République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud Mammeri de Tizi-Ouzou

Faculté de : Génie électrique et d'informatique

Département : Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de
Master en Informatique

Spécialité : Ingénierie des systèmes d'information

Présenté par :

HABBAS Yamina et HACHOUR Lynda

Thème : « Intégration de la dimension temps dans la représentation du profil de l'utilisateur pour une recherche d'information personnalisée »

Proposé et dirigé par : ACHEMOUKH Farida

Soutenu le 10 octobre 2019 devant le jury composé de:

Mme. FELLAG Samia	Présidente du Jury
Mme. BENTAYEB Mouna	Membre du Jury
Mme. ACHEMOUKH Farida	Directrice de mémoire

Remerciements

Tout d'abord, nous remercions Dieu le tout puissant de nous avoir aidé et donné la force de rédiger ce mémoire.

Nous adressons notre plus profonde gratitude à notre promotrice Mme ACHEMOUKH Farida, qui a toujours su orienté nos recherches, et pour tous ses conseils et le temps qu'elle nous a accordée.

Nous tenons à remercier également les membres de jury d'avoir accepté d'évaluer notre travail.

Nos remerciements vont également à nos amis(es) qui nous ont toujours encouragés et à toute personne ayant contribué de près ou de loin.

Dédicaces

Je dédie ce travail à :

Mes très chers parents,

Mon frère et ma sœur,

Mes grands-parents,

Et mes amis(es).

Yamina

Je dédie ce travail à :

Mes très chers parents,

Mes deux frères et ma sœur,

Mon Mari,

Et mes amis(es).

Lynda

Résumé

Notre travail s'inscrit dans le domaine de la recherche d'information (RI), plus particulièrement dans l'accès personnalisé a l'information. Tel qu'un même utilisateur peut avoir différents besoins à différents instants.

La surabondance de l'information ainsi que sa large accessibilité à travers le web ont engendré une dégradation des performances des systèmes de recherche d'information. L'origine de ces problèmes réside en partie dans le caractère non personnalisé à l'information. La personnalisation est une solution appropriée pour améliorer la qualité de la recherche de ces systèmes en prenant en compte le besoin spécifique de l'utilisateur représenté par son profil.

Comme le facteur temps a gagné beaucoup d'importance ces dernières années, la dynamique temporelle est introduite pour étudier l'évolution du profil utilisateur qui consiste principalement à saisir les changements des intérêts de l'utilisateur en fonction du temps.

La fraîcheur d'un centre intérêt définissant le profil est supposée être résolu uniquement par le nombre de termes dans ce profil mais pas par leurs positions dans le temps. Comme Les intérêts des utilisateurs évoluent dans le temps, et que les termes utilisés récemment dans les interactions de recherche de l'utilisateur contiennent de nouveaux intérêts et doivent être pris en considération plus que les anciens intérêts surtout que de nombreux travaux antérieurs ont prouvé que l'intérêt de l'utilisateur diminue avec le temps.

Nous avons proposé une approche de personnalisation qui intègre la dimension temps dans la représentation du profil utilisateur. Dans cette approche nous avons utilisé une représentation vectorielle qui prend en compte la dimension temps mesurée en combinant la fréquence normalisée des termes et leurs fraîcheurs en utilisant une fonction à noyaux gaussien.

Mots clés : Recherche d'information, Recherche d'information personnalisée, Modélisation d'un utilisateur, profil utilisateur, fraîcheur, fonction temporelle.

Abstract

Our work is in the field of information retrieval (IR), more particularly in personalized access to information. As a single user can have different needs at different times.

The overabundance of information and its wide accessibility across the web has led to a deterioration in the performance of information retrieval systems. The origin of these problems lies partly in the non-personalized nature of the information. Personalization is an appropriate solution to improve the search quality of these systems by taking into account the specific need of the user represented by his profile.

As the time factor has gained a lot of importance in recent years, the temporal dynamics are introduced to study the evolution of the user profile that consists mainly of capturing changes in the user's interests over time.

The freshness of an interest center defining a profile is supposed to be solved only by the counts of terms in this profile but not by their positions in time. As user interests evolve over time, and the terms recently used in the user's search interactions contain new interests and must be taken into consideration more than the old interests. In fact, many prior works have proved that the user interest decreasing as time goes by.

We proposed a personalization approach that incorporates a temporal dimension into the representation of the user profile. In this approach we used a vector representation that takes into account the measured temporal dimension by combining the normalized frequency of the terms and their freshness using a Gaussian kernel function.

Key-words: Information Retrieval, Personalized Search, User modeling, User profile, Freshness, Temporal function.

Table des matières

Table des tableaux et des figures

Introduction générale

1 Contexte du travail

Aujourd'hui, l'information joue un rôle primordial dans le quotidien des individus. Pour retrouver les informations adaptées à leurs besoins, les utilisateurs se servent de différents outils de recherche en particulier le web. Cependant, avec l'essor d'internet et son évolution au cours du temps, en particulier avec la croissance du nombre d'utilisateurs qui atteignait plus de 4.5 billions en juin 2019[1], il est de plus en plus difficile de trouver l'information qui correspond au mieux à l'attente de l'utilisateur.

La recherche d'information (RI) est le domaine par excellence qui s'intéresse à répondre à ce type d'attente. L'objectif principal de la RI est de fournir des techniques et des outils pour rechercher, organiser des masses d'informations et localiser celles qui seraient pertinentes relativement à un besoin en information d'un utilisateur, souvent, exprimé à travers une requête. Ces outils sont appelés des Systèmes de Recherche d'Information (SRI).

De manière générale, le fonctionnement d'un SRI consiste à construire une représentation des documents et de la requête et d'établir une comparaison entre ces deux représentations (requête, documents) pour retourner les documents pertinents. Les utilisateurs de ces systèmes font face à une surcharge informationnelle qui les désoriente car ces systèmes ne prennent pas en considération le contexte dans lequel la requête a été soumise.

En clair, le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte d'utilisation spécifique : une même requête est soumise par deux utilisateurs ayant des préférences et des intérêts différents, notamment lorsque les requêtes sont courtes et ambigües. Parmi les exemples des requêtes ambigües, on cite : perche (poisson ou instrument), java (café, langage de programmation ou île), jaguar (Apple software, animal ou voiture).

La personnalisation de la recherche d'informations a été proposée pour pallier à ces problèmes. L'objectif des systèmes de RI personnalisés est de fournir des résultats de recherche qui correspondent aux intérêts et aux besoins en information de chaque utilisateur, au lieu de toujours fournir les mêmes résultats à une requête quel que soit l'utilisateur qui l'a soumise. Pour atteindre cet objectif, l'emploi d'un profil a été adopté. En fait, le profil de l'utilisateur a été considéré comme l'élément contextuel le plus important qui permet d'améliorer la pertinence de la recherche.

[1]www.internetworldstats.com/stats.htm

Etudier le profil utilisateur nous amène à étudier les informations qui permettent de représenter ses intérêts. Cette représentation nécessite la collecte d'informations sur l'utilisateur d'une manière explicite en se basant sur les informations fournies par l'utilisateur où implicitement à partir de l'historique de ses recherches. Quant à l'exploitation du profil dans le processus de RI, les techniques les plus élaborées se basent sur son intégration dans l'une des phases de ce dernier, notamment la reformulation de requêtes, l'appariement requête-document ou le ré-ordonnancement des résultats.

La personnalisation de l'information a engendré le problème d'évolution du profil de l'utilisateur au cours du temps : Un intérêt utilisateur définissant son profil, n'est pas lié seulement, à un domaine et à un sujet particulier, mais aussi il est lié au temps (quand est-ce qu'un utilisateur montre un intérêt / désintérêt pour un centre d'intérêt particulier ? l'intérêt est-il récent ?).

Pour remédier à cette problématique, il est important d'introduire le facteur temps dans la modélisation du profil utilisateur car les utilisateurs sont plus intéressés par le contenu récent et leurs intérêts peuvent être pertinent a un moment donné et ne plus être significatif ultérieurement.

2 Problématique et motivations

Le processus de personnalisation dans les systèmes de recherche d'information est principalement confronté à la question de la définition des facteurs nécessaires intervenants dans la représentation du profil utilisateur. Cependant, pour améliorer la pertinence de la recherche, les SRI ne doivent pas se limiter à la fréquence de termes dans le profil, étant donné que les intérêts des utilisateurs évoluent de manière particulièrement rapide.

Par conséquent, nous nous intéressons à deux problématiques :

- **Comment représenter le profil utilisateur en se basant sur les données issues de ses interactions avec le SRI ?**

  La représentation des données concernant l'utilisateur et leur intégration dans le processus de recherche d'information est toujours d'actualité. Nous constatons que l'une des

principales raisons du manque de performances des techniques de personnalisation est liée aux types de données utilisées pour dériver les intérêts de l'utilisateur.

Le profil utilisateur explicite est considéré la plupart du temps comme une charge pour les utilisateurs qui ne sont pas toujours disposés à spécifier leurs informations personnelles. La création implicite du profil utilisateur est une solution permettant de recueillir plus d'informations sur l'utilisateur.

Notre objectif est de représenter implicitement le profil utilisateur comme un vecteur de termes pondérés qui correspond à un centre d'intérêt de l'utilisateur extrait de ses interactions de recherche. Nous considérons les documents pertinents sélectionnés par l'utilisateur lors de chaque interaction de recherche comme source de données utilisée pour inférer un centre d'intérêt intervenant dans la définition d'un profil utilisateur lié à une session de recherche.

- **Peut-on intégrer un facteur temps dans la représentation du profil utilisateur ?**

De nombreuses approches assignent plus d'importance aux termes fréquents peu importe leurs moments d'apparition. Comme le profil utilisateur évolue au fil du temps et certains centres d'intérêts récents ne représenteront plus ses besoins actuels. Dans d'autres cas, les centres d'intérêts récents de l'utilisateur peuvent être liés à un besoin spécifique et temporaire qui ne représente pas ses intérêts récurrents.

Pour cela, nous proposons d'intégrer un facteur temps afin de suivre l'évolution des centres d'intérêts des utilisateurs et donner plus d'importance aux centres d'intérêts récents, sans négliger les centres d'intérêts anciens.

Par exemple, un architecte peut faire des recherches sur le virus de la grippe lorsqu'il a un rhume. Ce besoin est juste temporaire et ne représente pas un centre d'intérêt récurent et pertinent.

Le temps est souvent utilisé pour discerner les profils utilisateurs à court et à long terme. Le premier type de profil est limité aux centres d'intérêts liés aux interactions de recherche courantes de l'utilisateur tandis que le second représente les centres d'intérêts persistants de l'utilisateur extraits de ses interactions de recherche. Le fait de discerner les centres d'intérêts à court et à long terme requiert l'utilisation d'un intervalle de temps pouvant inclure plusieurs centres d'intérêts (Dumais et al, 2003)

3 Contribution

Dans notre travail, nous nous basons sur un modèle multidimensionnel qui permet de représenter le profil selon deux dimensions. Nous nous sommes basés sur l'hypothèse que les termes utilisés récemment lors des interactions de recherche de l'utilisateur contiennent des informations supplémentaires expliquant mieux ses intérêts et que la fréquence d'un terme ne reflète pas nécessairement son importance, étant donné que les utilisateurs s'intéressent de plus en plus au contenu récent.

Nous avons utilisé une représentation vectorielle qui prend en compte la fréquence temporelle mesurée en combinant la fréquence normalisée des termes et leurs fraicheurs en utilisant une fonction à noyaux gaussien. Par la suite, nous étudions l'évolution des centres d'intérêts au cours du temps en s'appuyant sur l'intégration du facteur temps dans la représentation du profil utilisateur.

4 Organisation du mémoire

Notre travail est réparti sur trois chapitres :

Chapitre 1 : le passage de la RI classique vers la RI personnalisée. Dans ce chapitre nous présentons les généralités sur la recherche d'information.

Chapitre 2 : modélisation et évolution du profil utilisateur, ce chapitre est un état de l'art sur les différentes approches de représentation, construction et évolution du profil utilisateur.

Chapitre3 : intégration de la dimension temps dans la représentation du profil utilisateur. Dans ce chapitre, nous expliquons l'approche que nous avons proposée.

Chapitre 1 : De la RI classique vers la RI personnalisée

1 Introduction

La surabondance de l'information sur le Web, ayant remis en cause les modèles classiques de la Recherche d'Information(RI), présente un souci d'abondance des données fournies à l'utilisateur en réponse à une requête. De ce fait, les systèmes de Recherche d'Information (SRIs) font face à de nouveaux défis liés à la pertinence de l'information.

Les travaux en RI classique se sont orientés vers des approches adaptatives. Ces dernières exploitent diverses sources afin d'aider l'utilisateur dans sa quête de l'information pertinente. Néanmoins, la RI adaptative présente des limitations principalement liées à la représentation de l'utilisateur et de ses besoins. Cette limitation a conduit vers l'émergence de la recherche d'information personnalisée. Ce domaine apporte principalement la prise en compte de l'utilisateur en tant que composante principale dans le processus de recherche, cependant la modélisation de l'utilisateur permet de mieux cibler les données fournies en fonction des intérêts de ce dernier ainsi que de ses besoins.

Ce chapitre traite en premier les concepts de base de la RI ainsi que le processus de recherche et les modèles de recherche d'information, ensuite nous mettons l'accent sur les facteurs d'émergence de la RI personnalisée où nous abordons la définition de la RI personnalisée et celle d'un système de recherche d'information personnalisée (SRIP) et on termine avec les approches d'évaluation d'un SRIP.

2 Définition de la Recherche d'Information

La recherche d'information (information retrieval en anglais) est un domaine qui fournit des techniques et des outils permettant la représentation, le stockage, l'organisation et la recherche, dans une masse documentaire existante, des documents contenant l'information qui répond au besoin informationnel exprimé par l'utilisateur sous forme de requête (ON-AT, 2017).

3 Définition d'un Système de Recherche d'Information

Un système de recherche d'information (SRI) est un ensemble de techniques qui assurent les fonctions nécessaires pour la RI. Il a pour rôle de sélectionner les documents qui peuvent répondre au besoin en information de l'utilisateur formulé par une requête de recherche (Hannech, 2018).

Ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur avec une représentation du contenu des documents au moyen d'une fonction de correspondance. Ces derniers doivent pouvoir traiter : de grandes masses d'informations en langage naturel et de façon rapide et pertinente.
Cette définition fait ressortir trois notions clés : document, requête, pertinence.

4 Les fondements de la Recherche d'Information

La recherche d'information a pour but de retrouver, parmi une collection de documents préalablement stockée, les documents pertinents qui répondent au besoin en information d'un utilisateur exprimé par une requête.

Notions de base :

- **Requête** : la requête est l'interprétation d'un besoin en information de l'utilisateur qui peut être exprimée comme un ensemble de mots clés, en langage naturel (sous forme de texte libre), ou booléen (mots clés reliés par des connecteurs logiques).
- **Document** : le document représente le conteneur élémentaire d'information, exploitable et accessible par le SRI. Il peut s'agir d'un texte, une page web, une image…
- **Collection de documents** : la collection de documents aussi appelée fond documentaire ou corpus est l'ensemble des données que le système de recherche d'information exploite (Fuhr, 2005)

- **Pertinence** : la pertinence est une notion centrale en RI et un critère primaire pour l'évaluation des systèmes de recherche d'information. Elle représente le degré de correspondance entre un document et une requête.

  On distingue deux types de pertinence :

  1) **La pertinence système** est l'évaluation par le SRI, de la correspondance entre des documents et une requête.
  2) **La pertinence utilisateur** est l'évaluation par l'utilisateur, de la pertinence vis-à-vis de son besoin en information, des documents retournés par le SRI. Ce type de pertinence est subjectif, car un même document retourné en réponse à une même requête, peut être jugé différemment par deux utilisateurs différents.
     La pertinence utilisateur est dite évolutive car, si à un instant t donné un document est jugé non pertinent, à l'instant t+1, il pourrait être jugé pertinent car la connaissance de l'utilisateur sur le sujet aura évolué.

- **Besoin en information *:*** la notion de besoin en information dans le domaine de la recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types ont été définis (Ingwersen et al. 1994) :

  1) **Besoin vérificatif :** l'utilisateur cherche à vérifier les informations qu'il possède déjà, c'est à dire Il recherche une donnée particulière, et il sait la plupart du temps comment y accéder. Un besoin de type vérificatif est dit stable, il ne change pas au cours de la recherche. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin.
  2) **Besoin thématique connu :** l'utilisateur cherche de nouvelles informations dans le but de clarifier ou de trouver de nouveaux concepts liés à un sujet ou à un domaine connu. Un besoin de ce type peut être stable ou variable et peut aussi s'exprimer de façon incomplète.
  3) **Besoin thématique inconnu :** 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.

5 Le processus de la recherche d'information

Le processus de RI permet de sélectionner les documents les plus proches du besoin en information de l'utilisateur décrit par une requête. Pour cela le SRI compare la représentation

interne de la requête utilisateur aux représentations internes des documents de la collection afin de retrouver les documents les plus pertinents pour la requête.

La réalisation d'un tel système consiste principalement à mettre en œuvre un processus clé "processus en U de la RI" (Croft, 1992). Ce processus est composé de trois principales phases comme le présente la figure 1.1 **:**

- L'indexation,
- L'appariement document-requête,
- Reformulation de la requête.

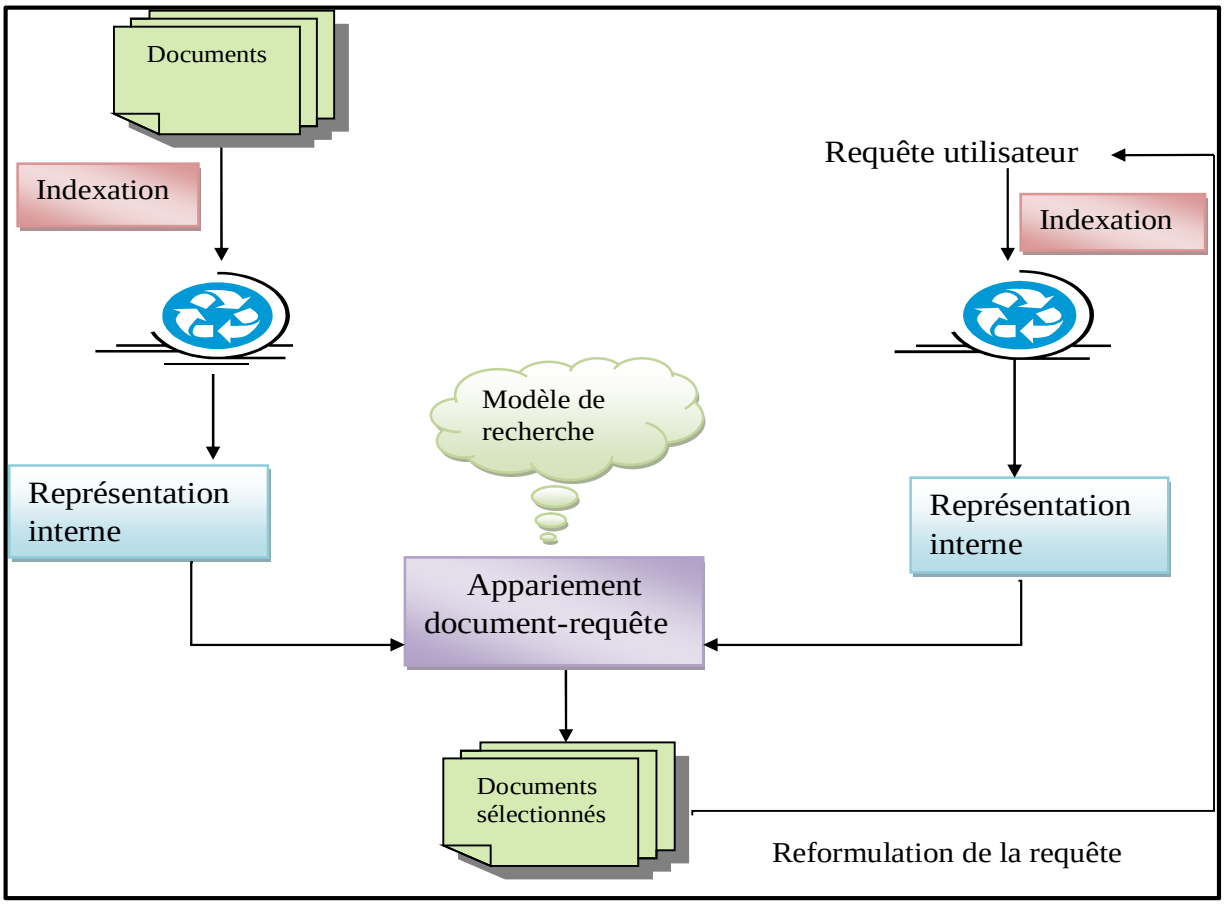


Figure 1.1 : processus de recherche d'information (Croft, 1992)

5.1 Indexation

L'indexation recouvre un ensemble de techniques visant à transformer les documents (ou requêtes) en substituts ou descripteurs capables de représenter leur contenu (Salton et al. 1983). Ces descripteurs forment le langage d'indexation représenté selon une structure souvent basée sur un ensemble de mots clés ou groupes de mots représentant le contenu textuel du document.

Dès lors, l'indexation consiste à sélectionner les termes les plus représentatifs du contenu du document ou d'une requête. Les performances et la qualité de réponse du système dépendent de cette phase.

On distingue différents types d'indexation :

- Indexation manuelle,
- Indexation automatique,
- Indexation hybride.

5.1.1 Indexation manuelle

C'est un indexeur humain qui se charge de définir les descripteurs (mots clés) représentatifs du contenu du document, cette approche est subjective, puisque le choix des termes d'indexation dépend de l'indexeur et de ses connaissances du domaine. Ce type d'indexation est pratiquement inapplicable aux corpus volumineux.

5.1.2 Indexation automatique

L'indexation automatique est un processus complètement automatisé qui se charge d'extraire les termes caractéristiques du document, particulièrement adaptée aux corpus volumineux.

Les techniques existantes se basent sur : l'analyse de texte du document, l'extraction des mots vides qui ne jouent qu'un rôle syntaxique, la suppression des mots qui apparaissent trop souvent et qui n'ont aucun intérêt, la normalisation et la pondération des mots en fonction de leurs apparitions, finalement création de l'index.

5.1.3 Indexation hybride

L'indexation hybride est une combinaison des deux types d'indexation manuelle et automatique, appelée aussi indexation supervisée. L'indexation automatique est d'abord lancée, elle extrait un ensemble de termes descripteurs du document, le choix final des termes d'indexation à partir du vocabulaire fourni est laissé à l'indexeur humain (généralement spécialiste du domaine).

5.2 L'appariement requête-document :

Le processus d'appariement requête-document permet de calculer le degré de pertinence de chaque document de la collection par rapport à une requête, puis retourne l'ensemble des documents les plus pertinents à l'utilisateur. On distingue deux types d'appariement :

- **Appariement exact :** Le résultat retourné est une liste de documents respectant exactement la requête spécifiée. Ces documents ne sont pas ordonnés.
- **Appariement approché :** Le résultat retourné est une liste de documents censés être pertinents pour la requête. Les documents sont triés selon leur degré de pertinence pour la requête.

Les différents types d'appariement dépendent du modèle de recherche utilisé, nous détaillerons les modèles de recherche en section 6.

5.3 Reformulation de la requête

La reformulation de la requête est l'une des méthodes adoptées pour l'adaptation du SRI aux besoins de l'utilisateur. Elle consiste, à partir d'une requête initiale formulée par l'utilisateur, des résultats initiaux fournis en réponse à cette requête et des jugements de pertinence utilisateur sur ces résultats, à construire une nouvelle requête qui répond mieux à son besoin informationnel en rajoutant de nouveaux termes et/ ou en supprimant des termes inutiles ; selon deux approches : manuelle ou automatique

- **Reformulation manuelle :** L'utilisateur soumet lui-même une nouvelle requête, en ajoutant ou en supprimant des termes de la requête initiale.
- **Reformulation automatique :** La nouvelle requête peut être obtenue lorsque le SRI utilise les n-tops documents les plus pertinents pour extraire les termes importants dans le but de reformuler la requête initiale (Kacem Sahraoui, 2017).

6 Les modèles de Recherche d'Information

Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche et un cadre théorique pour la modélisation de la mesure de pertinence.

Comme le montre la figure 1.2, on peut distinguer trois grandes classes de modèle, regroupés selon les fondements mathématiques sur lesquels ils se basent (Baeza-Yates et al. 1999).

- **Les modèles ensemblistes :** sont basés sur la théorie des ensembles. Ils englobent le modèle booléen et le modèle booléen étendu.
- **Les modèles algébriques :** sont basés sur l'algébrique. Ils comprennent le modèle vectoriel et le modèle connexionniste
- **Les modèles probabilistes :** sont basés sur la théorie des probabilités. Ils comprennent le modèle probabiliste et le modèle de langue.

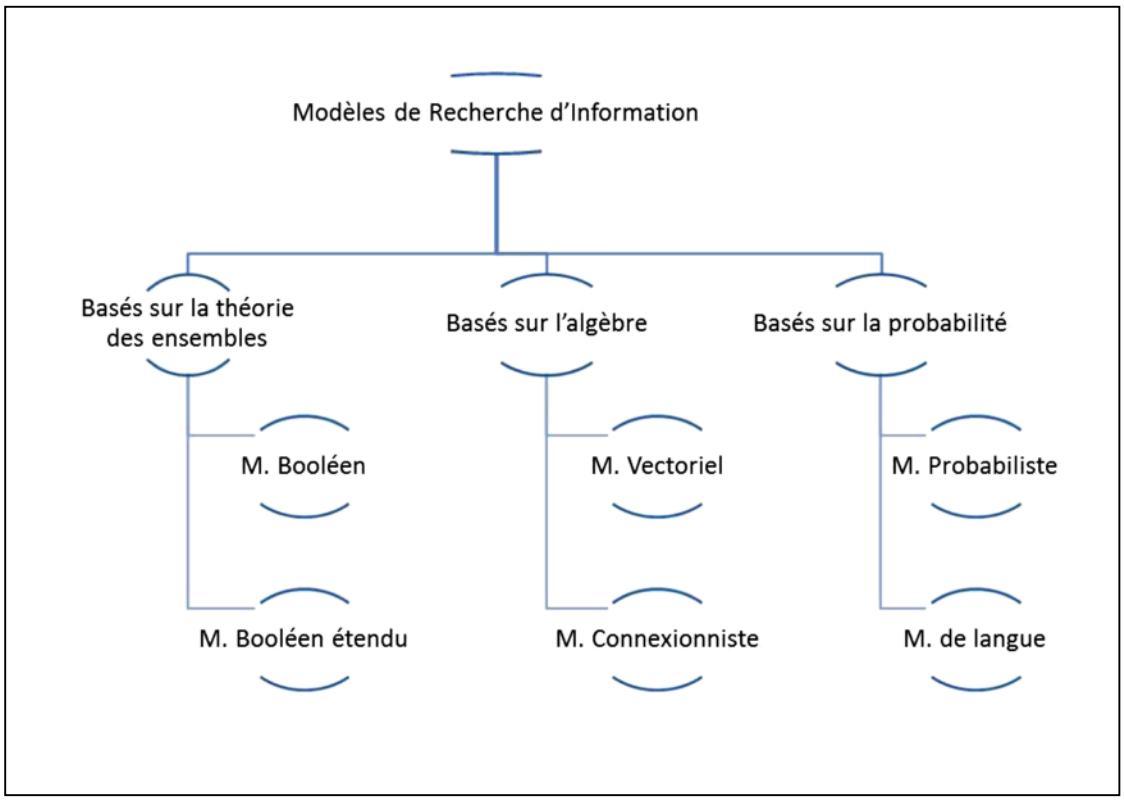


Figure 1. 2 : Modèles de RI (Baeza-Yates et al. 1999).

Nous présentons très brièvement dans ce qui suit, les principaux modèles de la RI :

6.1 Le modèle booléen

Le modèle booléen est le premier qui s'est imposé dans le monde de la recherche d'information. Il repose sur la théorie des ensembles et l'algèbre de bool (Salton, 1971). Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté comme une conjonction logique de termes (non pondérés) $T = \{t1, \dots, tn\}$, par exemple, $d = t1 \wedge t2 \wedge \dots \wedge tn$.

Une requête q est une expression logique quelconque de termes assemblés par les opérateursET $(\wedge), OU$ $(\vee)$ $etNON$ $(\neg)$.
Par exemple: $q = (t1 \wedge t2) \vee (t3 \wedge \neg t4)$ (Daoud, 2009).

Le processus de recherche mis en œuvre, consiste à effectuer des opérations sur les ensembles de documents définis par la présence et l'absence de termes d'indexation afin de réaliser un appariement exact {0, 1} de la requête comme suit :

$$\begin{cases} RSV(d,q) = 1 & si(q \in d) \\ RSV(d,q) = 0 & sinon \end{cases} \quad \textbf{(1.1)}$$

Donc il ne restitue que les documents répondant exactement à la requête ce qu'on appelle l'ensemble idéal. Pour remédier à cette limitation, une extension de ce modèle a été effectuée dans le modèle booléen étendu par l'intégration des poids dans l'expression d la requête et des documents (Salton, 1983).

6.2 Le modèle vectoriel

Ce modèle introduit par Gérard Salton (Salton, 1983) est basé sur la théorie de l'algèbre et plus précisément sur le calcul vectoriel (Hammache., 2013).

Dans ce modèle, les requêtes et le contenu des documents sont représentés dans un espace vectoriel construit par les termes d'indexation.

Un document dj est représenté par un vecteur de poids w_{ij} de dimension n, dans l'espace vectoriel composé de tous les termes d'indexation : $dj = (w_{1j}, w_{2j}, \dots, w_{nj})$.

Une requête q est aussi représentée par un vecteur de poids wiq défini dans le même espace vectoriel que le document : $\mathrm{q} = (w_{1q}, w_{2q}, \dots, w_{nq})$ tel que wiq, w_{ij} représentent respectivement le poids du terme ti dans la requête q et le poids terme ti dans le document dj.

Formellement, la pertinence du document *dj* pour la requête *q* est exprimée par l'une des mesures suivantes :

Mesures	**Formules**
Le produit scalaire	$\mathrm{Sim}(Q, dj) = \sum_{k=1}^{n} (wiQ * wij)$
La mesure de cosinus	$\mathrm{Sim}(Q, dj) = \frac{\sum_{k=1}^{n} (wiQ * wij)}{\left(\sum_{k=1}^{n} wiQ^2\right)^{1/2} * \left(\sum_{k=1}^{n} wij^2\right)^{1/2}}$
La mesure de Dice	$\mathrm{Sim}(Q, dj) = \frac{2 * \sum_{k=1}^{n} wiQ * wij}{\sum_{k=1}^{n} wiQ^2 + \sum_{k=1}^{n} wij^2}$
La mesure de Jaccard	$\mathrm{Sim}(Q, dj) = \frac{\sum_{k=1}^{n} wiQ * wij}{\sum_{k=1}^{n} wiQ^2 + \sum_{k=1}^{n} wij^2 - \sum_{k=1}^{n} wiQ * wij}$

Tableau 1.1: les mesures de similarité utilisée dans le modèle vectoriel

6.3 Modèle probabiliste

Ce modèle est basé sur des calculs probabilistes pour estimer la pertinence d'un document pour une requête. Dans ce modèle, documents et requête sont représentés par des vecteurs de poids dans l'espace vectoriel des termes d'index (Hammache., 2013).
La pertinence d'un document pour une requête est calculée comme suit :

$$RSV(q, dj) = \frac{P(per/d)}{P(\overline{per}/d)} \quad \textbf{(1.2)}$$

Où :

$P(per/d)$: est la probabilité qu'un document *dj* soit pertinent pour la requête *q* ;

$P(\overline{per}/d)$: est la probabilité qu'un document *dj* soit non pertinent pour la requête *q*.

7 Evaluation des SRI

L'évaluation constitue une étape importante dans la mise en œuvre d'un SRI. Elle permet de mesurer les caractéristiques du système en termes de qualité, de service et de facilité d'utilisation.

Les premiers protocoles adoptés en RI sont initiés par Cleverdon (Cleverdon, 1967) dans le cadre du projet Cranfield. Ils se basent sur une approche de type laboratoire. Cette approche constitue le cadre de référence dans lequel s'inscrivent les expérimentations et la validation des systèmes classiques (Hannech, 2018).

7.1 Mesures de l'évaluation :

L'évaluation est une composante critique et intégrale d'un SRI, elle peut être abordée selon deux angles : l'efficience qui mesure la qualité de recherche en terme de critères liés au rendement (temps de réponse et/ou quantité de ressources utilisés) et l'efficacité qui mesure les performances des SRIs en comparant les documents retournés par le système avec les documents que l'utilisateur souhaite retrouver.

Les mesures communément utilisées pour évaluer un SRI sont essentiellement basées sur le rappel et la précision comme le montre la figure (1.3) :

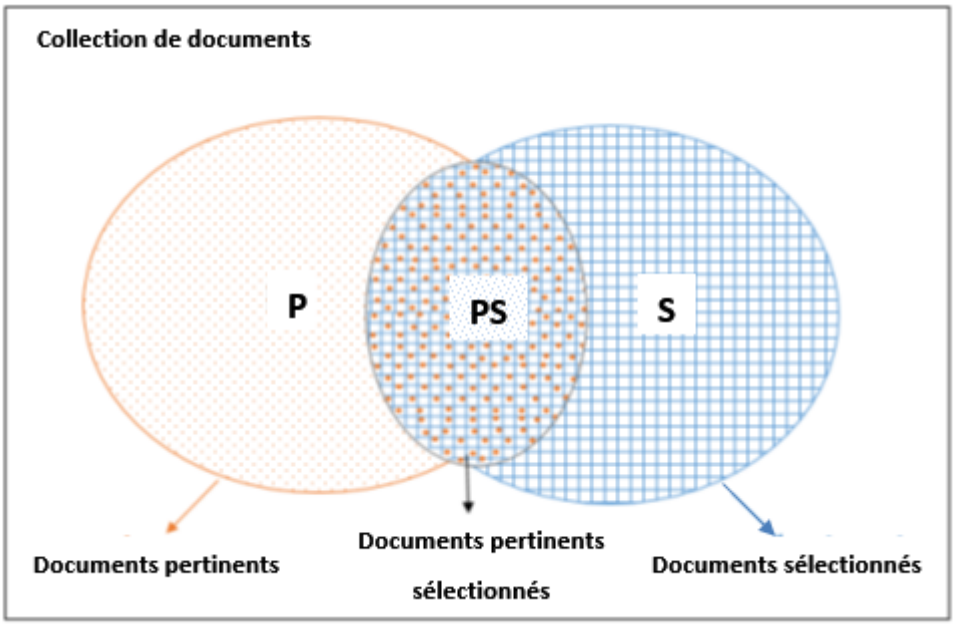


Figure 1.3 : le Rappel et la Précision (Kacem Sahraoui, 2017).

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

$$P = \frac{nombre\ de\ documents\ pertinents\ retournés}{nombre\ total\ de\ documents\ retournés\ par\ le\ systéme} \quad \textbf{(1.3)}$$

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

$$R = \frac{nombre\ de\ documents\ pertinents\ retournés}{nombre\ de\ documents\ pertinents\ total} \quad \textbf{(1.4)}$$

- **La F-mesure** permet d'agréger le rappel et la précisons dans une mesure unique et elle est définie comme la moyenne harmonique pondérée du rappel et de la précision, soit :

$$\mathrm{F} = \frac{1}{\left(\alpha * \frac{1}{\mathrm{P}}\right) + (1-\alpha) * \frac{1}{\mathrm{R}}} \quad \textbf{(1.5)}$$

$\alpha \in [0..1]$ est un paramètre qui permet de pondérer le rappel ou la précision.

8 De la RI classique vers la RI adaptative

La recherche d'information classique est principalement basée sur l'appariement requête-document, pour qu'un document soit retourné à l'utilisateur il doit contenir une partie ou la totalité des mots formulés dans la requête soumise par ce dernier. En effet, l'appariement requête-document dans les SRI classiques considèrent que les termes exacts et non ceux similaires. De ce fait, les performances d'un SRI ne seraient plus uniquement dépendantes de l'indexation des documents et de l'appariement requête-document mais aussi de sa capacité à prendre en compte les besoins de l'utilisateur. De ce constat est apparu un nouvel axe de recherche, celui de la RI adaptative.

La RI adaptative présente un ensemble de techniques ayant pour objectif de permettre la reformulation des requêtes dans un but d'adaptation des résultats aux besoins de l'utilisateur.

Elle tente d'utiliser les informations extraites des interactions de l'utilisateur avec le système pour améliorer la performance de la recherche. Deux principales classes de techniques ont été développées en RI adaptative : les techniques de reformulation de requête et de désambiguïsation du sens des mots de la requête (Daoud M. , 2009).

Cependant, les systèmes de RI adaptative présentent des limitations qui sont principalement liées à la représentation limitée du contexte de l'utilisateur, ce qui a conduit à l'émergence de la RI personnalisée qui tient en compte l'utilisateur en tant que composante principale dans le processus de la recherche (Hadjouni Krir, 2012).

9 Facteurs d'émergence de la RI personnalisée

Une problématique cruciale avec la RI classique et la RI adaptative est l'écart considérable qui peut exister entre les univers de représentation utilisés pour interpréter d'une part le besoin informationnel des utilisateurs, et d'autre part la collection de documents disponible pour la recherche.

Les facteurs d'émergence de la RI personnalisée sont principalement liés à la prolifération des ressources d'information hétérogènes, la diversité et l'ambigüité des besoins en information des utilisateurs ainsi que la non prise en compte de l'utilisateur par le SRI.

9.1 Faiblesse dans la représentation de l'information et dans la correspondance « requête-document »

Le phénomène de polysémie qui se manifeste lorsqu'un mot dans la requête ou un terme d'indexation a plusieurs significations et qui est connu aussi sous le nom d'ambigüité peut induire à des résultats non pertinents.

Cette dégradation est due à l'insuffisance de l'appariement sur lequel se base la recherche classique pour sélectionner les documents, cette recherche se base uniquement sur la ressemblance exacte ou lexicale entre les mots. De ce fait une autre limitation peut être soulevée, tel que l'absence de relations sémantiques entre les mots qui réduit l'accès à l'information pertinente.

9.2 Le manque d'expertise de l'utilisateur

Les requêtes sont souvent courtes et ambiguës, en particulier lorsque cet utilisateur n'a pas assez de connaissances sur le domaine de sa recherche, ou il a du mal à traduire ses besoins sous la forme de mots clés. Ceci est connu sous le problème de l'inadéquation des besoins réels de l'utilisateur avec sa requête.

9.3 Volume de l'information

Avec L'augmentation exponentielle des données dans les systèmes d'information, il devient de plus en plus difficile pour les utilisateurs de retrouver les informations qui correspondent précisément à leurs besoins, par conséquent l'utilisateur se retrouve face au problème de la surcharge cognitive qui le désoriente, et dans laquelle il ne sait plus quel chemin suivre lors de la navigation pour trouver des informations pertinentes.

9.4 La non-reconnaissance de l'utilisateur par le système et l'absence de son contexte de recherche

Par-dessus toutes ces contraintes citées, les systèmes non personnalisés ne permettent pas de reconnaitre les utilisateurs donc ils retournent les mêmes résultats pour la même requête envoyée par différents utilisateurs sans tenir compte de leurs contextes spécifiques.

10 La Recherche d'information personnalisée

La RI personnalisée est une discipline apparue dans le but d'améliorer la qualité des interactions homme-machine par inférence et prédiction des buts, préférence et contexte des utilisateurs à partir de faits observés (Hannech, 2018).

L'objectif fondamental de la RI personnalisée est d'exploiter des informations concernant l'utilisateur, en plus de la requête donnée, pour sélectionner le contenu correspondant aux besoins spécifiques de l'utilisateur. Cependant, la RI personnalisée vise à améliorer les problèmes de surcharge cognitive de systèmes de RI classiques en intégrant le profil de l'utilisateur dans le processus de recherche.

10.1 Le système de recherche d'information personnalisée

Un système de recherche d'information personnalisé (SRIP) est un système qui intègre l'utilisateur, en tant que structure informationnelle, tout au long de la chaîne d'accès à l'information.

Le but fondamental d'un SRI personnalisée est d'offrir des moyens permettant de retourner les informations pertinentes relatives à un besoin en information d'un utilisateur à travers une collection de documents. Ce système doit contenir :

- Des structures permettant de représenter l'utilisateur, ces structures traduisent essentiellement les centres d'intérêt, les préférences et contexte de l'utilisateur ;
- Des techniques pour collecter des informations descriptives de l'utilisateur ;
- Un processus d'accès à l'information intégrant les structures descriptives de l'utilisateur ainsi un mécanisme d'évolution de ces structures (Lechani Tamine, 2005).

10.2 Architecture d'un système de RI personnalisée

L'architecture d'un système de RI personnalisée est centrée autour de l'utilisateur et met en évidence :

1. Un gestionnaire de profil pour représenter, construire et faire évoluer les profils des utilisateurs.
2. Les étapes du cycle de vie de la requête où l'on intègre le profil utilisateur dans :
   a) La phase de reformulation de la requête afin de mieux cibler le contexte de la recherche de l'utilisateur.
   b) La phase de réduction de l'espace de recherche pour restreindre l'espace de recherche aux documents qui ciblent les besoins de l'utilisateur.
   c) La phase d'appariement pour calculer la pertinence des documents en fonction des caractéristiques spécifiques de l'utilisateur.
   d) La phase de présentation des résultats pour restituer les informations selon le contexte et les préférences de l'utilisateur (Hadef, 2014).

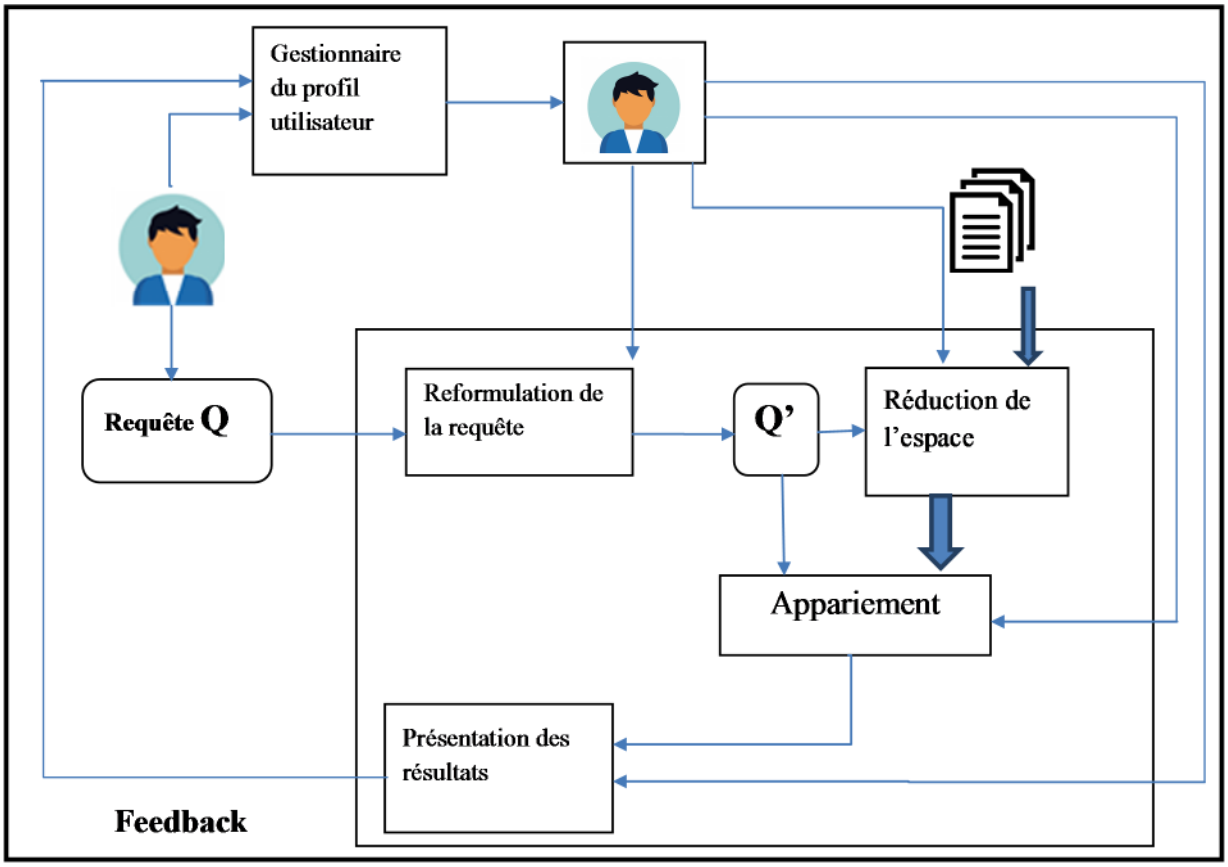


Figure1.4 : Un système de RI personnalisé (Zemirli, 2008).

11 Evaluation des systèmes de RI personnalisés

L'évaluation d'un SRI personnalisé consiste à mesurer ses performances et à estimer son aptitude à répondre aux besoins de l'utilisateur. La performance est estimée en comparant les réponses du SRI personnalisée fournies à un utilisateur pour une requête à celles que le même utilisateur voudrait idéalement avoir.

Les premières tentatives effectuées dans le cadre de l'évaluation des SRI en présence de contexte ont été proposé dans TREC, à travers les tâches Interactive et HARD.

- La tâche interactive de TREC consiste à étudier les interactions de l'utilisateur avec le système
- La tâche HARD de TREC permet aux systèmes d'atteindre une grande précision de recherche en intégrant le contexte de recherche ou le contexte de l'utilisateur dans le processus de recherche (Daoud M. , 2009). Cette proposition est effectuée en vue

d'améliorer la performance du système pour des requêtes difficiles, en particulier les requêtes courtes et ambigües.

En effet, ces tâches ne permettent pas d'évaluer un SRI personnalisée intégrant des aspects contextuels plus larges, tel qu'un profil utilisateur à centres d'intérêts multiples. Ceci a conduit à l'émergence des approches d'évaluation fondées sur l'utilisation des contextes de recherche simulés ou des contextes réelles en intégrant le profil utilisateur comme étant une composante principale de la collection de test.

11.1 Approche d'évaluation par simulation de contexte

L'évaluation d'un SRI par simulation de contextes consiste à simuler des utilisateurs et leurs interactions avec le système à travers une ou plusieurs requêtes liées à un centre d'intérêt de l'utilisateur et exploiter des jugements de pertinence préalablement donnés ou considérés pertinents s'ils sont classifiés dans le domaine d'intérêt simulé (Daoud, 2009).

11.2 Approche d'évaluation par utilisation de contextes réels

L'évaluation par utilisation de contextes réels fait appel à de vrais utilisateurs pour une étude de cas basée sur des contextes de recherche et des interactions réelles de l'utilisateur avec le système (Hannech, 2018). Cette approche permet de prendre en compte la nature dynamique du besoin en information et considérer des jugements de pertinence selon la perception de pertinence de l'utilisateur qui a émis la requête dans des situations de recherche réelles et bien spécifiées (Daoud, 2009).

Deux types d'évaluations peuvent être adoptés :

- Le premier type consiste à reformuler des requêtes par l'utilisateur afin de définir celles qui sont reliées à un même besoin en information définissant une session de recherche.
- Le deuxième type consiste à utiliser une interface de recherche (telle que Google API) pour formuler des requêtes selon des besoins spécifiques. Dans ce cas, les documents pertinents sont extraits par une analyse du comportement implicite des utilisateurs en vue d'extraire des fichiers logs, tels que l'analyse des clics, la considération du temps passé sur une page, etc… (Hannech, 2018).

L'évaluation par utilisation des contextes réels est efficace en termes d'utilité et d'utilisabilité réelles, car l'utilisateur est en interaction directe avec le système ce qui va lui permettre d'obtenir une information la plus juste possible pour son besoin en information. Mais cette

approche d'évaluation reste difficile à réaliser à cause du facteur temps qui est énormément demandé.

L'objectif des approches dévaluation d'un SRIP en intégrant le profil utilisateur dans le processus d'accès à l'information est alors de mesurer l'adéquation des profils utilisateur construit par le système avec les centres d'intérêts effectifs de l'utilisateur, ainsi que d'estimer la performance du système en fonction du comportement d'interaction de l'utilisateur et de la dynamique de besoin en information.

12 Conclusion

Nous avons présenté au cours de ce chapitre les concepts de base de la RI classique, ainsi que le processus et les modèle de recherche d'information. Ensuite nous avons met l'accent sur les facteurs d'émergence de la RI personnalisée où nous avons abordé la définition de la RI personnalisée et celle d'un SRIP, enfin nous avons terminé avec les approches d'évaluation d'un SRIP.

Avec l'augmentation et l'évolution considérable de données, les résultats retournés par les SRI classiques sont devenus de moins en moins satisfaisants. Cependant, la satisfaction des besoins en information d'un utilisateur demeure un but très important à atteindre pour les SRI actuels. Dans ce contexte la prise en compte de profil de l'utilisateur pour la personnalisation de la recherche d'information a été adoptée.

Dans le chapitre suivant nous présentons la modélisation du profil utilisateur et son évolution au cours du temps.

Chapitre 2 : Modélisation et évolution du profil utilisateur

1 Introduction

L'augmentation et l'évolution considérables des données soulèvent d'importants problèmes pour les utilisateurs notamment pour l'accès aux documents les plus pertinents à leurs requêtes de recherche. Afin de cibler la recherche à des besoins en information de l'utilisateur et améliorer la pertinence de recherche, il est devenu indispensable d'intégrer le profil de l'utilisateur dans le processus de recherche. Cependant, La prise en compte de l'utilisateur dans le processus de recherche est nécessite la modélisation de son profil.

La personnalisation de l'information engendre le problème de l'évolution de profil utilisateur au cours du temps. Les besoins de l'utilisateur évoluent au fil du temps et peuvent s'éloigner de ses intérêts antérieurs stockés dans son profil. De ce fait, le profil de l'utilisateur peut être mal exploité pour extraire ou inférer ses nouveaux besoins en information.

Dans ce deuxième chapitre, nous traitons de façon générale la modélisation du profil utilisateur. Nous présentons d'abord la notion de profil utilisateur et son intégration dans le processus de recherche. Ensuite nous présentons les approches de représentation et les techniques de construction du profil utilisateur. La prise en compte de l'évolution du profil utilisateur sera également abordée.

2 Notion de contexte

D'après (Dey, 2000) le contexte est lié aux informations pouvant être utilisées pour caractériser la situation des entités (par exemple : une personne, un lieu ou un objet) et qui sont jugées pertinentes pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et l'application eux-mêmes.

Un contexte multidimensionnel a également été défini par (Fuhr, 2000). Cette définition ajoute de nouvelles caractéristiques liées d'une part à l'aspect temporel du besoin en information et d'autre part au type de recherche demandé.

Les trois principales dimensions retenues pour le contexte sont : sociale, application et temps.

- **La dimension sociale** définit l'appartenance possible de l'utilisateur : individuel, groupe ou communauté.
- **La dimension application** définit le but de la tâche accomplie.
- **La dimension temps** permet de définir le contexte temporel du besoin : temps passé (batch), intérêt à court terme ou intérêt à long terme.

  Le contexte à court terme ou courant est associé aux besoins et préférences de l'utilisateur lors d'une session de recherche, alors que le contexte à long terme traduit les besoins et les préférences persistants de l'utilisateur tout au long de diverses sessions de recherche.

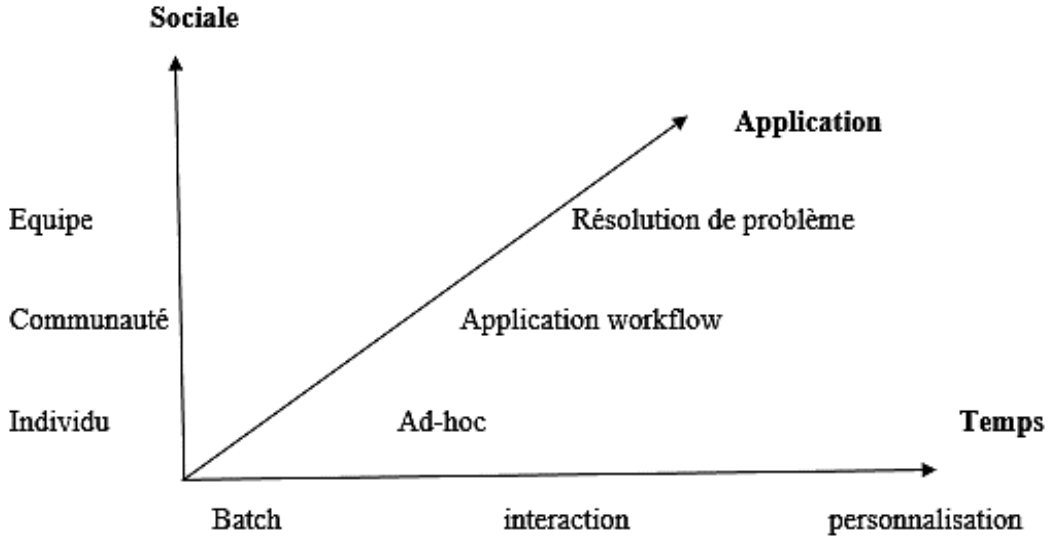


Figure 2.1 : Dimensions du contexte multidimensionnel (Fuhr, 2000).

D'après ces définitions on note que le contexte est souvent lié à l'utilisateur en caractérisant sa situation et en représentant ses informations, en effet le contexte ne dépend non seulement de la collection des données personnelles mais aussi en la manière d'intégrer ces données dans le profil de l'utilisateur.

3 Profil utilisateur

Le concept de profil utilisateur a été introduit pour l'accès à l'information en premier lieu dans les travaux de filtrage d'information pour décrire une structure représentative de l'utilisateur, en l'occurrence ses centres d'intérêts, puis ré-exploitée en recherche d'information personnalisée pour former les composantes du contexte dépendantes de l'utilisateur (Kobsa, 2001).

Dans (Myaeng et al, 1986) il était reconnu que les systèmes de vérification de l'information pouvaient être personnalisés pour les utilisateurs au moyen de profils. Au cours des dernières décennies, de nombreuses recherches ont été investies dans le domaine des profils d'utilisateurs. Ces derniers sont utilisés dans le but de délivrer à l'utilisateur une information pertinente et appropriée à ses préférences et à ses centres d'intérêts.

3.1 Définition du profil utilisateur

Un profil c'est toute structure qui permet de modéliser et de stocker des informations relatives à l'utilisateur.

Dans (Speretta., 2004), l'auteur divise les profils utilisateurs en deux groupes : les profils qui représentent les préférences de l'utilisateur et ceux qui représentent ses intérêts. Les profils représentant les intérêts de l'utilisateur sont plus répandus que ceux qui représentent les préférences (Thanh Trung, 2008).

Le profil utilisateur peut contenir :

- **Les données personnelles** telles que, son identité (nom, prénom, etc.), ses données démographiques (âge, genre, adresse, situation familiale, etc.) et ses données Professionnelles.
- **L'historique** qui regroupe l'ensemble des informations collectées sur son comportement, de façon explicite ou implicite (par exemple, le nombre de clics qu'il a effectués sur le lien d'une page ou le nombre de requêtes qu'il a émises)

- **Les annotations** associées par l'utilisateur aux documents qui peuvent être sous différentes formes (par exemple, les annotations textuelles, les signets qui mémorisent les liens vers d'autres documents, les tags qui sont les références sous forme d'un ensemble de mots-clés choisis librement par l'utilisateur pour identifier le document visité…).
- **Les préférences** qui désignent les caractéristiques de l'utilisateur en termes de présentations ou d'interactions avec les informations (par exemple, des couleurs et/ou les styles de présentation des pages web préférés, etc.).
- **Les intérêts** qui expriment son domaine d'expertise ou son périmètre d'exploration. Ils sont généralement définis par un ensemble de mots clés ou concepts, le plus souvent pondérés.

Les données personnelles sont relativement stables dans le temps et ne demandent pas a priori de mise à jour automatique, alors que les préférences et les intérêts tendent à changer au fil du temps (ON-AT, 2017).

4 Intégration du profil utilisateur dans le processus de recherche d'information

La personnalisation du processus d'accès à l'information consiste à intégrer le profil utilisateur dans le processus de recherche d'information. Le but fondamental des modèles d'accès personnalisée à l'information est de restituer, en haut de la liste des résultats, des documents qui intéressent l'utilisateur dans sa recherche, en d'autres termes qui semblent les plus similaires à son profil (Daoud, 2009).

Nous présentons dans ce qui suit les 3 principales techniques d'intégration du profil utilisateur dans le processus de recherche d'information (ON-AT, 2017) :

- **Modification (la reformulation de requête)** consiste à introduire dans la requête de l'utilisateur, les termes ou une partie des termes provenant du profil utilisateur. C'est la méthode la plus répandu.

- **La sélection d'information personnalisée (l'appariement requête-document)** consiste à intégrer les informations du profil utilisateur pendant l'étape de l'appariement entre la requête de l'utilisateur et chaque document indexé.

- **Le ré-ordonnancement des résultats (présentation des résultats)** consiste à intégrer les informations du profil utilisateur pour réordonnancer les résultats trouvés après l'étape d'appariement requête-documents. Le principe de base est d'affiner la recherche en ne présentant que les résultats en corrélation avec le contenu du profil.

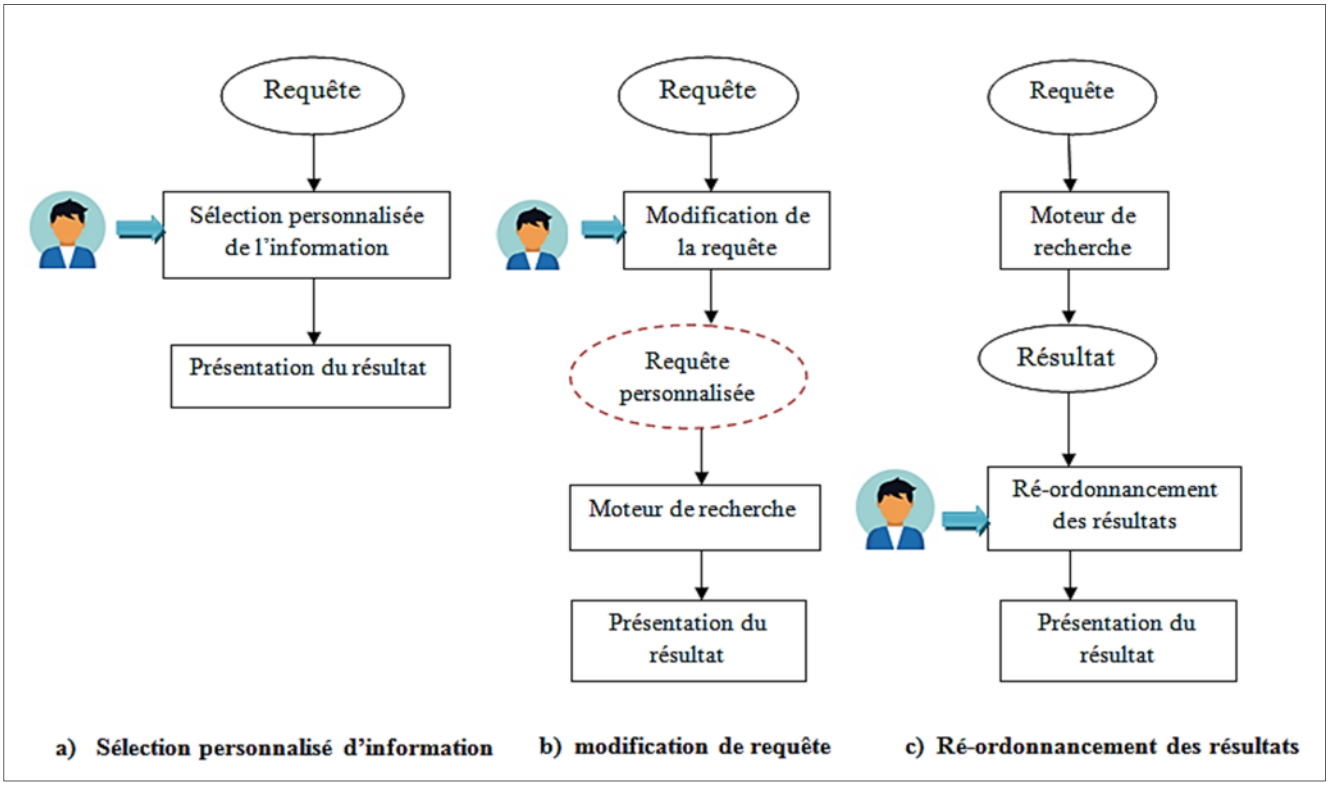


Figure 2.2 : Intégration du profil utilisateur dans le processus de recherche (ON-AT, 2017)

4.1 Intégration du profil utilisateur dans l'appariement requête-document

La fonction classique de calcul du score du document se base sur la requête comme la seule ressource d'information qui représente l'utilisateur. Dans le cadre de la RI personnalisée, l'appariement requête-document consiste à exploiter le profil utilisateur dans la fonction de calcul du score du document vis-à-vis d'une requête. Tel que le calcul du score du document est une fonction qui assigne au document un score de pertinence en fonction non seulement de la requête mais aussi du profil utilisateur U , $RSV(Q,D) \rightarrow RSV(Q,D,U)$.

Les travaux qui s'inscrivent dans ce cadre sont les modèles probabilistes d'analyse sémantique latente (PLSA). Ces modèles se basent sur le principe de corrélation sémantique entre des ensembles d'objets (Hofmann., 1999), leur but c'est d'identifier l'intention de recherche de l'utilisateur à partir de son comportement. En RI, l'intégration du profil dans ces

modèles est basé sur l'apprentissage d'une fonction de calcul de pertinence d'un ensemble de pages web, étant donnée une requête et un modèle utilisateur.

Dans ces travaux, le profil est construit implicitement et le modèle contient la requête soumise et les documents (pages web) sélectionnés, nommés clicked documents, pour chaque requête ainsi que l'utilisateur. Où, l'utilisateur $u \in U = \{u1, u2, \dots, un\}$, la requête $q \in Q = \{q1, q2, \dots, qm\}$ et les pages web associées $p \in P = \{p1, p2, \dots, pl\}$. Les relations sont associées aux variables latentes $z \in Z = \{z1, z2, \dots, zm\}$.

Le traitement de ces données s'effectue de deux manières. Pour une requête déjà soumise, le système doit retrouver les pages fréquemment sélectionnées dans le profil de l'utilisateur. Par contre, si c'est une nouvelle requête, le problème revient à calculer un degré de pertinence d'une page p' non existante dans le profil de l'utilisateur et qui a une forte probabilité d'être sélectionnée (Daoud, 2009).

Le système se base sur un modèle de probabilité qui calcule la pertinence d'une page p sachant l'utilisateur u et la requête q comme suit :

$$p(p|u,q) = \frac{\sum_{z \in Z} n(u,q,p) P(z|u,q,p)}{\sum_{p'} \sum_{z \in Z} n(u,q,p') P(z|u,q,p')} \qquad \textbf{(2.1)}$$

Où, $\mathrm{n(u, q, p)}$ est le nombre de fois que l'utilisateur u sélectionne la page web p pour la requête q. Puis, le système ordonne les résultats sur la base des valeurs de ces probabilités et retourne les pages web ayant les valeurs les plus élevées.

4.2 Intégration du profil utilisateur dans la phase de reformulation de la requête

Dans le cadre de la personnalisation, les données utilisées sont principalement relatives à l'utilisateur. Il s'agit alors d'un enrichissement de la requête qui consiste en l'ajout et/ou le retrait de termes de la requête initiale.

Le but fondamental de la reformulation de requêtes par utilisation de profil consiste à cibler la recherche des documents pertinents par augmentation de la requête par des termes issus du profil utilisateur afin de mieux répondre au besoin en information de l'utilisateur (Daoud, 2009).

Dans l'approche présentée dans le système de recherche ARCH (Sieg et al, 2004), le profil de l'utilisateur se compose de plusieurs vecteurs de termes pondérés.

Le fonctionnement du système est le suivant :

- Il collecte implicitement un ensemble de documents pertinents en prenant en compte plusieurs facteurs : fréquence de visite d'une page, le temps utilisé pour consulter la page etc…,
- Puis il utilise un algorithme de clustering pour regrouper ces documents dans des catégories différentes afin de calculer les vecteurs centroïdes de ces catégories. Chaque vecteur centroïde représente un profil individuel (un domaine d'intérêt de l'utilisateur).
- Chaque fois qu'un utilisateur soumet une requête $q1$ au système, le contenu de cette requête est comparé avec ces vecteurs pour trouver les vecteurs les plus similaires avec la requête.
- Le système va également comparer la requête avec les concepts dans une hiérarchie de concepts qui représente les domaines de connaissances pour trouver les concepts les plus similaires avec cette requête. Ces concepts sont aussi représentés par des vecteurs de termes pondérés.
- Enfin, les vecteurs sélectionnés sont comparés avec les concepts sélectionnés. La requête $q1$ est reformulée en utilisant la méthode de Rocchio :

$$q_2 = \alpha . q_1 + \beta . \sum T_{sel} - \gamma . \sum T_{dsel} \qquad \textbf{(2.2)}$$

Où :T_{sel}les concepts les plus similaires avec les vecteurs profils et T_{dsel}les concepts les plus différents avec ces vecteurs. Les facteurs α, β, γ sont des poids associés respectivement à la requête originale, au concept pertinent et au concept non pertinent.

Après cette étape, la requête reformulée $q2$ sera utilisée au lieu de la requête originale $q1$ (Thanh Trung, 2008).

4.3 Intégration du profil utilisateur dans le ré-ordonnancement des résultats

La personnalisation à ce stade du processus de recherche offre une solution en réordonnant les résultats pour ne présenter à l'utilisateur que les documents pertinents en réponse à son besoin en information. Le système de RI personnalisée envoie une requête à un moteur de

recherche, reçoit des résultats et puis retrie les résultats selon leurs similarités avec le profil d'utilisateur.

Dans l'approche présenté dans (Gowan, 2003), le ré-ordonnancement des résultats de recherche consiste à combiner le score de similarité entre le document et le centre d'intérêt courant, représenté par un vecteur de termes pondérés, avec le score d'appariement original du document.

Dans les travaux de (Speretta et al, 2004), les auteurs utilisent un modèle à base d'ontologie pour représenter les profils utilisateurs. Chaque concept dans l'ontologie a un poids représentant l'intérêt de l'utilisateur avec ce concept. Ces poids sont accumulés avec le temps en utilisant l'historique de recherche de l'utilisateur. Les informations prises en compte pour mettre à jour le profil sont les anciennes requêtes et les extraits des résultats sélectionnés par l'utilisateur.

Le fonctionnement du système est le suivant :

- Il utilise un wrapper pour le moteur de recherche Google. Ce wrapper est construit en utilisant le Google API et surveille les actions de l'utilisateur (Requêtes soumise, clics sur les résultats, etc.).
- Chaque fois qu'une requête est soumise, la similarité entre le profil de l'utilisateur et un document retourné est calculée par la formule suivante :

$$similarité(u,d) = \sum_{k=1}^{N} wt_{uk} \times wt_{dk} \qquad \textbf{(2.4)}$$

Où : wt_{uk} est le poids du concept *k* dans le profil de l'utilisateur *u* et wt_{dk} est le poids du concept k dans le document d.

- Les documents sont triés par leur similarité avec le profil utilisateur et ce rang est appelé rang-concept pour le distinguer du rang original de Google.
- Le classement final des documents est calculé en utilisant une combinaison de ces deux rangs :

$$rang_{final} = \alpha \times rang\_concept + (1-\alpha) \times rang\_google \qquad \textbf{(2.5)}$$

5 Modélisations du profil de l'utilisateur

La modélisation de l'utilisateur permet de mieux cibler les données fournies en fonction des intérêts de ce dernier ainsi que de ses besoins en information.

Pour modéliser l'utilisateur il faut définir en premier la structure de son profil qui permet non seulement de stocker les informations le concernant, mais aussi de les exploiter d'une manière optimale. En second, il faut déterminer les techniques de construction et de mise à jour de ce profil.

En général, dans la littérature, les termes « modèle utilisateur » ou « profil utilisateur » signifient la même appellation. Mais d'après (Koch, 2000) une différence existe entre le profil utilisateur et le modèle utilisateur. Il définit le profil de l'utilisateur comme une version simple de modèle utilisateur.

Dans ce travail, nous utiliserons indifféremment les deux termes « modèle utilisateur » et « profil utilisateur ».

5.1 Représentation du profil utilisateur

La représentation de l'utilisateur à travers la notion de profil permet de mieux comprendre certains mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence et au-delà, cibler ses besoins spécifiques dans le but d'améliorer les performances de recherche.

Cependant, on distingue quatre principales approches de représentation : ensembliste, connexionniste, conceptuelle et multidimensionnelle.

5.1.1 La représentation ensembliste

La représentation ensembliste dite aussi vectorielle fait partie des premières représentations du profil utilisateur qui ont été proposées et reste largement utilisée. Le profil est représenté par un ensemble de termes (mots-clés) pondérés où chaque terme représente un centre d'intérêt (Lieberman, 1997) ou par un vecteur de termes pondérés représentant un centre d'intérêt ou par un ensemble de vecteurs de termes pondérés dont chacun représente un centre d'intérêt (Gowan, 2003).

La représentation ensembliste du profil utilisateur apporte l'avantage de la simplicité de mise en œuvre. Néanmoins, même si les modèles de représentation permettent de traduire une multiplicité de centres d'intérêts en utilisant plusieurs vecteurs, cette représentation manque de

structuration et de cohérence (Daoud, 2009) elle ne met en évidence ni la dimension liée au temps marquant l'évolution des profils, ni l'organisation des informations pour hiérarchiser les centres intérêts.

5.1.2 La représentation connexionniste

La représentation connexionniste du profil utilisateur consiste à représenter le profil par un réseau de nœuds pondérés dans lequel chaque nœud est un concept traduisant un centre d'intérêt 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 (Daoud, 2009).

En effet, les problèmes tels que la polysémie des termes et l'incohérence éventuelle entre les centres d'intérêts peuvent être résolus par cette représentation qui apporte de la sémantique au modèle de l'utilisateur. Cette représentation présente néanmoins certaines limitations. En effet, la source de données du réseau sémantique représentant le modèle de l'utilisateur n'est autre que l'historique de recherche de l'utilisateur qui est souvent assez limité (Hadjouni Krir, 2012).

5.1.3 La représentation conceptuelle

Les centres d'intérêt de l'utilisateur sont représentés sous forme de réseau de nœuds conceptuels reliés entre eux en suivant la topologie des liens définis dans les hiérarchies et les ontologies de domaines. Chaque concept décrivant un centre d'intérêt est représenté par un vecteur de termes pondérés où le poids traduit le degré d'intérêt de l'utilisateur pour le concept de profil (Daoud, 2009).

Dans ce type de représentation, la modélisation du profil utilisateur est fondée sur l'élaboration d'une ontologie personnelle. L'ensemble des caractéristiques de l'utilisateur est organisé dans une structure hiérarchique de catégories (concepts), où chaque catégorie représente la connaissance d'un domaine d'intérêt de l'utilisateur.

La représentation conceptuelle est semblable à la représentation précédente, 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'intérêts de l'utilisateur, contrairement à la représentation

connexionniste, où les centres d'intérêts sont représentés par un mot ou un ensemble des mots (Achemoukh, 2018). De même, cette représentation peut être assimilée à une représentation ensembliste. En effet, ceci revient au fait que les domaines y sont généralement représentés par des vecteurs de termes pondérés.

Cette approche peut engendrer certains problèmes d'hétérogénéité et de diversité des intérêts. D'ailleurs, les utilisateurs peuvent avoir différentes perceptions d'un même concept, cela peut engendrer des imprécisions lors de la représentation de l'utilisateur (ON-AT, 2017).

5.1.4 La représentation multidimensionnelle

La représentation multidimensionnelle permet de structurer le profil selon un ensemble de dimensions représentées selon divers formalismes, elle a pour objectif de capturer toutes ces caractéristiques informationnelles. Les utilisateurs sont divers et complexes, les informations les caractérisant ne sont pas factuelles mais multidisciplinaires. Cependant, cette diversité n'est généralement pas fidèlement représentée par les modèles de profil présentés précédemment.

Le travail présenté dans (Amato et al, 1999) est l'un des premiers travaux vers la construction d'un modèle multidimensionnel pour représenter des profils utilisateurs. Cette représentation donne une description globale des utilisateurs en prenant en compte plusieurs dimensions différentes. Les informations concernant les utilisateurs peuvent être classifiées dans cinq catégories différentes, chaque catégorie est une dimension :

- **Catégorie de données personnelles :** informations concernant l'identité de l'utilisateur.
- **Catégorie de données de la source :** informations nécessaires pour décrire les préférences et restrictions sur les documents. Elle est divisée en trois sous catégories (Hadjouni Krir, 2012):
    1) **Le contenu** (des informations sur le sujet du document, la langue, etc.)
    2) **La structure** (format, type, date de publication, dimensions, etc.).
    3) **La source (**provenance, auteurs, éditeurs, etc.).

- **Catégorie de données de livraison :** informations sur la manière de transmettre des résultats à l'utilisateur. Ces informations sont regroupées selon deux sous catégories :
    1) **Le moyen** (mode de livraison par exemple email, fax téléphone, etc.)

2) **Le moment** (contient des informations temporelles sur le moment de livraison, etc.)

- **Catégorie de données de comportement :** enregistrements sur les interactions de l'utilisateur avec le système (URLs des pages visitées, documents lus et pertinence, etc.).
- **Catégorie de données de sécurité :** informations sur les conditions d'accès aux données du profil.

Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Ainsi, les propositions de standards **P3P** pour la sécurisation des profils ont défini des classes distinguant :

- Les attributs démographiques des utilisateurs (identité, données personnelles).
- Les Attributs professionnels (employeur, adresse, type).
- Les attributs de comportement (trace de navigation).

Dans ce même cadre, (Kostadinov, 2007) a poursuivi la classification d'Amato en proposant un ensemble de dimensions ouvertes, pouvant contenir la plupart des informations susceptibles de caractériser l'utilisateur, où il distingue principalement huit dimensions décrites brièvement dans ce qui suit :

- **Les données personnelles** sont la partie statique du profil et contiennent des informations qui décrivent l'utilisateur. Elles sont relativement stables. Elles comprennent l'identité de l'utilisateur (nom, prénom, numéro de sécurité sociale, etc.), les données démographiques (âge, genre, adresse, situation familiale, nombre d'enfants, etc.), les contacts personnels et professionnels de l'utilisateur et d'autres informations comme le numéro de la carte bancaire ou de la carte Vitale.
- **Les centres d'intérêts** expriment le domaine d'expertise de l'utilisateur ou son périmètre d'exploration. Il peut être défini par un ensemble de mots clés (concepts) ou un ensemble d'expressions logiques (requêtes).
- **L'ontologie du domaine** complète la définition du centre d'intérêt en explicitant la sémantique de certains termes ou de certains opérateurs employés par l'utilisateur dans son profil ou dans ses requêtes.
- **La qualité attendue des résultats délivrés :** La qualité est un des facteurs clés de la personnalisation, elle permet d'exprimer des préférences extrinsèques sur l'origine de

l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source.

- **Les données de livraison (customisation) :** Ces données concernent tout ce qui est lié à la présentation des résultats en fonction de la plateforme de l'utilisateur, de la nature et du volume des informations délivrées, des préférences visuelles de l'utilisateur.
- **Les données de sécurité :** les données de sécurité représentent les droits d'accès aux données. Elles peuvent concerner les droits de visualisation ou de modification des données. Elles peuvent représenter également le droit de visualisation des données du profil.
- **Le retour de préférence (feedback) :** le retour de préférence représente l'ensemble des informations collectées sur le comportement de l'utilisateur par exemple le nombre de clics qu'il a effectué sur le lien d'une page.
- **Les informations diverses :** Certaines applications demandent des informations spécifiques ne pouvant être incluses dans aucune des dimensions précédentes comme par exemple la bande passante attribuée au gestionnaire du profil (Zemirli, 2008).

5.2 Construction du profil utilisateur

L'approche de construction dépend fortement de la représentation choisie pour le profil de l'utilisateur. Cependant, La construction du profil utilisateur repose sur deux phases principales : la phase de collecte d'informations sur l'utilisateur à partir de sources d'informations diverses et la phase d'exploitation de ces informations pour construire le profil utilisateur (ON-AT, 2017).

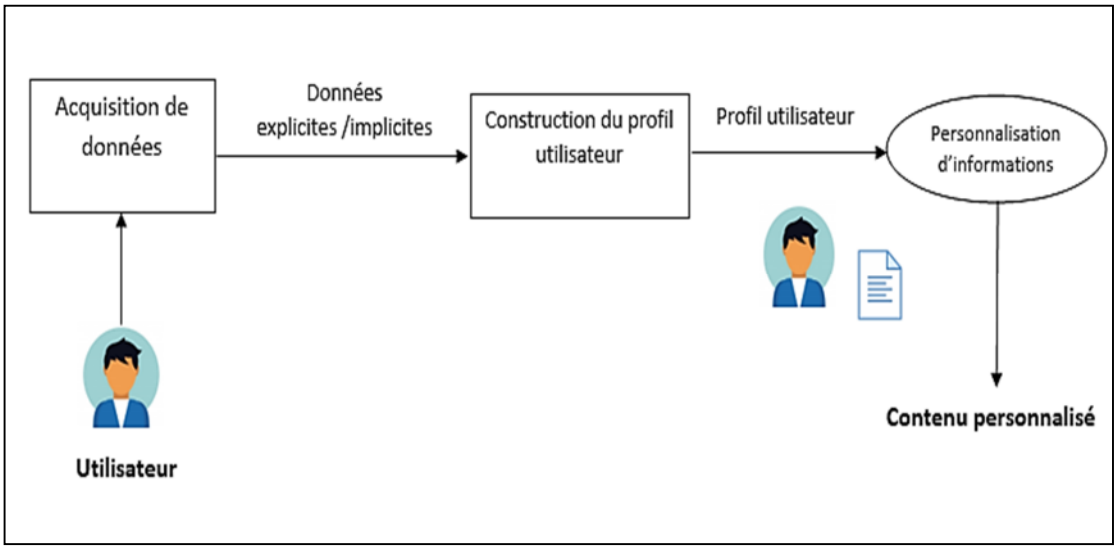


Figure 2.3 : Les phases de construction du profil (ON-AT, 2017).

5.2.1 Acquisition de données

Pour personnaliser la recherche sur le Web, nous devons d'abord tirer parti des informations sources fréquemment utilisées quotidiennement par l'utilisateur. Le profil de l'utilisateur peut être explicitement construit en demandant à l'utilisateur de fournir ses propres informations, ou implicitement en observant les activités de l'utilisateur (Kacem Sahraoui, 2017). Nous détaillons ces deux approches dans ce qui suit :

5.2.1.1 Acquisition explicite

La technique d'acquisition de données explicite est une technique simple, qui consiste à interroger l'utilisateur, pour lui demander des informations personnelles, démographiques, et/ou intérêts…etc (Gauch et al, 2007), Cela peut se faire en demandant à l'utilisateur de remplir un formulaire d'informations personnelles pour construire son profil comme le demande par exemple MyYahoo![1].

L'acquisition explicite repose principalement sur les techniques de feedback explicite de l'utilisateur exprimé lors de son interaction avec le système, par exemple les notes de l'utilisateur sur les documents trouvés, les films regardés ou les produits achetés sur Internet. Ce type d'acquisition est utilisé dans plusieurs systèmes de RI personnalisée, tels que Google

[1] https://my.yahoo.com

personalized search version[2] 1.1 (2004), l'utilisateur peut saisir un ensemble de catégories représentatives de ses centres d'intérêts (Daoud, 2009).

L'avantage des approches explicites est que les profils ainsi construits sont plus précis que ceux obtenus par acquisition implicite. Mais, cette technique peut entrainer un désintéressement et un possible abandon de l'utilisateur lors de la saisie de données, ce qui engendre la réduction de pertinence du système. Les limitations dans l'acquisition des données explicites, ont orienté les travaux vers des techniques d'acquisition de données implicites de l'utilisateur.

5.2.1.2 Acquisition implicite

Dans cette technique il s'agit de ne plus demander à l'utilisateur de fournir explicitement ses données, mais de trouver des sources d'informations permettant d'extraire des connaissances sur l'utilisateur et de construire son profil. En effet, les données utilisées sont issues de l'observation des comportements et des interactions de l'utilisateur avec le système lors de ses recherches, ces données sont généralement : Les dernières pages web consultés, les contenus publiés sur le web ou sur les réseaux sociaux (annotations, commentaires), le contenu des documents consultés et/ou imprimés et/ou sauvegardés,

L'avantage de cette approche, est qu'il n'y a aucun effort requis de l'utilisateur. Cependant avec cette technique d'acquisition de données, on peut faire face au problème d'informations biaisées ou de manque d'informations. En effet, avec les données acquises sans vérification de la part de l'utilisateur, il se peut que ces dernières ne soient pas pertinentes, il est donc nécessaire d'appliquer un prétraitement de données (ON-AT, 2017).

5.2.2 Prétraitement de données

Le prétraitement de données est une étape cruciale et nécessaire pour améliorer la qualité des données et les rendre exploitables.

Les données collectées peuvent contenir de nombreuses inconsistances telles que : des données incomplètes (valeurs manquantes ou agrégées), des données biaisées (présence d'erreurs produites lors des saisies ou de la collection automatique de données), des incohérences (divergence entre attributs) (Tchuente, 2013).

[2]http ://www.google.com/psearch

On utilise plusieurs types de prétraitement selon les données :

- **Nettoyage de données :** dans le cas de données incomplètes, biaisées ou incohérentes, on peut ignorer les données manquantes ou utiliser la valeur moyenne d'un attribut en remplacement ou encore à utiliser la valeur la plus probable (formule bayésienne ou arbre de décision) en remplacement.
- **Discrétisation des données :** convertir des attributs continus vers des attributs nominaux ordinaux.
- **Réduction des données :** peut être appliquée pour obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit (ou presque) les mêmes résultats.
- **Transformation de données :** Pour rendre les données conformes au modèle ou à l'algorithme utilisé, on peut appliquer cette technique qui permet par exemple, de ne conserver qu'un résumé d'un texte au lieu du texte entier, traduction d'un texte d'une langue à une autre, etc.

5.2.3 Techniques de construction

Le processus de construction consiste à organiser et à extraire les éléments qui constituent le profil à partir des données de l'utilisateur collectées lors de l'étape précédente.

La construction s'appuie sur différentes techniques selon la représentation de profil utilisateur. Cependant, On distingue trois principales techniques détaillées dans les paragraphes suivants : l'extraction des termes, l'extraction de réseaux de termes et l'extraction de concepts.

5.2.3.1 Extraction d'ensemble de termes

La construction d'un profil ensembliste se base sur des techniques d'extraction de termes. Généralement, l'extraction de termes comprend les phases de traitements automatisés suivantes : extraction de termes (segmentation), élimination des mots vides, normalisation et pondération.

Par exemple, dans le cadre de cette approche, les termes vont être pondérés pour former des vecteurs de termes représentant les centres d'intérêts. La fonction de pondération appliquée par la majorité des systèmes est issue du schéma TF -IDF. Le nombre de termes extraits est souvent fixé selon un seuil de pondération de sorte que seul les termes dépassant cette valeur contribuent à la construction du profil. Ceci permet d'obtenir des profils plus concis et plus représentatifs des centres d'intérêts de l'utilisateur.

Des systèmes tels que WebMate (Chen et al, 1998) et Alipes (Widyantoro et al, 1999) appliquent cette approche de construction.

5.2.3.2 Extraction de réseaux de termes

Cette technique est similaire à la technique précédente, où les termes sont extraits des documents jugés pertinents par l'utilisateur. Néanmoins, la différence réside dans la représentation des termes qui est sous forme de réseau de nœuds. Pour construire le profil de l'utilisateur, il est nécessaire d'exploiter des relations préexistantes entre les termes et les concepts. Ces relations peuvent se trouver dans des dictionnaires de données tels que WordNet (Miller, 1995). Des systèmes tels que SiteIF (Stefani, 1998) utilisent cette approche de construction.

5.2.3.3 Extraction de concepts

Cette construction se base sur la technique de l'extraction de concepts, qui utilise une taxonomie de concepts de référence comme profil de base.

L'approche de construction présente de manière générale les deux étapes suivantes :

- Identification des concepts et les niveaux de l'ontologie à exploiter
- Extraction des centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie.

Des systèmes tels que ARCH (Sieg et al, 2004) (approche hybride combinant vecteurs de termes et hiérarchie de concepts), le système du projet OBIWAN (Ontology Based Informing Web Agent Navigation) (Pretschner et al, 1999) utilisent cette approche de construction.

5.3 Evolution du profil utilisateur

La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction du profil utilisateur et désigne l'adaptation du profil à la variation des intérêts et aux variations des besoins en information de l'utilisateur au cours du temps (Zemirli, 2008).

Les techniques de collecte d'informations utilisées dans la gestion de l´évolution du profil utilisateur sont relativement dépendantes de la portée temporelle du profil. On distingue le profil à court terme, le profil à long terme ou les deux à la fois (Bennett et al, 2012). Le premier représente les centres d'intérêts liés aux activités de recherche courantes de l'utilisateur. Le second représente les centres d'intérêts persistants de l'utilisateur et issus de son historique de recherche tout entier (Daoud, 2009).

5.3.1 Evolution du profil utilisateur à court terme

Le profil à court terme représente les centres d'intérêts liés aux sessions de recherche courantes de l'utilisateur. Le principal objectif est d'améliorer la précision de recherche d'information en utilisant le profil le plus approprié sans le bruit causé par les centres d'intérêts qui ne sont pas liés au contexte de recherche.

Dans (Gowan, 2003) le profil à court terme est représenté par un vecteur de termes pondérés. Il évolue par agrégation de l'information issue à partir des interactions de recherche courantes de l'utilisateur. En effet, dans un intervalle de temps précis, un vecteur de termes pondérés est créé à partir des documents issus des interactions de recherche de l'utilisateur puis comparé à l'ensemble des centres d'intérêts préalablement appris, selon une mesure de similarité vectorielle. Le centre d'intérêt qui excède un seuil de similarité est combiné avec le vecteur de termes pondérés et permet de définir le profil à court terme.

Dans (Zemirli, 2008) le profil utilisateur à court terme est défini dans une session de recherche par un besoin en information unique. L'évolution du profil dans ce cas nécessite des mécanismes de délimitation des sessions de recherche, où une session est définie par un ensemble de requêtes liées à un même besoin en information.

5.3.2 Évolution du profil à long terme

Le profil de l'utilisateur à long terme modélise des centres d'intérêts récurrents de l'utilisateur. Son évolution consiste à ajouter, ou modifier un profil préalablement appris selon des changements éventuels des centres d'intérêts de l'utilisateur au cours des sessions de recherche (Tamine et al, 2007), (Daoud, 2009). Ce profil peut être exploitable dans le but d'améliorer la recherche pour toute requête soumise par l'utilisateur (Achemoukh, 2018).

Certains auteurs tels que (Tan et al, 2006) ont étudié la représentation des intérêts des utilisateurs à long terme sur la base de requêtes, documents et clics. Ils ont considéré différents historiques de recherche et constaté que l'historique récente était la plus importante pour les nouvelles requêtes, mais pour les requêtes récurrentes l'historique à plus long terme était plus significatif.

L'approche dans (Daoud, 2009), intègre la gestion de l'évolution du profil à long terme basée sur son augmentation par des profils construits à court terme. L'exploitation du profil à long terme est à la base de la détection du basculement des centres d'intérêts au cours des

sessions de recherche. Cette détection est basée sur une similarité conceptuelle entre requêtes successives qui permet de scruter le changement de l'importance des concepts récurrents d'une requête à une autre selon une méthode statistique. Cela afin de détecter, au cours du temps, les différents changements des centres d'intérêts.

5.3.3 Evolution du profil à court terme et à long terme

Dans (Bennett et al, 2012), les auteurs ont proposé un nouveau cadre unifié pour étudier la dynamique du comportement de l'utilisateur. Ils ont personnalisé une requête émise en utilisant à la fois des profils d'utilisateurs à court et à long terme basés sur trois vues temporelles : une session (court terme), l'historique de recherche (long terme) et une combinaison des deux. Ils ont considéré des requêtes émises, des résultats précédents et toute action effectuée par l'utilisateur tel que l'affichage d'un résultat, son ignorance explicite ou son absence (Kacem Sahraoui, 2017)

Dans (Tamine et al, 2008), ils ont proposé un classement personnalisé des documents basé sur un profil utilisateur. Premièrement, ils ont modélisé le profil de l'utilisateur qui contient les concepts d'intérêts déduits de son historique de navigation. Deuxièmement, ils ont appris les intérêts à long terme de l'utilisateur en gérant les intérêts à court terme. Ils ont défini le profil à court terme comme un nombre limité de sessions et l'utiliser pour mettre à jour le profil en utilisant une mesure de corrélation permettant de détecter les changements dans le comportement de l'utilisateur.

6 Session de recherche

Dans (Zemirli, 2008), une session de recherche est décrite par une requête et un ensemble de documents associés, jugés explicitement ou implicitement pertinents par l'utilisateur.

Une autre définition a été proposée dans (Achemoukh, 2012) où une session est considérée comme une ou plusieurs activités de recherche correspondante au même centre d'intérêt. Une activité de recherche est l'association d'une requête et le centre d'intérêt correspondant avec l'ensemble des documents jugés pertinents pour cette requête et les termes qui les indexent.

D'autre part, au niveau temporel, la durée d'une session peut varier de moins d'une minute (Spink et al, 2006) quelques minutes (Göker et al, 2000), à quelques heures (Spink et al, 2006).

Dans ces différents cas, la durée d'une session reste courte et inférieure à une journée (Leva et al, 2014).

Malgré ces points de divergence, la majorité des définitions s'accordent sur le fait qu'une session permet de regrouper des requêtes soumises par un même utilisateur, liées à un même besoin en information. Ces requêtes peuvent être envisagées de manière séquentielle (Silverstein et al, 1999), (Göker et al, 2000), (Jansen et al, 2007) ou imbriquée (Jones et al, 2008), Ce dernier correspond à une recherche multitâche et peut se traduire par une alternance entre des requêtes visant chacune un besoin en information distinct, et donc par des sessions imbriquées entre elles (Leva et al, 2014) (Achemoukh, 2018).

6.1 Approches de délimitation des sessions de recherche

L'évolution du profil utilisateur à court terme nécessite des techniques d'identification et de collecte d'informations utiles et fortement liées aux interactions de recherche courantes de l'utilisateur. Ces techniques se basent souvent sur des mécanismes de délimitation des sessions de recherche définies par un intervalle de temps ou une séquence de requêtes liées à un même besoin en information. Sur un intervalle de temps, un utilisateur peut faire une ou plusieurs sessions de recherche.

Dans le but de définir les sessions de recherche, plusieurs approches ont été introduites. Ces approches peuvent être classifiés en trois catégories : les approches basé-temps, les approches basé-contenu et les approches sémantiques.

6.1.1 Les approches basées temps

La première méthode de détection automatique des sessions s'appuie sur la dimension temporelle des sessions, et envisage donc leur structure comme séquentielle.

Les premières approches de définition des sessions de recherche sont basées sur la spécification d'un intervalle de temps moyen pour une session, appelé Timeout (Daoud, 2009). Cette méthode repose sur l'observation que plus la durée entre deux requêtes consécutives est longue, moins il est probable que ces requêtes renvoient à un même besoin en information, et donc elles appartiennent à une même session. Tout l'enjeu réside alors dans le choix d'un seuil temporel approprié fixant la durée maximale entre deux requêtes successives appartenant à la

même session : 5 minutes (Silverstein et al, 1999), 15 minutes (Göker et al, 2000), ou encore 30 minutes (Jansen et al, 2007).

L'analyse est faite sur deux fichiers logs et montre qu'un intervalle de temps entre 10 et 15 minutes est identifiés comme le seuil optimal d'identification des sessions de recherche basé temps.

Malgré sa forte utilisation due à sa simplicité de mise en œuvre, cette approche ne détecte ni les sessions très courtes résultant d'un changement soudain du besoin en information, ni les sessions très longues au cours desquelles l'utilisateur peut effectuer des pauses importantes entre chaque requête. La prise en compte de cette approche nécessite en effet de s'appuyer sur d'autres indices de lien entre les requêtes (Leva et al, 2014).

6.1.2 Les approches basé-contenu

Les limitations de l'approche temporelle, ont orienté les travaux vers une approche basée contenu qui exploite le lien lexical entre les requêtes visant un même besoin en information.

Ces approches sont basées sur des mesures de similarité textuelle qui se catégorisent en des mesures basées mots clés ou phrases ou alors des mesures basées sur la distance d'édition des chaines de caractères entre deux requêtes successives. En effet, plus les requêtes ont un contenu lexical en commun, plus il est probable qu'elles appartiennent à une même session.

Néanmoins, l'approche lexicale possède deux principaux inconvénients : elle nécessite la présence d'au moins un mot commun entre les requêtes, et se heurte aux phénomènes de changement sémantique (synonymie, hyperonymie, hyponymie…) (Leva et al, 2014).

6.1.3 Les approches sémantiques

Ces approches sont basées sur des mesures de similarité sémantique qui se catégorisent en des mesures basées sur le feedback utilisateur et des mesures basées sur l'information mutuelle. Les mesures basées sur le feedback utilisateur consistent à calculer le nombre de pages visitées en commun pour deux requêtes successives. L'intuition derrière cette mesure est que deux requêtes ayant des documents en commun visitées par l'utilisateur partagent le même sujet. Cette mesure permet de grouper des requêtes sémantiquement liées dans une même session (Daoud, 2009).

Les mesures basées sur l'information mutuelle consistent à calculer le nombre de documents indexés par les termes provenant des deux requêtes successives. Le but dans cette

étude est de développer un SRI basé-session où le contexte est représenté par l'ensemble de requêtes et ses résultats associés dans une même session de recherche.

7 Synthèse des approches de modélisation du profil utilisateur

Nous présentons dans cette section une synthèse des approches de représentation, construction et évolution du profil utilisateur abordées précédemment ainsi que des exemples des systèmes. Ces approches sont groupées dans le tableau 2.1 :

Représentation du profil	**Construction du profil**	**Evolution du profil**	**Exemples de systèmes**
Ensembliste	Classification non supervisée des pages Web visitées ou pertinentes dans des classes/centres d'intérêts.	Le profil est construit de plusieurs centres d'intérêts sans subir une évolution	(Gowan J. , 2003)
	Ensemble de documents collectés	à court terme	(Dumais et al, 2003)
	Agrégation des centres d'intérêts appris au cours des sessions de recherche	à court et à long terme.	(Tamine et al, 2008)
Conceptuelle	Pages web visités et l'ODP	à court terme	(White et al, 2010)
	classification des documents dans des ontologies de domaines prédéfinies.	à court terme	(Daoud, 2009)
Connexionniste	Extraction d'un graphe de requêtes documents à l'intermédiaire des sessions de recherche.	Le processus d'évolution est similaire au processus de construction par ajout des relations de corrélations pondérées entre les requêtes documents.	(RongWen et al, 2004)
	Agrégation des contextes de recherche au cours de l'historique de recherche.	à court terme	(Ustinovskiy et al, 2013)

Tableau 2.1: Synthèse des différents modèles de modélisation du profil utilisateur (Daoud, 2009)

8 Conclusion

Dans ce chapitre nous avons présenté la notion du profil utilisateur et son intégration dans le processus de recherche. Nous avons abordé les approches et techniques de modélisation du profil utilisateur, à savoir sa représentation, sa construction et son évolution au cours de temps. Ensuite nous avons abordé les différentes méthodes de délimitations de sessions de recherche.

Nous avons pu constater que les défis majeurs pour faire asseoir une personnalisation efficace dépendent du modèle de représentation du profil utilisateur, sa construction et son évolution au cours du temps.

Chapitre 3 : Intégration de la dimension temps dans la représentation du profil utilisateur.

1 Introduction

Dans les approches de personnalisation, la fraîcheur d'un terme dans le profil utilisateur est supposée être résolue uniquement par sa fréquence mais pas par sa position dans le temps.

Plusieurs travaux de la recherche d'information personnalisée se sont orientés vers les approches qui considèrent la dimension temps pour améliorer la recherche. Elle a été utilisée pour modéliser le profil de l'utilisateur en considérant la distribution temporelle des termes en prenant en compte leurs fréquences normalisées et leurs positions dans le temps. La motivation principale de ce type d'approche réside dans le fait que les centres d'intérêts des utilisateurs évoluent avec le temps.

Dans ce chapitre, nous nous intéressons à l'intégration de la dimension temps dans la représentation du profil utilisateur. Pour cela, nous nous basons sur le modèle multidimensionnel qui prend en compte deux dimensions. Nous avons pondéré les termes du profil en combinant leurs fréquences normalisées avec un critère temporel.

Nous essayons de répondre aux problématiques suivantes :

1. Comment représenter le profil utilisateur en intégrant la dimension temps dans sa représentation ?
2. Comment intégrer le profil résultant dans le processus de recherche ?

2 Terminologie et notations

2.1 Interaction de recherche

Une interaction de recherche est représentée par une requête *q* soumise à un instant *t* par un utilisateur, la liste de résultats D retournés par le système correspondant à la requête *q* et la sous-listes de résultats D_r jugés pertinents implicitement par l'utilisateur.

2.2 Session de recherche

Une session de recherche est définie par une séquence d'interactions de recherche liées à un même besoin en information. De plus, la session peut être définit par un intervalle qui varie de quelques minutes à quelques heures, ainsi la durée de cette session reste courte et inférieure à une journée.

2.3 Centre d'intérêt

Un centre d'intérêt est l'ensemble des besoins en information courants et récurrents de l'utilisateur. Chaque centre d'intérêt est représenté par un vecteur de termes pondérés.

2.4 Profil utilisateur à court terme

Le profil à court terme traduit généralement un centre d'intérêt de l'utilisateur construit sur la base d'une ou plusieurs interactions recherche de la même session traitant un même besoin en information

2.5 Profil utilisateur à long terme

Le Profil utilisateur à long terme représente les intérêts persistants de l'utilisateur extrait de ses interactions de recherche antérieurs.

2.6 Fraîcheur de l'information

Le concept de la fraîcheur des données introduit l'idée de l'âge des données : la donnée est-elle suffisamment récente par rapport aux attentes de l'utilisateur ? Est-ce que les données de la source sont les données les plus récentes (Peralta et al, 2004). La fraîcheur concerne principalement les documents qui traitent des nouvelles informations (ON-AT, 2017).

3 Approche de personnalisation

La motivation principale dans la plupart des approches temporelle réside dans le fait que les centres d'intérêts des utilisateurs évoluent avec le temps, pour cela nous visons à pondérer les termes du profil en combinant leurs fréquences normalisées et leurs fraîcheurs.

Notre approche de personnalisation est basée sur la représentation, la construction et l'évolution du profil utilisateur au cours du temps et ainsi que l'exploitation de ce profil dans le processus de la recherche d'information. Elle consiste à :

1) La représentation vectorielle du profil utilisateur comme un vecteur de termes pondérés traduisant un centre d'intérêt selon une récolte implicite des données, à partir de ses interactions de recherche.
2) L'intégration du facteur temps dans la représentation du profil.
3) La construction du profil utilisateur et son évolution au cours d'une session de recherche.
4) L'intégration du profil, dans la phase de ré-ordonnancement.

3.1 Représentation vectorielle du profil

Le profil utilisateur est représenté initialement sous forme vectorielle où nous attribuons à chaque terme du document un poids $w_{i,u}$,

$$U = \left(t_{i,1}: w_{i,1}, t_{i,2}: w_{i,2}, \dots t_{i,n}: w_{i,n}\right) \quad \textbf{(3.1)}$$

$w_{i,u}$ est le poids du terme t_i qui correspond à son degré d'importance dans le profil U et qui peut être intuitivement obtenu comme suit :

$$Wi,u = \frac{1}{|Dr|} \sum_{dj \in Dr} \frac{freq\ (t_i, dj)}{\sum_{\forall t_k \in d} freq\ (t_k, dj)} \quad \textbf{(3.2)}$$

Où : $freq\ (t_i, dj)$ est la fréquence d'un terme t_i dans dj, $\sum_{\forall t_k \in d} freq\ (t_k, dj)$représente la somme des fréquences de tous les termes apparus dans dj et $|Dr|$ représente le nombre de documents pertinents.

3.2 Intégration du facteur temps dans la représentation

En tenant compte du fait que les centres d'intérêts des utilisateurs évoluent au cours du temps, nous supposons que plus le centre d'intérêt est proche d'un temps courant, plus sa fréquence temporelle serait significative.

Notre objectif est de mesurer la fraicheur d'un centre d'intérêt représentant le profil en revisitant la notion de fréquence normalisée des termes qui le compose.

Nous utilisons une fonction à noyaux gaussien[1] qui permet de pondérer les termes du profil selon leur fréquence et leur fraîcheur, après avoir prouvé son efficacité par des travaux antérieurs dans e domaine de positionnement du terme (lv et al, 2009) (Gerani et al, 2010) (Kacem Sahraoui, 2017).

Elle est présentée comme suit :

$$k(S_C, S_j) = \frac{1}{\sqrt{2.\pi}\,.\,\sigma}.\exp\left[\frac{-(S_C - S_j)^2}{2.\sigma^2}\right] \quad \textbf{(3.3)}$$

Où : σ : est un coefficient d'interpolation.

S_C : est le temps courant, S_j :est le temps où le terme apparait.

Par conséquent, le profil utilisateur sera représenté par un vecteur de termes, où chaque terme possède un poids dépendant du temps qui reflète sa fraîcheur :

$$\vec{U} = \left(t_1{}^{S_j} : W_1{}^{S_j}, t_2{}^{S_j}: W_2{}^{S_j}, \ldots, t_m{}^{sj}: W_m{}^{S_j}\right) \quad \textbf{(3.4)}$$

Où le poids d'un terme $W(t_i)^{S_c}$ après intégration du temps dans le profil correspond à la somme de ses fréquences normalisées (poids) en fonction du temps sur le nombre de documents pertinents Dr définie comme suit :

$$W(ti)^{S_c} = \frac{1}{|Dr|}\sum_{dj \in Dr} TF(ti)^{Sj}.\ K(S_C, S_j) \quad \textbf{(3.5)}$$

[1] https://www.em-consulte.com/em/autopromo/Pass_sante_470981_chap07.pdf
http://pages.stat.wisc.edu/~mchung/teaching/MIA/reading/diffusion.gaussian.kernel.pdf.pdf

Où : $|Dr|$: c'est le nombre de documents pertinents. $TF(ti)^{Sj}$: représente le poids d'un terme dans le document à un temps *Sj*. Estimé comme suit :

$$TF(ti)^{Sj} = \frac{freq^{S_j}(ti)}{\sum_{\forall k \in Dr^{S_j}} freq^{S_j}(tk)} \quad \textbf{(3.6)}$$

Où $freq^{S_j}(ti)$ est la fréquence d'un terme ti dans Dr^{S_j} et $\sum_{\forall k \in Dr^{Sj}} freq^{S_j}(tk)$ représente la somme des fréquences de tous les termes apparus dans Dr^{S_j}

3.3 Algorithme de construction d'un profil utilisateur basé temps

Algorithme : construction d'un profil utilisateur ;

// Entrée : Dr : ensemble de documents pertinents, Sj : le temps où le terme apparait

Sc : le temps où le terme apparait

// Sortie : $\vec{U}$: Vecteur du profil utilisateur

Début

$Sc \leftarrow$ Temps courant

Répéter pour chaque temps *Sj* **faire**

Pour tous dj ∈ Dr **faire**

Pour tous les termes de dj **faire**

$W(ti)^{Sc} \leftarrow nTF(ti)^{Sj}.\ K(\mathrm{Sc}, \mathrm{Sj})$

Fin pour ;

Fin pour ;

Jusqu'à $(Sj > Sc)$

$\vec{U} \leftarrow (t, w(t))$

Fin.

3.3.1 Illustration

Dans le but d'illustrer notre approche, on suppose un scénario de recherche comportant onze interactions de recherche effectuées par un utilisateur dans une même session.

Dans cet exemple, on suppose qu'à chaque temps t les documents ayant été jugés pertinents par l'utilisateur u sont D= (d1, d2) tel que T= (t1, t2, t3) sont les termes indexant ces documents.

1) Construction du profil à l'instant t = 1 :

- La première interaction de recherche est caractérisé par la requête q= {t1, t2}
- A partir de l'ensemble de documents retournés, on suppose que l'utilisateur a jugé deux documents pertinents, qui sont : d1 = {t1, t2}, d2 = {t1, t2}.

a) Calcul des poids des termes dans le profil

Le tableau 3.1 représente la fréquence des termes et le leurs poids dans les documents en appliquant la formule (3.6) :

	Fréquences		Poids	
	T1	T2	T1	T2
D1	3	1	0.75	0.25
D2	3	1	0.75	0.25

Tableau 3.1: la fréquence des termes et le leurs poids.

Le centre d'intérêt représentant le profil utilisateur est construit comme un vecteur de termes pondérés selon la formule (3.2) : $U = (t1 : 0.75 ; \ t2 : 0.25)$.

b) Calcul de des poids des termes dans le profil en intégrant la dimension temps

En appliquant la formule (3.3) la valeur du facteur temps k (Sc, Sj) est estimée comme suit :

$$k(11,1) = \frac{1}{\sqrt{2.3.14}.4}.\exp\left[\frac{-(11-1)^2}{2.4^2}\right] = 0.0043$$

Où : $S_c = 11$: représente le temps courant.

$S_j = 1$: représente le temps où le terme *ti* apparait.

$\sigma = 4$: est un coefficient d'interpolation.

Pour obtenir le poids d'un terme dans le profil après intégration du facteur temps, on applique la formule (3.5) :

$$W(t1)^{Sc} = \frac{1}{2}\left[\left(\frac{3}{4}+\frac{3}{4}\right)\right] * 0.0043 = 0.003$$

$$W(t2)^{Sc} = \frac{1}{2}\left[\left(\frac{1}{4}+\frac{1}{4}\right)\right] * 0.0043 = 0.001$$

Le vecteur profil utilisateur résultant est le suivant : $\boldsymbol{U} = (\mathbf{t1} : \mathbf{0.003}, \mathbf{t2} : \mathbf{0.001})$

D'une manière similaire, nous avons construit le profil utilisateur à différents temps. Les tableaux ci-dessous (3.2) et (3.3) représentent respectivement les résultats obtenus après le calcul des poids des termes dans les profils sans et avec intégration du temps :

temps	Profil	T1	T2	T3
1	U1	0,75	0,25	0
2	U2	0.657	0.342	0
3	U3	0,585	0,414	0
4	U4	0,688	0,311	0
5	U5	0,590	0,408	0
6	U6	0,543	0,290	0,157
7	U7	0,568	0,252	0,178
8	U8	0,489	0,233	0,230
9	U9	0,411	0,181	0,407
10	U10	0,374	0,179	0,445
11	U11	0,314	0,186	0,498

Tableau 3.2: le poids des termes sans intégration du temps.

temps	Profil	T1	T2	T3
1	U1	0,003	0,001	0
2	U2	0,004	0,002	0
3	U3	0,005	0,003	0
4	U4	0,007	0,003	0
5	U5	0,007	0,005	0
6	U6	0,009	0,005	0,007
7	U7	0,012	0,005	0,009
8	U8	0,014	0,006	0,013
9	U9	0,014	0,006	0,027
10	U10	0,014	0,006	0,033
11	U11	0,014	0,008	0,041

Tableau 3.3: le poids des termes après intégration du temps

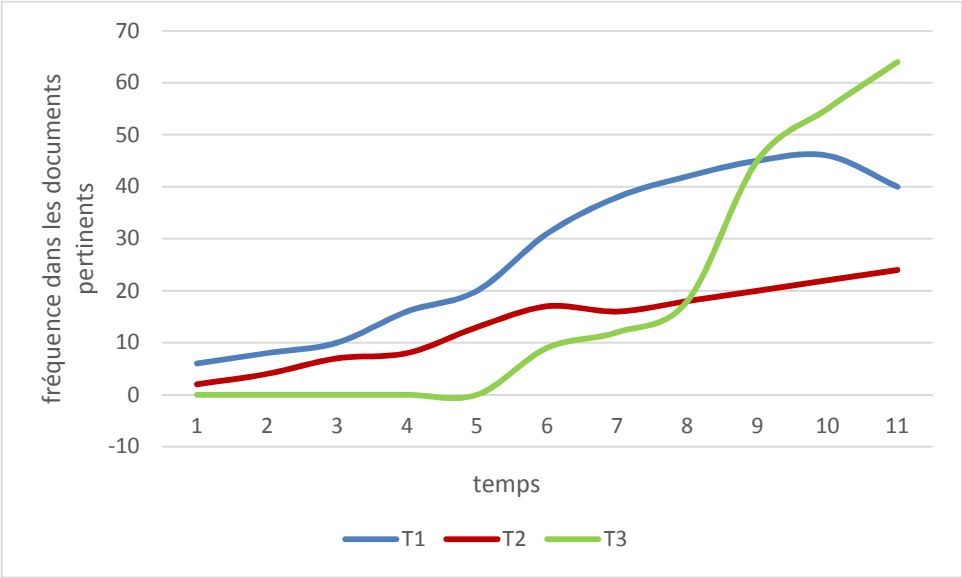
Après l'intégration du facteur temps, on remarque qu'à partir du temps t=9 la valeur du poids de terme T3 augmente jusqu'à dépasser le poids de T1, malgré que la fréquence du terme T1 dans les documents pertinents est élevée par rapport à la fréquence du terme T3. Le poids de T1 reste stable tandis que la valeur de poids de T3 continue d'augmenter.

3.3.2 Interprétation des résultats

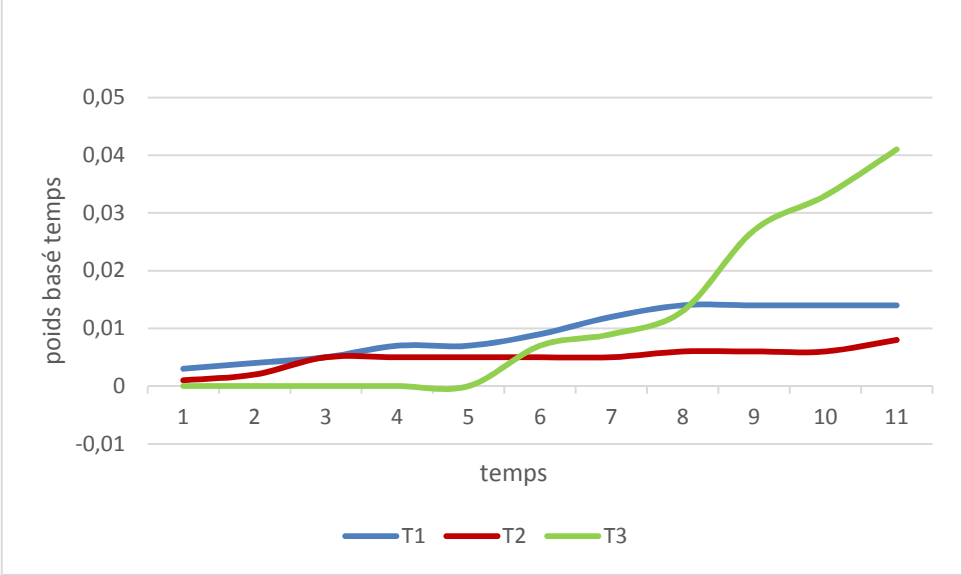
Pour mieux illustrer nos résultats les figures (3.1-a) et (3.1-b) montrent la distribution de trois termes composant les documents pertinents d1 et d2 au cours des différents temps. La figure (3.1) illustre les résultats de la distribution de trois termes en utilisant d'abord leur fréquence dans les documents pertinents (voir la figure 3.1-a) et leur fréquence normalisée combinées avec le temps en utilisant la fonction à noyaux gaussien (voir la figure 3.1-b).

D'après les résultats obtenus nous remarquons que le terme T1 commence par une haute fréquence, et lorsqu'on utilise une fonction temporelle T1 augmente lentement. Cependant le terme T3 apparu récemment commence à partir d'une basse fréquence (0 dans ce cas) et continue d'augmenter jusqu'à atteindre la même fréquence que T1, lorsqu'on utilise une fonction temporelle le terme T3 dépasse T1. Le terme T2 qui a une distribution uniforme continue d'augmenter uniformément.

Ainsi nous concluons que la formule (3.5) permet l'atténuation de l'importance des termes apparus pour longtemps et donne plus d'importance aux termes apparus récemment.



(a)



(b)

Figure 3.1:Exemple de distribution de termes en utilisant la fréquence des termes dans les documents pertinents (a) et le poids basé temps (b)

3.4 Intégration du profil dans la phase de ré-ordonnancement

Le principe de ce modèle consiste à identifier le besoin informationnel de l'utilisateur derrière sa requête de recherche et le projeter sur le contenu de son profil**.** Un SRI de base retourne à l'utilisateur un ensemble de documents ordonnés selon leur degré de pertinence en réponse à sa requête.

Le profil utilisateur construit est exploité dans le ré-ordonnancement des résultats de recherche. Cependant la personnalisation à ce stade du processus de recherche offre une solution en réordonnant les résultats pour ne présenter à l'utilisateur que les documents pertinents en prenant en compte son centre d'intérêt.

La fonction de ré-ordonnancement présenté dans notre travail consiste à combiner le score de similarité entre le document et la requête et le score de similarité entre document et le profil, dans le but de personnaliser les résultats de recherche pour toutes les requêtes utilisateurs durant une session de recherche.

$$score(\vec{U}, \vec{Q}) = \alpha . Sim(\vec{D}, \vec{Q}) + (1 - \alpha) . Sim(\vec{U}, \vec{D}) \quad \textbf{(3.7)}$$

Où $\vec{U}$ un vecteur du profil utilisateur, $\vec{Q}$ un vecteur requête, $\vec{D}$ un vecteur document,

Sim $(\vec{D}, \vec{Q})$est le score obtenu des résultats originaux en combinant entre le document et la requête, et Sim $(\vec{U}, \vec{D})$ est la similarité entre le profil utilisateur et le document. Ainsi la similarité entre deux vecteurs est calculée en utilisant la mesure de cosinus comme suit :

$$\cos(\vec{Q}, \overrightarrow{D_j}) = \frac{\vec{Q} * \overrightarrow{D_j}}{\|\vec{Q}\| * \|\overrightarrow{D_j}\|} = \frac{\sum_j^n w_{iQ} * w_{ij}}{\sum_{i=1}^n (w_{iQ}{}^2)^{1/2} * \sum_{i=1}^n (w_{ij}{}^2)^{1/2}} \quad \textbf{(3.8)}$$

Où : w_{ij} est le poids su terme ti dans dj, w_{iQ} est le poids du terme ti dans la requête.

3.4.1 Algorithme de ré ordonnancement

Algorithme du ré-ordonnancement dans le profil utilisateur basé temps

Algorithme : personnalisation

// Entrée : $\overrightarrow{U}$: Vecteur de profil utilisateur ; $\overrightarrow{Q}$: Vecteur de requête ; $\overrightarrow{dJ}$: Vecteur document ;

// Sortie : vecteur de profil utilisateur réordonnancé ;

α : Paramètre de lissage ;

Début

1. Une requête utilisateur est soumise au système de recherche d'information.
2. Le moteur de recherche retourne des documents.

Pour chaque $\overrightarrow{Q_k}$ **faire**

Pour chaque $\overrightarrow{d_j}$ **faire**

Calculer la similarité $(\overrightarrow{Q_k}, \overrightarrow{d_j}) \leftarrow \cos(\overrightarrow{Q_k}, \overrightarrow{d_j})$

Fin pour

Fin pour

3. Les documents sont ordonnés de plus pertinent au moins pertinent.

4. Le système calcule la similarité de ces documents avec le profil utilisateur.

Pour chaque $\overrightarrow{U_i}$ **faire**

Calculer la similarité $(\overrightarrow{U_i}, \overrightarrow{d_j}) \leftarrow \cos(\overrightarrow{U_i}, \overrightarrow{d_j})$

Fin pour

Score $(\overrightarrow{U_i}, \overrightarrow{Q_k}) \leftarrow \alpha \cos(\overrightarrow{Q_k}, \overrightarrow{d_j}) + (1-\alpha) \cos(\overrightarrow{U_i}, \overrightarrow{d_j})$

5. Enfin, la liste réordonnée de documents les plus pertinents est présentée à l'utilisateur.

$\overrightarrow{d_j}$ Réordonnés.

Fin

4 Conclusion

Dans ce chapitre, nous avons exploré le problème de la RI personnalisée. Nous avons proposé une approche de personnalisation qui intègre la dimension temps dans la représentation du profil utilisateur.

Nous avons pondéré les termes du profil utilisateur en combinant leur fréquence normalisée et leur fraicheur en utilisant une fonction à noyaux gaussien. Par la suite nous avons illustré notre approche par un exemple et nous avons discuté les résultats obtenus.

Ainsi nous concluons que le fait de revisiter la notion de fréquence des termes normalisées en la biaisant avec une fonction à noyaux gaussien permet d'atténuer l'importance des termes en particulier ceux qui sont apparus pour longtemps.

Conclusion générale

1 Conclusion :

Notre travail s'inscrit dans le domaine de la recherche d'information (RI), plus particulièrement dans l'accès personnalisé à l'information. L'objectif étant d'intégrer la dimension temps dans la représentation du profil utilisateur pour la RIP.

Tout d'abord, nous avons commencé par introduire les généralités liées à la RI classique puis nous avons traité le passage vers la RI personnalisée, par la suite nous avons présenté la notion de profil utilisateur ainsi les approches de représentation, construction et évolution d'un profil utilisateur, enfin nous avons proposé notre approche.

Dans notre approche nous avons exploré le problème de la RI personnalisée, nous avons combiné deux critères, la fréquence et la fraicheur des termes composant le profil.

Le profil de l'utilisateur correspond à son centre d'intérêt, il est implicitement représenté comme un vecteur de termes pondérés où un vecteur représente un centre d'intérêt. Plus précisément, nous nous sommes intéressés à la distribution temporelle des centres d'intérêts d'un utilisateur. Nous avons revisité la notion de fréquence des termes normalisée en la biaisant avec une fonction à noyaux gaussien qui permet d'atténuer l'importance des termes en particulier ceux qui sont apparus pour longtemps. Par la suite nous avons illustré notre approche par un exemple et nous avons discuté les résultats obtenus.

Ce travail nous a permis d'explorer le domaine de la recherche d'information, d'enrichir nos connaissances et de développer nos propres capacités que ce soit sur le plan pratique ou même personnel.

Nous envisageons comme perspectives :

- L'implémentation du processus de recherche qui intègre le profil de l'utilisateur appris afin de personnaliser les résultats de recherche.
- Effectuer des tests sur une collection d'évaluation.

Bibliographie

Achemoukh, F. (2018). "Modèle de Recherche d'information personnalisée basé sur les réseaux bayésiens", thése de doctorat.

Achemoukh, F., Ahmed-Ouamer, R. (2012). "Modélisation d'évolution de profil utilisateur en recherche d'information personnalisée".

Ahu Sieg, Bamshad Mobasher, Robin Burke. (2007). Ontological User Profiles for Personalized Web Search.

Amato G, Straccia U. (1999). user profile modelling and applications to digital libraries.

Bennett, P.N. WHITE, R.W. CHU, W. DUMAIS, S.T. BAILEY, P. BORISYUK, F. CUI, X. (2012). "Modeling the Impact of Short- and Long-term Behavior on Search Personalization", Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 185–194.

Chen, L. and Sycara, K. (1998). 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.

Cheng, Y. QIU, G. BU, J. LIU, K. HAN, Y. WANG, C. CHEN, C. (2008). "Model Bloggers: Interests Based on Forgetting Mechanism", Proceedings of the 17th International Conference on World Wide Web, p. 1129–1130.

Cleverdon, C. 1967.

Croft, N. Belkin and W. (1992). "Information filtering and information retrieval : Two sides of the same coin ?". Communication of the ACM, 35(12):29-38,. 1992. 29-38.

D. Vallet, M. Fern´andez, P. Castells, P. Mylonas, and Y. Avrithis. (2006). personalized information retrieval in context. proceedings of the 21st National conference on artificial intelligence- 3rd international workshop on modeling and retrieval of context, . Bston, USA, : s.n.

Daoud, M. (2009). "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",.

Dey, Abowd and. 2000.

Dumais, S, Cuttrel, E, Cadiz, J.J., Jancke, G, R. Sarin, et D.C Robbins. (2003). "A system for a personal information retrieval and re-use". Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development Toronto, Canada, pp: 72–79.

Fuhr, N. (2000). Information retrieval : introduction and survey.

Fuhr, N. (2005). "Information retrieval – From Information Acess To Contextual Retrieval",.In M.Eibl,C.Wolf, and C.Womser-Hacker, editors,Designing Information Systems. , pp: 47-57.

Gauch, S., Chaffee, J., and Pretschner, A. 2003. "ontology-based personalized search and browsing". web intelligence and agent systems, pp: 219– 234.

Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.,. 2007. "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.

Gauch., S. Speretta and S. 2005. Personalized search based on user search histories. In Web Intelligence. IEEE Computer Society. france, : s.n., 2005. pages 622-628, .

Gerani, Shima, Mark James Carman, and Fabio Crestani. 2010. “Proximity based opinion retrieval". In: Proceedings of the 33rd International ACM SIGIR, Conference on Research and Development in Information Retrieval. SIGIR, ’10. Geneva, Switzerland: ACM.

Göker, A. et D. 2000. " Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning ". In P. Brusilovsky, O. Stock, et C. Strapparava (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems.

Gowan, J. 2003. "A multiple model approach to personalised information access". Master thesis in computer science, Faculty of science, Université de college dublin.

H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. 2003.. Query expansion by mining user logs. IEEE Trans. Knowl. 2003. 829–839, .

Hadef, M. 2014. "Prise en compte du profil utilisateur Dans un système de recherche d'information".

Hadjouni Krir, M. 2012. "Un Système de Recherche d'Information personnalisée basé sur un la modélisation multidimensionnelle de l'utilisateur".

Hammache., Arezki. 2013. "Recherche d'Information : un modèle de langue combinant mots simples et mots composés".

Hofmann., T. 1999. Probabilistic latent semantic analysis. In Proceedings of Uncertainty in Artificial Intelligence,. Stockholm.

Ingwersen, P. 1994. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. 1994. pp. 101-110.

Jansen, B. J., Spink A., Blakely C., Koshman. S. 2007. " Defining a Session on Web Search Engines". Journal of the American Society for Information Science and Technology 58(6) , pp: 862–871.

Jones, R. Klinkner, K. L. 2008. "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs". In Proceedings of the 17th ACM Conference on Information and Knowledge Management , pp: 699–708.

Kacem Sahraoui, A. (2017). "Personalized information retrieval based on time-sensitive user profile",.

Kacem, A. Boughanem, M. Faiz, R. (2014). "Time-User Profile for Optimizing Search Personalization", 22 nd User Modeling, Adaptation and Personalization.

Kobsa, A. (2001). "generic user modeling systems. user modeling and user adapted interaction".

Kobsa, A. (2005). "User modeling and user-adapted interaction. User Modeling and User-Adapted Interaction". 2005. 185-190.

Koch, N. (2000). "software engineering for adaptative hypermedia systems- reference model, modelling techniques and development process". Ph.D Thesis, Fakultät der Mathematik und Informatik, Ludwig-Maximilians-Universität München.

Kostadinov, D. (2007). "Personnalisation de l'information : une approche de gestion de profils et de reformulation de requêtes". PhD thesis, L'université De Versailles Saint-Quentin-En-Yvelines.

Lechani Tamine, L. Boughanem, M. (2005). "Accès personnalisé à l'inforamtion : Approches et Techniques".

Leva, S., Faessel, N. (2014). "Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe".

Lieberman, H. (1997). "Autonomous interface agents". In CHI '97 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 67–74, New York, NY, USA. ACM.

Lv, Yuanhua and ChengXiang Zhai (2009). "Positional Language Models for Information Retrieval". In: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09. Boston, MA, USA: ACM, pp. 299–306. ISBN: 978-1-60558-483-6.

Maloof, M.A. MICHALSKI, R.S. (2000). "Selecting Examples for Partial Memory Learning", Machine Learning, 41, 1, p. 27 52.

Miller, G. (1995). "Wordnet :a lexical database for english". Commun. ACM, 38(11) :39–41.

Myaeng, S. H. and Korfhage, R. R. (1986). "Towards an intelligent and personal-ized retrieval system". InProceedings of the ACM SIGART international sym-posium on Methodologies for intelligent systems, pages 121–129, Knoxville, Ten-nessee, United States. AC.

ON-AT, S. (2017). "Temporalité et réseaux sociaux : prise en compte de l'évolution dans la construction du profil utlisateur",.

Ourdia, RESSAD-BOUIDGHAGHEN. (2011). 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.

Peralta, V., Ruggia, R and Bouzeghoub, M. (2004). "analyzing and Evaluating Data Freshness in Data Integration Systems". In: Ingénierie des Systèmes d'Information 9.5-6, pp. 145–162.

Pretschner, A., Gauch.S. (1999). "Ontology Based Personalized Search". In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI),.

R. Kraft, F. Maghoul, and C. Chang. (2005). "contextual search at the point of inspiration". In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. 816-823.

Rong Wen, j. Lao, N. and Ma, W.-Y. 2004. "Probabilistic model for contextual retrieval". In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pages 57–63. ACM Press. 2004. pp. 57–63.

Salton, G. 1971. "The SMART Retrieval System--Experiments in Automatic Document Processing". Prentice-Hall Inc, NJ.

Salton, G. McGill, M. (1983). "Introduction to Modern Information Retrieval". McGraw-Hill, New York.

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

Sieg, A., Mobasher, B., and Burke, R. (2007). "web search personnalization with ontological user profiles". IN CIMK407 : proceedings of sixteeth ACM conference on conference on information and knowledge management,. NEW YORK, NY, USA, : s.n., 2007. pages 525-534,.

Sieg, A., Mobasher, B., Lytinen, S., and Burke, R. 2004. "Using concept hierarchies to enhance user queries in web-based information retrieval" . In Proceedings of the International Conference on Artificial Intelligence and Applications, IASTED .

Silverstein, C., H. Marais, M. Henzinger, et M. Moricz. 1999. " Analysis of a Very Large Web Search Engine Query Log". SIGIR Forum 33(1) , pp: 6–12.

Sontag D., Collins Thompson K., Bennet P.N., White R. W., Dumais S.T., and Billerbeck B. 2012. "Probabilistic models for personalizing web search". Proc WSDM , pp: 433-442.

Speretta, S. Gauch, S. 2004. "Personalizing search based on user search histories". In Proceedings of the 13th International Conference on Information Knowledge and Management, CIKM , pp: 238–239.

Speretta, S., Gauch, S. (2004). "Personalizing search based on user search histories". In Thirteenth International Conference on Information and Knowledge Management, CIKM , pp: 238–239.

Spink, A., Park M., Jansen B. J., Pedersen J. (2006). «Multitasking During Web Search Sessions". Information Processing and Management 42(1), pp: 264–275.

Stefani, C. Strappavara, A. (1998). "Personalizing access to web sites: The SiteIF project". In Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia, Pittsburgh, June, pp: 20-24.

Tamine, L., Zemirli, W. N., and Bahsoun, W. (2007). " Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information". Information - Interaction – Intelligence.

Tamine-Lechani, L, Boughanem,M and Zemirli, N. (2008). "Personalized document ranking : exploiting evidence from multiple user interests for profiling and retrieval". In Journal of Digital Information Management.

Tan, Bin, Xuehua Shen, and ChengXiang Zhai. (2006). "Mining Long-term Search History to Improve Search Accuracy". In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06. Philadelphia, PA, USA: ACM, pp. 718–723. ISBN: 1-59593-339-5.

Tanudjaja, F and Mui, L. (2002). "Persona : A contextualized and personalized web search". In Proc 35th Hawaii International Conference on System Sciences, page 53, Big Island,Hawaii.

Tchuente, D. (2013). "Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques".

Thanh Trung, Van. (2008). "Utilisation de profils utilisateurs pour l'accès à une bibliothèque numérique".

Ustinovskiy, Y., Serdyukov, P. (2013). "Personalization of web-search using short-term browsing context". In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp: 1979-1988.

White, R. W., Bennett, P.N. and Dumais. S.T. (2010). Predicting short-term interests using activity based search context" . Proc .CIKM, pp.1009-1018.

Widyantoro, D. H., Yin, J., Nasr, M. S. E., Yang, L., Zacchi, A., and Yen, J. (1999). "Alipes: A swift messenger in cyberspace." In Spring Symposium on Intelligent Agents in Cyberspace, , pp: 62-67.

X. Shen, B. Tan, and C. Zhai. (2005),. Context-sensitive information retrieval using implicit feedback. In SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, : s.n., 2005,. 43–50,.

Zemirli, W, N. (2008). "Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur".

Zheng, N. LI, Q. (2011). "A recommender system based on tag and time information for social tagging systems", Expert Systems with Applications, 38, 4, p. 4575 4587.