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

Mémoire de fin d'études

Pour l'obtention du Diplôme de Master en Informatique

Spécialité : Ingénierie des systèmes et technologie du web

Thème :

Exploitation des centres d'intérêts à court-terme de l'utilisateur pour la personnalisation de la recherche d'information

Réalisé par :

- Benkaidali Amira
- Benzekkour Zohra

Proposé et dirigé par :

- Mr RAFA Tahar

Année 2022 – 2023

Remerciements

On remercie dieu le tout puissant de nous avoir donné la santé et la volanté d'entamer et de terminer ce mémoire. Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mr Rafa Tahar, on le remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire. Nos remerciements s'adressent à Mr Rahim Messaoud pour avoir accepté de présider la commission d'évaluation de notre travail. Nous remercions aussi Mme Sadouki Fatiha d'avoir accepté d'examiner notre travail. Nos remerciements s'adressent également à tous nos enseignants pour leurs générosités et la grande patience dont ils ont fait preuve malgré leurs charges académiques et professionnelles.

Dédicace

Je tiens à dédier ce mémoire de fin d'études à toutes les personnes qui ont été présentes à mes côtés durant cette période de ma vie académique.

À ma famille, mes parents et mes frères 'Abdou', 'chaouki' et le petit 'Hamza' je vous suis profondément reconnaissante pour votre soutien indéfectible, vos encouragements constants et votre amour inconditionnel. Votre soutien moral et vos sacrifices ont été les piliers qui m'ont permis d'atteindre ce moment important de ma vie.

À mes cousines 'Zola' et 'Kawter' et amis 'Meriem' et 'soulaf", vous avez été bien plus que des compagnons de vie. Vos discussions, vos échanges et votre entraide ont été des sources d'inspiration et de motivation tout au long de ce parcours. Merci d'avoir partagé cette aventure avec moi et d'avoir rendu ces années d'études si mémorables.

À toi, mon binôme 'Zohra', je souhaite exprimer ma gratitude pour ta présence constante, ta collaboration précieuse et ton soutien indéfectible. Nous avons formé une équipe solide et complémentaire, et notre travail commun a été essentiel à la réalisation de ce projet.

À mon professeurs et encadrant 'Mr RAFA', je tiens à exprimer ma gratitude pour votre enseignement, votre guidance et vos conseils avisés. Votre expertise et votre dévouement ont été des éléments déterminants dans l'acquisition de mes connaissances et compétences. Je suis honorée d'avoir eu la chance de bénéficier de votre accompagnement tout au long de ce mémoire.

À mes grands-mères, ma tante 'Biba' et toutes les personnes qui ont contribué de près ou de loin à mon parcours académique et surtout ma maitresse 'Dalila', je vous adresse mes sincères remerciements. Vos précieux conseils, vos encouragements et votre soutien m'ont permis de surmonter les difficultés et de persévérer jusqu'à la réalisation de ce mémoire.

Enfin, je dédie ce mémoire à moi-même, en reconnaissance de mon engagement, de ma détermination et de ma persévérance. Ce travail est le reflet de mon investissement et de ma passion pour mes études. Je suis fière de cette réalisation et je suis reconnaissante envers moi-même pour avoir repoussé mes limites et atteint mes objectifs.

Que cette dédicace soit l'expression de ma gratitude envers toutes les personnes qui ont contribué à ma réussite et à mon épanouissement académique. Ce mémoire est le fruit de nos efforts collectifs et je suis honorée de pouvoir le dédier à chacun d'entre vous.



Dédicace

Avec l'expression de ma reconnaissance, je dédie ce modeste travail a ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect : mon cher père : abd elkader.

A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non à mes exigences et qui n'a épargné aucun effort pour me rendre heureuse : mon adorable mère Hadda.

A mes chers sœur Fadhila, Meriem et mes frères Hicham, Sid Ali, lakheder, qui n'ont pas cessée de me conseiller, encourage et soutenir tout au long de mes études. Que dieu les protège et leurs offre la chance et le bonheur.

A mes neveux sif aldin, noor, dima, assil qui savent toujours apporter la joie et le bonheur pour toute la famille.et a ma grande mère, et toute la famille.

A tous les cousins, les amis que j'ai connu jusqu'à maintenant merci pour leurs amours.

Sans oublier mon binôme Amira Benkaidali pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

Zahra.

RESUME

En tant qu'un être humain, l'utilisateur du système de recherche d'information cherche toujours la nouveauté dans sa vie, pour cela ses intérêts changent au fil de temps, parfois rapidement et d'autres fois progressivement. De ce fait, l'utilisateur est souvent guidé par ses intérêts a court-terme qui représentent son besoin en information courant. Dans ce mémoire, nous proposons une approche de personnalisation de recherche d'information basée sur l'exploitation de centre d'intérêts a court-terme de l'utilisateur. Dans ce centre d'intérêts, nous exploitons les termes ayant plus d'importance pour l'utilisateur dans le but d'identifier et représenter ses intérêts a court-terme et leur changement. L'exploitation de ces intérêts a court-terme permet l'enrichissement des requêtes d'utilisateur pour les aider à bien exprimer ses besoins en information et guider ses nouvelles recherches d'information.

Mots-clés : Recherche d'information personnalisée, centre d'intérêts à court-terme, centre d'intérêts à long-terme, profil d'utilisateur, pertinence utilisateur.

ABSTRACT

As a human being, the user of the information retrieval system always seeks novelty in their life, and therefore their interests change over time, sometimes rapidly and other times gradually. As a result, the user is often guided by their short-term interests, which represent their current information needs. In this document, we propose an approach to personalize information retrieval based on the exploitation of the user's short-term interests. Within these interests, we leverage terms that hold greater importance to the user in order to identify and represent their short-term interests and their changes. Exploiting these short-term interests enriches the user's queries, helping them to effectively express their information needs and guide their new information searches.

Keywords: Personalized information retrieval, short-term interests, long-term interests, user profile, user relevance.

Table des matières

1 **Introduction Générale** 2

Chapitre 1: **La recherche d'information classique** 5

1 Introduction: 5

2 Les concepts de base d'un système de RI: 6

2.1 Base documentaire: 6

2.2 Document: 6

2.3 Requête: 6

2.4 Besoin d'information: 7

2.5 Pertinence: 7

2.6 Système de recherche d'information (SRI): 8

3 Principales phases du processus de recherche d'information: 9

3.1 Indexation: 9

3.2 Reformulation de la requête: 12

3.3 La reformulation du besoin informationnel *:* 13

4 Modèles de recherche d'information : 13

4.1 Modèle booléen: 14

4.2 Modèle Vectoriel: 15

4.3 Modèle Probabiliste : 16

4.4 Autre Modèle de recherche d'informations : 17

5 Evaluation : 17

5.1 Composants de collection de test 17

5.2 Mesures de l'évaluation 18

6 Limites de la RI classique : 19

7 Conclusion : 20

Chapitre 2: **La recherche d'information personnalisée** 22

1 Introduction 22

2 L'apparition de la RI personnalisée 22

3 Modélisation de l'utilisateur 25

3.1 Pertinence contextuelle 26

3.1.1. Pertinence cognitive: 26

3.1.2. Pertinence situationnelle: 26
3.2. Profil utilisateur: 26
3.3. Source des données de profil 27
3.4. Acquisition: 27
3.5. Représentation du profil utilisateur: 27
3.5.1. La représentation basée sur l'historique de recherche : 27
3.5.2. La représentation ensembliste: 28
3.5.3. La représentation connexionniste: 28
3.5.4. La représentation sémantique: 28
3.6. Evolution de profil utilisateur: 28
3.6.1. Evolution du profil utilisateur à court terme: 29
3.6.2. Evolution du profil utilisateur à long terme: 29
4 Les dimensions du contexte de recherche : 30
4.1. Dimension localisable 30
4.2. Dimension temporelle 31
4.2.1. Dimension sociale 32
4.2.2. Dimension sémantique 33
5 Intégration de profile dans le processus de recherche : 35
6 Évaluation d'accès personnalisée à l'information : 35
6.1. Les limites de l'évaluation classique : 35
6.2. Les mesures de l'évaluation des systèmes personnalisée: 36
7. Conclusion: 37
Chapitre 3: **Exploitation des centres d'intérêts à court terme de l'utilisateur pour la personnalisation de la recherche d'information** 39
1 Introduction 39
2 Modélisation du profil d'utilisateur 40
2.1 Les données personnelles statiques 40
2.2. Base de sessions de recherche 40
2.2..1 Pondération des termes dans une session de recherche 41
2.3. Centre d'intérêts 42
2.3.1. Centre d'intérêt à long terme 42
2.3.2. Centre d'intérêtsà court terme 43
2.3.3. Construction et mise à jour du centre d'intérêts à court-terme 45

2.4. Exemple d'illustration 48
3 Processus de recherche 55
3.1. Sélection de centre d'intérêts le plus adéquat à la requête 56
3.2. Reformulation de la requête 57
4Présentation de l'expérimentation 57
4.1. Création des profiles 57
4.2. Création de la BD 60
4.2.1. Le système de gestion de bases de données MySQL 60
4.2.2. PhpMyAdmin 60
5 Presentation de l'application 62
6 Conclusion 68
Conclusion générale et prespectives 69

Table des figures:

Figure 1:architecture générale d'un systéme de recherche d'information ---------- 9
Figure 2: 1er résultat Google sur « Eclipse » ----------23
Figure 3: 2emeet 3emerésultatGoogle sur « Eclipse »----------24
Figure 4: 4emerésultat Google sur« Eclipse »----------24
Figure 5: 4emerésultat Google sur« Eclipse »----------25
Figure 6: ProfilUtilisateur ----------40
Figure 7: Session de recherche----------41
Figure 8: Base de Cooccurrence Terme-Terme ----------43
Figure 9 : Intérêt croissant----------44
Figure 10: Intérêt stable----------44
Figure 11 : Intérêt décroissant----------45
Figure 12 :Construction et mise à jour du CCT----------46
Figure 13: insertion du terme T1 dans CCT.----------48
Figure 14 : suppression du terme t2 dans CCT ----------50
Figure 15 : surpression de terme t2 dans CCT----------51
Figure 16: maintenance du terme t3 dans CCT ----------52
Figure 17: T4 reste dans CCT ----------53
Figure 18:suppression de terme T5 dans CCT ----------55
Figure 19 : processus de recherche----------56
Figure 20 : Base de données ----------60
Figure 21:Table de CLT ----------61
Figure 22: table de sessions----------61
Figure 23 :Table de CCT ----------62
Figure 24: Interface de recherche----------63
Figure 25: Résultat de la recherche classique (1)----------63
Figure 26: Résultat de la recherche classique (2)----------64
Figure 27 : Résultat de la recherche classique (3) ----------64
Figure 28 : Similarité entre la requête et les centres d'intérêts ----------65
Figure 29 : La requête reformulée.----------66
Figure 30F : Résultat de la recherche personnalisée (1) ----------66

Figure 31 : Résultat de la recherche personnalisée (2) ----------67
Figure 32 : Résultat de la recherche personnalisée (3) ----------67

Liste d'acronymes :

RI : Recherche d'information ;

RIP : Recherche d'information personnalisée ;

SRI : Systèmes de recherche d'information ;

SR : Session de recherche ;

CCT : Centre d'intérêts à court-terme ;

CLT : Centre d'intérêts à long-terme.

Introduction générale

1 Introduction Générale

L'objectif de la recherche d'information RI est de faciliter l'accès a l'information pertinente pour un utilisateur ayant un besoin en information. Le besoin en information de l'utilisateur est souvent exprimé en langage naturel par une requête composée d'un ensemble de mots clés. La requête d'utilisateur est soumise à un système de recherche d'information (SRI) qui permet de retrouver, à partir d'une collection de documents, un sous-ensemble de documents susceptibles d'être pertinents pour la requête de l'utilisateur. Les approches classiques de recherche d'information renvoient, en réponse a une requête utilisateur, un très grand nombre de résultats qui semble naïvement correspondre à cette requête, même si ces résultats ne sont pas tous importants selon les intentions de l'utilisateur. D'où le besoin d'avoir des connaissances sur l'utilisateur et ses intérêts pout pouvoir adapter les résultats de recherche a ses intentions, et par conséquent améliorer la pertinence des résultats de la recherche, et c'est ainsi que la recherche d'information personnalisée est apparait.

La recherche d'information personnalisée (RIP) vise à prendre en compte les intérêts de l'utilisateur dans le processus de recherche d'information pour l'amélioration de la pertinence des résultats. La personnalisation de la recherche nécessite la construction d'une structure pour stocker les données qui caractérisent l'utilisateur et qui représentent des indicateurs d'intérêts pour lui, cette structure s'appelle profil d'utilisateur.

Le profil utilisateur contient généralement des informations (mots-clés, concepts, annotations, dates, locations, URLs, …etc.) qui représentent les centres d'intérêt de l'utilisateur. Ce profil est utilisé pour adapter les résultats retournés par le système de recherche aux intérêts de l'utilisateur, notamment pour la reformulation de la requête utilisateur pour mieux exprimer son besoin en information.

La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction du profil et désigne son adaptation à la variation des centres d'intérêts de l'utilisateur au cours du temps. Le but fondamental de cette gestion est d'affiner la connaissance que le système possède sur l'utilisateur en mettant à profit l'expérience passée lors de ses activités de recherche. Cette gestion comporte deux phases : (i) capturer les changements des centres d'intérêt de l'utilisateur, et (ii) intégrer ces changements au niveau de la représentation du modèle de l'utilisateur.

L'observation de l'évolution des intérêts de l'utilisateur au cours du temps, permet de distinguer deux types de centres d'intérêts à mettre en évidence dans le profil d'utilisateur selon leurs portées :

(i) Centre d'intérêts à court-terme : qui représente les intérêts liés aux activités de recherche courantes de l'utilisateur, et qui couvre une période de quelques jours à quelques semaines. Le but fondamental d'identifier les intérêts à court terme est d'améliorer la précision de recherche en inférant les besoins en information appropriés à la requête et non bruité par des intérêts hors le contexte de recherche.

(ii) Centre d'intérêts à long-terme qui représente les intérêts persistants de l'utilisateur qui sont issus de son historique de recherche tout entier. Ce centre d'intérêts peut être exploitable dans le but d'améliorer la recherche pour les requêtes qui ne sont pas en relation avec les intérêts courants de l'utilisateur

L'objectif de ce travail est de proposer une nouvelle approche de personnalisation de la recherche d'information, basée sur la construction d'un profil d'utilisateur. Ce profil comporte un centre d'intérêt à court terme dans lequel sont sauvegardés les termes pertinents ayant une tendance croissante ou stable de pondération dans les dernières sessions de recherche, et un centre d'intérêts à long-terme qui sert pour sauvegarder tous les termes pertinents de l'historique de recherche.

L'intégration du profil d'utilisateur dans la phase de recherche permet d'inférer le besoin en information de l'utilisateur à partir de sa requête, et de déterminer s'il s'agit d'un intérêt courant pour lequel on doit utiliser le centre d'intérêts à court terme qui est plus utile et plus approprié pour reformuler la requête, ou d'un intérêt d'ordre général pour lequel on doit utiliser le centre d'intérêts à long terme.

Le document est organisé en 3 chapitres comme suit :

- *Chapitre 1,* est réservé pour la recherche d'information classique. Dans ce chapitre nous avons mis en évidence : les notions de base de la RI, le processus de recherche d'information, les différents modèles de recherche les plus connus, ainsi que les techniques d'évaluations des SRI. Une section importante est consacrée à la recherche d'information sémantique avec ses notions de base, les différentes opérations sémantiques liées à la recherche d'information, ainsi que les différentes ressources sémantiques externes exploitées.
- *Chapitre 2,* est dédié à la recherche d'information personnalisée. Dans ce chapitre nous projetons la lumière sur l'émergence de la RI personnalisée, la modélisation de l'utilisateur et sa pertinence, le profil d'utilisateur de sa conception et construction jusqu'à son intégration au processus d'accès à l'information et on termine par l'évaluation des systèmes de recherche personnalisés.
- *Chapitre 3,* présente notre contribution. Où nous présentons une approche de la personnalisation de la recherche d'information basée sur la construction d'un profil d'utilisateur qui permet de mettre en évidence le centre d'intérêts à long terme et le centre d'intérêts à court terme. La présentation du profil est suivie par la présentation de processus de recherche qui permet d'intégrer ce profil dans la phase de recherche.

Le document se termine avec une conclusion et des perspectives.

Chapitre 1

La recherche d'information classique

Chapitre 1 : La recherche d'information classique

1 Introduction

La recherche d'information (RI) est un ensemble des méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information pertinente pour un utilisateur. La recherche d'informations classiques (RI) est un sous-domaine de l'informatique qui se concentre sur la conception et la mise en œuvre des outils informatiques appelés Systèmes de Recherche d'Information (SRI). Elle est axée sur la compréhension des algorithmes et modèles fondamentaux utilisés pour le classement et la récupération des documents.

Le nom de « recherche d'information » (information retrieval) fut donné par Calvin N. Mooers en 1948. La recherche d'information a évolué au fil des années, passant d'une activité manuelle à une activité largement automatisée, pendant les années 1940-1950 la recherche d'information était principalement manuelle, avec des bibliothécaires et des archivistes effectuant des recherches dans des catalogues et des fichiers d'index, Et avec l'avènement des ordinateurs, les systèmes de recherche d'information ont commencé à être développés pour automatiser le processus de recherche entre 1960 et 1970. Au début des années 1990, les moteurs de recherche ont commencé à se développer rapidement, avec l'émergence de sociétés comme Yahoo!, AltaVista et Google. Depuis lors, la recherche d'information a continué à évoluer, avec l'apparition des nouvelles technologies telles que l'apprentissage automatique, le traitement du langage naturel et l'intelligence artificielle, qui ont permis de développer des systèmes de recherche d'information plus intelligents et plus performants.

Les SRI classiques visent à fournir des informations pertinentes en réponse aux requêtes des utilisateurs de manière efficace et évolutive. Ils utilisent la requête de l'utilisateur comme une seule ressource pour définir le besoin en information de ce dernier. L'ensemble de documents récupérés et le même pour la même requête envoyée par des utilisateurs ayant des intérêts différents. En effet, quel que soit le système de recherche utilisé, le résultat de recherche ne peut être pertinent si la requête ne décrit pas explicitement et clairement les besoins de l'utilisateur.

Ce chapitre vise à introduire le domaine de la RI classique. Dans la première partie, nous introduit les concepts de base de la RI, les modèles de RI. Dans la dernière partie nous avons présenté l'évolution des performances des SRIs ainsi que les limites de la RI classique.

2 Les concepts de base d'un système de RI

Les utilisateurs expriment généralement leur besoin d'information avec un ensemble de mots clés définis dans la requête que le système de RI traite pour retourner les documents nécessaires qui correspondent aux besoins des utilisateurs.

Dans cette section, nous allons présenter les concepts de bases d'un système de RI, les modèles de RI, les algorithmes de la recherche ainsi que l'évaluation de la qualité.

2.1 Base documentaire

La base documentaire (ou fonds documentaire) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas général et pour des raisons d'optimalité, la base constitue des représentations très simplifiées mais suffisantes de ces documents. Ces représentations sont étudiées de telle sorte que la gestion (ajout, suppression d'un document) et l'interrogation (recherche) de la base se font dans les meilleures conditions de coût. [Samir Kechid 2009]

2.2 Document

Le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document. Nous utiliserons dans la suite les termes information ou document pour désigner un granule documentaire. [Ourdia]

2.3 Requête

Une requête est une formulation du besoin d'information d'un utilisateur. Elle peut être vue comme étant une description sommaire des documents ciblés par la recherche. Pour une recherche documentaire donnée, l'utilisateur doit soumettre une requête au moteur de recherche dans laquelle il spécifie les mots clés représentant son besoin en information. La requête est exprimée dans un langage d'interrogation parmi lesquels nous citons :

- Des requêtes simples de mots clés ;
- Des requêtes booléennes ou les mots clés sont liés par des opérateurs booléens (AND, OR, NOT) ;
- Des requêtes structures basées sur des attributs tels que les noms d'auteurs, la date de parution… etc. ;
- Des requêtes sous forme d'expressions régulières ;

- Des requêtes complexes qui englobent les types précédents.

2.4 Besoin d'information

La notion de besoin en informatique en recherche d'information est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis par [Ingwersen 92] :

- *Besoin de vérification* : L'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérification est dit stable, c'est-à -dire qu'il ne change pas au cours de la recherche.
- *Besoin thématique connu* : L'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connu. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label.
- *Besoin thématique inconnu* : Cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations en dehors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

2.5 Pertinence

La pertinence est la correspondance entre un document et une requête, une mesure de l'informativité du document à la requête, un degré de relation (chevauchement) entre le document et la requête etc. La notion de pertinence est le critère primaire pour l'évaluation des systèmes de recherche d'informations. Le processus de jugement de la pertinence de l'information est basé sur le degré de similitude de la représentation de la requête avec le contenu du document retrouvé par le système. [Daoud Mariam]

- *Pertinence système* : c'est le degré de pertinence de l'information calculé par le système de RI.
- *Pertinence utilisateur* : c'est le degré de pertinence de l'information correspondant au jugement de l'utilisateur. [Samir Kechid 2009]

2.6 Système de recherche d'information (SRI)

Un système de recherche d'information (SRI) est un ensemble de technologies et de méthodes conçues pour permettre aux utilisateurs de trouver et de récupérer l'information souhaitée. Les SRIs peuvent inclure des bases de données, des moteurs de recherche, des bibliothèques en ligne, des archives en ligne, etc. Les fonctionnalités clés des SRIs incluent la collecte, l'indexation et la catégorisation de l'information, la fourniture de moyens pour formuler des requêtes de recherche, la recherche dans les documents, le calcul de la pertinence et la présentation des résultats. Les Souris peuvent être utilisés pour trouver des informations sur un large éventail de sujets, allant des actualités à l'histoire en passant par la science et les données financières.

Un système de recherche d'Information (SRI) intègre différents acteurs : le besoin informationnel d'un utilisateur ou requête, le corpus documentaire, ainsi qu'un ensemble de modèles et de processus qui permettent de sélectionner des résultats pertinents répondant au besoin en information de l'utilisateur représenté par une requête. Selon Salton [Salton, 1989] "Un système de recherche d'information (RI) est un système qui permet de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse." [Zina Guedria]

Les systèmes de RI suivent une approche simple, communément appelée : Processus en U de RI. Présenté dans la figure suivante :

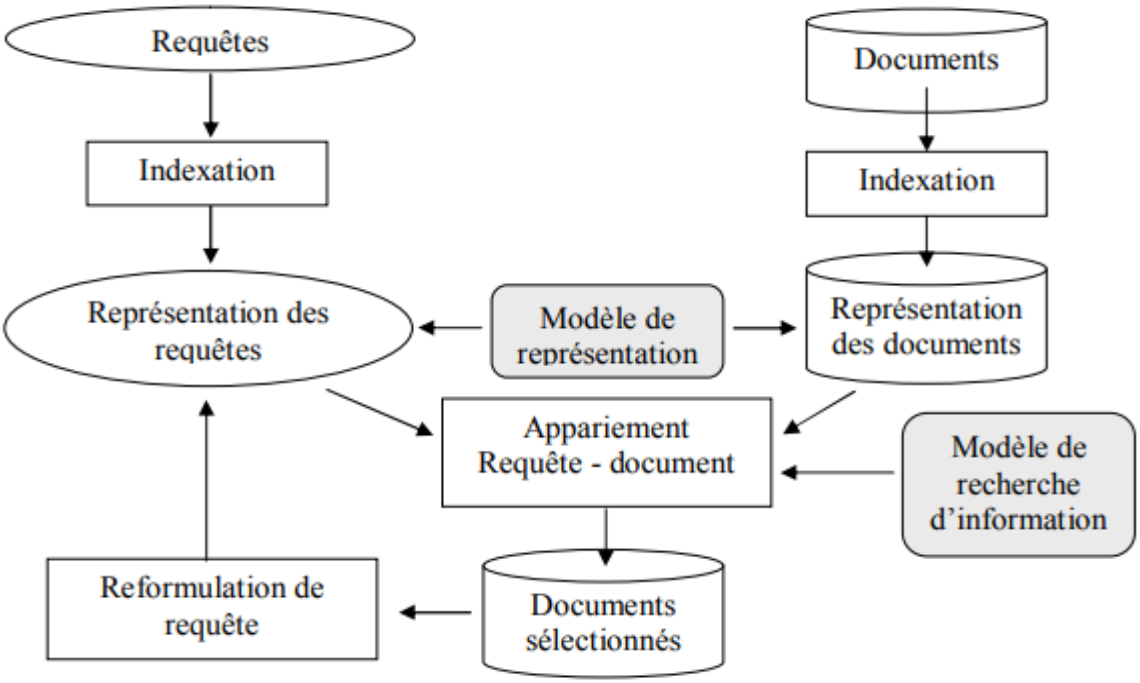


Figure 1:architecture générale d'un système de recherche d'information

3 Principales phases du processus de recherche d'information

Les étapes fondamentales du processus RI ce sont les suivantes :

3.1 Indexation

L'indexation recouvre un ensemble de techniques visant à transformer les documents (ou requêtes) en substituts ou descripteurs capables de représenter leur contenu. [Salton & McGill, 1986]

L'indexation fait référence à la tâche qui construit les descripteurs ou la présentation de chaque document d'une collection de document pour les recherches ultérieures. Ces descripteurs seront par la suite comparés avec la représentation de la requête dans le but de retourner les documents susceptibles de répondre au besoin de l'utilisateur. L'indexation sert alors à extraire les termes les plus représentatifs pour un document ou une requête. Pour cela, il est important de bien indexer un document puisque le résultat de la recherche se base en réalité sur les termes indexant le document.

Il existe différents types d'indexation en recherche d'information :

- **L'indexation manuelle** : L'indexation est réalisée d'une façon manuelle. En effet, chaque document est analysé par un spécialiste du domaine, généralement un documentaliste. Pour chaque document, il synthétise l'idée ou les idées du contenu de ce

document. Il choisit les termes les plus pertinents pour représenter le contenu d'un document.

- **L'indexation automatique** : L'indexation automatique vient résoudre le problème du coût de l'indexation manuelle. En effet, l'indexation automatique est réalisée par un processus entièrement automatisé. Il n'y a donc pas d'intervention d'expert pendant le processus. Une indexation automatique est réalisée sur deux étapes :
  - Rechercher les termes descriptifs qui caractérisent le contenu d'un document ;
  - Évaluer le pouvoir de caractérisation des termes sélectionnés dans l'étape précédente.

  Généralement en recherche d'information, l'indexation automatique consiste à :
  - Sélectionner des termes simples (des groupes de mots dans d'autres approches) ;
  - Supprimer les mots vides (en se basant sur un anti-dictionnaire qui contient les mots outils de la langue) ;
  - Radicaliser ou normaliser les termes restants (en supprimant les variantes morphologiques. Une possibilité est utilisée de l'algorithme de Porter, [Porter, 1980]) ;
  - Pondérer les radicaux obtenus en utilisant leur fréquence relative dans le document courant ainsi que leur fréquence absolue dans la collection (tf.idf pour terme frequency-inverse.document frequency est la pondération la plus connue [Robertson & Jones, 1976]) ;
  - Présenter le résultat de l'indexation pour chaque document généralement, en un ensemble de couple (terme d'indexation, poids).

- **L'indexation semi-automatique (mixte)** : C'est la combinaison de l'indexation manuelle et l'indexation automatique :
  - Premièrement, l'indexation automatique est appliquée : Elle permet d'extraire les termes de chaque document d'une manière automatique ;
  - Ensuite, l'utilisation de l'indexation manuelle se résume dans le fait que c'est le spécialiste du domaine, l'expert ou le documentaliste qui a le choix final pour choisir les termes significatifs et représentatifs ainsi qu'établir les relations entre les termes.

- **Analyse lexicale :** Permet de convertir le texte du document en glossaire de termes, Elle a pour objectif l'identification des mots d'un texte.

- **Elimination des mots vides :** Les mots vides ne sont pas des mots importants dans le document, car ils n'abordent pas le sujet de document.

  On distingue deux techniques pour éliminer les mots vide :

  - L'utilisation d'une liste préétablie de mots vides (aussi appelée anti dictionnaire ou stoplist)
  - L'élimination des mots ayant une fréquence qui dépasse un certain seuil dans la collection. L'élimination des mots vides réduit la taille de l'index, ce qui améliore le temps de réponse du système. [Dounas tarik 2015]

- **Normalisation :** La normalisation consiste à représenter les différentes variantes d'un terme dans un format unique appelé lemme ou racine. Cela a pour effet de réduire la taille de l'index. Plusieurs stratégies de normalisation sont utilisées : table de correspondance, suppression des suffixes.

  Le principal inconvénient de ce procédé est qu'il supprime dans certains cas sémantique des termes originaux, c'est le cas par exemple des termes dérivationnels/dérivés, normalisés par l'algorithme de Porter.

- **Pondération :** La pondération (en anglais weighting) fait référence au processus d'attribution de poids à différents termes dans un document pour refléter leur importance relative dans la recherche.

  Plus précisément, la pondération est utilisée pour calculer le score de pertinence d'un document par rapport à une requête de recherche donnée. Les termes considérés comme plus importants pour la requête reçoivent une pondération plus élevée, tandis que les termes moins pertinents reçoivent une pondération plus faible. La fonction de pondération d'un terme dans un document est connue sous la forme ***Tf.Idf,*** elle est reprise dans différentes versions par la plupart des SRI. Nous distinguons :

  - *Tf (term frequency) :* Cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée de base est que plus un terme apparaît fréquemment

dans un document, plus il est important dans la description de ce document. Tf est généralement exprimé avec l'une des déclinaisons suivantes :

a. *Tf:* utilisation totale.
b. 0.5+0.5TfMax (Tf)

- *Idf (Inverse of Document Frequency) :* mesure l'importance d'un terme dans toute la collection. L'idée de base est que les termes qui apparaissent dans un petit nombre de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection. Cette mesure est exprimée avec l'une des déclinaisons suivantes :
  a. *Idf* = logNdf
  b. *Idf* = logN-dfdf.

Le score de pertinence global du document est ensuite calculé en combinant les poids de tous les termes qu'il contient.

La pondération est une étape clé dans la recherche d'information car elle permet de classer les documents en fonction de leur pertinence par rapport à une requête de recherche donnée.

3.2 Reformulation de la requête

Reformuler un besoin d'information fait référence à l'étape à laquelle l'utilisateur modifie ou re-identifie son besoin d'information au cours de la session de recherche. Cette reformulation peut être faite automatiquement où le système exploite les termes ayant plus de poids dans les documents pertinents ou les documents visités, soit d'une manière manuelle où l'utilisateur définit son besoin en soumettant une nouvelle requête.

L'appariement document-requête : Le processus d'appariement document-requête (matching) permet d'évaluer la pertinence d'un document par rapport à une requête. Plus précisément, lorsque l'utilisateur introduit sa requête le système calcule un score de pertinence entre cette requête et les documents. Le score de pertinence est calculé en utilisant une fonction de correspondance dans le but de calculer un score de similarité RSV (Q, D) (pour Relevance status Value) entre une requête indexée notée Q et un document indexé noté D. Cette similarité est exprimée en fonction du modèle

d'indexation et de pondération des termes. Nous pouvons distinguer deux méthodes d'appariement :

• **Appariement exact** : Dans ce cas, la requête spécifie d'une manière exacte et précise les critères recherchés. Le résultat de la correspondance est une liste non ordonnée de documents qui respectent exactement la requête [Salton, 1971].

• **Appariement approché** : Contrairement à l'appariement exact, l'appariement rapproché fournit une liste ordonnée de documents qui sont sélectionnés en fonction d'un degré de pertinence de la requête par rapport aux documents. La requête décrit d'une sorte les critères recherchés dans un document [Robertson & Jones, 1976].

3.3 La reformulation du besoin informationnel

La reformulation du besoin informationnel fait référence à l'étape où l'utilisateur modifie ou redéfinit son besoin informationnel pendant la session de recherche. Cette reformulation peut être réalisée d'une manière automatique où le système exploite les termes ayant le poids le plus important dans les documents pertinents ou les documents visités, ou alors d'une manière manuelle où l'utilisateur redéfinit son besoin en introduisant une nouvelle requête. [Zina Guedria]

4 Modèles de recherche d'information

Un modèle de recherche d'information (RI) sert pour la modélisation de la mesure de pertinence des documents par rapport à une requête formulée par un utilisateur. Il utilise une combinaison d'algorithmes et de techniques pour évaluer la similitude entre les documents et la requête, et pour déterminer les documents les plus pertinents.

Un grand nombre de modèles de RI est développé dans la littérature. Ils ont en commun le vocabulaire d'indexation basée sur la forme mots clés et ils se diffèrent principalement par le modèle d'appariement requête-document. Selon [R. Baeza-Yates 1999], un modèle de RI est défini par un quadruplet (D, Q, F, R (q, d)) : où

- D est l'ensemble de documents
- Q est l'ensemble de requêtes
- F est le schéma du modèle théorique de représentation des documents et requêtes
- R(q,d) est la fonction de pertinence du document de la requête q

Nous allons présenter dans la suite de ce chapitre les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

4.1 Modèle booléen

Le modèle booléen de la recherche d'information est basé sur l'utilisation des opérateurs (AND, OR, NOT) pour effectuer des recherches compliquées. Ce modèle consiste à autoriser les utilisateurs à combiner plusieurs termes pour former une requête et ensuite récupérer seulement les documents ayant les conditions spécifiées.

Un exemple de représentation d'un document est comme suit :

$$d = t1 \wedge t2 \wedge t3 ... \wedge tn.$$

Un exemple de représentation d'une requête est comme suit :

$$q = (t1 \wedge t2) \vee (t3 \wedge t4).$$

La fonction de correspondance est une fonction binaire se base principalement sur la présence ou l'absence des termes de la requête dans le document. Elle est décrite comme suit :

$$\text{Appariement } (q, d) = \text{RSV } (q, d) = 1 \text{ ou } 0.$$

Par exemple, soit la requête "recherche documentaire". L'utilisateur est intéressé par les résultats qui contiennent "recherche" et "documentaire".

En langage booléen la requête s'écrit de la manière suivante :

$$Q = \text{Recherche} \wedge \text{documentaire}$$

La recherche d'information pour cette requête suit les étapes suivantes :

- **Étape 1** : chercher les documents dont le terme "recherche" fait partie de son index et mettre les résultats dans un ensemble A1.
- **Étape 2** : chercher les documents dont le terme "documentaire" fait partie de son index et mettre les résultats dans un ensemble A2.
- **Étape 3** : enregistrer l'intersection de A1 et A2 dans un nouvel ensemble A3.
- **Étape 4** : trouver les documents qui correspondent aux éléments de l'ensemble A3.

Supposons que l'utilisateur souhaite avoir des documents qui contiennent soit le terme "recherché" OU le terme "documentaire".

La requête devient :

Q = Recherche ∨ documentaire.

Dans ce cas, l'étape 1 et 2 restent inchangées. Ensuite l'étape 3 devient :

- **Étape 3 :** supprimer de l'ensemble A1 les éléments qui sont aussi présents dans A2 et mettre les résultats dans un ensemble A3.
- **Étape 4** : trouver les documents qui correspondent aux éléments de l'ensemble A3

Le modèle booléen utilise l'appariement exact, les documents pertinents correspondent exactement à la requête et sont retournés sans ordre de pertinence. Le modèle booléen étendu [Salton & McGill, 1986] vient rajouter un ordre de pertinence dans les résultats. Un modèle booléen présente des inconvénients : Les documents sont sélectionnés en se basant sur une décision binaire. En plus, la formulation de la requête en langage booléen est difficile et n'est pas évidente pour tous les utilisateurs. L'appariement exact élimine les documents qui sont susceptibles d'être pertinents. [Zina Guerdia]

4.2 Modèle Vectoriel

Initialement proposés par Salton dans le système SMART [204], ces modèles représentent les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel à M dimensions, ces dimensions étant les termes du vocabulaire d'indexation. Chaque document est représenté par un vecteur : dj = (w_{1j}, w_{2j}. . ., w_{Mj}). De même, chaque requête est représentée par un vecteur : qi = (w_{1i}, w_{2i}. . ., w_{Mi}). Avec : w correspond au poids d'un terme dans le document dj ou dans la requête qi. La pondération des composantes de la requête est soit la même que celle utilisée pour les documents, soit donnée par l'utilisateur lors de sa formulation. La pertinence requête-document est traduite par la similarité de leurs vecteurs associés. Le mécanisme de recherche consiste donc à trouver les vecteurs documents qui s'approchent le plus du vecteur requêtes. [Ourdia]

Les principales mesures de similarité utilisées sont :

Le produit scalaire :

$$RSV(qi, di) = \sum_{k=1}^{M} w_{ki} . w_{kj}$$

La mesure de Jaccard :

$$\text{RSV (qi, di)} = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\sum_{k=1}^{M} w_{ki}^2 + \sum_{k=1}^{M} w_{kj}^2 - \sum_{k=1}^{M} w_{ki}.w_{kj}}$$

La mesure cosinus :

$$\text{RSV (qi, di)} = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{(\sum_{k=1}^{M} w_{ki}^2)^{\frac{1}{2}} . (\sum_{k=1}^{M} w_{kj}^2)^{\frac{1}{2}}}$$

4.3 Modèle Probabiliste

L'objectif de modèle probabiliste est d'estimer la similarité entre une requête et un document à partir des probabilités des termes. Ce modèle permet de favoriser les documents ayant la plus forte probabilité d'être pertinents et la plus faible probabilité d'être non-pertinents à la requête. On décrit la pertinence du document *d* par rapport à la requête utilisateur *q* comme suit :

$$\text{RSV (d, q)} = \frac{p(p|d)}{p(\overline{p}|d)}$$

Avec p(p|d) et p(p|d), respectivement représente la probabilité de pertinence et de non pertinence, par rapport à la requête q. Plusieurs méthodes ont été proposées pour estimer cette probabilité, telles que le modèle binaire BIR [C. T. YU et G. SALTON 1976] ou le modèle BM25 [S. E. ROBERTSON et S. WALKER 1994]. Ce dernier est le plus utilisé et ses atouts majeurs consistent en la considération de la longueur des documents dans le calcul du score de pertinence.

Le score RSV en employant le modèle BM25 est calculé comme suit :

$$\text{RSV (d,q)} = \sum_{t_i \in q} idf_{t_i} \times \frac{tf(t_i,d) \,.\, (k1+1)}{tf(t_i,d) + k_1(1-b+b\frac{|d|}{avg_{dl}})}$$

Avec id fti est la fréquence inverse de document pondérant le terme ti de la requête, tf(ti , d) représente la fréquence d'apparition du terme ti de la requête dans le document d, |d| représente la longueur du document et avgdl représente la longueur moyenne des documents. Le paramètres b permet de contrôler la normalisation par la longueur des documents et le paramètre k1 contrôle l'effet de la saturation au niveau des occurrences des termes du document. Les valeurs par défaut des deux paramètres sont k1 ∈ [1.2; 2.0] et b = 0.75.[Ouled Amer 2020]

4.4 Autre Modèle de recherche d'informations

Par ailleurs, il existe d'autres modèles de recherche classique tels que le modèle de langage qui suppose que chaque document a une probabilité de générer une requête donnée, ces documents sont classés en fonction de leur probabilité de générer la requête de l'utilisateur. En plus le modèle de la recherche d'information distribuée qui est utilisé pour la recherche d'information sur les réseaux distribués, tels que le web. Il permet aux utilisateurs de former des requêtes distribuées sur différents serveurs, et les résultats sont ensuite combinés pour fournir une réponse complète à la requête. Ainsi que le modèle LSI qui utilise des techniques mathématiques pour analyser la signification latente (ou cachée) des mots dans un document. Il fonctionne en créant une représentation vectorielle des termes présents dans un corpus de documents. Ces vecteurs représentent la fréquence des termes et leur importance relative dans chaque document. Ensuite, la matrice de termes-document est décomposée en valeurs singulières (SVD), ce qui permet d'extraire les relations sémantiques entre les termes et les documents.

5 Evaluation

L'évaluation d'un système de recherche d'information consiste à mesurer ses performances en termes de pertinence des résultats retournés à l'utilisateur pour une requête donnée. Elle est généralement effectuée à l'aide d'un ensemble de requêtes préétablies et d'un ensemble de documents pertinents pour chaque requête.

Ces données de référence permettent de comparer les performances de différents systèmes de recherche et de mesurer leur efficacité en termes de pertinence des résultats.

5.1 Composants de collection de test

Pour évaluer la performance d'un système de recherche d'information une collection de documents et de requêtes est utilisées et elle représente la collection de test, Les éléments de base de cette dernière comprennent généralement les suivants :

a. *Collection de documents :* est l'ensemble de documents utilisés pour tester le système de recherche. Les documents peuvent être de différents types : des articles, des pages web, des images etc.
b. *Requêtes :* Ce sont les demandes de l'utilisateur pour trouver des documents pertinents dans la collection. Elles peuvent être formulées sous forme de mots clés, des phrases ou des questions.

c. *Évaluation de pertinence :* est l'annotation manuelle de chaque document de la collection de test, indiquant la pertinence de chaque document pour chaque requête.

5.2 Mesures de l'évaluation

Le principal objectif d'un système de recherche d'information est de restituer à l'utilisateur tous les documents pertinents et de rejeter tous les documents non pertinents. Cet objectif est évalué à l'aide de différentes mesures d'évaluation [Sanderson, 2010]. On présente ci-dessous les plus utilisées.

a. La précision : c'est la proportion des documents retrouvés qui sont pertinents, relativement à l'ensemble des documents retournés par le système. Une précision égale à 1 signifie que le système n'a retrouvé que des documents pertinents. Exprimée ainsi:

$$P = \frac{\text{Nombre de documents pertinents retournés}}{\text{Nombre totale de documents retournés par le système}}$$

b. Le rappel : c'est la proportion des documents pertinents qui sont retrouvés par le SRI relativement à l'ensemble des documents contenus dans le corpus.
Un rappel égal à 1 signifie que tous les documents pertinents ont été retrouvés. Exprimé ainsi :

$$R = \frac{\text{Nombre de documents pertinents retournés}}{\text{Nombre de documents pertinents total}}$$

c. F-Mesure : Mesure tenant compte à la fois du rappel et de la précision. Exprimé ainsi :

$$F = \frac{2PR}{P+R}$$

d. E-Mesure (F-Mesure paramétrique) : Une variante de F-Mesure qui tient compte du poids accordé à la précision vis-à-vis du rappel. Exprimé ainsi :

$$E = \frac{(1+\beta^2)PR}{\beta^2 PR}$$

Contrôle le compromis R, P :

- =1 : même poids de précision et rappel (E=F)
- >1 : privilégie la précision au rappel
- <1: plus d'importance au rappel.

e. Average précision (AP) : Mesure la précision moyenne de tous les documents pertinents retournés par le système de recherche pour une requête donnée.

Exprimée ainsi:

$$AP_Q = \frac{\sum Precision(Q)}{Nombre\ de\ documents\ pertinents(Q)}$$

Q: la requête donnée.

f. MAP (Mean Average Precision) : Cette mesure peut être qualifiée de globale puisqu'elle combine différents points de mesure. Elle est moins sensible au nombre de documents pertinents que les mesures Pn. Exprimée ainsi:

$$\text{MAP}= \frac{\sum AP(Q)}{Nombre\ total\ de\ Q}$$

Q: la requête donnée.

6 Limites de la RI classique

La principale limite de la RI classique est qu'elle repose sur une approche généraliste qui évalue toujours les requêtes des utilisateurs et fournit des résultats, quels que soient les critères spécifiques de l'utilisateur émettant la requête ou le contexte de la recherche.

La RI classique utilise la requête de l'utilisateur comme une seule ressource pour définir son besoin en information, sachant que cette requête peut être utilisée par des différents utilisateurs ayant des différents centres d'intérêt. Prenez le terme Java par exemple. Le mot Java est ambigu. Cette requête est ambiguë. Il peut faire référence à un langage de programmation ou au nom d'une île indonésienne, ou aux grains de café de l'île. IR doit tenir compte des besoins potentiels des utilisateurs.

Dans le même exemple, le contexte est également utile pour deux utilisateurs faisant la même demande avec des besoins d'informations différents. Le résultat fourni par le SRI risque de ne pas répondre au besoin d'un des deux, disons un informaticien attendant les résultats d'un cours de langue JAVA, et la description de l'île de java n'est pas pertinente.

Ces limitations ont ouvert la voie à une RI personnalisée. Le domaine qui prend en compte principalement les considérations des utilisateurs comme une composante majeure du processus de recherche.

7 Conclusion

Dans ce chapitre nous avons présenté les principales notions et concepts de la recherche d'information, des systèmes de recherche d'information, les modèles utilisés dans la RI classique, l'évaluation des SRI ainsi que les limites de la RI classiques. A travers les différentes sections que nous avons présentées, nous concluons que la recherche d'information vise à définir des modèles et des systèmes facilitant l'accès à un ensemble de documents trouvés dans des référentiels documentaires ou sur le Web. L'objectif est de permettre aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information, il s'agit donc de retourner l'ensemble de documents pertinents. Cependant, nous constatons que la notion de pertinence dépend de la satisfaction de l'utilisateur d'une part, et des différentes significations portées par les termes de requête d'autre part. Ce constat constitue une faiblesse de la recherche d'information classique et représente également un point de départ pour de nouveaux paradigmes de recherche. Dans le cadre de ce mémoire, nous nous intéressons à une nouvelle direction de l'IR : la recherche d'informations personnalisée.

Chapitre 2

La recherche d'information personnalisée

Chapitre 2 : La recherche d'information personnalisée

1 Introduction

Compte tenu des limites des SRI traditionnels face aux besoins des utilisateurs et la croissance contenue du nombre et types des documents sur le web, il est de plus en plus difficile pour les utilisateurs de trouver du contenu pertinent pour leurs requêtes, cela a incité certains efforts à se concentrer sur l'amélioration de l'accès a l'information qui mène aux approches de RI à s'orienter vers de nouvelle génération basée sur un accès personnalisé aux informations

La personnalisation de l'accès à l'information peut être définie comme un processus fondé sur les caractéristiques sociodémographiques déclarées de l'utilisateur (sexe, âge, lieu de résidence, etc.) et/ou sur son comportement observé.

Le but de la personnalisation est de faciliter l'expression des besoins des utilisateurs, de rechercher des informations sur un certain sujet en écartant les informations non pertinentes, réduisant ainsi l'espace de recherche.

Dans ce chapitre, nous éclairons l'émergence de la RI personnalisée, la modélisation de l'utilisateur et sa pertinence, les informations sur le profil de l'utilisateur depuis sa conception et sa construction jusqu'à son intégration dans le processus d'entretien, et concluons par l'évaluation du système de recherche personnalisé.

2 L'apparition de la RI personnalisée

Au cours de la dernière décennie, la RI personnalisée a pris l'importance grâce à la quantité exponentielle des informations sur le web qui limite la RI classique et la rendre non efficace pour donner les résultats satisfaisants aux utilisateurs. Pour améliorer la pertinence de ces résultats, les moteurs de recherche ont commencé à utiliser des techniques de personnalisation basées sur les données utilisateur (ses préférences et intérêts, les jugements de pertinence, l'environnement social …etc.)

L'essor des réseaux sociaux, des applications mobiles et des objets connectés a permis aux moteurs de recherche de collecter des données sur les utilisateurs à une échelle sans précèdent, ces données aident à la création des modèles d'utilisateurs plus précis pour personnaliser les résultats de recherche en fonction des besoins de chaque utilisateur.

Cet exemple peut illustrer le problème de la personnalisation, si on a deux étudiants formulant la même requête « Eclipse » un de ces étudiants est un astronome et l'autre est un informaticien, les résultats obtenus sont présentés dans les figures (N° 2, 3, 4, 5) comme suit :

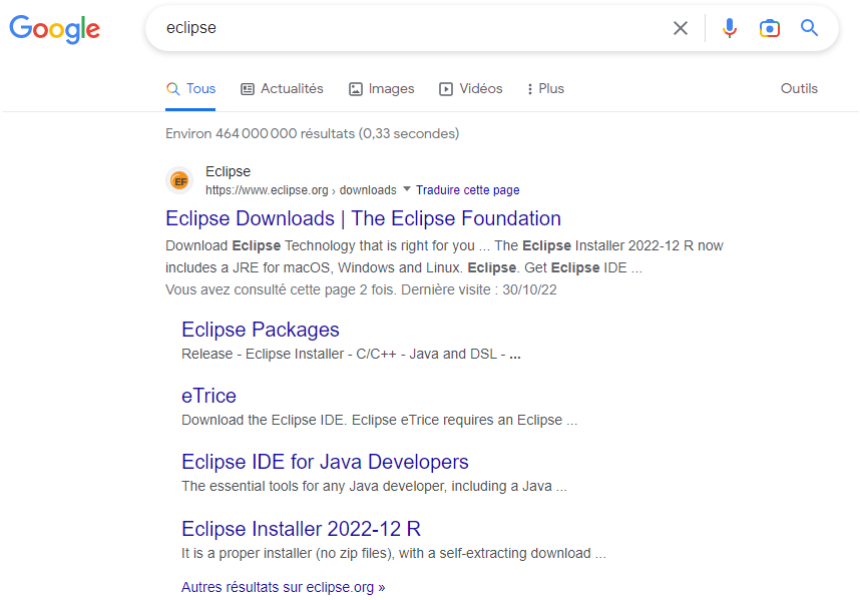


Figure 2: 1er résultat Google sur « Eclipse »

Le premier résultat retourné (Figure 2) concerne le logiciel de développement Eclipse en informatique ainsi que le deuxième et le troisième résultat (Figure 3).

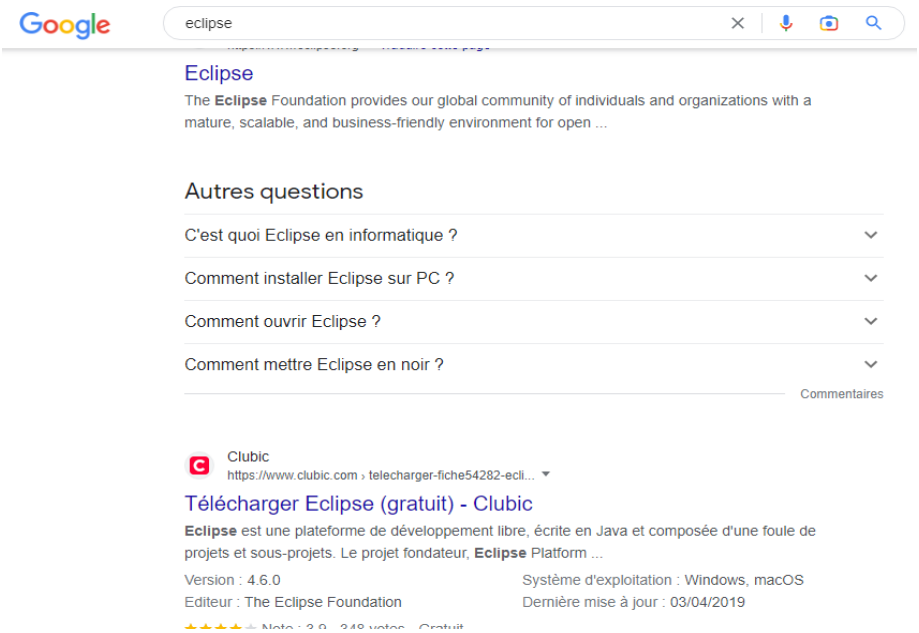


Figure 3: 2emeet 3emerésultat Google sur « Eclipse »

Le quatrième résultat retourné (Figure 4) concerne l'Eclipse de soleil en astronomie.

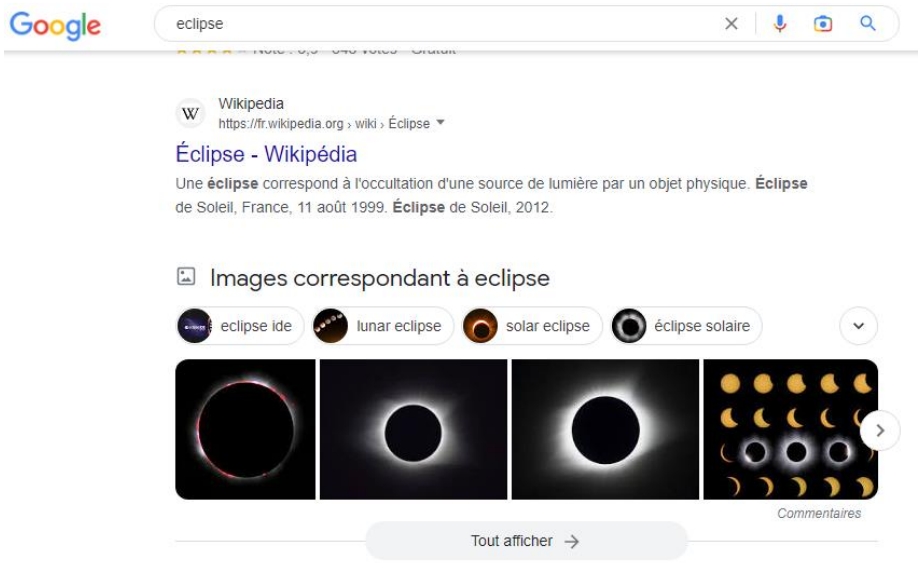


Figure 4: 4emerésultat Google sur « Eclipse »

Les trois résultats suivants (du 5eme au 8eme) sont toujours sur le logiciel de développement Eclipse en informatique (Figure 5).

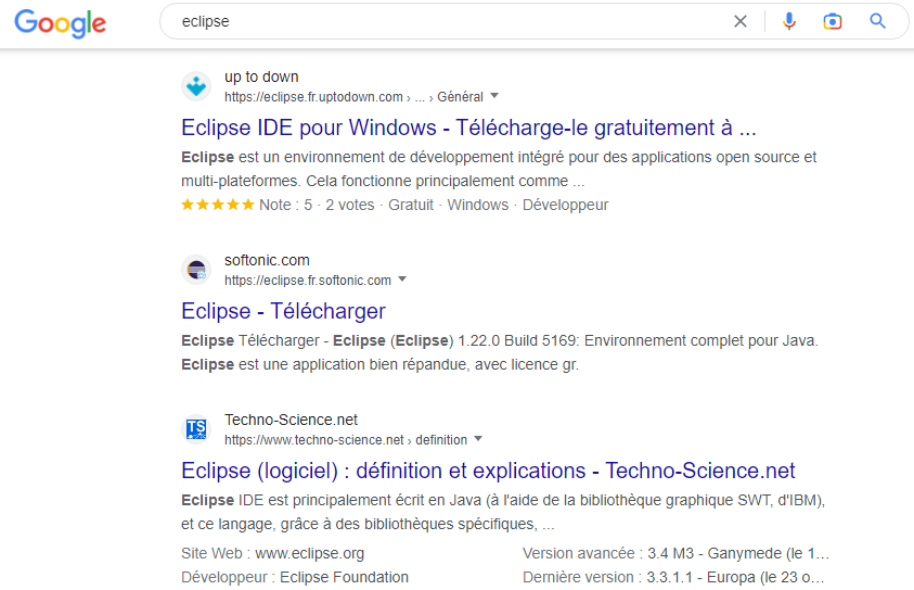


Figure 5: 4emerésultat Google sur « Eclipse »

A partir de ses résultats, l'utilisateur pourra trouver ce qu'il cherche s'il est intéressé au développement, dans notre cas l'informaticien. Cependant l'astronome sera déçu car parmi les huit résultats retournés dans la première page on a seulement un résultat qui l'intéresse.

Si le système de recherche a pris en compte les informations du contexte de l'utilisateur lors de la recherche, il aurait certainement fourni des réponses mieux satisfaisantes pour l'utilisateur.

La prise en compte de ces informations du contexte est généralement appelée modélisation de l'utilisateur, et nous la traitons ci-dessous.

3 Modélisation de l'utilisateur

La modélisation de l'utilisateur est le processus d'acquisition, d'extraction et de représentation des données caractérisant l'utilisateur. Sans modèle d'utilisateur, un système de recherche d'information se comportera exactement de la même manière avec tous les utilisateurs. Mais ces derniers sont différents : ils ont des connaissances différentes, des préférences différentes, ainsi que des buts et des centres d'intérêts différents. La prise en compte de l'utilisateur et les données lui caractérisant est au cœur du processus de personnalisation de la recherche d'information. A

cet effet, les recherches dans le domaine de la modélisation de l'utilisateur se sont principalement concentrées sur la définition d'approches servant à cette modélisation [Ould Amer 2020].

On peut résumer la modélisation de l'utilisateur par la question suivante : quelles sont les informations significatives sur l'utilisateur pouvant représenter des capteurs de pertinence pour lui, que doit contenir son profil pour guider ses nouvelles recherches ?

3.1 Pertinence contextuelle

La pertinence est sans doute la question fondamentale posée lors de la sélection des documents à retourner en réponse d'une requête d'utilisateur. La pertinence est un concept subjectif et ne peut être généralisé à aucun type informatif et efficace pour tous les types d'utilisateurs. Cela dépend surtout du centre d'intérêt ou le domaine d'application, l'heure, le lieu et le support sélectionnés par l'utilisateur qui a accès à l'information et aux systèmes qui fournissent cette information.

3.1.1. Pertinence cognitive

Il s'agit d'une pertinence associée à un sujet de requête, "pondérée" par la perception ou la connaissance de l'utilisateur sur le même sujet.

3.1.2. Pertinence situationnelle

Il s'agit d'une corrélation liée à la tâche de recherche. Ce gars La pertinence reflète essentiellement l'utilité de l'information par rapport à sa finalité recherche d'utilisateurs.

3.2. Profil utilisateur

Le terme profil d'utilisateur a été largement débat tu dans le domaine de la modélisation des utilisateurs. Depuis le début des années 1970, le concept de portraits d'utilisateurs a été introduit en affichant le centre d'intérêt de l'usager, accédez à des informations pour décrire la structure présentative de l'usager.

La notion de profile est directement liée à l'utilisateur, elle est à décrire la structure présentative de l'utilisateur, plus précisément son centre d'intérêt. Dans la plupart des systèmes où les utilisateurs doivent être modélisés pour adapter les informations à leurs besoins spécifiques, les utilisateurs sont représentés par des profils composés généralement de leurs centres d'intérêt. Les

intérêts des utilisateurs se construisent et s'enrichissent au fil du temps à partir de leurs interactions avec le système de recherche. [Tchuente 2013]

Un profil d'utilisateur est défini comme "contient l'acquisition de tous les aspects d'un utilisateur qui peuvent être des sources utiles de connaissances sur le comportement du système". [Bouramoul, 2011]

3.3. Source des données de profil

Selon les tâches à effectuer, un profil d'utilisateur peut contenir des informations très différentes. Dans une recherche d'informations, le contenu d'un profil d'utilisateur est résumé : ses caractéristiques personne elles, ses intérêts et préférences, ses compétences, ses objectifs actuels et enfin l'historique d'interaction avec le système.

3.4. Acquisition

Étape très importante dans le développement des systèmes personnalisés de recherche d'informations. Les données de l'utilisateur peuvent être obtenues implicitement ou explicitement. Dans la récupération explicite, les données sont récupérées directement par l'utilisateur lui-même. Dans l'acquisition implicite, les données de l'utilisateur sont obtenues à partir de son comportement. L'acquisition et la collecte des données des utilisateurs est une, des interactions avec le système et réalisée à partir de leur historique de recherche.

3.5. Représentation du profil utilisateur

La représentation du profil utilisateur est de représenter ces centres d'intérêts intérieurement. Nous présentons dans cette section les différents modèles de représentation possibles : la représentation basée sur l'historique de recherche, la représentation ensembliste, connexionniste ou conceptuelle ainsi que la représentation sémantique.

3.5.1. La représentation basée sur l'historique de recherche

La représentation basée sur l'historique de recherche peut simplement inclure les types de requêtes de recherche effectuées par l'utilisateur, les mots-clés utilisés dans les requêtes, les sources d'information consultées, les résultats de recherche préférés, les filtres de recherche utilisés, les dates et heures des requêtes.

3.5.2. La représentation ensembliste

La représentation ensembliste se base sur un ensemble de mots clés (ou vecteurs de termes) pondérés représentés souvent selon le modèle vectoriel de Salton. Le poids d'un terme est souvent calculé selon le schéma *TF*IDF*. Les paquets de termes représentent généralement les centres d'intérêts de l'utilisateur. La représentation ensembliste est simple et facile à implémenter. Cependant, cette représentation manque de cohérence, de structuration et ne permet pas de représenter les différentes relations de corrélation entre les centres d'intérêts de l'utilisateur [Rafa 2023].

3.5.3. La représentation connexionniste

La représentation connexionniste des profils utilisateurs consiste à représenter le centre d'intérêt de l'utilisateur à travers un réseau de nœuds pondérés, où chaque nœud représente un concept qui traduit le centre d'intérêt de l'utilisateur. Cette représentation peut combler les lacunes des représentations ensemblistes en établissant des relations sémantiquement pertinentes entre les centres d'intérêt dans les profils. En effet, la richesse sémantique de cette représentation permet d'aborder les problèmes d'ambiguïté terminologique inhérents aux représentations ensembliste, d'éventuelles incohérences entre centres d'intérêt et d'identification des relations par la sémantique aux profils appropriés des sujets de requête [Daoud 2009].

3.5.4. La représentation sémantique

La représentation sémantique met davantage en relief les relations de sens entre unités d'information représentant le profil en apportant des solutions aux problèmes de dissémination et synonymie. La direction proposée dans ce contexte, est la construction hiérarchique de concepts plutôt qu'une liste de structures indépendantes, à partir d'informations issus des fichiers *logs*. La hiérarchie peut rendre compte des niveaux de préférences de l'utilisateur et des associations latentes entre concepts et donc un raisonnement sémantique pour la dérivation du profil s'prête aisément. Ce type de représentation utilise généralement des ontologies [Kechid 2009].

3.6. Evolution de profil utilisateur

L'évolution de la représentation du profil utilisateur décrit le changement des degrés d'intérêts dans certains domaines présentés par une mise à jour de la structure/contenu des centres d'intérêts appris ou bien l'apparition d'un nouveau besoin en information qui implique un nouveau centre d'intérêts au profil utilisateur.

« La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction d'un profil utilisateur et désigne leur adaptation à la variation des centres d'intérêts des utilisateurs au cours du temps » [Daoud 2009].

Le rôle principal de la gestion de l'évolution du profil utilisateur est de capter les changements des centres d'intérêts de l'utilisateur dans un premier temps et les transmettre à l'échelon de la représentation du profil.

3.6.1. Evolution du profil utilisateur à court terme

« Le profil utilisateur à court terme décrit des centres d'intérêts et des besoins utilisateurs lies aux activités et les taches de recherche courante. »[Daoud 2009].

L'évolution du profil utilisateur à court terme fait référence aux changements dans les intérêts et comportements d'un utilisateur au fil de temps, sur une période relativement courte. Ces changements peuvent être influencés par divers facteurs, tels que les évènements récents, les expériences personnelles, les tendances culturelles, les nouvelles technologies, etc.

Les moteurs de recherche peuvent utiliser ces changements dans le profil utilisateur pour personnaliser les résultats de recherche en temps réel. Cela peut inclure la recommandation de contenu pertinent en fonction des dernières recherches de l'utilisateur, de son historique de navigation, de ses interactions sur les réseaux sociaux et de tout autre comportement en ligne qui peut fournir des indications sur ses intérêts actuels.

On distingue parmi les SRI personnalisés des systèmes qui s'adaptent bien aux besoins courants de l'utilisateur, tels que Les portails Web (MyYahoo, InfoQuest, etc.), *Yahoo Contextual Search* (Y ! Q), WebPersonae et d'autre systèmes.

Yahoo ContextualSearch (Y! Q) est un outil de recherche personnalisée conçu pour aider les utilisateurs à trouver plus des informations pertinentes liée sa leur recherche courante. [Daoud 2009].

3.6.2. Evolution du profil utilisateur à long terme

« Le profil utilisateur à long terme modélise des centres d'intérêts généraux, persistants, ou récurrent sets issus de son historique de recherche tout entier. Ce profil peut être exploitable dans le but d'améliorer la recherche pour toute requête soumise par l'utilisateur » [Daoud 2009].

Le profil utilisateur à long terme est un profil détaillé d'un utilisateur qui est construit sur une longue période de temps, Il peut être également utilisé pour améliorer l'expérience utilisateur en fournissant des résultats plus pertinents et plus personnalisés.

En suivant l'évolution du profil utilisateur à long terme, les systèmes peuvent comprendre comment les intérêts et les besoins de l'utilisateur ont changée au fil du temps et adapter leurs recommandations en conséquence.

« L'acquisition implicite des données utilisateurs pour l'évolution du profil utilisateur à long terme se fait selon des approches proactives ou réactives. Dans les approches proactives, les données sont collectées et traitées dans une étape préliminaire, souvent basée sur une analyse des relations entre requêtes et pages web visitées par l'utilisateur. Dans certaines approches réactives, la collecte des données est conjointement liée à des mécanismes de délimitation des sessions de recherche dans le cas où le profil à long terme est construit par agrégation des profils appris à court terme. Sinon la collecte des données est indépendante de tout intervalle de temps et est faite durant les interactions de l'utilisateur avec le système ». [Daoud 2009].

4 Les dimensions du contexte de recherche

Le contexte de recherche en RI se réfère à l'ensemble des facteurs et des conditions qui entourent une requête de recherche effectuée par un utilisateur. Cela peut inclure les caractéristiques de l'utilisateur, le but de la recherche, le domaine de recherche, le temps disponible, les ressources disponibles, etc. L'importance du contexte de recherche se voit dans son impact significatif sur la façon dont les résultats de recherche sont interprétés et utilisés par l'utilisateur. Parmi ses dimensions, on cite la dimension spatiale ou localisable, temporelle, sociale et sémantique.

Nous présentons dans cette section les différentes dimensions contextuelles des informations prenant en considération comme c'est affirmé dans [Buidghaghen 2011], qu'il n'est souvent pas facile de trouver un SRI personnalisé qui gère toutes les dimensions contextuelles en même temps.

4.1. Dimension localisable

Cette dimension se réfère à l'emplacement géographique ou la recherche est effectuée, la localisation de l'utilisateur mobile est souvent considérée comme un facteur d'une grande

importance et un concept seulement caractérisé par les coordonnées physiques par rapport à un système de coordonnées géographiques.

Toutefois, les coordonnées physiques, même étant pertinentes pour caractériser une localisation ne sont pas suffisantes pour définir cette dernière. Certainement, comme est mentionné dans [Ourdia] il existe différentes façons possibles et correctes pour caractériser la localisation de l'utilisateur mobile :

- Position absolue dans un système de coordonnées géographiques,
- Relative (a cote de, …),
- Nom de place
- Classe nommée qui représente le type de la place, eg. Ecole (démographique), Plage (environnement), monument (historique), domicile (personnel), ….

La représentation en 'classe nommée' ou 'type de la place' était retenu dans un objectif de caractériser la dimension spatiale de la situation de l'utilisateur. Ce type de classification reflète généralement une caractérisation fonctionnelle de la place.

'Le type de la place' peut être récupérer à partir de classifications manuelles de places qui sont pratiquées par la plupart des systèmes d'informations géographiques, tel que *The Alexandria Digital Library* et *GeoNames servers*, qui donnent à chaque coordonnée géographique (adresse ou nom de place) une classification standard de catégorie de type de localisation.

4.2. Dimension temporelle

L'information contextuelle temporelle en RI est très importante pour la description des intérêts de l'utilisateur et elle fait référence aux éléments du contexte qui sont liés à la dimension temporelle de la requête de recherche. Le contexte de recherche peut varier en fonction de l'instant où la recherche est effectuée, la durée de la recherche, l'évolution des préférences de l'utilisateur et les changements dans le domaine de recherche au fil du temps etc.

La recherche d'information temporelle est un domaine de recherche qui a principalement un but d'améliorer des modèles classiques de la RI en combinant la notion de pertinence traditionnelle avec la dimension temporelle. En effet, dans certains cas le besoin réel d'information d'une requête utilisateur est incité par un événement important qui occupe l'opinion public sur le réseau social. Ainsi, la prise en compte des évènements et des informations temporelles par

l'extraction à partir des textes et la représentation dans le profil utilisateur peut bien aider à améliorer la pertinence des résultats de recherches [rafa 2023].

Dans [Ourdia] est proposé une approche recherche d'information permet d'abstraire l'instant temporel dans certaines périodes spécifiques et significatives qu'il est estimé d'avoir un effet sur le comportement de l'utilisateur (e.g. matin, weekend, été). Il était choisi de représenter le temps sur trois dimensions : journée, semaine et saison. La variation de chacune de ces dimensions peut altérer l'activité de l'utilisateur mobile. Les auteurs proposent pour une bonne représentation de l'information temporelle et sa manipulation des approches sémantiques avec des ontologies temporelles.

4.2.1. Dimension sociale

Un contexte informationnel social de l'utilisateur peut être extrait de ses actions et activités au sein du réseau social. Il peut être composé de différents types d'informations : annotations, commentaires, notes, relations sociales, etc.

Deux informations sociales principales sont souvent utilisées :

- ***Les relations sociales***

Les relations sociales entre les utilisateurs d'un système de recherche peuvent avoir un impact significatif sur les résultats de la recherche et leur exploitation permet d'enrichir le profil utilisateur ou la pertinence prédite d'un document donné peut être améliorée en explorant les actions du voisinage de l'utilisateur.

Parmi les exemples de relations sociales qui peuvent influencer la recherche d'information :

a. *Relations de confiance* : les utilisateurs sont plus susceptibles de faire confiance à des sources d'information recommandées par des personnes de confiance,
b. *Influence sociale :* les utilisateurs peuvent être influencés par les opinions et les comportements de recherche de leurs pairs,
c. *Collaboration* : les utilisateurs peuvent collaborer pour trouver et partager des informations pertinentes,

Les liens d'amitié avec les autres utilisateurs de système de recherche peuvent aider à deviner le besoin en information de l'utilisateur, de fait que les besoins cognitives des individus peuvent

être influencés par ce qui intéresse et attire l'attention de leurs communautés proches (amis, voisins, collègues, …) [Badache et al. 2019]. En effet, certaines études selon [Smyth 2011] affirment que (66 %) de nos recherches sont similaires à celles de nos collègues. [Rafa 2022]

- ***Les annotations sociales (Tags)***

Les annotations sociales, également appelé es 'tags', sont une forme de méta données générées par les utilisateurs pour décrire et organiser les informations, souvent utilisées pour déterminer les centres d'intérêts des utilisateurs. Les annotations sociales sont des mot clés ou des phrases courtes très significatives pour caractériser l'utilisateur de fait qu'ils expriment un avis, un jugement ou une impression sur les contenus des documents [Kamer et al. 2014; Ould Amer 2020].

Dans [Ould Amer 2020], les auteurs ont discuté le défi de proposer une meilleure estimation des poids des tags des utilisateurs, permettant de refléter de façon pertinente le degré d'intérêt de chaque tag pour l'utilisateur. Le poids de chaque tag suit un schéma de pondération basé sur la fréquence d'utilisation de ce tag par l'utilisateur (TF), ou par un *TF-IDF*. Pour la représentation du profil basé sur les tags, ils ont proposé une liste pour chaque utilisateur appelée $Profil(u)_{tags}$ ayant chaque tag avec son poids comme suit :

$$Profil(u)_{tags} = \{< tg1, w_{tg1} >, < tg2, w_{tg2} >, \ldots, < tgY, w_{tgy} >\}$$

Ou *u* désigne un utilisateur.

4.2.2. Dimension sémantique

En RI traditionnelle, les résultats de recherche sont souvent basés sur des correspondances de mot clés entre la requête de recherche et les documents indexés. Cependant, cette approche semble approximativement à une opération de reconnaissance de forme et peut produire des résultats imprécis ou non pertinents, en particulier lorsque les termes de recherche ont plusieurs significations ou sont ambigus [Aimé et al. 2010]. La sémantique dans la RI personnalisée mène à comprendre la signification des termes de recherche en contexte, en prenant en compte les synonymes, les termes connexes et les concepts liés à la requête de recherche pour une compréhension profonde du besoin de l'utilisateur, sa requête et le document cible.

La sémantique aide à établir des relations entre les concepts et les termes de recherche, en créant des ontologies ou des taxonomies pour organiser les termes de recherche en fonction de leur signification et de leur relation avec d'autres concepts. Comme cité dans [Zargayouna et al. 2015], Les approches de recherche d'information personnalisée orientée sémantique, s'appuient sur l'exploitation des ressources sémantiques externes tels que les ontologies, les thesaurus, les dictionnaires…etc. pour extraire les liens sémantiques entre les termes.

Récemment, plusieurs travaux en RI personnalisée visent à introduire la sémantique dans le profil d'utilisateur pour améliorer la pertinence des résultats de recherche par la compréhension des intentions de l'utilisateur. Les auteurs dans [Ould Ameur 2020] ont discuté les approches qui ont proposé de tirer parti des concepts des bases de connaissances tels que DBpedia pour représenter les centres d'intérêts des utilisateurs. L'un des avantages de l'exploitation des bases de connaissances est qu'il est possible d'exploiter les connaissances de base de ces concepts pour déduire le centre d'intérêts des utilisateurs qui pourraient ne pas être capturés utilisant les approches basées sur des mots clés

Différentes stratégies de représentation des profils à l'aide des concepts ont été explorées :

- **Représentation par les entités** : une approche qui extrait les entités à partir des données des utilisateurs et utilise ces entités pour représenter les centres d'intérêts d'un utilisateur [F. ZARRINKALAM. 2015].
- **Représentation par les catégories :** une approche qui exploite les catégories de DBpedia qui permettent d'avoir une représentation plus générale des centres d'intérêts des utilisateurs à la différence de la représentation par les entités [S. FARALLI, G. STILO].
- **Représentation hybride** : une approche qui combine les deux représentations précédentes (par les entités et par catégories) [P. KAPANIPATHI et al.2014, G. PIAO et J.J.G. BRESLIN. 2016].

Ainsi, [T. Rafa S. Kechid 2019] présentent une approche de personnalisation sémantique de la recherche basée sur l'ontologie ODP (open directory project). Les concepts de l'ontologie sont utilisés pour l'annotation des documents pertinents de l'historique de l'utilisateur, afin de créer un profil sémantique. Ces concepts sont utilisés aussi dans l'indexation sémantique des documents pour créer un espace de recherche sémantique. Dans [Schenkel et al. 2008], en plus de la dimension sociale sur laquelle les auteurs se basent pour construire un contexte social de

l'utilisateur, ils proposent de prendre en compte une dimension sémantique pour sélectionner des termes. Cela se fait sur la base d'une similarité sémantique entre ces termes et ceux de la requête de l'utilisateur.

5 Intégration de profile dans le processus de recherche

L'intégration des profils utilisateurs revient à utiliser ces derniers dans l'une des principales phases d'évaluation de la demande : reformulation, calcul de la note de pertinence de l'information ou encore présentation de certaines des méthodes que nous avons évoquées :

- Intégrez des profils d'utilisateurs dans la fonctionnalité de correspondance de documents de requête. Le centre d'intérêt est utilisé dans la fonction d'appariement du modèle IR "modèle bayésien". La valeur de pertinence d'un document par rapport à une requête n'est plus en fonction de la requête elle-même, mais plutôt de l'intérêt de l'utilisateur qui l'a soumise. [Tamine 2007]
- Intégrez des profils d'utilisateurs dans la présentation des résultats, ils sont basés sur la réorganisation des résultats de recherche. Le reclassement est basé sur une combinaison entre le classement initial du document et le classement résultant d'une mesure de similarité entre le document et le profil de l'utilisateur. [C. Asrar et A. sara 2020]
- Le reclassement des résultats de recherche consiste à combiner les scores bruts de correspondance des documents avec les scores de similarité entre les documents et le centre d'intérêt de l'utilisateur. Ce dernier est calculé en appliquant une mesure de similarité vectorielle basée sur le cosinus entre les documents et les profils des utilisateurs. [Vallet et al. 2006]

6 Évaluation d'accès personnalisée à l'information

6.1. Les limites de l'évaluation classique

L'évaluation est une étape dans la mesure de la capacité d'un système de recherche documentaire à satisfaire les besoins d'information des utilisateurs. Enrèg le générale, les systèmes IR classiques sont évalués par une série de tests composés de 3 éléments de base, qui sont :

- Un ensemble de requêtes représentant les besoins d'information de l'utilisateur. Ce groupe est fourni pour évaluer la capacité de chaque système de requête à trouver des documents pertinents.
- Un ensemble de documents contenant des informations pertinentes en réponse à toutes les requêtes des utilisateurs.
- Jugement de pertinence, qui permet d'identifier les dossiers pertinents pour chaque demande. Ces jugements sont généralement fournis par les utilisateurs.

Bien que la performance de l'évaluation classique ait été démontrée par divers es activités d'évaluation (telles que TREC, CLEF ou autres), et soit généralement indiquée par des métriques de précision/rappel, cette stratégie d'évaluation présente certaines limites liées au sujet. Les tests effectués n'ont tenu compte ni du contexte dans lequel la recherche a été effectuée, ni de la perception de l'utilisateur de la pertinence dans ce contexte, ni de la diversité des centres d'intérêt de l'utilisateur. Il est donc difficile d'identifier un ensemble de tests qui reflète les aspects subjectifs du concept de pertinence et le centre d'intérêt de l'utilisateur Enfaite, les modèles classiques ne traitent pas des exigences dynamiques en information, mais traitent les exigences comme des concepts statiques pleinement reflétés par les requêtes.

6.2. Les mesures de l'évaluation des systèmes personnalisée

- ***La mesure de précision P@K :***

  Les performances du système sont habituellement évaluées en termes de **moyenne** de la précision à la position k P@k, de la moyenne de la précision AP et de la moyenne de la précision de toutes les requêtes MAP, définies comme suit :

$$p@k = \frac{nombre\ de\ document\ pertinents\ sélectionés\ jusqu'rangk}{nombre\ totale\ document\ pertinents\ jusqu'rangk}$$

- ***F_Mesure:***

  Est une mesure qui estime la moyenne harmonique pondérée de la précision et du rappel, décrite par la formule suivante :

$$F - mesure = \frac{2 * précision * rappel}{précision + rappel}$$

- ***La MAP (Mean Average precision):***

  Est une mesure qui estime la moyenne de la précision moyenne de toutes les requêtes, et mesure ainsi les performances globales du système, et est décrite par la formule suivante :

$$MAP = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{r}\sum_{r\in R} précision(q)@r$$

- ***Le NDCG (Normlized Discounted Cumulative Gain):***

  Il évalue la capacité de l'IRS à renvoyer les documents pertinents en haut de la liste des résultats. Il est défini comme suit:

$$NDCG@K = \frac{DCG}{IDCG} OUDCG@K = \sum_{i=1}^{k} \frac{2^{rel(k)} - 1}{\log_2 i}$$

  Et IDCG@k est la valeur maximale possible de DCG@k.

7. Conclusion

Dans ce chapitre, nous avons présenté les principaux concepts et concepts de la Recherche d'Informations Personnalisées (RIP), tels que les concepts de contexte, de profil d'utilisateur et de pertinence contextuelle. Nous citons également divers services de personnalisation, à savoir le filtrage des résultats, le reclassement des résultats de requête et la recommandation. Nous étudions ensuite l'architecture générale d'un système de recherche d'informations (SRIP), où nous identifions deux fonctions de base de gestion de profil et de sélection d'informations, ainsi qu'un prototype d'un système d'accès personnalisé et une approche de modélisation de profil utilisateur. Et nous étudions l'évaluation des systèmes d'accès à l'information personnalisés, dans laquelle nous énumérons les questions d'évaluation et quelques recommandations.

Enfin, nous pouvons accéder aux apports des études sémantiques et de personnalisation en ciblant la pondération terminologique et la représentation sémantique des profils utilisateurs.

Chapitre 3

Exploitation des centres d'intérêts à court terme de l'utilisateur pour la personnalisation de la recherche d'information

Chapitre 3 : Exploitation des centres d'intérêts à court terme de l'utilisateur pour la personnalisation de la recherche d'information

1 Introduction

La recherche d'information personnalisée défend l'hypothèse de base que deux utilisateurs différents peuvent formuler la même requête pour exprimer des besoins en information différents. Et de fait, leur retourner la même réponse peut ne pas les satisfaire tous deux. Et ainsi, l'objectif des systèmes de recherche d'information est d'inférer le réel besoin en information de l'utilisateur caché derrière sa requête, et de lui retourner par la suite les résultats les mieux correspondants à ce besoin. [Rafa 2023]

Dans la recherche d'information, les nouvelles recherches des utilisateurs sont souvent en relation avec leurs recherches précédentes. Par conséquent le besoin en information de l'utilisateur peut être influencé par son historique de recherche. Pour cela les profils d'utilisateurs donnent plus d'importance aux données dynamiques implicitement extraites à partir de l'historique pertinent de l'utilisateur.

Pour améliorer la représentativité des profils et leur capacité à décrire les intérêts des utilisateurs et d'inférer leurs besoins en information, certaines approches exploitent des données thématiques (documents et termes), d'autres approches [Daoud 2009] exploitent le profil a court-terme dans le processus d'ordonnancement personnalisé des résultats de recherche des requêtes traitant le même domaine d'intérêts porté par ce profil. Il s'agit d'exploiter le profil à court terme pour des requêtes inscrites dans une même session de recherche ou le profil à long terme le plus proche d'une requête initiant une nouvelle session de recherche.

Dans ce chapitre, nous proposons une approche de personnalisation basée sur l'exploitation de centre d'intérêts à court-terme sans trop baser sur le temps mais plutôt sur l'évolution de l'importance de terme dans l'historique de recherche récent de l'utilisateur. Notre objectif dans cette approche est d'exploiter le centre d'intérêts à court-terme pour reformuler les requêtes initiales de l'utilisateur dans le but de les désambigüiser et les aider à mieux exprimer le réel besoin en information de l'utilisateur, et par conséquent pour améliorer la pertinence des résultats de recherche.

Le reste de ce chapitre est organisé comme suit : la section 2 représente la modélisation du profil d'utilisateur ; la section 3 représente le processus de recherche suivi par la section 4 présentation de l'expérimentation, finalement une conclusion.

2 Modélisation du profil d'utilisateur

Le profil utilisateur que nous proposons dans ce travail se compose des composantes illustrées suivant la figure 3.1 :

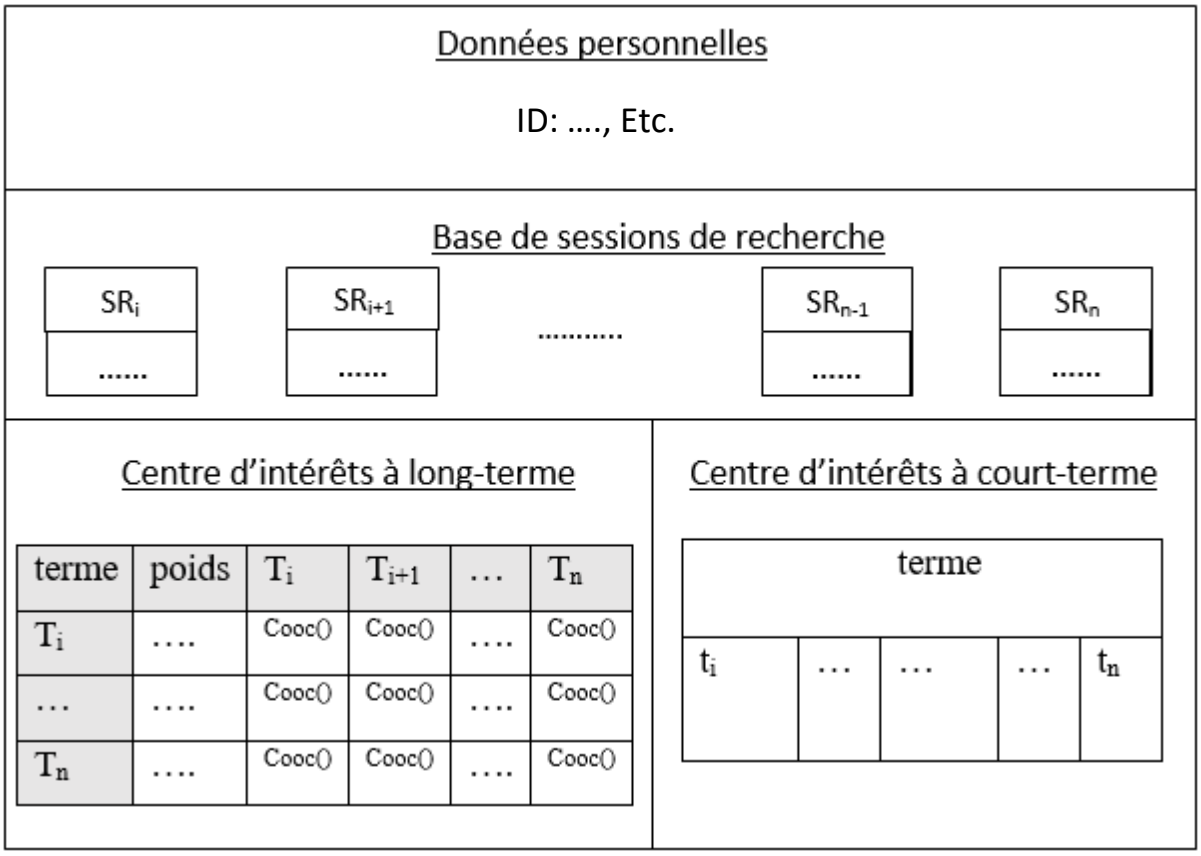


Figure 6: Profil Utilisateur

2.1 Les données personnelles statiques

C'est l'ensemble de données introduites explicitement par l'utilisateur et qui permettent de l'identifier. Nous pouvons trouver le nom, l'âge, l'adresse, le domaine, les préférences, …etc.

2.2. Base de sessions de recherche

Une session de recherche est une série de requêtes successives lancées dans le même emplacement. Dans la littérature, la durée d'une session peut varier de quelques minutes [Göker et al, 2000], à quelques heures [Spink et al, 2006], ou inférieure à une journée [Leva et al, 2014].

Pour nous, une session de recherche représente les recherches faites en une journée.

L'ensemble de sessions de recherche qui représente l'historique de recherche de l'utilisateur est sauvegardé dans une base appelée base de session de recherche.

Pour une session de recherche (figure 3.2), on sauvegarde :

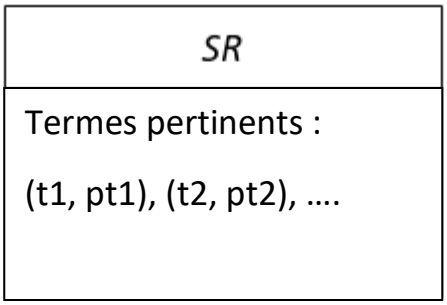


Figure 7: Session de recherche

- L'ensemble de documents retournés par le système de recherche et jugés pertinents par l'utilisateur (longuement visités, imprimés, …etc.)
- L'ensemble de termes pertinents extraits à partir de ces documents

2.2.1 Pondération des termes dans une session de recherche

La pondération des termes aide à définir l'importance du terme dans la session de recherche ou il apparait, et permet d'observer, à travers le temps, la tendance (croissante, stable ou décroissante) du poids du terme pour l'utilisateur en le comparant avec les sessions de recherche précédentes.

Le poids d'un terme t est calculé selon la formule suivante :

$$P(t) = \left(0.5 + 0.5 * {}^{tf}/_{max\ tf}\right) * log({}^{N}/_{n}) \qquad (1)$$

Avec, *tf* est la fréquence totale de terme t dans l'historique de la session de recherche concernée ; *maxtf* est la plus grande fréquence des termes dans l'historique de la même session de recherche ; *N* est le nombre de documents de l'historique de session de recherche ; et *n* est le nombre de documents contenant le terme t dans l'historique de la session.

2.3. Centre d’intérêts

Nous comptons deux types de centre d’intérêts :

2.3.1. Centre d’intérêt à long terme

Apres chaque session de recherche, les nouveaux documents pertinents sont traités pour extraire les termes pertinents dont le poids est calculé par rapport à tout l’historique de recherche du profil d’utilisateur (pondération *tf-idf* du modèle vectoriel mentionnée précédemment).

La collection de termes pertinents représente le premier niveau d'utilisation de l'historique de recherche pour décrire les utilisateurs et leurs intérêts.

Le second niveau d'utilisation des termes pertinents consiste à introduire une pseudo-sémantique basée sur les relations de cooccurrence entre ces termes.

Nous souhaitons mettre en évidence les paires de termes pertinents qui apparaissent ensemble dans de nombreux documents pertinents dans l'historique de recherche de l'utilisateur.

Pour nous, la présence fréquente de deux termes dans de nombreux documents indique qu'ils coopèrent pour décrire le même sujet. Par exemple, l‘apparition de deux termes « éclipse » et « Programmation » peut indiquer qu‘il s‘agit d‘un sujet sur l‘Informatique, tandis que l‘apparition de « éclipse » et « Eclipse de soleil » indique un sujet d’astronomie. Et si « éclipse » figure dans une nouvelle requête utilisateur il serait facile de l‘enrichir par l‘ajout de terme ayant la relation la plus forte de cooccurrence avec « éclipse ».

La valorisation des relations de cooccurrence entre les termes pertinents est représentée par une similarité thématique, notée $Cooccur(t_i, t_j)$, qui exprime la liaison entre chaque terme pertinent t_i avec les autres termes pertinents t_j, et elle est calculée selon la formule suivante :

$$Cooccur(t_i, t_j) = \frac{|d(t_i,t_j)|}{|d(t_i)|+ |d(t_j)|} \qquad (2)$$

Avec :

- $|d(t_i, t_j)|$: nombre de documents contenant $t_i$et $t_j$conjointement dans tout l’historique de recherche.
- $|d(t_i)|$: nombre de documents contenant t_i.

- $|d(t_j)|$: nombre de documents contenant t_j.

La base de relations de cooccurrence des termes pertinents comme l'affiche la figure suivante, sert à sauvegarder les liens de cooccurrence entre les termes pertinents de l'historique de profil utilisateur.

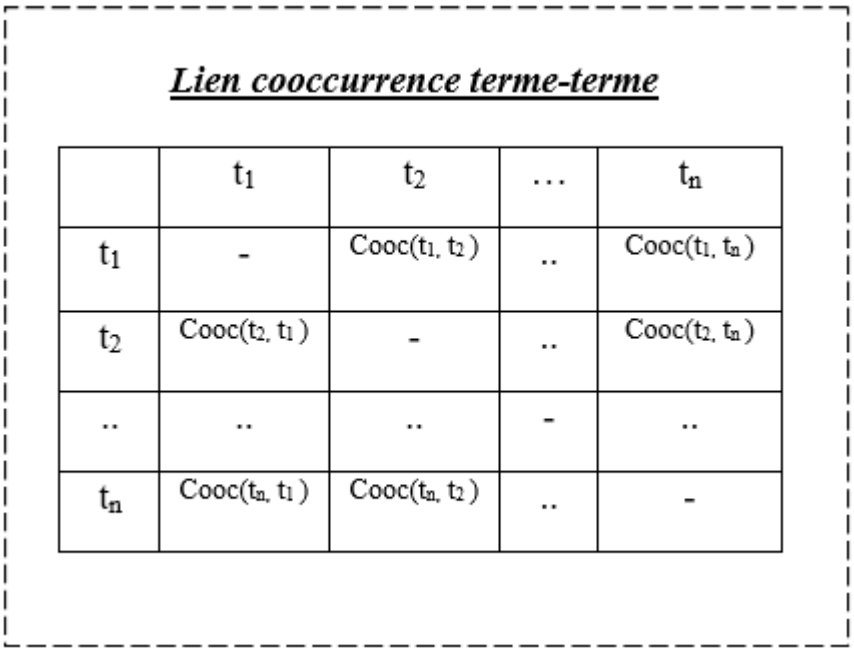


Figure 8: Base de Cooccurrence Terme-Terme

2.3.2. Centre d'intérêts à court terme

Comme c'est défini précédemment dans le deuxième chapitre sur le profil à court terme, on admet que ce profil sert à mieux cibler la recherche vu qu'il contient des données considérées spécifiques et pertinentes au besoin en information courant de l'utilisateur.

L'observation de l'évolution des poids des termes au fil de temps, notamment dans les dernières sessions de recherche, permet de distinguer trois types d'intérêts à court-terme de l'utilisateur selon les tendances : intérêt croissant, intérêt stable et intérêt décroissant.

a) Intérêt croissant.

Si au fil de temps, un terme figure de plus en plus dans les nouvelles sessions de recherche, on observe une augmentation de l'intérêt de l'utilisateur pour le domaine qui est représenté par ce terme (figure 9)

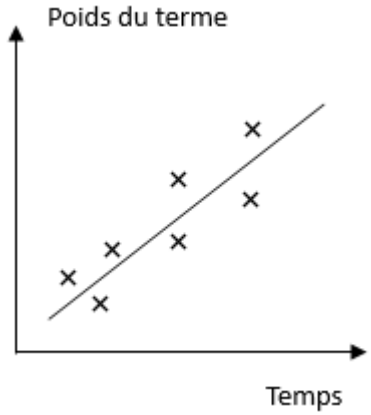


Figure 9 : Intérêt croissant

b) Intérêt stable.

Les préférences des utilisateurs pour un domaine qui évolue peu (rythme stable d'évolution des poids des termes) pendant le temps sont traduites comme des intérêts constants. Une illustration de cette tendance est représentée dans (la figure 10)

c)

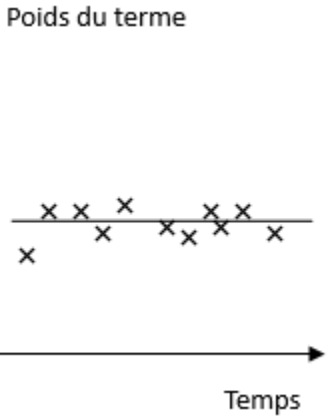


Figure 10: Intérêt stable

c) Intérêt en décroissance.

Cette tendance se traduit par le fait que l'individu perd son intérêt pour le domaine représenté par les termes utilisés au fil de temps (La figure11) illustre cette évolution.

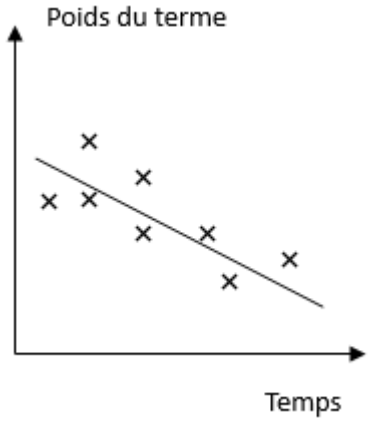


Figure 11 : Intérêt décroissant

Dans le cadre de ce travail, le centre d'intérêts à court terme sert pour regrouper les termes ayant une tendance croissante ou stable de leur poids, et qui représentent par conséquent des intérêts croissants ou stables pendant les dernières sessions de recherche.

2.3.3. Construction et mise à jour du centre d'intérêts à court-terme

Apres chaque session de recherche, les nouveaux documents pertinents sont traités pour extraire les termes pertinents avec leurs poids par rapport à l'historique de la session concernée (pondération du modèle vectoriel mentionnée précédemment *tf.idf*).

Les questions fondamentales qui seront abordées et sur lesquelles se focalise notre approche sont alors les suivantes :

- Comment insérer un nouveau terme dans le centre d'intérêts à court terme (CCT) ?
- Comment mettre à jours les poids des termes existants dans le CCT ?
- Comment retirer un terme du CCT ?

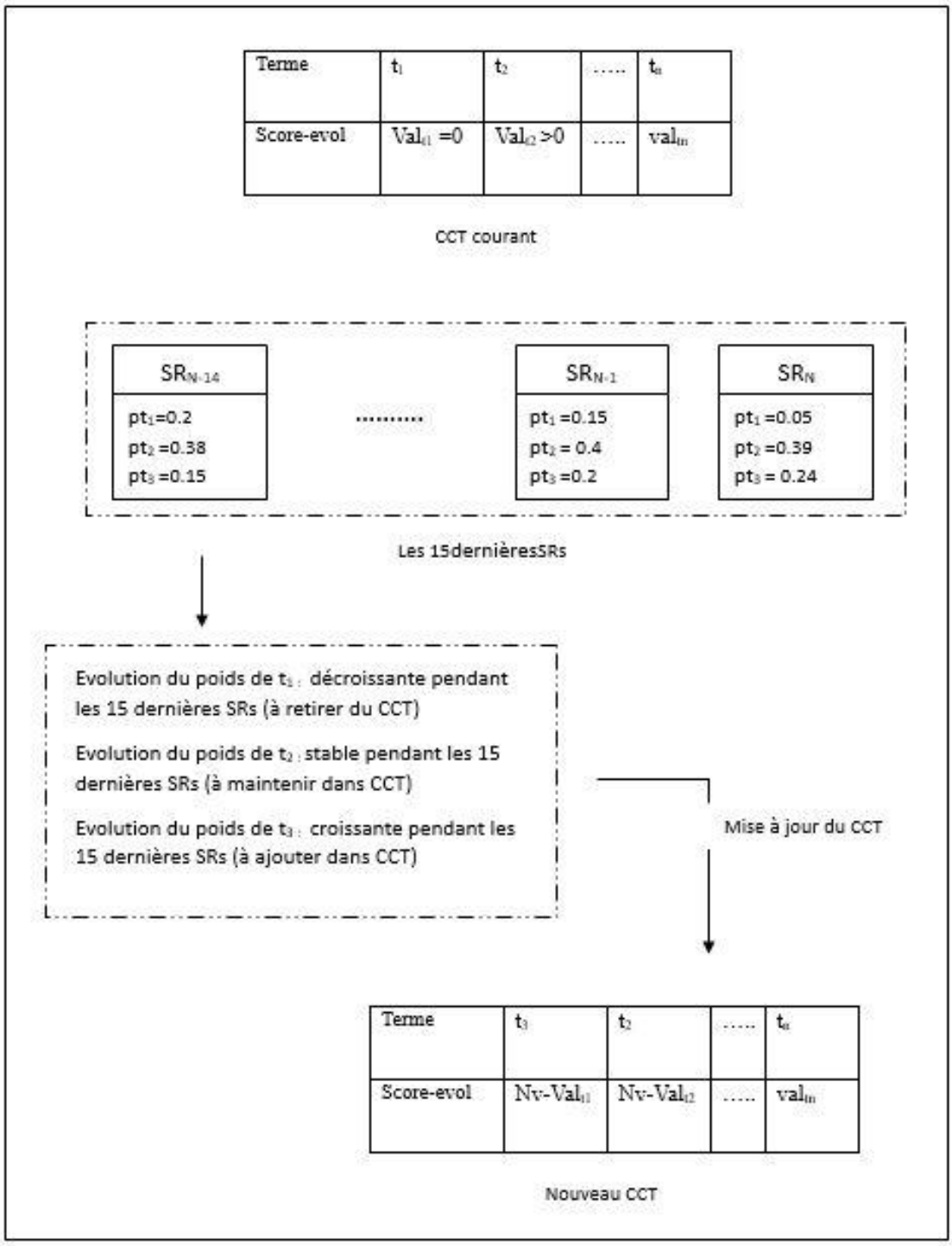


Figure 12 : Construction et mise à jour du CCT

- **Insertion des termes dans le centre d'intérêts à court-terme**

Pour insérer un nouveau terme qui figure dans la dernière SR (Session de Recherche) et qui n'existe pas dans le centre d'intérêts à court-terme, nous calculons la somme des différences des poids de ce terme dans les sept (7) dernières SRs avec le poids moyen (*pm)* dans les même SRs

considérées. Cette valeur, notée $score_evol(t)$, représente le score d'évolution des poids du terme concerné dans les dernières SRs, et elle est calculée selon la formule suivante :

$$score_evol(t) = \sum_{i=N-6}^{N}(p_i(t) - pm\,(t)) \qquad (3)$$

Avec:

- *Pm(t)* représente le poids moyen du terme t dans les 7 dernières SR ;
- $P_i(t)$ représente le poids du terme *t* dans la i[ème] session de recherche.

Ce terme doit répondre à la condition suivante :

$$score_evol(t) \geq 0$$

La satisfaction de cette condition permet l'insertion du terme concerné dans le centre d'intérêts à court-terme avec sa valeur de $score_evol(t)$.

- **Mise à jour des poids des termes existants dans le centre d'intérêts à court-terme**

Pour tous les termes qui existent dans le centre d'intérêts à court-terme et qui figurent dans la nouvelle SR, nous calculons le score d'évolution des poids du terme $score_evol(t)$ pour les quinze (15) dernières SRs.

Dans cette étape deux cas se présentent :

a) En cas de $score_evol(t) \geq 0$, le terme concerné reste dans le centre d'intérêts à court-terme avec mise à jour de sa valeur de $score_evol(t)$.
b) En cas de $score_evol(t) < 0$, le terme est supprimé du centre d'intérêts à court-terme.

- **Epuration périodique du centre d'intérêts à court-terme**

Périodiquement, le centre d'intérêts à court-terme est revu pour retirer les termes qui perdent leurs importances dans les dernières SR. Pour chaque terme qui existent dans le centre d'intérêts à court-terme, nous calculons le score d'évolution des poids des termes $score_evol(t)$ dans les quinze (15) dernières SRs.

Les poids du terme *t* dans les SRs où il ne figure pas sont considérés 0. Tous les termes dont la valeur $score_evol(t) < 0$ sont supprimés du centre d'intérêts à court-terme. Pour les autres termes, la valeur de $score_evol(t)$est mise à jour.

2.4. Exemple d'illustration

- Dans le cas de l'insertion

L'évolution de poids du terme t1 dans les 7 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7
Pt1=0.05 Pt2= … Pt3= …	Pt0= … Pt1=0.07 Pt2= …	Pt0= … Pt1=0.02 Pt3= …	Pt1=0.09 Pt4= … Pt2= …	Pt3= … Pt2= … Pt1=0.08	Pt1=0.1 Pt5= … Pt3= …	Pt1=0.1 Pt4= … Pt0= …

Le poids moyen de t1 est calculé comme suit :

$$pm = \frac{0.05 + 0.07 + 0.02 + 0.09 + 0.08 + 0.1 + 0.1}{7} = 0.072$$

Le score d'évolution de poids de terme t1 est calculé comme suit :

$$score_evol(t_1) = -0.022 - 0.002 - 0.052 + 0.018 + 0.008 + 0.028 + 0.028 = 0.006$$

Alors $score_evol \geq 0$.:

T_1 : rentre dans le CCT.

La figure suivante (*Figure 13*) représente graphiquement l'évolution du poids de terme t1 dans les 7 dernières SRs.

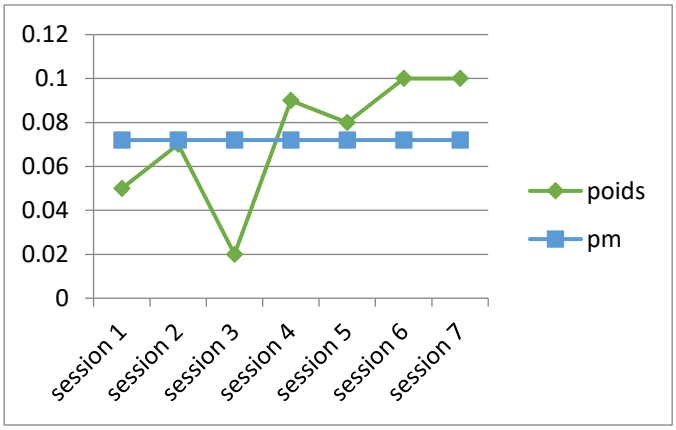


Figure 13: insertion du terme T1 dans CCT.

- En cas de suppression du terme T_2 dans CCT :

L'évolution de poids du terme T2 dans les 15 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7	Sr8
Pt1= … Pt2= 0.1 Pt3= …	Pt2=0.09 Pt4= … Pt0= …	Pt6= … Pt2=0.05 Pt20= …	Pt7= … Pt11= … Pt2=0.03	Pt5= … Pt2=0.02 Pt14= …	Pt2=0.01 Pt19= … Pt10= …	Pt8= … Pt1= … Pt2= 0	Pt2= 0 Pt9= … Pt12= …

Sr9	Sr10	Sr11	Sr12	Sr13	Sr14	Sr15
Pt14= … Pt2= 0.03 Pt3= …	Pt2=0.015 Pt5= … Pt22= …	Pt7= … Pt16= … Pt2= 0	Pt16 … Pt18= … Pt2=0.009	Pt2=0.006 Pt10= … Pt17= …	Pt8= … Pt2=0.005 Pt11= …	Pt2=0.001 Pt20= … Pt15= …

Le poids moyen de T2 est calculé comme suit :

$$pm = \frac{0.1 + 0.09 + 0.05 + 0.03 + 0.02 + 0.01 + 0 + 0.03 + 0.015 + 0 + 0.009 + 0.006 + 0.005 + 0.001}{15}$$

$$= 0.0244$$

Le score d'évolution de poids de terme T2 est calculé comme suit :

$$score_{evol} = 0.0756 + 0.0656 + 0.0256 + 0.0056 - 0.0044 - 0.0144 - 0.0244 - 0.0244 + 0.0056 - 0.0094 - 0.0244 - 0.0154 - 0.0184 - 0.0194 - 0.0234 = 0$$

Alors : $score_evol \geq 0$

Donc T2 reste dans le CCT.

La figure suivante (*Figure 14*) représente graphiquement l'évolution du poids de terme T2 dans 15 SRs avant la suppression.

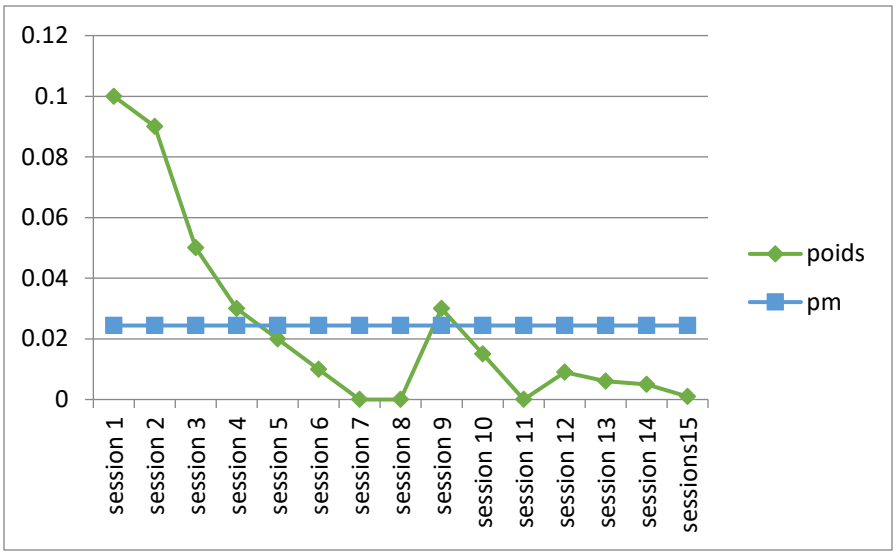


Figure 14 : Evolution du terme t2 dans CCT avant la suppression

- L'évolution de poids du terme T2 dans les 15 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7	Sr8
Pt8= … Pt13= … Pt2= 0.09	Pt2= 0.05 Pt10= … Pt21= …	Pt1= … Pt2= 0.03 Pt3= …	Pt5= … Pt9= … Pt2= 0.02	Pt14= … Pt2= 0.01 Pt20= …	Pt11= … Pt2= 0 Pt9= …	Pt2= 0 Pt7= … Pt3= …	Pt8= … Pt16= … Pt2= 0.03

Sr9	Sr10	Sr11	Sr12	Sr13	Sr14	Sr15
Pt2= 0.015 Pt17= … Pt25= …	Pt7= … Pt2= 0 Pt0= …	Pt1= … Pt2= 0.009 Pt4= …	Pt2= 0.006 Pt6= … Pt3= …	Pt18= … Pt2= 0.005 Pt23= …	Pt30= … Pt6= … Pt2= 0.001	Pt7= … Pt2= 0.001 Pt4= …

Le poids moyen de T2 est calculé comme suit :

$$pm = \frac{0.09 + 0.05 + 0.03 + 0.02 + 0.01 + 0 + 0 + 0.03 + 0.015 + 0 + 0.009 + 0.006 + 0.005 + 0.001 + 0.001}{15}$$

$$= 0.0178$$

Le score d'évolution de poids de terme T2 est calculé comme suit :

$$score_{evol} = 0.0722 + 0.0322 + 0.0122 + 0.0022 - 0.0078 - 0.0178 - 0.0178 + 0.0122 - 0.0028 - 0.0178 - 0.0088 - 0.0118 - 0.0128 - 0.0169 - 0.0169 = -0.0002$$

Alors : $score_evol < 0$

Donc T2 a retiré du CCT.

La figure suivante (*Figure 15*) représente graphiquement l'évolution du poids de terme T2 dans les 15 dernières SRs.

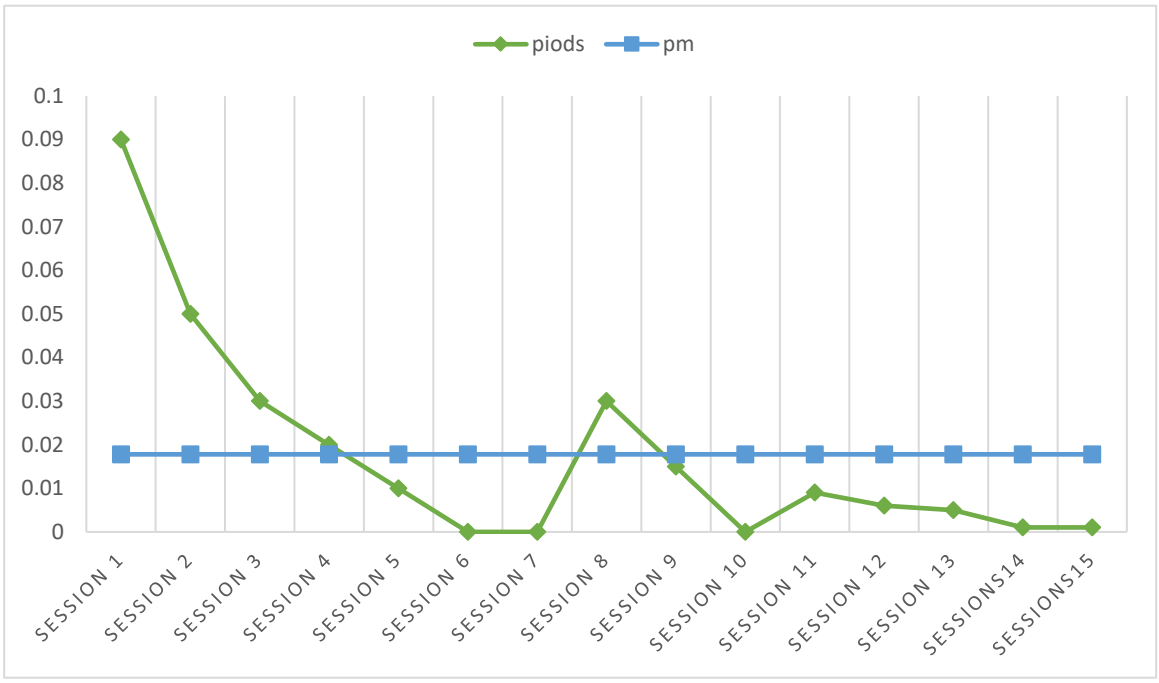


Figure 15 : suppression de terme t2 dans CCT

- En cas du terme T3 dans CCT :

L'évolution de poids du terme T3 dans les 15 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7	Sr8
Pt11= … Pt17= … Pt3= 0.2	Pt3= 0.19 Pt8= … Pt5= …	Pt19= … Pt3= 0.23 Pt24= …	Pt3= 0.17 Pt5= … Pt4= …	Pt0= … Pt3= 0.2 Pt15= …	Pt25= … Pt6= … Pt3= 0.15	Pt3= 0.17 Pt18= … Pt2= …	Pt1= … Pt3= 0.21 Pt8= …

Sr9	Sr10	Sr11	Sr12	Sr13	Sr14	Sr15
Pt5= … Pt3= 0.25 Pt19= …	Pt26= … Pt20= … Pt3= 0.22	Pt3= 0.18 Pt12= … Pt27= …	Pt9= … Pt29= … Pt3= 0.15	Pt18= … Pt3= 0.15 Pt27= …	Pt5= … Pt3= 0.18 Pt0= …	Pt3= 0.2 Pt1= … Pt7= …

Le poids moyen de T3 est calculé comme suit :

$$\boldsymbol{pm} = \frac{\mathbf{0.2 + 0.19 + 0.23 + 0.17 + 0.2 + 0.15 + 0.17 + 0.21 + 0.25 + 0.22 + 0.18 + 0.15 + 0.15 + 0.18 + 0.2}}{\mathbf{15}}$$

$$= \mathbf{0.19}$$

Le score d'évolution de poids de terme T3 est calculé comme suit :

$$score_{evol} = 0.01 + 0 + 0.04 + 0.01 - 0.02 - 0.04 - 0.02 + 0.02 + 0.06 + 0.03 - 0.01 - 0.04 - 0.04 - 0.01 + 0.01 = 0$$

Alors $scor_evol \geq 0$, donc T_3 reste dans le CCT.

La figure suivante (*Figure 16*) représente graphiquement l'évolution du poids de terme T2 dans les 15 dernières SRs.

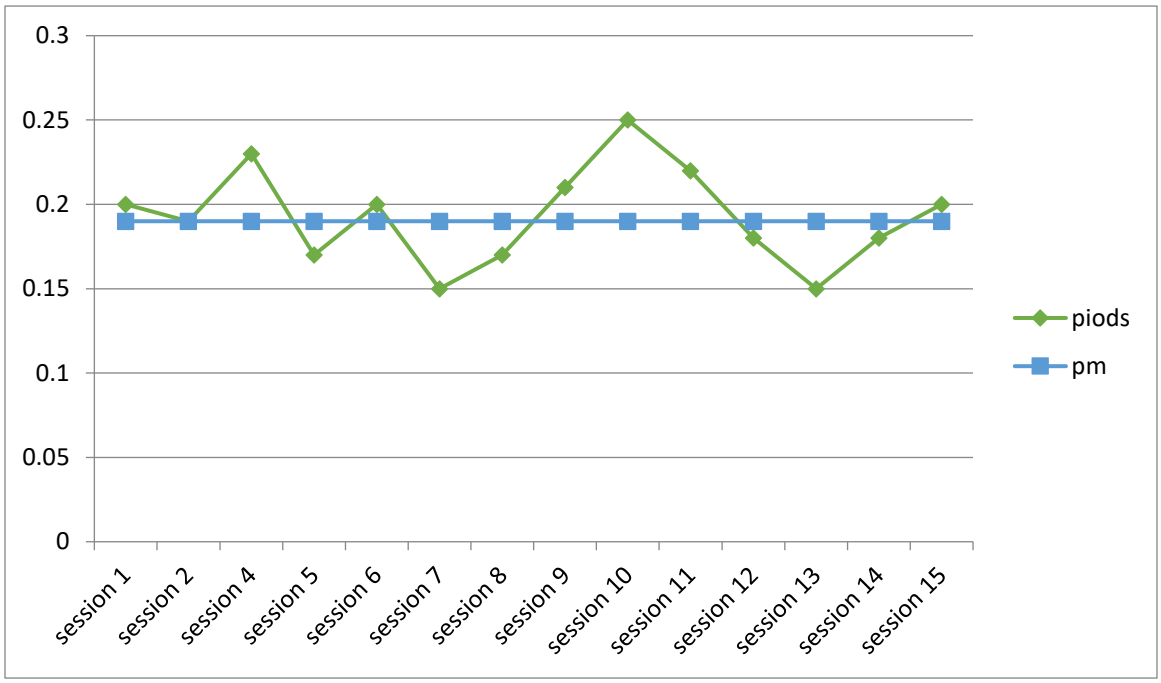


Figure 16: maintenance du terme t3 dans CCT

- Le cas d'épurations périodique du terme T4dans CCT :

L'évolution de poids du terme T4dans les 15 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7	Sr8
Pt5= … Pt4= 0.11 Pt19= …	Pt6= … Pt12= … Pt4= 0.1	Pt0= … Pt4= 0.05 Pt1= …	Pt4= 0.03 Pt9= … Pt21= …	Pt5= … Pt14= … Pt4= 0.02	Pt17= … Pt5= … Pt4= 0.01	Pt4= 0 Pt5= … Pt19= …	Pt7= … Pt4= 0 Pt11= …

Sr9	Sr10	Sr11	Sr12	Sr13	Sr14	Sr15
Pt6= … Pt4= 0.015 Pt31= …	Pt8= … Pt14= … Pt4= 0.009	Pt4= 0 Pt18= … Pt9= …	Pt26= … Pt10= … Pt4= 0.015	Pt1= … Pt20= … Pt4= 0.007	Pt4= 0.006 Pt14= … Pt2= …	Pt4= … Pt27= … Pt4= 0

Le poids moyen de T4 est calculé comme suit :

$$\boldsymbol{pm} = \frac{\mathbf{0.1 + 0.1 + 0.05 + 0.03 + 0.02 + 0.01 + 0 + 0 + 0.015 + 0.009 + 0 + 0.015 + 0.007 + 0.006 + 0}}{\mathbf{15}}$$

$$= \mathbf{0.0241}$$

Le score d'évolution de poids de terme T5 est calculé comme suit :

$$scor_{evol} = 0.0759 + 0.0756 + 0.0259 + 0.0059 - 0.0041 - 0.0241 - 0.0243 - 0.0091 - 0.0151 - 0.0241 - 0.0091 - 0.0171 - 0.0181 - 0.0241 = 0.0141$$

Alors $scor_evol \geq 0$, donc T4 reste dans le CCT.

La figure suivante (*Figure 17*) représente graphiquement l'évolution du poids de terme T4 dans les 15 dernières SRs.

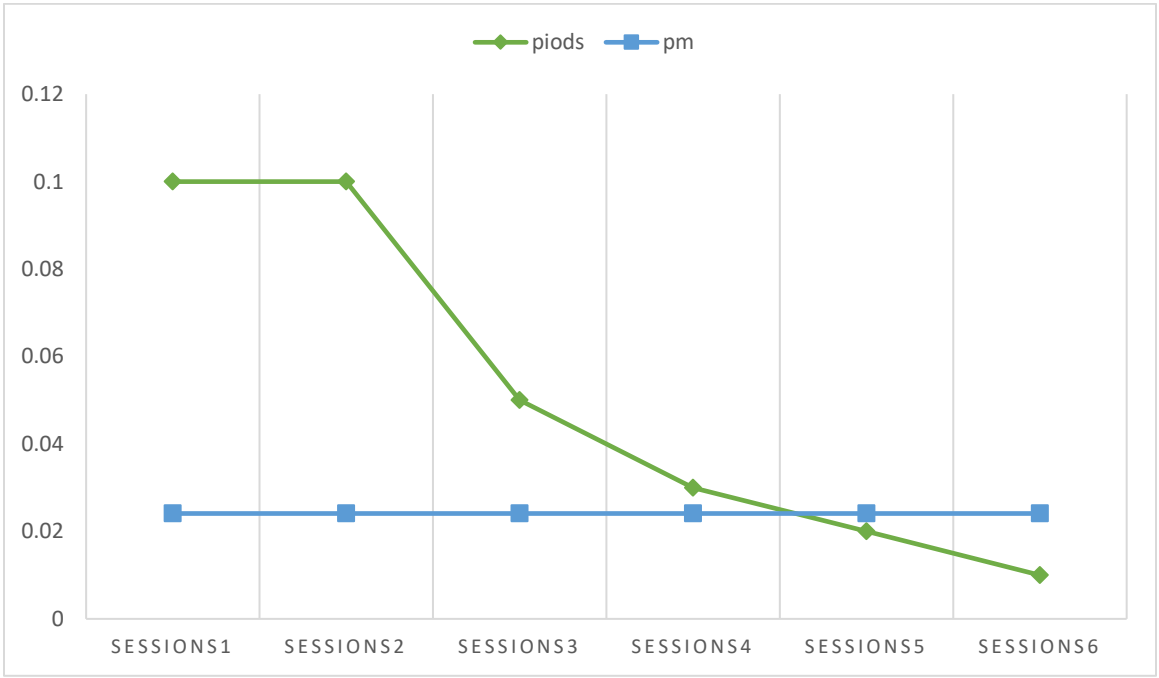


Figure 17: T4 reste dans CCT

L'évolution de poids du terme T5dans les 15 dernières SRs est mentionnée dans ce tableau :

Sr1	Sr2	Sr3	Sr4	Sr5	Sr6	Sr7	Sr8
Pt15= … Pt5= 0.1 Pt19= …	Pt20= … Pt9= … Pt5= 0.05	Pt5= 0.03 Pt6= … Pt1= …	Pt5= 0.02 Pt16= … Pt3= …	Pt22= … Pt5= 0.01 Pt10= …	Pt2= … Pt5= 0 Pt11= …	Pt5= 0 Pt16= … Pt9= …	Pt5= 0.015 Pt6= … Pt9= …

Sr9	Sr10	Sr11	Sr12	Sr13	Sr14	Sr15
Pt18= … Pt5= … Pt5= 0.009	Pt5= 0 Pt18= … Pt9= …	Pt5= 0.006 Pt20= … Pt4= …	Pt5= 0.003 Pt11= … Pt14= …	Pt2= … Pt5= 0.001 Pt0= …	Pt21= … Pt5= 0 Pt10= …	Pt2= … Pt5= 0 Pt0= …

Le poids moyen de T5 est calculé comme suit :

$$pm = \frac{0.1 + 0.05 + 0.03 + 0.02 + 0.01 + 0 + 0 + 0.015 + 0.009 + 0 + 0.006 + 0.003 + 0.001 + 0 + 0}{15}$$

$$= 0.0162$$

Le score d'évolution de poids de terme T5 est calculé comme suit :

$$scor_{evol} = 0.0837 + 0.0337 + 0.0034 - 0.0062 - 0.0162 - 0.0162 - 0.0012 - 0.0072 - 0.0162 - 0.0102 - 0.0132 - 0.0152 - 0.0162 - 0.0162 = -0.001$$

Alors $scor_evo < 0$, donc T5 à retirer du CCT.

La figure suivante (*Figure 18*) représente graphiquement l'évolution du poids de terme T5 dans les 15 dernières SRs.



Figure 18: Suppression de terme T5 dans CCT

3 Processus de recherche

Notre processus de recherche suit les étapes suivantes illustrées dans la figure 3.14:

- Le système reçoit la requête utilisateur Q.

- Le système sélectionne le centre d'intérêts le plus correspondant à la requête parmi les deux centres à long-terme ou à court-terme

- Selon le centre d'intérêts sélectionné, la requête est reformulée pour produire une requête reformulée Q'.

- la requête reformulée Q' est envoyée au moteur de recherche classique.

- Le système observe les réactions et les comportements de l'utilisateur avec les résultats de recherche retournés pour mettre à jour son profil.

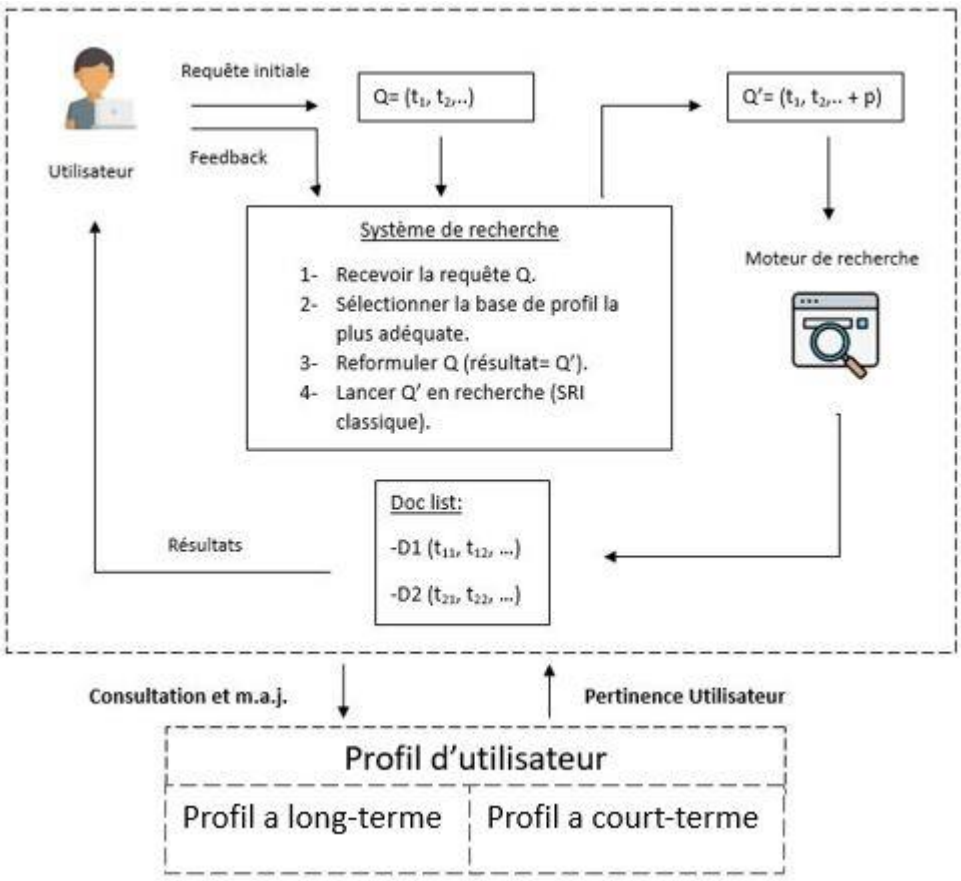


Figure 19 : processus de recherche

3.1. Sélection de centre d'intérêts le plus adéquat à la requête

L'objectif de cette étape est de choisir parmi les deux centres d'intérêts de l'utilisateur, ce qui est le plus proche au sujet de la requête et qui servira mieux à sa reformulation.

Pour réaliser cette sélection, notre système calcule des scores de similarité entre la requête et chacun des deux centres d'intérêts :

$$sim(CCT, Q) = \left(\sum_{ti \in Q} \sum_{tj \in CCT} cooccur(ti, tj)\right) \qquad (4)$$

$$sim(CLT, Q) = \log\left(\sum_{ti \in Q} \sum_{tj \in CLT} cooccur(ti, tj)\right) \qquad (5)$$

L'application de ces formules exige que le contenu de CCT doit représenter au minimum 0.5 % de CLT. Par exemple si le CLT compte 5000 termes, le CCT doit compter au moins 25 termes.

Le centre d'intérêts dont la similarité avec la requête est max sera choisi comme contexte courant de la recherche et sera utiliser pour reformuler la requête.

3.2. Reformulation de la requête

L'objectif de la technique de reformulation des requêtes par retour de pertinence est de générer de nouvelles requêtes plus adaptées aux besoins d'information des utilisateurs. La reformulation de la requête consiste à ajouter de nouveaux termes à la requête initiale ou à répondre à ses terme sa fin de cibler les recherches sur les documents pertinents. La génération de nouvelles requêtes se fait soit par reformulation automatique, soit par reformulation interactive.

En respectant ce principe, nous allons enrichir les requêtes des utilisateurs Q pour générer de nouvelles requêtes Q' en ajoutant des termes en fonction du contexte de recherche actuel exprimé par les centres d'intérêts sélectionnés (CCT ou CLT).

Ainsi, selon le centre d'intérêts choisi, une nouvelle requête reconstruite Q' est obtenue à partir de la requête initiale Q comme suit :

$$\boldsymbol{Q' = Q + T}$$

Avec T est le terme qui maximise la relation de co-occurrence avec les termes de la requête. En cas d'égalité, on peut différencier entre les termes par le poids ou la somme de différence des poids

4 Présentation de l'expérimentation

4.1. Création des profiles

Vu l'absence d'un cadre d'évaluation dédié au contexte de personnalisation, nous avons construit notre cadre d'évaluation comme suit :

Nous avons construit 5 profils pour 5 utilisateurs appartenant à des domaines professionnels différents pour avoir des centres d'intérêts différents. Nous avons construit chaque profil avec la collaboration de l'utilisateur concerné. Les poids des termes et les liens de cooccurrence entre eux sont calculés selon une simulation.

Dans le tableau suivant (Tableau N 3.1), nous présentons quelques exemples des profils d'utilisateurs et leurs centres d'intérêts (CCT, CLT) :

Tableau N 3.1 Exemples de profils conçus

Utilisateur	CCT	CLT
U1 : étudiant en informatique	Réseau, information, informatique, personnalisée, technologie, matériel, système, Smartphone, Apple, Androïde, protocole, sécurité, recherche, panne, processus, logiciel, application,	Informatique, logiciel, langage, PHP, SQL, programmation, java, éclipse, SGBD, application, algorithme, information, système, recherche d'information, sécurité, vulnérabilité
U2 : étudiant en math	Matlab, algèbre,statistique, analyse,proba, équations différentielles,géométrie,logique,symétrie,précision, produit, scalaire, champ, numérique ,méthode , Formule.	Exponentielle, intégral, nombre, théorème, Euler, fonction, euclidien, Galilée, calcule, logique, produit, scalaire.
U3 : étudiant en historique	époque, biographie, géologie patrimoine, révolution, seiche, idéalisme, muséum, culture, satellite, spatiale, peuple.	Capitalisme, guerre, coloniale, témoin, civilisation, littérature, héritage, événement, , satellite, spatiale,

		peuple.
U4 : étudiant en biologie	Bactéries, virus, microscope, protozoaires, mycètes, procaryotes, eucaryotes, algues, ADN, hépatite	Poliomyélite, antibiotiques, ARN, écologisme, ADN circulaire, anti-hormone , adénine,
U5 : étudiant en économie	Finance, marché, entreprise, croissance, production, consommation, investissement, ressource, commerce, monnaie	Budget, économiste, emploi, crise, développement, échange, industrie, fiscalité, Inflation, PIB(Produit Intérieur Brut)
…..	…..	…..
…..	…..	…..

Les données des profils sont organisées comme c'est indiqué dans la figure suivante :

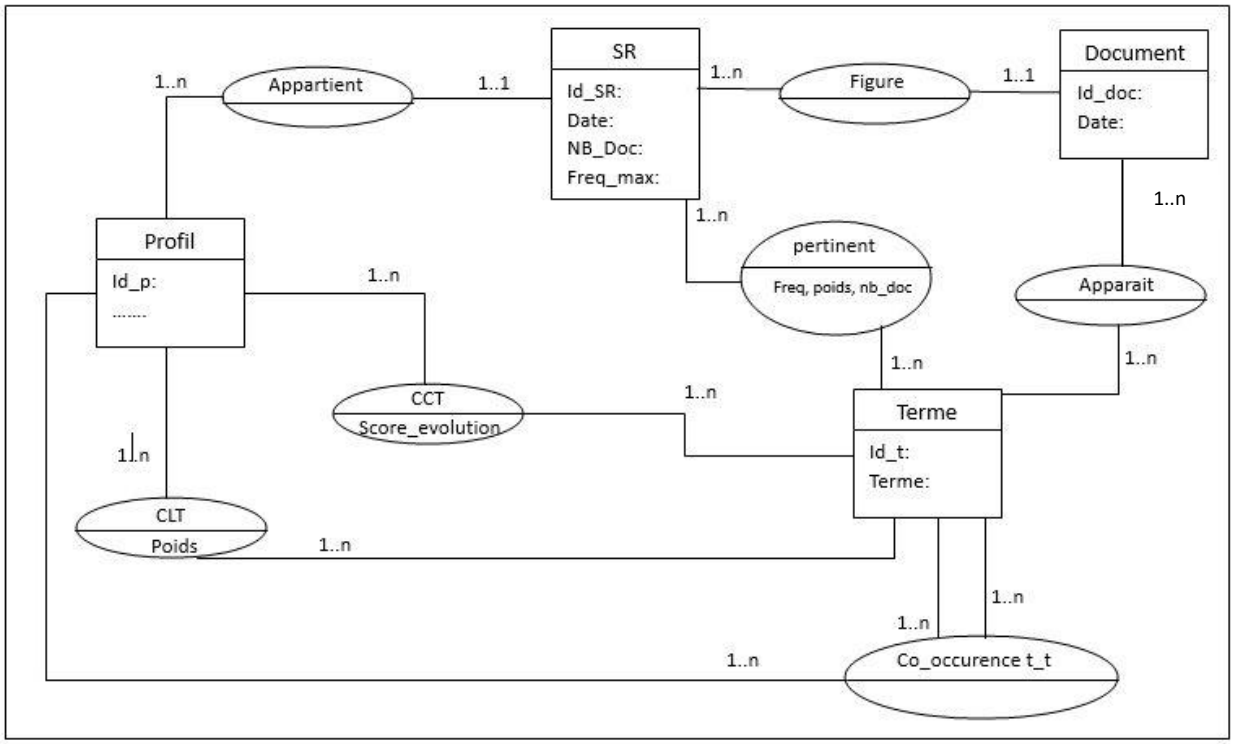


Figure 20 : Base de données

4.2. Création de la BD

4.2.1. Le système de gestion de bases de données MySQL

MySQL est un serveur de base de données relationnelle, basé sur le langage de requête SQL, il est implémenté sur un mode client/serveur, et se caractérise par sa rapidité et sa facilité d'utilisation.

4.2.2. PhpMyAdmin

L'outil phpMyAdmin est développé en PHP (ensemble de scripts PHP), il offre une interface graphique pour l'administration des bases de données MYSQL via un navigateur Web. Les fonctions principales de PhpMyAdmin sont:

- Création de nouvelles bases de données.
- Création/suppression/modification des tables.
- L'édition, l'ajout et la suppression de champs.
- L'exécution de commandes SQL et de requêtes.

Quelques exemples des tables :

Tables de CLT :

id_u	id_t	freq	nb_doc	poids
1	12	10	5	0.66
1	13	15	6	0.69
1	14	17	8	0.623
1	15	13	7	0.611
1	16	20	10	0.582
1	17	21	12	0.526
1	18	26	15	0.488
1	19	30	13	0.585
1	20	15	4	0.822
1	21	18	6	0.736
1	22	25	10	0.64
1	23	20	11	0.54
1	1	10	9	0.496
1	9	9	9	0.484

Figure 21:Table de CLT

Table de sessions de rechreche :

id_s	freq_max_s	nb_doc_s	id_u	date_s
7	8	20	1	2023-05-01
8	12	29	1	2023-05-02
9	7	15	1	2023-05-03
10	9	18	1	2023-05-04
11	14	25	1	2023-05-05
12	16	30	1	2023-05-06
13	11	22	1	2023-05-07
14	10	20	1	2023-05-08
15	5	10	1	2023-05-09
16	6	17	1	2023-05-10
17	13	26	1	2023-05-11
18	11	21	1	2023-05-12
19	15	31	1	2023-05-13
20	8	16	1	2023-05-14
21	10	24	1	2023-05-15

Figure 22: table de sessions

Table de CCT :

id_u	id_f	score_evolution
1	1	0.03
1	2	0.015
1	3	0.01
1	5	0.04
1	6	0.1
1	7	0.019
1	8	0.15
1	9	0.05
1	10	0.09
1	11	0

Figure 23 : Table de CCT

5 Presentation de l'application

Pour effectuer ses recherches, l'utilisateur a besoin d'une interface utilisateur où il peut interagir avec le système de recherche, cela nécessite des différentes étapes qui sont expliquées ci-dessous :

a. Lorsque l'utilisateur fait son login, une interface d'utilisateur apparait contenant une barre de recherche où il peut taper sa requête. Pour faire ça, l'utilisateur a 2 choix : (1) faire une recherche classique en cliquant sur le buton « Classic », (2) faire une recherche personnalisée en cliquant sur le buton « Personalized ».
b. Le clic sur le buton « Classic » permet d'afficher les premiers résultats classiques basés seulement sur la requête de l'utilisateur.
c. Le buton « Personalized » permet d'afficher une petite fenêtre où laquelle on trouve les valeurs de similarités entre la requête d'utilisateur et ses centres d'intérêts, cela permet de nous informer du centre d'intérêts qui a été choisi pour reformuler la requête.
d. Le clic sur le buton « Continue » permet d'afficher une autre barre de recherche contenant la requête reformulée.
e. Finalement, le buton « Search » sert à guider la recherche vers le moteur de recherche Google pour afficher les nouveaux résultats de la nouvelle requête.

Nous donnons un exemple qui illustre les étapes mentionnées avant :

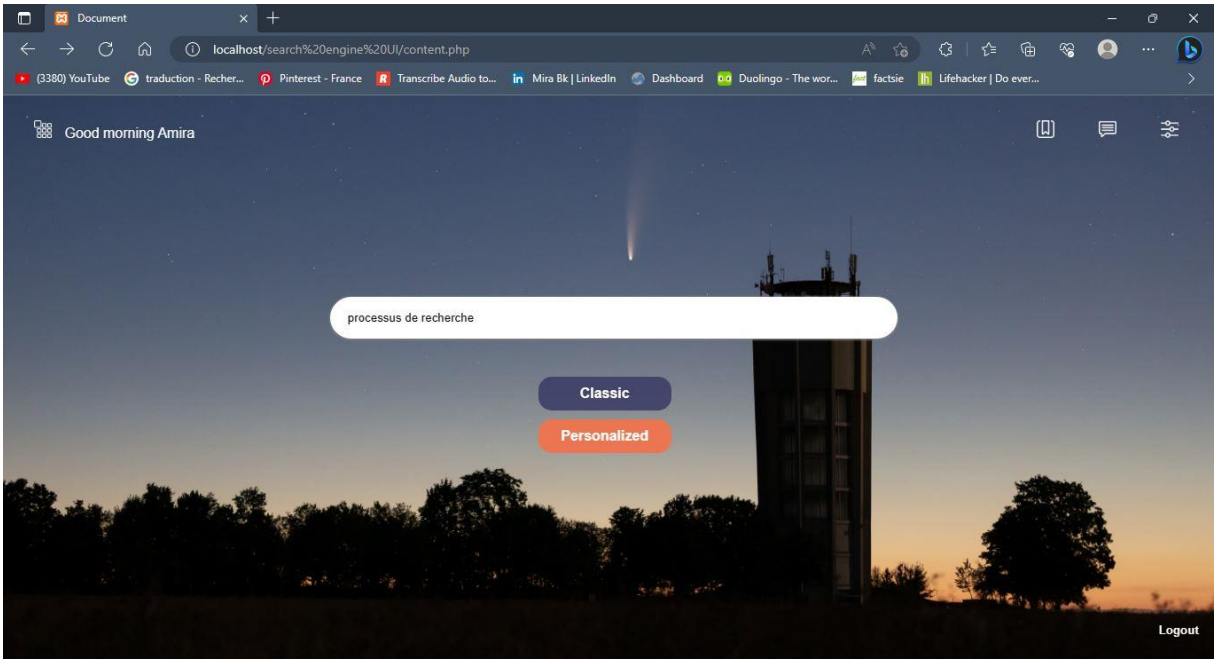


Figure 24: Interface de recherche

Dans cet exemple, la requête de l'utilisateur est définie comme « processus de recherche », commençant avec la recherche classique l'utilisateur clic sur le buton « Classic », les résultats sont les suivants :



Figure 25: Résultat de la recherche classique (1)



Figure 26: Résultat de la recherche classique (2)

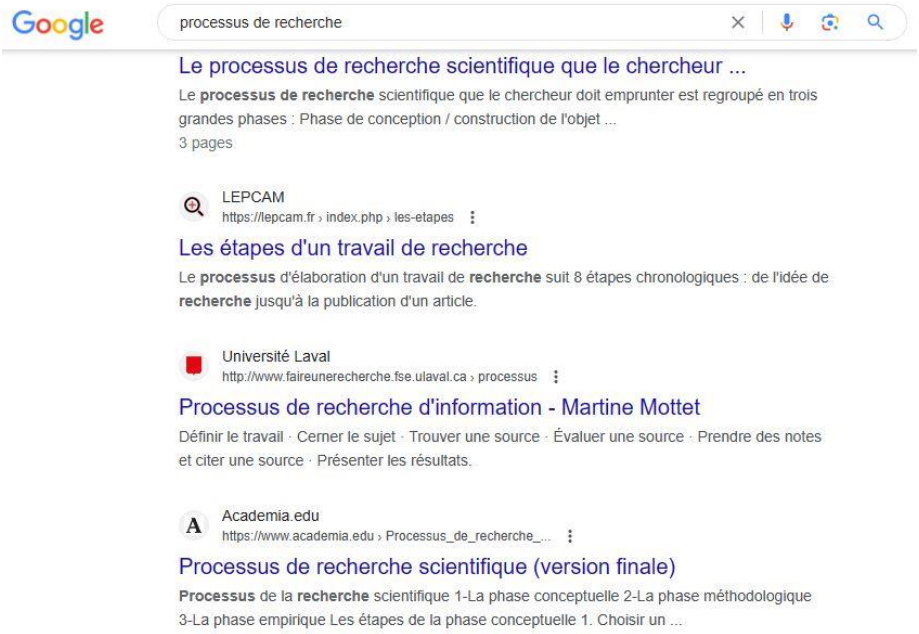


Figure 27 : Résultat de la recherche classique (3)

Voyons que la requête « processus de recherche » est ambiguë selon les résultats fournis par le moteur de recherche Google qui nous a retourné des documents de différents domaines (processus de recherche marketing, processus de recherche scientifique et processus de recherche d'information) pour un utilisateur (U1) dont le centre d'intérêts parle de l'informatique.

Pour avoir une recherche mieux personnalisée (utilisant la même requête), l'utilisateur clic sur le buton « Personalized », la figure suivante illustre ce processus :

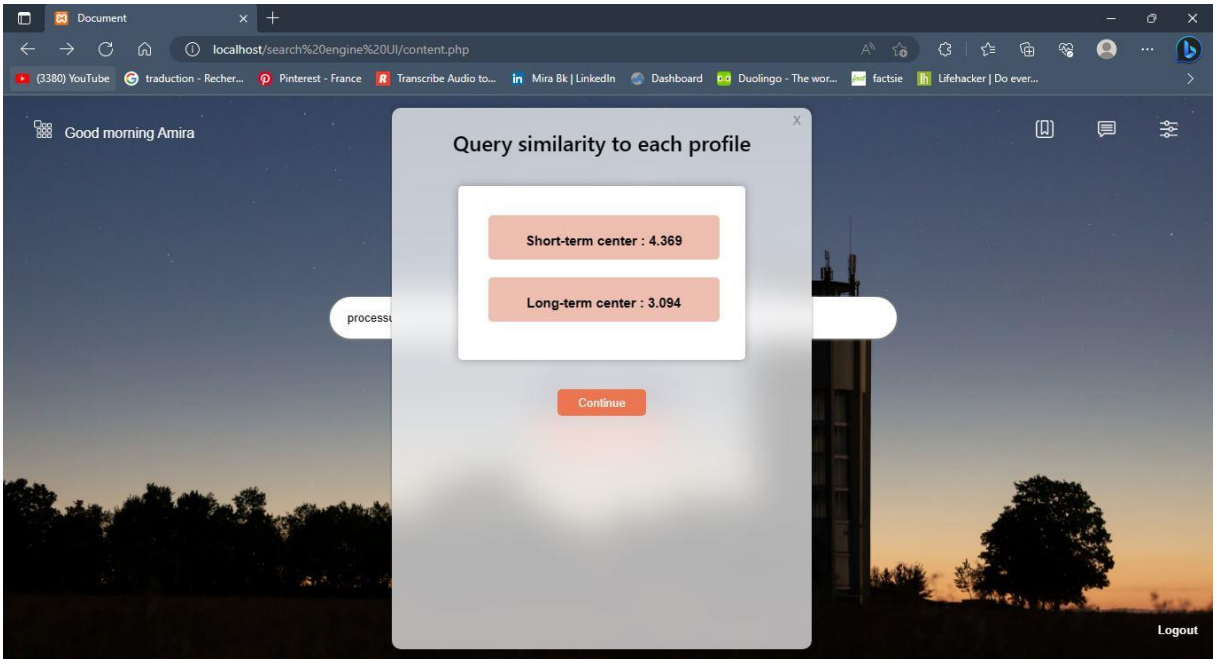


Figure 28 : Similarité entre la requête et les centres d'intérêts

Dans ce cas, nous remarquons que la requête est plus similaire avec le contenu du centre d'intérêts à court-terme et ça nous dit que cette requête appartient aux intérêts courants de l'utilisateur. Nous déduisons alors que le centre d'intérêts à court-terme est le plus adéquat pour reformuler la requête.

Pour obtenir la nouvelle requête l'utilisateur clic sur le buton « Continue », l'affichage de la requête reformulée est le suivant :

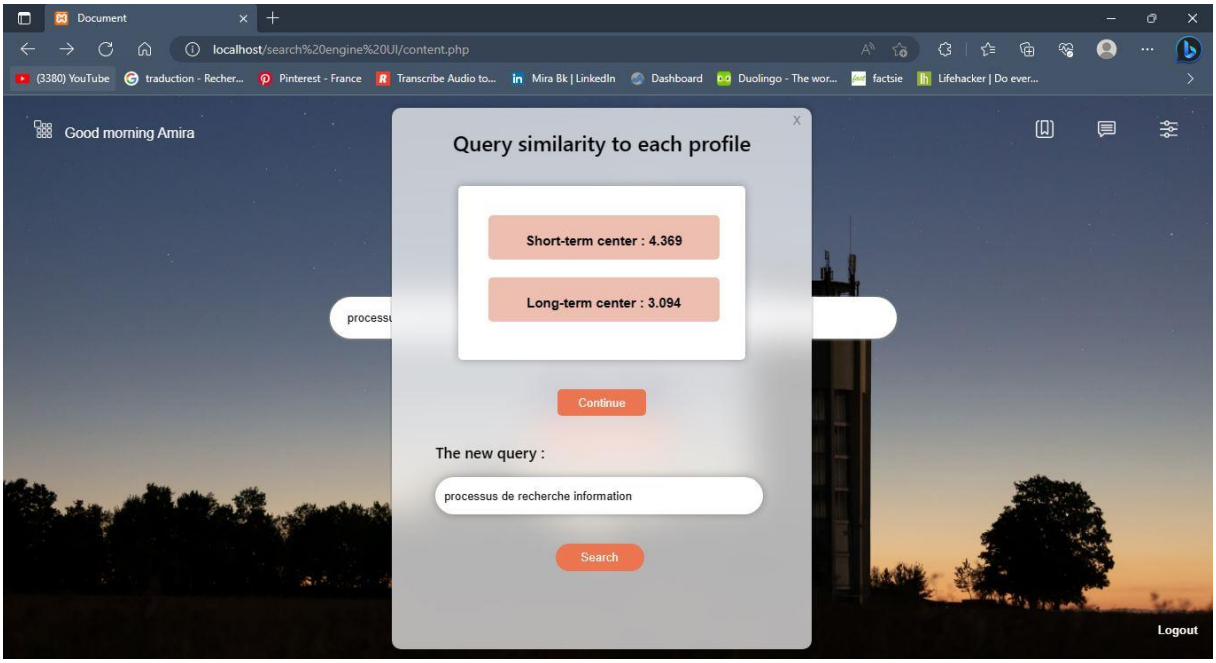


Figure 29 : La requête reformulée.

La requête initiale de l'utilisateur a été reformulée par l'injection du terme « information » pour guider la recherche de l'utilisateur d'une façon qu'il aura les résultats les plus pertinents.

Comme dernière étape, l'utilisateur clic sur le buton « Search » pour envoyer la requête reformulée au moteur de recherche Google et les résultats se présentent ci-dessous :

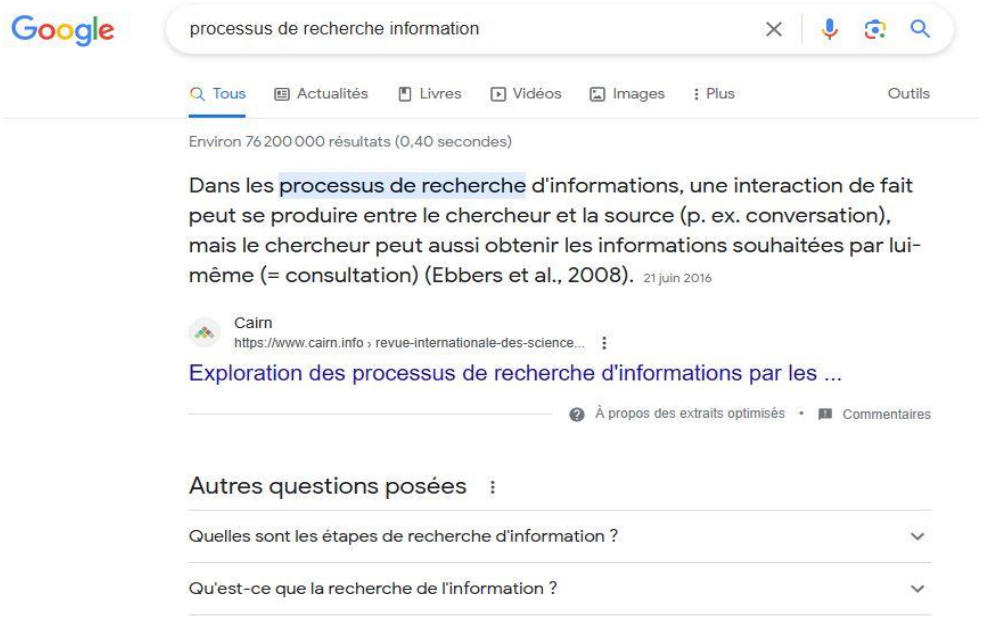


Figure 30 : Résultat de la recherche personnalisée (1)



Figure 31 : Résultat de la recherche personnalisée (2)

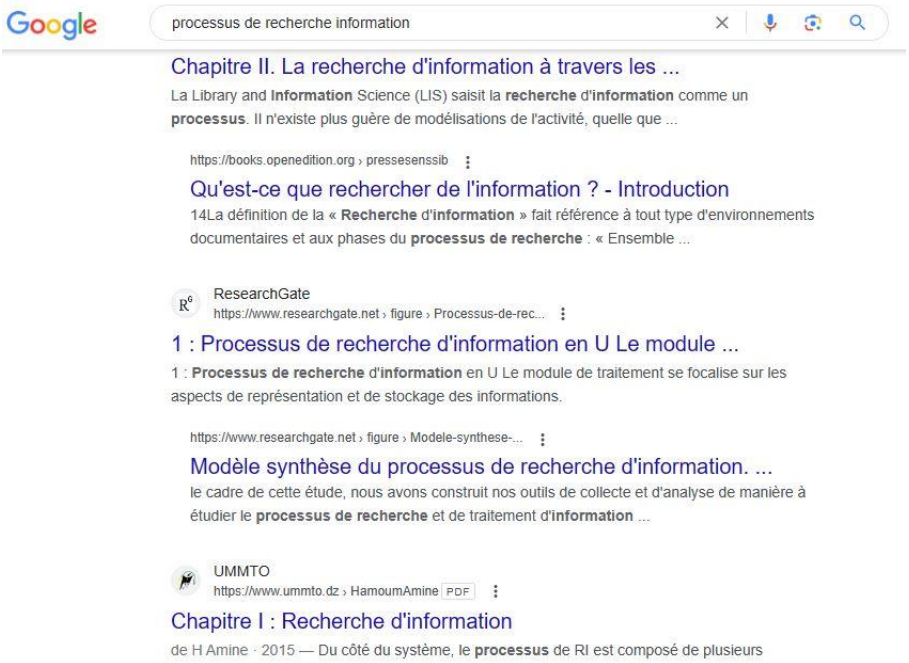


Figure 32 : Résultat de la recherche personnalisée (3)

Nous pouvons maintenant comparer entre ces résultats et les résultats précédents de la recherche classique (sur la même requête), il est clair que les résultats ont été orientés de plus vers un seul domaine (la recherche d'information) qui doit répondre aux besoins en information de l'utilisateur (U1) intéressé de l'informatique.

6 Conclusion

Dans ce chapitre nous avons présenté notre approche de personnalisation de recherche. Nous avons construit un profil d'utilisateur qui représente deux centres d'intérêt (CTL, CCT). Un centre d'intérêts à court-terme dont l'objectif est de mettre en évidence les intérêts courants de l'utilisateur, et un centre d'intérêts à long-terme qui permet de suivre les intérêts de l'utilisateur tout au long de son historique de recherche.

Une pseudo-sémantique basée sur les liens de cooccurrence entre les termes pertinents de profil est introduite. Comme nous avons proposé plusieurs formules pour calculer les différents poids et les scores de similarité.

Conclusion générale et perspectives

Conclusion générale et perspectives

Le travail présenté dans ce mémoire se situe dans le contexte général de l'accès personnalisé a l'information et plus précisément dans la recherche d'information personnalisée a l'aide de profils d'utilisateurs a court-terme.

Nos objectifs de recherche initiaux étaient de comprendre comment exploiter efficacement les centres d'intérêts à court-terme des utilisateurs pour améliorer la pertinence des résultats de recherche en se concentrant sur les besoins immédiats des utilisateurs plutôt que sur des informations datées.

En soulignant la contribution de ce mémoire, nous avons identifié une nouvelle approche pour la personnalisation de la recherche d'information basée sur l'exploitation des centres d'intérêts à court-terme en mettant l'accent sur l'importance du terme sans se fier à la dimension temporelle. La construction du profil utilisateur a court-terme consiste à collecter les documents jugés pertinents implicitement pour un ensemble de requêtes liées à un même besoin en information définissant ainsi une session de recherche et ensuite les exploiter pour extraire les termes les plus pertinents par rapport à leurs poids. L'évolution du profil à court-terme est basée sur l'augmentation de l'importance des termes récurrents dans la session de recherche, les centres d'intérêts à court-terme permettent ainsi de définir le centre d'intérêts à long-terme. La personnalisation du processus de recherche consiste finalement à la reformulation des requêtes d'utilisateurs utilisant le centre d'intérêts le plus adéquat à la requête.

Cette approche présente plusieurs avantages, notamment en permettant de mieux répondre aux besoins spécifiques des utilisateurs en fournissant des résultats adaptés à leurs centres d'intérêts actuels et en réduisant la surcharge informationnelle liées à l'inclusion du facteur temporel.

Cependant, notre recherche a également mis en évidence des défis importants à prendre en compte dans l'exploitation des centres d'intérêts à court terme. La compréhension précise des centres d'intérêts et la capture des préférences changeantes des utilisateurs constituent des défis complexes. De plus, il est essentiel de maintenir un équilibre entre la personnalisation et la diversité des résultats, afin de garantir une expérience d'exploration et de découverte satisfaisante pour les utilisateurs.

En définitive, l'exploitation des centres d'intérêts à court terme dans la recherche d'information personnalisée offre des possibilités considérables pour améliorer l'expérience de recherche des utilisateurs. Cependant, des recherches supplémentaires sont nécessaires pour résoudre les défis techniques et pour évaluer l'impact réel de cette approche sur l'expérience utilisateur.

Comme perspectives, ce travail peut être suivi par les sujets de recherche suivants :

- Etudier la stabilité et l'évolution dans le temps des intérêts de l'utilisateur, pour estimer à quel point son profil peut être utile à inférer ses futurs besoins en informations, et par conséquent de guider ses nouvelles recherches.

- Etudier les tendances de dépendance et d'indépendance des intérêts de l'utilisateur par rapport à son environnement social (influence sociale).

- Introduire les techniques d'apprentissage automatique et d'apprentissage profond dans la modélisation de l'utilisateur pour améliorer la représentativité du profil d'utilisateur.

- Étudier le problème d'identification des intérêts dans le cas d'un nouvel utilisateur (le problème de démarrage à zéro).

- Etudier la sensibilité de la requête à la reformulation, et le problème dépendant de l'expertise de l'utilisateur dans le domaine de la requête.

Références Bibliographique

[Aimé et al. 2010] Aimé, X, F Fürst, P Kuntz and F Trichet (2010), « Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées. In Actes de l'atelier 10ieme Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010), pp. 19–26. Hammamet, Tunisie.

[Badache et al. 2019] Badache, I, A Abu-Thaher, M Hamdan and L Abu-Jaish (2019), "Recherche d'information sociale en langue arabe: Cas de facebook", In 16th CORIA'19, Lyon-France.

[Bouramoul, 2011] MR. ABDELKRIM BOURAMOUL, Recherche d'information Contextuelle et Sémantique sur le web, 2011.

[C. T. YU et G. SALTON.]«Precision Weighting—An Effective Automatic In- dexing Method». In : t. 23. 1. New York, NY, USA : Association for Com- puting Machinery, jan. 1976, p. 76-88. DOI : 10.1145/321921.321930. URL : https://doi.org/10.1145/321921.321930.

[Daoud ,2009] Mariam DAOUD, Accès personnalisée a l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines `a travers l'historique des sessions de recherche, 2009.

[dounas tarik] mémoire de master, université mouloud Mammeri, Tizi-Ouzou, « implémentation d'un modèle d'appariement pour un système de recherche d'information personnalisé,2015.

[F. ZARRINKALAM] «Semantics-Enabled User Interest Mining». In : Lecture Notes in Computer Science. T. 9088. In : Gandon, 2015, p. 817-828.

[G. SALTON et M.J. MCGILL]. «Introduction to Modern information Retrie- val». In : McGraw-Hill, Inc, 1986.

[Guedria zina] thèse Doctorat,normandie université, « la recherche documentaire par un approche adaptative a base d'agents et d'artefacts »,2018. **[Ingwersen, 1992]** Peter Ingwersen, From Information Retrieval to Information Interaction, January1992.

[Kechid, 2009] Samir Kechid, Intégration du modèle utilisateur dans un système de recherche d'information distribuée, 2009.

[Kumar et al. 2014] : Kumar H., Lee S., G.Kim H., (2014). Exploiting social bookmarking services to build clustered user interest profile for personalized search, Inf. Sci. 281399– 417.

[Ould Amer 2020] Nawal OULD AMER, (2020), «Recherche Sociale et Personnalisée d'Information », thèse de doctorat, université de Grenoble Alpes, France.

[Ourdia, 2011] Mme RESSAD-BOUIDGHAGHEN Ourdia, 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, 2011.

[P. KAPANIPATHI et al]. «User Interests Identification on Twitter Using a Hie- rarchical Knowledge Base. In : The Semantic Web : Trends and Challenges». In : Crete, Greece : Springer, Anissaras, 2014, p. 99-113.

[Rafa tahar] thèse doctorat, université des scinces et de la technologie houari Boumediene, « combinaison d'un profile social et un profil situationel pour la recherche d'information »,2023.

[Robertson and Jones 1976] Robertson, S. E. and Jones, S. K., "Relevance weighting of search terms", Journal of the American Society for Information Science, 27(3), p. 129- 146. (1976).

[Porter1980], Martin F. 1980. An algorithm for suffix stripping. Program, 14(3), 130–137.

[Salton 1971] Salton, G. "The SMART retrieval system—experiments in automatic document processing.

[SaltonGerard. 1989]. Automatic text processing : The transformation, analysis, and retrieval of. Reading : Addison-Wesley.

[Schenkel et al. 2008]:Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., and Weikum, G. (2008). ―Efficient top-k querying over social-tagging networks‖. In 31st Conference on Research and Development in Information Retrieval, SIGIR'08, pages 523–530.

[Smyth 2011] Smyth, B. (2011). Web search: Social & collaborative, in: Proceedings of 8th French Information Retrieval Conférence en Recherche d'Infomations et Applications, Avignon, France, March16–18,, CORIA,2011,p.3.

[S.E.Robertson et al] :S.E., Robertson, S. Walker Jones, S., Hancock-Beaulieu, M. and Gatford, M. Okapi at trec-3. TREC, pp. 109-126, 1994.

[Tamine2007]. " Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information". Information - Interaction–Intelligence.

[Tchuente 2013] Tchuente Dieudonné (2013), " modélisation et dérivation des profils utilisateurs à partir des réseaux sociaux : approche à partir de communauté de réseaux k-egocentriques ", thèse de doctorat, Université Paul Sabatier - Toulouse III, France, 2013.

[**Vallet et al. 2006]** D. Vallet, M. Fernandez, P. Castells, P. Mylonas, and Y. Avrithis. (2006), "Personalized information retrieval in context". In Proceedings of the 21st National Conference on Artificial Intelligence - 3rd International Workshop on Modeling and Retrieval of Context, American Association for Artificial Intelligence, 2006

[Zargayouna et al. 2015] Zargayouna, H, C Roussey and J-P Chevallet (2015). "Recherche d'information sémantique: état des lieux ». Traitement Automatique des Langues TAL, 56(3), p 49–73.

[Sanderson M. (2010)] Sanderson M. (2010b). Évaluation basée sur la collection de tests des systèmes de recherche d'informations. Fondements et tendances de la recherche d'informations, vol. 4, n° 4, p. 247–375.

[C. Asrar et A. sara 2020] Reformulation sémantique et personnalisée de la requête d'utilisateur pour améliorer la pertinence des résultats de recherche d'information.

[S. FARALLI, G. STILO] Acquisition automatique d'une taxonomie des intérêts des utilisateurs de microblogs.

[Baeza-Yates and Ribeiro-Neto, eds. (1999)] Modern Information Retrieval Addison Wesley.

[T. Rafa S. Kechid 2019] Introduire une Sémantique Personnelle dans la Modélisation du Profil Utilisateur pour une Recherche d'Information Personnalisée.