République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia Farès de Médéa

Faculté de Sciences

Département Mathématique et Informatique (MI)

Mémoire de Master

Filière : Informatique

Spécialité: Ingénierie des systèmes et technologie du web

Thème

Profil d'utilisateur guidé par les évènements pour une recherche d'information personnalisée

Sujet encadré par :

- M. Rafa Tahar

Réalisé par :

- Otmani Sara
- Arafa Sara

Année : 2022 – 2023

Remerciements

Grace à dieu le tout puissant le travail porte ses fruits

Premièrement, nous remercions Dieu pour la force qu'il nous a donnée pour commencer et terminer ce mémoire.

Personne ne peut ignorer les obstacles qui se dressent dans la réalisation d'un travail soigneusement préparé dans une période donnée afin d'atteindre l'objectif que tout le monde attend.

Tout d'abord, nous n'aurions pas pu produire ce travail sans l'aide et la supervision de ***M. Rafa Tahar****, nous le remercions beaucoup pour la qualité de son encadrement, sa patience avec nous, sa précision et son accompagnement tout au long de notre préparation nous le remercions également pour les précieux conseils et informations qu'il nous a donnés, qui ont contribué à enrichir le sujet de notre étude sous ses différents aspects.*

Nous tenons à remercier Monsieur Abdoun Nabil et Monsieur Mazari Redha d'avoir accepté de participer à l'évaluation de notre travail.

Nous remercions notre collègue Amine pour son aide dans ce travail.

Dédicace

Je dédie cette mémoire à mes parents en reconnaissance de leur soutien tout au long de ma carrière scolaire, en leur disant que si j'ai réussi aujourd'hui, C'est grâce à vous. Et C'est le fruit de votre éducation.

À mon cher frère Youcef, ma source d'énergie, que Dieu vous bénisse

À mes deux sœurs Hadjer et Djoumana, la source de ma force, que Dieu vous aide dans votre chemin scolaire.

À mon amie, ma partenaire dans la mémoire, Sarah.

Sans oublier, la famille Otmani et Tlemsani, ainsi qu'à tous mes amis et connaissances.

Otmani sara

Dédicace

Je tiens à dédier cet humble travail à :

Ma tendre mère et mon très cher père pour leur soutien et leur patience et leur amour, qu'ils voient dans ce travail comme le fruit de leur amour et de leurs grands sacrifices, ils trouvent ici ma profonde gratitude et mon grand respect. À ma seule précieuse sœur Fatima zahra et à sa petite famille, à mes trois frères : Sidali, Mohamed, Walid.

À Toute la famille d'Arafa surtout ma seule grand'mère.

à ma binôme : Otmani Sara, et à mes meilleurs amis : Ikram, Meriem, Khawla .., À Tous mes amis d'enfance et du long parcours scolaire et universitaire. Et en fine à Tous ceux qui m'aiment et que j'aime.

Arafa sara

Résumé

L'objectif de la recherche d'informations personnelles est d'obtenir des informations pertinentes et personnalisées pour répondre de manière efficace aux besoins et à l'exigence individuelle des utilisateurs. En collectant des informations personnelles telles que les préférences, les intérêts et les historiques de recherche précédents.

Le but de ce mémoire est de proposer une approche personnalisée de la recherche d'informations basée sur la création de profils d'utilisateurs événementiels. Dans ce profil, nous nous combinons l'aspect thématique (termes pertinents) et l'aspect évènementiel. Le profil proposé permet d'enrichir les requêtes des utilisateurs afin de les aider à exprimer correctement leurs besoins d'information et ainsi orienter leurs nouvelles recherches d'informations.

Mots-clés : Recherche d'information, Recherche d'information Personnalisée, profil d'utilisateur, recherche d'information évènementiel.

ABSTRACT

The objective of personal information research is to obtain relevant and personalized information to effectively meet the needs and individual requirement of users. By collecting personal information such as preferences, interests and previous search histories.

The purpose of this dissertation is to propose a personalized approach to information retrieval based on the creation of event-driven user profiles. In this profile, we combine the thematic aspect (relevant terms) and the event aspect. The proposed profile makes it possible to enrich the requests of the users in order to help them to correctly express their needs for information and thus to direct their new searches for information.

Keywords: Research information, Personalized information search, user profile, event information search.

ملخص

الهدف من البحث عن المعلومات الشخصية هو الحصول على المعلومات ذات الصلة والشخصية لتلبية الاحتياجات والمتطلبات الفردية للمستخدمين بشكل فعال. من خلال جمع المعلومات الشخصية مثل التفضيلات والاهتمامات وتاريخ البحث السابق.

الغرض من هذه الرسالة هو اقتراح نهج شخصي لاسترجاع المعلومات بناءً على إنشاء ملفات تعريف المستخدمين المدفوعة بالحدث. في هذا الملف الشخصي، نجمع بين الجانب الموضوعي (المصطلحات ذات الصلة) وجانب الحدث. يتيح الملف الشخصي المقترح إثراء طلبات المستخدمين لمساعدتهم على التعبير بشكل صحيح عن احتياجاتهم من المعلومات وبالتالي توجيه عمليات بحثهم الجديدة عن المعلومات.

الكلمات المفتاحية: معلومات البحث، البحث عن المعلومات الشخصية، الملف الشخصي للمستخدم، البحث عن معلومات الحدث.

Tables des Matières

Introduction générale ---------- 2

Chapitre I : Recherche d'Information Classique ----------5

1 Introduction ---------- 5

2 Notions de base de la recherche d'information ---------- 5

2.1 Termes ---------- 5

2.2 Requête ---------- 5

2.3 Document ---------- 6

2.4 Collection de documents ---------- 6

2.5 Besoin d'information ---------- 6

2.5.1 Besoin vérificatif ---------- 6

2.5.2 Besoin thématique connu ---------- 6

2.5.3 Besoin thématique inconnu ---------- 6

2.6 Pertinence ---------- 6

2.6.1 Pertinence algorithmique ---------- 7

2.6.2 Pertinence thématique ---------- 7

2.6.3 Pertinence cognitive ---------- 7

2.7 SRI ---------- 7

3 Processus de recherche d'information ---------- 8

3.1 Expression du besoin en information ---------- 8

3.2 Processus d'indexation des documents ---------- 9

3.2.1 Analyse lexicale (extraction des mots simples) ----------10

3.2.2 Elimination des mots vides ----------10

3.2.3 Normalisation ----------10

3.2.4 Pondération des termes ----------10

3.3 Reformulation de la requête ----------11

3.4 Appariement document requête ----------11

4 Modèles de recherche d'information ----------11

4.1 Le modèle booléen ----------11

4.2 Le modèle vectoriel ----------12

4.3 Le modèle probabiliste ----------13

4.4 Autres modèles de recherche d'information ----------14

5 Evaluation des systèmes de recherche d'information ----------15

5.1 Collection de test ----------15

5.2 Mesures d'évaluation----------15
5.2.1 Rappel et précision ----------16
5.2.2 F-mesure ----------16
5.2.3 Précision @k ----------16
5.2.4 R précision----------16
5.2.5 La précision moyenne AP ----------16
5.2.6 La moyenne des précisons moyennes MAP ----------17
5.2.7 Autres mesures d'évaluation d'un SRI ----------17
6 Limitation de RI classique----------17
7 Conclusion ----------17
Chapitre II : La recherche d'Information personnalisée ---------- 19
1 Introduction----------19
2 Emergence de la recherche d'information personnalisée ----------19
3 Modélisation de l'utilisateur ----------21
3.1 Pertinence utilisateur----------22
3.2 Profil d'utilisateur----------22
3.3 Sources des données de profil ----------23
3.4 Acquisition des données de profil ----------24
3.4.1 Acquisition explicite ----------24
3.4.2 Acquisition implicite----------25
3.5 Techniques de construction de centres d'intérêts----------25
3.6 Représentation des données de profil d'utilisateur ----------26
3.6.1 Représentation basée sur l'historique de recherche----------27
3.6.2 Représentation ensembliste ----------27
3.6.3 Représentation connexionniste ----------28
3.6.4 Représentation multidimensionnelle----------29
3.6.5 Représentation conceptuelle ----------30
3.7 Evolution de profil utilisateur ----------31
4 Les dimensions du contexte de recherche----------31
4.1 Dimension sociale----------32
4.2 Dimension localisable ----------33
4.3 Dimension temporelle ----------34
4.4 Dimension sémantique ----------35
5 Intégration de profil utilisateur dans le processus de recherche----------36
5.1 Reformulation de requête ----------36

5.2 Reclassement de résultats ----------36
6 L'évaluation des systèmes de recherche d'information personnalisés ----------36
6.1 Les limites de l'évaluation classique ----------36
6.2 La problématique de l'évaluation personnalisée----------37
6.3 Protocole d'évaluation personnalisée ----------38
6.4 Les mesures d'évaluation des systèmes personnalisés----------39
7 Conclusion ----------40
Chapitre III : Un profil d'utilisateur guidé par les évènements ---------- 42
1 Introduction----------42
2 Modélisation de l'évènement----------42
2.1 Définition de l'évènement ----------42
2.2 Classification des évènements----------42
3 Profil d'utilisateur guidé par les évènements ----------43
3.1 L'historique de recherche thématique----------44
3.1.1 Les documents pertinents----------44
3.1.2 Les termes pertinents----------44
3.1.3 La base de liens de cooccurrence terme-terme ----------45
3.2 L'historique de recherche évènementiel----------46
3.2.1 Modèle d'évènement proposé----------46
3.2.2 Pondération des évènements passés----------46
3.2.3 Pondération des évènements actuels ----------47
3.3 Processus de recherche d'information----------49
3.3.1 Sélection de la base de profil la plus adéquate pour la requête utilisateur-------50
3.3.2 Reformulation de requête----------50
3.3.3 Mise à jour du profil----------51
4 Simulation ----------51
4.1 Cadre d'évaluation----------51
4.2 Présentation de base de données ----------53
4.3 Présentation de l'application----------56
5 Conclusion ----------61
Conclusion Générale ---------- 63
Bibliographie ---------- 65

Tables des Figures

Figure 1.1: Système de Recherche d'Information. 08
Figure 1.2: Modèle vectoriel "VSM". 13
Figure 2.1 : de 1ere page au 5eme résultat Google sur « Virus ». 20
Figure 2.2 : 2eme page de résultat Google sur « Virus ». 20
Figure 2.3 : 3eme page de résultat Google sur « Virus ». 21
Figure 2.4 : Profil source. 23
Figure 2.5 : Exemple de profil représenté par des vecteurs de mots clés. 28
Figure 2.6 : Un extrait d'un profil utilisateur sémantique. 29
Figure 2.7 : Métamodèle de profil utilisateur. 30
Figure 3.1 : Profil d'utilisateur. 43
Figure 3.2 : Historique de recherche. 44
Figure 3.3 : Base de cooccurrence terme-event-a. 45
Figure 3.4 : Modèle de l'évènement. 46
Figure 3.5 : base de cooccurrence terme-event-p. 47
Figure 3.6 : Base de cooccurrence terme-event-a. 49
Figure 3.7 : Processus de recherche. 50
Figure 3.8 : Représentation de modèle conceptuel des données (MCD). 52
Figure 3.9 : Table des profiles. 53
Figure 3.10 : Table des termes (terms). 54
Figure 3.11 : Table des événements passés (event_passes). 54
Figure 3.12 : La base de liens de cooccurrence terme-terme (profile_term_terms). 55
Figure 3.13 : Pondération des liaisons des termes pertinents à un évènement passé (profile_term_event_passes). 55
Figure 3.14 : Se connecter au compte. 56
Figure 3.15 : Inscription au compte. 57
Figure 3.16 : Mettre à jour le profil. 57
Figure 3.17 : Page de recherche. 58
Figure 3.18 : Les résultats de recherche classique. 58
Figure 3.19 : L'étape de la sélection de la base la plus adéquate a Q. 59
Figure 3.20 : Les calculs de similarité. 59
Figure 3.21 : La reformulation de Q. 60
Figure 3.22 : Les résultats de recherche avec Q'. 60

Introduction générale

Introduction Générale

L'objectif de la recherche d'information personnalisée RIP est de permettre aux systèmes de recherche d'information de retourner, en réponse d'une requête utilisateur, les résultats les plus pertinents pour les réels besoins en information cachés derrière cette requête. La personnalisation de la recherche d'information est basée sur la prise en compte, lors de la recherche, d'un ensemble de données à propos de l'utilisateur et ses intérêts.

Les données sur l'utilisateur, qui peuvent être explicitement exprimées par lui-même ou implicitement extraites à partir de son historique de recherche, couvrent plusieurs aspects contextuels :

- L'aspect thématique représenté par les documents jugés pertinents par l'utilisateur pendant ses activités de recherche précédentes ainsi que les termes pertinents extraits à partir de ces documents.
- L'aspect social représenté par les différentes activités sociales identifiées dans l'historique de recherche l'utilisateur (annotations, commentaires, recommandations, partage, liens d'amitié…etc.),
- L'aspect situationnel représenté par les différentes informations concernant la localisation (les emplacements de recherche et leurs types, les cordonnées GPS, …etc.).

Un autre aspect très important pour la description des intérêts de l'utilisateur est l'aspect temporel. Dans la littérature de la recherche d'information personnalisée, l'information temporelle souvent prise en considération est celle concernant la création, l'actualisation et les mouvements des documents. L'objective étant de pouvoir classer les documents en fonction de ces informations temporelles.

Dans ce présent travail, différemment de la plupart des approches qui traitent le sujet de l'information temporelle, nous nous intéressons à l'aspect temporel inclus dans les contenus des documents. L'information temporelle que nous allons prendre en compte est en particulier celle qui couvre la dimension évènementielle des besoins en information de l'utilisateur.

Nous défendons l'hypothèse suivante : l'utilisateur peut s'intéresser aux évènements actuels, imminents ou récemment passés qui attirent l'attention d'une bonne partie de sa communauté proche, comme il peut s'intéresser aux événements passés ayant une périodicité de reproduction ou de célébration. Et ainsi, la prise en compte des évènements 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.

La contribution de notre nouvelle approche proposée pour la personnalisation de la recherche se résume à :

(i). La proposition d'un modèle de l'évènement qui facilite son identification et extraction à partir des textes,
(ii). La proposition d'un modèle pour le profil d'utilisateur prenant en compte l'aspect évènementiel,
(iii). L'intégration du profil d'utilisateur dans le processus de recherche d'information.

(iv). La définition de certaines formules de pondération et de scores de similarité.

Le document est organisé en 3 chapitres comme suit :

- Chapitre I : recherche d'information classique.
- Chapitre II : recherche d'information personnalisée.
- Chapitre III : Un profil d'utilisateur guidé par les évènements.

Le document se termine avec une conclusion et des perspectives.

Chapitre I :
Recherche d'Information Classique

1 Introduction

La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information. [Boubekeur F, 08]

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur.

L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents (Indexe Ou Fichier Inverse) au moyen d'une fonction de comparaison (ou de correspondance). L'essor du web a remis la RI face à de nouveaux défis d'accès à l'information, il s'agit cette fois de retrouver une information pertinente dans un espace diversifié et de taille considérable. Ces difficultés ont donné naissance à une nouvelle discipline appelée Recherche d'Information sur le Web [Abdelkarim B, 11]. Ces trois éléments indissociables qui sont la RI, les SRI et la RI sur le Web, font l'objet de ce chapitre.

Ce chapitre est divisé en cinq parties, la première comprend les notions de base de recherche d'information, La deuxième partie décrit le processus de RI qui contient l'expression du besoin en information , la reformulation de la requête et l'appariement document requête, La troisième partie sera consacrée les modèles de recherche d'information comme le modèle booléen ,vectoriel et probabiliste , la quatrième partie nous discuterons d'évaluation des systèmes de recherche d'information en parlant de la collection de test et les mesures d'évaluation , au bout de nos parlerons , la limitation de RI classique.

2 Notions de base de la recherche d'information

Dans le domaine de la recherche d'informations, il existe de nombreux concepts de base, notamment :

2.1 Termes

Un modèle de représentation est un processus permettant d'extraire d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique. Ce processus de conversion est appelé indexation. Le résultat de l'indexation constitue le descripteur du document ou de la requête, qui est une liste de termes ou groupes de termes (concepts), significatifs pour l'unité textuelle correspondante, auxquels sont associés généralement des poids, pour différencier leurs degrés de représentativité du contenu sémantique de l'unité en question. L'ensemble des termes reconnus par le SRI est rangé dans une structure appelée dictionnaire constituant le langage d'indexation. Ce type de langage garantit le rappel de documents lorsque la requête utilise dans une large mesure les termes du dictionnaire. En revanche, il y a risque important de perte d'informations lorsque la requête s'éloigne de ce vocabulaire. [ABBASSI et al, 13]

2.2 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 [Mariam D, 09], Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique. [Abdelkarim B, 11]

2.3 Document

On appelle document toute unité d'information qui peut constituer une réponse à un besoin en information/requête d'un utilisateur. Un document peut être un texte, un morceau de texte, une image, une bande vidéo, etc. [Mariam D, 09]

2.4 Collection de documents

La collection de documents (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 un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. Ces représentations sont étudiées de telles sortes que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base se font dans les meilleures conditions de coût. [Abdelkarim B, 11]

2.5 Besoin d'information

La notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis par [P. Ingwersen, 94] :

2.5.1 Besoin vérificatif

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 vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.

2.5.2 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.

2.5.3 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.

2.6 Pertinence

Selon les premières définitions de la pertinence [G. Salton et M. McGill, 70], 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, etc.) 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. Cependant, de nombreuses études ont été menées [L. Barry, 94] [B. Pia, 03] autour de la notion de pertinence et montrent que la pertinence n'est pas une relation isolée entre le document et la requête et qu'elle est définie par un ensemble de critères et de préférences qui varient selon les utilisateurs. Ces critères sont des facteurs qui déterminent la pertinence accordée à l'information retrouvée par l'utilisateur dans un contexte de recherches précises. Les facteurs qui affectent les jugements de pertinence font l'objet de recherche de plusieurs études [K. Carlos et A. Cuadro, 67] [L. Barry, 94] [B. Pia, 03]. Nous citons les critères définis par [L. Barry, 94] et groupés dans sept catégories : (1) le contenu informationnel des documents ; (2) le niveau d'expertise et de connaissances de l'utilisateur ; (3) les croyances et préférences de l'utilisateur ; (4) autres informations liées à l'environnement ; (5) les sources des documents ; (6) les documents comme des entités physiques ; et (7) la situation de l'utilisateur. Compte tenu de ces facteurs, il existe plusieurs "'pertinences"' possibles entre un document et un besoin, nous en citons les quatre (4) les plus importantes [L. Tamine and S. Calabretto, 08] :

2.6.1 Pertinence algorithmique

Cette pertinence (appelée aussi pertinence système) est traduite par une mesure algorithmique basée sur le calcul de la pertinence de l'information par rapport à la requête en utilisant des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte.

2.6.2 Pertinence thématique

Cette pertinence est traduite par la topicalisé et est définie par le degré de couverture du document du thème évoqué par le sujet de la requête. C'est le seul type de pertinence appliqué dans les campagnes d'évaluation tels que TREC.

2.6.3 Pertinence cognitive

C'est la pertinence liée au thème de la requête, selon la perception ou les connaissances de l'utilisateur sur ce même thème ; cette pertinence est caractérisée par une dynamique qui permet d'améliorer la connaissance de l'utilisateur via l'information renvoyée au cours de sa recherche. 4. pertinence situationnelle : cette pertinence est définie par l'utilité ou l'usabilité de l'information jugée relativement à la tâche de recherche exprimée par le besoin en information selon la perception de l'utilisateur. C'est une pertinence potentiellement dépendante du contexte de recherche et est vue comme étant une pertinence dynamique. Il est à noter qu'un SRI idéal doit supporter un modèle de recherche.

2.7 SRI

Un SRI est un système qui gère une collection d'informations organisée sous forme d'une représentation intermédiaire qui reflète aussi fidèlement que possible le contenu des documents. Ceci est réalisé grâce à un processus préalable d'indexation manuelle ou automatique. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée. Le résultat est donné par la mise en œuvre d'un mécanisme

d'appariement entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents. [Ouaneche L, 13]

- La notion de document : L'ensemble des documents sur lesquels portera la recherche est stocké dans une banque de données (sur le Web). Un document est le type d'objet de base géré par le système.
- La notion de besoin d'information d'un utilisateur (requête) : Ce besoin est exprimé par une requête spécifiée dans un formalisme propre au système. Le formalisme de spécification de la requête peut être en langage naturel.
- Notion de correspondance entre la requête et les documents : Une fois la requête spécifiée, le système tente de retrouver les documents qui correspondent à la requête en se basant sur une mesure de similarité.
- La notion de contexte de l'application : Le contexte de l'application représente l'univers dans lequel le système fonctionne. L'univers est nécessaire aux SRI pour une bonne compréhension des besoins des utilisateurs. Un SRI doit être capable de retrouver les documents pertinents à partir d'une banque de données, satisfaisant la requête posée par un utilisateur et traduisant un besoin d'information donné.

3 Processus de recherche d'information

Un système de recherche d'information est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple), et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information [C. Tambellini, 07].

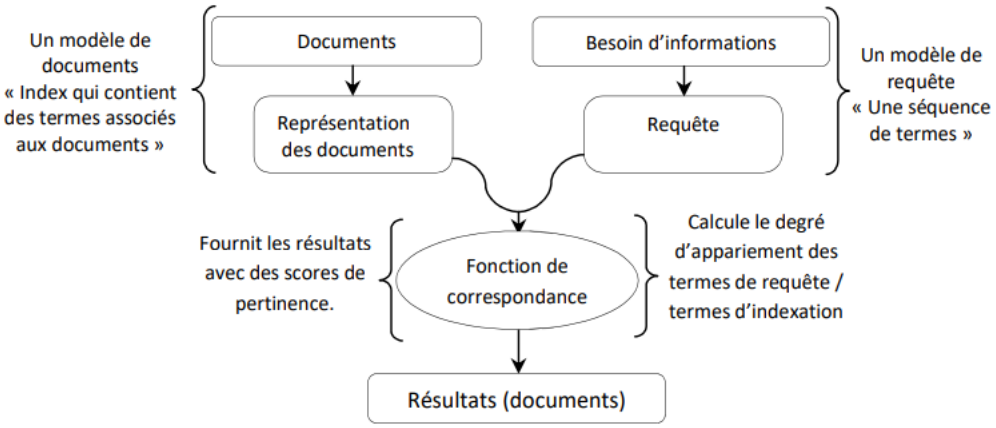


Figure 1.1 – Système de Recherche d'Information.

3.1 Expression du besoin en information

Le besoin en information est l'expression mentale de ce que l'utilisateur recherche.

Trois types de besoin ont été définis dans [P. Ingwersen, 94] :

- Le besoin vérificatif : l'utilisateur cherche à vérifier une information particulière dont il sait comment y accéder (Exemple : La recherche d'un article à partir d'une adresse web connue). Ce type de besoin est stable et ne change pas au cours de la recherche.

- Le besoin thématique connu : l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un domaine connu (Exemple : l'utilisateur cherche les remèdes d'une maladie qu'il connait déjà). Ce type de besoin s'affine généralement au cours de la recherche.
- Le besoin thématique inconnu : l'utilisateur cherche de nouvelles connaissances dans un domaine ou un sujet qu'il ne connait pas. Un besoin de ce type est généralement exprimé de manière incomplète et imprécise.

Une requête est une expression approximative du besoin en information de l'utilisateur. Elle peut donc être considérée comme une description partielle d'un besoin d'information à un instant donné.
Les requêtes soumises au SRI par les utilisateurs peuvent ne pas refléter leurs besoins en information. Cela est dû, au fait que l'utilisateur ignore le fonctionnement interne du SRI, et il n'a qu'une vision restreinte des documents disponibles dans la collection. D'autre part le SRI n'a souvent aucune connaissance a priori de ses utilisateurs (centres d'intérêts, niveaux, etc.). Ce biais entre la requête et le besoin en information est une des difficultés majeures de tout système de recherche d'information. Afin de remédier partiellement à ce problème, un mécanisme de reformulation de requêtes peut être intégré dans les SRI.

3.2 Processus d'indexation des documents

L'indexation est une phase très importante pour un SRI car de sa qualité dépend la qualité des réponses du système et donc les performances de ce dernier. Une bonne indexation doit permettre de retrouver tous les documents pertinents au besoin de l'utilisateur et pas (ou peu) de documents non pertinents pour celui-ci. [Fatiha B, 08]
Selon Philipe Lefevre :« l'indexation est la représentation formalisée et condensée du contenu thématique ou sémantique d'un document, à l'aide d'indicateur sémantique qui peuvent être des indices de classification ou des descripteurs, contrôlés ou non. Les indicateurs appartiennent au langage naturel et/ou à un langage documentaire ». [Paris, 04] L'indexation peut être :

- **Indexation manuelle :**

Dans l'indexation manuelle, chaque document est analysé par un spécialiste du domaine ou par un documentaliste. Ce type d'indexation permet une :

- ✓ Meilleure précision dans les documents que le système de RI retourne en réponse aux requête des utilisateurs.
- ✓ Meilleure correspondance entre les documents et les termes choisis pour les représenter.

Cependant, elle présente certains inconvénients, comme :

- ✓ Méthode inapplicable aux corpus de textes volumineux.
- ✓ Elle est lourde à mettre en œuvre lorsque le document est grand.
- ✓ L'effort intellectuel qu'elle exige (en temps et en nombre de personnes).

- **Indexation semi-automatique (supervisée) :**

L'indexation semi-automatique est une combinaison entre les deux méthodes (manuelle, automatique), elle utilise un vocabulaire contrôlé sous forme de thesaurus ou de base terminologique.
Les premiers termes d'indexation sont obtenus grâce à l'indexation automatique, il vient alors le rôle de l'indexeur qui corrige puis complète les informations obtenues et établi un ensemble final des termes d'indexation.

- **Indexation automatique :**

Elle est sans doute celle qui a été la plus étudiée en recherche d'information. Dans cette approche, chaque document est analysé à l'aide d'un processus entièrement automatisé. Elle permet de traiter les textes plus rapidement que l'indexation manuelle et elle est particulièrement adaptée aux corpus volumineux.

3.2.1 Analyse lexicale (extraction des mots simples)

Il s'agit du processus de conversion de texte en une séquence de mots. Elle convertit un texte de document en une liste de termes.
Un terme est un groupe de caractères constituant un mot significatif. L'analyse lexicale permet de reconnaître les espaces de séparations des mots, les chiffres, etc. [HAMOUM A et ABERKANE F]

3.2.2 Elimination des mots vides

Les mots vides sont des mots peut significatifs et porteurs de peu de sens, augmentant ainsi la taille de l'index et rendant la recherche plus lente. De ce fait, leur élimination est impérative.
L'élimination des mots vides est importante dans la mesure où il représente un facteur qui a une grande influence sur la précision de la recherche. En effet, le fait de ne pas éliminer les mots vides provoque inévitablement du bruit (les documents non pertinents à la requête). On distingue deux techniques pour éliminer les mots vides [HAMOUM A et ABERKANE F] :

- L'utilisation d'une liste de mots vides.
- L'élimination des mots dépassant un certain nombre d'occurrence dans la collection.

3.2.3 Normalisation

La normalisation consiste à représenter les différentes variantes d'un terme par un format unique appelé lemme ou racine, Il s'agit de ramener le mot à son radical. Ce qui a pour effet de réduire la taille de l'index. Plusieurs stratégies de normalisation sont utilisées : la table de correspondance, l'élimination des affixes, la troncature.
L'inconvénient majeur de cette opération est qu'elle supprime dans certain cas la sémantique des termes originaux, c'est le cas par exemple activate/active [HAMOUM A et ABERKANE F].

3.2.4 Pondération des termes

La pondération d'un terme d'indexation est l'association d'une valeur numérique à ce terme de manière à indiquer sa représentativité, estimée à travers la fréquence d'apparition du terme au sein d'un document et son pouvoir de discrimination dans le corpus, obtenu par le biais de la fréquence d'apparition globale du terme au sein de tout le corpus [Manning C et al. 08]

Stalon et Buckley décrivent et comparent différentes fonctions de pondération qui permettent de distinguer différents modèles de représentation comme par exemple : le modèle binaire, le modèle vectoriel… [Samir K, 09]

L'élément fondamental dans un système de recherche d'information est la technique utilisée pour pondérer les termes [Sparck J, 79] [Fatiha B, 08]. La pondération des termes est utilisée dans l'opération d'indexation, permettant d'associer à chaque terme son poids dans le document. La plupart des techniques de pondération des termes sont basées sur les facteurs Tf et Idf :

Tf (terme frequency) : cette mesure est proportionnelle à la fréquence du terme dans le document. Elle est souvent exprimée selon l'une des déclinaisons suivantes :

- Tf : $utilisation\ brute$;
- $a + log(Tf)$;
- $[0/1]$: $Présence, absence$;
- $0.5 + 0.5 * (tf\ /\ Max(tf))$;

Idf (Inverse of Document Frequency) : mesure l'importance d'un terme dans toute la collection. Un terme trop fréquent dans la collection ne doit pas avoir le même impact sur la collection qu'un terme moins fréquent. Elle est exprimée selon l'une des déclinaisons suivantes :

- $Log\ (N\ /df)$;
- $Log\ ((N - df)\ /\ df)$.

Où df est le nombre de documents contenant le terme et N est le nombre dedocuments dans toute la collection.

3.3 Reformulation de la requête

Compte tenu des volumes croissants des bases d'information, retrouver les informations pertinentes en utilisant seulement la requête initiale de l'utilisateur est une tâche quasi-impossible. Afin de rapprocher au mieux la pertinence système de la pertinence utilisateur, une étape de reformulation de la requête est souvent utilisée dans les systèmes de recherche d'information. La reformulation de la requête consiste à modifier la requête de l'utilisateur par ajout de termes significatifs et/ou réestimation de leurs [Sparck J, 71] poids.

3.4 Appariement document requête

Une fois l'indexation des requêtes et des documents est effectuée, l'appariement entre la requête et le document peut désormais s'effectuer. L'appariement consiste à détecter les documents pertinents pour la requête posée, et éventuellement calculer le degré de cette pertinence appelé aussi score ou RSV (Retrieval Status Value). Ce score sert à trier la liste des documents retournés. [Sparck J, 71]

4 Modèles de recherche d'information

Un modèle de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence. Il existe un grand nombre de modèles de RI textuelle d´enveloppées dans la littérature. Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requêté-document. Le vocabulaire d'indexation V = {ti}, i ∈ {1, ..., n} est constitué de n mots ou racines de mots qui apparaissent dans les documents. Selon [R. Baeza et A. Ribeiro, 99], un modèle de RI est défini par un quadruplet (D, Q, F, R(q,d)) :

- 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 d à la requête q.

Nous présentons dans la suite les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

4.1 Le modèle booléen

Le modèle booléen [G. Salton, 71] est basé sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document. Un exemple de représentation d'un document est comme suit :

$$d = t_1 \wedge t_2 \wedge t_3 \ldots \wedge t_n \qquad (01)$$

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Un exemple de représentation d'une requête est comme suit :

$$q = (t_1 \wedge t_2) \vee (t_3 \wedge t_4) \qquad (02)$$

La fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document dj implique l'expression logique de la requête q. Le résultat de cette fonction est donc binaire décrite comme suit :

$$RSV\ (q,d)\ =\ \{1,0\} \qquad (03)$$

4.2 Le modèle vectoriel

Le modèle vectoriel est un modèle qui a été proposé par Gerard Salton dans les années 1970 [G. Salton, 70]. C'est un modèle algébrique, il représente les documents ainsi que les requêtes par des vecteurs de termes dans un espace multidimensionnel (N dimensions, avec N : le nombre des termes d'indexation de la collection).
Les composants des vecteurs sont le poids du terme ti d'indexation considéré dans le document ou la requête, ils sont définis de la manière suivante :

- Un vecteur document $\boldsymbol{Dj\ =\ (d_{1j}, d_{2j},\ d_{3j}, \dots, d_{nj})}$ (04)
- Un vecteur requête $\boldsymbol{Q\ =\ (q_1, q_2, q_3, \dots, q_n)}$ (05)

Où : dij est le poids du terme ti dans le document dj
La pondération est une fonction très importante en RI, elle permet de mesurer l'importance d'un terme dans un document.
Les techniques de pondérations les plus utilisées sont basées sur les facteurs $\boldsymbol{Tf}$ et $\boldsymbol{Idf}$, qui permettent de combiner les pondérations locale et globale d'un terme :

- $\boldsymbol{Tf}$ **(TermFrequency) :** C'est la fréquence du terme dans le document c'est-à-dire le nombre d'occurrences d'un terme dans le document (pondération locale).
- $\boldsymbol{Idf}$ **(Inverse of Document Frequency) :** C'est la fréquence du terme dans l'ensemble des documents (pondération globale). Elle doit être d'autant plus fort (faible) que le terme permet de mieux (moins bien) discriminer le document parmi les autres documents du corpus.

La mesure $\boldsymbol{Tf * Idf}$ donne une bonne approximation de l'importance du terme dans le document, particulièrement dans les corpus de documents de taille homogène. Cependant, elle ne tient pas compte d'un aspect du document : sa longueur. En général, les documents les plus longs ont tendance à utiliser les mêmes termes de façon répétée, ou à utiliser plus de termes pour décrire un sujet. Par conséquence, les fréquences des termes dans les documents seront plus élevées, et les similarités à la requête seront également plus grandes.
La comparaison de la requête au document dans ce modèle est effectuée en évaluant la similarité entre leurs vecteurs respectifs. Les documents considérés comme les plus pertinents sont ceux dont le vecteur est le plus proche de celui de la requête, en tenant compte d'une mesure de similarité définit au préalable. Plusieurs mesures de similarité ont été définies, les principales utilisées sont les suivantes :

- Le produit scalaire : $SIM\ (Q, D_j) = \sum_{i=1}^{N} q_i * d_{ij}$ (06)
- Mesure de Jaccard : $SIM\ (Q, D_j) = \frac{\sum_{i=1}^{N} q_i * d_{ij}}{\sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} d_{ij}^2 - \sum_{i=1}^{N} q_i * d_{ij}}$ (07)
- Mesure cosinus : $SIM\ (Q, D_j) = \frac{\sum_{i=1}^{N} q_i * d_{ij}}{(\sum_{i=1}^{N} q_i^2)^{1/2} * (\sum_{i=1}^{N} d_{ij}^2)^{1/2}}$ (08)

Le modèle vectoriel est simple à comprendre et facile à implémenter. Il permet une représentation uniforme des documents et des requêtes, il prend en considération le poids des termes dans les documents et les mesures de similarités qu'il utilise permettent de considérer un document comme pertinent même s'il ne contient pas tous les termes de la requête. De plus, ce modèle classe les résultats par ordre décroissant de pertinence.
Ce modèle présente quelques limitations comme le fait qu'il ne respecte pas l'ordre des mots, et il suppose l'indépendance entre les termes d'indexation, or que ces termes dans les documents sont sémantiquement liés.

Les documents et la requête sont représentés comme des vecteurs dans l'espace vectoriel (figure 1.2).

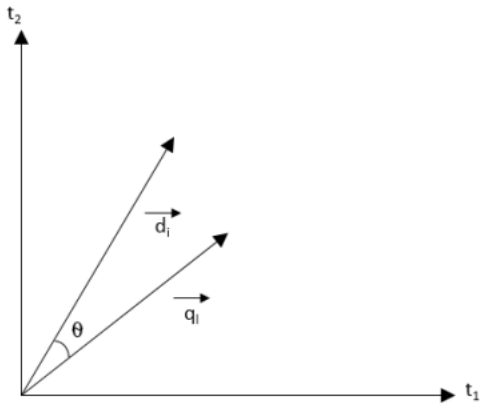


Figure 1.2 – Modèle vectoriel "VSM".

4.3 Le modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête [S. Robertson et K. Sparck, 76] [G. Salton et M. McGill, 83] [E. Maron et L. Kuhns, 60]. Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Etant donné une requête utilisateur Q et un document D, il s'agit de calculer la probabilité de pertinence du document pour cette requête. Deux possibilités se présentent : R, D est pertinent pour q et R, D n'est pas pertinent pour q. Les documents et les requêtes sont représentés par des vecteurs booléens dans un espace à n dimensions. Un exemple de représentation d'un document dj et une requête q est le suivant :

- $Dj = (w_{1,j}, w_{2,j}, w_{3,j}, \dots, w_{n,j})$ (09)
- $Q = (w_{1,q}, w_{2,q}, w_{3,q}, \dots, w_{n,q}). Avec\ w_{k,j} \in [0\ 1]\ et\ w_{k,q} \in [0\ 1]$ (10)

La valeur de $w_{k,j}$ (resp. $w_{k,q}$) désigne si le terme t_k apparaît ou non dans le document Dj (resp. Q). Le modèle probabiliste évalue la pertinence du document dj pour la requête q. Un document est sélectionné si la probabilité que le document d soit pertinent, notée p(R/D), est supérieure à la probabilité que d soit non pertinent pour q, notée p(R/D) ou R est l'événement de pertinence et R est l'´événement de non pertinence. Le score d'appariment entre le document D et la requête Q, noté $RSV(Q, D)$ est donné par :

$$RSV(Q, D) = \frac{P(R/D)}{P(\overline{R}/D)} \quad (11)$$

Ces probabilités sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent. Cette mesure de similarité requête document peut se calculer par différentes formules.

Ce modèle a donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI. Le modèle Okapi BM25 a été développé par (Robertson et al., 1994) dans lequel le calcul du poids d'un terme dans un document intègre des aspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents :

$$w_{td} = \log\left(\frac{N-df+0.5}{df+0.5}\right) \times \frac{(K_1+1)*TF}{K_1 \times \left((1-b)+b\times\frac{dl}{avgdl}\right)+tf} \qquad (12)$$

Ou $\boldsymbol{dl} = \boldsymbol{moy_{di\in D}(dli)}$ et $\mathbf{K_1}$, **b** sont des paramètres qui dépendent de la collection.

4.4 Autres modèles de recherche d'information

Parmi les Autres modèles de recherche d'information il y a :

Le modèle connexionniste : Les SRI basés sur l'approche connexionniste utilisent le fondement des réseaux de neurones, tant pour la modélisation des unités textuelles que pour la mise en œuvre du processus de RI. L'idée de base est que la RI est un processus associatif qui peut être représenté par les mécanismes de propagation d'activation des réseaux de neurones. De plus, les capacités d'apprentissage de ces modèles peuvent permettre d'obtenir des SRI adaptatifs.

Deux modèles théoriques ont été utilisés : les modèles à auto-organisation et les modèles à couches.

Les modèles à auto-organisation [X. Soergei, G. Marchionini, 91] permettent à partir de la description des documents, d'en réaliser une classification par l'apprentissage du réseau de neurones. Ces modèles sont basés sur les cartes auto-organisatrices de Kohonen [T. Kohonen, 98].

Les modèles à couches : Les SRI basés sur un modèle connexionniste à couches [Kwok K. L, 89] [Belew R, 89] [Mohand Boughanem, 92] [Mothe J, 94] sont représentés par un minimum de trois couches de neurones interconnectées : la couche requête (Q), la couche termes (T) et la couche documents (D). Le mécanisme de recherche est basé sur une activation initiale des neurones termes induite par une requête, et qui se propage vers les documents à travers les connexions du réseau. Dans le modèle MERCURE [Mohand Boughanem, 92], une requête Q est représentée par un vecteur de poids sous forme :

$$Q_u^{(t)} = (q_{u1}^{(t)}, q_{u2}^{(t)}, \dots, q_{uT}^{(t)}) \qquad (13)$$

Les poids des termes dans la requête sont affectés aux liens requête-termes. L'activité initiale du réseau correspond à l'activation d'un nœud requête en envoyant un signal de valeur 1 à travers les liens requête-termes. Chaque neurone terme t_j affecté par la requête, reçoit une entrée $In(t_j)$ et fournit une sortie $Out(t_j)$ respectivement définies par :

$$\ln(t_j) = q_{uj}^{(t)} \qquad (14) \qquad Out(t_j) = g(\ln(t_i))) \qquad (15)$$

Un document di qui a des termes t_j en commun avec la requête recevra une entrée $In(d_i)$ et calculera sa sortie $Out(d_i)$ telles que :

$$\ln(d_i) = \sum_{i=1}^{T} Out(t_j) * w_{ij} \quad (16) \qquad Out(d_i) = g(\ln(d_i)) \quad (17)$$

Où w_{ij} est le poids du terme t_j dans le document d_i.

Les valeurs de sortie des différents documents correspondent à leurs degrés de pertinence pour la requête donnée.

Modèle LSI : est une approche vectorielle qui Exploite la coexistence des termes et réduit l'espace des termes, en regroupant les termes communs dans les mêmes dimensions, puis les documents et les requêtes sont représentés dans espace plus petit, composé de concepts de niveau supérieur. [Boughanem M]

5 Evaluation des systèmes de recherche d'information

L'évaluation des méthodes et des modèles a été un centre d'intérêt important et cela dès l'apparition des premiers SRI. Et pour cela, des moyens comme les mesures de qualité, des systèmes de recherche et un ensemble de données, ont été développés dans le but d'évaluer ces systèmes. L'évaluation des SRI peut être vue sous deux angles différents : paradigme système et paradigme usager. Le but du paradigme système consiste à mesurer les capacités du système à satisfaire les besoins en information des utilisateurs, c'est-à-dire la pertinence des résultats retournés. Le paradigme usager par contre, vise à satisfaire l'utilisateur, et non les capacités du système. [HAMOUM A et ABERKANE F]

5.1 Collection de test

Une collection de test représente le moyen d'évaluation des SRI. Elle comprend généralement :

- Un corpus de documents : c'est un ensemble de documents à indexer, sur lesquels le système sera évalué.
- Une liste de requêtes prédéfinies.
- Les jugements de pertinence : ils sont établis manuellement et ils contiennent la liste des documents pertinents pour chaque requête.

On peut juger l'efficacité et la performance d'un système de RI, en comparant les résultats retournés par ce dernier avec les jugements de pertinence. Depuis les années 70, les collections de test se sont multipliées. Elles sont mises en place dans le cadre de compagne d'évaluation des SRI, dont la collection CACM, la collection CISI, la compagne CLEF et la compagne TREC qui constitue à ce jour la compagne de référence pour ce qui concerne l'évaluation des SRI et cela depuis l'année de son lancement 1992. L'objectif de la compagne TREC consiste à réunir sous une plate-forme des collections de test et des protocoles d'évaluations pour des tâches spécifiques pour mesurer les différentes stratégies de recherche. Les tâches proposées se sont diversifiées d'une compagne à une autre, parmi celles proposées dans TREC 2012 pour citer : recherche d'informations sur le web, recherche d'informations dans les micros blogs, recherche d'informations médicales, etc. [HAMOUM A et ABERKANE F]

5.2 Mesures d'évaluation

L'objectif principal des SRI est de présenter à l'utilisateur tous les documents pertinents et de rejeter tous les documents non pertinents. Cet objectif est évalué avec différentes mesures

d'évaluation. Les plus utilisées sont : la précision, le rappel, et les mesures alternatives. [HAMOUM A et ABERKANE F]

5.2.1 Rappel et précision

Le rappel (R) : le rappel mesure le pourcentage de la proportion du nombre de documents pertinents restitué parmi le nombre total de documents pertinents dans la collection il est exprimé comme suit :

$$R = \frac{Nomber\ de\ documents\ pertinents\ restitués}{Nomber\ total\ de\ documents\ pertinents} \qquad (18)$$

Quand le rappel vaut 1 c'est que le nombre total de documents pertinents ont tous été retrouvés par le système. Quand il vaut 0 alors aucun document pertinent n'a été retrouvé. Des mesures complémentaires au rappel et précision ont été définies, il s'agit de bruit et de silence. [HAMOUM A et ABERKANE F]

La précision (P) : elle mesure le pourcentage de la proportion du nombre de documents pertinents restitués parmi tous les documents restitués par le système, elle est exprimée comme suit :

$$P = \frac{Nomber\ de\ documents\ pertinents\ restitués}{Nomber\ total\ de\ documents\ restitués} \qquad (19)$$

La précision mesure la capacité du système à trouver seulement les documents pertinents. [HAMOUM A et ABERKANE F] Elle vaut 1 si tous les documents retrouvés sont pertinents, et elle vaut 0 si aucun document retrouvé n'est pertinent.

5.2.2 F-mesure

Est une mesure qui estime une moyenne harmonique pondérée de la précision et du rappel et est décrite par la formule suivante [Nawal O] :

$$F - mesure = \frac{2 \times Precision \times Rappel}{Precision \times Rappel} \qquad (20)$$

5.2.3 Précision @k

La précision à un rang donné (P@k) c'est la précision pour les k premiers documents retournés [Chahrazed Bouhini].

5.2.4 R précision

C'est une précision qui est calculée sur les R premiers documents retournés par le système, sachant que la requête admet R documents pertinents. [HAMOUM A et ABERKANE F]

5.2.5 La précision moyenne AP

La précision moyenne pour AP_l est calculée pour chaque requête q_l. Il s'agit de la moyenne des Précisions à chaque rang de document pertinent pour cette requête. Elle est donnée par la formule [Chahrazed Bouhini] :

$$AP_l = \frac{\sum_{k=1,K}(rel(k).P@k)}{\sum_{k=1,K} rel(k)} \qquad (21)$$

Avec K le nombre de documents retrouvés, P@k : la précision au rang k et rel(k) calculé comme suit :

$$rel(k) = \begin{cases} 1: si\ le\ k^{ieme}\ document\ est\ pertinent \\ 0: sinon \end{cases}$$

5.2.6 La moyenne des précisons moyennes MAP

La moyenne des précisions moyennes (Mean Average Precision "MAP") pour l'ensemble des requêtes est calculée à l'aide des précisions moyennes AP_l des requêtes, afin de mesurer la performance globale du système ([Manning et al., 2008]) comme présenté dans l'équation [Chahrazed Bouhini] :

$$MAP = \frac{1}{|Q|}\sum_{l=1,|Q|} AP_l \qquad (22)$$

5.2.7 Autres mesures d'évaluation d'un SRI

D'autres mesures d'évaluation d'un SRI existent. Ainsi, des mesures complémentaires au rappel et à la précision, respectivement le bruit et le silence ont été définies comme suit :

$$\text{Bruit} : B = \frac{|DNPR|}{|DPR \cup DNPR|} \qquad (23) \qquad \text{Silence} : S = \frac{|DPNR|}{|DP|} \qquad (24)$$

L'indice de Fallout (ou Hallucination d'après C. Berrut2) peut être utilisé à la place du rappel [Ishioka, T, 03]. Il définit le pourcentage de documents non pertinents qui ont été retrouvés ([E. Kraft et A. Bookstein, 78] cité dans [Ishioka, T, 03]). Il exprime l'erreur du système. Formellement :

$$\text{Fallout} : a = \frac{|DNPR|}{|DNP|} \qquad (25)$$

6 Limitation de RI classique

La limite majeure de la RI classique est qu'elle est basée sur une approche généraliste qui évalue invariablement les requêtes des utilisateurs et délivrent des résultats sans tenir compte des critères spécifiques de l'utilisateur qui a émis la requête ou du contexte de de recherche. Par exemple, la requête "gouvernement" est une requête locale qui doit renvoyer des résultats concernant le gouvernement du pays dans lequel est localisé l'utilisateur. Par exemple, cette requête doit renvoyer en France le site du gouvernement français. La requête "Java magazines" est une requête ambigüe du fait qu'elle est liée à différentes thématiques de recherche, telles que dès le langage de programmation Java, l'île de Java ou un magazine ayant le nom Java. Pour ce type de requêtes, le système doit renvoyer des résultats dépendant des centres d'intérêts de l'utilisateur lors de la recherche. [Mariam D, 09]

7 Conclusion

Dans ce chapitre, nous avons parlé des notions et les concepts de la recherche d'information classique, suivi par les systèmes de RI ou nous l'avons expliqué le processus d'indexation des documents, la reformulation de la requête et l'appariement document requête. Et aux points dont nous avons présenté les modèles de RI, et on a évalué des systèmes de RI par deux points : collection de test et les mesures d'évaluation.

Dans la recherche d'information classique ça ne prend pas en considérations l'utilisateur et généralement les requêtes sont ambigües cela nous donne des résultats qui ne suffisent pas les besoins en information de l'utilisateur et c'est ce qui nous amène à la recherche d'information personnalisés. Et à partir d'ici et dans le chapitre suivant, nous aborderons la RI personnalisés.

Chapitre II :
La recherche d'Information Personnalisée

1 Introduction

La récupération d'informations personnelles est une méthode de recherche qui vise à fournir des résultats de recherche adaptés aux besoins spécifiques de l'utilisateur. Il s'appuie sur les informations recueillies auprès de l'utilisateur, telles que les préférences, les intérêts et les habitudes de recherche, pour personnaliser les résultats de recherche et les rendre plus pertinents pour eux.

Cette méthode de recherche d'informations permet aux utilisateurs de trouver plus facilement et plus rapidement les informations qu'ils recherchent en limitant les résultats de recherche à ceux qui les concernent le plus. Il peut également permettre aux utilisateurs de découvrir de nouvelles informations pertinentes qui, autrement, ne seraient pas apparues dans les résultats de recherche non personnalisés.

Les technologies de recherche d'informations personnelles utilisent souvent des algorithmes d'apprentissage automatique et des systèmes de recommandation pour analyser les données des utilisateurs et fournir des résultats de recherche adaptés à leurs besoins. Ces technologies sont couramment utilisées dans des domaines tels que la recherche sur le Web, la recherche de produits en ligne, la recherche de contenu sur les réseaux sociaux, etc.

Cependant, il est important de noter que la recherche d'informations personnelles peut soulever des problèmes de confidentialité et de sécurité des données. Les utilisateurs doivent être conscients des informations qu'ils fournissent et de la manière dont elles sont utilisées pour personnaliser les résultats de recherche.

Dans ce chapitre, nous projetons la lumière sur l’émergence de la RI personnalisée, la modélisation de l’utilisateur qui contiennent la pertinence, le profil d’utilisateur, l’acquisition des sources d’information et des données utilisateur …. Et on termine par l’évaluation des systèmes de recherche personnalisés.

2 Emergence de la recherche d’information personnalisée

Le développement des technologies de l'information et de la communication a fait émerger la recherche d'informations personnelles. Qui vise à fournir des résultats de recherche plus pertinents et personnalisés en fonction des intérêts et des besoins de chaque utilisateur.

Par conséquent, les moteurs de recherche tels que Google et Bing ont commencé à incorporer des algorithmes de personnalisation pour améliorer les résultats de recherche en fonction de l'historique de recherche, des intérêts et des préférences d'un utilisateur.

Par exemple, lorsque nous avons formulé la requête suivante « virus » sur le moteur de recherche Google, nous avons remarqué, comme le montrent les figures (n° 2.1, 2.2 et 2.3), que les dix premiers résultats sont Comme suit :



Figure 2.1 : de 1ere page au 5eme résultat Google sur « Virus ».

Les cinq résultats suivants (d'un à cinq) sont des sites Web qui parlent de types de virus répandus dans le monde qui mettent en danger la santé de l'organisme (Fig. N 2.1).



Figure 2.2 : 2eme page de résultat Google sur « Virus ».

Les cinq résultats suivants (d'un &à cinq) e de la 2eme page, sont toujours sur les types de virus qui menacent la santé de l'organisme (Fig. N 2.2).

Dans ce cas, si l'utilisateur qui a soumis cette demande est médecin ou intéressé par le domaine de la santé, alors il sera satisfait de ces résultats de recherche. Mais si cet utilisateur est un informaticien et veut des informations sur les virus liés aux ordinateurs, il trouvera très mauvaise performance dans les 5 premiers résultats dans la 1ere et la 2eme page car ils ne sont pas pertinents selon les besoins de l'utilisateur respectif.

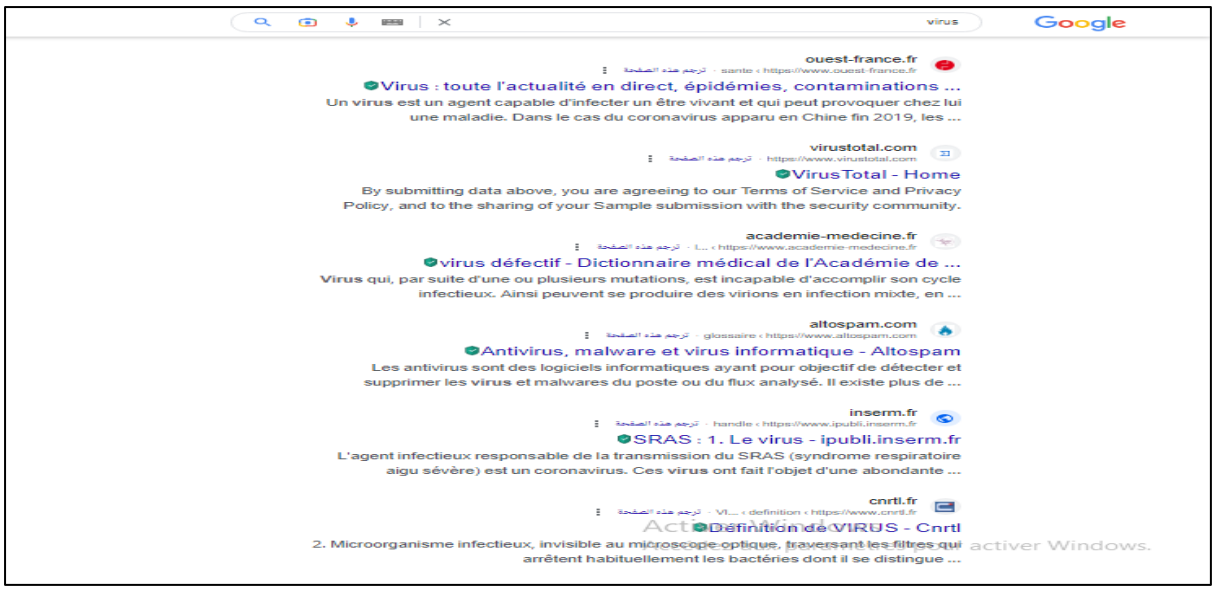


Figure 2.3 : 3eme page de résultat Google sur « Virus ».

Au vu des résultats renvoyés, cet utilisateur trouvera ce qu'il veut (virus informatiques) dans la quatrième réponse de la troisième page, mais il sera satisfait s'il trouve ce qu'il cherche en haut de la liste des résultats renvoyés par le système (Fig. N 2.3).

En bref, l'émergence de la recherche d'informations personnelles vise à répondre aux besoins spécifiques des utilisateurs dans un monde où l'abondance et la complexité de l'information ne cessent de croître.

3 Modélisation de l'utilisateur

La modélisation de l'utilisateur est une discipline de recherche datant des années 70 et évoquant en premier lieu les travaux d'Allen, Cohen et Perrault [All, 79] [CP, 79]. Dans le but d'améliorer la qualité des interactions homme-machine.

Les systèmes qui adaptent leur comportement aux besoins de l'utilisateur individuel ont souvent une structure explicite de représentation qui contient des informations sur leurs utilisateurs ; cette structure s'appelle généralement un modèle d'utilisateur.

Le modèle utilisateur est une source de connaissances, une base de données sur un utilisateur.

Plus précisément il représente un ensemble de données persistantes qui caractérisent un utilisateur ou un groupe d'utilisateurs particuliers en contenant des caractéristiques sur les préférences, les connaissances, les objectifs, les centres d'intérêt, etc. de l'utilisateur.

En général, dans la littérature, les termes « modèle utilisateur » ou « profil utilisateur » signifient la même appellation. Mais d'après Koch [KOC, 00], une différence existe entre le profil utilisateur et le modèle. Il définit le profil utilisateur comme une version simple du modèle utilisateur, Le profil utilisateur peut être vu comme une collection d'informations personnelles caractérisant un utilisateur telles que les buts, besoins, compétences, préférences, contexte d'utilisation, etc. Ces informations sont stockées sans aucune interprétation ou description. Ces informations sont généralement représentées par des couples (attribut, valeur). Ces valeurs sont soit stables soit évolutives. Le profil utilisateur est utilisé pour construire le

modèle utilisateur. Donc ce dernier est vu comme une représentation des croyances du système sur un utilisateur donné.

[KOC, 00] définit l'application du modèle utilisateur comme suit: « Users are different: they have different background, different knowledge about a subject, different preferences, goals and interests. To individualize, personalize or customize actions a user model is needed that allows for selection of individualized responses to the user."

Donc le modèle utilisateur et le profil utilisateur sont nécessaires dans le processus de personnalisation.

Dans ce qui suit on va donner un aperçu des différentes approches et technique de modélisation.

3.1 Pertinence utilisateur

La pertinence est incontestablement la question fondamentale posée lors de l'accès à l'information.

La pertinence est une notion subjective, non généralisable à tout type d'information et valide pour tout type d'utilisateurs. Elle dépend notamment du centre d'intérêt ou domaine d'application, du moment, du lieu et du support que l'utilisateur a choisi pour accéder à l'information et du système qui délivre cette information.

Cette notion subjective, dépendant essentiellement du point de vue de l'utilisateur, a de nouveau été l'objet d'investigations dans le cadre de l'accès personnalisé à l'information de manière générale et de la RI personnalisée de manière particulière. Dans ce contexte, la pertinence est spéculée comme étant un concept multidimensionnel [P. Borlund, 03], dont on distingue

Principalement quatre types :

- Pertinence algorithmique : la pertinence est traduite par une mesure algorithmique dépendant des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte ;
- Pertinence thématique : la pertinence traduit le degré d'adéquation de l'information à couvrir, en partie, le thème évoqué par le sujet de la requête. C'est le type de pertinence adressé par les assesseurs de la campagne d'évaluation TREC ;
- Pertinence cognitive : c'est la pertinence liée au thème de la requête, « pondérée » par la perception ou les connaissances de l'utilisateur sur ce même thème ;
- Pertinence situationnelle : c'est la pertinence liée à la tâche de recherche. Ce type de pertinence traduit essentiellement l'utilité de l'information relativement au but de recherche de l'utilisateur.

La RI personnalisée explore essentiellement la pertinence cognitive et la pertinence situationnelle.

3.2 Profil d'utilisateur

Dans le but de personnaliser l'information, le SRIP doit disposer des éléments ayant une incidence concrète sur la recherche en cours. Ces éléments représentent les données contenues dans le profil utilisateur et les métadonnées des documents. En effet, l'utilisateur et le document ne sont pas assimilés seulement à des descripteurs exprimés à l'aide de mots comme c'est le cas dans les SRI classiques, ils possèdent tous deux des caractéristiques propres.

Le profil utilisateur est toute structure qui permet de modéliser et de stocker les données caractérisant l'utilisateur au sein d'un système. Grâce à ce profil, le système possède une connaissance sur les centres d'intérêts, les préférences, et les besoins en informations de l'utilisateur ou un groupe d'utilisateur grâce à lesquelles il peut lui proposer des repenses plus adaptés. Pour modéliser le profil il faut décrire le « quoi » c'est-à-dire ce qui doit être représenté et le « comment » c'est-à-dire comment représenter le « quoi » au sein du profil [Amato 99] [Amadieu et al. 09].

- Le Quoi : déterminer ce que doit représenter le contenu du profil.
- Le Comment : déterminer la structure de contenu du profil et les techniques utiliser pour la construction de ce profil.

3.3 Sources des données de profil

Le contenu de ce profil constitue le représentant d'une source d'information. Il décrit les informations que la source contient et leur nature. La nature des informations représente les critères (caractéristiques) de la source. Nous définissons le profil d'une source par les attributs suivants :

- Identité : est un attribut qui permet d'identifier la source, nous pouvons utiliser l'url et le nom,
- Contenu : est un attribut qui donne une vue du contenu de l'ensemble des documents dans la source.
- Critères-documents : est un multi-attribut qui nous informe sur la nature des documents dans la source, à savoir, les types des documents, les langues des documents, la fraîcheur des documents et le coût des documents.

La figure suivante (figure 5.1) montre le schéma d'un profil source :

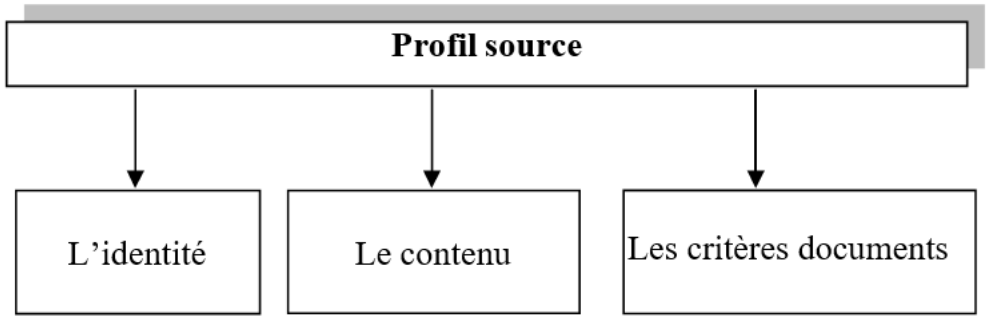


Figure 2.4: Profil source.

Instanciation du profil source

L'étape de l'instanciation du profil source consiste à initialiser ces différents attributs :

- Identité : l'url et le nom d'une source sont fixés au moment de l'implémentation du prototype.
- Contenu : le contenu d'une source est représenté par un vecteur (ensemble) de termes pondérés, en utilisant le modèle vectoriel [Salton et al 1975], qui représente la source comme un gigantesque document. Nous utilisons le schéma $tf * idf$ pour la

pondération des termes. Ces termes sont extraits des k premiers documents retournés par la source pour la requête de l'utilisateur.

- Critères-documents : est un multi-attribut qui permet de mesurer l'importance de chaque critère dans la source. Chaque critère possède un ensemble de valeurs, pour chaque valeur nous associons un poids. Le poids d'une valeur d'un critère mesure son importance dans la source.

3.4 Acquisition des données de profil

L'acquisition des sources d'information et des données utilisateur est l'une des problématiques importantes à aborder dans le but de d´développer un SRI personnalisé efficace. On distingue deux modes d'acquisition des données utilisées dans la construction du profil utilisateur : acquisition explicite et acquisition implicite.

3.4.1 Acquisition explicite

L'acquisition explicite des données utilisateur repose principalement sur les techniques de feedback explicite largement utilisées dans les systèmes de filtrage et de reformulation de requêtes par réinjection de pertinence. Ces techniques sont utilisées dans le cadre de la RI personnalisée dans le but de maintenir et faire évoluer le profil de l'utilisateur. L'acquisition explicite permet à l'utilisateur de saisir manuellement ses domaines d'intérêts ou alors juger de façon explicite les documents renvoyés pour une requête servant à la construction de son profil. Ce type d'acquisition est utilisé dans plusieurs systèmes de RI personnalisée en ligne. Googles Alertés permet à l'utilisateur de saisir une requête profil servant comme critère de sélection de l'information. Google personalized search dans sa version 1.1 (2004) permet à l'utilisateur de saisir explicitement un ensemble de catégories/concepts qui représentent ses domaines d'intérêts. La recherche contextuelle mise en œuvre par yahoo dans sa version Y ! Q [R. Kraft et al. 2005] permet à l'utilisateur de sélectionner explicitement un texte d'une page en cours de lecture et qui sert à représenter un profil textuel lié `à la recherche courante. L'approche dans [A. Sieg et al. 2004] requiert que l'utilisateur sélectionne parmi l'ensemble des concepts de l'ontologie de l'ODP une paire de concepts, un adéquat et un autre inadéquat au besoin en information, utilisés dans un processus de reformulation personnalisée de la requête.

Plusieurs autres systèmes en RI personnalisée adoptent ce mode d'acquisition. [F. Liu et al. 2004] propose à l'utilisateur de sélectionner les concepts du profil issus de l'ODP, dans un processus de clarification du besoin derrière la requête et qui servent comme contexte de recherche. [G. Koutrika et Y.I.A, 05] et les systèmes ifWeb, Wifs et InfoWeb utilisent un feedback explicite de l'utilisateur permettant d'identifier les documents pertinents utilisés dans la dérivation du réseau de termes représentant le profil utilisateur. Même si ces techniques servent à la construction contrôlée du profil utilisateur, cependant elles présentent des limitations. D'une part, l'acquisition explicite ne permet pas d'améliorer la représentation du modèle de l'utilisateur efficacement, plus particulièrement si les caractéristiques de l'interface d'acquisition explicite ne sont pas très bien assurées [A. Waern, 04]. D'autre part, l'effort supplémentaire imposé à l'utilisateur à spécifier explicitement ses besoins est indésirable pour des utilisateurs Web. Ces limitations ont orienté les travaux vers des techniques d'acquisition implicite des données utilisateur dans le processus de construction du profil utilisateur.

3.4.2 Acquisition implicite

Une approche alternative remplaçant l'acquisition explicite des besoins en information de l'utilisateur, consiste à développer des algorithmes d'acquisition implicite de ces besoins. L'acquisition implicite des données utilisateurs consiste à collecter les données de l'utilisateur en observant ses comportements et interactions avec le système durant les activités de recherche. L'avantage de cette approche est qu'elle ne n´nécessite aucune implication directe de l'utilisateur. Nous présentons dans cette section la collecte implicite des données utilisateurs exploités dans la dérivation du profil utilisateur ainsi que les indicateurs de pertinence implicite utilisés dans cette collecte implicite.

Collecte implicite données utilisateurs : Le processus clé dans la plupart des approches de personnalisation consiste à exploiter des sources d'évidence issues de l'historique de recherche de l'utilisateur, afin d'inférer son profil. Letizia [H. Lieberman, 95], WebMate [L. Chen et K. Sycara, 98], Per- sonalWebWatcher [D. Mladenic, 98] et OBIWAN [S. Gauch et al. 2003] construisent le profil en analysant les pages web visitées par l'utilisateur lors de sa recherche. D'autres sources d'information sont également exploitées, telles que les annotations dans Basar [C. G. Thomas et G. Fischer, 97], les requêtes et ses résultats de recherche dans [E. Rich, 98], Syskill and Webert [M. J. Pazzani et al. 1996], et Persona [F. Tanudjaja et L. Mui, 02], les requêtes et les résumés textuels de ses résultats [X. Shen et al. 2005]. La combinaison des sources d'évidence multiples, telles que les pages web, les emails et les documents textes, est adoptée dans [S. Dumais et al. 2003]. Comme source de personnalisation, plusieurs systèmes de re- cherche sur ordinateur, notamment, Copernic, Google Desktop Search, Mac Tiger, Windows Desktop Search, Yahoo ! Desktop Search ou X1, et d'autres outils de recherche fournissent un accès simple à l'index de l'information créé, copié, visualisé par les utilisateurs. Microsoft's Stuff I've Seen (SIS) [J. Teevan et al. 2005] fournit aussi ce type de personnalisation.

3.5 Techniques de construction de centres d'intérêts

La construction du profil utilisateur consiste à organiser et extraire les éléments qui constituent le profil à partir des données de l'utilisateur collectées lors de l'étape précédente, de l'étape précédente. On distingue différentes technique s de construction du profil utilisateur, basées sur l'analyse statistique de termes, le filtrage collaboratif et les techniques d'apprentissage [Gauch et al., 2007].

- **Analyse statistique de termes**

Cette approche permet d'analyser le contenue d'un document et d'extraire les mots clés significatifs qui décrivent son contenu. Ces mots clés sont stockés pour être utilisés ensuite afin de comparer des éléments entre eux ou avec les préférences de l'utilisateur lorsqu'elles sont exprimées sous la forme de mots clés.

Il existe différentes structures de stockage et de représentation des mots clés, on distingue trois principales techniques : l'extraction des termes, l'extraction de réseaux de termes et l'extraction de concepts.

- **Extraction d'ensemble de termes :** se base sur des techniques d'extraction des termes à partir des données visitées par l'utilisateur. Ces termes vont servir dans l'algorithme d'apprentissage du modèle de l'utilisateur.

- Parmi les systèmes appliquant cette approche de construction, on peut citer les systèmes WebMate [Chen et Sycara, 1998] et Alipes [Widyantoro et al., 1999].
- **Extraction de réseaux de termes :** les termes sont extraits des documents jugés par l'utilisateur et représenter sous forme de réseau de nœuds. Afin de construire le profil utilisateur, il est nécessaire d'exploiter des relations préexistantes entre les termes et les concepts. Le système SiteIF [Stefani et Strappavara, 1998] utilise cette technique en se basant sur le dictionnaire WordNet.

Extraction de concepts : concerne principalement les profils utilisateurs représentés par une hiérarchie de concepts pondérés, elle utilise une taxonomie de concepts de référence comme profil de base. Les principales ressources sémantiques utilisées sont ODP 3 un annuaire de concepts hiérarchique open source et la hiérarchie de concepts Yahoo, L'approche de construction présente de manière générale les deux étapes suivantes :

(1) identification des concepts et niveaux de l'ontologie à exploiter

(2) extraction des centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie.

- **Filtrage collaboratif**

Cette approche consiste à comparer les profils des utilisateurs entre eux afin de trouver des similarités dans leur comportement et dans leurs préférences et de faire des recommandations sur la base de ces similarités. Le filtrage collaboratif est utilisé le plus souvent dans les systèmes de recommandation d'éléments lorsque les profils des utilisateurs sont composés par un ensemble de couples (objet, pertinence de l'objet).

Ce type de filtrage est caractérisé par sa flexibilité lors du passage à l'échelle (…) et l'indépendance du format des objets manipulés

- **Techniques d'apprentissage**

Le principe de base de ces techniques est l'étude du comportement de l'utilisateur et la classification de ses caractéristiques ou celles des objets recherchés. Ces techniques sont caractérisées par leur fraîcheur et l'exactitude des données dérivées. Leur inconvénient est lié à la complexité temporelle des algorithmes utilisés Des exemples de telles techniques sont les réseaux de neurones, les méthodes de classification (raisonnement par cas, classificateurs bayésiens...), les règles d'association.

3.6 Représentation des données de profil d'utilisateur

La représentation de l'utilisateur à travers un profil permet de cibler ses besoins spécifiques et à prendre en considérations ses préférences à travers l'amélioration du comportement des systèmes informatiques. Un modèle de représentation de profil permet d'organiser ces éléments afin de faciliter leur exploitation par les applications. On distingue quatre principales approches de représentation de profil utilisateur : Historique, Ensembliste, Connexionniste, Multi-dimensionnelle et Conceptuelle.

Category behavior	Segment	Object	Class
Examine	View Listen	Select	
Retain	Print	Bookmark Save Delete	Subscribe

		Purchase	
Reference	Copy-and-pasteQuote	Forward Reply Link Cite	
Annotate	Mark up	Rate Publish	Organize

TAB. 2.1 Classification des comportements utilisés comme mesures de pertinence implicite par [D. W. Oard and J. Kim, 01]

Représentation implicite d'un profil utilisateur basée sur l'historique de recherche. Le profil est ainsi représenté par requêtes et les sites Web que l'utilisateur a sélectionné de l'ensemble des résultats retournés en réponse à la requête.

3.6.1 Représentation basée sur l'historique de recherche

Cette représentation est la plus répondue dans le domaine de la recherche d'informations. L'historique de recherche consiste en un ensemble des requêtes, des pages Web visitées ou cliqués ou des résumés textuels des résultats associés accumulés au cours des sessions de recherche de l'utilisateur.

Nous citons dans cette catégorie, les travaux des auteurs [V.Raghavan et H.Sever, 95] qui représentent le profil utilisateur par son historique de recherche en se servant comme une base de données des requêtes soumises précédemment ainsi que leurs résultats associés. Cette base de données est utilisée pour sélectionner les requêtes les plus similaires à une requête en cours d'évaluation. Aussi l'approche proposée par l'article [B.Tan et al. 06], où le profil utilisateur est représenté par l'historique des requêtes passées et l'historique des clicks sur les résultats sélectionnés par l'utilisateur.

3.6.2 Représentation ensembliste

La représentation ensembliste du profil utilisateur est basée sur un ensemble de mots clés (ou vecteurs de mots clés) pondérés représentés souvent selon le modèle vectoriel [G.Salton et C.Yang, 97]. Ce type de représentation est le premier conçu pour modéliser le profil utilisateur.

Les paquets de termes traduisent les centres d'intérêts de l'utilisateur. Le poids d'un terme est souvent calculé selon le schéma $TF * IDF$ communément utilisé en RI et représente le degré d'intérêt de l'utilisateur dans le profil. Dans [M.Daoud, 09], l'auteur divise la représentation ensembliste en trois sous-modèles de représentation :

- Un ensemble de termes pondérés où chaque terme représente un centre d'intérêt possible de l'utilisateur,
- Un vecteur de termes pondérés représentant un centre d'intérêt [H.Lieberman, 97],
- Un ensemble des vecteurs de termes pondérés dont chacun représente un centre d'intérêt [J.Gowan, 03].

La figure 2.5 montre un exemple de profil utilisateur représenté par un ensemble de vecteurs de mots clés.

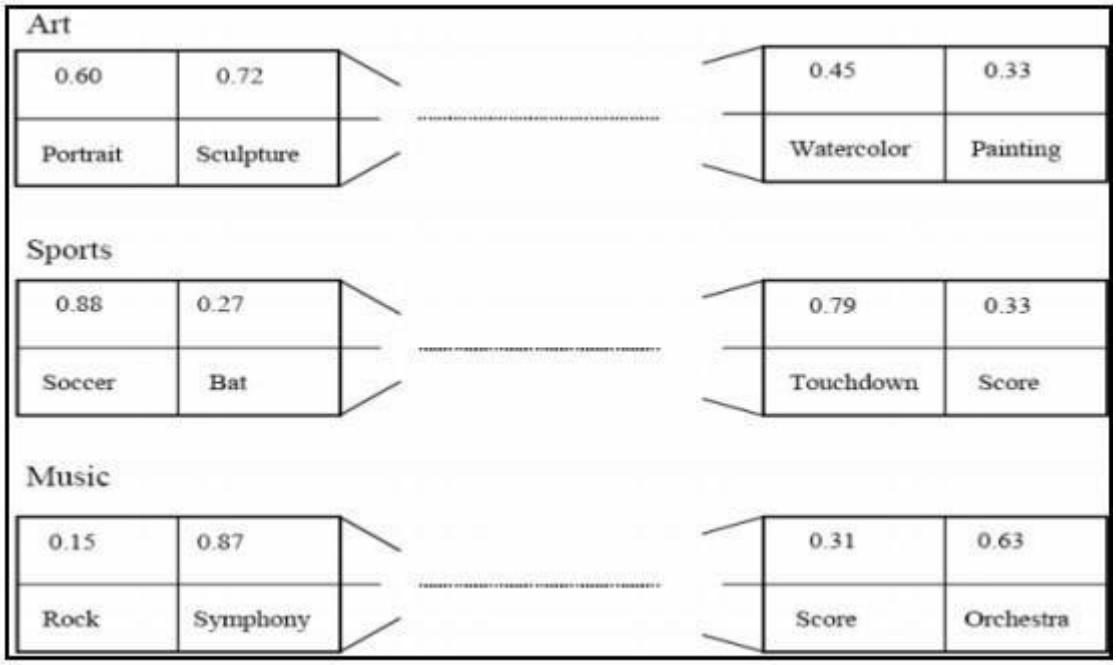


Figure 2.5 – Exemple de profil représenté par des vecteurs de mots clés. [M.Daoud, 09].

La construction d'un profil ensembliste repose sur des techniques d'extraction des termes à partir des documents pertinents, jugés implicitement ou explicitement par l'utilisateur. Plusieurs systèmes d'accès personnalisé à l'information adoptent ce type de représentation. Tel est le cas des portails web tels que MyYahoo, InfoQuest, Anatagonomy qui est un système personnalisé de consultation de nouvelles et de journaux en ligne [H.Sakagami et T.Kamba, 97].

3.6.3 Représentation connexionniste

La représentation connexionniste du profil utilisateur consiste à représenter les centres d'intérêts de l'utilisateur par un réseau de nœuds pondérés dont chaque nœud représente un concept traduisant un centre d'intérêt de l'utilisateur. Cette représentation permet de résoudre les failles de la représentation ensembliste par la mise en place des relations de corrélation sémantiques entre les centres d'intérêts du profil. En effet, la richesse sémantique dans cette représentation permet de résoudre le problème de la polysémie des termes inhérents à la représentation ensembliste, l'incohérence possible entre les centres d'intérêts et l'identification d'un profil adéquat au sujet de la requête via les relations sémantiques [M.Daoud, 09]. Certains auteurs tels que [A.Stefani et C.Strapparava, 98] [G.Gentili et al. 03], se sont intéressés à la représentation du profil utilisateur par un réseau de nœuds pondérés dans lequel chaque nœud représente un concept traduisant un centre d'intérêt utilisateur. Ce type de représentation offre le double avantage de la structuration et de la représentation associative permettant de considérer l'ensemble des aspects représentatifs du profil. Les concepts composant le profil sont souvent représentés par des relations de paires de nœuds. L'auteur dans [M.Pazienza, 97], suggère de créer les arcs reliant deux nœuds sur la base des co-occurrences entre ses termes. La figure 2.6 représente un extrait d'un profil utilisateur sémantique.

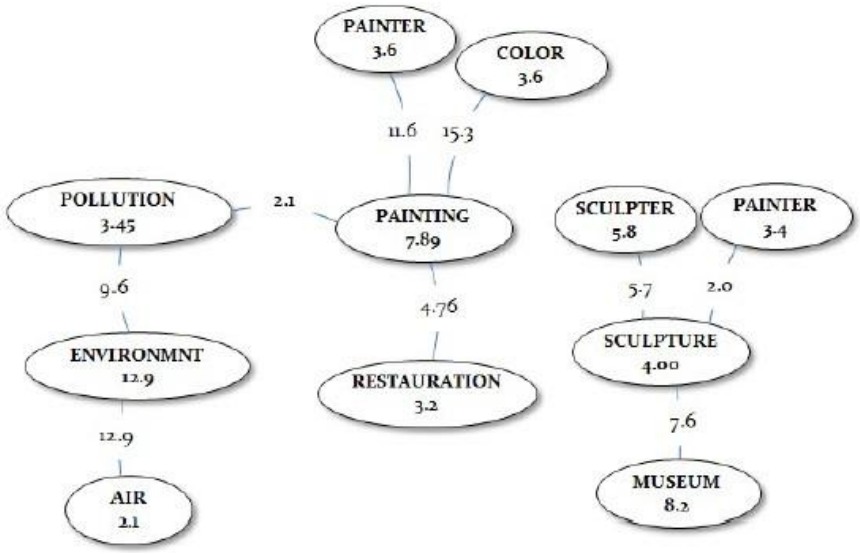


Figure 2.6 – Un extrait d'un profil utilisateur sémantique. [M.Pazienza, 97].

3.6.4 Représentation multidimensionnelle

Les auteurs dans [D.Kostadinov, 08] [G.Amato et U.Straccia, 99] ont adopté ce type de représentation pour le profil utilisateur. Il consiste en un ensemble de dimensions représentant chacune un aspect particulier (comme par exemple les données personnelles, le domaine d'intérêt). Son contenu est représenté par un modèle structuré de dimensions (ou catégories) prédéfinies dans [G.Amato et U.Straccia, 99] ou bien par une structure générique [D.Kostadinov, 08] (2.3) où chaque dimension est constituée d'un ensemble d'attributs éventuellement organisés en sous dimensions offrant ainsi, la possibilité d'élargir l'ensemble des données représentées en fonction de domaine d'application et des besoins utilisateur. Les attributs peuvent être simples ou composés. Une sous dimension regroupe un ensemble d'attributs simples qui sont liés sémantiquement (par exemple l'adresse est composée du numéro de la rue, du nom de la rue, du code postal etc.…).

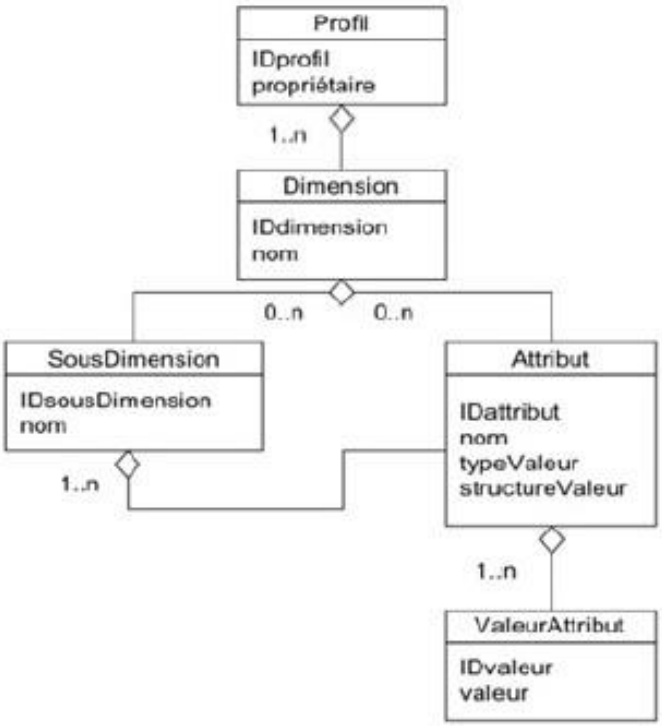


FIGURE 2.7 – Métamodèle de profil utilisateur [D.Kostadinov, 08].

3.6.5 Représentation conceptuelle

La représentation du profil met en évidence, dans cette approche, une ou plusieurs relations sémantiques entre les informations du profil. Suivant une direction proposée dans un contexte plus général [M.N. Huhn et L. Stephens, 99], cette représentation offre une alternative intéressante à l'approche des réseaux sémantiques. En effet, les travaux actuels tendent à représenter le profil sous forme d'une ontologie de concepts personnels en se basant sur les connaissances contenues dans les ontologies plutôt que de construire les profils utilisateur seulement à partir des données implicites collectées de ses interactions [J. Chaffee et S. Gauch].

La représentation est essentiellement basée sur l'utilisation d'ontologies ou de réseaux probabilistes.

Dans ce type d'approche, les liens entre les concepts sont explicitement induits de l'ontologie et le profil résultant inclura des relations informationnelles plus diverses et spécifiques. La représentation conceptuelle est semblable à la représentation par réseaux sémantiques, dans le sens où elle représente les centres d'intérêts de l'utilisateur par un réseau de nœuds conceptuels. Cependant, dans l'approche conceptuelle, les nœuds correspondent à des domaines abstraits représentant les centres d'intérêt de l'utilisateur, contrairement à l'approche sémantique où les centres d'intérêts sont représentés par des mots spécifiques, ou ensemble de mots relatifs. La représentation conceptuelle peut également être assimilée à une approche ensembliste (vectorielle) du fait que les domaines sont souvent représentés comme les vecteurs de termes pondérés. Néanmoins, les termes de ces vecteurs sont regroupés pour former un domaine spécifique et non de simples mots clés.

De l'association des centres d'intérêts de l'utilisateur aux concepts des domaines de l'ontologie, on obtient un profil représenté sous forme d'une hiérarchie de concepts. Les données implicites issues des interactions de l'utilisateur sont classifiées dans ces concepts et

l'intérêt de l'utilisateur pour de tels concepts est enregistré. Plusieurs mécanismes peuvent être appliqués pour exprimer le degré d'intérêt de l'utilisateur pour chaque domaine. Néanmoins la technique la plus utilisée est l'affectation d'un poids d'une valeur numérique pour chaque domaine [S. Gauch et al. 03].

3.7 Evolution de 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'intérêt des utilisateurs au cours du temps. L'évolution du profil utilisateur se fait sou- vent selon un processus incrémental basé sur l'addition de nouvelles informations dans la représentation du profil.

La gestion de l'évolution du profil utilisateur consiste principalement à capturer les changements des centres d'intérêts de l'utilisateur dans une première phase et propager ces changements au niveau de la représentation du profil. Autrement dit, L'évolution de la représentation du profil utilisateur implique un changement des degrés d'intérêts dans certains domaines qui se traduit par une mise à jour de la structure/contenu des centres d'intérêts préalablement appris ou alors l'apparition d'un nouveau besoin en information qui se traduit par un ajout d'un nouveau centre d'intérêt au profil utilisateur [M.Daoud, 09].

4 Les dimensions du contexte de recherche

De nombreux travaux en recherche d'information [Borlund 2003] [Taylor et al. 2007] [Saracevic 2007], ont prouvé que le concept de pertinence a un caractère multidimensionnel, et ont identifié plusieurs critères, au-delà de la dimension thématique, qui peuvent affecter la perception de pertinence de l'utilisateur. Ce qui a poussé vers une direction de recherche qui consiste à considérer, à la fois, tous les critères ayant un impact sur la notion de pertinence et pouvant influencer les jugements de pertinence des utilisateurs, afin d'améliorer la précision des systèmes de recherche.

Pour améliorer la représentativité des profils et leur capacité à décrire les intérêts des utilisateurs et d'inférer leurs besoins en information, les approches de la recherche d'information personnalisée exploitent, comme capteurs de pertinence utilisateur :

En premier lieu, les données thématiques issues de l'historique de recherche pertinent de l'utilisateur. En effet, le contenu (thèmes, titres, termes, mots-clés, …etc.) des documents jugés pertinents par l'utilisateur est une source large et riche d'information qui peut mieux aider à identifier les centres d'intérêts de l'utilisateur. La dimension de pertinence thématique ou tropicale qui traduit la proximité des sujets des requêtes et de documents est la plus couramment utilisée, et plusieurs travaux ont souligné que ce critère thématique est la dimension principale de pertinence et que tous les autres critères en sont dépendants [Moulahi 2015].

En second lieu, d'autres capteurs supplémentaires de pertinence liés aux différentes dimensions contextuelles de recherche sont introduits pour mieux cerner la pertinence utilisateur. Un nombre considérable d'études ont souligné l'aspect multidimensionnel de la pertinence utilisateur. On cite la dimension sociale, localisable, temporelle, et même sémantique.

Nous présentons dans ce qui suit, les différentes dimensions contextuelles des informations représentant des capteurs de pertinence utilisateur, que le profil d'utilisateur doit les identifier et les exploiter pour l'amélioration de la performance des résultats de recherche. Il est à noter,

comme c'est affirmé dans [Buidghaghen 2011], qu'il est difficile de trouver un système de recherche personnalisé qui gère toutes les dimensions contextuelles en même temps. Les dimensions à considérer lors de la prise en compte du contexte peuvent être plus ou moins pertinentes selon l'application de recherche.

4.1 Dimension sociale

Dans le cadre de la recherche d'information personnalisée, les informations sociales sont utilisées pour construire les profils des utilisateurs. Deux informations sociales principales sont utilisées :

Les annotations sociales (Tags)

Les tags employés par les utilisateurs pour annoter les documents sont souvent utilisés pour déterminer leurs centres d'intérêts. Les annotations sociales sont 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].

Plusieurs travaux ont été proposés pour estimer le poids de chaque tag, où ce poids reflète son importance pour l'utilisateur. Certaines approches proposent une représentation vectorielle des tags de l'utilisateur. Le poids de chaque tag peut suivre un schéma de pondération basé sur la fréquence d'utilisation de ce tag par l'utilisateur (TF) [Noll et Meinel 2007], ou par un TF-IDF [Xu et al. 2008], ou une combinaison du modèle TF-IUF et du modèle BM25 [Vallet et al. 2010].

Selon [Heymann et al. 2008], il est constaté que 50% des tags utilisés pour annoter les documents sont des termes du document. Ceci démontre que certains utilisateurs ont tendance à employer des termes du document comme tags pour les annoter car de leur point de vue ces termes décrivent bien le sujet du document. Une autre étude [Bischoff et al. 2008], a montré que les tags des utilisateurs sont corrélés avec leurs motivations d'annotation. Un utilisateur peut annoter un document dans le but de le retrouver par la suite (par exemple tag : à lire), ou pour l'envoyer à quelqu'un (par exemple tag : à partager), …etc.

Les relations sociales entre utilisateurs du système de recherche

L'exploitation des relations sociales entre utilisateurs du système de recherche permet d'enrichir le profil de l'utilisateur, où la pertinence prédite d'un document donné peut être améliorée en explorant les actions du voisinage de l'utilisateur. La notion d'importance sociale des utilisateurs est aussi introduite, et elle suppose qu'un document pertinent est produit par un auteur important d'où l'importance scientifique des ressources informationnelles peut être dérivée à partir de l'importance sociale des auteurs associés [Bouhini 2014] [Benjabeur et al. 2010].

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.

Plusieurs approches de la RI personnalisée, ces dernières années, exploitent les annotations et les relations sociales pour la construction des profils des utilisateurs. [Badache et al. 2019]

proposent, dans un cadre de recherche d'information sociale en langue arabe sur Facebook, une approche qui sélectionne et exploite les meilleurs traces des utilisateurs (polarité, partage, j'aime, haha) laissées sur des publications Facebook pour estimer leur importance sociale. L'objectif est de montrer comment ces signaux peuvent jouer un rôle vital dans l'amélioration de la recherche en langue arabe sur Facebook. [Monteserin et Armentano 2019] proposent une approche de prédiction de liens entre les utilisateurs d'un réseau sociale. Ils définissent la notion d'influence sociale pouvant être exercée sur un utilisateur donné par les autres utilisateurs de réseau par tendance de l'utilisateur à accepter leurs suggestions et recommandations. Ils exploitent la transitivité pour prédire les liens indirects. [Bouhini et al. 2016], proposent une approche de personnalisation de la recherche basée sur la combinaison des annotations sociales de l'utilisateur et les termes pertinents de son profil. [Saoud et Kechid 2016] présentent, dans un cadre de recherche d'information distribuée, une approche de personnalisation combinant les annotations et les relations sociales de l'utilisateur pour la construction de son profil. [Ould Amer et al. 2015] proposent une approche de recherche de conversations dans les tweets basée sur la combinaison de pertinence textuelle et sociale des tweets. Plusieurs facteurs sont considérés dans la pertinence sociale du tweet comme les hashtags (#) et les URLs qu'elle contient, ainsi que l'expertise du blogueur qui la publié et son influence sociale représentée par le nombre de ses tweets favorisées ou retweetées par les autres blogueurs. [Shafik et al. 2015] ont proposé une nouvelle approche de personnalisation de la recherche sur le Web, qui extrait les préférences des utilisateurs en se basant sur les activités social des utilisateurs sur un réseau social. Les préférences de l'utilisateur sont exploitées pour reclasser les résultats retournés par le moteur de recherche.

4.2 Dimension localisable

L'orientation vers l'introduction de la dimension localisation de recherche est motivée par la prolifération de la technologie mobile (PDAs, téléphones mobiles, tablettes, …) qui a rendu l'accès à une grande masse de documents sur le web, possible à toute place et à tout moment. En effet, vu les contraintes et particularités techniques des appareils mobiles (des difficultés de saisie de requêtes, zone d'affichage limitée, …), nous assistons à une pratique de recherche différente de celle des requêtes traditionnelles de bureau.

L'hypothèse de base de cette orientation est que la pertinence d'un document pour une requête donnée dépend de l'emplacement de l'utilisateur au moment de formulation de la requête. Des études [Buidghaghen 2009], montrent que 25 % des besoins informationnels des utilisateurs mobiles sont liés à la localisation. On peut imaginer, par exemple, si un utilisateur lance la requete « echourouk » en Algérie, fort possible qu'il cherche le journal algérien Echourouk, mais s'il l'utilisateur est en Egypte, la réponse serait mieux le journal Echourouk de l'Egypte. Un autre exemple, si un utilisateur en déplacement qui émet la requête ―restaurant pas cher‖ s'intéresse plus aux restaurant proches de sa localisation que ceux qui sont dans une autre zone géographique. Donc le critère emplacement est très important dans ce contexte de recherche.

Plusieurs travaux en recherche d'information personnalisée exploitent l'information situationnelle ou de localisation (coordonnées GPS, nom de localisation, type de localisation, …etc.) et proposent de filtrer l'information sur la base de la proximité géographique pour améliorer les résultats de recherche d'information. Les auteurs dans [Akermi et al. 2015], proposent une approche de recommandation personnalisée dans un environnement mobile guidée par un profil situationnel de l'utilisateur. Ce profil est basé sur la localisation et les

préférences de l'utilisateur, et vise à identifier la bonne information à recommander en fonction de la situation dans laquelle se trouve l'utilisateur. [O. Bouidghaghen 2011], propose une approche de personnalisation de la recherche basée sur la construction d'un profil situationnel d'utilisateur composé des différentes situations de recherche et les centres d'intérêts attachées, et elle propose un modèle de classification des requêtes pour identifier leurs sensibilités à la localisation. Dans [Bila et al. 2008], les auteurs proposent un profil d'utilisateur mobile dans lequel les situations de l'utilisateur sont représentées par les lieux les plus visités par l'utilisateur, et les relations entre elles sont modélisées à base d'arbre. Les centres d'intérêt de l'utilisateur sont explicitement acquis à l'aide d'un questionnaire. Dans [Hattori et al. 2007], les auteurs traitent le problème de l'ambiguïté des requêtes en utilisant une méthode d'expansion basée sur la localisation. Leur méthode consiste à récupérer les coordonnées spatiales de l'utilisateur, les transformer en des mots contextuels (noms de places, d'activités liés à cette place) en utilisant un système d'information géographique et des techniques d'apprentissage des weblog, puis à pondérer ces mots contextuels sur la base de la comparaison de la probabilité globale du mot contextuel dans l'ensemble des documents du corpus cible et sa probabilité locale dans les documents retournés par la requête originale, en fin, la requête originale est étendue par les mots contextuels ayants les poids les plus élevés.

4.3 Dimension temporelle

Une autre dimension très importante pour la description des intérêts de l'utilisateur est la dimension temporelle ou évènementielle. La recherche d'information temporelle est un domaine de recherche dont l'objectif principal est l'amélioration des modèles classiques de la recherche d'information en exploitant les informations temporelles qui peuvent exister dans les requêtes et les documents. D'une façon générale, les modèles de recherche d'information combinent 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. Et 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 [Moulahi 2015].

L'information temporelle peut concerner le temps de soumission des requêtes utilisateurs ou le temps de publication des documents ciblés, comme elle peut concerner les expressions temporelles ou celles relatives aux évènements dans les contenus des documents.

Dans [Vandenbussche et al. 2011] est proposée une approche recherche d'information qui combine les informations temporelles sur le document et celles de son contenu. Les auteurs de [Kanhabua et al. 2012] propose un système temporel-sémantique pour l'apprentissage de l'extraction d'évènement et d'entités nommées à partir de texte. Dans [Arikan et al. 2009] la requête utilisateur est considérée porteuse de deux sens thématique et temporel. Le travail présenté dans [Basil et al. 2014] propose un système pour l'extraction et l'indexation d'évènements et la récupération des sujets liés à ces évènements à partir des textes. Dans [Panayiotou et al. 06], les auteurs exploitent l'importance du temps et de l'expérience (au travail, en vacances, etc.) dans la personnalisation d'un portail de recherche de services web pour un utilisateur mobile. Ils proposent de construire un profil dynamique où les centres d'intérêt sont pondérés selon des zones temporelles apprises par l'étude de la routine journalière de l'utilisateur et ses activités dans chaque zone.

4.4 Dimension sémantique

Plusieurs travaux de la recherche d'information personnalisée représentent le profil utilisateur à l'aide de l'ensemble de termes ou de mots clés issus de l'historique de recherche pertinent [Uddin et al. 2012]. Néanmoins, un processus de recherche basé uniquement sur la syntaxe (mots clés/termes), semble plus proche à une opération de reconnaissance de formes, qu'à une opération d'assimilation et de satisfaction d'un besoin cognitif de l'utilisateur exprimé par une requête souvent courte et ambigüe [Aimé et al. 2010]. Et sans l'exploitation des liens et des relations sémantiques entre les termes, la recherche d'information reste incapable de bénéficier de la richesse de la langue humaine pour une compréhension profonde de la requête, du besoin de l'utilisateur et du document cible [Fouad et al. 2012]. En effet, beaucoup de documents dont le contenu est sémantiquement similaire au sujet de la requête peuvent ne pas être sélectionnés, de fait qu'ils utilisent d'autres termes pour décrire différemment le même contenu informationnel.

La sémantique ne considère pas les documents comme un simple ensemble de termes sans liens entre eux, mais plutôt elle tient compte de la signification véhiculée par ces termes. 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 [Zargayouna et al. 2015].

Plusieurs travaux en recherche d'information personnalisée s'orientent, ces dernières années, vers l'introduction de la sémantique dans le profil d'utilisateur pour améliorer l'exactitude et la pertinence des résultats de recherche en mieux comprenant les intentions de l'utilisateur. [Mohammed et Abdelmoty 2017] proposent un profil spatial sémantique dans un réseau social basé sur les localisations. Ce profil est construit par la combinaison des informations concernant les différentes localisations visitées par l'utilisateur, et les annotations sociales qu'il a utilisées pour annoter ces localisations ou pour annoter les activités qu'il a effectuées dans ces localisations. Les auteurs utilisent la base de localisation « Foursquare » pour extraire les informations sémantiques des localisations (nom, catégorie, région…), et la base lexicale de l'anglais « WordNet » pour filtrer et catégoriser les annotations de l'utilisateur. [T. H. Duong et al. 2013] 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. [Daoud et al. 2011], dans leur approche de personnalisation sémantique proposent de modéliser le profil d'utilisateur et les documents sous forme d'un graphe de concepts. Ces concepts sont issus de l'ontologie ODP. Ils proposent aussi un modèle de ré-appariement (doc/profil) basé sur une mesure de distance sémantique qui combine le maximum et le minimum sous graphe de concept commun. 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 profil utilisateur dans le processus de recherche

A partir des informations continues dans le profil le SRIP va cibler un ensemble de documents. Ces informations sont représentées par une liste de termes, cette liste va être comparée aux descripteurs des documents de la collection. Les documents ayant un degré de similarité élevé seront sélectionnés et formeront le sous espace de recherche sur lequel portera la requête.

5.1 Reformulation de requête

La requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Elle représente le besoin à court terme de l'utilisateur, alors que le profil représente son besoin à long terme.

La requête initiale de l'utilisateur contient des valeurs de paramètres du profil utilisateur fournies directement par l'utilisateur. Ces paramètres représentent les classes d'intérêt de haut niveau dans la structure hiérarchique de profil utilisateur En choisissant une classe d'intérêts l'utilisateur exprime son but de recherche. A partir des choix le système va extraire les documents correspondants et effectuer la reformulation de la requête. En effet, le SRIP va considérer ces documents comme jugés pertinents afin de reformuler la requête initiale.

5.2 Reclassement de résultats

Dans cette phase l'interaction entre le profil et le module de représentation du résultat se fait à travers les catégories « Customisation » et « Catégories des données de l'environnent ». En fonction des informations fournies par ces catégories le SRIP va déployer des mécanismes adaptation afin de bien restituer ces résultats selon le contexte et les préférences de l'utilisateur.

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

Les méthodes d'évaluation largement adoptées en recherche d'information, sont souvent basées sur une évaluation d'avantage quantitative que qualitative. En effet, les résultats obtenus sont issus de la comparaison de mesures et de métriques en termes de rappel et précision, sur les réponses fournies par le système relativement à celles issues des réponses attendues qui constituent le référentiel. Ce type d'évaluation, est orienté vers une approche comparative de plusieurs systèmes reposant sur le principe d'évaluation des collections de tests. Bien qu'adopté par des campagnes d'évaluation de référence en recherche d'information, tel que TREC, il est cependant contesté [Balpe et 1996] notamment en raison de la non considération ni du contexte dans lequel se fait la recherche, ni de la perception de la pertinence des utilisateurs dans ce même contexte. L'introduction de la dimension utilisateur dans le processus d'accès à l'information, accentue d'avantage la difficulté d'évaluation. En effet, en raison du caractère subjectif des utilisateurs d'une part et du caractère dynamique ou adaptatif du système, d'autre part, il est difficile de fournir des valeurs absolues aux métriques. On cite dans ce qui suit les principaux problèmes liés à l'évaluation des systèmes personnalisés puis présentons quelques recommandations qui augmentent la fiabilité des résultats qui en sont issus [Lechani et Boughanem 2005].

6.1 Les limites de l'évaluation classique

Malgré la popularité et la reconnaissance de ces deux campagnes d'évaluation que sont TREC et CLEF, ces approches pour l'évaluation des systèmes de recherche d'informations présentent

certaines limites notamment au sujet de la prise en compte de l'usager, de la constitution des corpus et des requêtes mais également de l'évaluation elle-même.

Pour de mieux cerner les limites de l'évaluation classique des SRI, Nous nous sommes basés sur les travaux de [Tamine, 10] [Chaudiron, 02] [Menegon, 09]. Une synthèse de ces travaux nous a permis de définir trois (03) classes de problèmes, chacune de ces classes est liée à un acteur qu'on trouve généralement autours d'un processus d'évaluation. Il s'agit des limites liées à l'absence de l'usager dans le processus d'évaluation, celles liées aux jugements de pertinence et enfin les limites liées aux corpus de documents et celui de requêtes.

- **Limites par rapport à l'utilisateur :** Nous pouvons reprocher à ces approches d'évaluation d'être artificielles et arbitraires. Si le TREC a effectivement amélioré l'efficacité du système, la notion d'utilisateur final implique des connaissances personnelles, une expérience et des capacités de recherche différentes, dont l'évaluation du système ne se soucie pas. En effet, ces évaluations ne tiennent pas compte du contexte dans lequel se fait la recherche puisqu'elles ne sont pas effectuées en situation d'utilisation réelle. Dans ce contexte [Chaudiron, 02] affirme que l'absence de l'usager dans le processus d'évaluation est l'une des premières et sans doute la plus importante critique des approches classiques qui utilisent des critères autres que le rappel et la précision.
- **Limites par rapport aux jugements de pertinence :** La pertinence est une notion subjective et il paraît impensable de pouvoir la mesurer sans être arbitraire. Nous pouvons noter également que les jugements de pertinence dans le TREC opèrent de façon binaire : un document est jugé pertinent ou ne l'est pas. Pourtant, ce n'est évidemment pas toujours le cas, certains documents sont plus pertinents que d'autres qui le sont quand même. Ces degrés de pertinence dépendent encore de la disposition d'esprit de la personne ayant réellement besoin de ces documents. Cette constatation est validée par les travaux de [Tamine, 10] montrant que la pertinence considérée dans l'évaluation classique des SRI est thématique, indépendante du contexte, de la situation de recherche et des centres d'intérêt des utilisateurs. De même les travaux de [Chaudiron, 02] ont montré que les jugements de pertinence doivent être nuancés dans le sens où ils sont stables et ne varient pas dans le temps et qu'ils sont attribués indépendamment les uns désastres.
- **Limites par rapport au corpus de documents et de requêtes :** Dans les corpus traditionnels, un document est un texte à part entière, et l'évaluation se fait par rapport au nombre de documents retrouvés, or en général, un utilisateur ne cherche pas des documents mais de l'information, et les documents ne contiennent jamais la même quantité d'information. De même pour les corpus de requêtes où la requête est un besoin d'information exprimé en langage naturel. Or, c'est tout le problème de la représentation du besoin d'information de l'usager. La tâche de RI se transforme en tâche de savoir poser les questions à ces systèmes car les écarts sont grands entre ce que nous pensons et ce qui est interprété. [Tamine, 10] constate que le protocole d'évaluation étant en mode batch, les requêtes sont supposées ainsi représenter à elles seules l'utilisateur. Par conséquence les utilisateurs directs ayant émis ces requêtes, leurs centres d'intérêt et interactions avec le SRI ne font pas partie intégrante de la collection.

6.2 La problématique de l'évaluation personnalisée

Dans les SRI classique l'évaluation des requêtes se fait sans tenir compte du contexte de l'utilisateur qui l'émise. Ils supposent que l'utilisateur est complètement représenté par sa

requête et que les résultats retournés pour une même requête sont identiques même si cette dernière est exprimée par des utilisateurs différents. Considérons à titre illustratif la requête « virus » émise par deux utilisateurs, un informaticien qui cherche des virus informatiques et un biologiste qui cherche des virus biologiques. Vu la non considération du contexte de l'utilisateur en cours du processus de recherche et face aux phénomènes actuels d'accroissement incessant d'informations (volume), ainsi qu'à leur hétérogénéité et leur disparité, cela pose des problèmes de l'ambiguïté du sens des mots, l'impossibilité de sélectionner des sources opportunes et l'inintelligibilité des résultats retournés. Ceci a pour corollaire la non pertinence des résultats de recherche et de fait, l'insatisfaction de l'utilisateur.

Volume : le volume d'information ne se mesure plus actuellement en giga-octets mais en téraoctets voire en pétaoctets et exa-octets. Malheureusement, tous les algorithmes en RI ne sont pas de complexité linéaire en fonction du volume des informations ceci fait émerger le problème de passage à l'échelle qui engendre des dégradations des performances des processus de recherche [Chevallet et al, 04].

Hétérogénéité : le web est caractérisé par une forte hétérogénéité des sources d'information.

Cette hétérogénéité porte sur divers aspects : langue (plus de cent langues actuellement sur le web), média (texte, image, vidéo) etc.

Disparité : est une caractéristique qui traduit l'occurrence disséminée de l'information dans de large collections de documents, généralement interconnecté. Les outils de navigation hypertextes sont, à ce titre, destinés à matérialisé la proximité des informations autour d'un besoin particulier. Cependant copte tenu du volume important d'informations disponible, les utilisateurs sont vite submergés par le nombre considérable de lien proposés, ce qui engendre les phénomènes fort connus de désorientation de l'utilisateur et de surcharge informationnelle.

Ces problèmes font émerger la personnalisation comme approche essentielle aux systèmes d'accès à l'information.

La personnalisation est considérée dès lors comme un aspect dominant dans plusieurs secteurs. Elle peut cibler les deux aspects d'un système (interface ou contenu) de manière spécifique ou simultanément :

Au niveau de la présentation : personnalisé des aspects de l'interface utilisateur, y compris les couleurs, les polices, le positionnement et l'affichage des données. Cet aspect correspond à ce que l'on nomme la customisation des systèmes. Les aspects de mobilité et d'environnement géographique des utilisateurs sont également pris en considération.

Au niveau du contenu : cibler la recherche en fonction des besoins et des centres d'intérêts des différents utilisateurs. Cet aspect correspond plus à ce que l'on nomme « personnalisation ».

6.3 Protocole d'évaluation personnalisée

Le protocole d'évaluation dans le modèle d'évaluation oriente laboratoire consiste à comparer plusieurs systèmes, stratégies de recherche ou algorithmes entre eux en spécifiant trois composantes non indépendantes qui sont le nombre de topics utilisés, les mesures d'évaluation utilisés et la différence de performance requise pour considérer qu'une stratégie de recherche est meilleure qu'une autre [C. Buckley et E. M. Voorhees, 00].

L'évaluation de l'efficacité de chaque stratégie de recherche consiste à évaluer la liste de résultats obtenus pour chaque requête de test. Cette évaluation est à la base de la correspondance entre la pertinence algorithmique calculée par le système et la pertinence donnée par les assesseurs. L'efficacité globale d'une stratégie de recherche est calculée comme étant la moyenne des précisions calculées selon une mesure donnée sur l'ensemble des requêtes dans la collection de test.

6.4 Les mesures d'évaluation des systèmes personnalisés

Différentes mesures d'évaluation ayant été proposées dans le cadre des travaux sur la recherche des systèmes interactifs. Ces mesures peuvent être également employées pour l'évaluation d'un système d'accès personnalisé à l'information [174].

(i). La mesure RR (Relative Relevance).

La mesure RR [P. Borlund et P. Ingwersen, 98] a pour objectif de considérer différents types de pertinence (pertinence non binaire) dans l'évaluation de l'efficacité d'un système d'accès contextuel à l'information. Cette mesure quantifie le degré de concordance entre les types de jugement de pertinence émis dans le cas de deux ensembles de jugements (soit R1 et R2) associés à une même liste de documents qui constitue les résultats d'une session de recherche. En pratique, R1 correspond généralement aux scores de pertinence algorithmique retournés par un SRI et R2 à des scores de pertinence contextuelle correspondant à un type de pertinence donné : situationnelle si elle est exprimée par un utilisateur, thématique si elle est exprimée par un assesseur etc. La valeur de corrélation entre R1 et R2 est généralement calculée en utilisant une mesure du cosinus ; elle quantifie globalement, la capacité du système à prédire le type de pertinence contextuelle considéré.

A la différence de la mesure classique de précision, cette mesure permet de considérer les différents types de pertinence ; néanmoins, elle pose un problème lors de l'évaluation comparative entre différents algorithmes de recherche voire entre différents SRI [P. Borlund, 03].

En effet les scores de pertinence algorithmique ne sont pas étalonnés à la même échelle entre différents SRI, ce qui rend la comparaison de mesures RR non significative.

(ii). Les mesure CG (Cumulative Gain) et DCG (Discount Cumulative Gain)

Les mesures CG et DCG [K. Jarvelin et J. Kekalainen, 00] [K. Jarvelin et J. Kekalainen, 02] sont des mesures orientées position définies dans le contexte d'une pertinence graduelle et dont l'objectif est d'estimer le gain de l'utilisateur en termes de pertinence cumulée en observant les documents situés jusqu'à un rang donné. Ces mesures sont définies comme suit :

$$CG[i] = \begin{cases} G[1], si\ i = 1 \\ CG[i-1] + G[i], sinon \end{cases} \qquad (1)$$

Où G[i] est la valeur de pertinence associée au document de rang i.

$$CG[i] = \begin{cases} G[1], si\ i = 1 \\ CG[i-1] + G[i]/log_q, sinon \end{cases} \qquad (2)$$

Comparativement à la mesure CG, la mesure DCG permet d'atténuer le gain de pertinence apporté par un document en fonction du rang associé. Ceci rejoint en effet l'hypothèse évidente que plus le rang d'un document est élevé, moins il est probable que l'utilisateur l'examine et donc moins il est à l'origine d'un gain effectif de pertinence.

(iii). La mesure GRP (Generalised Recall and Precision)

La mesure GRP [K. Jarvelin et J. Kekalainen, 00] est également une mesure orientée position qui généralise les mesures classiques de rappel et précision en considérant une pertinence graduelle. Le rappel généralisé (GR) et la précision généralisée (GP) sont calculés comme suit :

$$GP = \sum_{d \in R} r(d)/|R| \qquad (3)$$

$$GR = \sum_{d \in R} r(d) / \sum_{d \in D} r(d) \qquad (4)$$

Où R est l'ensemble des documents retournés par le SRI, D est l'ensemble des documents de la collection, r(d) est la valeur de pertinence graduelle associée au document d.

De manière analogue aux mesures classiques de rappel/précision, ces mesures offrent la possibilité d'être agrégées pour plusieurs requêtes ou plusieurs niveaux de rappel et donnent ainsi la possibilité de tracer des courbes de performances.

7 Conclusion

En conclusion, la recherche d'informations personnelles permet aux utilisateurs de trouver plus facilement et plus rapidement les informations qu'ils recherchent. Il utilise des algorithmes d'apprentissage automatique pour personnaliser les résultats de recherche en fonction des préférences, des intérêts et des habitudes de recherche des utilisateurs.

La recherche d'informations personnalisée a ses avantages et ses inconvénients. Elle améliore considérablement l'expérience de recherche des utilisateurs, mais elle pose des problèmes de confidentialité et de sécurité des données. Les utilisateurs doivent être conscients des informations qu'ils fournissent et de la manière dont elles sont utilisées pour personnaliser les résultats de recherche.

Chapitre III :
Un profil d'utilisateur guidé par les évènements

1 Introduction

L'objectif de ce travail est de prendre en compte, lors de la recherche d'information, la dimension évènementielle des besoins en information cachés derrière la requête de l'utilisateur. Nous défendons l'hypothèse suivante : l'utilisateur peut s'intéresser aux évènements actuels, récemment passés ou à venir ; qui attirent l'attention d'une bonne partie de sa communauté, comme il peut s'intéresser aux événements passés ayant une périodicité de reproduction ou de célébration.

Dans la littérature de la recherche d'information personnalisée, l'information temporelle souvent prise en considération est celle concernant la création, l'actualisation et les mouvements des documents. L'objective étant de pouvoir classer les documents en fonction de ces informations temporelles. Néanmoins, très peu d'approches qui s'intéressent à l'aspect évènementiel et temporel de contenu des documents [Basile et al. 2014].

Ce que nous présentons dans ce travail se résume à :

(i). La proposition d'un modèle de l'évènement qui permet de définir et de classifier les évènements, et facilite leur identification dans les textes.
(ii). La proposition d'un profil d'utilisateur qui permet de représenter ses intérêts thématiques et temporels.
(iii). Le processus de recherche dans lequel sont intégrées les informations contextuelles fournies par le profil le de l'utilisateur.
(iv). La proposition de certaines formules de calcul des poids et des similarités concernant les évènements et leurs termes associés.

2 Modélisation de l'évènement

2.1 Définition de l'évènement

La notion d'événement ne conduit pas à une définition stricte, précise et consensuelle [Arikan et al. 2009]. Une variété de définitions est donnée à l'évènement dans la littérature. Ces définitions se diffèrent selon le domaine d'étude (histoire, philosophie, linguistique, …), néanmoins elles renferment une caractéristique commune de l'évènement « ancré dans le temps » [Arnulphy 2012].

Dans notre cas, nous adoptons la définition proposée dans [Arikan et al. 2009] : « les évènements sont des entités spatiotemporelles Finies », et celle proposée dans [Ludovic et al. 2013] : « un événement E est la combinaison de 3 composantes E (S, SP, I) : une propriété sémantique S, un intervalle temporel I, et une entité spatiale SP ».

2.2 Classification des évènements

Les modèles d'évènement proposés par les différentes approches d'extraction d'évènements permettent d'extraire à partir des textes tout ce qui peut être jugé comme évènement quoi que ce soit son importance pour l'utilisateur. Ce qui va plus loin que nos besoins dans ce travail qui se focalise en particulier sur les évènements importants et qui attirent massivement l'attention des utilisateurs, et qui ont fait l'objet de recherche pour une bonne partie de la communauté de l'utilisateur.

Dans notre cas, et pour fuir au problème de diversité de classes et d'attributs proposés dans les différents travaux pour décrire ou identifier un évènement, nous allons classifier les évènements seulement en deux classes :

(i). Evènement passé :

Nous qualifions d' «évènement passé important» :

- Un évènement ayant une périodicité de reproduction (la coupe du monde de football par exemple).
- Un évènement ayant une périodicité de célébration (la guerre mondiale par exemple).

(ii). Evènement actuel :

Nous qualifions d' «évènement d'actualité »:

- Un évènement qui est passé très récemment et qui occupe encore l'opinion d'une bonne partie de la communauté de l'utilisateur.
- Un évènement en cours (arrivé et n'est pas encore achevé).
- Un évènement qui va venir très prochainement (prévu) mais il attire l'attention dès maintenant.

3 Profil d'utilisateur guidé par les évènements

Pour nous, la pertinence utilisateur peut être déduite, en plus de sa requête, à partir de :

(i). L'aspect thématique, représenté par l'historique de recherches qui comporte les documents pertinents ayant intéressés l'utilisateur (longuement visités, trop cliqués, imprimés…), et les termes pertinents extraits à partir de ces documents.
(ii). L'aspect évènementiel, représenté par les évènements extraits à partir de son historique de recherche.

Et ainsi, le profil d'utilisateur que nous proposons se compose des composantes suivantes :

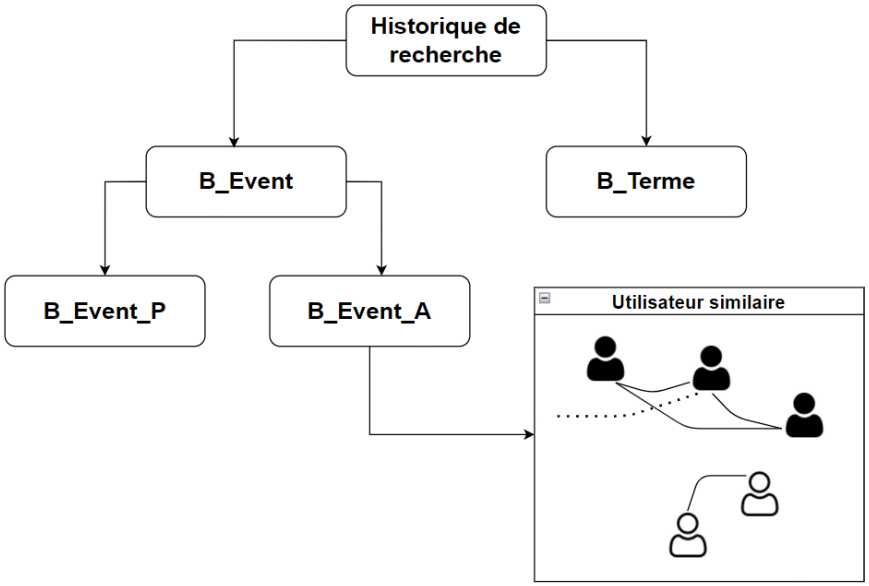


Figure 3.1 : Profil d'utilisateur.

- **B_Terme :** Base des termes.
- **B_Event_P :** Base des évènements passés.
- **B_Event_A :** Base des évènements actuels.

3.1 L'historique de recherche thématique

Cette composante (figure 3.2) sert pour sauvegarder les traces des activités de recherche les plus significatives pour l'utilisateur. Elle se compose des 2 sous-ensembles suivants :

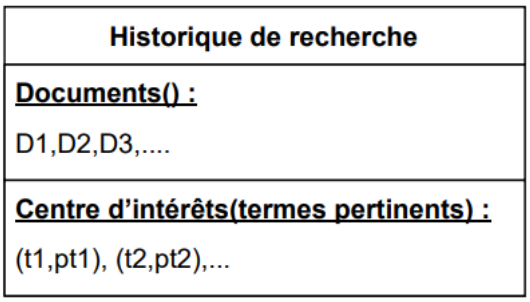


Figure 3.2 : Historique de recherche.

3.1.1 Les documents pertinents

Cette composante regroupe l'ensemble de documents jugés pertinents par l'utilisateur (longuement visités, imprimés, partagés, annotés, marqués…etc.) dans les différentes opérations de recherche.

3.1.2 Les termes pertinents

Cette composante regroupe l'ensemble de termes pertinents extrait à partir des documents pertinents. Ces termes pertinents de l'historique forment le centre d'intérêts de l'utilisateur, qui est représenté par un vecteur noté :

$$Vt = \{(t_1, pt_1), (t_2, pt_2), \dots, (t_i, pt_i), \dots, (t_k, pt_k)\} \qquad (01)$$

Où :

- t_i : représente le $i^{ème}$ terme du centre d'intérêts.
- pt_i : représente le poids de t_i par rapport à l'historique de recherche.

Le poids d'un terme t dans l'historique de recherche reflète son importance pour l'utilisateur dans cet historique. Il est calculé selon la formule de $tf * idf$ du modèle vectoriel [Salton et Buckley 1988] :

$$P(t) = (0.5 + 0.5 * tf / Max\ tf) * \log(N / n) \qquad (02)$$

Avec :

- tf : fréquence totale de terme t dans les documents de l'historique.
- $Max\ tf$: la plus grande fréquence des termes dans les documents de l'historique.
- N : nombre de documents de l'historique.
- n : nombre de documents contenant le terme t de l'historique.

3.1.3 La base de liens de cooccurrence terme-terme

La collection de termes pertinents dans un centre d'intérêt représente une utilisation de premier niveau de l'historique de recherche pour décrire l'utilisateur et ses intérêts. Le deuxième niveau d'utilisation des termes pertinents consiste à identifier et évaluer leurs relations de cooccurrence. Dans certains travaux de recherche d'information [Mousset et al. 2019] [Bernardo et al. 2013], la relation de cooccurrence entre les termes est considérée comme une relation sémantique.

Nous souhaitons mettre en évidence les occurrences fréquentes de paires de termes dans de nombreux documents de l'historique de recherche de l'utilisateur. Pour nous, l'occurrence fréquente des deux termes dans de nombreux documents peut signifier qu'ils coopèrent pour décrire le même sujet. Par exemple, l'apparition de deux termes « Java » et « Programmation » peut indiquer qu'il s'agit d'un sujet sur l'informatique, tandis que l'apparition de « Java » et « ile » indique un sujet de tourisme. Et si « Java » 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 « Java ».

La valorisation des relations de cooccurrence entre les termes pertinents est représentée par une similarité thématique, notée sim_terme (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 :

$$Sim_terme(t_i, t_j) = |d(t_i, t_j)|/(|d(t_i)| + |d(t_j)|) \qquad (03)$$

Avec :

- $|d\ (t_i, t_j)\ |$: nombre de documents contenant t_i et t_j conjointement.
- $|d\ (t_i)|$: nombre de documents contenant t_i.
- $|d\ (t_j)|$: nombre de documents contenant t_j.

La formule (03) favorise, pour un terme donné t_i, les termes répondant aux 2 critères : (i) apparaissant dans le maximum de documents contenant t_i, et (ii) apparaissant dans le minimum de documents ne contenant pas t_i.

La base de relations de cooccurrence des termes pertinents (B_Terme) comme l'affiche (figure 3.3), sert à sauvegarder les liens de similarité entre les termes pertinents de l'historique de profil utilisateur.

B_Terme
Lien(terme-terme)

	t_1	t_2	....	t_n
t_1	-	$Sim(t_1,t_2)$	....	$Sim(t_1,t_n)$
t_2	$Sim(t_2,t_1)$	-	....	$Sim(t_2,t_n)$
....	....	....	-	....
t_n	$Sim(t_n,t_1)$	$Sim(t_n,t_2)$	....	-

Figure 3.3 : Base de cooccurrence terme-event-a.

3.2 L'historique de recherche évènementiel

3.2.1 Modèle d'évènement proposé

le modéle que nous proposons pour l'un évènement (figure 3.4) se compose de deux (02) composantes :

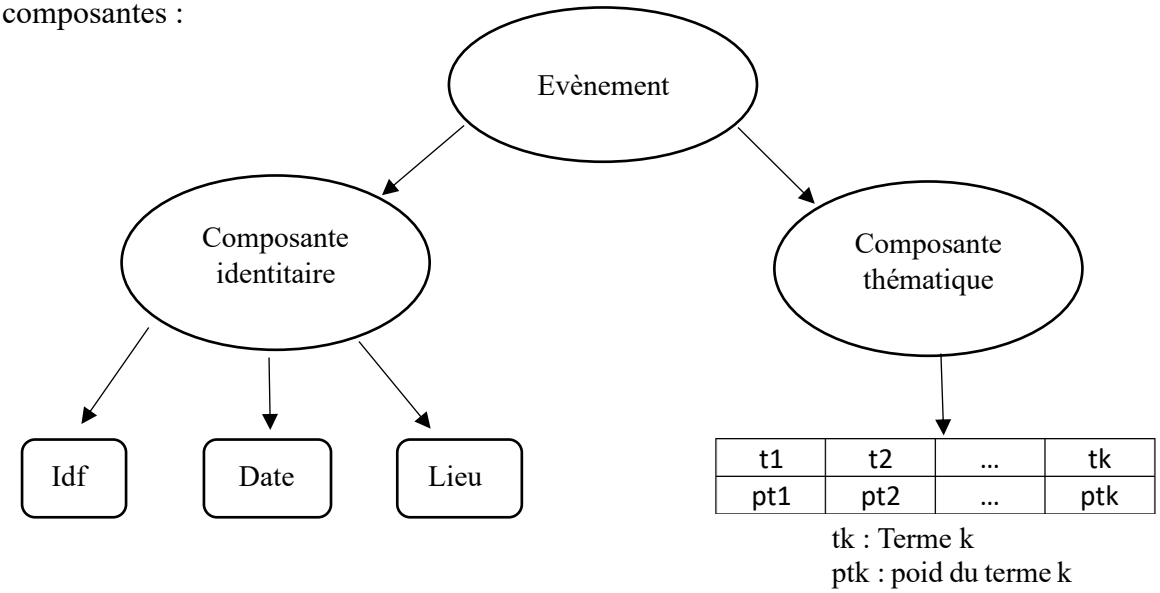


Figure 3.4 : Modèle de l'évènement.

Composante d'identité :

Un évènement est décrit par trois propriétés :

- L'identifiant de l'évènement qui peut être représenté par un nom propre, entité nommée, ou même par une phrase.
- La date de l'évènement qui peut être sous une forme complète ou partielle d'une date.
- Le lieu de l'évènement qui peut être représenté par le nom d'une ville, une région, un pays …etc.

Une composante thématique :

Regroupe l'ensemble de termes pertinents extraits de l'historique de recherche de l'utilisateur, ayant une relation de cooccurrence avec l'évènement.

3.2.2 Pondération des évènements passés

Les évènements passés sont sauvegardés dans une base notée B_Event_P. Et pour chaque évènement est associé un ensemble de termes les plus liés.

Pour nous, l'évènement présent dans de nombreux documents de l'historique de recherche, est considéré plus pertinent. Ainsi, le poids d'un évènement E est une fonction de la proportion du nombre de documents contenant l'événement par rapport au nombre total de documents de l'historique de recherche dans le profil.

Ce poids, noté P(Ep), est calculé comme suit :

$$P(Ep) = \frac{|D(Ep)|}{|D|} \qquad (04)$$

Avec :

- $|D(Ep)|$: nombre de documents dans l'historique de recherche contenant l'évènement Ep.
- $|D|$: nombre total de documents dans l'historique de recherche.

- **Pondération des liaisons des termes pertinents à un évènement passé**

Une liaison d'un terme à un évènement est calculée par rapport à tout l'historique de recherche du profil d'utilisateur. Elle favorise les termes apparaissant conjointement avec l'évènement dans de nombreux documents de l'historique de recherche. Cette liaison est notée sim(t, Ep), et elle calculée comme suit:

$$sim(t, Ep) = \frac{P(t) * P(Ep) * |D(t, Ep)|}{|D(t)| + |D(Ep)|} \quad (05)$$

Où :

- $|D(t, Ep)|$: nombre de documents de l'historique de recherche contenant le terme t et l'évènement Ep.
- $|D\ (t)\ |$: nombre de documents de l'historique de recherche contenant le terme t.
- $|D(Ep)|$: nombre de documents de l'historique de recherche contenant l'évènement Ep.
- $P(t)$: Le poids d'un terme t dans l'historique de recherche selon la formule (02).
- $P(Ep)$: Le poids d'un évènement passé dans l'historique de recherche selon la formule (04).

B_Event_P
Lien(terme-Ep)

	Ep_1	Ep_2	....	Ep_n
t_1	-	$Sim(t_1,Ep_2)$	....	$Sim(t_1,Ep_n)$
t_2	$Sim(t_2,Ep_1)$	-	....	$Sim(t_2,Ep_n)$
....	....	....	-	....
t_n	$Sim(t_n,Ep_1)$	$Sim(t_n,Ep_2)$	....	-

Figure 3.5 : base de cooccurrence terme-event-p.

3.2.3 Pondération des évènements actuels

Les évènements d'actualité peuvent ne pas figurer dans l'historique de recherche de l'utilisateur. Néanmoins, ils peuvent l'intéresser de fait qu'ils intéressent ses amis. Pour aider l'utilisateur en matière d'évènements actuels, nous allons doter son profil par une base pour stocker les évènements actuels les plus fréquents dans les historiques de recherche de ses amis. A chaque évènement est associé l'ensemble des termes les plus liés.

- **Pondération d'un utilisateur similaire**

Dans le cas de nouveaux ou de pauvre profil ainsi que dans le cas de changement d'intérêt, le centre d'intérêt de l'utilisateur ne peut pas guider les nouvelles de recherches. Pour cela les

approches de la RIP demandent de l'aide auprès des utilisateurs similaires les plus proches de l'utilisateur et qui partagent les mêmes intérêts avec lui. Le but est d'aider à désambigüiser la requête utilisateur et d'identifier le besoin en information qui peut être dérivé à partir des profils de ses utilisateurs similaires.

Pour nous, l'importance d'un évènement pertinent dans l'historique de recherche d'un utilisateur similaire de l'utilisateur, dépond de l'importance de l'utilisateur similaire lui-même pour l'utilisateur. Pour cela, nous dotons notre profil d'utilisateur par une liste dynamique de ses utilisateurs similaires les plus importants pour lui.

Pour un utilisateur U, l'importance d'un utilisateur similaire A est pondérée par une similarité, $Sim_us(U,A)$, basée sur l'intersection des centres d'intérêts thématiques et évènementiels des deux utilisateurs.

La formule se présente comme suit :

$$Sim_us(U,A) = \frac{\sum_{t \in \boldsymbol{CI_com(U,A)}} Pu(t) + \sum_{Ep \in CI_com(U,A)} Pu(Ep)}{\left|\{v_t(U) \cup v_{Ep}(U)\}\right|} \qquad (06)$$

Tel que :

- $\mathrm{CI_com}(U,A) = \{v_t(U) \cup v_{Ep}(U)\} \cap \{v_t(A) \cup v_{Ep}(A)\}$

Où :

- $Pu(t)$: le poids du terme t dans l'historique de l'utilisateur U selon la formule 2.
- $Pu(Ep)$: le poids de l'évènement passé dans l'historique de l'utilisateur U selon la formule 4.
- $CI_com(U,A)$: le centre d'intérêts commun entre l'utilisateur u et son utilisateur similaire A.

La formule 6 permet de classer en premier les utilisateurs similaires avec lesquels ; l'utilisateur partage un nombre important de termes et d'évènements passés, qui sont bien pondérés dans son historique de recherche.

Poids d'un évènement actuel

Soit A la liste des utilisateurs similaires accompagnés par leurs degrés d'importance pour lui :

$$A = \{(A1, sim_us(U,A1)), (A2, sim_us(U,A2)), \dots, (An, sim_us(U,An))\}$$

Le poids d'un évènement actuel pour l'utilisateur U, noté $P(Ea)$, est donné par :

$$P(Ea) = \frac{1}{|A|} * \sum_{i=1..|A|} \left(P_i(Ep) * Sim_us(U,A_i)\right) \qquad (07)$$

Avec :

- $|A|$: nombre des utilisateurs similaires.
- $P_i(Ep)$: poids de l'évènement passé Ep dans l'historique de recherche récent de l'utilisateur similaire A_i (calculé selon la formule 04).
- $Sim_us(U, A)$: poids d'utilisateur similaire (formule 06).

- **Pondération des liaisons des termes pertinents à un évènement actuel**

Le degré de liaison d'un terme t à un évènement actuel Ea, noté $sim(t, Ea)$, est donné par :

$$sim(t, Ea) = \frac{1}{|A|} * \sum_{i=1..|A|} \big(sim(t, Ep) * Sim_us(U, Ai)\big) \qquad (08)$$

Avec :

- $sim(t, Ep)$: liaison de terme t à l'évènement passé Ep dans l'historique restreint de l'utilisateur similaire A (Ep pour l'utilisateur similaire A= Ea pour l'utilisateur U)

Pour chaque évènement actuel Ea, introduit dans la base des évènements actuels de l'utilisateur B_event_A, juste un nombre très restreint de termes très liés à cet évènement dans l'historique récent de l'utilisateur similaire d'où il est trouvé, qui seront introduits dans le profil de l'utilisateur concernés.

B_Event_A
Lien(terme-Ea)

	Ea_1	Ea_2	....	Ea_n
t_1	-	$Sim(t_1,Ea_2)$	....	$Sim(t_1,Ea_n)$
t_2	$Sim(t_2,Ea_1)$	-	....	$Sim(t_2,Ea_n)$
....	....	....	-	....
t_n	$Sim(t_n,Ea_1)$	$Sim(t_n,Ea_2)$	....	-

Figure 3.6 : Base de cooccurrence terme-event-a.

3.3 Processus de recherche d'information

Notre processus de recherche est le suivant :

- Le système reçoit la requête utilisateur Q.
- Le système sélectionne la base de cooccurrence la plus correspondante à la requête parmi les trois bases (B_Terme, B_Event_P ou B_Event_A)) la plus appropriée pour la requête Q.
- Selon la base choisie, la requête Q est reformulée en Q'.
- La requête reformulée Q' est envoyée au moteur de recherche classique.
- Le système transmet les résultats de recherche à l'utilisateur et observe ses réactions et ses comportements pour mettre à jour son profil.

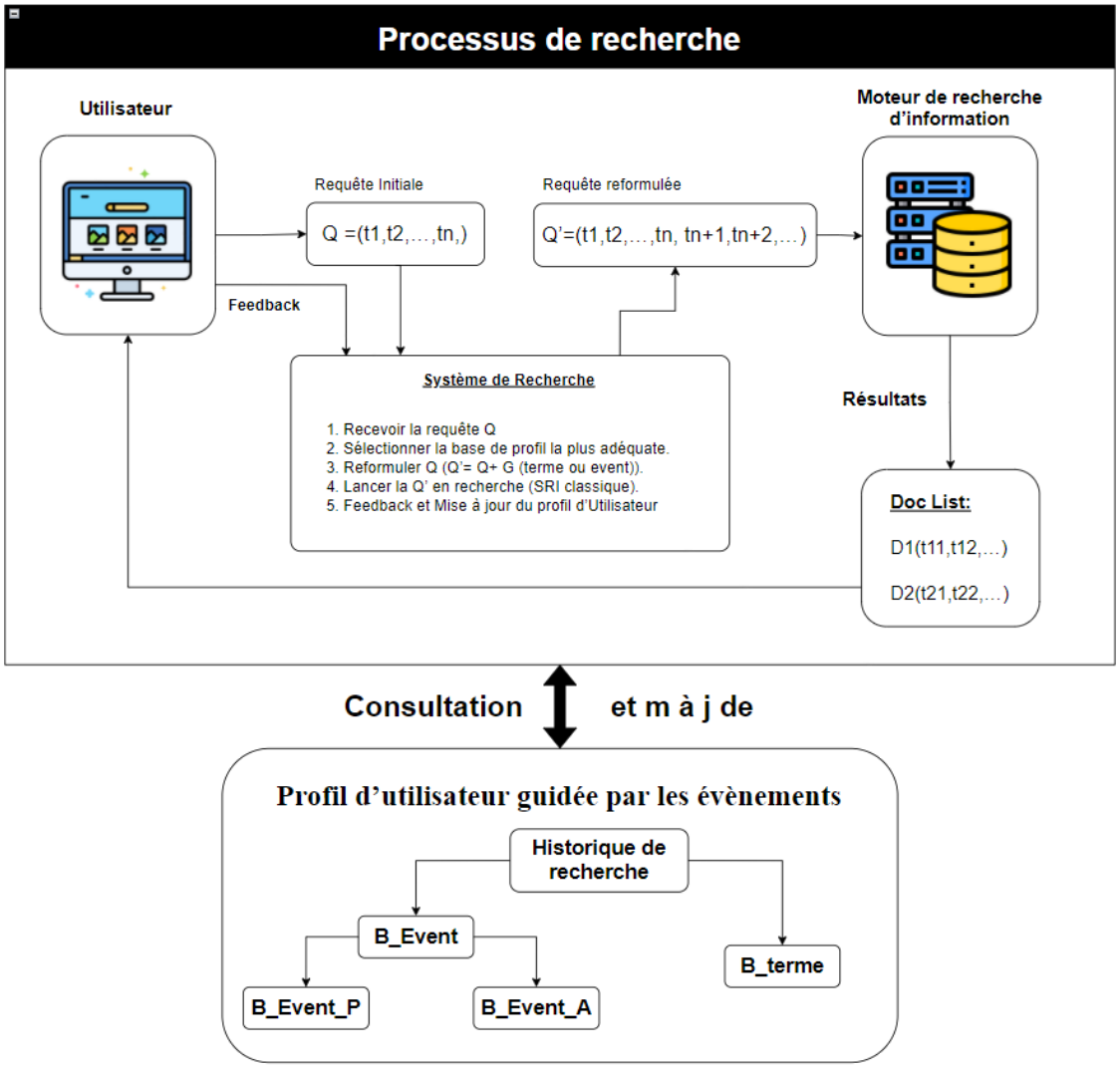


Figure 3.7 : Processus de recherche.

3.3.1 Sélection de la base de profil la plus adéquate pour la requête utilisateur

Le but de cette étape est de sélectionner l'une des bases B_Terme, B_Event_P ou B_Event_A en fonction des similitudes, notre système calcule un score de similarité entre la requête et chacune des trois bases.

Soient :

- $Sim\ (B_Terme/Q) = \sum_{t_i \in Q} \sum_{t_j \in B_Terme} sim_terme(t_i, t_j)$ (09)
- $sim\ (B_Event_P/Q) = \sum_{t \in Q} \sum_{Ep \in B_Event_P} sim_event_p(t, Ep)$ (10)
- $sim\ (B_Event_A/Q) = \sum_{t \in Q} \sum_{Ea \in B_Event_A} sim_event_a(t, Ea)$ (11)

La base qui a la similarité max avec la requête, sera choisie comme contexte courant de la recherche.

3.3.2 Reformulation de requête

Dans l'objectif de bien cibler le champ de la recherche et ne pas aller plus loin que les intérêts de l'utilisateur, la requête initiale de l'utilisateur Q sera enrichie par l'injection de quelques

termes ou d'évènements pertinents issus de la base de profil sélectionnée pour produire une nouvelle requête Q', tels que :

$$Q' = Q + G \qquad (12)$$

Avec :

- En cas où la base *B_Terme* est sélectionnée : G est le terme qui maximise la similarité (3) avec les termes de Q
- En cas où la base *B_event_P* est sélectionnée : G est l'évènement passé Ep qui maximise la similarité (5) avec les termes de Q
- En cas où la base *B_event_A* est sélectionnée : G est l'évènement actuel Ea qui maximise la similarité (8) avec les termes de Q.

3.3.3 Mise à jour du profil

La mise à jour du profil d'utilisateur guidé par les évènements doit assurer d'un côté, son enrichissement et son évolution dans le temps par l'ajout des nouveaux documents, termes et évènements pertinents, et la mise à jour des bases thématique et évènementiel. Et d'un autre côté, elle doit veiller sur l'optimisation de l'espace occupé et le temps de réponse pris par les différents calculs, et ce par la suppression des documents, des termes et des évènements qui ne sont pas pertinents depuis un certain temps.

4 Simulation

Comme il n'existe pas de cadre d'évaluation personnalisé pour le contexte de personnalisation, nous avons construit notre cadre d'évaluation comme suit :

4.1 Cadre d'évaluation

Dans cette étape, nous avons construit 15 profils de 15 utilisateurs appartenant à différents domaines (informatique, médecine, politique…) pour avoir des centres d'intérêts différents. Nous avons construit chaque profil en collaboration avec l'utilisateur respectif. Les termes pertinents, les événements pertinents et les différentes pondérations, liens de cooccurrence et similarités sont définies par une simulation, et sont enregistrés dans les deux bases B_Terme ou B_Event_P.

Les données des profils sont organisées comme indique la figure (figure 3.8) suivante :

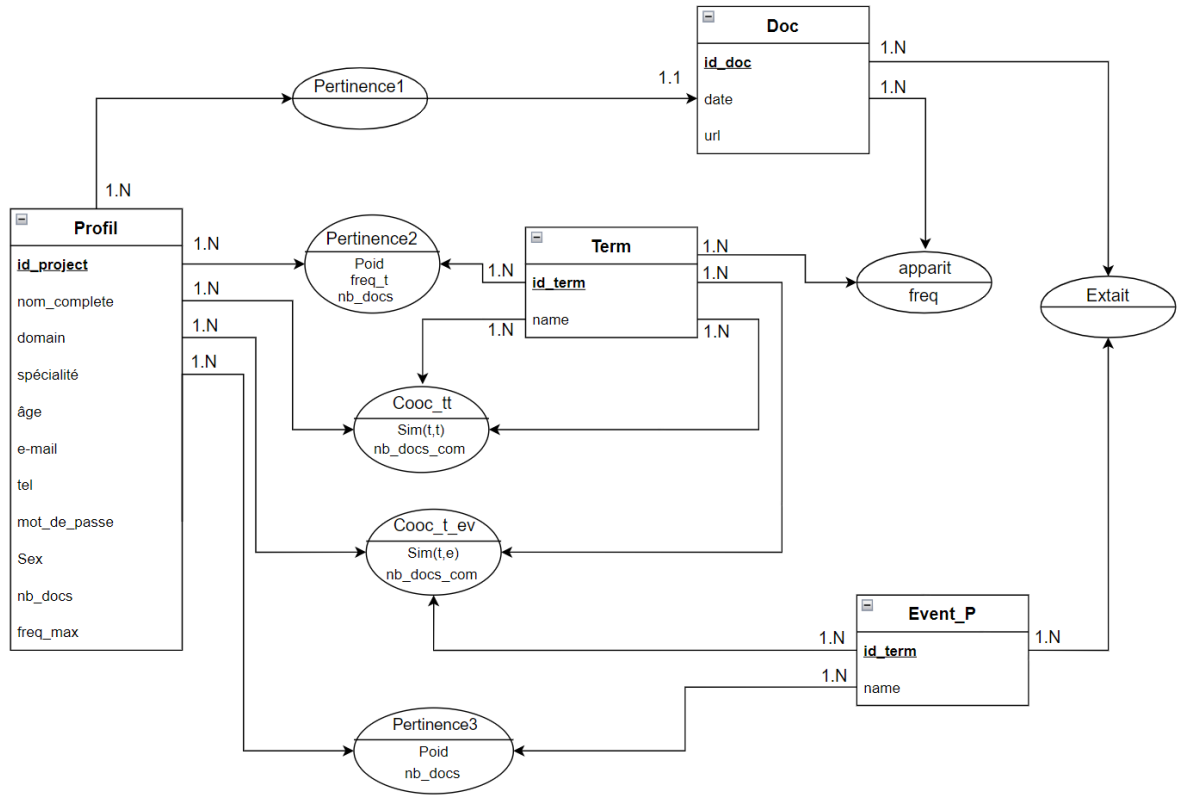


Figure 3.8 : Représentation de modèle conceptuel des données (MCD).

Dans le tableau suivant, nous présentons quelques exemples des contenus des profils :

	Utilisateurs	**Centres d'intérêts**	**Requêtes envoyées**
Informatique	U1 : étudient	**Termes :** Programmation, système d'exploitation, informatique, réseaux, programmation orientée objet, sécurité, java swing, MySQL, UML, système binaire, algorithme, site web, Développement, virus, raccourci USB, …. **Évènements :** Hackathon dz, soutenance 2023, …	• Développement • Virus • Sécurité
Médecine	U2 : médecin	**Termes :** Hôpital, Diagnostic Traitement, Prescription, Chirurgie, Anesthésie, Cabinet, médicale, Spécialiste, Stéthoscope, Radiographie, Prélèvement sanguin, Virus, Vaccination, analyse, Ordonnance, … **Évènements :** corona 19, malaria, …	• Virus • Ordonnance
	U1 : Chercheur en politique	**Termes :** politiques, comportements politiques, institutions gouvernementales, publiques, Analyse, Théorie, libéralisme, marxisme, féminisme, électorales, vote, partis politiques, sondage d'opinion,	• Politiques • La guerre

Politique		organisation internationale, conflits, diplomatie, la coopération mondiale, publique, médias, manipulation de l'information, la guerre, …. **Évènements :** Russie vs Ukraine	
	…….	**…….**	•
	…….	**…….**	•

Table 3.1: exemples des contenus des profils.

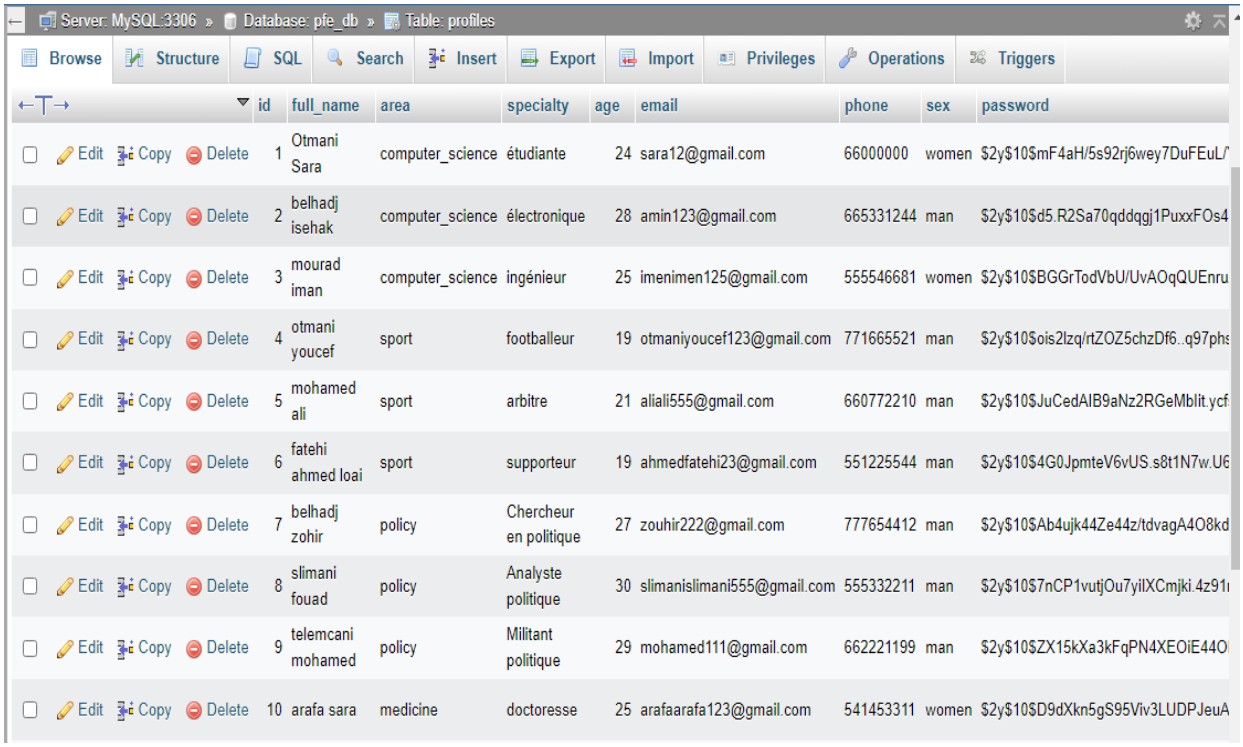
4.2 Présentation de base de données

Pour créer des bases de données, nous avons utilisé MySQL

MySQL est un SGBD (Système de Gestion de Base de Données) relationnel open-source, ce qui signifie qu'il est conçu pour stocker, organiser et gérer des données de manière structurée en utilisant des relations entre différentes tables.

Voici quelques captures d'écran des tables et des relations pour cette base de données :

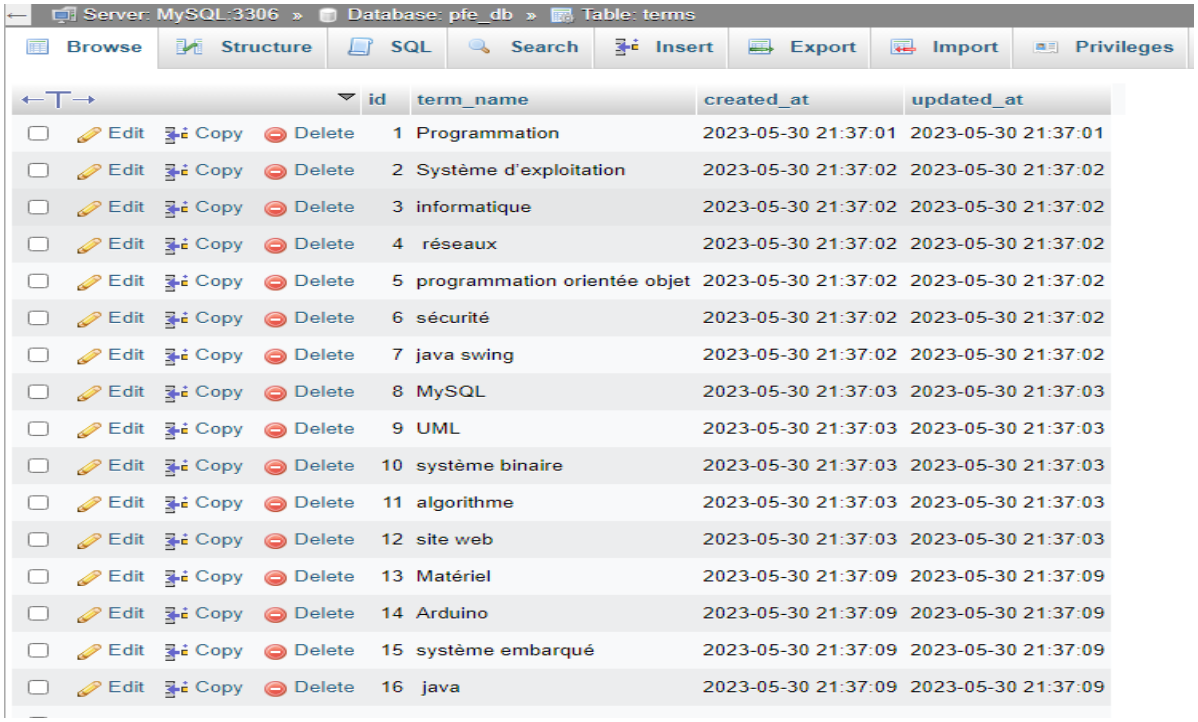
- **Table Profil**



	id	full_name	area	specialty	age	email	phone	sex	password
Edit Copy Delete	1	Otmani Sara	computer_science	étudiante	24	sara12@gmail.com	66000000	women	$2y$10$mF4aH/5s92rj6wey7DuFEuL/
Edit Copy Delete	2	belhadj isehak	computer_science	électronique	28	amin123@gmail.com	665331244	man	$2y$10$d5.R2Sa70qddqgj1PuxxFOs4
Edit Copy Delete	3	mourad iman	computer_science	ingénieur	25	imenimen125@gmail.com	555546681	women	$2y$10$BGGrTodVbU/UvAOqQUEnru
Edit Copy Delete	4	otmani youcef	sport	footballeur	19	otmaniyoucef123@gmail.com	771665521	man	$2y$10$ois2lzq/rtZOZ5chzDf6..q97phs
Edit Copy Delete	5	mohamed ali	sport	arbitre	21	alialí555@gmail.com	660772210	man	$2y$10$JuCedAIB9aNz2RGeMblit.ycf
Edit Copy Delete	6	fatehi ahmed loai	sport	supporteur	19	ahmedfatehi23@gmail.com	551225544	man	$2y$10$4G0JpmteV6vUS.s8t1N7w.U6
Edit Copy Delete	7	belhadj zohir	policy	Chercheur en politique	27	zouhir222@gmail.com	777654412	man	$2y$10$Ab4ujk44Ze44z/tdvagA4O8kd
Edit Copy Delete	8	slimani fouad	policy	Analyste politique	30	slimanislimani555@gmail.com	555332211	man	$2y$10$7nCP1vutjOu7yilXCmjki.4z91
Edit Copy Delete	9	telemcani mohamed	policy	Militant politique	29	mohamed111@gmail.com	662221199	man	$2y$10$ZX15kXa3kFqPN4XEOiE44O
Edit Copy Delete	10	arafa sara	medicine	doctoresse	25	arafaarafa123@gmail.com	541453311	women	$2y$10$D9dXkn5gS95Viv3LUDPJeuA

Figure 3.9 : Table des profiles.

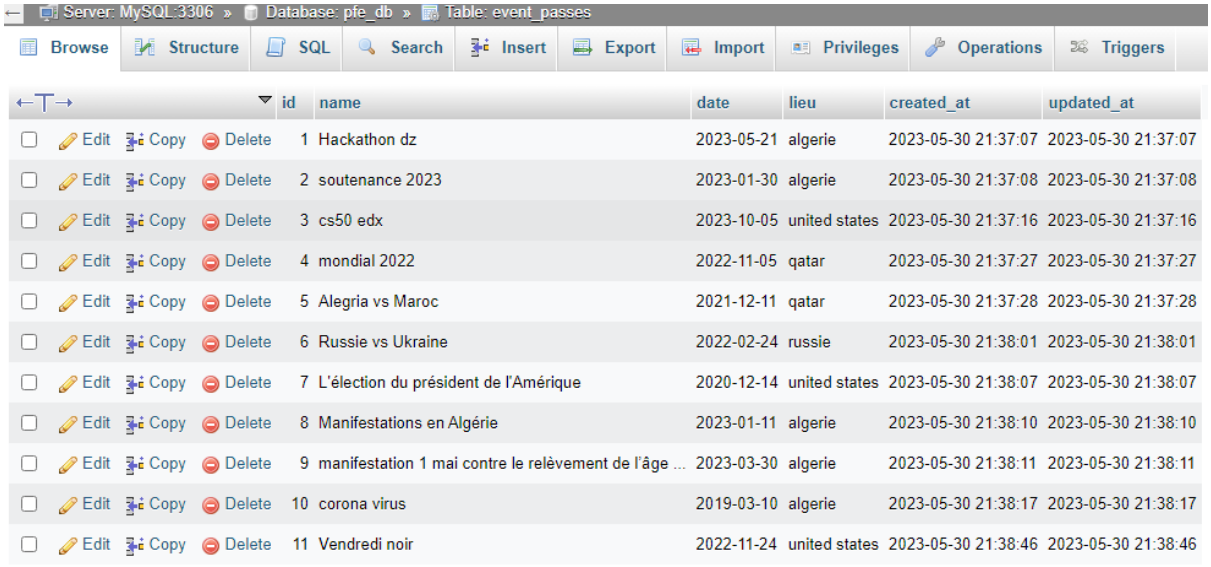
❖ Table Terme



	id	term_name	created_at	updated_at
Edit Copy Delete	1	Programmation	2023-05-30 21:37:01	2023-05-30 21:37:01
Edit Copy Delete	2	Système d'exploitation	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	3	informatique	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	4	réseaux	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	5	programmation orientée objet	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	6	sécurité	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	7	java swing	2023-05-30 21:37:02	2023-05-30 21:37:02
Edit Copy Delete	8	MySQL	2023-05-30 21:37:03	2023-05-30 21:37:03
Edit Copy Delete	9	UML	2023-05-30 21:37:03	2023-05-30 21:37:03
Edit Copy Delete	10	système binaire	2023-05-30 21:37:03	2023-05-30 21:37:03
Edit Copy Delete	11	algorithme	2023-05-30 21:37:03	2023-05-30 21:37:03
Edit Copy Delete	12	site web	2023-05-30 21:37:03	2023-05-30 21:37:03
Edit Copy Delete	13	Matériel	2023-05-30 21:37:09	2023-05-30 21:37:09
Edit Copy Delete	14	Arduino	2023-05-30 21:37:09	2023-05-30 21:37:09
Edit Copy Delete	15	système embarqué	2023-05-30 21:37:09	2023-05-30 21:37:09
Edit Copy Delete	16	java	2023-05-30 21:37:09	2023-05-30 21:37:09

Figure 3.10 : Table des termes (terms).

❖ Table des évènements passés



	id	name	date	lieu	created_at	updated_at
Edit Copy Delete	1	Hackathon dz	2023-05-21	algerie	2023-05-30 21:37:07	2023-05-30 21:37:07
Edit Copy Delete	2	soutenance 2023	2023-01-30	algerie	2023-05-30 21:37:08	2023-05-30 21:37:08
Edit Copy Delete	3	cs50 edx	2023-10-05	united states	2023-05-30 21:37:16	2023-05-30 21:37:16
Edit Copy Delete	4	mondial 2022	2022-11-05	qatar	2023-05-30 21:37:27	2023-05-30 21:37:27
Edit Copy Delete	5	Alegria vs Maroc	2021-12-11	qatar	2023-05-30 21:37:28	2023-05-30 21:37:28
Edit Copy Delete	6	Russie vs Ukraine	2022-02-24	russie	2023-05-30 21:38:01	2023-05-30 21:38:01
Edit Copy Delete	7	L'élection du président de l'Amérique	2020-12-14	united states	2023-05-30 21:38:07	2023-05-30 21:38:07
Edit Copy Delete	8	Manifestations en Algérie	2023-01-11	algerie	2023-05-30 21:38:10	2023-05-30 21:38:10
Edit Copy Delete	9	manifestation 1 mai contre le relèvement de l'âge ...	2023-03-30	algerie	2023-05-30 21:38:11	2023-05-30 21:38:11
Edit Copy Delete	10	corona virus	2019-03-10	algerie	2023-05-30 21:38:17	2023-05-30 21:38:17
Edit Copy Delete	11	Vendredi noir	2022-11-24	united states	2023-05-30 21:38:46	2023-05-30 21:38:46

Figure 3.11 : Table des événements passés (event_passes).

❖ **Table de liens de cooccurrence terme-terme**

Server: MySQL:3306 » Database: pfe_db » Table: profile_term_terms

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges

	id	profile_id	term_id	term2_id	nb_docs_com	sim_tt	created_
Edit Copy Delete	1	1	1	2	1	0.14285714285714	NULL
Edit Copy Delete	2	1	1	3	1	0.2	NULL
Edit Copy Delete	3	1	1	4	0	0	NULL
Edit Copy Delete	4	1	1	5	2	0.28571428571429	NULL
Edit Copy Delete	5	1	1	6	0	0	NULL
Edit Copy Delete	6	1	1	7	2	0.25	NULL
Edit Copy Delete	7	1	1	8	1	0.2	NULL
Edit Copy Delete	8	1	1	9	1	0.16666666666667	NULL
Edit Copy Delete	9	1	1	10	0	0	NULL
Edit Copy Delete	10	1	1	11	1	0.16666666666667	NULL
Edit Copy Delete	11	1	1	12	2	0.25	NULL
Edit Copy Delete	12	1	2	3	0	0	NULL
Edit Copy Delete	13	1	2	4	1	0.2	NULL
Edit Copy Delete	14	1	2	5	1	0.16666666666667	NULL

Figure 3.12: La base de liens de cooccurrence terme-terme (profile_term_terms).

❖ **Table de Pondération des liaisons des termes pertinents à un évènement passé**

Server: MySQL:3306 » Database: pfe_db » Table: profile_term_event_passes

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges

	id	profile_id	term_id	event_pass_id	nb_docs_com	sim_te
Edit Copy Delete	1	1	1	1	1	0.013379110918399
Edit Copy Delete	2	1	2	1	1	0.01490890562953
Edit Copy Delete	3	1	3	1	0	0
Edit Copy Delete	4	1	4	1	1	0.022577249674799
Edit Copy Delete	5	1	5	1	1	0.014198957742409
Edit Copy Delete	6	1	6	1	1	0.042645916052398
Edit Copy Delete	7	1	7	1	1	0.010452430404999
Edit Copy Delete	8	1	8	1	0	0
Edit Copy Delete	9	1	9	1	2	0.047663082646797
Edit Copy Delete	10	1	10	1	1	0.021322958026199
Edit Copy Delete	11	1	11	1	2	0.055188832538396
Edit Copy Delete	12	1	12	1	1	0.011706722053599
Edit Copy Delete	13	1	1	2	1	0.0080274665510395
Edit Copy Delete	14	1	2	2	0	0
Edit Copy Delete	15	1	3	2	1	0.031984437039298

Figure 3.13 : Pondération des liaisons des termes pertinents à un évènement passé (profile_term_event_passes).

4.3 Présentation de l'application

❖ Outils utilisés pour créer cette application

PHP Laravel (BackEnd) est un framework de développement web open-source écrit en PHP. Il fournit une structure et des outils puissants pour créer des applications web robustes et évolutives. Laravel suit le modèle MVC (Modèle-Vue-Contrôleur), ce qui facilite l'organisation du code et la séparation des préoccupations.

Vuetify.js (FrontEnd) est une bibliothèque open-source de composants d'interface utilisateur basée sur Vue.js. Elle est conçue pour faciliter le développement d'applications web modernes et réactives en fournissant des composants prêts à l'emploi et esthétiquement agréables.

❖ Page d'accueil

Cette page permet à l'utilisateur de se connecter à l'application en introduisant son login et son mot de passe.

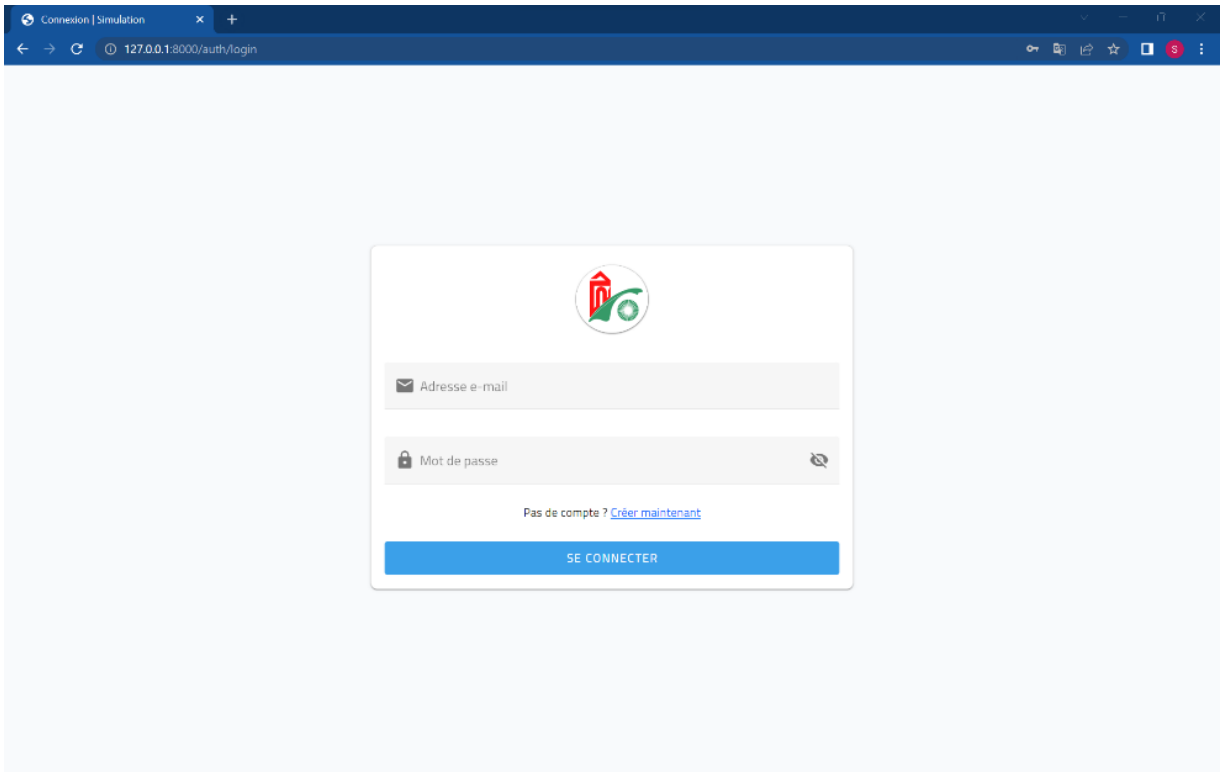


Figure 3.14 : Se connecter au compte.

Un nouvel utilisateur peut s'inscrire pour lui créer un nouveau profil.

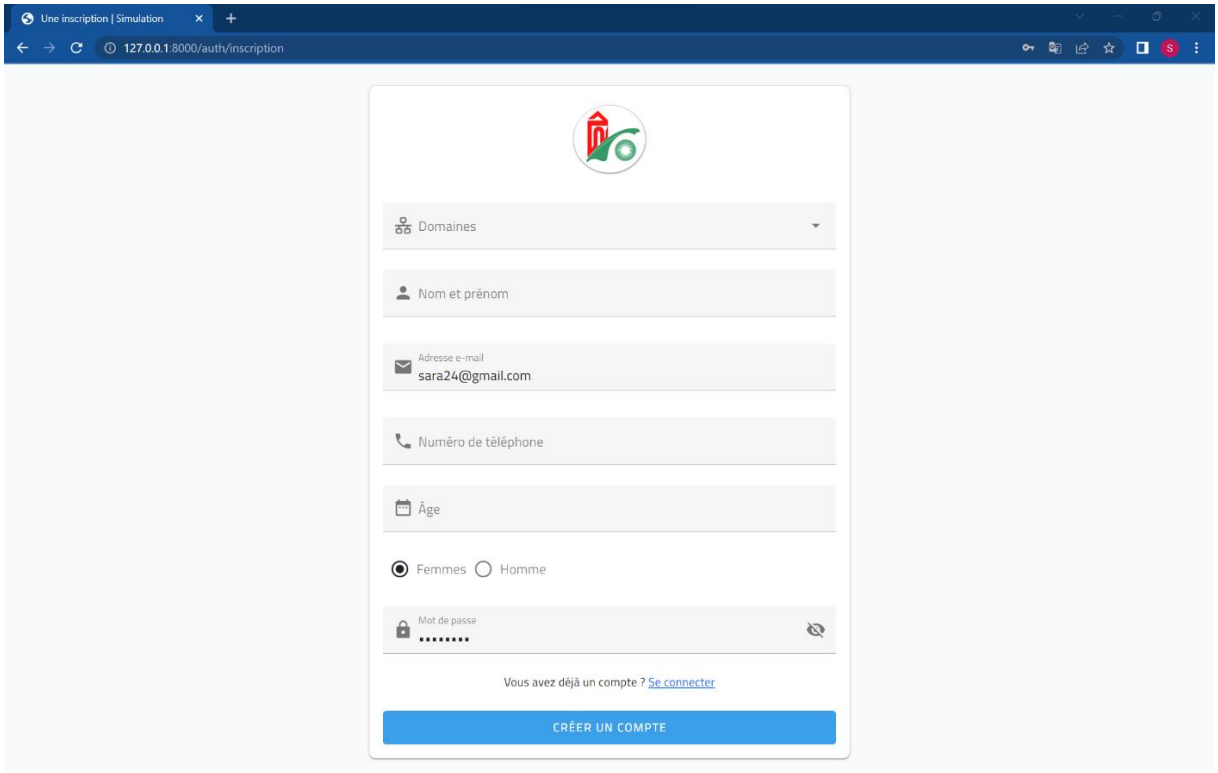


Figure 3.15 : Inscription au compte.

Cette page permet à l'utilisateur de modifier ses informations personnelles.

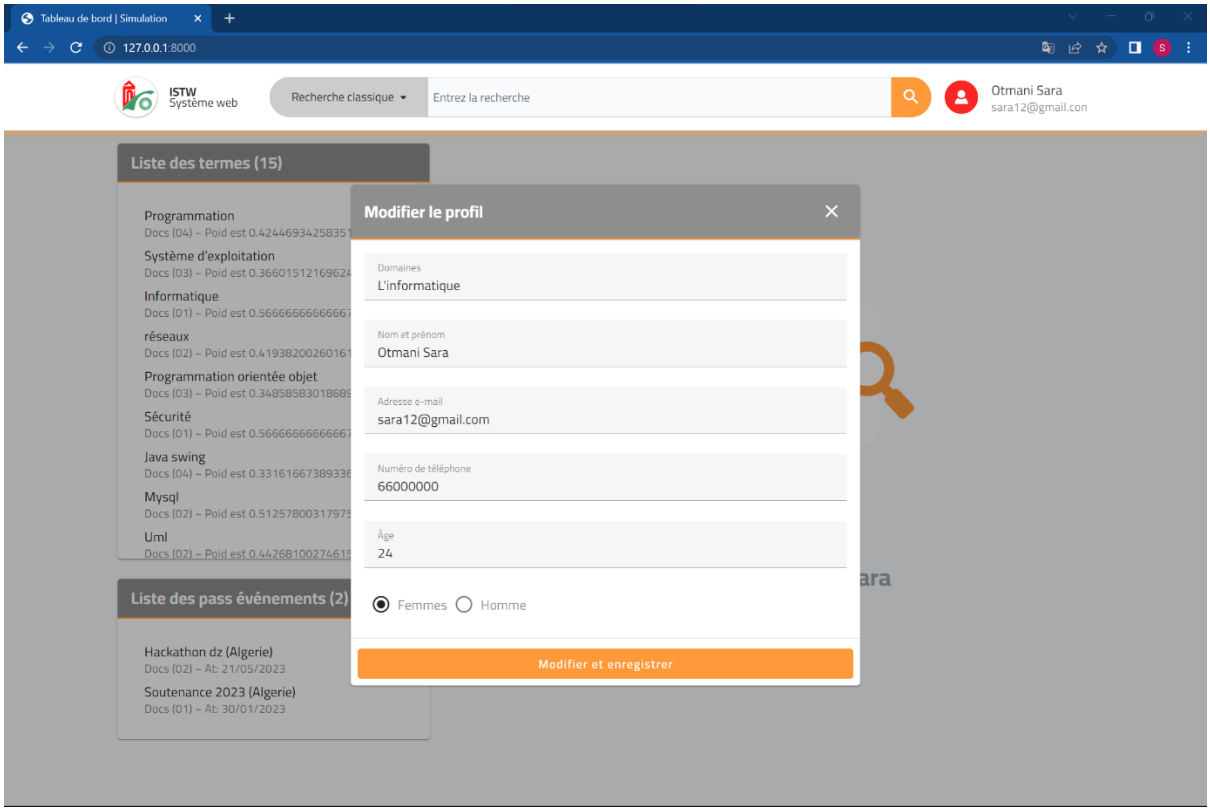


Figure 3.16 : Mettre à jour le profil.

❖ Page de recherche

L'utilisateur peut saisir sa requête Q, et lancer la recherche.

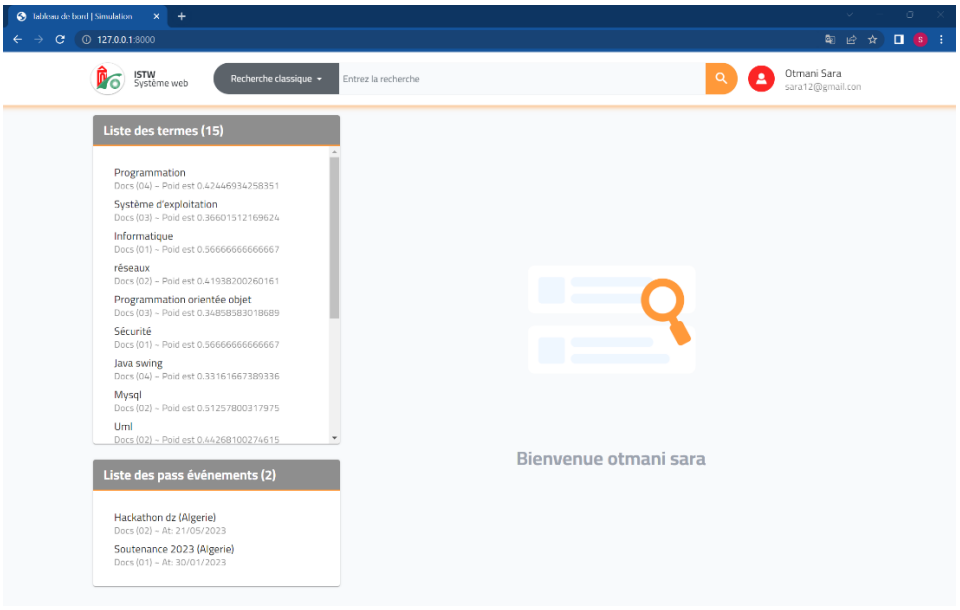


Figure 3.17 : Page de recherche.

La recherche classique passe la requête au moteur de recherche sans reformulation.

La recherche personnalisée permet de passer sur le profil utilisateur pour reformule la Q et relancer la recherche par la suite.

❖ Exemple d'application

Cette figure présente les résultats de recherche classique pour la requête "développement" envoyée par l'utilisateur U1 (étudiant en informatique).

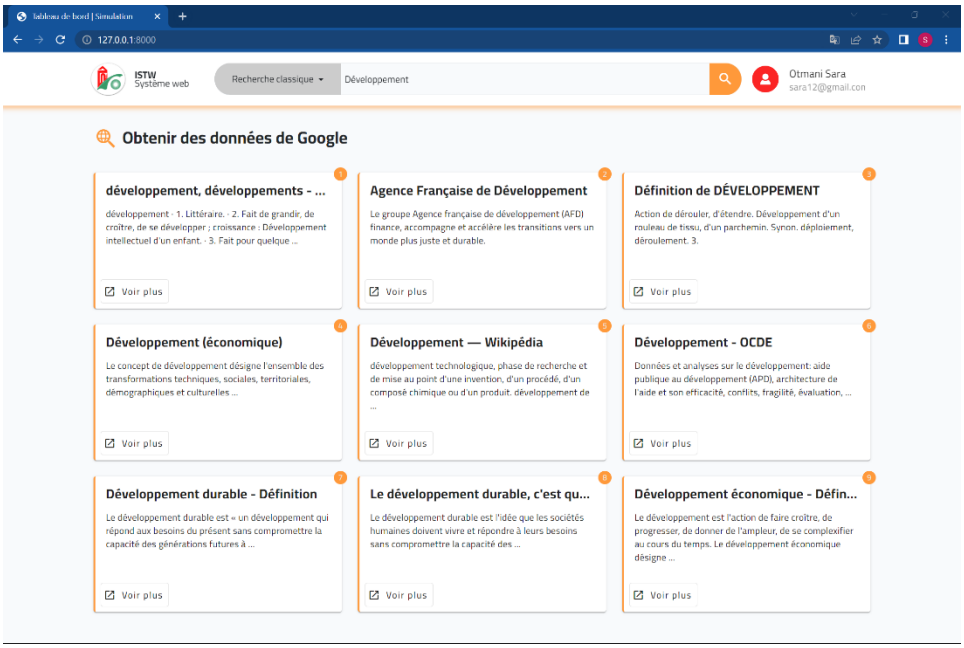


Figure 3.18 : Les résultats de recherche classique.

Pour une recherche personnalisée, cette figure présente l'étape de la sélection de la base la plus adéquate a Q.

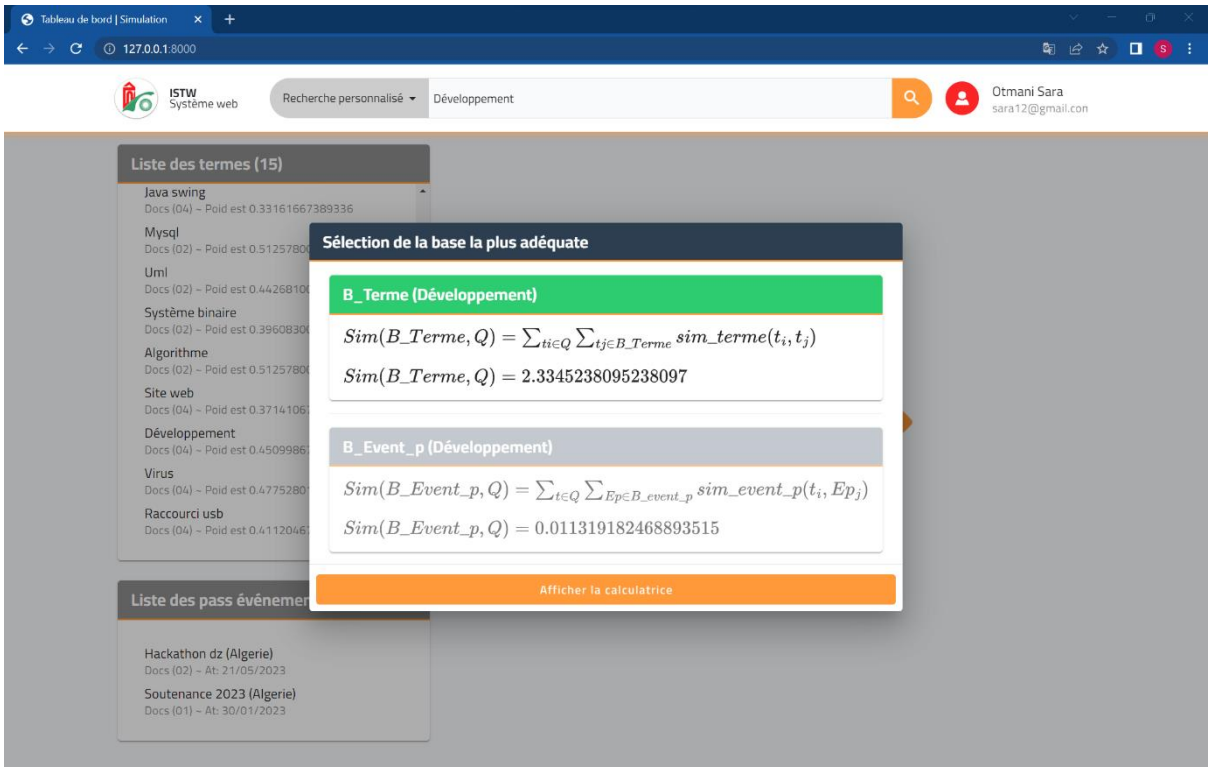


Figure 3.19 : L'étape de la sélection de la base la plus adéquate a Q.

Cette figure présente les calculs de similarité pour reformule Q.

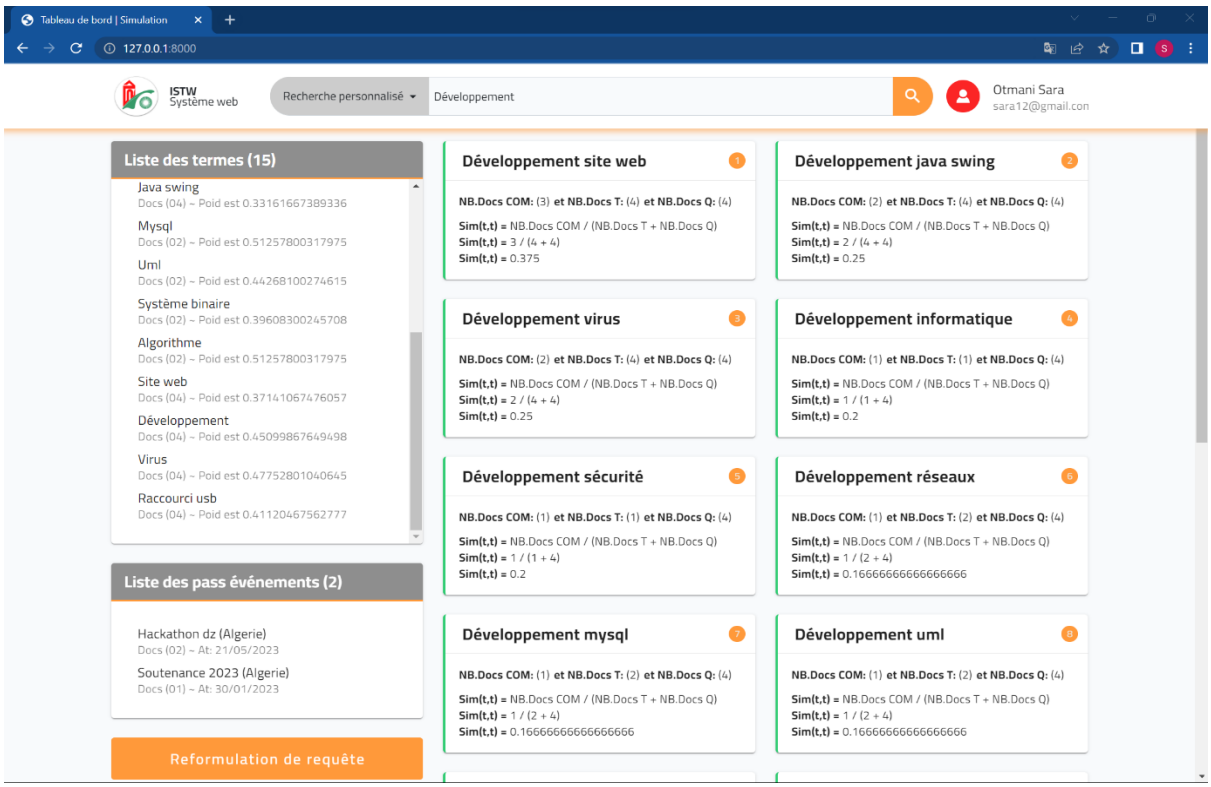


Figure 3.20 : Les calculs de similarité.

Cette figure présente Q reformulée (Q').

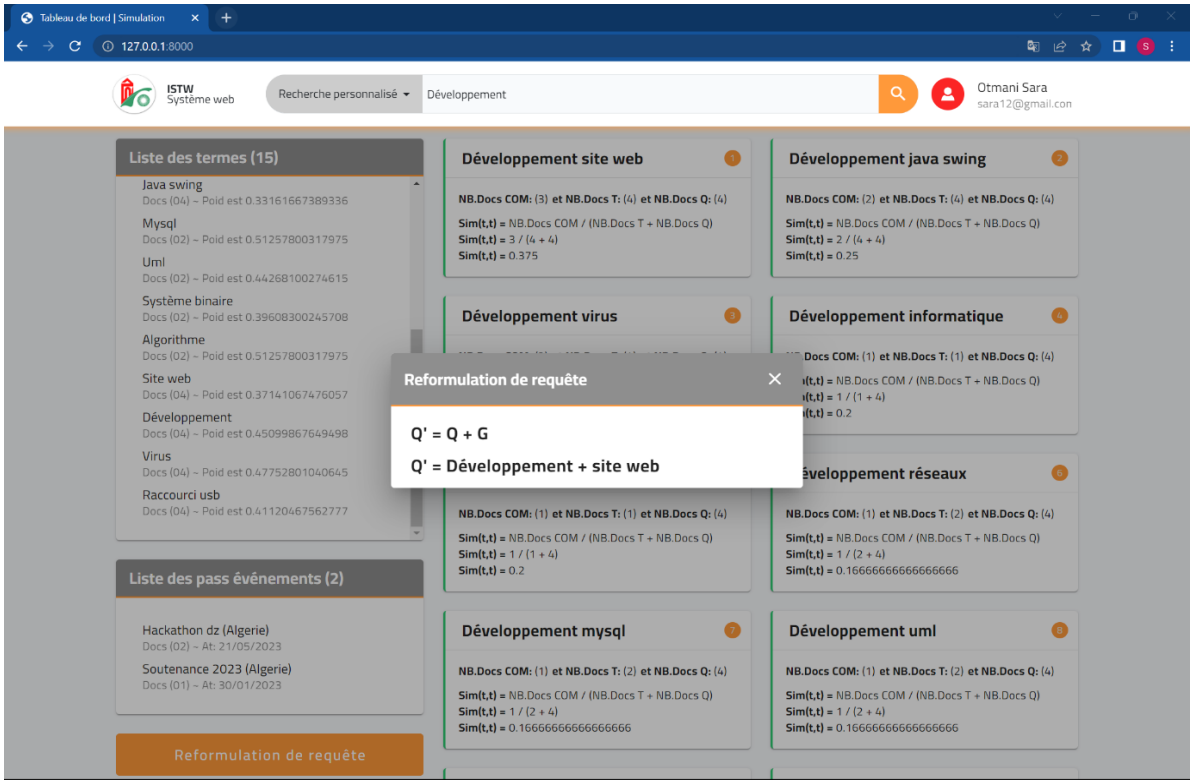


Figure 3.21 : La reformulation de Q.

Cette figure présente les résultats de recherche avec Q' à la recherche classique.

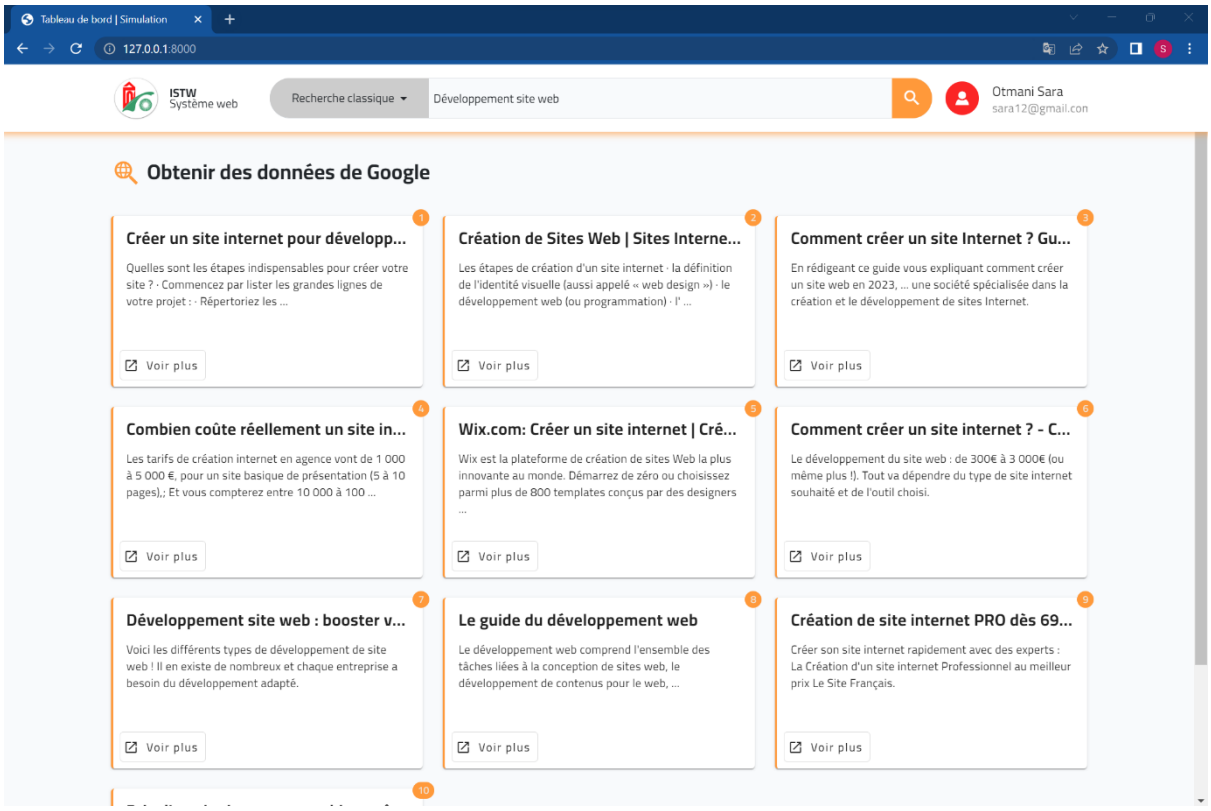


Figure 3.22 : Les résultats de recherche avec Q'.

5 Conclusion

Dans ce chapitre, nous avons présenté notre approche de la personnalisation de la recherche.

Nous avons créé un profil d'utilisateur qui représente l'utilisateur en combinant des informations thématiques et évènementielles.

Une pseudo-sémantique basée sur des liens de cooccurrence est introduite pour représenter les composants du profil. Nous avons également proposé plusieurs formules de pondération et des scores pour mesurer les différents liens de similarité.

Conclusion Générale

Conclusion Générale et Perspectives

Dans ce travail, nous avons proposé une nouvelle approche pour la personnalisation de la recherche d'information, basée sur la définition d'un profil d'utilisateur qui combine les indicateurs d'intérêts de l'utilisateur issus des contextes thématique et évènementiel de ses recherches précédentes.

Nous avons proposé un ensemble de formules de pondération et de scores de similarités qui nous permettent la sélection des capteurs de pertinences les plus important pour l'utilisateur. Ainsi que nous avons introduit une pseudo-sémantique qui s'intéresse aux liens de cooccurrence entre les termes eux-mêmes et entre les termes et les évènements.

Cette approche est proposée dans l'objectif d'améliorer la pertinence des résultats de recherche retournés, par la reformulation de la requête initiale pour mieux exprimer le réel besoin en information de l'utilisateur.

Ce travail ouvre voie vers d'autres pistes de recherche telles que :

- La réflexion à l'implémentation d'un cadre d'évaluation en ligne dédié à la personnalisation de la recherche d'information, avec des comptes utilisateurs et leurs profils, des collections de tests et des préjugements des utilisateurs pour des requêtes prédéfinies, …etc.
- Etudier la stabilité et l'évolution des intérêts de l'utilisateur, pour estimer à quel point son profil peut être utile à inférer ses futurs besoins informationnels.

Bibliographie

[A. Sieg et al. 2004], A. Sieg, B. Mobasher, R. Burke, G. Prabu, and S. Lytinen. Using concept hierarchies to enhance user queries in web-based information retrieval. In The IASTED International Conference on Artificial Intelligence and Applications. Innsbruck, Austria, 2004.

[A. Waern, 04] A. Waern. User involvement in automatic filtering : An experimental study. User Modeling and User-Adapted Interaction, 14(2-3) :201–237, 2004.

[A.Stefani et C.Strapparava, 98] A.Stefani and C.Strapparava. Personalizing access to web sites : The siteif project. In Proceedings of the 2 nd Workshop on Adaptive Hypertext and Hypermedia, USA, pages 69–75, June 1998.

[ABBASSI et al, 13] Un modèle de reformulation des requêtes pour la recherche d'information sur le Web

[Abdelkarim B, 11] ABDELKRIM BOURAMOUL, RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB, Le : 25 / 09 / 2011

[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.

[Akermi et al. 2015] Akermi, I., Boughanem, M., Faiz, R.: Une approche de recommandation proactive dans un environnement mobile. In: INFORSID 2015, France, pp. 301–316 (2015)

[All, 79] Allen J.F,A plan based approch to speech act recognition.Technical report, dept of computer science,University of Toronto,Canada,79.

[Amadieu et al. 09] Amadieu, effets of prior knowledye deversity of learning.2009

[Amato 99] G.amato, U.stroccia, user profil modeling and application to digital librarie.1999

[Arikan et al. 2009] Arikan, I., Bedathur, S.J., Berberich, K. (2009), "TimeWill Tell: Leveraging Temporal Expressions in IR". In Proceedings of the 2ND International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, ACM (2009).

[Arnulphy 2012] Beatrice Arnulphy. (2012), « Désignations nominales des événements : étude et extraction automatique dans les textes. Thèse de doctorat. Universities Paris Sud - Paris XI, 2012.

[B. Pia, 03] B. Pia. The concept of relevance in ir. Journal of the American Society for Information Science, 54(10) :913–925, 2003.

[B.Tan et al. 06] B.Tan, X.Shen, and C.Zhai. Mining long-term search history to improve search accuracy. In Proceedings of the 12th international conference on Knowledge Discovery and Data Mining, University of Illinois at Urbana-Champaign, pages 718–723, August 2006.

[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.

[Balpe et 1996] Balpe J.P., Lelu A., Papy F., Techniques avancées pour l'hypertexte. Hermes (Eds) 1996.

[Basil et al. 2014] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Lucia Siciliani, (2014) "Extending an Information Retrieval System through Time Event Extraction", in 8th International Workshop on Information Filtering and Retrieval, DART 2014, Co-located with XIII AI*IA Symposium on Artificial Intelligence, AI*IA 2014 36-47 CEUR Workshop Proceedings

[Basile et al. 2014] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Lucia Siciliani, (2014) "Extending an Information Retrieval System through Time Event Extraction", in 8th International Workshop on Information Filtering and Retrieval, DART 2014, Co-located with XIII AI*IA Symposium on Artificial Intelligence, AI*IA 2014 36-47 CEUR Workshop Proceedings.

[Belew R, 89] Belew R (1989): Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In: Belkin and Rijsbergen 1989. pp. 11-20.

[Benjabeur et al. 2010] Benjabeur, L, L Tamin and M Boughanem (2010), "A social model for Literature Access: Towards a weighted social network of authors", In International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010), pp. 32–39. France

[Bernardo et al. 2013] Bernardo Pereira Nunes, Stefan Dietze, Marco A. Casanova ,Ricardo Kawase, BesnikFetahu, and Wolfgang Nejd (2013) ―Combining a co-occurrencebased and a semantic measure for entity linking », 10th (ESWC'13) Extended Semantic Web ConferenceAt: Montpellier, FranceVolume: Lecture Notes in Computer Science, Vol. 7882. DOI: 10.1007/978-3-642-38288-8_37.

[Bila et al. 2008] Bila, N, J Cao, R Dino, T Ho, R Hull, B Kumar and P Santos (2008), "Mobile user profile acquisition through network observables and explicit user queries", In 9th International Conference on Mobile Data Management, pp. 98–107, Beijing, China.

[Bischoff et al. 2008] Kerstin BISCHOFF et al. (2008), « Can All Tags Be Used for Search ? » In Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08. California, USA (2008), p. 193-202. DOI : 10.1145/1458082.1458112.

[Borlund 2003] Borlund, P. (2003), "The concept of relevance in IR". Journal of the American Society for Information Science and Technology, 54(10), pp. 913–925.

[Boubekeur F, 08] Boubekeur F. "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets", thèse de doctorat en informatique, Université Paul Sabatier. 2008

[Boughanem M] Boughanem Mohand : Indexation sémantique Latente Latent Semantic Indexing.

[Bouhini 2014] Chahrazed Bouhini, (2014) « Impact des réseaux sociaux sur le processus de recherche d'information », thèse de doctorat, Université Jean Monnet de SaintÉtienne et Universités de Lyon, France.

[Bouhini et al. 2016] Bouhini C, M Gery and C Largeron (2016). "Personalized information retrieval models integrating the user's profile", IEEE 10th International Conference on Research Challenges in Information Science (RCIS 2016), pp. 1–9, Grenoble, France.

[C. Buckley et E. M. Voorhees, 00] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pages 33–40, New York, NY, USA, 2000. ACM.

[C. G. Thomas et G. Fischer, 97] C. G. Thomas and G. Fischer. Using agents to personalize the web. In IUI '97 : Proceedings of the 2nd international conference on Intelligent user interfaces, pages 53–60, New York, NY, USA, 1997. ACM.

[C. Tambellini, 07] C. Tambellini. "Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue". Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, 2007.

[Chahrazed Bouhini] Chahrazed Bouhini: Impact des réseaux sociaux sur le processus de recherche d'information

[Chaudiron, 02] Chaudiron S, Ihadjadenem "Quelle place pour l'usager dans l'évaluation des SRI ? " In : V. Couzinet, G. Regimbeau. Recherches récentes en sciences de l'information : Convergences et dynamiques. Actes du colloque international MICS LERASS, Toulouse. p. 211-232. Mars 2002.

[Chen et Sycara, 1998] Chen L. and Sycara K., Webmate : A personal agent for browsing and searching. In Proceedings of the 2nd international conference on autonomous agents and multi agent systems, Minneapolis, pages 10–13, 1998.

[Chevallet et al, 04] J.P Chevallet, J.Martinez, M.Boughanem, L.LechaniTamine, S.Calabretto, Rapport final de l'AS Passage a l'échelle dans la taille des corpus, janvier 2004.

[CP, 79] cohen P.R Perrault,C,R :Elements of a plan based theory of speech acts cognitive science 3,1979.

[D. Mladenic, 98] D. Mladenic. Personal webwatcher : design and implementation. In Technical Report IJS-DP-7472. J. Stefan Institute, Department for Intelligent Systems, 1998.

[D. W. Oard and J. Kim, 01] D. W. Oard and J. Kim. Modeling information content using observable behavior. In Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology, pages 38–45, USA, 2001.

[D.Kostadinov, 08] D.Kostadinov. Personnalisation de l'information : une approche de gestion de profils et de reformulation de requêtes. Thèse de Doctorat en informatique présentée à l'Université de Versailles Saint-Quentinen- Yvelines, France, Septembre 2008.

[Daoud et al. 2011] Daoud, M, L Tamine and M Boughanem (2011), "A personalized search using a semantic distance measure in a graph-based ranking model". Journal of Information Science, 37(6), 614–636, doi: 10.1177/0165551511420220.

[E. Kraft et A. Bookstein, 78] D. E. Kraft, A. Bookstein. Evaluation of Information Retrieval System: A Decision Theory approach, Journal of the American Society for Information Science, 29 : 31–40, 1978.

[E. Maron et L. Kuhns, 60] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. J. ACM, 7(3) :216–244, 1960.

[E. Rich, 98] E. Rich. User modeling via stereotypes. pages 329–342, 1998.

[F. Liu et al. 2004] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16(1) :28–40, 2004.

[F. Tanudjaja et L. Mui, 02] F. Tanudjaja and L. Mui. Persona : A contextualized and personalized web search. In Proc 35th Hawaii International Conference on System Sciences, page 53, Big Island, Hawaii, January 2002.

[Fatiha B, 08] Fatiha BOUBEKEUR-AMIROUCHE. "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets", soutenue Le 01/ 07/ 2008

[Fouad et al. 2012] Fouad, KM, AR Khalifa, NM Nagdy and HM Harb (2012), "Web-based semantic and personalized information retrieval

[G. Koutrika et Y.I.A, 05] G. Koutrika and Y. Ioannidis. A unified user profile framework for query disambiguation and personalization. In Proceedings of Workshop on New Technologies for Personalized Information Access, July 2005.

[G. Salton et M. McGill, 70] G. Salton and M. McGill. The concept of "relevance" in information science: A historical review. R.R. Bowker, New York, 1970

[G. Salton et M. McGill, 83] G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

[G. Salton, 70] G. Salton. The SMART retrieval system: Experiments in automatic document processing. Prentice Hall. 1970.

[G. Salton, 71] G. Salton. A comparison between manual and automatic indexing methods. Journal of American Documentation, 20(1) :61–71, 1971.

[G.Amato et U.Straccia, 99] G.Amato and U.Straccia. User profile modelling and applications to digital librairies. In Proceedings of the 3 rd European Conference on Research and advanced technology for digital libraries, pages 184–187, 1999.

[G.Gentili et al. 03] G.Gentili, A.Micarelli, and F.Sciarrone. Infoweb : An adaptive information filtering system for the cultural heritage domain. Applied Artificial Intelligence (AAI), 17(8-9) :715–744, 2003.

[G.Salton et C.Yang, 97] G.Salton and C.Yang. On the specification of term values in automatic indexing. Journal of documentation, 29 :351–372, 1973.

[Gauch et al., 2007] Gauch S., Speretta M., Chandramouli A., Micarelli A., « User Profiles for Personalized Information Access », dans Brusilovsky P., Kobsa A., Nejdl W. (dirs.), The Adaptive Web, Springer Berlin Heidelberg (Lecture Notes in Computer Science), p. 54 89, 2007.

[H. Lieberman, 95] H. Lieberman. Letizia : an agent that assists web browsing. In IJCAI 95 : Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence., pages 924–929, 1995.

[H.Lieberman, 97] H.Lieberman. Autonomous interface agents. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, New York USA, pages 67–74, April 1997.

[H.Sakagami et T.Kamba, 97] H.Sakagami and T.Kamba. Learning personal preferences on online newspaper articles from user behaviors. In Proceedings of the 6 th International WWW Conference, Santa Clara, California, pages 7–11, April 1997.

[HAMOUM A et ABERKANE F] HAMOUM Amine, ABERKANE Ferhat : Implémentation et évaluation d'une méthode de sélection de documents d'expansion basée sur le score de clarté de la requête.

[Hattori et al. 2007] Hattori S., Tezuka T., Tanaka K. (2007), «Context-aware query refinement for mobile web search», Proc. of the 2007 International Symposium on Applications and the Internet Workshops.

[Heymann et al. 2008] Paul Heymann, Daniel Ramage et Hector Garcia-Molina (2008), « Social Tag Prediction ». In Proceedings of the 31st ACM SIGIR'08, Singapore, DOI: 10. 1145 / 1390334.1390425.

[Ishioka, T, 03] Ishioka, T. Evaluation of criteria for information retrieval. Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on Volume, Issue, 13-17 Oct. 2003.

[J. Chaffee et S. Gauch] J. Chaffee and S. Gauch, "Personal ontologies for web navigation. In Proceedings of the ninth international conference on Information and knowledge management, , McLean,"

[J. Teevan et al. 2005] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 449–456, New York, NY, USA, 2005. ACM.

[J.Gowan, 03] J.Gowan. A multiple model approach to personalised information access. Mémoire de Master en informatique présenté à l'Université de College Dublin, February 2003.

[K. Carlos et A. Cuadro, 67] R. V. K. Carlos A. Cuadro. Opening the black box of relevance. Journal of Documentation, 23(4) :291–303, 1967.

[K. Jarvelin et J. Kekalainen, 00] K. Jarvelin and J. Kekalainen. Ir evaluation methods for highly relevant documents. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 41–48. Belkin and al Eds, 2000.

[K. Jarvelin et J. Kekalainen, 02] K. Jarvelin and J. Kekalainen. Cumulative gain-based evaluation of ir techniques. ACM Transactions on Information Systems, 20(4) :422–446, 2002.

[Kanhabua et al. 2012] Kanhabua, N., Nørvag, K. (2012), "Learning to Rank Search Results for Time sensitive Queries", In Proceedings of the 21st ACM International Conference on Information and Knowledge Management CIKM '12, pp. 2463–2466, 2012.

[Koc, 00] Koch N., Software Engineering for Adaptative Hypermedia Systems – Reference Model ,Modeling techniques and development process, Ph.D Thesis, Fakultat der Mathematik und Informatik, Ludwig-Maximilians-Universität München, December 2000.

[Kwok K. L, 89] Kwok K. L. (1989): A Neural Network for Probabilistic Information Retrieval. In: Belkin and Rijsbergen 1989. pp. 21-30.

[L. Barry, 94] L. Barry. User-defined relevance criteria: an exploratory study. Journal of the American Society for Information Science, 45 :149–159, 1994.

[L. Chen et K. Sycara, 98] L. Chen and K. Sycara. Webmate : A personal agent for browsing and searching. In Proceedings of the 2nd international conference on autonomous agents and multi agent systems, Minneapolis, pages 10–13, 1998.

[L. Tamine and S. Calabretto, 08] L. Tamine and S. Calabretto. Recherche d'information contextuelle et web. In M. Boughanem and J. Savoy, editors, Recherche d'information : état des lieux et perspectives, volume 1, chapter 7, pages 201–224. Hermès, http ://www.editions-hermes.fr/, avril 2008.

[Lechani et Boughanem 2005] Lechani L., Boughanem M. Accès personnalisé à l'information : Approches et Techniques. Equipe SIG/RFI , IRIT/Toulouse, Rapport Interne, Janvier 2005.

[Ludovic et al. 2013] Ludovic Jean-Louis, Romaric Besançon et Olivier Ferret. (2013). « Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires », in TAL. Volume 54 – n° 1/2013, pages 139 à 170.

[M. J. Pazzani et al. 1996] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill & webert : Identifying interesting web sites. In Proceedings of the 30th National Conference on Artificial Intelligence, pages 54–61, Portland, 1996.

[M.Daoud, 09] M.Daoud. Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche. Thèse de Doctorat en informatique présentée à l'Université Paul Sabatier, France, Decembre 2009.

[M.N. Huhn et L. Stephens, 99] M.N. Huhn and L. Stephens, Personal ontologies. Internet Computing 3(5)., October 1999.

[M.Pazienza, 97] M.Pazienza. Information extraction : a multidisciplinary approch to an emerging information technology. In proceedings of International Summer School, SCIE-97 Frascati, Italy, July 1997.

[Manning C et al. 08] Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press; 1st edition.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press ; 1st edition.

[Mariam D, 09] Mariam DAOUD. "Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche", thèse de doctorat en informatique, Université Paul Sabatier. (2009)

[Menegon, 09] Menegon D, Mizzaro S, Nazzi E, Vassena L "Benchmark evaluation of context-aware Web search". Workshop on Contextual Information Access, Seeking and Retrieval Evaluation, Toulouse, France, Springer, April 2009.

[Mohammed et Abdelmoty 2017] Mohamed, S and AI Abdelmoty (2017), "Spatio-semantic user profiles in location-based social networks". IN International Journal of Data Science and Analytics, 4, 127–142, 2017. DOI: 10.1007/ s41060-017-0059-9.

[Mohand Boughanem, 92] Mohand Boughanem: les Systèmes de Recherche d'Information : d'un modèle classique à un modèle connexionniste. Thèse de Doctorat de l'Université Paul Sabatier, 1992.

[Monteserin et Armentano 2019] Ariel Monteserin Marcelo and G. Armentano (2019) « Influence me! Predicting links to influential users », Information Retrieval Journal (2019) 22, pp 32–54, https://doi.org/10.1007/s10791-018-9335-0.

[Mothe J, 94] Mothe J., Modèle Connexionniste Pour la Recherche d 'Informations. Expansion Dirigée de Requêtes et Apprentissage. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Octobre 1994.

[Moulahi 2015] Bilel Moulahi (2015), « Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information », thèse de doctorat, Université Toulouse III Paul Sabatier, 2015.

[Mousset et al. 2019] P Mousset, Y Pitarch, L Tamine (2019). « Régularisation Spatiale de Représentations Distribuées de Mots‖, In proceedings of 16th CORIA'19 conference, Lyon, France. Doi:10.24348/coria.2019.CORIA_2019_paper_23.

[Nawal O] Nawal OULD AMER : Recherche Sociale et Personnalisée d'Information

[Noll et Meinel 2007] M. G. Noll et Christoph Meinel, (2007) « Web Search Personalization via Social Bookmarking and Tagging ». In Proceedings of the 6th International conference of the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference. 2007, p. 367-380.

[O. Bouidghaghen 2011] O. Buidghaghen., « 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 ». thèse de doctorat, Univ Toulouse 3, France, (2011).

[Ouaneche L, 13] Melle. Ouaneche Linda, Système de recherche d'information pour la langue arabe : Utilisation du modèle vectoriel à base d'ontologie. Thèse de master, université Hassiba Ben Bouali, (2013)

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

[Ould Amer et al. 2015] Nawal Ould Amer, Philippe Mulhem and Mathias Géry (2015), "Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter », 12th CORIA'15 conférence, Paris-France. doi:10.24348/coria.2015.6.

[P. Borlund et P. Ingwersen, 98] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life : performance indicators in interactive ir. In Proceedings of the 21st annual

international ACM SIGIR conference on Research and development in information retrieval, pages 324–331, Melbourne, Australia, August 1998. Croft W.B et al Eds, ACM.

[P. Borlund, 03] P. Borlund. The iir evaluation model : a framework for evaluation of interactive information retrieval systems. Information Research, 8(3) :152–179, Avril 2003.

[P. Ingwersen, 94] P. Ingwersen. "Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction". In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994.

[Panayiotou et al. 06] Panayiotou C., Samaras G., (2006), «Mobile User Personalization with Dynamic Profiles: Time and Activity», On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, LNCS 4278, pp.1295-1304, 2006.

[Paris, 04] La recherche d'information, du texte intégral au thésaurus, ed. HermesScience, paris 2004.

[R. Baeza et A. Ribeiro, 99] R. Baeza-Yates and R. A. Ribeiro-Neto. Modern Information Retrieval. New York: ACM Press; Harlow England: Addison-Wesley, cop., 1999.

[R. Kraft et al. 2005] R. Kraft, F. Maghoul, and C. Chang. Y !q : contextual search at the point of inspiration. In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management, pages 816–823, New York, NY, USA, October 2005. ACM Press.

[S. Dumais et al. 2003] S. Dumais, E. C. J. J. Cadiz, G. Jancke, R. Sarin, and C. R. Daniel. Stuff i've seen : a system for personal information retrieval and re-use. In SIGIR'03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pages 72–79. ACM Press, 2003.

[S. Gauch et al. 03] S. Gauch, J. Chaffe, and A. Pretschner., "Ontology-based personalized search and browsing.," in Web Intelligence and Agent System, 2003, pp. 219–234.

[S. Gauch et al. 2003] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 1(3-4) :219– 234, 2003

[S. Robertson et K. Sparck, 76] S. Robertson and K. Sparck Jones. Relevance weighting for search terms. Journal of The American Society for Information Science, 27(3) :129–146, 1976.

[Salton et al 1975] Salton G., Wong A., Yang C., (1975) A vector space model for automatic indexing. Communications of the ACM, 18, 1975 pp. 613-620.

[Salton et Buckley 1988] Salton, G. and Buckley, C., (1988) — Term Weighting Approaches in Automatic Text Retrival‖, Information Processing and Management. 513-523.

[Samir K, 09] Samir KECHID. "Intégration du modèle utilisateur dans un système de recherche d'information distribuée", diplôme de Doctorat, Soutenue le 11/04/2009.

[Saoud et Kechid 2016] Saoud, Z and S Kechid (2016), "Integrating social profile to improve the source selection and the result merging process in distributed information retrieval". Information Sciences, 336, 115–128, http://dx.doi.org/10.1016/j.ins.2015.12.012.

[Saracevic 2007] Saracevic, T. (2007), “Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii : Behavior and effects of relevance”, Journal of the American Society for Information Science, 58(13):2126–2144

[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 socialtagging networks”. In 31st Conference on Research and Development in Information Retrieval, SIGIR’08, pages 523–530.

[Shafik et al. 2015] O.Shafiq, R.Alhajj, J.G.Rokne, (2015) “On personalizing web search using social network analysis”, Inf.Sci. 314(2015)55–76. http://dx.doi.org/10.1016/j.ins.2015.02.029

[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.

[Sparck J, 71] Sparck Jones K., Automatic keyword classification for information retrieval, 1971.

[Sparck J, 79] Sparck Jones K., Experiments in relevance weighting of search terms. IPM, 15(3), 1979.

[T. H. Duong et al. 2013] Duong, TH, MN Uddin and CD Nguyen (2013), “Personalized semantic search using ODP: A study case in academic domain”. In ICCSA 2013, Part V, LNCS 7975, pp. 607–619.

[T. Kohonen, 98] T. Kohonen. Self_Organization and Associative Memory. Springer Verlag. ISBN 0387513876, 1989.

[Tamine, 10] Tamine L, Boughanem M, Daoud M “Evaluation of contextual information retrieval effectiveness: overview of issues and research”. Journal of Knowledge and Information Systems. Volume 24 Issue 1, pp. 1-34. Springer, Londres, United Kingdom. July 2010.

[Taylor et al. 2007] Taylor, A. R., Cool, C., Belkin, N. J. et Amadio, W. J. (2007), “Relationships between categories of relevance criteria and stage in task completion”. Information Processing and Management, 43(4):1071–1084.

[Uddin et al. 2012] Uddin, MN, TH Duong, V Sean and G-S Jo (2012), “Construction of semantic user profile for personalized web search”. In ICCCI 2012, Part II, LNAI7654, pp. 99–108. Springer.

[V.Raghavan et H.Sever, 95] V.Raghavan and H.Sever. On the reuse of past optimal queries. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, NY, USA, pages 344–350, 1995.

[Vallet et al. 2010] David Vallet, Iván Cantador et Joemon M. Jose. « Personalizing Web Search with Folksonomy-Based User and Document Profiles ». In Advances in Information Retrieval, 32nd ECIR 2010, UK, p 420-431.

[Vandenbussche et al. 2011] Vandenbussche, P.Y., Teissèdre, C., (2011), "Events Retrieval Using Enhanced SemanticWeb Knowledge". In Workshop DeRIVE 2011, in conjunction with 10th International Semantic Web Conference 2011 (ISWC 2011).

[Widyantoro et al., 1999] Widyantoro D., Yin J., El Nasr M., Yang L., Zacchi A., and Yen J., Alipes : A swift messenger in cyberspace. In Proceedings of Spring Symposium Workshop on Intelligent Agents in Cyberspace, pages 62–67, Stanford, March 22-24 1999.

[X. Shen et al. 2005] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management, pages 824–831, New York, NY, USA, 2005. ACM.

[X. Soergei, G. Marchionini, 91] Lin, X., Soergei, D., and Marchionini, G. (1991) "A self-organizing semantic map for information retrieval". Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, (Chicago, IL, USA), pp. 262-269.

[Xu et al. 2008] Shengliang XU et al. (2008) « Exploring Folksonomy for Personalized Search ». In Proceedings of the 31st ACM SIGIR'08, Singapore, URL: https://doi.org/10.1145/1390334. 1390363.

[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.