information, personnalisation, requêtes, reformulation de requêtes.

Abstract :

This thesis examines the challenges of information retrieval (IR) amidst the exponential growth of data. It highlights the limitations of traditional approaches and proposes a solution through personalized information retrieval (PIR). A query database was designed to aid query reformulation by leveraging user profiles and search histories, delivering more relevant results. A Django-based application was developed to demonstrate this approach, featuring automatic suggestions and dynamic profile updates.

Keywords: Information retrieval, personalization, query, query reformulation

---

ملخص :

تستكشف هذه الأطروحة تحديات البحث عن المعلومات في ظل النمو المتزايد للبيانات. وتسلط الضوء على قيود المناهج التقليدية وتقترح حلاً يستند إلى البحث المخصص عن المعلومات. حيث تم تصميم قاعدة بيانات للاستعلامات لتسهيل إعادة صياغة الاستعلامات بناءً على ملفات تعريف المستخدمين وتاريخ بحثهم، وبالتالي تقديم نتائج أكثر صلة. كما تم تطوير تطبيق قائم على مكتبة ـ دجانغو ـ لتوضيح هذا النهج، مع ميزات مثل الاقتراح التلقائي والتحديث الديناميكي للملف الشخصي.

الكلمات المفتاحية : البحث عن المعلومات، الشخصنة، الإستفسارات، إعادة صياغة الاستفسارات

Table des matières

Liste des figures **ix**

Liste des tableaux **x**

Liste des acronymes **xi**

Introduction générale **1**

I Recherche d'Information Classique **3**

I.1 Introduction 5

I.2 Notion de base de la recherche d'information 5

I.3 Processus de recherche d'information 6

I.3.1 Expression du besoin en information 7

I.3.2 Processus d'indexation des documents 7

I.3.3 Reformulation de la requête 9

I.3.4 Appariement document-requête 11

I.4 Modèles de la recherche d'information 11

I.4.1 Modèle Booléen 11

I.4.2 Modèle vectoriel 12

I.4.3 Modèles Probabilistes 12

I.5 Évaluation des systèmes de recherche d'information 13

I.5.1 Collection de test 14

I.5.2 Protocole d'évaluation 15

I.5.3 Mesures d'évaluation 15

I.6 Conclusion 18

II Recherche d'Information Personnalisée **20**

II.1 Introduction 22

II.2 Émergence de la recherche d'information personnalisée 22

II.3 Modélisation de l'utilisateur 23

II.3.1 Pertinence utilisateur . . . 23
II.3.2 Profil utilisateur . . . 23
II.3.3 Sources des données de profil . . . 23
II.3.4 Acquisition des données de profil . . . 23
II.3.5 Techniques de construction des centres d'intérêts . . . 24
II.3.6 Représentation des données de profil . . . 24
II.3.7 Évolution du profil utilisateur . . . 24
II.4 Les dimensions du contexte de recherche . . . 24
II.4.1 Contexte géographique . . . 25
II.4.2 Contexte temporel . . . 25
II.4.3 Contexte social . . . 25
II.4.4 Contexte comportemental . . . 25
II.4.5 Contexte technologique . . . 25
II.4.6 Contexte cognitif . . . 26
II.5 Intégration du Profil Utilisateur dans le Processus de Recherche . . . 26
II.5.1 Reformulation de la requête . . . 26
II.5.2 Reclassement des Résultats . . . 28
II.6 L'évaluation des systèmes de recherche d'information personnalisés . . . 29
II.6.1 Les limites de l'évaluation classique . . . 29
II.6.2 La problématique de l'évaluation personnalisée . . . 30
II.6.3 Protocole d'évaluation personnalisée . . . 30
II.6.4 Mesures d'évaluation des systèmes personnalisés . . . 31
II.7 Conclusion . . . 33

III Conception d'une Banque de Bonnes Requêtes Comme Outil d'Aide a la Reformulation de Nouvelles Requêtes de Recherche d'Information 35

III.1 Introduction . . . 37
III.2 Profil d'utilisateur . . . 37
III.2.1 Données Statiques . . . 37
III.2.2 Donnes dynamiques . . . 38
III.3 Amis d'utilisateur . . . 40
III.3.1 Mesure de similarité entre utilisateurs . . . 40
III.4 Processus de reformulation . . . 41
III.4.1 Similarité entre requêtes . . . 42
III.4.2 Reformulation de requête . . . 42
III.4.3 Mise à jour du profil utilisateur . . . 43

III.5 Conception de la base de donnes 44
III.6 Réalisation du Système 45
III.6.1 Outils de Développement 45
III.6.2 Architecture de projet django 45
III.6.3 Présentation de L'application 47
III.7 Conclusion 50

Conclusion générale et perspectives **52**

Bibliographie **54**

Liste des figures

I.1 Processus de recherche d'information. 7
I.2 Modele vectoriel "VSM". 12
I.3 Rappel et précision. 16
I.4 Forme générale de la courbe de précision-rappel d'un SRI. 16

III.1 Schéma conceptuel du processus de reformulation de requête 41
III.2 le modèle conceptuel de données 44
III.3 Architecture d'un projet Django 46
III.4 Structure Générale du Projet 47
III.5 Login de l'application 48
III.6 Fenêtre d'Accueil 48
III.7 Page de recherche et reformulation 49
III.8 Fenêtre de confirmation de sauvegarde de requête 50

Liste des tableaux

I.1 Tableau de contingence 13

III.1 Table utilisateur 38
III.2 Structure simplifiée d'une requête dans le profil utilisateur 39
III.3 Structure simplifiée d'un tableaux terme dans le profil utilisateur 39

Liste des acronymes

CG	**C**gain **G**umulée
DCG	**G**ain **C**umulé **D**écoté
IDF	**F**réquence **I**nverse de **D**ocument
MAP	**P**récision **M**oyenne **A**rgée
MVC	**M**odèle-**V**ue-**C**ontrôleur
MVT	**M**odèle-**V**ue-**T**emplate
QRS	**S**ystème de **R**eformulation de **Q**uêtes
RI	**R**echerche d'**I**nformation
RSV	**V**aleur de **S**tatut de **R**écupération
SRI	**S**ystème de **R**echerche d'**I**nformation
SRIP	**S**ystème de **R**echerche d'**I**nformation **P**ersonnalisée
TF	**F**réquence du **T**erme
TF-IDF	**F**réquence du **T**erme - **F**réquence **I**nverse du **D**ocument
TREC	**C**onférence sur la **R**écupération de **T**exte
VSM	**M**odèle **S**patial **V**ectoriel

Introduction générale

Dans un contexte où les utilisateurs interagissent quotidiennement avec une quantité massive d'informations disponibles en ligne, la **recherche d'information personnalisée** est devenue un enjeu central pour améliorer la pertinence des résultats obtenus par les systèmes de recherche d'information. Contrairement aux approches classiques de recherche qui s'appuient uniquement sur le contenu de la requête, les systèmes de recherche personnalisée s'efforcent d'intégrer des dimensions contextuelles liées à l'utilisateur : ses préférences, son historique, son profil, ou encore son réseau social. Cette personnalisation vise à mieux cerner l'intention réelle de l'utilisateur derrière une requête souvent courte, ambiguë ou imprécise, afin de proposer des résultats plus adaptés à ses besoins d'information spécifiques.

Un obstacle majeur persistant souvent en face des approches de personnalisation est la formulation initiale des requêtes. De nombreux utilisateurs peinent à exprimer clairement leurs besoins d'information, ce qui entraîne des résultats peu satisfaisants. D'où l'importance de la reformulation de requêtes, une stratégie qui consiste à enrichir ou à corriger les requêtes initiales pour en améliorer l'efficacité. Dans ce contexte, concevoir une banque de bonnes requêtes, c'est-à-dire un ensemble structuré de requêtes jugées efficaces ou pertinentes, constitue une piste prometteuse. Un tel outil pourrait servir de guide pour aider les utilisateurs à reformuler leurs requêtes en s'appuyant sur des exemples issus de recherches similaires ou bien réussies.

Dans ce travail, nous proposons une approche de personnalisation de recherche d'information basée sur la construction d'un profil d'utilisateur pour modéliser et représenter l'utilisateur et ces intérêts. Les deux composantes les plus importantes de ce profil sont :

— la banque de bonnes requêtes des recherches précédentes de l'utilisateur. Une bonne requête est celle dont les résultats retournés par le système comportent au moins un résultat pertinent dans les 5 Top résultats.

— La liste des amis les plus similaires à l'utilisateur. Le principe d'amitié que nous adoptons est complètement implicite, et il est basé sur les intérêts communs entre les utilisateurs du système de recherche, sans exiger aucun échange explicite entre

eux.

Le processus de recherche propose, pour chaque nouvelle requête de l'utilisateur, un ensemble de requêtes candidates pour reformuler sa requête initiale.la sélection de ces requêtes candidates est faite à partir des banques des requêtes de l'utilisateur et de celles de ses amis les plus proches.

Des formules de calcul des poids des termes dans les profils, des similarités entre les requêtes et des similarités entre les utilisateurs sont proposées dans ce travail.

Le document est organisé en trois chapitres comme suit :

— **Le premier chapitre** explore les fondements de la recherche d'information classique, en détaillant ses concepts, processus, modèles et méthodes d'évaluation, tout en mettant enlumière ses limites

— **Le deuxième chapitre** se focalise sur la recherche d'information personnalisée, en examinant la modélisation des utilisateurs,l'intégration du contexte, et les défis liés à l'évaluation des systèmes personnalisés

— **Le troisième chapitre** présente la conception et la réalisation d'une banque de données de requêtes, en décrivant son architecture, les mécanismes de reformulation, et l'application développée pour soutenir ce processus.

Enfin, le mémoire se conclut par une synthèse des principales conclusions tirées des études menées et propose des perspectives pour de futures recherches dans ce domaine.

Chapitre I

RECHERCHE D'INFORMATION CLASSIQUE

I.1 Introduction

Dans un contexte marqué par une croissance exponentielle des données, la recherche d'information occupe une place cruciale pour accéder, organiser et exploiter les connaissances. Avec le développement rapide des technologies numériques, la quantité d'informations disponibles en ligne a considérablement augmenté, rendant la recherche à la fois indispensable et plus complexe.

La recherche d'information classique repose sur des méthodes et des modèles bien établis, conçus pour répondre aux attentes des utilisateurs en termes de pertinence et de précision des résultats. Ces méthodes s'appuient sur divers outils, tels que les moteurs de recherche, les bases de données et les systèmes de gestion documentaire, afin de localiser les informations pertinentes au sein de vastes ensembles textuels.

Parmi les principales approches classiques figurent les modèles booléens, vectoriels et probabilistes, qui se concentrent sur l'adéquation entre les termes des requêtes des utilisateurs et les documents disponibles dans les bases de données. Bien que ces approches aient démontré leur efficacité, elles présentent certaines limites, notamment dans leur capacité à satisfaire pleinement les besoins des utilisateurs, à intégrer le contexte, et à optimiser les requêtes pour améliorer la pertinence des résultats.

Ce chapitre vise à examiner les fondements de la recherche d'information classique tout en mettant en évidence ses limites, lesquelles ont conduit à l'émergence de nouvelles approches et méthodologies dans ce domaine. .

I.2 Notion de base de la recherche d'information

La recherche d'information est une démarche organisée et méthodique dont l'objectif principal est de répondre à un besoin spécifique en identifiant, évaluant et exploitant des ressources pertinentes. Ce processus s'appuie sur des concepts clés bien définis, qui ont été étudiés et formalisés dans le domaine des sciences de l'information et de la communication, ainsi qu'en informatique tel-que :

- **Document** : Un document est une unité d'information qui peut prendre diverses formes : texte, image, audio, vidéo, ou même une combinaison de ces éléments. Dans le contexte de la recherche d'information, un document est considéré comme une source potentielle de connaissances. Il peut s'agir d'un livre, d'un article scientifique, d'une page web, d'un rapport, ou d'une base de données. La qualité et la structure du document influencent son utilité dans le processus de recherche. Selon Buckland [1], un document est défini comme "tout ce qui peut être utilisé pour fournir une preuve ou une information", ce qui inclut aussi bien des supports physiques que numériques.

- **Requête** : Une requête est une expression formalisée du besoin en information, souvent formulée sous forme de mots-clés, de phrases ou de critères de recherche. Elle sert d'interface entre l'utilisateur et le système de recherche d'information. Une requête bien construite est essentielle pour obtenir des résultats pertinents. Par exemple, dans un moteur de recherche, une requête comme "impact du changement climatique sur l'agriculture en Afrique" permet de cibler des documents spécifiques. Ingwersen et Järvelin [2] soulignent que la formulation de la requête est un élément clé dans l'efficacité de la recherche, car elle influence directement la qualité des résultats retournés par le système.
- **Collection de documents** : La collection de documents, également appelée fonds documentaire ou corpus, représente l'ensemble des informations (ou documents) disponibles et accessibles pour consultation. C'est au sein de cette collection que s'effectuent la recherche et la sélection des documents considérés comme pertinents pour répondre à la requête de l'utilisateur. [3]
- **Besoin en information** : Le besoin en information représente la raison pour laquelle une recherche est initiée. Il peut être explicite (par exemple, trouver des statistiques sur le chômage en France) ou implicite (comme explorer un sujet pour se familiariser avec un domaine). Identifier clairement ce besoin est crucial pour orienter la recherche et formuler des requêtes efficaces. Le besoin en information évolue souvent au fil de la recherche, en fonction des découvertes et des nouvelles questions qui émergent. Kuhlthau [4].
- **Pertinence** besoin en information est à l'origine de toute recherche. Il évolue au fil du processus et doit être compris et affiné pour formuler des requêtes efficaces [4].

  On distingue deux formes principales :

  — Pertinence système : évaluée automatiquement par le système de recherche d'information (SRI), elle repose sur une fonction d'appariement notée $RSV(q, d)$, mesurant la similarité entre la requête q et le document d ;

  — Pertinence utilisateur : jugement subjectif de l'utilisateur après consultation du document. Elle constitue l'évaluation finale, mais reste difficile à modéliser automatiquement.

I.3 Processus de recherche d'information

Le processus de recherche d'information est une séquence d'étapes structurées qui permet de transformer un besoin en information en résultats pertinents. Voici une explication détaillée des étapes clés de ce processus comme le montre la figure I.1 .

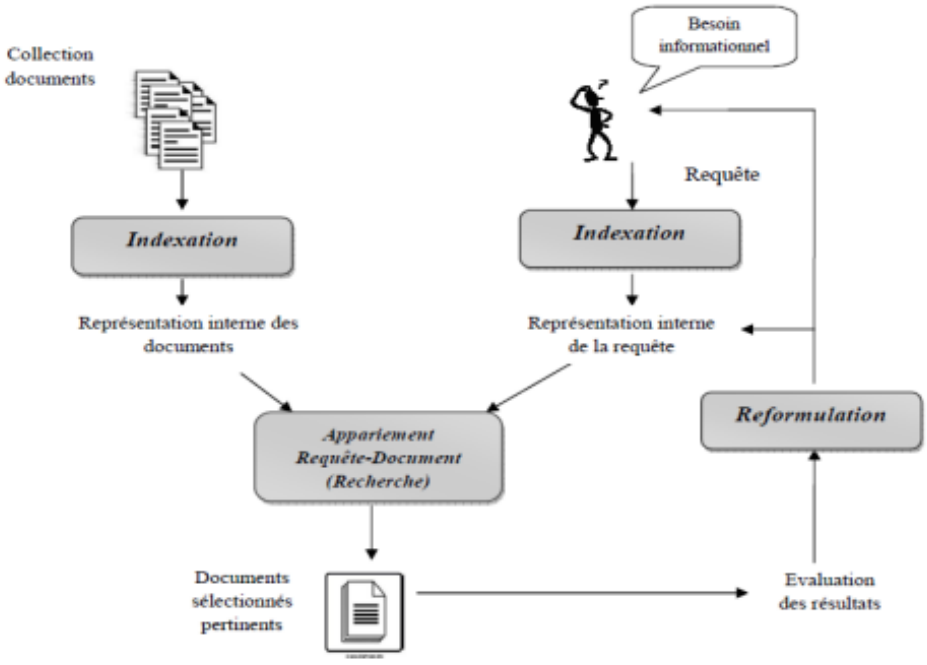


Figure I.1 — Processus de recherche d'information.

I.3.1 Expression du besoin en information

L'expression du besoin en information est la première étape du processus de recherche. Elle consiste à définir clairement ce que l'utilisateur cherche, en tenant compte du contexte, des objectifs et des contraintes. Selon Kuhlthau [4], cette étape est cruciale car elle influence toutes les étapes suivantes. Par exemple, un chercheur en sciences sociales pourrait exprimer son besoin en cherchant des "études récentes sur l'impact des réseaux sociaux sur la santé mentale des adolescents".

I.3.2 Processus d'indexation des documents

L'indexation est une étape technique qui permet de structurer et d'organiser les documents pour faciliter leur recherche. Elle comprend plusieurs sous-étapes :

- **Analyse lexicale :** L'analyse lexicale est une étape qui consiste à transformer le texte d'un document en un ensemble de termes, également appelés tokens en anglais. Un terme représente une unité ou entité lexicale. L'analyseur lexical identifie et traite les éléments tels que les espaces entre les mots, les chiffres, les ponctuations, etc., pour ensuite les éliminer ou les normaliser en vue d'un traitement ultérieur [3].
- **Élimination des mots vides :** L'élimination des mots vides consiste à retirer les termes non significatifs d'un texte afin de simplifier et de réduire la quantité de données à indexer [Daoud] [5]. Un mot vide est un mot qui n'apporte pas de sens spécifique et qui apparaît avec une fréquence similaire dans tous les textes d'une

collection. Par conséquent, il ne permet pas de différencier les documents les uns des autres. Un mot vide peut être un pronom personnel, une préposition ou tout autre mot grammatical. Il peut également s'agir d'un mot athématique, c'est-à-dire un terme qui peut apparaître dans n'importe quel document sans en traiter le sujet de manière spécifique (par exemple : contenir, appartenir, etc.). Ces mots, bien qu'utiles pour la structure linguistique, ne contribuent pas à la distinction ou à la caractérisation du contenu des documents [3].

- **Normalisation :** La normalisation vise à regrouper les mots similaires partageant la même racine en un seul terme, appelé lemme ou racine (souade) [6]. Pour identifier les racines des mots, plusieurs techniques sont utilisées, reposant généralement sur la suppression des terminaisons (comme celles liées au genre, à la conjugaison ou à la déclinaison). La normalisation s'appuie sur deux stratégies principales :
  - Lemmatisation : Elle consiste à représenter les différentes formes d'un mot appartenant à une même catégorie grammaticale par sa forme canonique, appelée lemme. Par exemple, les verbes "est, serons, fut,..." sont lemmatisés en leur forme infinitive "être" ;
  - Racinisation (ou stemming) : Elle consiste à réduire un ensemble de mots à leur racine commune (appelée stem) en supprimant leurs suffixes variés. Par exemple, les mots "marchand, marchande, marchandises" peuvent être racinisés en "marchand".

  Ces deux approches permettent de simplifier et d'unifier les termes pour faciliter leur traitement dans les systèmes de recherche et d'analyse.
- **Pondération des termes :** La pondération permet d'évaluer l'importance des termes dans les textes en leur attribuant des poids. Ces poids sont calculés à l'aide de formules de pondération, généralement basées sur la fréquence des termes dans le texte [3].
  - Pondération par fréquence simple (TF) : Certaines formules utilisent la fréquence simple de terme (TF - Term Frequency), qui mesure l'importance locale d'un terme dans un document. L'idée sous-jacente est qu'un terme qui apparaît fréquemment dans un document est plus pertinent pour décrire son contenu. Ainsi, un document est considéré comme pertinent pour une requête contenant un terme t si ce document présente un nombre élevé d'occurrences de t ;
  - Pondération par fréquence inverse de document (IDF) : D'autres formules reposent sur la fréquence inverse de document (IDF - Inverse Document Frequency), qui évalue l'importance globale d'un terme dans l'ensemble de la col-

lection de documents. Un terme très fréquent dans la collection est considéré comme moins important qu'un terme rare. L'IDF est calculée selon l'une des formules suivantes (la formule I.1 et la formule I.2) :

$$\mathrm{IDF}_t = \log\left(\frac{N}{\mathrm{df}_t}\right) \tag{I.1}$$

ou

$$\mathrm{IDF}_t = \log\left(\frac{N - \mathrm{df}_t}{\mathrm{df}_t}\right) \tag{I.2}$$

où df_t représente le nombre de documents contenant le terme t, et N est le nombre total de documents dans la collection ;

— Combinaison TF-IDF : Enfin, il existe des formules combinant les deux mesures, comme le TF-IDF, qui s'exprime par la formule I.3 :

$$\mathrm{TF} \times \mathrm{IDF} = \log(1 + \mathrm{TF}) \times \log\left(\frac{N}{\mathrm{df}_t}\right) \tag{I.3}$$

I.3.3 Reformulation de la requête

La reformulation de la requête (Query Reformulation) constitue une étape essentielle dans le processus de recherche d'information. Elle vise à améliorer la qualité des résultats en ajustant la requête initiale pour la rendre plus précise et plus adaptée aux besoins de l'utilisateur. Selon Ruthven [7], cette pratique permet de modifier la requête en fonction des retours d'expérience des résultats initiaux, augmentant ainsi l'efficacité de la récupération d'information. Plusieurs méthodes peuvent être utilisé pour y parvenir :

- **L'inclusion de synonymes des termes :** associés est une stratégie efficace pour élargir la portée de la recherche. Par exemple, lors d'une recherche sur "l'impact des réseaux sociaux", il est possible d'ajouter des termes tels que "médias sociaux" ou "effets psychologiques de Facebook". Comme le soulignent Baeza-Yates et Ribeiro-Neto [8] , l'utilisation de plusieurs termes connexes augmente les chances de trouver des informations pertinentes, en particulier dans les bases de données académiques.
- **Utilisation des opérateurs booléens (AND, OR, NOT) :** Les opérateurs booléens permettent d'affiner la recherche en combinant ou en excluant des termes spécifiques. Par exemple :

  L'opérateur AND (ex. : "réseaux sociaux AND santé mentale") restreint les

résultats aux documents contenant les deux termes.

L'opérateur OR (ex. : "apprentissage automatique OR intelligence artificielle") élargit la recherche pour inclure des documents contenant l'un ou l'autre terme.

L'opérateur NOT (ex. : "réseaux sociaux NOT marketing") exclut les documents contenant un terme indésirable. Selon Grossman [9], ces opérateurs sont indispensables pour structurer les requêtes et améliorer la pertinence des résultats.

- **Élargissement ou restriction du champ de recherche :** L'ajustement du champ de recherche en fonction des résultats obtenus est une pratique courante. Si les résultats sont trop larges, il est possible de restreindre la recherche en ajoutant des critères spécifiques. Inversement, si les résultats sont trop étroits, on peut élargir la recherche en utilisant des termes plus généraux. Par exemple :
  - Pour restreindre : "impact des réseaux sociaux sur les adolescents" → "impact des réseaux sociaux sur la santé mentale des adolescents" ;
  - Pour élargir : "réseaux sociaux" → "médias sociaux et communication numérique".

  Salton et Buckley [10] soulignent que cette flexibilité est cruciale pour atteindre un équilibre entre précision et rappel.
- **Utilisation de la troncature et des caractères génériques :** La troncature et les caractères génériques (ex. : *, ?) permettent de capturer des variations d'un mot. Par exemple : "informati*" pour inclure "informatique", "information", "informatisé", etc, et "réseau ?" pour inclure "réseaux" ou "réseau".

  Ces techniques sont particulièrement utiles pour gérer les variations linguistiques ou orthographiques [11].
- **Reformulation interactive avec rétroaction de pertinence :** La reformulation interactive repose sur l'analyse des résultats jugés pertinents par l'utilisateur. Par exemple, si un utilisateur identifie des documents pertinents, le système peut suggérer des termes supplémentaires ou modifier la requête pour mieux correspondre à ces documents. [12], dans son algorithme de rétroaction de pertinence, démontre que cette approche permet d'améliorer significativement la qualité des résultats.
- **Utilisation de thésaurus ou d'ontologies :** Les thésaurus et les ontologies sont des outils précieux pour structurer la recherche en reliant les termes entre eux. Par exemple, un thésaurus peut suggérer des termes hiérarchiquement liés, comme "réseaux sociaux" → "Facebook", "Instagram", "Twitter". [13] montre

que ces ressources améliorent à la fois la précision et le rappel des résultats.

I.3.4 Appariement document-requête

L'appariement est le processus qui consiste à comparer la requête de l'utilisateur avec les documents indexés pour identifier les résultats les plus pertinents. Selon Baeza-Yates et Ribeiro-Neto [14], Les documents sont ensuite classés par ordre de pertinence et présentés à l'utilisateur.

I.4 Modèles de la recherche d'information

Un modèle de Recherche d'Information (RI) a pour objectif de formaliser le processus de RI et de fournir un cadre théorique pour modéliser la mesure de pertinence. De nombreux modèles de RI textuelle ont été développés dans la littérature. Ces modèles partagent un vocabulaire d'indexation basé sur le formalisme des mots-clés, mais diffèrent principalement par leur approche d'appariement entre les requêtes et les documents.

Le vocabulaire d'indexation $V = \{t_i\}$, où $i \in \{1, ..., n\}$, est constitué de n mots ou racines de mots apparaissant dans les documents. Selon [5], un modèle de RI est défini par un quadruplet $(D, Q, F, R(q, d))$, où :

- D représente l'ensemble des documents,
- Q représente l'ensemble des requêtes,
- F est le schéma théorique de représentation des documents et des requêtes,
- $R(q, d)$ est la fonction de pertinence qui évalue la correspondance entre un document d et une requête q.

Dans ce qui suit, nous présentons les principaux modèles de RI :

I.4.1 Modèle Booléen

Le modèle booléen, introduit par [15], constitue le premier modèle formel de recherche d'information (RI) [15]. Il est fondé sur la théorie des ensembles, où chaque document est représenté comme un ensemble de termes.

Les requêtes sont formulées en combinant des mots-clés à l'aide d'opérateurs booléens classiques : **AND ()**, **OR ()** et **NOT (¬)**. L'appariement entre un document d et une requête q est binaire [16], défini par la fonction de pertinence suivante I.4 :

$$RSV(q,d) = \begin{cases} 1, & \text{si le document satisfait entièrement la requête} \\ 0, & \text{sinon} \end{cases} \tag{I.4}$$

I.4.2 Modèle vectoriel

Le modèle vectoriel de base, également appelé Vector Space Model (VSM), a été introduit par Salton dans le cadre du système SMART [15]. Dans ce modèle, un document est représenté par un vecteur à $|T|$ dimensions, où T désigne l'ensemble des termes de l'index. La requête est exprimée sous la forme d'un ensemble de mots-clés. Ainsi, les documents et la requête sont représentés comme des vecteurs dans un espace vectoriel (Voir la figure I.2) [3].

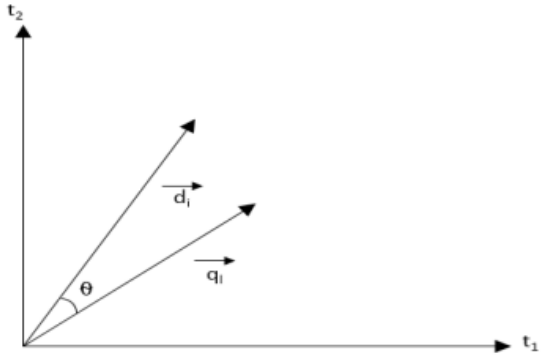


Figure I.2 — Modele vectoriel "VSM".

Cette représentation permet de calculer la similarité entre les documents et la requête en utilisant des mesures telles que le cosinus de l'angle entre les vecteurs. La similarité cosinus est définie comme suit I.5 :

$$\text{similarité}(q,d) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \cdot \|\vec{d}\|} \tag{I.5}$$

où :

- $\vec{q}$ est le vecteur de la requête ;
- $\vec{d}$ est le vecteur du document ;
- $\vec{q} \cdot \vec{d}$ est le produit scalaire des deux vecteurs ;
- $\|\vec{q}\|$ et $\|\vec{d}\|$ sont les normes (longueurs) des vecteurs.

I.4.3 Modèles Probabilistes

Le modèle probabiliste est basé sur la théorie des probabilités. Il repose sur l'hypothèse qu'il existe deux classes de documents pour une requête donnée, une classe de documents pertinents et une classe de documents non pertinents [17].

Le score de pertinence de système est calculé en comparant deux probabilités :

— La probabilité $p(d|q)$ qu'un document d soit pertinent pour une requête q ;

— La probabilité $p(d|\overline{q})$ qu'un document d soit non pertinent pour la même requête q.

Ces probabilités sont estimées en fonction de la présence ou de l'absence des termes de la requête dans les documents :

— $P(t \in d|d$ pertinent) : Probabilité p qu'un terme t de la requête apparaisse dans un document pertinent ;

— $P(t \in d|d$ non pertinent) : Probabilité q qu'un terme t de la requête apparaisse dans un document non pertinent.

Le score de pertinence $RSV(d, q)$ est donné par la formule suivante :

$$RSV(d, q) = \sum_{t \in q} \log \left(\frac{p_t(1 - q_t)}{q_t(1 - p_t)} \right) \quad (2.2)$$

où :

— p_t est la probabilité que le terme t apparaisse dans un document pertinent ;

— q_t est la probabilité que le terme t apparaisse dans un document non pertinent ;

— n est le nombre total de termes dans la requête.

Pour estimer les probabilités p_t et q_t, on utilise un tableau de contingence qui compte la présence ou l'absence des termes dans les documents pertinents et non pertinents [18]. Voici une représentation du tableau I.1 :

	Documents pertinents $Pert$	Documents non pertinents $\overline{Pert}$	Total
Terme t présent	N_t	$df_t - N_t$	df_t
Terme t absent	$\|Pert\| - N_t$	$(N - df_t) - (\|Pert\| - N_t)$	$N - df_t$
Total	$\|Pert\|$	$N - \|Pert\|$	N

Table I.1 — Tableau de contingence

Dans les modèles probabilistes initiaux, Les fréquences d'occurrence des termes ont été intégrées pour améliorer la précision du RSV, [19] menant au développement du modèle 2-Poisson et de la pondération BM25 [20].

I.5 Évaluation des systèmes de recherche d'information

L'évaluation d'un système de recherche d'information (SRI) vise à mesurer son efficacité à retrouver des documents pertinents et à évaluer sa performance globale.

L'objectif principal d'un SRI est de répondre de manière satisfaisante au besoin informationnel de l'utilisateur, en identifiant les documents pertinents tout en écartant ceux qui ne le sont pas.on distingue deux types de pertinence :

— La pertinence utilisateur : elle est déterminée par l'utilisateur lui-même, en fonction de son besoin spécifique en information ;

— La pertinence système : elle est estimée automatiquement par le SRI à l'aide d'une fonction de score appelée RSV(q, d), qui évalue la pertinence d'un document d par rapport à une requête q.

Pour que le système soit efficace, la pertinence estimée par le système doit se rapprocher au maximum de celle perçue par l'utilisateur [16].

I.5.1 Collection de test

Une collection de test est un ensemble structuré de document utilisé pour évaluer les performances d'un système de recherche d'information. Elle se compose généralement de trois éléments essentiels :

- **Corpus de documents :** Un corpus de documents est une collection de documents destinés à être indexés. Il représente l'ensemble des informations accessibles et exploitables sur lesquelles un système sera évalué.
- **Collection de requêtes :** Une collection de requêtes regroupe un ensemble de requêtes représentant les besoins informationnels des utilisateurs. Ces requêtes peuvent être artificielles, formulées par des participants à une campagne d'évaluation, ou réelles, extraites de logs de recherche provenant d'utilisateurs réels [21]. Elles servent à simuler ou à refléter les intentions de recherche des utilisateurs.
- **Jugements de pertinence :** Les jugements de pertinence sont des évaluations manuelles qui identifient les documents pertinents pour chaque requête. Ces jugements sont établis par des utilisateurs ou des experts, qui examinent chaque document pour déterminer sa pertinence par rapport à une requête donnée. Dans le cadre de grandes collections, comme celles utilisées dans les campagnes TREC (qui peuvent contenir plus d'un million de documents), il est impossible d'évaluer exhaustivement tous les documents. Par conséquent, les jugements de pertinence sont souvent construits à l'aide de la technique de pooling, qui se concentre sur les 1 000 premiers documents retournés par les systèmes participants [3]. Cette méthode permet de limiter l'effort d'évaluation tout en assurant une couverture représentative des résultats.

I.5.2 Protocole d'évaluation

Le protocole d'évaluation dans le modèle d'évaluation orienté-laboratoire établit une méthodologie rigoureuse et efficace pour comparer plusieurs systèmes de recherche d'information (SRI), stratégies de recherche ou algorithmes sur une même base. Ce protocole spécifie trois composants interdépendants : le nombre de sujets (topics) utilisés, les mesures d'évaluation utilisé, et la différence de performance nécessaire pour considérer qu'une stratégie de recherche est supérieure à une autre [22].

L'évaluation de l'efficacité de chaque stratégie de recherche repose sur l'analyse des résultats obtenus pour chaque requête de test. Cette évaluation se fonde sur la correspondance entre la pertinence algorithmique calculée par le système et la pertinence déterminée par les évaluateurs humains. L'efficacité globale d'une stratégie de recherche est calculée comme la moyenne des précisions obtenues selon une mesure donnée, appliquée à l'ensemble des sujets dans la collection de test.

Les protocoles d'évaluation utilisés s'appuient sur des métriques spécifiques. Dans ce qui suit, nous présentons les mesures les plus couramment utilisées, notamment les classiques "Rappel" et "Précision".

I.5.3 Mesures d'évaluation

Parmi ces critères, la précision et le rappel sont les plus couramment utilisés pour estimer l'efficacité d'un SRI, notamment sa capacité à sélectionner tous les documents pertinents et à rejeter les documents non pertinents. Des performances importantes.

- **Rappel et Précision :** Étant donné une requête q, les documents d'une collection peuvent être classés en fonction de leur pertinence par rapport à la requête (pertinents ou non pertinents), comme illustré dans la figure I.3.

  On considère $|S|$ comme le nombre de documents sélectionnés par un système de recherche d'information (RI) pour la requête q. De plus, on note $|P|$ le nombre de documents pertinents dans la collection pour cette requête, et $|P \cap S|$ le nombre de documents pertinents sélectionnés par le système [22].

  La mesure de **précision** évalue la capacité du système à rejeter les documents non pertinents pour une requête. Elle est définie comme le rapport entre le nombre de documents pertinents sélectionnés ($|P \cap S|$) et le nombre total de documents sélectionnés ($|S|$) :

$$\text{Précision} = \frac{|P \cap S|}{|S|}$$

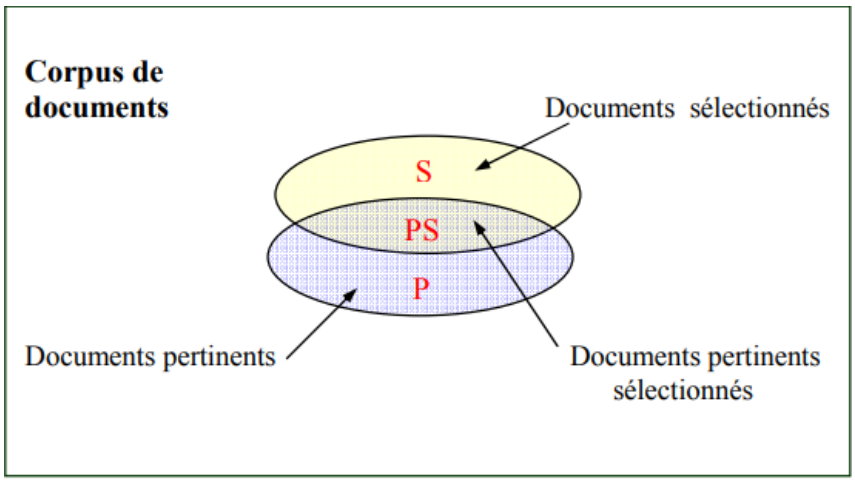


Figure I.3 — Rappel et précision.

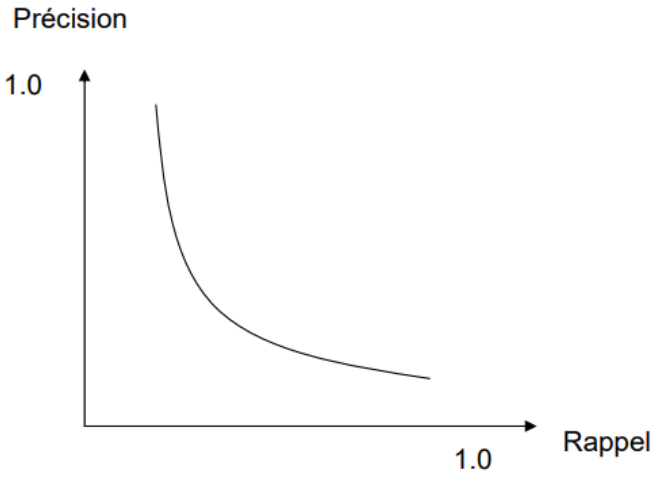


Figure I.4 — Forme générale de la courbe de précision-rappel d'un SRI.

D'autre part, la mesure de **rappel** évalue la capacité du système à retourner tous les documents pertinents disponibles dans la collection pour une requête donnée. Elle est définie comme le rapport entre le nombre de documents pertinents sélectionnés ($|P \cap S|$) et le nombre total de documents pertinents dans la collection ($|P|$) :

$$\text{Rappel} = \frac{|P \cap S|}{|P|}$$

Ces deux mesures sont couramment utilisées pour évaluer la performance d'un système de recherche d'information .

- **F-mesure :** La F-mesure (ou F-score) est une métrique qui combine la précision et le rappel en une seule valeur. Elle est définie comme la moyenne harmonique de la précision et du rappel, et est calculée par la formule

suivante :

$$F = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Cette mesure permet d'évaluer de manière équilibrée la performance d'un système en tenant compte à la fois de la capacité à retrouver des documents pertinents (rappel) et à éviter les documents non pertinents (précision).

- **Précision @X :** La Précision @X (notée P@X) est une mesure qui évalue la précision à différents niveaux de la liste des documents retournés par le système. Elle représente la proportion de documents pertinents parmi les X premiers documents restitués par le système. Par exemple, P@10 correspond à la proportion de documents pertinents retrouvés parmi les 10 premiers documents de la liste de résultats. Cette mesure est particulièrement utile pour évaluer la performance d'un système de recherche d'information dans un contexte où l'ordre des résultats est important, comme dans les moteurs de recherche.
- **R-précision :** La R-précision est une métrique qui mesure la précision du système après avoir retrouvé R documents pertinents pour une requête donnée. Par exemple, la 2-précision correspond à la précision calculée juste après que le système a retrouvé le deuxième document pertinent. Cette mesure est utile pour évaluer la performance d'un système en se concentrant sur un nombre spécifique de documents pertinents.
- **Précision moyenne (Mean Average Precision-MAP :** La Mean Average Precision (MAP) correspond à la moyenne des précisions moyennes (*Average Precision - AP*) obtenues sur l'ensemble des requêtes. Elle est calculée chaque fois qu'un document pertinent est retrouvé et s'exprime par la formule suivante :

$$MAP = \frac{\sum_{q \in Q} AP_q}{|Q|}$$

où :

— AP_q représente la précision moyenne pour une requête q ;

— Q est l'ensemble des requêtes ;

— $|Q|$ désigne le nombre total de requêtes.

Cette mesure est considérée comme globale, car elle intègre plusieurs points de mesure. Contrairement aux métriques P_n, elle est moins sensible au nombre total de documents pertinents.

La précision moyenne peut également être calculée à différents niveaux de rappel (0 %, 10 %, 20 %, ..., 100 %). Dans ce cas, elle est appelée précision moyenne interpolée (MAiP). Pour chaque niveau de rappel, les valeurs obtenues sont moyennées sur l'ensemble des requêtes.

- **La moyenne des réciproques des MRR :** La Mean Reciprocal Rank (MRR) est une métrique basée sur le rang qui permet d'évaluer le nombre de documents à examiner avant de trouver le premier document pertinent. Elle correspond à la moyenne, calculée sur l'ensemble des requêtes, de l'inverse du rang du premier document pertinent retrouvé.

  Elle est définie par la formule suivante :

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

  où :

  — $|Q|$ représente le nombre total de requêtes ;

  — rank_i est le rang du premier document pertinent retourné pour la requête i.

  Si aucun document pertinent n'est retourné pour une requête donnée, la valeur de MRR est nulle. En revanche, un MRR élevé indique que le système positionne les documents pertinents en haut des résultats présentés à l'utilisateur.

  Cette mesure est particulièrement utilisée dans les systèmes de questions-réponses, où l'objectif est d'afficher la bonne réponse en premier.

I.6 Conclusion

Ce premier chapitre a permis d'explorer les fondements de la recherche d'information classique, en mettant en lumière les concepts clés, les modèles et les méthodes qui sous-tendent ce domaine. Nous avons vu que la recherche d'information repose sur un processus structuré, allant de l'expression du besoin en information à l'appariement entre les requêtes et les documents, en passant par l'indexation et la reformulation des requêtes. Les modèles classiques, tels que le modèle booléen, le modèle vectoriel et les modèles probabilistes, offrent des approches variées pour évaluer la pertinence des documents par rapport aux requêtes des utilisateurs.

Cependant, malgré leur efficacité, ces modèles présentent certaines limites. Ils peinent à intégrer pleinement le contexte de l'utilisateur, à gérer l'ambiguïté des

termes et à optimiser les requêtes pour des résultats plus pertinents. Ces lacunes ont conduit à l'émergence de nouvelles approches, notamment dans le domaine de la recherche d'information personnalisée, qui sera abordée dans le chapitre suivant.

Chapitre II

RECHERCHE D'INFORMATION PERSONNALISÉE

II.1 Introduction

La recherche d'information constitue un pilier fondamental de l'accès à la connaissance dans l'ère numérique contemporaine. Face à l'explosion du volume des données disponibles en ligne, les utilisateurs sont confrontés à un double défi : formuler des requêtes précises et identifier les informations réellement pertinentes parmi une multitude de résultats. Les systèmes de recherche classiques, tels que les moteurs de recherche traditionnels, reposent sur des techniques standardisées de traitement des requêtes et d'indexation, qui s'appliquent de manière uniforme à tous les utilisateurs, sans distinction de leurs préférences ou contextes spécifiques.

Cependant, cette approche universelle montre ses limites lorsqu'il s'agit de répondre à des besoins complexes, variés, ou évolutifs. L'utilisateur se retrouve souvent confronté à des résultats trop génériques, voire inadéquats, en raison de l'absence de personnalisation. De plus, la formulation d'une requête efficace n'est pas toujours intuitive, surtout pour les non-spécialistes de la recherche documentaire. C'est dans ce cadre que se développent de nouvelles approches visant à améliorer la qualité et la pertinence des résultats, notamment par l'intégration de la personnalisation.

Ce travail s'inscrit dans cette dynamique en explorant les apports de la recherche d'information personnalisée, qui vise à adapter les résultats en fonction du profil, des préférences et du comportement antérieur de l'utilisateur. En exploitant les requêtes jugées pertinentes et en construisant des modèles d'utilisateurs, ces systèmes de nouvelle génération permettent une meilleure adéquation entre les intentions réelles des individus et les réponses proposées.

II.2 Émergence de la recherche d'information personnalisée

Ces dernières années, pour améliorer les performances des systèmes de recherche d'information (SRI), il est devenu nécessaire de prendre en compte des aspects implicites liés à l'utilisateur, tels que ses besoins, préférences, jugements de pertinence, et son contexte social ou spatio-temporel. Cela a conduit à l'émergence de la recherche d'information personnalisée (RIP), où l'utilisateur est placé au centre du processus de recherche. L'objectif de la RIP est d'adapter les résultats de recherche aux intérêts spécifiques de l'utilisateur.

Les systèmes de recherche classiques, décrits par le quadruplet (Documents, Requêtes, Schéma de représentation, Fonction de correspondance) [23], ne prennent pas en compte l'utilisateur et son contexte. Par exemple, une requête comme " Algérie vs Maroc " sur Google peut retourner des résultats variés (sport, politique,

etc.), mais ces résultats ne seront pas pertinents si l'utilisateur cherche une comparaison socio-économique. Cela montre la nécessité d'intégrer le contexte utilisateur pour améliorer la pertinence des résultats.

II.3 Modélisation de l'utilisateur

La modélisation de l'utilisateur est un processus clé pour la RIP(la recherche d'information personnelle). Elle consiste à acquérir, extraire et représenter les données caractérisant l'utilisateur. Sans un modèle d'utilisateur, un SRI traite tous les utilisateurs de la même manière, ignorant leurs différences en termes de connaissances, préférences et objectifs.

II.3.1 Pertinence utilisateur

La pertinence est une notion subjective qui dépend des besoins et du contexte de l'utilisateur. Deux utilisateurs peuvent juger différemment les mêmes résultats de recherche en fonction de leurs objectifs et de leurs connaissances. La modélisation de l'utilisateur permet de mieux comprendre ces mécanismes cognitifs et d'améliorer la pertinence des résultats [21, 23].

II.3.2 Profil utilisateur

Le profil utilisateur est une représentation structurée des centres d'intérêts, des connaissances et des objectifs de l'utilisateur. Il est essentiel pour adapter les résultats de recherche et suivre l'évolution des intérêts de l'utilisateur au fil du temps [5].

II.3.3 Sources des données de profil

Les données pour construire le profil utilisateur proviennent de diverses sources, telles que l'historique de navigation, les pages visitées, les interactions avec le système, et des informations contextuelles (localisation, temps, etc.) [22].

II.3.4 Acquisition des données de profil

L'acquisition des données peut être explicite (via des formulaires ou des feedbacks directs) ou implicite (via l'analyse des comportements de navigation). L'acquisition implicite, bien que moins précise, est plus dynamique et moins intrusive pour l'utilisateur [24].

II.3.5 Techniques de construction des centres d'intérêts

La construction des centres d'intérêts d'un utilisateur repose sur diverses approches qui visent à modéliser ses besoins informationnels de manière fine et évolutive. Parmi les techniques les plus répandues figurent l'extraction automatique de termes significatifs, l'identification de réseaux de cooccurrence entre ces termes, ainsi que la structuration de concepts issus de documents jugés pertinents. Ces méthodes permettent de capter non seulement les mots-clés fréquemment utilisés, mais également les relations sémantiques et les thématiques récurrentes propres à l'utilisateur [25]. Ainsi, elles facilitent la construction de profils riches, capables d'alimenter des systèmes de recommandation ou de recherche d'information personnalisée.

II.3.6 Représentation des données de profil

Le profil utilisateur peut être représenté de différentes manières :

- Historique de recherche : basé sur les requêtes et pages visitées.
- Ensembliste : utilisant des vecteurs de termes pondérés.
- Connexionniste : basé sur des réseaux de nœuds représentant des concepts.
- Multidimensionnelle : organisé en catégories (données personnelles, comportement, etc.).
- Conceptuelle : utilisant des ontologies ou hiérarchies de concepts [22].

II.3.7 Évolution du profil utilisateur

Le profil utilisateur doit évoluer pour refléter les changements dans les centres d'intérêts de l'utilisateur. On distingue le profil à court terme (lié aux activités récentes) et le profil à long terme (reflétant les intérêts persistants) [5].

II.4 Les dimensions du contexte de recherche

La recherche d'information personnalisée repose sur l'intégration de différentes dimensions contextuelles pour adapter les résultats aux besoins spécifiques de l'utilisateur. Ces dimensions permettent de capturer des informations variées sur l'environnement, les préférences et les comportements de l'utilisateur, afin d'améliorer la pertinence des résultats.

II.4.1 Contexte géographique

La localisation de l'utilisateur est une dimension clé pour personnaliser les résultats de recherche. Par exemple, un utilisateur cherchant un restaurant préférera des résultats proches de sa position actuelle. Les systèmes de recommandation et de recherche exploitent souvent des données de géolocalisation pour fournir des résultats adaptés à la situation géographique de l'utilisateur [26, 27].

II.4.2 Contexte temporel

Le temps est une autre dimension importante. Les besoins d'un utilisateur peuvent varier en fonction du moment de la journée, de la saison, ou d'événements spécifiques. Par exemple, une recherche sur les " festivals " en été retournera des résultats différents de ceux en hiver. Les systèmes de recherche peuvent utiliser des données temporelles pour ajuster les résultats en fonction du contexte temporel de l'utilisateur [28].

II.4.3 Contexte social

Les interactions sociales et les préférences collectives jouent un rôle crucial dans la personnalisation. Les systèmes de recommandation exploitent souvent les données sociales, telles que les avis des amis ou les tendances communautaires, pour affiner les résultats. Par exemple, un utilisateur peut préférer des recommandations basées sur les choix de son réseau social [29].

II.4.4 Contexte comportemental

Les comportements passés de l'utilisateur, tels que son historique de navigation, ses clics, et ses interactions avec les résultats de recherche, fournissent des indications précieuses sur ses préférences. Ces données permettent aux systèmes de prédire les besoins futurs de l'utilisateur et d'adapter les résultats en conséquence [26].

II.4.5 Contexte technologique

Le type d'appareil utilisé (smartphone, tablette, ordinateur) et les caractéristiques techniques (taille de l'écran, connexion Internet) influencent également la recherche d'information. Par exemple, les résultats peuvent être adaptés pour s'afficher correctement sur un petit écran ou pour limiter l'utilisation de données mobiles [27].

II.4.6 Contexte cognitif

Enfin, le contexte cognitif prend en compte les connaissances, les intérêts et les objectifs spécifiques de l'utilisateur. Cette dimension est souvent modélisée à travers le profil utilisateur, qui capture les centres d'intérêts et les préférences de l'utilisateur pour guider les résultats de recherche [29].

II.5 Intégration du Profil Utilisateur dans le Processus de Recherche

L'intégration du profil utilisateur dans le processus de recherche d'information permet d'améliorer la pertinence des résultats retournés en tenant compte des préférences et du comportement passés de l'utilisateur. Le profil de l'utilisateur, une fois construit, est intégré dans des phases spécifiques du processus de recherche, telles que la reformulation de la requête initiale ou le reclassement des résultats.

II.5.1 Reformulation de la requête

Les utilisateurs expriment souvent leurs besoins d'information avec des requêtes très courtes et parfois ambiguës. Par exemple, une requête simple comme "Apple" peut faire référence à une pomme dans le domaine de l'agriculture ou à la célèbre entreprise technologique. Cette ambiguïté nécessite souvent une reformulation de la requête pour mieux comprendre l'intention de l'utilisateur.

La reformulation de requête est l'une des applications les plus courantes de l'intégration du profil utilisateur dans le processus de recherche. Le système utilise les informations contenues dans le profil de l'utilisateur pour affiner ou élargir la requête afin d'obtenir des résultats plus précis. La transformation de la requête peut se faire de deux manières :

- Raffinement de la requête : Cette approche vise à réduire la requête en éliminant des termes inutiles.
- Expansion de la requête : Cette approche consiste à ajouter des termes supplémentaires afin d'élargir la recherche et de diminuer l'ambiguïté de la requête.

II.5.1.1 Approche de Schenkel et al. (2008) et Xie et al. (2012)

- Type de profil utilisateur : Profil vectoriel ou liste de préférences (termes ou entités d'intérêt extraits de l'historique de navigation, de clics, ou de documents consultés).

— Utilisation du profil : Les termes du profil sont directement ajoutés à la requête de l'utilisateur. Il n'y a pas de traitement sémantique ou pondération : c'est une fusion brute des termes.

Principe :"fusion simple des termes", Cette méthode repose sur une intégration directe des termes issus du profil utilisateur à ceux de la requête initiale, sans recours à des mécanismes de pondération complexes. L'objectif est de personnaliser la requête en fonction des intérêts ou de l'historique de navigation de l'utilisateur.

Exemple : une requête " football " associée au profil d'un utilisateur intéressé par " Cristiano Ronaldo " deviendra " football Cristiano Ronaldo ".

II.5.1.2 Approche de Wen et al. (2004)

— Type de profil utilisateur : Historique de recherche ou documents jugés pertinents par l'utilisateur (implicite).
— Utilisation du profil : Le système analyse les cooccurrences entre les termes des requêtes et ceux des documents que l'utilisateur a trouvés pertinents (clics, temps passé, etc.), pour enrichir la requête avec des termes pertinents.

Principe : "corrélation entre les termes", Cette approche s'appuie sur l'analyse des relations statistiques entre les termes de la requête et ceux figurant dans les documents jugés pertinents. Les termes fortement corrélés à ceux de la requête sont détectés, puis utilisés pour améliorer ou reformuler la requête, par ajout ou suppression de mots.

Exemple : une recherche sur "météo " pourrait être enrichie avec "pluie", "température", etc., si ces mots apparaissent fréquemment dans les résultats pertinents.

II.5.1.3 Approche de Biancalana et al. (2013)

— Type de profil utilisateur : Profil sémantique basé sur les termes souvent utilisés par l'utilisateur ou présents dans les documents qu'il consomme.
— Utilisation du profil : Les termes du profil et de la requête sont utilisés pour calculer des cooccurrences, permettant d'identifier des termes contextuellement proches pour enrichir la requête.

Principe : "analyse par cooccurrence" ,Cette méthode exploite des matrices de cooccurrence, qui mesurent la fréquence d'apparition conjointe de termes dans un même contexte documentaire. Elle permet ainsi d'identifier les termes contextuellement

proches de ceux présents dans la requête ou le profil utilisateur, en vue de l'enrichissement.

But : repérer et ajouter des termes fréquemment associés aux mots de la requête dans les documents.

II.5.1.4 Approche de Ould Amer et al. (2016) avec word2vec

- Type de profil utilisateur : Profil vectoriel sémantique, où les mots du profil sont représentés sous forme de vecteurs (embedding) grâce à word2vec.
- Utilisation du profil : Les vecteurs des termes de la requête et du profil sont comparés dans l'espace sémantique pour trouver des mots similaires, qui sont ensuite ajoutés à la requête.

Principe : "représentation vectorielle des mots", Cette approche fait appel à des plongements lexicaux (notamment word2vec) pour représenter les mots sous forme de vecteurs dans un espace sémantique. Elle permet d'identifier automatiquement des termes sémantiquement proches, c'est-à-dire ayant un sens voisin, en fonction de leur contexte d'usage.

Exemple : une requête sur "avion" pourrait être complétée par des termes comme "aéroport" , " vol " ou " décollage".

II.5.2 Reclassement des Résultats

Le reclassement des résultats vise à réorganiser les documents retournés par le moteur de recherche en fonction de leur pertinence vis-à-vis du profil utilisateur. Dans de nombreux cas, les utilisateurs ne consultent que les premiers résultats proposés. Par conséquent, il est crucial de classer les documents les plus pertinents en tête de la liste.

Le processus de reclassement repose généralement sur deux étapes :

- Calcul de la similarité : La première étape consiste à calculer un score de pertinence entre la requête de l'utilisateur et les documents retournés par le moteur de recherche classique.
- Reclassement des résultats : Dans la seconde étape, les documents sont réorganisés en fonction de leur similarité avec le profil utilisateur. Le reclassement peut être effectué en combinant les scores classiques avec des informations issues du profil utilisateur, comme le centre d'intérêts.

Certaines approches proposent de combiner les résultats de la recherche classique avec ceux issus d'une recherche personnalisée. Par exemple, Vallet et al. (2006)

proposent d'utiliser une mesure de similarité vectorielle basée sur le cosinus pour évaluer la pertinence d'un document par rapport au profil de l'utilisateur [30].

II.6 L'évaluation des systèmes de recherche d'information personnalisés

Divers travaux de recherche ont cherché à établir un cadre d'évaluation rigoureux adapté aux Systèmes de Recherche d'Information Personnalisés (SRIP). L'objectif principal de ces protocoles est de mesurer l'efficacité de l'apprentissage du profil utilisateur tant dans sa construction que dans son évolution ainsi que d'évaluer l'impact de l'intégration de ce profil dans le processus de recherche d'information.

Selon Zemirli [31], un protocole d'évaluation efficace doit répondre à deux exigences fondamentales :

— Valider l'approche de personnalisation, en mesurant la pertinence du profil utilisateur généré et la performance de la méthode de sa construction ;

— Comparer les performances du système avec et sans l'intégration du profil utilisateur, afin d'évaluer l'apport réel de la personnalisation dans le processus d'accès à l'information.

II.6.1 Les limites de l'évaluation classique

L'évaluation des systèmes de recherche d'information (SRI) classiques repose sur trois éléments clés [18] :

— Un ensemble de requêtes : représentant les besoins en information des utilisateurs.

— Une collection de documents : contenant les informations pertinentes pour répondre aux requêtes.

— Des jugements de pertinence : réalisés par des utilisateurs pour identifier les documents pertinents pour chaque requête.

Bien que cette approche ait prouvé son efficacité à travers des campagnes d'évaluation comme TREC et CLEF, elle présente des limites importantes :

— Absence de contexte : les tests ne prennent pas en compte le contexte de recherche ou la perception subjective de la pertinence par l'utilisateur.

— Diversité des besoins : les besoins en information sont traités comme des concepts statiques, sans tenir compte de leur dynamique ou de la diversité des centres d'intérêts des utilisateurs.

Ces limites rendent difficile l'évaluation des systèmes de recherche d'information personnalisée (RIP), où la pertinence est intrinsèquement liée à l'utilisateur et à son contexte.

II.6.2 La problématique de l'évaluation personnalisée

La pertinence dans un cadre de recherche personnalisée est subjective et dépend entièrement de l'utilisateur [32]. Cela pose un défi majeur pour l'évaluation, car il est nécessaire de disposer de :

— Un profil utilisateur : contenant des informations sur l'utilisateur, son historique de recherche et ses interactions avec le système.

— Une collection de documents et de requêtes : associée à des jugements de pertinence réalisés par les utilisateurs concernés.

Cependant, il n'existe pas de collections de tests standardisées pour la RIP, car la collecte de ces données est complexe et coûteuse en temps et en ressources [33].

II.6.3 Protocole d'évaluation personnalisée

En l'absence de standards, les chercheurs ont développé des protocoles spécifiques pour évaluer les systèmes de RIP. Voici quelques exemples :

- **Collecte des informations sur l'utilisateur :** Pour la collecte des données utilisateur en vue de construire les profils, différentes approches ont été adoptées.
  — Historique de bookmarking : utilisé par "bouhini2016 et saoudkechi2016" [34, 35] via le site social Delicious.
  — Fichiers logs de recherche : exploités par "Daoud2011" [36] sur le site commercial Exalead.
  — Données géographiques : utilisées par"MohammedAbdelmoty2017" [37] à partir des check-ins sur Foursquare.
  — Réseaux sociaux : utilisés par [38] via Google Scholar Citations et Facebook.
  — Interactions utilisateur-requête : proposées par "Buidguaguen2011" [39] pour simuler le profil utilisateur.
  — Questionnaires : utilisés par "Bouramoul2011" [40] pour interroger directement les utilisateurs.
- **Construction des datasets :** En l'absence de jugements de pertinence préétablis, les chercheurs ont recours à des solutions alternatives :

- Documents retournés par les moteurs de recherche : par exemple, [36] utilisent les 50 premiers résultats de Yahoo pour 50 requêtes.
- Recherche distribuée : [41] utilisent les 20 premiers résultats de 10 moteurs de recherche pour 90 requêtes.
- Combinaison de moteurs de recherche : [40] utilisent les 20 premiers résultats de Google, Yahoo et Bing pour 30 requêtes.

- **Évaluation des performances :** En l'absence de jugements de pertinence préétablis, les chercheurs font appel à des participants externes pour évaluer les résultats. Par exemple :
  - (5) participants dans "Daoud2011" [36] et "Duong2013" [38].
  - (6, 15 et 24) participants respectivement dans (Buidguaguen2011,SaoudKechid2016,Bouramoul2011) [39], [41] et [40].

Ces participants fournissent des post-jugements de pertinence, qui servent de référence pour évaluer les performances des systèmes de RIP.

II.6.4 Mesures d'évaluation des systèmes personnalisés

Plusieurs mesures sont utilisées pour évaluer les performances des systèmes de recherche d'information personnalisée (RIP). Voici les principales :

- **Précision à X (P@X) :**
  La précision à X mesure la proportion de documents pertinents parmi les X premiers résultats retournés par le système.
  Formule :

$$P@X = \frac{\text{Nombre de documents pertinents parmi les } X \text{ premiers}}{X}$$

  Exemple : Si 2 documents sur les 5 premiers sont pertinents, alors :

$$P@5 = \frac{2}{5} = 0.4$$

- **R-précision :**
  La R-précision mesure la proportion de documents pertinents parmi les R premiers résultats, où R est le nombre total de documents pertinents pour la requête.
  Formule :

$$R\text{-précision} = \frac{\text{Nombre de documents pertinents parmi les } R \text{ premiers}}{R}$$

Exemple : Si le 1er document pertinent est à la 3e position et le 2e à la 5e position, alors :

$$\text{1-précision} = \frac{1}{3} \approx 0.33 \quad \text{et} \quad \text{2-précision} = \frac{2}{5} = 0.4$$

- **Précision moyenne (MAP) :**

  La précision moyenne (Mean Average Precision) est la moyenne des précisions moyennes pour toutes les requêtes.

  Formule :

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q} \sum_{k=1}^{r_q} \text{Précision}(q)@k$$

  Où :

  — Q : Ensemble des requêtes.

  — r_q : Nombre de documents pertinents pour la requête q.

  — Précision$(q)@k$: Précision à la position k pour la requête q.

- **F-mesure :**

  La F-mesure est une moyenne harmonique pondérée de la précision et du rappel.

  Formule :

$$F\text{-mesure} = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

- **Pertinence relative (RR) :**

  La pertinence relative mesure la correspondance entre la pertinence algorithmique et la pertinence perçue par l'utilisateur.

  Formule (cosinus) :

$$RR(R_1, R_2) = \frac{\sum(R_1 \times R_2)}{\sqrt{\sum R_1^2} \times \sqrt{\sum R_2^2}}$$

  Formule (Jaccard) :

$$RR(R_1, R_2) = \frac{\sum(R_1 \times R_2)}{\sum R_1 + \sum R_2 - \sum(R_1 \times R_2)}$$

- **Gain cumulatif (CG) et Gain cumulatif actualisé (DCG)**

  Ces mesures estiment le gain de pertinence cumulé en fonction du rang des documents.

Formule pour CG :

$$CG[i] = \begin{cases} G[1] & \text{si } i = 1 \\ CG[i-1] + G[i] & \text{sinon} \end{cases}$$

Formule pour DCG :

$$DCG[i] = \begin{cases} G[1] & \text{si } i = 1 \\ DCG[i-1] + \frac{G[i]}{\log_2(i)} & \text{sinon} \end{cases}$$

Où $G[i]$ est le gain de pertinence au rang i.

Explication :

— CG mesure le gain total de pertinence sans tenir compte du rang.

— DCG atténue le gain de pertinence en fonction du rang, reflétant l'idée que les documents en haut de la liste sont plus importants.

II.7 Conclusion

Dans ce chapitre, nous avons exploré les limites des approches classiques de recherche d'information, qui ne parviennent pas à répondre aux besoins évolutifs et personnalisés des utilisateurs. Ces limites ont conduit à l'émergence de la recherche d'information personnalisée (RIP), qui place l'utilisateur et ses intérêts au cœur du processus de recherche.

Nous avons commencé par aborder la modélisation de l'utilisateur, un processus essentiel pour identifier, collecter, organiser et exploiter les informations qui caractérisent l'utilisateur et sa perception de la pertinence. Cette modélisation se concrétise à travers la construction d'un profil utilisateur, qui représente les centres d'intérêts, les préférences et le contexte de l'utilisateur.

Ensuite, nous avons détaillé les sources d'informations utilisées pour construire ce profil, ainsi que les méthodes d'acquisition (explicites et implicites) et les techniques de représentation des données (ensembliste, connexionniste, multidimensionnelle, etc.). Nous avons également souligné l'importance des différentes dimensions contextuelles, telles que la dimension thématique, sociale, mobile et sémantique, qui enrichissent le profil utilisateur.

Par ailleurs, nous avons examiné l'intégration du profil utilisateur dans le processus de recherche, notamment à travers la reformulation des requêtes et le reclassement

des résultats, afin d'adapter les réponses aux besoins spécifiques de l'utilisateur.

Enfin, nous avons abordé les défis liés à l'évaluation des systèmes de RIP, en mettant en avant les limites des méthodes classiques et en proposant des protocoles d'évaluation adaptés, ainsi que des mesures de performance spécifiques (P@X, MAP, F-mesure, DCG, etc.).

En résumé, la recherche d'information personnalisée représente une avancée majeure pour répondre aux besoins dynamiques et subjectifs des utilisateurs. Cependant, elle nécessite des approches innovantes pour la modélisation de l'utilisateur, l'intégration du contexte et l'évaluation des performances, afin de garantir des résultats pertinents et adaptés à chaque utilisateur.

Chapitre III

CONCEPTION D'UNE BANQUE DE BONNES REQUÊTES COMME OUTIL D'AIDE A LA REFORMULATION DE NOUVELLES REQUÊTES DE RECHERCHE D'INFORMATION

III.1 Introduction

La recherche d'information est un processus itératif où les utilisateurs expriment leur besoin informationnel à travers des requêtes successives. Toutefois, il est fréquent que les requêtes initiales soient imprécises, incomplètes ou mal formulées, ce qui limite l'efficacité des systèmes de recherche. Afin d'améliorer l'adéquation entre les requêtes et les résultats retournés, il est essentiel de proposer des mécanismes de reformulation adaptés aux utilisateurs et à leurs habitudes de recherche. Dans ce contexte, nous proposons la conception d'une banque de bonnes requêtes permettant de capitaliser les interactions passées des utilisateurs avec le système, dans le but de leur suggérer des reformulations pertinentes.

Ce chapitre décrit l'architecture et les composants de cette banque de bonnes requêtes, à travers la modélisation du profil d'utilisateur, l'analyse des requêtes et des termes utilisés, l'identification des requêtes efficaces, ainsi que la mise en relation avec des utilisateurs similaires (appelés " amis ").

L'objectif est de fournir un support décisionnel pour la reformulation automatique ou assistée des requêtes, en s'appuyant sur des métriques de pondération et de similarité entre requêtes et profils d'utilisateurs. Cette banque de bonnes requêtes constitue ainsi un outil pour orienter les utilisateurs vers des requêtes plus efficaces, tout en enrichissant dynamiquement leur profil à mesure de leurs interactions avec le système.

III.2 Profil d'utilisateur

Le profil d'utilisateur est une représentation structurée des besoins informationnels d'un individu, construite à partir de ses requêtes antérieures jugées pertinentes. Il combine des informations thématiques (termes fréquemment utilisés) et sociales (utilisateurs similaire) afin de mieux cerner ses préférences et d'adapter les réponses à ses intentions de recherche [42]. Il est structuré en deux composantes principales : les données statiques et les données dynamiques.

III.2.1 Données Statiques

Les données statiques regroupent les informations contextuelles et associées à chaque utilisateur, collectées lors de la création du compte utilisateur. Elles incluent :

— Identifiant : Un identifiant unique permettant de distinguer chaque utilisateur dans le système ;

— Informations personnelles : Nom, prénom, âge, sexe, langue préférée..... ;

— Utilisateur ami : Des liens avec d'autres utilisateurs, définis comme des relations d'amitié, qui permettent d'explorer les requêtes des amies.

Ces données sont stockées dans une table relationnelle, comme illustré dans le tableau III.1. Selon Elmasri [43], une structuration claire des données statiques est essentielle pour une gestion efficace des profils.

Attribut	**Type**
`id_utilisteur`	integer
`Nom`	string
`Prénom`	string
`nom d'utilisateur`	string
`Mot de passe`	string
`E-mail`	string
`Age`	integer
`Sexe`	string
`Adresse`	string
`Langue`	string

Table III.1 — Table utilisateur

III.2.2 Donnes dynamiques

Contrairement aux données statiques, les données dynamiques évoluent avec le temps et traduisent les interactions continues de l'utilisateur avec le système. Elles englobent principalement les requêtes formulées et les termes fréquemment utilisés, reflétant ainsi les besoins informationnels actualisés de l'utilisateur.

III.2.2.1 Requêtes

Nous considérons qu'une requête est jugée "bonne" ou efficace lorsque au moins l'un des cinq premiers résultats retournés par le système répond de manière satisfaisante aux besoins de l'utilisateur. Cette satisfaction peut être déduite de comportements implicites observés chez l'utilisateur, tels que le partage de l'article, l'impression du contenu, l'enregistrement du fichier, ou toute autre forme d'interaction positive. Ces signaux d'interaction sont interprétés comme des indicateurs de la pertinence du résultat, et par extension, de la qualité de la requête soumise. Les requêtes jugées pertinentes sont alors conservées dans la banque de bonnes requêtes, et viennent

enrichir la partie dynamique du profil utilisateur. Chaque requête est associée à un ensemble de caractéristiques clés suivantes :

— Identifiant de la requête : Un identifiant unique permettant de distinguer chaque requête dans le système.

— Contenu textuel : Le texte brut de la requête, qui exprime le besoin d'information de l'utilisateur

— Métadonnées : Des informations complémentaires, telles que la date et l'heure de soumission, permettant d'analyser l'évolution des besoins informationnels dans le temps et d'identifier les contextes temporels susceptibles d'influencer le comportement de recherche de l'utilisateur.

Les requêtes sont collectées et stockées dans la banque de bonnes requêtes, comme illustré dans le tableau III.2 :

Attribut	**Type**
`id_requete`	integer
`texte`	string ou text
`date`	datetime ou timestamp

Table III.2 — Structure simplifiée d'une requête dans le profil utilisateur

III.2.2.2 Termes et leur pondération

représentent les mots ou expressions clés traduisant les besoins informationnels de l'utilisateur, et servent de base à la comparaison ainsi qu'à la reformulation des requêtes. En décomposant les requêtes en termes, le système identifie les éléments significatifs du profil de recherche, facilitant ainsi la personnalisation des proposée.

Ces termes sont organisés et enregistrés dans une structure appelée Tableau Terme, qui recense chaque terme unique accompagné de ses caractéristiques, comme illustré dans le tableau III.3.

Attribut	**type**
`id_terme`	integer
`terme`	string ou text

Table III.3 — Structure simplifiée d'un tableaux terme dans le profil utilisateur

Afin de quantifier l'importance de chaque terme dans l'ensemble des requêtes d'un utilisateur, une pondération est attribuée à chaque entrée. cette pondération $w(t)$

d'un terme t est calculée selon la formule :

$$\mathrm{w}(t) = \frac{freq(t)}{|Q|}$$

où :

— $freq(t)$ est la fréquence du terme t dans toutes les requêtes du profil ;

— Q est le nombre totale de requêtes.

Cette méthode permet de mesurer l'importance relative d'un terme au sein du profil, contribuant ainsi à une reformulation plus ciblée et plus pertinente des requêtes futures.

III.3 Amis d'utilisateur

Le concept d'ami désigne des utilisateurs avec des intérêts similaires dont les requêtes des uns peuvent être reformulées à partir des profils des autres.

III.3.1 Mesure de similarité entre utilisateurs

La similarité entre deux utilisateurs (U_i et U_j) est déterminée par le calcul de la similarité entre les deux ensembles de termes de chacun, qui ont été précédemment stockés dans ces banques de requêtes. Cette dernière est définie par la formule suivante :

$$sim(U_i, U_j) = \frac{|T_i \cap T_j|}{|T_i|}$$

Où

— T est l'ensemble des termes ;

— T_i est l'ensemble des termes de U_i ;

— T_j est l'ensemble des termes de U_j.

Cette mesure s'inspire partiellement du coefficient de Jaccard [44] qui évalue la similarité entre deux ensembles.

Cette mesure de similarité est utilisée exclusivement pour trier les amis de l'utilisateur du plus similaire au moins similaire. Par la suite on ne prend en considération qu'un nombre limité des amis les plus similaire pour le processus de reformulation des requêtes.

III.4 Processus de reformulation

Le diagramme ci-dessous [III.1] illustre le processus de reformulation des requêtes dans le cadre d'un système d'aide à la recherche d'information. Le processus commence par la soumission d'une requête originale, suivie par une phase de recherche des requêtes similaires à travers l'exploration du "profil personnel de l'utilisateur" et des "profils de ses amis".

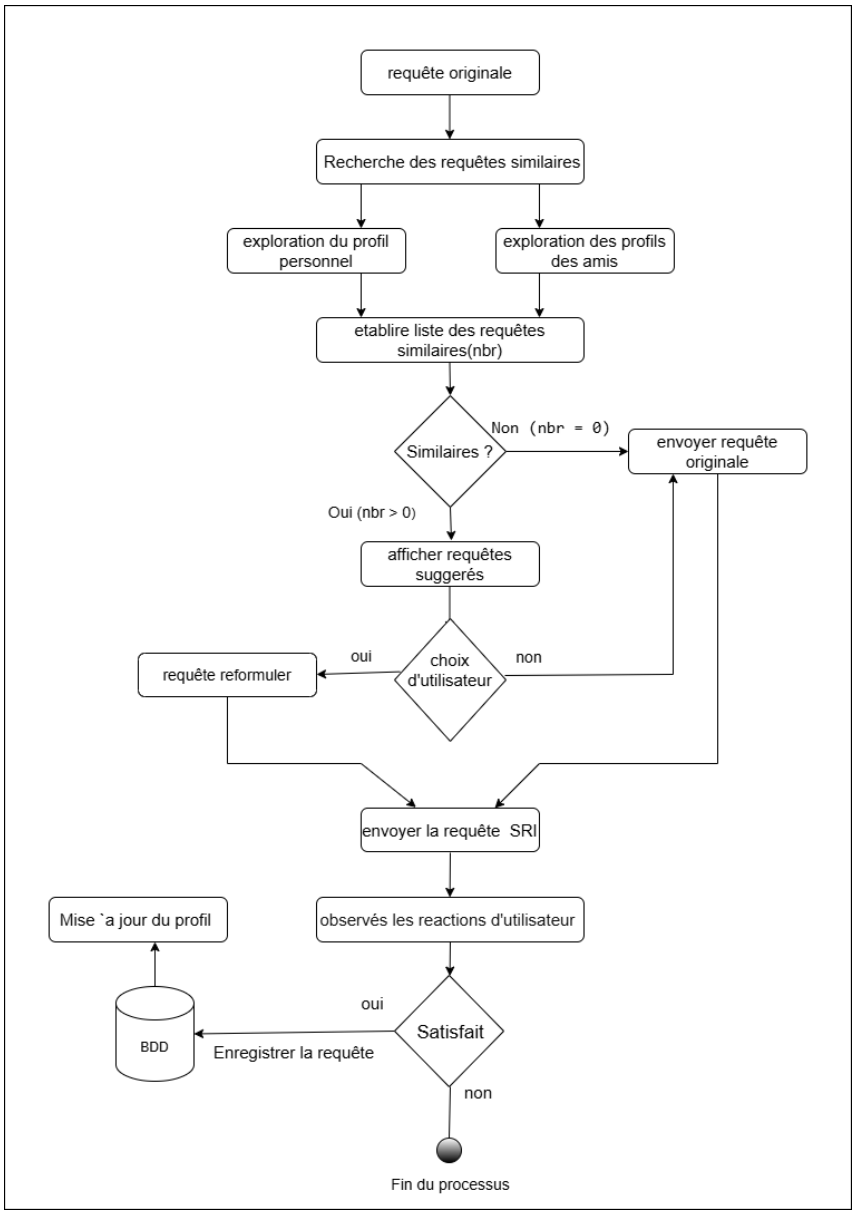


Figure III.1 — Schéma conceptuel du processus de reformulation de requête

À cette étape, le système établit une liste de requêtes similaires, dont le nombre est représenté par le variable nbr. Si nbr = 0, cela signifie qu'aucune requête similaire n'a été trouvée, et la requête originale est envoyée directement au système de recherche d'information (SRI). En revanche, si nbr > 0, c'est-à-dire qu'il existe une ou plusieurs

requêtes similaires, celles-ci sont affichées à l'utilisateur sous forme de suggestions.

L'utilisateur peut alors choisir d'utiliser l'une des requêtes proposées pour reformuler sa demande ou bien conserver la requête initiale. Une fois la requête envoyée, le système voir les réactions d'utilisateur afin d'évaluer sa satisfaction.

Si le système détecte que l'utilisateur est satisfait, la requête est enregistrée dans la banque de bonnes requêtes (`BDD`) et sert à enrichir son profil. En cas d'insatisfaction, le processus se termine sans mise à jour.

III.4.1 Similarité entre requêtes

Pour mesurer la similarité entre deux requêtes Q_1 et Q_2, nous utilisons une formule qui prend en compte à la fois les termes communs et la proximité de leurs positions respectives dans chaque requête. Cette méthode permet de capturer à la fois le contenu lexical et la structure syntaxique des requêtes, en s'appuyant sur la formule suivante :

$$\mathrm{Sim}(Q_1, Q_2) = \frac{\sum_{t_i \in Q_1 \cap Q_2} \frac{1}{1+|\mathrm{pos}_{Q_1}(t_i) - \mathrm{pos}_{Q_2}(t_i)|}}{|Q_1| \cup |Q_2|}$$

où :

- t_i est le terme commun entre les deux requêtes ;
- $\mathrm{pos}_{Q_k}(t_i)$ désigne la position du terme t_i dans la requête Q_k $(k \in \{1, 2\})$;
- $|Q|$ indique le nombre total de termes dans la requête Q.

Plus les termes communs apparaissent à des positions similaires, plus la similarité entre les requêtes est élevée.

Cette approche s'inscrit dans la lignée des travaux sur la similarité sémantique positionnelle, tels que ceux de Mihalcea et Tarau [45] dans TextRank, ou encore Bouchoucha [46] pour la reformulation de requêtes en RI, qui intègrent des informations de position pour améliorer la précision des comparaisons.

III.4.2 Reformulation de requête

Dans le cadre du processus de reformulation de requêtes, la première étape intervient après la soumission d'une requête initiale par l'utilisateur. À ce stade, le système procède à la recherche des **requêtes similaires**, en explorant deux sources principales :

- le profil personnel de l'utilisateur, qui regroupe l'historique de ses propres requêtes bonnes ;

— les profils de ses amis, c'est-à-dire les utilisateurs similaires identifiés à l'aide de la fonction de similarité entre utilisateurs.

Les requêtes extraites sont ensuite comparées à la requête initiale à l'aide de la mesure de similarité requêtes, qui intègre la pondération des termes supplémentaire. Cette comparaison permet de sélectionner les requêtes les plus pertinentes et plus similaire au nouvelle requête pour la reformulation.

Au total, six requêtes suggérés sont proposées à l'utilisateur :

— deux issues de son propre profil ;

— quatre extraites des profils de ses amis .

Ce choix vise à établir un équilibre entre la pertinence personnalisée et la diversité des suggestions, deux axes complémentaires pour enrichir le processus de reformulation.

En cas d'égalité de similarité entre plusieurs requêtes candidates, le classement entre eux est fait sur la base des poids des termes supplémentaires dans le profil d'utilisateur par rapport à la requête originale,si ces deriniers aussi sont egaux on passe a une autre critere qu'est la date de modification

Enfin, l'utilisateur conserve la liberté de choix il peut adopter l'une des requêtes proposées, ou continuer la recherche avec sa requête initiale.

III.4.3 Mise à jour du profil utilisateur

Une requête est enregistrée si l'utilisateur montre de l'intérêt pour l'une des suggestions, à condition qu'elle provienne du profil d'un de ses amis (parmi les quatre proposées). En revanche, si elle vient de son propre profil (parmi les deux proposées), elle n'est pas sauvegardée une deuxième fois pour éviter les doublons.

Le profil de l'utilisateur est mis à jour à chaque fois qu'une nouvelle requête est enregistrée, qu'il s'agisse d'une suggestion ou de la requête initiale.

Lors de cet enregistrement, les poids des termes du profil sont recalculés, et la liste des amis (utilisateurs similaires) est réajustée en fonction des nouvelles données. Ce processus permet au système de s'adapter progressivement aux préférences de l'utilisateur.

Le profil de l'utilisateur est mis à jour à chaque enregistrement d'une nouvelle requête dans sa banque de bonnes requêtes, soit l'une des requêtes proposées ou la requête initiale.

Les poids des termes dans le profil de l'utilisateur sont recalculés à chaque enregistrement d'une nouvelle requête pertinente, et la liste de ses amis (utilisateurs

similaires) est ajustée en fonction des nouvelles données. Ce mécanisme permet au système de s'adapter progressivement aux préférences évolutives de l'utilisateur.

III.5 Conception de la base de donnes

Afin d'assurer une structuration cohérente et exploitable des données utilisateurs qu'elles soient statiques ou dynamiques une base de données relationnelle a été conçue. Cette dernière repose sur un ensemble Des tables reliées entre elles représentant les différentes entités du système, telles que l'utilisateur, les requêtes, les termes.

La figure III.2 illustre le modèle conceptuel de données adopté pour la représentation de notre système .

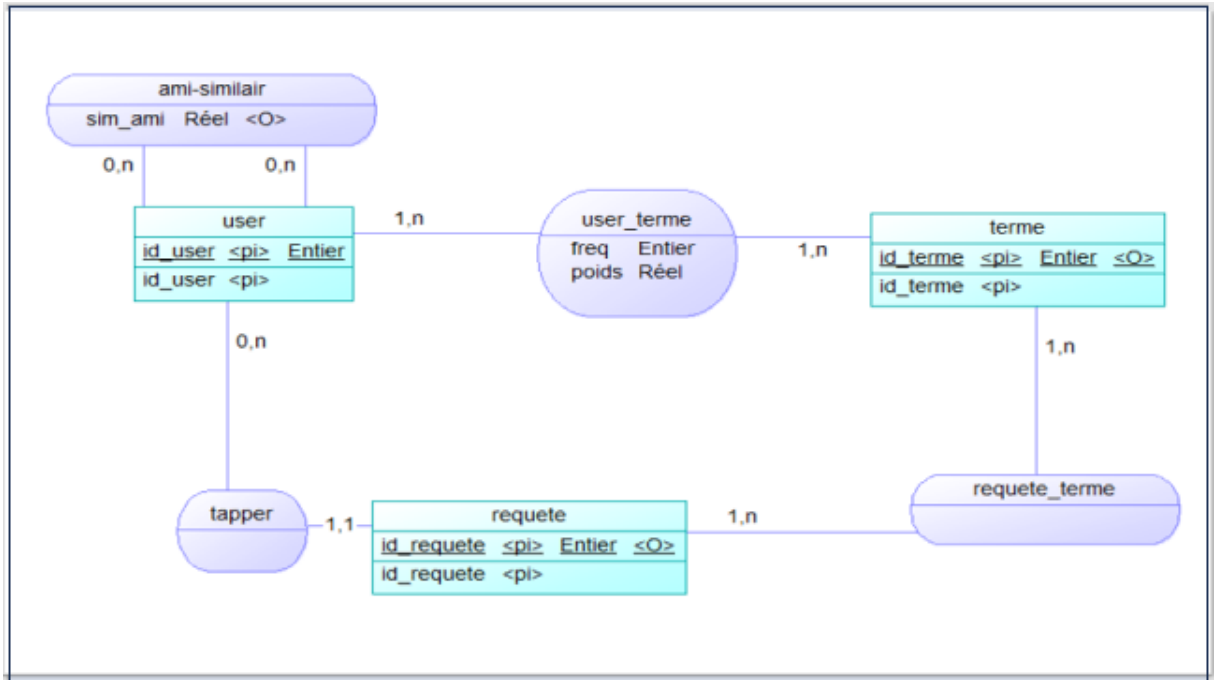


Figure III.2 — le modèle conceptuel de données

Elle met en évidence les principales relations entre les entités :

- Il y a deux relations de type "plusieurs à plusieurs " : l'une entre utilisateur et terme, et l'autre entre requête et terme ;
- Chaque utilisateur peut avoir plusieurs requêtes, mais une requête ne correspond qu'à un seul utilisateur ;
- De même, on a une relation récursive sur la table " user " qui peut répertorier les amis.

Ce modèle relationnel permet une structuration cohérente des données, facilitant

ainsi la personnalisation des recherches à travers l'analyse des termes fréquemment utilisés par l'utilisateur et par des profils similaires.

III.6 Réalisation du Système

Dans cette section , nous allons mettre en pratique la conception et la modélisation de notre solution pour entamer sa réalisation. Pour ce faire, nous présenterons les outils de développement utilisés ainsi que la méthodologie adoptée pour le développement. L'objectif principal est de produire une solution fonctionnelle et exploitable. Nous allons détailler également l'architecture du projet Django.

III.6.1 Outils de Développement

Les outils suivants ont été sélectionnés pour leur compatibilité avec les besoins du projet :

- Visual Studio Code (VS Code) : Éditeur léger et puissant, utilisé pour écrire et déboguer le code. Extensions installées : Python, Django, HTML/CSS/JS ;
- Django : est un Framework web open-source écrit en Python. Il est conçu pour faciliter le développement rapide de sites web sûrs, maintenables et évolutifs, suivant le modèle MVC ;
- Python : est un langage de programmation interprété, choisi pour sa simplicité et sa compatibilité avec Django ;
- MySQL : Système de Gestion des Base de données pour stocker les données ;
- HTML : Standards pour structurer les pages web ;
- java script :
- CSS : est un langage de feuille de style utilisé pour décrire la présentation et la mise en forme d'un document écrit en HTML ;

III.6.2 Architecture de projet django

L'architecture de l'application repose sur le framework Django, qui adopte une structure MVT (Modèle-Vue-Template), dérivée du modèle MVC (Modèle-Vue-Contrôleur). Cette architecture repose sur une séparation stricte des responsabilités entre les composants : le Modèle gère les données et la logique métier, la Vue (dans Django, c'est le contrôleur logique) traite les requêtes et la logique de l'application, tandis que le Template constitue la couche de présentation (les interfaces grafiques).

Cette organisation favorise une meilleure maintenabilité du code, une évolutivité naturelle de l'application et une structuration claire des traitements (voir figure III.3).

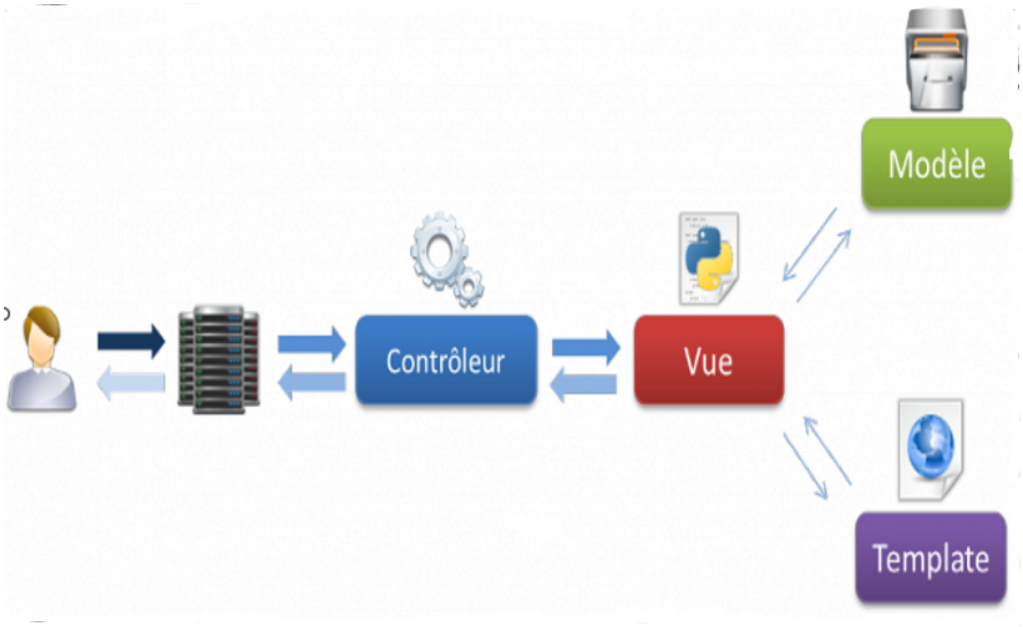


Figure III.3 — Architecture d'un projet Django

Dans cette optique, la structure générale du projet a été conçue de manière à refléter ces principes fondamentaux. L'arborescence du projet Django est organisée en modules distincts, chacun jouant un rôle bien défini : les dossiers models.py, views.py, templates/ et urls.py sont répartis selon les conventions de Django, assurant ainsi une cohérence entre la structure physique du projet et l'architecture logique.

La figure III.4 présente un schéma global de la structure de notre projet nomme (QRS), en illustrant les relations entre les composants et en montrant comment le projet s'appuie sur le modèle MVT pour gérer les requêtes, interagir avec la base de données, et générer dynamiquement les interfaces destinées à l'utilisateur.

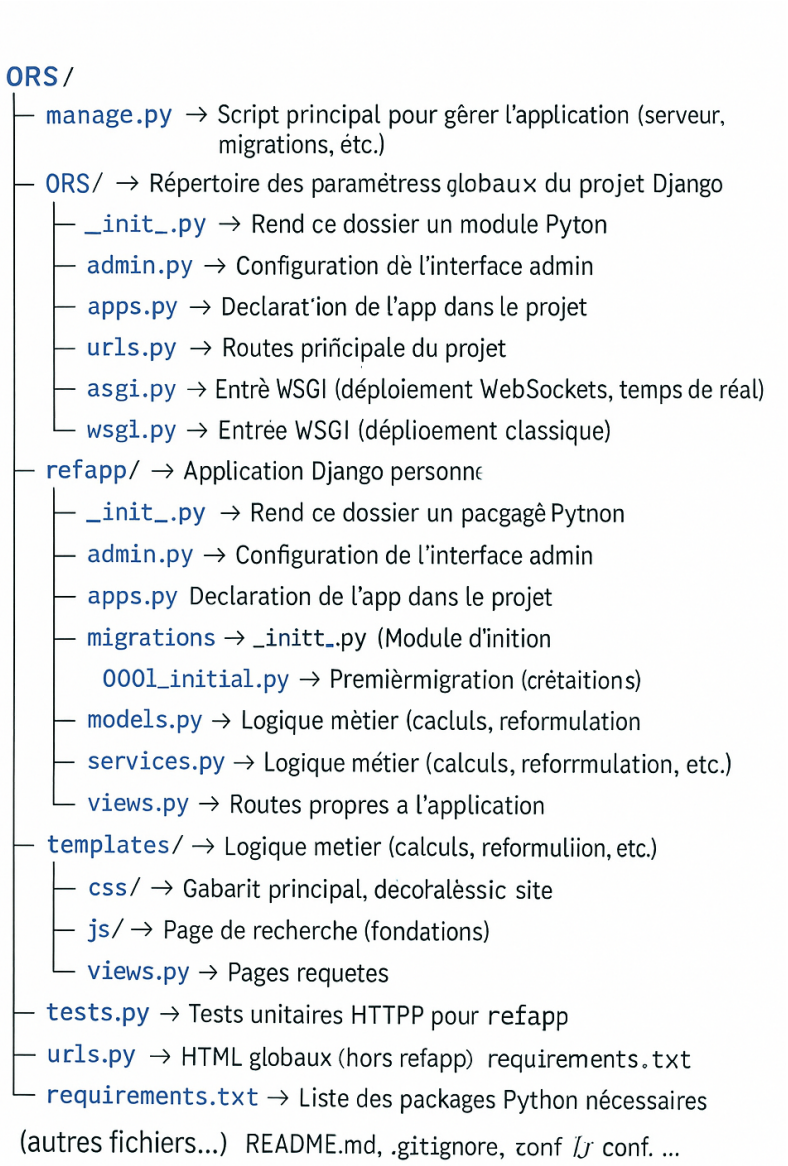


Figure III.4 — Structure Générale du Projet

III.6.3 Présentation de L'application

Le développement, la maîtrise, la cohérence et l'efficacité du fonctionnement d'une conception se montre par sa réalisation. Pour cela, En utilisant les outils cités précédemment, nous avons abouti à l'application suivante :

III.6.3.1 Fenêtre d'authentification de l'application

C'est une fenêtre protégée par mot de passe à travers laquelle on accède à la fenêtre principale (voir figure III.5) :

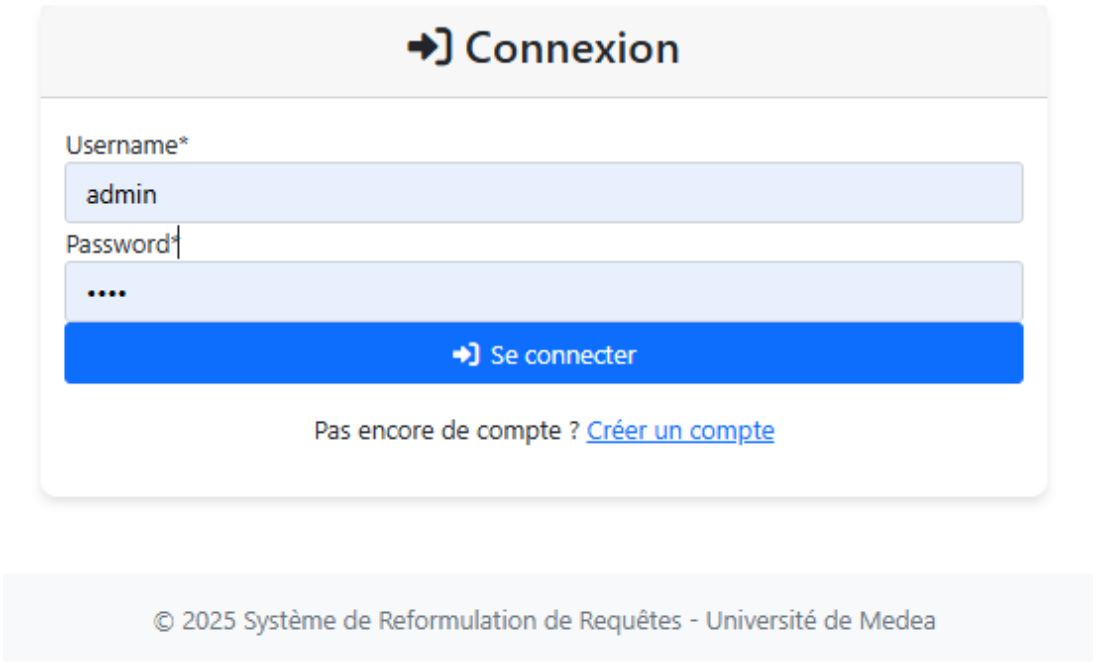


Figure III.5 — Login de l'application

III.6.3.2 Fenêtre d'Accueil

Une fois connecté, l'utilisateur est redirigé vers son tableau de bord personnel, comme illustré dans la figure III.6. Cette interface centrale lui permet de visualiser rapidement l'état de son activité et d'accéder à des fonctionnalités essentielles. Elle se compose principalement des éléments suivants :

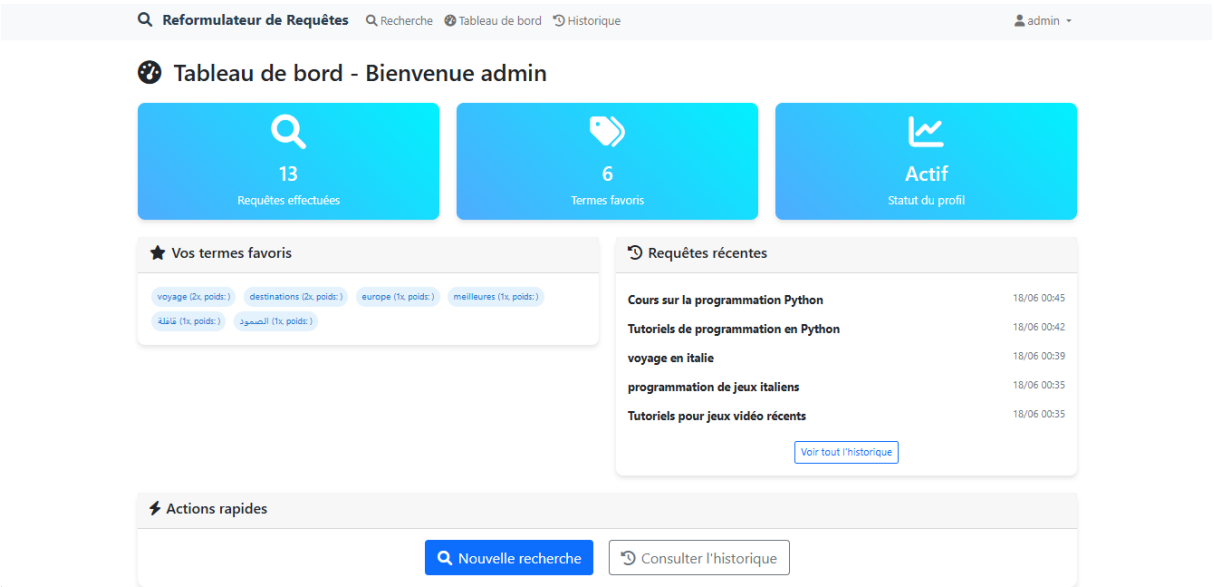


Figure III.6 — Fenêtre d'Accueil

— Indicateurs clés : Trois cartes en haut de la page présentent des statistiques synthétiques : le nombre total de requêtes effectuées, le nombre de termes favoris identifiés dans le profil, et le statut actuel du profil (actif/inactif).

— Termes favoris : Cette section affiche les mots-clés les plus représentatifs de l'utilisateur, accompagnés d'indications sur leur fréquence et leur poids, ce qui reflète l'importance de chaque terme dans son profil personnalisé.

— Requêtes récentes : Une liste chronologique des dernières requêtes soumises permet à l'utilisateur de suivre son historique de recherche de manière concise et structurée.

— Actions rapides : En bas de la page, deux boutons donnent un accès immédiat aux fonctionnalités de recherche : soumettre une nouvelle requête ou consulter l'historique complet.

Ce tableau de bord a été conçu pour offrir une vue d'ensemble intuitive et dynamique, facilitant la gestion du profil utilisateur et l'accès aux fonctions du système.

III.6.3.3 Page de recherche et reformulation

Lorsque l'utilisateur saisit sa requête et lance une recherche, le système propose automatiquement des reformulations. Comme il est mentionné dans la figure III.7.

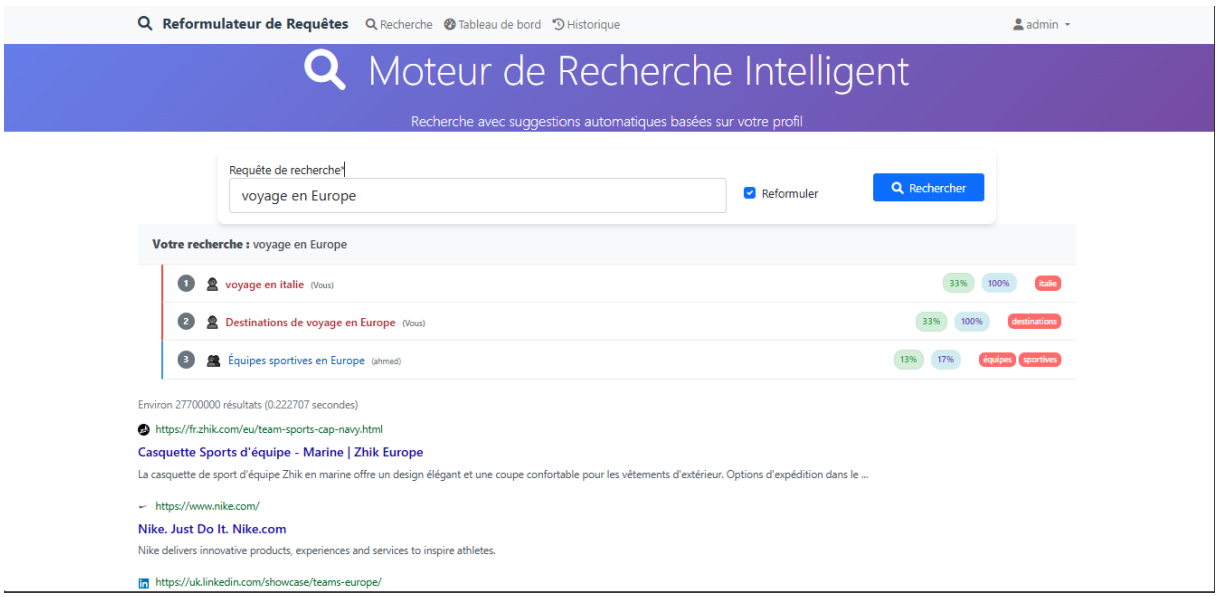


Figure III.7 — Page de recherche et reformulation

— La case verte montre le taux de similarité entre la requête originale et la suggestion ;

— La case bleue indique la source : 100% signifie que la suggestion vient du profil personnel, sinon elle provient d'un ami ;

— Les mots en rouge sont les termes ajoutés par rapport à la requête d'origine

III.6.3.4 Fenêtre de confirmation de sauvegarde de requête

Lorsque l'utilisateur sélectionne une suggestion et lance une recherche, une fenêtre contextuelle s'affiche pour confirmer la sauvegarde de la requête (voir figure III.8).

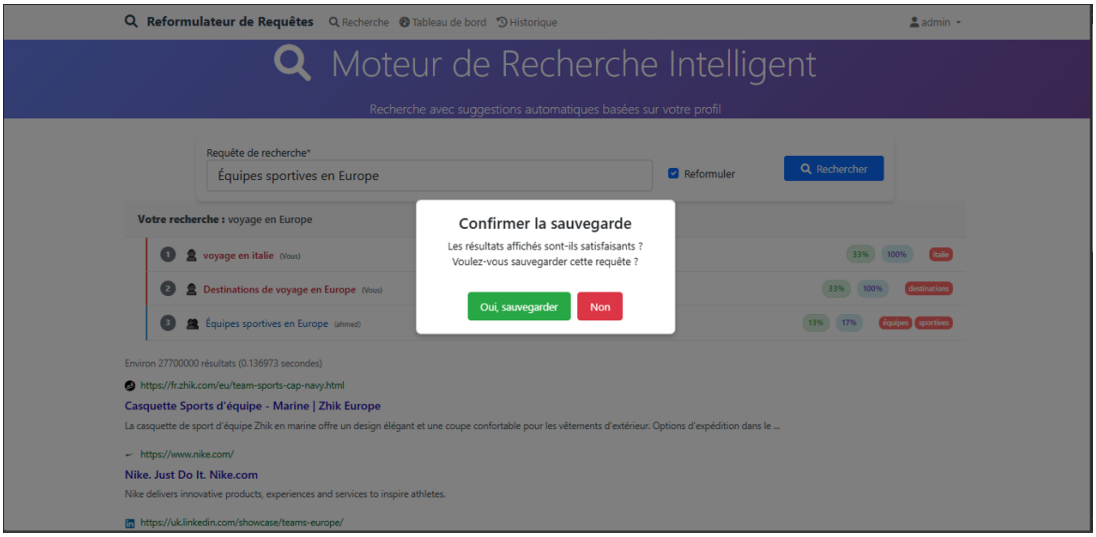


Figure III.8 — Fenêtre de confirmation de sauvegarde de requête

Cette boîte de dialogue demande si les résultats affichés sont satisfaisants et propose de sauvegarder la requête via deux boutons clairs :

— Oui, sauvegarder (en vert) pour enregistrer la requête dans le profil utilisateur ;

— Non (en rouge) pour ignorer l'enregistrement.

III.7 Conclusion

Dans ce chapitre, nous avons présenté l'architecture conceptuelle et la modélisation de notre système de reformulation de requêtes. À travers l'analyse du profil utilisateur, constitué de données statiques et dynamiques, nous avons pu établir une représentation fine des besoins informationnels. La prise en compte des relations sociales via la notion d'amis permet d'enrichir les suggestions proposées, en s'appuyant sur des profils similaires.

Nous avons défini les métriques de similarité entre requêtes et entre utilisateurs, ainsi que les mécanismes de pondération des termes, éléments essentiels pour assurer la pertinence des reformulations suggérées. Le processus de reformulation a été détaillé, depuis l'émission d'une requête jusqu'à la mise à jour du profil utilisateur, en intégrant les retours implicites liés à la satisfaction.

Enfin, nous avons concrétisé cette approche à travers la modélisation relationnelle de la base de données et la réalisation d'une application fonctionnelle basée sur le framework Django, illustrant la faisabilité et l'opérabilité de notre solution. Le chapitre suivant sera consacré à l'évaluation expérimentale du système et à l'analyse des résultats obtenus.

Conclusion générale et perspectives

Ce mémoire a exploré les enjeux de la recherche d'information (RI) dans un contexte marqué par des données numériques, en mettant en avant les limites des approches classiques et l'émergence de la recherche d'information personnalisée (RIP). À travers une analyse approfondie, nous avons d'abord établi les fondements théoriques de la RI classique, en détaillant ses modèles et ses méthodes d'évaluation. Ces bases ont permis de souligner les insuffisances des systèmes traditionnels, notamment leur incapacité à intégrer le contexte et les préférences individuelles des utilisateurs.

Le deuxième chapitre a mis en lumière l'importance de la RIP, qui place l'utilisateur au centre du processus en s'appuyant sur des profils dynamiques et des dimensions contextuelles. Cette approche permet de répondre de manière plus précise aux besoins spécifiques, tout en posant des défis en termes de modélisation et d'évaluation.

Le troisième chapitre constitue le cœur de notre contribution, avec la conception d'une banque de données de requêtes comme outil d'aide à la reformulation. En combinant des données statiques et dynamiques des profils utilisateurs, des métriques de similarité (entre utilisateurs et requêtes), et des relations sociales via la notion d'amis, ce système propose des reformulations personnalisées et pertinentes. L'implémentation de cette solution dans une application Django, soutenue par une base de données relationnelle, a démontré sa viabilité pratique

Perspectives

À la lumière de ces résultats obtenus, plusieurs perspectives de recherche future s'ouvrent pour approfondir cette :

— Amélioration des métriques de similarité : Intégrer des techniques d'apprentissage automatique, comme les modèles d'embeddings sémantiques (BERT, par exemple), pour capturer des relations plus complexes entre les termes et les requêtes ;

— Évaluation à grande échelle : Tester le système sur une collection de données

réelle et diversifiée, avec un plus grand nombre d'utilisateurs, pour valider sa robustesse et son efficacité dans des scénarios variés ;

— Prise en compte du contexte multimodal : Incorporer des données contextuelles sup plémentaires, comme les images ou les vidéos, pour enrichir les profils utilisateurs et les suggestions de reformulation ;

— Protection des données : Renforcer les mécanismes de confidentialité pour garantir la sécurité des profils utilisateurs, notamment en anonymisant les données dynamiques col lectées ;

— Interaction utilisateur avancée : Développer des interfaces plus intuitives, avec des visualisations interactives des suggestions, pour améliorer l'expérience utilisateur

Ce projet ouvre la voie à des systèmes de recherche d'information toujours plus adaptatifs et centrés sur l'utilisateur, contribuant à l'évolution des technologies de recherche dans un monde numérique en constante transformation.

Bibliographie

[1] M. K. Buckland, "What is a "document"?" *Journal of the American society for information science*, vol. 48, no. 9, pp. 804–809, 1997.

[2] K. Ingwersen, Peter andaoud Järvelin, *The turn : Integration of information seeking andaoud retrieval in context.* Springer Science & Business Media, 2005, vol. 18.

[3] T. Rafa, "Combinaison d'un profil social et un profil situationnel pour la recherche d'information," Thèse de Doctorat en Sciences, Université des Sciences et de la Technologie Houari Boumediene (USTHB), Janvier 2023, spécialité : Intelligence Artificielle.

[4] C. C. Kuhlthau, "Inside the search process : Information seeking from the user's perspective," *Journal of the American society for information science*, vol. 42, no. 5, pp. 361–371, 1991.

[5] M. Daoud, "Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche," Ph.D. dissertation, Université Paul Sabatier-Toulouse III, 2009.

[6] Z. Saoud, "Exploitation des réseaux sociaux pour la recherche d'information," Ph.D. dissertation, Faculté d'Electronique et d'Informatique, 2017.

[7] I. Ruthven, "Re-examining the potential effectiveness of interactive query expansion," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 213–220.

[8] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval : The Concepts and Technology behind Search*, 2nd ed. Addison-Wesley, 2011.

[9] D. A. Grossman and O. Frieder, *Information Retrieval : Algorithms and Heuristics*, 2nd ed. Springer, 2004.

[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[11] C. D. Manning, P. Raghavan, and H. Schütze, "Boolean retrieval," *Introduction to information retrieval*, pp. 1–18, 2008.

[12] J. J. Rocchio, "Relevance feedback in information retrieval," *The SMART Retrieval System : Experiments in Automatic Document Processing*, pp. 313–323, 1971.

[13] E. M. Voorhees, "Query expansion using lexical-semantic relations," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 61–69, 1994.

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463, no. 1999.

[15] G. Salton, *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971.

[16] F. Douma and F. Ramdani, "Reformulation interactive personnalisée de la requête utilisateur pour améliorer la pertinence des résultats de recherche d'information," Mémoire de Master, Université Yahia Fares de Médéa, Algérie, 2024, spécialité : Ingénierie des Systèmes et Technologies du Web.

[17] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM (J. ACM)*, vol. 7, no. 3, pp. 216–244, 1960.

[18] C. Bouhini, "Impact des réseaux sociaux sur le processus de recherche d'information," Ph.D. dissertation, Université Jean Monnet de Saint-Étienne, Octobre 2014, soutenue le 21 Octobre 2014, Discipline : Informatique.

[19] W. B. Croft and D. J. Harper, "Using probabilistic models of document retrieval without relevance information," *Journal of Documentation*, vol. 35, no. 4, pp. 285–295, 1979.

[20] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," *SIGIR Forum*, vol. 28, no. 2, pp. 232–241, 1994.

[21] B. Moulahi, "Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information," Ph.D. dissertation, Université Toulouse III Paul Sabatier, 2015.

[22] O. Ressad-Bouidghaghen, "Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes," Ph.D. dissertation, Université Toulouse III - Paul Sabatier, Toulouse, France, December 2011.

[23] S. Kechid, "Modélisation des profils d'utilisateur pour la recherche d'information personnalisée," Ph.D. dissertation, Université [à compléter], 2009, [Source fictive, à remplacer par une référence réelle si disponible].

[24] M. Tchuente, N. Bensaid, and S. Chabridon, "Une approche de personnalisation de la recherche d'information basée sur le profil utilisateur et les contextes de recherche," *Revue des Nouvelles Technologies de l'Information (RNTI)*, vol. E-24, pp. 95–110, 2013.

[25] S. Krir, *Personnalisation de la Recherche d'Information : Profil utilisateur, filtrage et recommandation.* Éditions universitaires européennes, 2012.

[26] H. Gao and J. Zhang, "Personalized search : From web search to context-aware search," *International Journal of Computer Applications*, vol. 118, no. 24, pp. 39–45, 2015.

[27] Y. Zhang and X. Ma, "Contextualization of web search : Challenges and applications," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–31, 2014.

[28] S. Chakrabarti, *Mining the Web : Discovering Knowledge from Hypertext Data.* Morgan Kaufmann, 2017.

[29] D. Jannach and G. Adomavicius, *Recommendation Systems : Challenges and Opportunities.* Springer, 2016.

[30] D. Vallet, M. Fernández, and P. Castells, "Re-ranking search results with user profile information," in *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, ser. Lecture Notes in Computer Science, vol. 3936. Springer, 2006, pp. 545–548.

[31] N. W. Zemirli, "Personnalisation de la recherche d'information : modèles de profil utilisateur et adaptation des résultats," Thèse de doctorat, Université Abou Bekr Belkaid - Tlemcen, 2008.

[32] L. Tamine, "User profiling in information retrieval : a survey," *Information Processing & Management*, vol. 54, no. 6, p. 102–123, 2018.

[33] L. Ould Amer, "Personnalisation de la recherche d'information : Modélisation du profil utilisateur et recommandations," *Revue des Technologies de l'Information*, vol. 12, no. 1, pp. 45–58, 2020.

[34] D. Bouhini and C. Khentout, "Un modèle de profil utilisateur basé sur l'analyse de la navigation sociale," *Revue des Nouvelles Technologies de l'Information*, vol. B-10, pp. 183–196, 2016.

[35] B. Saoud Kechid and C. Khentout, "Personnalisation de la recherche d'information basée sur les intérêts de l'utilisateur," *Revue de l'Information Scientifique*

et Technique, vol. 23, pp. 25–34, 2016.

[36] M. Daoud, L. Tamine, and M. Boughanem, "A session-based personalized search using an ontological user profile," in *Proceedings of the 2011 ACM Symposium on Applied Computing*. ACM, 2011, pp. 1143–1148.

[37] A. Abdelmoty and C. B. Jones, "Personalised spatial web search with location histories," in *International Conference on Geographic Information Science*. Springer, 2017, pp. 1–14.

[38] A. Duong, B. Pham, and L. Cao, "User profiling based on social network information," in *2013 International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2013, pp. 167–172.

[39] B. Buidguaguen and G. Belalem, "Profilage d'utilisateur dans un système de recommandation d'informations basé sur la sémantique," in *Colloque International sur l'Optimisation et la Recherche Opérationnelle*, 2011.

[40] A. Bouramoul, "Recherche d'information contextuelle et sémantique sur le web," Thèse de doctorat, Université Mentouri, Constantine, Constantine, Algérie, September 2011.

[41] B. Saoud Kechid and C. Khentout, "Modélisation des préférences utilisateur pour améliorer la recherche personnalisée," in *Colloque sur la Recherche d'Information et Applications (CORIA)*, 2016, pp. 145–156.

[42] S. Kechid and M. Djoudi, "Profil utilisateur dans un système d'information adaptatif," in *Actes de la Conférence sur l'Adaptation des Systèmes d'Information et des Applications (ASIA)*, 2006.

[43] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*, 7th ed. Pearson, 2015.

[44] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[45] R. Mihalcea and P. Tarau, "Textrank : Bringing order into texts," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2005, pp. 404–411.

[46] A. Bouchoucha, N. Aussenac-Gilles, and J.-M. Pinon, "Une méthode de reformulation de requêtes fondée sur une ontologie dans un système de recherche d'information," in *Conférence en Recherche d'Information et Applications (CORIA)*, 2014.