Un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur

Myriam Hadjouni Hadjouni Krir

Université de La Manouba
Ecole Nationale des Sciences de l'Informatique

Université Paris-Sud 11
Ecole Doctorale Informatique Paris-Sud

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

THESE EN COTUTELLE

Présentée en vue de l'obtention du Diplôme de Doctorat en Informatique

Par
Myriam Hadjouni Krir

Sujet

Un Système de Recherche d'Information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur

Préparée au sein du laboratoire

Soutenue le 21 Septembre 2012 à l'ENSI devant le jury d'examen :

M. F. KAMMOUN	Professeur émérite, Tunisie	Président
M. F. GARGOURI	Professeur à l'Institut Supérieur d'Informatique et de Multimédia de Sfax, Tunisie	Rapporteur
M. M. BOUGHANEM	Professeur à l'Université Paul Sabatier, Toulouse, France	Rapporteur
Mme. H. HAJJAMI BEN GHEZALA	Professeur à l'Ecole Nationale des Sciences de l'informatique, Tunisie	Directrice de thèse
Mme. M.A. AUFAURE	Professeur à l'Ecole Centrale Paris, France	Directrice de thèse

b

Remerciements

C'est avec une grande satisfaction que j'écris cette page de remerciements, signe de l'achèvement de ce travail mené sur plusieurs années. Plus forte que la tradition, l'existence de cette page est liée à un devoir moral et à une reconnaissance sincère.

Je tiens à remercier M. Farouk Kammoun, Professeur Emérite, pour avoir accepté de présider le jury. Je remercie également M. Faiez Gargouri, Professeur à l'Institut Supérieur d'Informatique et de Multimédia de Sfax, et M. Mohand Boughanem, Professeur à l'Université Paul Sabatier, Toulouse, pour avoir accepté d'être les rapporteurs de cette thèse et pour l'honneur qu'ils me font en participant au jury.

Un grand merci à Mme Henda Ben Ghezalla, Professeur à l'Ecole Nationale des Sciences de l'Informatique et Directrice du laboratoire RIADI-GDl, pour m'avoir chaleureusement accueillie au sein de son équipe. Je tiens à lui exprimer ma profonde reconnaissance pour la confiance qu'elle m'a témoignée en me confiant ce travail.

Je remercie chaleureusement Docteur Hajer Baazaoui, Maître-assistante à l'Institut Supérieur des Arts et MultiMedia de la Manouba pour avoir assuré l'encadrement de cette thèse. Je la remercie pour avoir suivi de très près ce travail. Ses conseils constructifs, son perpétuel encouragement et son souci constant de l'avancement de ma thèse ont permis l'aboutissement de mes travaux et la rédaction de ce mémoire.

Un grand merci à Mme Marie Aude Aufaure, Professeur à l'Ecole Centrale de Paris, pour avoir été ma co-directrice de thèse. Cette thèse, effectuée en cotutelle entre l'Université

Paris XI et l'Université de la Manouba a été l'occasion pour moi de travailler au sein d'une équipe dynamique lors de mes différents séjours en france.

Cette thèse rentre dans le cadre du projet STIC entre le laboratoire RIADI-GDL, l'équipe Axis de l'INRIA de Rocquencourt et l'Ecole Navale de Brest.
Je remercie à cette occasion, M. Yves Le Chevallier pour m'avoir accueillie à l'INRIA, m'avoir prodigué ses conseils ainsi que ses précieuses remarques.

Mes remerciements s'adressent aussi à M. Christophe Claramunt, Professeur en Informatique et Directeur de l'Institut de REcherche de l'académie NAVale (IRENAV) à l'Ecole Navale de Brest. Merci pour m'avoir accueillie à l'école Navale et de m'avoir donné de son temps tout au long de cette thèse. J'adresse mes remerciements à toutes les personnes de l'institut pour leur accueil chaleureux. Mes pensées vont particulièrement à David Brosset, Meriem Horchani, Marie Coz et Mathieu Petit.

Je tiens aussi à remercier Mme Lynda Tamine-Lechani, Maître de Conférences HDR à l'Université Paul Sabatier, pour le temps qu'elle m'a consacré ainsi que pour ses précieux conseils.

C'est avec sympathie que je souhaite témoigner ma reconnaissance à Mohamed Ramzi Haddad et Nesrine Ben Mustapha pour leur précieuse collaboration. Travailler avec eux est pour moi un grand plaisir.
Merci aussi à toutes les personnes du laboratoire qui m'ont accompagnée par leurs encouragements et leur continuel sourire.

Parallèlement à l'élaboration de ma thèse, j'ai travaillé à l'Ecole Supérieure de Commerce Electronique à l'Université de Manouba en tant qu'enseignante, ce qui m'a permis de connaître un Directeur hors du commun, M. Malek Ghenima, que je remercie pour sa disponibilité, son écoute et aussi pour m'avoir donné une grande place au sein de son équipe. Cette expérience m'a beaucoup enrichie.

Je remercie tous mes collègues et amis, et tout particulièrement Chaker Ben Mahmoud, Mohamed Ali Labidi, Mehdi Ben Ghanem, Naafa Hamza, Mme Wafa Ben Khaled, ...

Une pensée particulière à Insaf et Nadia Amri pour leur perpétuel soutien.
Je suis très heureuse d'avoir croisé sur ma route des personnes qui m'ont donné envie de faire ce métier.

Merci au personnel de l'ENSI pour sa grande collaboration. Une pensée particulière à Nabil pour tout le temps qu'il consacre aux thésards.

J'aimerais exprimer tous mes remerciements à mes amis et ma famille.
Une pensée affectueuse à ma soeur qui avait hate que je finisse, alors ça y est s'est fait, on retrouvera un peu du temps perdu.
Merci à ma tante Najet, à ma belle maman Nabiha pour leur perpétuel encouragement et pour leur grande et immense sagesse.

J'adresse un merci tout particulier à mon mari, Samy Krir, pour sa patience et son soutien tout au long de cette thèse.

Je dédie cette thèse à mes fils, Mohamed Chahyn et Mohamed Chady, que vous puissiez un jour accomplir vos propres rêves, et un peu de ceux que je fais pour vous...

A la mémoire de

Ma mère (16-05-2004)

Mon père (19-03-2007)

Et mon fils (12-11-2008)

Qu'ils reposent en paix

Résumé

Depuis l'explosion du Web, la Recherche d'Information (RI) s'est vue étendue et les moteurs de recherche sur le Web ont vu le jour. Une des conséquences de cette évolution est que les méthodes classiques de la RI, surtout destinées à des recherches textuelles simples, se sont retrouvées face à des documents de différents formats et des contenus riches. D'un autre côté, l'utilisateur du Web devient de plus en plus exigeant quant aux résultats retournés par les systèmes de RI. La personnalisation tente de répondre à ces exigences en ayant pour objectif principal l'amélioration des résultats retournés à l'utilisateur en fonction de sa perception et de ses intérêts ainsi que de ses préférences.

Le présent travail de thèse se situe à la croisée des différents aspects présentés et couvre cette problématique. Notre principal objectif est de proposer des solutions nouvelles et efficaces à cette problématique. Pour atteindre cet objectif, un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur a été proposé. Ce système comprend deux volets : 1/ la modélisation multidimensionnelle de l'utilisateur 2/ la collaboration implicite des utilisateurs à travers la construction d'un réseau de modèles utilisateurs, construit itérativement lors des différentes recherches effectuées en ligne. Un prototype supportant le système proposé a été développé afin d'expérimenter et d'évaluer l'ensemble de la proposition. Ainsi, nous avons effectué un ensemble d'évaluation :

- l'évaluation de l'efficacité de la recherche
- l'évaluation de l'efficacité de la recherche d'information intégrant les informations spatiales
- l'évaluation de la recherche exploitant le réseau d'utilisateurs

Les expérimentations menées montrent une amélioration de la personnalisation des résultats présentés par rapport à ceux obtenus par d'autres moteurs de recherche.

Mots clés : modélisation utilisateur, modélisation multidimensionnelle de l'utilisateur, personnalisation, recherche d'information, données spatiales

Abstract

The web explosion has led Information Retrieval (IR) to be extended and web search engines emergence. The conventional IR methods, usually intended for simple textual searches, faced new documents types and rich and scalable contents. The user, facing these evolutions, asks more for IR systems search results quality. In this context, the personalization main objective is improving results returned to the end user based sing on its perception and its interests and preferences.

This thesis context is concerned with these different aspects. Its main objective is to propose new and effective solutions to the personalization problem. To achieve this goal, a multidimensional user model based personalized search system is proposed. This system has two components : 1/ multidimensional user modeling /2 implicit users' collaboration through the construction of a users' models network.

A system prototype was developed for the evaluation purpose that contains :

- information retrieval quality evaluation
- information retrieval quality evaluation with the spatial user model data
- information retrieval quality evaluation with the whole user model data and the users' models network.

Experiments showed amelioration in the personalized search results compared to a baseline web search.

Keywords : User modeling, multidimensional user model, personalization, information retrieval, spatial data

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

"Quel document choisir afin de retrouver l'information que je cherche ? Quels sont les documents susceptibles de me répondre ? " Et bien d'autres questions sont posées par l'utilisateur au moment d'effectuer une recherche. Dans ce contexte, la multitude de ressources disponibles rend assez difficile leur exploitation. En effet, la propagation de l'outil informatique, du multimédia et surtout l'essor d'Internet ont beaucoup affecté la diversité et l'expansion de l'information. La perception de l'utilisateur quant à l'information a évolué. En effet, l'explosion de la quantité de données et des ressources, a rendu l'information plus disponible mais moins pertinente aux yeux de l'utilisateur. Ce dernier doit faire face à la difficulté de trouver la bonne information qui répond à son besoin.

En réponse à cette difficulté, les outils de recherche d'information ont évolué pour intégrer l'utilisateur dans leur processus de recherche. Par conséquent, l'utilisateur a plus de chance d'avoir une réponse adéquate à sa recherche.

Contexte de la thèse

Comme susmentionné, l'information utile et répondant aux besoins de l'utilisateur est de plus en plus difficile à fournir. La recherche d'information (RI), qui remonte aux années cinquante, tente de pallier ce problème. Son objectif principal est de sélectionner à travers une collection de documents (préalablement connus et dont les termes significatifs et descriptifs ont été extraits) ceux qui correspondent le mieux à la requête de l'utilisateur. Néanmoins, la RI se retrouve face au problème de l'explosion de l'information, en particulier dans le cadre du Web ainsi qu'à son perpétuel changement. Les méthodes classiques du processus de la RI, principalement utilisées pour des recherches textuelles, se retrouvent

face à une multitude de formats et de contenus. Ceci a conduit à l'émergence de nouvelles thématiques telles que la RI contextuelle.

Dans le présent travail, nous nous intéressons à la RI dans le cadre du Web. Ce dernier se caractérise par le volume de plus en plus croissant des documents qui le constituent ainsi que par la diversité de leurs types. A cet effet, la RI sur le Web peut être considérée telle une extension de la RI classique.

La personnalisation de l'information a pour objectif de répondre à l'utilisateur en tenant compte de ses caractéristiques et de ses préférences. Ces données représentent un élément important du processus de recherche. En effet, afin de permettre au SRI d'exploiter les données des préférences et des intérêts de l'utilisateur, ces dernières sont représentées par un modèle de l'utilisateur. A ce titre, la personnalisation dans les Systèmes de Recherche d'Information (SRI) prenant en charge la modélisation de l'utilisateur représente un atout considérable. Le modèle de l'utilisateur pourrait contenir ses centres d'intérêt, ses préférences, son contexte géographique et spatial, ses déplacements ou encore ses informations personnelles.

Par ailleurs, et compte tenu de l'évolution des technologies et des outils de communication, la donnée géographique est devenue de plus en plus présente dans les documents du Web. Elle est, de ce fait, demandée dans les recherches des utilisateurs et aussi dans les données qui leurs sont fournies. La donnée géographique pourrait être décomposée en trois facettes (Usery, 1996) ; (Palacio et al., 2010) : le spatial, le temporel et le thématique. En se basant sur cette décomposition, un utilisateur, en parcourant un document résultat d'une recherche sur le Web pourrait connaître la provenance de ce document. Il connait aussi le sujet dont traite le document ainsi que la date de sa publication ou sa mise en ligne. Cependant, l'utilisateur ne connait généralement pas sa provenance en termes de localisation. Ainsi, les données présentes sur le Web contiennent de plus en plus d'informations spatiales qui doivent être intégrées afin de mieux répondre aux besoins de l'utilisateur. En effet, des termes comme ceux des noms de villes ou des indications de chemin (près de, à droite, etc.), ou encore sa position à un moment donné (information fournie par les systèmes de localisation GPS [1]) sont autant de données spatiales utiles et pouvant servir

1. Global Positioning System : traduit par système de positionnement mondial

à l'amélioration des résultats de recherche fournis à l'utilisateur.
Ce travail de thèse se situe ainsi à la croisée des différents domaines présentés ; à savoir la recherche d'information, la personnalisation impliquant la modélisation de l'utilisateur et l'intégration des données spatiales.

Problématique et objectifs de la thèse

L'état de l'art a porté sur les axes de recherche suivants :

- La recherche d'information sur le Web
- La personnalisation de la Recherche d'Information ainsi que la modélisation de l'utilisateur
- L'intégration des données spatiales dans le processus de personnalisation

Cette étude a permis de constater que le processus de personnalisation dans les systèmes de recherche d'information est principalement confronté à la question de la définition des données nécessaires pour la connaissance de l'utilisateur. En effet, la représentation des données concernant l'utilisateur et leur intégration dans le processus de l'extraction de l'information est toujours d'actualité. La prise en compte de l'utilisateur dans ce processus nécessite de ce fait la modélisation et la représentation de ses données ainsi que son évolution à travers les différentes recherches. (Kobsa, 2005). En effet, nous constatons que l'une des principales raisons du manque de performances des techniques de personnalisation est typiquement l'application d'un modèle de l'utilisateur trop général (Gauch et al., 2007). Les utilisateurs peuvent avoir des préférences générales, récurrentes et stables, cependant, l'ensemble des informations contenues dans le profil n'est pas forcément approprié à toutes les situations de recherche. Le plus souvent, les systèmes n'utilisent qu'un sous-ensemble de ces informations, qu'ils supposent pertinent pour la recherche en cours. Dans ce cas, ces systèmes ont recours à des informations explicitement fournies par les utilisateurs eux-mêmes.
Par ailleurs, une autre constatation porte sur l'utilisation des données géographiques lors des recherches sur le Web. En effet, les données spatiales, temporelles et thématiques ne

sont pas assez exploitées lors des recherches. Des informations telles que la longitude et la latitude pouvant être contenues dans les documents ne sont pas très exploitées : lors des recherches : un même contenu parlant d'un musée d'art contemporain et ne présentant aucune information sémantique relative à un endroit peut être associé à deux lieux différents. Une manière de détecter cette différence est de considérer l'information spatiale relative au document même telle que l'appartenance géographique du publieur. Dans ce cadre, et dans le contexte de notre travail, nous exploitons une seule des trois facettes des données géographiques : l'information spatiale.

Notre objectif est donc de proposer un système de recherche d'information personnalisée sur le Web intégrant les données spatiales et les données sémantiques. Ce système repose sur la modélisation multidimensionnelle de l'utilisateur en considérant les informations spatiales émanant de ses centres d'intérêts et de ses recherches ainsi que les informations sémantiques pouvant en découler. Nous basons notre proposition sur la collaboration implicite des utilisateurs en utilisant leurs modèles et à travers la construction d'un réseau. Cette collaboration a pour objectif d'améliorer la pertinence des résultats fournis à l'utilisateur en élargissant le champ de recherche et en le variant du plus global au plus personnalisé. En effet, la collaboration est ainsi basée sur la prise en compte des modèles des utilisateurs les plus similaires au modèle de l'utilisateur en cours de recherche. L'utilisateur est ainsi au coeur du processus de recherche proposé.

Nos travaux font partie du projet de coopération tuniso-française DGRST [2]-INRIA [3] STIC, intitulé "Modélisation d'un environnement logiciel pour la personnalisation de la recherche d'information géographique sur le Web". L'objectif du projet est de proposer un environnement permettant la personnalisation de la recherche en tenant compte des profils et des préférences des utilisateurs.

Contributions

Pour répondre à cette problématique ainsi qu'à l'objectif de la thèse, la proposition considère principalement les aspects suivants :

2. Direction Générale de la Recherche Scientifique et Technique, Tunisie
3. Institut National de Recherche en Informatique et en Automatique, France

- La définition d'un modèle multidimensionnel de l'utilisateur caractérisé par la collecte implicite des données exploitées.
- La prise en compte des informations spatiales demandées et/ou fournies par l'utilisateur. Ces données peuvent faire partie des requêtes utilisateur ou encore être directement sélectionnées parmi les résultats fournis.
- L'exploitation des modèles des utilisateurs dans la construction d'un réseau de modèles afin de fournir une collaboration implicite pour la personnalisation de la recherche d'information.

Les trois aspects considérés par notre proposition, sont fondés sur une estimation implicite du degré d'intérêt de l'utilisateur pour les données visitées ainsi que du degré de rapprochement que peuvent avoir les modèles des utilisateurs dans le graphe.

Comparativement aux travaux présents dans le domaine, le système proposé se distingue par :

- La modélisation multidimensionnelle de l'utilisateur,
- L'intégration des données spatiales aussi bien dans la construction du modèle de l'utilisateur que dans le processus global de la recherche personnalisée,
- La construction d'un réseau d'utilisateurs qui a pour vocation de permettre une collaboration implicite entre les modèles des utilisateurs. Le réseau est traité en tant que graphe afin d'y effectuer les recherches nécessaires.

Un prototype a été développé afin d'expérimenter le système proposé. Ce prototype permet d'effectuer une recherche à partir de données du web en intégrant les modèles des utilisateurs dans le processus de recherche et de personnalisation des résultats.

L'expérimentation se déroule en deux principales phases : la récolte de données de navigation d'utilisateurs expérimentaux et l'étude de la personnalisation. Nous procédons dans l'évaluation par étapes en partant de la composante minimale du système proposé. , à savoir le modèle de l'utilisateur avec la dimension intérêt pour arriver à l'évaluation du système dans sa totalité. Les résultats expérimentaux sont donc évalués par niveaux :

- Le premier niveau concerne l'évaluation de la qualité du modèle de l'utilisateur ne contenant que la dimension intérêt, et de ce fait la composante sémantique du modèle.
- Le second niveau d'évaluation est relatif à l'efficacité de la recherche personnalisée

sémantique.

- Dans le troisième niveau, nous procédons à l'évaluation de l'efficacité de la recherche intégrant les informations spatiales du modèle de l'utilisateur.
- Et finalement, nous expérimentons le système dans sa totalité. Cette étape est décomposée en deux parties, la première expérimente la classification effectuée des utilisateurs par rapport aux intérêts réels de ces derniers. La seconde, évalue la personnalisation proposée par le système.

Organisation du rapport

Ce mémoire est constitué de six chapitres :

Le chapitre 1 présente les notions et concepts de base de la Recherche d'Information. Nous y présentons les principaux modèles de RI.

Le chapitre 2 présente l'émergence de la personnalisation de la RI. Nous y présentons les limites de la RI classique dans un environnement hétérogène et volumineux qui ont conduit à l'émergence de la personnalisation de la RI. Les approches de représentation, de construction et d'évolution du modèle de l'utilisateur de l'utilisateur y sont aussi présentées. Nous terminons ce chapitre par présenter les approches de personnalisation.

Le chapitre 3 présente notre contribution à la personnalisation de la recherche d'information et à la modélisation de l'utilisateur à travers la proposition d'un système de personnalisation spatiale et sémantique. Nous y introduisons la modélisation de l'utilisateur proposée ainsi que le réseau de modèles utilisateur.

Le chapitre 4 présente le modèle multidimensionnel de l'utilisateur. Les dimensions constituant ce modèle et leur construction y sont décrits.

Dans le chapitre 5, nous présentons le réseau de modèles utilisateurs ainsi que le processus global de la personnalisation.

Le chapitre 6 est consacré aux expérimentations et à leur évaluation. Nous y décrivons aussi le prototype développé à cet effet.

En conclusion, nous dressons le bilan de nos travaux réalisés dans le cadre de la personnalisation de la recherche d'information. Nous proposons ensuite quelques perspectives de ces travaux.

Chapitre 1

Concepts de base de la Recherche d'Information classique

1.1 Introduction

La Recherche d'Information (RI) est une discipline dont le but est de faciliter l'accès à l'information pertinente à travers l'exploitation de modèle et de techniques d'acquisition, d'organisation et de recherche de données.Ce chapitre traite les concepts de base de la RI classique.
La section 1.3 présente la recherche d'information et ses concepts de base. La section 1.4 présente le processus de la recherche d'information. Les modèles de RI sont présentés dans la section 1.5. Nous abordons ensuite la RI dans le web, section 1.6. Finalement, dans la section 1.7, nous présentons l'évaluation des systèmes de RI ainsi que les mesures exploitées.

1.2 Présentation de la recherche d'information

Il y a quelques années, rechercher une information nécessitait l'accès et la consultation directe des livres, ou bien l'accès à des bibliothèques. Dans ce dernier cas, il fallait parcourir les notices bibliographiques des documents. Ces notices, traitées manuellement, étaient, pour la plupart, classées par mots clefs, par auteurs ou encore par titres. Aujourd'hui, l'accès à l'information est beaucoup plus aisé. En effet, rechercher une information ne se limite plus au simple parcours des données qui concernent les documents. Le contenu de ces derniers est parcouru afin d'en dégager le nécessaire à la recherche. Dans ce contexte,

la discipline de la Recherche d'information (RI) a pour but de faciliter l'accès aux informations pertinentes répondant au mieux aux besoins des utilisateurs. Une information pertinente étant communément présentée comme "une information adaptée aux besoins de l'utilisateur".

Dans ce contexte, les systèmes de recherche d'information (SRI) intègrent les modèles et techniques permettant d'accomplir la tâche principale de tout SRI, de retrouver, à partir d'une collection de documents, ceux susceptibles d'être pertinents pour un utilisateur. Un système de recherche d'information est de ce fait présenté comme un système dont l'objectif est de retrouver les documents pertinents répondant à une requête de l'utilisateur. Dans la définition d'un SRI, on distingue trois notions clés : le document, la requête et la pertinence. Le document peut aussi bien être un texte, une partie de texte, une page Web, un fichier multimédia, etc. La requête est une interprétation du besoin de l'utilisateur. Quant à la pertinence, notion assez complexe, elle exprime le degré de correspondance du document aux besoins de l'utilisateur. Cette pertinence est liée à un besoin en information d'un utilisateur exprimé sous forme de requête.

1.3 Concepts de base de la recherche d'information

La Recherche d'Information (RI), (Salton, 1971), (van Rijsbergen, 1979), englobe des approches, systèmes et outils ayant pour objectif de permettre à l'utilisateur de choisir à partir d'un ensemble de documents ceux qui répondent le plus à ses besoins. La gestion de l'ensemble des documents doit permettre leur stockage, leur extraction à travers la recherche, et l'exploration des documents pertinents. Plusieurs concepts s'articulent autour de cette définition :

- La collection de documents, appelée aussi corpus de documents, est l'ensemble des données que le système de recherche d'information va exploiter. Chaque document est assimilé à un conteneur élémentaire d'information.
- Le document représente la donnée basique de la collection documents.
- Besoin d'information : nous pouvons distinguer trois principaux types de besoins de l'utilisateur (Ingwersen and Belkin, 2004) :

  - Besoin de vérification : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il est à la recherche d'une donnée particulière dont il connait les détails d'accès telle que la recherche d'une page web connaissant son URL. Un besoin de type vérificatif est dit stable, vu qu'il n'évolue pas au cours de la recherche (Baziz et al., 2007).
  - Besoin thématique connu : l'utilisateur a une idée claire de ce qu'il cherche mais a besoin de la clarifier ou de trouver de nouvelles informations dans un sujet et domaine connus. Un tel besoin peut aussi bien être stable qu'évolutif au fur et à mesure de la recherche.
  - Besoin thématique inconnu : dans ce cas, l'utilisateur effectue une recherche sur un domaine qu'il ne connait pas forcément.
- La requête permet à l'utilisateur d'exprimer ses besoins. Elle est souvent représentée par un ensemble de mots clefs exprimés en langage naturel, booléen ou par des sélections graphiques.
- La pertinence, notion assez subjective, est dépendante de l'utilisateur. Saracevic, dans (Saracevic, 1975) puis dans une version révisée (Saracevic, 1996), a distingué différents types de manifestation de la pertinence qui ont été reprises par (Simonnot, 2008) :
  - La pertinence du système : représente un degré de relation entre le document et la requête.
  - La pertinence du sujet : représente la relation entre le sujet de la requête (topic) et le document.
  - La pertinence cognitive : est présentée comme la relation entre l'état des connaissances de l'utilisateur et son besoin informationnel.
  - La pertinence situationnelle : comme son nom l'indique, ce type de pertinence est relié à la tâche courante. Il s'agit de la relation entre la recherche courante et des textes retrouvés.

1.4 Le processus de recherche d'information

La Recherche d'Information (RI) traite la représentation, le stockage, l'organisation et l'accès à l'information. La figure 1.1 représente le processus de la RI qui est construit selon les étapes suivantes :

- Un utilisateur formule son besoin d'information sous la forme d'une requête qui est ensuite indexée par le système,
- Le système construit les représentations de la requête et du document et a recourt à la collection de documents préalablement indexée. L'indexation des données étant l'identification des termes et concepts significatifs qui serviront pour la recherche dans la collection. Les représentations, indépendantes les unes des autres, sont ensuite mises en correspondance.
- Le système retourne une liste de documents considérés par le SRI comme pertinents par rapport à la requête utilisateur.

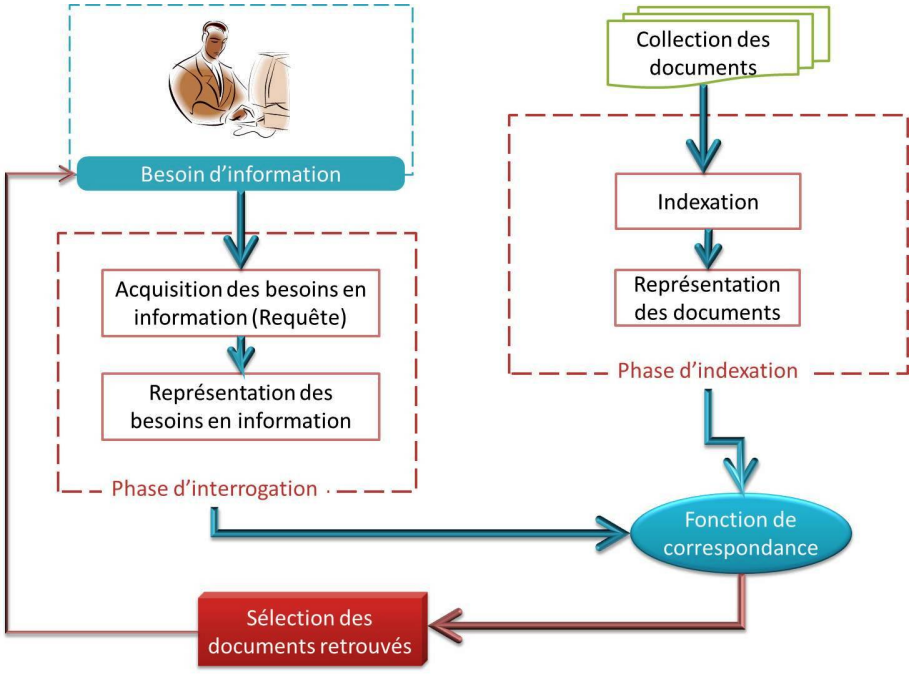


FIGURE 1.1 – Processus de Recherche d'information

1.4.1 L'indexation des documents et des requêtes

L'indexation consiste à analyser les documents que le système utilisera afin d'en extraire les mots clefs, appelés aussi descripteurs, qui serviront à la recherche. D'après la norme AFNOR [4] NF Z 47-102 1996 , il s'agit d'une opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts qui y sont contenus. L'indexation des documents peut être effectuée de manière manuelle, automatique ou semi-automatique. Dans le premier cas, un spécialiste du domaine d'application du système de recherche d'information est impliqué pour l'analyse. Pour le second cas, le processus d'indexation est totalement informatisé. Quant au troisième cas, il débute par une indexation automatique qui sera validée par la suite par le spécialiste du domaine.

1.4.2 Fonction de correspondance

La fonction de correspondance document/requête est représentée par un processus d'appariement entre le document et la requête. L'exploitation de cette fonction implique la connaissance de la requête ainsi que celle de la collection de documents dans laquelle la recherche est effectuée. A la soumission d'une requête de l'utilisateur, le SRI en effectue une représentation et calcule le score de correspondance relatif aux documents et à la requête. Ce score traduit un degré de pertinence système tel que défini par (Saracevic, 1975), il tient compte des poids des termes déterminés, de manière générale, en fonction d'analyses statistiques et probabilistes. Pour une requête donnée, le système retourne des documents classés par ordre décroissant selon le score de pertinence.

1.5 Les modèles de recherche d'information

Le modèle de recherche d'information joue un rôle principal dans la RI : il détermine le comportement du système de RI. Il se définit principalement par sa manière de modéliser la fonction de correspondance document/requête ainsi que par la représentation des documents et des requêtes. On peut distinguer trois grandes classes de modèles :

– Les modèles ensemblistes : sont basés sur la théorie des ensembles et ont été les

4. http ://www.afnor.org

premiers à avoir été mis en place.

- Les modèles algébriques : calculent des distances entre les représentations.
- Les modèles probabilistes : sont basés sur la théorie des probabilités. Ils ont pour objectif d'estimer des probabilités de pertinence d'un document en fonction de la requête.

Nous présentons dans la figure 1.2 la taxonomie des modèles de la RI selon (Baeza-Yates and Ribeiro-Neto, 1999) :

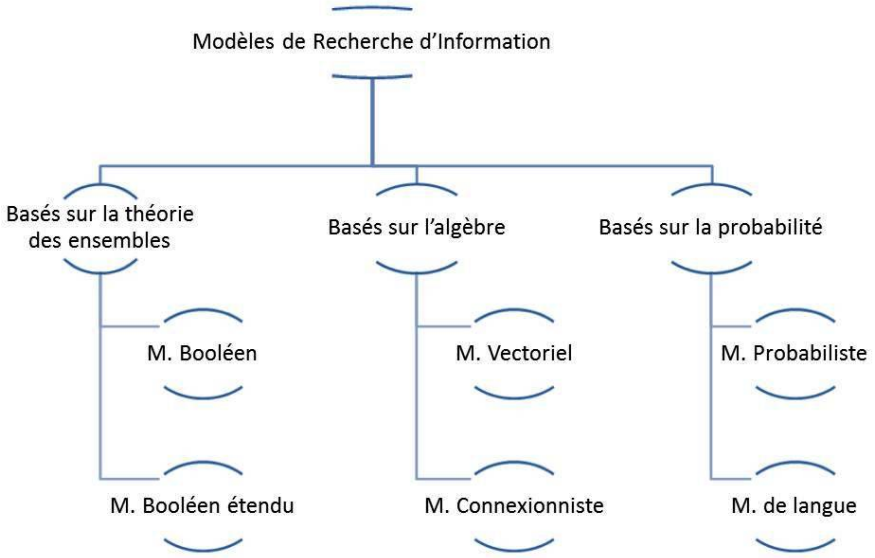


FIGURE 1.2 – Modèles de RI selon (Baeza-Yates and Ribeiro-Neto, 1999)

Nous présentons dans la suite de cette section ces trois modèles.

1.5.1 Le modèle ensembliste

Les modèles ensemblistes ne retournent que des documents qui répondent de manière exacte à la requête émise. Ces modèles sont basés sur la théorie des ensembles (Salton, 1971). En effet, les documents à retrouver sont exactement décrits par la requête : ils contiennent tous les termes de la requête émise (intersection des ensembles). Les documents, de ce fait, sont de deux types : pertinents ou non pertinents.

Dans ce modèle, documents et requêtes sont représentés par des ensembles de mots clés. Les documents sont indexés par un ensemble de termes non pondérés. Ils sont de ce fait représentés par une conjonction logique de ces termes. Quant aux requêtes, elles sont sous forme d'expressions booléennes dont les termes sont reliés par des opérateurs logiques OR, AND, NOT.
Pour un document $d = t_1, t_2, ...t_n$, et une requête q, la fonction de correspondance $rsv(d, q)$, pour Retrieval status value, est déterminée comme suit :
$rsv(d, t_i) = 1$ si $t_i \in d$; 0 sinon
$rsv(d, q_1 \wedge q_2) = 1$ si $rsv(d, q_1) = 1$ et $rsv(d, q_2) = 1$; 0sinon.
$rsv(d, q_1 \vee q_2) = 1$ si $rsv(d, q_1) = 1$ ou $rsv(d, q_2) = 1$; 0 sinon.
$rsv(d, \neg q_1) = 1$ si $rsv(d, q_1) = 0$; 0 sinon.

Ce modèle a l'avantage d'être facile à implémenter et de permettre aux utilisateurs d'exprimer des contraintes structurelles et conceptuelles (Marcus, 1991). Néanmoins, ce modèle présente les problèmes suivants :

- Les opérateurs logiques peuvent être mal manipulés par les usagers ce qui rend difficile la formulation de requêtes adéquates à l'aide d'expressions booléennes (Belkin and Croft, 1992). En effet, dans le langage courant l'expression de recherche "recherche d'information et personnalisation" veut dire recherche les deux ensembles de termes. Or, en logique booléenne, le "et" représente l'intersection de ces deux ensembles de termes.
- La correspondance entre document et requête étant soit 1, soit 0, il n'est pas possible à un système de recherche d'information de classer les documents retournés par ordre de pertinence.

1.5.2 Le modèle algébrique ou vectoriel

Dans le modèle vectoriel, les requêtes et le contenu des documents sont représentés dans un espace vectoriel construit par les termes d'indexation (Salton et al., 1983). La requête ainsi que le document sont donc représentés par deux vecteurs respectifs. La fonction de

correspondance est alors calculée en se basant sur la similarité entre ces deux vecteurs. Pour cela, des mesures comme le produit scalaire de deux vecteurs, la distance métrique et la mesure cosinus ont été proposées. Soient R l'espace vectoriel défini par l'ensemble des termes t_1,t_2,...,t_N, le document d et la requête q peuvent être représentés par des vecteurs de poids tels que : $\overrightarrow{d}$=(w_1,w_2,...,w_N) ; w_i correspond au poids associé au terme d'indice i de R dans le document d et $\overrightarrow{Q}$=(q_1,q_2,...,q_N) ; q_i correspond à un coefficient. Nous présentons dans le tableau 1.1 quelques mesures de similarité des vecteurs.

TABLE 1.1 – Mesures de similarités entre vecteurs

Mesure	**Equation**
Le produit scalaire	$rsv(\overrightarrow{d}, \overrightarrow{q}) = \sum_{k=1}^{n}(d_k * q_k)$ (2.1)
La mesure de Jaccard	$Jaccard(D_j, Q) = \frac{\sum_{i=1}^{n}(d_{ij} * q_i)}{\sum_{i=1}^{n}(q_i^2) + \sum_{i=1}^{n}(d_{ij}^2) - \sum_{i=1}^{n}(d_{ij} * q_i)}$ (2.2)
La mesure de cosinus	$\cos(D_j, Q) = \frac{\sum_{i=1}^{n} d_{ij} * q_i}{\sqrt{\sum_{i=1}^{n} q_i^2 * \sum_{i=1}^{n} d_i^2}}$ (2.3)

Pour la mesure de cosinus, une représentation vectorielle complète est utilisée avec la fréquence des mots. Dans ce cas, deux documents sont similaires si leurs vecteurs sont confondus. Et le cas échéant, leurs vecteurs forment un angle dont le cosinus est représenté par l'équation 1.1.

1.5.3 Le modèle probabiliste

Les modèles probabilistes se basent sur un calcul de similarité probabiliste (Salton et al., 1983). En effet, ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête (Robertson and Jones, 1976) et son principe de base est

de retrouver, pour une requête donnée, les documents ayant à la fois une forte probabilité d'être pertinents, et une faible probabilité de ne pas l'être. La similarité entre une requête q et un document d est donc fonction de la probabilité de pertinence du document d pour la requête q : il est soit pertinent soit non pertinent pour la requête. De ce fait, un document est sélectionné si la probabilité de sa pertinence et supérieure à la probabilité de sa non pertinence.

En notant R l'évènement de pertinence de d pour q et $\overline{R}$ l'évènement de non pertinence, la fonction de correspondance document requête, appelée aussi pertinence globale, RSV(Q,D), est calculée par :

$$RSV(q,d) = \frac{P(R/q)}{P(\overline{R}/d)} \tag{1.1}$$

Où $P(R/q)$ est la probabilité que le document soit pertinent,

et $P(\overline{R}/d)$ la probabilité que le document soit non pertinent.

Le modèle probabiliste a vu plusieurs extensions, la plus connue est BM25. Dans le modèle Okapi BM25, (Robertson and Walker, 1994), le calcul du poids d'un terme dans un document prend en considération la fréquence des termes, leur rareté ainsi que la longueur des documents. Le poids d'un terme dans un document est donc calculé par la fonction (Robertson and Jones, 1976) :

$$w = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{1.2}$$

Avec :

r : nombre de documents pertinents contenant le terme t,

R : nombre total de documents pertinents,

n : nombre de documents contenant le terme t,

N : nombre total de documents.

1.6 La recherche d'information dans le web

Comme décrit dans la section précédente, le but principal de la recherche d'information classique est de fournir des données ou du texte contenant des informations utiles à l'utilisateur. Une des principales différences entre la RI classique et la RI sur le web réside dans l'hétérogénéité des utilisateurs du web et de leurs besoins. En effet, la propagation

mondiale de l'information à travers la toile permet à des utilisateurs géographiquement, socialement et culturellement dissemblables d'accéder à l'information. Cette diversification des utilisateurs induit une diversification de leurs requêtes qui comportent un minimum de deux termes. En effet, la longueur moyenne des requêtes des utilisateurs sur le Web est estimée à environ 3 mots d'après (Jansen et al., 2000), (jason zien, 2001).

Se basant sur le fait que toute recherche sur le web part d'un besoin informationnel, plusieurs études se sont portées sur le passage de ce besoin vers la formulation de la requête. Dans une étude, (Broder, 2002) a repris l'exploitation en RI de la "forme verbale" dans le modèle de RI sur le web comme passage entre le besoin en information, associé à une tâche, et la requête (cf. figure 1.3). La "forme verbale" est généralement la transcription mentale d'un besoin en information. Cependant, les trois modèles de la RI classique présentés dans la section 1.5 ne tiennent pas compte de la sémantique des termes, ni du contexte de la requête. En effet, vu la diversité des tâches de recherche ainsi que des différents besoins des utilisateurs du web, plusieurs études ont été effectuées dans l'objectif d'extraire une taxonomie des requêtes des utilisateurs. Il en découle trois types de besoins des utilisateurs du web, introduits par (Ingwersen, 1992) et repris par (Rose and Levinson, 2004) et (Broder, 2007) : les besoins informationnels, les besoins navigationnels et les besoins transactionnels.

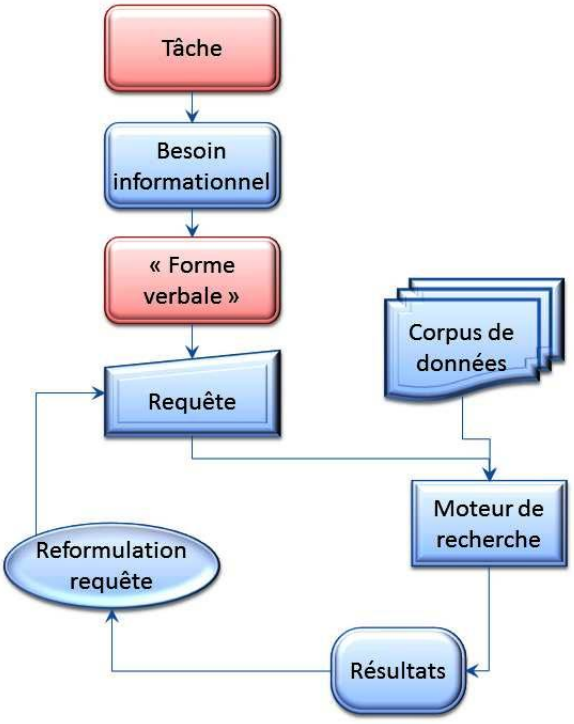


FIGURE 1.3 – Illustration d'un modèle de RI classique augmentée par le besoin utilisateur

La tâche de recherche informationnelle, associée à la tâche de recherche d'information, est supportée depuis les premières générations de moteurs de recherche : Altavista, WebCrawler, etc. Les requêtes associées à ce type de tâche représentent entre 40% et 65% des tâches utilisateurs. Les tâches navigationnelles représentent, quant à elles, entre 15% et 25% des requêtes émises par les utilisateurs. L'objectif de l'utilisateur est d'accéder à une page. Le reste des requêtes, soit 20% à 35% concerne les requêtes à tâches transactionnelles. Cette différenciation dans les types de requêtes émises a pour objectif principal la mise en place et l'utilisation de stratégies de recherche permettant d'avoir les ressources les plus à même de correspondre à chaque type de besoin.

L'évolution de la RI sur le web n'a pas seulement affecté les requêtes, elle a aussi amplifié la prise en considération de l'utilisateur et de ses attentes. En effet, traiter uniquement la requête utilisateur ne permet pas de deviner ses attentes et de ce fait de créer son profil (Hölscher and Strube, 2000). Ceci est illustré par le cas où un doctorant et un directeur des thèses effectuent la même recherche mais n'ayant pas les mêmes attentes (le directeur

aura tendance à avoir besoin de données pédagogiques alors que le doctorant chercherait probablement les évolutions du domaine de recherche). Connaitre donc l'utilisateur, en acquérir les données relatives (Teevan et al., 2008), lui construire un profil (Sieg et al., 2007a) et l'exploiter (Chirita et al., 2007) constituent un vrai challenge pour la RI.
La personnalisation de la recherche d'information tente de répondre en offrant des approches nécessaires à l'adaptation des résultats à l'utilisateur. Ces derniers doivent alors être connus et de ce fait modélisés par les systèmes de recherche d'information.

1.7 Evaluation des systèmes de recherche d'information

Evaluer un SRI consiste à en mesurer les performances et à estimer son aptitude à répondre aux besoins des utilisateurs. La performance est estimée en comparant les réponses du SRI fournies à un utilisateur pour une requête et celles que ce même utilisateur voudrait idéalement avoir. L'évaluation des systèmes de RI a fait l'objet de plusieurs travaux depuis les années 50.Le premier protocole d'évaluation a été proposé dans le cadre du projet Cranfield (Cleverdon et al., 1966). Sur ce modèle repose la plupart des modèles d'évaluation actuellement utilisés (Hersh et al., 1995), (Kekäläinen and Järvelin, 2002), (Borlund, 2003).

1.7.1 Hypothèses d'évaluation

Afin de réaliser une évaluation, les éléments expérimentaux suivants doivent être présents :

- Un ensemble de requêtes
- Un ensemble de documents.
- Etablir pour chaque requête la liste des documents pertinents

Les points principaux suivants sont considérés lors d'une campagne d'évaluation (Voorhees, 1998) :

- Les documents sont triés par scores décroissants,
- Lors de son parcours des résultats, l'utilisateur commence du premier au dernier, et ne procède jamais de façon aléatoire
- Deux documents non pertinents ne formeront pas d'unité informationnelle pertinente

– Un jugement de pertinence doit pouvoir se ramener au mieux à un nombre réel généralement borné. Cette hypothèse est réduite à un jugement de pertinence binaire : un document est alors soir pertinent pour l'utilisateur soit non pertinent.

1.7.2 Mesures d'évaluation

Les mesures d'évaluation se sont enrichies au fil des ans, la plupart se basent sur l'hypothèse que les documents non jugés sont non pertinents. A cet effet, plusieurs métriques ont été proposées pour évaluer la performance d'un SRI. La majorité d'entre elles tentent d'évaluer la capacité d'un SRI à sélectionner les documents pertinents en réponse à une requête. (Baccini et al., 2010) présentent une étude dans laquelle Ils ont défini sept groupes de mesures, extraites de 27 de départ. Nous présentons dans ce qui suit quelques-unes des mesures les plus utilisées dans l'évaluation de systèmes de RI.
Soit D l'ensemble des documents présents dans le corpus de recherche. Pour une requête donnée, un document est soit restitué ou non restitué et soit pertinent ou non pertinent. La figure 1.4 représente la répartition des documents lors d'une interrogation d'un système de RI.

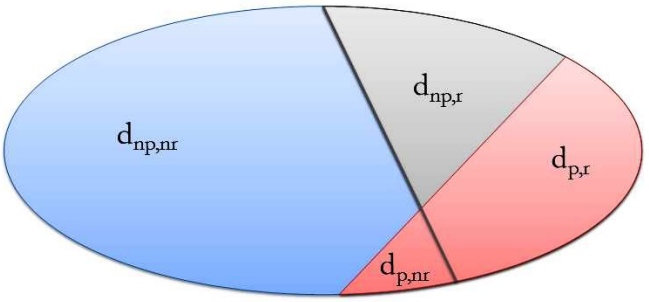


FIGURE 1.4 – Représentation des documents lors d'une interrogation du système de RI

Avec :

d_np, nr : l'ensemble des documents non pertinents et non restitués lors de la recherche,

d_np, r : l'ensemble des documents non pertinents et restitués, il s'agit du bruit,

d_p, r : l'ensemble des documents pertinents et restitués à l'utilisateur,

d_p, nr : l'ensemble des documents pertinents et non restitués, il s'agit du silence.

Nous reprenons les mêmes notations de cette figure afin de présenter les mesures d'évaluation suivantes :

Le rappel est la proportion des documents pertinents retournés parmi tous les documents pertinents disponibles. Il est calculé par la formule :

$$R = \frac{d_{p,r}}{d_{p,r} + d_{p,nr}} \tag{1.3}$$

La précision est la proportion des documents pertinents restitués parmi tous les documents restitués, la formule est la suivante :

$$P = \frac{d_{p,r}}{d_{p,r} + d_{p,nr}} \tag{1.4}$$

Les mesures de rappel et de précision sont inversement proportionnelles : lorsque l'une augmente l'autre diminue. Le but du système de RI étant l'augmentation de la précision sans altérer le rappel, et inversement.

La précision à différents niveaux de coupe , Pn, est le calcul de la précision après que n documents aient été retrouvés (pertinents ou non) ; n peut être dans l'ensemble $\{5, 10, 15, 20, 30, 100, 200, 500, 1000\}$.

La précision moyenne AP (Average Precision) est la précision moyenne calculée pour une requête. Il s'agit de la précision moyenne obtenue chaque fois qu'un document pertinent est retrouvé. La formule utilisée est :

$$AP = \frac{\sum_{k=1}^{n}(Pk * rel(k))}{d_{p,r}} \tag{1.5}$$

Où :

k le rang du document retourné

n le nombre de documents retournés

Pk précision au niveau de coupe k.

$Rel(k) = 1$ si le k^{me} document restitué est pertinent,

$Rel(k) = 0$ si le k^{me} document restitué est non pertinent.

La moyenne des précisions moyennes , plus utilisée sous la terminologie anglaise "Mean Average Precision", MAP. Il s'agit de la moyenne des précisions pour l'ensemble des requêtes considérées.

$$MAP = \frac{\sum_{q \in Q} AP_q}{|Q|} \tag{1.6}$$

Où : Q est l'ensemble des requêtes et |Q| est le nombre de requêtes.

La F-mesure correspond à une moyenne harmonique de la précision et du rappel. La valeur de F-measure diminue quand l'un de ses paramètres est petit et elle augmente quand les deux paramètres sont proches tout en étant élevés (van Rijsbergen, 1979).

$$F - mesure = \frac{(1 + \beta^2) * P * R}{(\beta^2 * P) + R} \tag{1.7}$$

Le paramètre β permet de pondérer la précision ou le rappel. β est généralement égal à 1.

1.8 Conclusion

Ce chapitre a été dédié à la présentation des principaux concepts de base de la RI classique. Nous avons aussi présenté le processus de RI, les étapes qui le composent ainsi que les mesures d'évaluation des résultats de la recherche. Nous avons cité certains des modèles de RI ainsi que l'impact de l'émergence du web dans leur exploitation. A cet effet, les résultats restitués par les SRIs classiques, tenant compte du volume et de l'hétérogénéité des ressources disponibles, sont devenus de moins en moins satisfaisants. Ce constat a amené l'émergence de la RI adaptative suivie de la RI personnalisée. La présentation de ces deux générations de SRI fera l'objet de la première partie du chapitre suivant (ch. 2). Ce chapitre a pour objectif de présenter la personnalisation de la recherche d'information.

Chapitre 2

Personnalisation de la recherche d'information

2.1 Introduction

Dans un contexte de continuel accroissement des sources d'information hétérogènes et de la diversité des utilisateurs, 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 à l'émergence de la RI personnalisée. Ce domaine apporte principalement la prise en compte de l'utilisateur en tant que composante principale dans le processus de la recherche.

Nous présentons dans ce chapitre la personnalisation de la recherche d'information. Nous commençons, dans la section 2.2 par présenter les évolutions qui ont conduit à l'émergence de la RI personnalisée. Dans la section 2.3, nous abordons les notions de base pour la personnalisation de la recherche d'information. Nous présentons ensuite, dans la section 2.4, les approches de représentation, de construction et d'évolution des modèles des utilisateurs. Les approches de personnalisations sont présentées dans la section 2.5. Nous terminons ce chapitre par une section de synthèse, section 2.6, et par conclure.

2.2 De la recherche d'information classique à la recherche d'information personnalisée

La recherche d'information classique est principalement basée sur l'appariement document requête, ce qui implique que 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. Or, la différence du vocabulaire de l'utilisateur et de l'auteur d'un document peut aboutir à une disparité entre les termes utilisés ou encore à l'existence d'ambiguïtés. Ainsi, un document ne contenant pas les mêmes termes utilisés dans une requête risque fort ne pas être retourné à l'utilisateur même s'il est pertinent. En effet, l'appariement document-requête dans les systèmes de recherche d'information classiques ne considère que les termes exacts et non ceux similaires.

De plus, pour une même requête soumise par deux utilisateurs distincts et présentant des besoins différents, le système de recherche d'information retournera les mêmes résultats de recherche. De ce fait, les performances d'un système de recherche d'information ne seraient plus uniquement dépendantes de l'indexation des documents et de l'appariement documents-requêtes mais aussi de sa capacité à prendre en compte les besoins de l'utilisateur. D'autres facteurs tels que les requêtes utilisateurs courtes, une formulation vague des besoins ou les volumes d'information de plus en plus importants viennent appuyer le manque de performance des SRI face à la prise en compte du besoin de l'utilisateur. Par exemple, la requête "article de loi" est une requête assez vague mais qui devrait retourner des résultats concernant les articles de loi du pays dans lequel est localisé l'utilisateur ayant formulé cette requête. Un autre exemple présentant une ambiguïté assez classique est la requête "Apple". En effet cette requête est liée à plusieurs thématiques de recherche, telles que le constructeur d'ordinateurs Apple ou le fruit lui-même. Pour ce type de requêtes, un SRI devrait renvoyer des résultats liés aux besoins de l'utilisateur lors de la recherche. 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 des utilisa-

teurs. En effet, la RI adaptative 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, de désambiguïsation du sens des mots de la requête.

Cependant, les systèmes de RI adaptative présentent des limitations. Ces limites sont principalement liées à la représentation limitée de l'utilisateur et de la notion de contexte utilisateur. En effet, lors d'une recherche, le contexte est généralement limité au besoin informationnel extrait des termes de la requête soumise. Des informations telles que la tâche courante de l'utilisateur, ses précédentes recherches, sa situation géographique ou autres informations utiles sont rarement utilisées lors de l'interprétation de la requête de l'utilisateur. De ce second constat, est apparue l'évolution de la RI adaptative vers la personnalisation de la RI qui reprend les techniques de la RI adaptative et la prise en compte de l'utilisateur en tant que composante principale dans le processus de la recherche.

2.3 Notions de base pour la personnalisation de la recherche d'information

Les éléments communs à tous les systèmes d'accès personnalisé à l'information incluent : (Mobasher et al., 2000)

- le traitement préalable des données de recherche,
- l'extraction des relations existantes entre ces différents types de données de recherche, et
- la détermination des actions que le système de personnalisation devra effectuer.

Ces trois éléments conduisent à la notion de contexte de la recherche. Cette notion, présentée dans la section 2.3.1, englobe aussi les informations sur les préférences, les besoins en information ou encore l'environnement de recherche de l'utilisateur. L'ensemble de ces informations est appelé le contexte de l'utilisateur ou, plus spécifiquement, le modèle de l'utilisateur, présenté dans la section 2.3.2.

2.3.1 La notion de Contexte

Dey et Abowd, (Dey and Abowd, 2000), ont effectué une revue des différentes définitions du contexte de laquelle découle leur propre définition : "le contexte est toute information pouvant être utilisée pour caractériser la situation des entités (par exemple, une personne, un lieu ou un objet) qui sont jugées pertinentes pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et l'application eux-mêmes". Cette définition correspond à la première perspective de Dourish. En effet, selon Dourish, (Dourish, 2004), le contexte peut être défini suivant deux perspectives : tel un problème de représentation, ou comme un problème d'interaction. Dans la première perspective, le contexte est considéré comme une "forme d'information" délimitée, stable et indépendante de l'activité. Ce contexte se compose alors d'attributs implicites décrivant aussi bien l'utilisateur que l'environnement de recherche d'information. La seconde perspective voit le contexte comme découlant des activités de l'utilisateur. D'autres définitions de contexte correspondant à la première perspective de Dourish sont présentées, (Lopes, 2009) :

- Marchionini, (Marchionini, 1995) : le contexte est défini comme un "cadre" qui a "des composants physiques et conceptuels/sociales, ainsi que l'information si la tâche se fait en collaboration ou non ainsi que les états physique et psychologique de l'utilisateur".
- Goker et Myrhaug, (Göker and Myrhaug, 2002) : "la description des aspects d'une situation",
- Johnson, (Johnson, 2003) : le contexte est défini comme une relation entre les informations spécifiques et les processus. Cette définition est plus proche de la seconde perspective de Dourish.
- Sato, (Sato, 2004) définit le contexte comme "un modèle de comportement ou de relations entre les variables qui sont en dehors des sujets de conception et qui affectent potentiellement le comportement des utilisateurs et les performances du système".
- Ingwersen et Jarvelin, (Ingwersen and Ja"rvelin, 2005), affirment que "les acteurs et leurs composants agissent en tant que contexte les uns pour les autres dans les processus d'interaction. Il y a des contextes sociaux, organisationnels, culturels ainsi que systémiques. Tous ces contextes évoluent au fil du temps.

Par ailleurs, des niveaux contextuels considérés plus significatifs en RI ont été présentés (Cool and Spink, 2002). Ces niveaux ont pour objectif de dissocier les entités qui interviennent dans le processus de recherche :

Niveau(1) : Environnement de recherche : sont considérés dans ce niveau les facteurs cognitifs, sociaux ou professionnels. Ces facteurs ont une influence sur le comportement de recherche de l'utilisateur.

Niveau(2) : Connaissances de l'utilisateur : ce niveau prend en compte les buts et les intentions de recherche de l'utilisateur.

Niveau(3) : Interaction utilisateur-système : Mise en évidence de l'impact de l'environnement sur la rétroaction ou les jugements de pertinence de l'utilisateur.

Niveau(4) : Niveau de requête du contexte : dans ce niveau, la performance du SRI est explorée dans l'interprétation des requêtes des utilisateurs.

Dans leur proposition, Ingwersen et Jarvelin, (Ingwersen and Ja"rvelin, 2005), ont décomposé le contexte en classes dépendantes les unes des autres, les plus significatives sont :

- La dimension des tâches de travail : couvre la tâche de travail établie par l'organisation, l'organisation sociale du travail, la collaboration entre les acteurs, l'environnement physique et le système.
- La dimension utilisateur : couvre les connaissances déclarées de l'utilisateur ainsi que d'autres caractéristiques personnelles, telles que son besoin en information.
- La dimension des tâches de recherche : couvre les pratiques nécessaires à la recherche et à l'extraction de l'information.
- La dimension perception de la tâche de travail : cette dimension concerne la perception qu'a l'utilisateur de la tâche de travail à exécuter.
- La dimension perception de la tâche de recherche inclut, en plus de la perception de l'utilisateur, les types de besoin d'informations concernant la tâche de recherche et le processus de performance associé.
- La dimension caractéristiques de système : couvre la représentation des documents ou des besoins d'information et de l'information elle-même. Cette dimension couvre également les outils et le soutien pour la formulation des requêtes et des méthodes

pour faire correspondre les documents et les représentations de la requête.

- La dimension des caractéristiques d'interaction. Cette dimension prend en compte les stratégies d'accès aux informations, l'interaction entre l'utilisateur et l'interface du système.
- La dimension couvrant les caractéristiques de la collection de documents (genre des documents, etc.).

Les points communs aux différentes définitions données du contexte sont principalement le besoin en information, les centres d'intérêts et l'interaction de l'utilisateur avec le système. Nous introduisons dans ce qui suit la notion de modèle de l'utilisateur, plus connu sous profil utilisateur.

2.3.2 Profil-Modèle utilisateur

La prise en compte des données de l'utilisateur est au coeur du processus de personnalisation RI. A cet effet, les recherches dans le domaine de la modélisation de l'utilisateur se sont principalement concentrées sur la définition d'approches servant à cette modélisation. On trouve dans la littérature plusieurs définitions de la modélisation utilisateur, en voici quelques-unes :

> *" A user model is a knowledge source in a natural-language dialogue system which contains explicit assumptions on all aspects of the user that may be relevant for the dialogue behavior of the system." (Wahlster and Kobsa, 1986)*

> *" User model is an explicit representation of the system of a particular user's characteristics that may be relevant for personalized interaction." (Razmerita, 2009)*

Le concept de modèle utilisateur, d'abord introduit dans les travaux de filtrage d'information (Belkin and Croft, 1992), a été ensuite exploité en RI personnalisée afin de constituer les composantes du contexte de l'utilisateur, à savoir les centres d'intérêts, les préférences, les besoins en informations de l'utilisateur (Bouzeghoub and Kostadinov, 2005). Le parcours de la littérature nous a permis de constater l'utilisation des termes "modèle utilisateur" ou "profil utilisateur". Ces deux termes sont généralement utilisés comme synonymes. Néanmoins, d'après Koch [KOC 00], il y a une différence entre ces deux termes :

le profil est une version simplifiée du modèle. En effet, il considère que le profil utilisateur est utilisé pour construire le modèle utilisateur. Ce dernier est de ce fait vu comme une représentation de ce que le système perçoit sur un utilisateur donné. Avant de écrire et de présenter plus en détail la modélisation de l'utilisateur, il est important de signaler que, lors de ce travail, nous utiliserons le terme "modèle de l'utilisateur" pour la représentation des données de l'utilisateur et le terme "profil de l'utilisateur" comme une partie du modèle. Plus de détails sur la modélisation de l'utilisateur sont donnés dans la section 2.4.

2.3.3 Personnalisation versus Adaptation

Au cours des interactions entre l'utilisateur et le système exploité, trois formes de changements de ce dernier pourraient avoir lieu : changement de l'apparence des interfaces, changement du mode d'interaction ou ajustement du système à l'utilisateur.
La première forme de changement permet à l'utilisateur de modifier l'apparence ou le contenu du système qui de ce fait se dit personnalisable ou encore adaptable. La seconde forme porte sur la modification automatique de l'interaction homme-machine par rapport aux préférences de l'utilisateur. Le système est alors dit système personnalisé ou adaptatif. Et la dernière forme est représentée par la capacité du système à s'ajuster en fonction du contexte d'utilisation. Il est alors dit système "sensible au contexte", mieux connu sous système "context aware".
Dans le cadre de notre thèse, nous abordons la seconde forme relative à la personnalisation et l'adaptation.
L'adaptation est définie par Dietrich et al. (Dieterich et al., 1993) comme "une tentative de modifier le comportement interactif d'un système en considérant à la fois les besoins individuels des utilisateurs humains et les conditions propres à l'environnement de l'application". L'adaptation prend éventuellement en charge les modifications techniques nécessaires à l'environnement matériel du système. Nous pouvons déduire de cette présentation que la personnalisation est liée à l'adaptation et qu'elle pourrait en représenter une sous-catégorie. De ce fait une personnalisation peut être de l'adaptation mais une adaptation n'offre pas forcément de la personnalisation.
La personnalisation est définie comme un processus qui change les fonctionnalités, l'inter-

face, le contenu ou l'aspect d'un système dans le but d'améliorer sa pertinence et cela en fonction caractéristiques l'utilisateur et/ou de ses navigations. En effet un système personnalisé tente de deviner quelles pourraient être les préférences de l'utilisateur (Kim and Allen, 2002). (Bazsalicza and Naim, 2001) définissent la personnalisation comme la capacité d'un système à produire des ressources en fonction de l'identité du demandeur.
L'utilisateur fait partie du processus de personnalisation dans lequel il peut décider de choisir parmi les résultats mis à sa disposition ceux qui répondent à son besoin. La personnalisation peut de ce fait être définie comme un apprentissage réalisé à partir des préférences retournées par les utilisateurs à l'issue de la présentation des résultats successifs du système.

2.4 La modélisation de l'utilisateur

En 1996, (Höök et al., 1996) a présenté le modèle de l'utilisateur comme étant "une connaissance à propos de l'utilisateur explicitement ou implicitement codée, utilisée par le système afin d'améliorer son interaction". Ce modèle contient des informations sur les buts, les besoins, les préférences ou les intentions des utilisateurs. De ce fait, modéliser l'utilisateur, ses centres d'intérêts et ses préférences est une tâche primordiale pour les concepteurs des systèmes de RI personnalisée. Construire un modèle de l'utilisateur nécessite la connaissance de la structure à mettre en place afin de stocker les informations le concernant et de les exploiter d'une manière optimale. Plusieurs approches et techniques ont été développées afin de modéliser l'utilisateur. Ces techniques diffèrent par leur représentation, construction et mise à jour du contenu du modèle de l'utilisateur. Ce dernier est aussi fortement dépendant du système dans lequel il évolue. En effet, généralement, les données exploitées par le système déterminent le contenu du modèle.

L'acquisition des données pour la construction du modèle de l'utilisateur peut être elle aussi explicite ou implicite. La première approche exploite des informations issues et fournies par l'utilisateur lui-même. Ces données peuvent être :

- Des jugements de pertinence pour les documents visités
- Des mots clés représentatifs des centres d'intérêts de l'utilisateur

– La sélection de liens ou de thèmes favoris

Ce mode d'acquisition est caractérisé par la simplicité de sa mise en place. Il présente le principal inconvénient d'être couteux en temps d'exécution. Un second inconvénient est le décalage pouvant exister entre l'intention de l'utilisateur et ce qu'il désire réellement au moment de sa recherche.

Quant à l'acquisition implicite de données utilisateur, elle a pour objectif principal de décharger ce dernier de la contrainte de saisie d'information sur ses centres d'intérêts. Cette acquisition est effectuée au cours de la navigation et des interactions de l'utilisateur avec le système de recherche. Les comportements de navigation sont observés, enregistrés et ensuite traités. Parmi ces traces, nous citons :

– L'historique des visites ainsi que leur fréquence : ici, le temps passé sur une page visitée peut aussi être pris en compte. Il est à noter que ce temps est généralement comptabilisé au-delà d'un seuil minimum.
– Les cliques et les défilements de la souris sur les différentes pages parcourues. L'ordre d'impression, de sauvegarde ou de copie d'un document peuvent aussi intervenir.

Cette méthode implicite présente le principal avantage d'être transparente pour l'utilisateur mais a néanmoins certains inconvénients : elle peut être moins précise que l'évaluation explicite par l'utilisateur mais cependant beaucoup plus dynamique dans le temps.

La surabondance de l'information sur le Web, ayant remis en cause les modèles classique de la RI, présente également un souci d'abondance des données fournies à l'utilisateur en réponse à une requête. Modéliser l'utilisateur permet de ce fait de mieux cibler les données fournies en fonction des intérêts de ce dernier ainsi que de ses besoins.

2.4.1 Approches de représentation du modèle de l'utilisateur

Le modèle de l'utilisateur peut être représenté en intégrant la sémantique ou encore en utilisant un formalisme de pondération vectorielle. Dans cette section, nous présentons les principales représentations des modèles utilisateurs.

Représentation ensembliste

La modélisation de l'utilisateur de manière ensembliste a pour objectif de formaliser les données du modèle en un ensemble de termes pondérés ou en vecteurs de termes pondérés (Lieberman, 1998) ou en classes de vecteurs de termes pondérés (Gowan, 2003) souvent représentés selon le modèle vectoriel de Salton (Salton and Yang, 1973). Ces termes traduisent les centres d'intérêt de l'utilisateur. Le poids d'un terme est souvent calculé selon le schéma TF*IDF.
Plusieurs systèmes d'accès personnalisé à l'information utilisent ce type de représentation tels que Fab (Balabanovic and Shoham, 1997), système de recommandation de page web et Letizia (Lieberman, 1998), système d'aide à la navigation, etc. Le moteur de recherche Google personalized search version 1.1, moteur qui n'est plus disponible actuellement, utilisait un ensemble de catégories représentées par un terme ou un ensemble de termes explicitement saisis par l'utilisateur. Quant à Yahoo contextual search (Y !Q), il utilise un ensemble de termes issus d'une sélection à partir de la page courante visualisée. Dans (Zemirli, 2008), l'approche présentée est basée sur une modélisation ensembliste avec une représentation vectorielle des centres d'intérêt de l'utilisateur par mots clefs pondérés. La session de recherche de l'utilisateur est représentée par une matrice document-termes avec leurs poids associés. L'historique des interactions de l'utilisateur est quant à lui représenté par une matrice agrégée des matrices des sessions de recherche. A partir de ces données, un vecteur de termes pondérés représentant le contexte courant est extrait, il représente le centre d'intérêt courant de l'utilisateur. Dans leur système de recherche d'information utilisant les données spatiales et temporelles, (Hinze et al., 2009) construisent le modèle de utilisateur sous la forme d'un ensemble de termes et/ou de couples attribut/valeur.
Ceci est assez simple à mettre en oeuvre mais présente l'inconvénient d'un manque de structuration et de l'absence de gestion des différents niveaux de généralités pouvant caractériser l'utilisateur. Structuration que peut apporter la modélisation connexionniste.

Représentation connexionniste

La modélisation de l'utilisateur par représentation connexionniste représente les centres d'intérêts par un réseau de noeuds (concepts du centre d'intérêt) pondérés. Les relations de dépendance sémantique entre les centres d'intérêt du modèle de l'utilisateur établies dans

cette représentation permettent de résoudre les failles de la représentation ensembliste. 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. Plusieurs SRI personnalisés adoptent ce type de représentation (Gentili et al., 2003), (Micarelli and Sciarrone, 2004), (Koutrika and Ioannidis, 2005).
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.

Représentation conceptuelle

La représentation conceptuelle est basée sur l'exploitation des ontologies de domaine ainsi que sur des hiérarchies de concepts (Gauch et al., 2003), (Liu et al., 2004), (Sieg et al., 2007b). Les centres d'intérêt de l'utilisateur sont représentés sous forme de réseau de noeuds conceptuels reliés entre eux en suivant la topologie des liens définis dans les hiérarchies et les ontologies de domaines (Daoud, 2009). Chacun des concepts est représenté par un vecteur de termes pondérés, la pondération traduit le degré d'intérêt de l'utilisateur.
La représentation conceptuelle peut être considérée semblable à la représentation sémantique de par le fait que les centres d'intérêts de l'utilisateur sont représentés par un réseau de noeuds conceptuels. De même, cette représentation peut être assimilée à une approche ensembliste. En effet, ceci revient au fait que les domaines y sont généralement représentés en vecteurs de termes pondérés (Zemirli, 2008).
Deux sources d'information sont utilisées afin de construire le modèle utilisateur : les données des navigations utilisateur collectées au cours de ses recherches et les ressources sémantiques prédéfinies. La construction du modèle de l'utilisateur commence par la spécification des niveaux de concepts de l'ontologie de domaine utilisée. L'étape suivante est l'application d'un processus de déploiement des données dans des techniques de pondération de ces concepts (Daoud, 2009).

Représentation sémantique

L'intégration de la sémantique a pour principal objectif de contourner les ambigüités sémantiques des mots. L'obtention d'informations sémantiques pour la représentation d'un modèle de l'utilisateur peut être effectuée de différentes manières :

- Association à un mot clef un contexte constitué d'autres mots (Marinilli et al., 1999). Mais une telle approche nécessite la construction de contextes.
- Utilisation d'un ensemble de documents servant à construire des relations inter-documents. Ceci a pour objectif de conserver les mots dans leur contexte documentaire (Billsus and Pazzani, 1999) ; (Cotter and Smyth, 2000).
- La troisième manière et la principale se base sur la mise des préférences utilisateurs sous forme d'une ontologie (Blanco-Fernandez et al., 2008) ; (Dolog and Nejdl, 2007) ; (Jrad et al., 2007) ; (Baldwin et al., 2000) ; (Di Lascio et al., 1999).

Par ailleurs, un système pour la modélisation des utilisateurs à base d'ontologie, OntobUMf (Razmerita, 2008), a été proposé. L'acquisition de données utilisateur est effectuée de manière explicite à travers un éditeur de profil, ou en utilisant différentes techniques de modélisation de l'utilisateur (techniques à bases d'heuristiques et de logique floue). Le modèle de l'utilisateur est sous forme d'ontologie générique. Cette ontologie utilisateur contient diverses caractéristiques d'un utilisateur. Ces caractéristiques sont présentées à base de concepts, sous concepts et relations taxonomiques et non taxonomiques entre les différents concepts. OntobUMf étant dans le domaine du e-learning, l'ontologie utilisateur a été conceptualisée à partir de la spécification "Information Management System Learner Information Package" . Le schéma de la figure 2.1 présente les techniques de modélisation et les mécanismes de personnalisation qui sont sous forme de services intelligents.

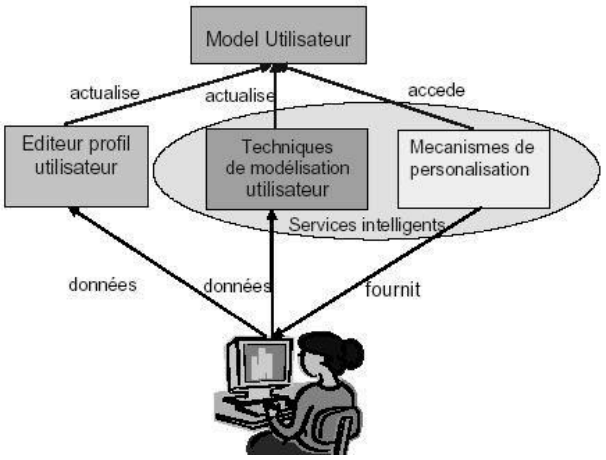


FIGURE 2.1 – Modélisation utilisateur et mécanismes de personnalisation (Razmerita, 2008)

Dans le contexte de systèmes basés sur la localisation spatiale, (Tryfona and Pfoser, 2005) construisent un modèle de l'utilisateur sous la forme d'ontologie pour laquelle ils exploitent, en plus des données utilisateur, les informations de l'appareil mobile accèdant au système. Des rôles de l'utilisateur sont associés aux mots clefs, et sont ensuite associés aux descripteurs de services du système afin de faciliter leux exploitation automatique.

Représentation multidimensionnelle

Un troisième type de modélisation se propose de structurer le profil selon un ensemble de dimensions représentées selon divers formalismes : c'est la modélisation multidimensionnelle (Bouzeghoub and Kostadinov, 2005), (Amato and Straccia, 1999).. Elle présente comme points forts de mieux interpréter la sémantique du profil et d'être indépendante de tout type d'application.

Amato (Amato and Straccia, 1999), dans le cadre du développement d'un service de bibliothèque numérique - système EUROgather service-, représente le modèle de l'utilisateur par des dimensions, qu'ils ont aussi qualifiés de catégories, prédéfinies. C'est la première approche qui offre un modèle général et dont les informations sont structurées. Les catégories sont au nombre de cinq :

- 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 :

contenu (des informations sur le sujet du document, la langue, etc.) structure, (format, type, date de publication, dimensions, etc.), 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 : moyen (mode de livraison par exemple email, fax téléphone, etc.) et moment (contient des informations temporelles sur le moment de livraison comme lors d'un changement, vers midi, entre 9h et 9h15, 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 sont données sur les conditions d'accès aux données du profil.

Poursuivant la classification de (Amato and Straccia, 1999), (Kostadinov et al., 2007) propose un ensemble de dimensions ouvertes où il distingue les cinq dimensions suivantes :

- Le domaine d'intérêt : regroupe tous les attributs qui concernent les objets de contenu (informations ciblées). Cette dimension peut aussi bien définir le domaine d'expertise et le niveau de qualification de l'utilisateur que le contenu auquel il s'intéresse.
- Les données personnelles : ces données représentent la partie statique du profil et contiennent des informations qui décrivent l'utilisateur et ne dépendent pas du système à interroger
- La qualité : ces données décrivent la qualité attendue l'utilisateur.
- Les données de livraison : ces données concernent tout ce qui est lié à la présentation des résultats en fonction de la plateforme de l'utilisateur.
- Les données de sécurité : La sécurité est une dimension fondamentale du profil. Elle peut concerner les données que l'on interroge ou modifie, les informations que l'on calcule, les requêtes utilisateurs ou les autres dimensions du profil. Cette représentation du modèle de l'utilisateur est appliquée au domaine des bases de données.

(Fink and Kobsa, 2002) ont proposé un système générique de modélisation de l'utilisateur pour la personnalisation des services du tourisme. Ce système analyse les comportements de l'utilisateur et effectue des prédictions quant aux actions futures de ce dernier.

(Bohnert, 2008), quant à eux, incluent dans leur système de visite de musées des approches de modélisation collaborative et basées sur le contenu afin de modéliser les intérêts utilisateur.

2.4.2 Construction du modèle de l'utilisateur

La construction du modèle de l'utilisateur repose sur deux phases. La première phase est la collecte d'informations sur l'utilisateur et la seconde est la construction proprement dite du modèle. Vient ensuite une phase d'évolution du modèle. Nous présentons dans cette section les techniques d'acquisition des données utilisateur ainsi que les techniques de construction du modèle. Ensuite nous introduisons la notion de gestion de l'évolution du modèle de l'utilisateur.

Acquisition des données utilisateurs

La phase d'acquisition des données utilisateur consiste à collecter les informations qui serviront à l'instanciation du modèle de l'utilisateur. Nous distinguons deux modes d'acquisition : acquisition explicite et acquisition implicite. Dans l'approche explicite, les informations sont directement obtenues de l'utilisateur. Dans l'approche implicite, ce sont les données de comportement de l'utilisateur qui sont exploitées. Nous détaillons ces deux approches dans ce qui suit.

- **Acquisition explicite :**

  Pour cette technique, l'utilisateur est directement interrogé à travers des formulaires à remplir, par exemple. Ce dernier peut aussi être amené à émettre un jugement d'intérêt en fournissant une valeur sur une échelle graduée. L'acquisition explicite a été largement utilisée pour la personnalisation des interfaces des sites web en fonction des préférences des utilisateurs.

  L'acquisition explicite repose essentiellement sur les techniques de feedback explicite. Ces techniques sont utilisées dans les systèmes de filtrage et de reformulation de requêtes par réinjection de pertinence. Ce type d'acquisition est utilisé dans plusieurs systèmes de RI personnalisée en ligne. Dans Google personalized search version 1.1, l'utilisateur peut saisir un ensemble de catégories représentatives de ses centres d'in-

térêt. Alors que dans yahoo contextual search, (Kraft et al., 2005), l'utilisateur peut sélectionner dans la page courante un texte qui servira à représenter un profil textuel lié à la recherche en cours. L'approche dans (Sieg et al., 2004b) sollicite de l'utilisateur une sélection d'une paire de concepts (adéquat, inadéquat) à travers l'ensemble des concepts de l'ontologie de l'ODP. Cette approche est utilisée dans un processus de reformulation personnalisée de la requête. Koutrika et Ioannidis (Koutrika and Ioannidis, 2005), ainsi que le système ifWeb (Asnicar and Tasso, 1997), acquièrent les données utilisateur en utilisant un feedback explicite.

Ce procédé d'acquisition présente le principal inconvénient de pouvoir induire au désintéressement de l'utilisateur lors de la saisie de données (Wu et al., 2008). De plus, la construction du modèle de l'utilisateur étant fortement liée au mode d'acquisition de données, elle se retrouve fortement liée au degré d'implication de l'utilisateur. En effet, si l'utilisateur ne fournit pas les informations, aucun modèle ne sera construit. Ces limitations ont mené vers l'utilisation des techniques d'acquisition implicite des données utilisateur pour la construction du modèle de l'utilisateur.

– **Acquisition implicite :**

Dans cette technique de collecte de données, l'utilisateur n'est pas directement impliqué. 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. Les activités que peut avoir un utilisateur peuvent correspondre à l'utilisation de moteurs de recherche à travers la soumission de requêtes et la sélection de documents, la navigation sur le web ou encore l'utilisation d'applications dans le contexte de sa recherche (applications de bureautique, éditeurs de texte, ...).

Une fois la collecte de données effectuée, le modèle de l'utilisateur est dérivé. Plusieurs techniques de dérivations existent telles que les techniques de data mining sur l'historique de recherche de l'utilisateur (Eirinaki and Vazirgiannis, 2003) ou encore l'utilisation de processus d'apprentissage des données utilisateurs (Webb et al., 2001). Des systèmes tels que Letizia (Lieberman, 1995), WebMate (Chen and Sycara, 1998) et Personal WebWatcher (Mladenic, 1999) utilisent les pages web visitées pour la construction du modèle de l'utilisateur. D'autres sources implicites d'information

sont aussi exploitées telles que les requêtes et leur association aux résultats de recherche dans (Rich, 1998) et Syskill and Webert (Pazzani et al., 1996) ou les requêtes et les résumés textuels des résultats associés (Shen et al., 2005). Certains combinent même les sources d'information en utilisant à la fois les pages web, les emails et les documents textuels (Dumais et al., 2003).

Cette approche présente le principal avantage de ne pas impliquer l'utilisateur et de ne pas lui imposer l'émission de jugements, ni un effort d'attention particulier lors de sa recherche. Néanmoins, la difficulté dans cette techniques est de définir un processus d'interprétation du comportement observé dans un contexte d'application spécifique (Daoud, 2009).

Techniques de construction

Les modèles des utilisateurs sont construits en utilisant les sources d'information utilisateur. A cet effet, une variété de techniques de construction basées sur l'apprentissage ou sur l'extraction d'information est utilisée (Gauch et al., 2007). Nous écrivons dans cette section les techniques les plus courantes utilisées dans la construction de modèles à base de mots-clés, de réseaux sémantiques ou de concepts.

- **Extraction d'ensemble de termes :** La technique d'extraction d'ensemble de termes est basée sur des techniques d'analyse statistique de mots clés. En effet, le contenu des documents visités par l'utilisateur est analysé pour en extraire les mots clés significatifs. Ces derniers vont servir dans l'algorithme d'apprentissage du modèle de l'utilisateur. Par exemple, dans le cadre d'une approche vectorielle, les termes extraits sont pondérés dans l'objectif de former des vecteurs de termes représentant les centres d'intérêts de l'utilisateur. Des systèmes tels que WebMate (Chen and Sycara, 1998) et Alipes (Widyantoro et al., 1999), appliquent cette approche de construction.
- **Extraction de réseaux de termes :** Similairement à la technique d'extraction d'ensemble de termes, les termes sont extraits des documents jugés par l'utilisateur. Cependant, la différence réside dans la représentation des termes qui est sous forme de réseau de noeuds. Cette technique concerne principalement les représentations

sémantiques du modèle de l'utilisateur. Pour construire le modèle 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. Des systèmes tels que SiteIF (Stefani, 1998) utilisent cette technique.

– **Extraction de concepts :** Cette technique de construction concerne principalement les modèles d'utilisateurs représentés par une hiérarchie de concepts pondérés. Le principe de base de cette technique est l'utilisation d'une taxonomie de concepts de référence comme profil de base. Cette dernière peut être aussi bien l'ODP du projet Open Directory Project [5], un annuaire de concepts hiérarchique open source ou encore la hiérarchie de concepts Yahoo [6]. L'approche de construction présente de manière générale les étapes deux suivantes : (1) identification des concepts et niveaux de l'ontologie à exploiter et (2) extraction des centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie. Des systèmes tels que Persona (Tanudjaja and Mui, 2002) (approche de coloration d'arbre), ARCH (Sieg et al., 2004a) (approche hybride combinant vecteurs de termes et hiérarchie de concepts) et le système du projet OBIWAN (association des documents collectés avec les noeuds de l'ontologie.

2.4.3 Evolution du modèle de l'utilisateur

Une des caractéristiques des systèmes de personnalisation est la gestion de l'évolution du modèle de l'utilisateur. Cette gestion consiste en premier lieu à capturer les transformations des centres d'intérêt de l'utilisateur et en second lieu à propager ces changements au niveau de la représentation du modèle e l'utilisateur. L'évolution du modèle peut s'effectuer lors de sa modélisation par deux modèles, l'un à court terme et l'autre à long terme (Billsus and Pazzani, 1999) ou encore en appliquant des algorithmes d'évolution génétique.
L'adaptation du modèle de l'utilisateur implique d'avoir des changements dans les centres d'intérêts de l'utilisateur ce qui peut comprendre aussi la suppression ou l'ajout de domaines. En effet, faire évoluer le contenu du modèle de l'utilisateur consiste à adapter aussi bien la structure que le contenu aux évolutions des centres d'intérêts de l'utilisateur. Peu de travaux ont abordé cet aspect du modèle de l'utilisateur (Zemirli, 2008). En effet, l'évo-

5. http ://www.dmoz.org/
6. Yahoo. Yahoo directory

lution n'est considérée que quand le modèle de l'utilisateur a une existence permanente, ce qui n'est pas le cas dans la majorité des systèmes.
L'évolution du modèle de l'utilisateur dans le cas d'une représentation ensembliste est effectuée par l'ajout de nouveaux vecteurs de termes extraits des documents pertinents pour l'utilisateur. L'inconvénient majeur, dans ce cas, est l'augmentation du nombre de centres d'intérêt due à l'absence de dépendance entre les vecteurs. Une approche novatrice consiste à utiliser théorie de la vie artificielle pour la construction et la mise à jour du contenu du modèle (Chen et al., 2001). Dans cette approche, les classes du modèle de l'utilisateur sont enrichies par une valeur d'énergie qui traduit le degré d'importance d'un centre d'intérêt par rapport à un autre.

2.5 Les approches de personnalisation

Afin de présenter les approches de personnalisation, nous les avons regroupé en trois grandes catégories : les approches basées sur les différents types de filtrage, les approches basées sur la recherche textuelle et finalement les approches hybrides. Ces approches sont utilisées aussi bien dans la RI classique que dans la RI sur le web. L'utilisateur faisant partie intégrante de la personnalisation, nous n'aborderons dans cette section que les données de l'utilisateur utilisées. La représentation de ce dernier fera l'objet de la section 2.4.

2.5.1 Le filtrage collaboratif

La technique de filtrage collaboratif regroupe l'ensemble des méthodes qui visent à la construction de systèmes de personnalisation utilisant les opinions, les évaluations et les navigations d'un groupe pour aider l'individu (Goldberg et al., 2001). Le filtrage collaboratif est perçu comme un processus qui exploite les corrélations et les similarités entre les utilisateurs ou entre leurs navigations. Selon la nature des informations étudiées, les approches de filtrage collaboratif peuvent être classées en deux catégories : le filtrage collaboratif utilisateur et le filtrage collaboratif Objet.
Le filtrage collaboratif utilisateur : les systèmes à filtrage collaboratif utilisateur effectuent une classification des utilisateurs selon leurs comportements et selon leurs votes. Les techniques de classification utilisées sont les réseaux de neurones, les réseaux de Bayesiens

ou des algorithmes de clustering comme le k-means ou fuzzy mean. Leur objectif étant de construire des groupes homogènes d'utilisateurs (Tan et al., 2005). Une association de l'utilisateur courant à une classe d'utilisateurs qui ont le même comportement vis-à-vis des ressources est ainsi faite. Le système à filtrage collaboratif recommande les items les plus appropriés à cette classe d'utilisateurs. Les principaux avantages du filtrage collaboratif sont l'aide apportée à l'utilisateur lors de la découverte de nouveaux items jugés intéressant par la communauté à laquelle il est associé et une favorisation de la navigation de l'utilisateur à travers les items résultats de recherche. L'inconvénient majeur est le regroupement général des items. En effet, les préférences de l'utilisateur sont considérées par rapport à la communauté et non pas par rapport à lui-même.
Le filtrage collaboratif Objet : les systèmes de filtrage collaboratif objet ont été popularisés par Amazon (Linden et al., 2003) avec la fonctionnalité "les gens qui ont acheté x ont aussi acheté y". Cette technique utilise les algorithmes de data mining (Tan et al., 2005) afin d'extraire des informations concernant les corrélations entre les items. Le processus consiste bâtir une matrice "item-item" déterminant des relations entre des objets (pairs) et à ensuite utiliser cette matrice pour proposer des objets. Cette technique est adoptée par plusieurs sites d'e-commerce. L'inconvénient majeur de ces systèmes est le fait que l'article est recommandé en se basant sur sa corrélation avec un article x et ne prend donc pas en considération les préférences de l'utilisateur.

Dans le cadre de la personnalisation prenant en compte les données spatiales, le filtrage collaboratif est de plus en plus exploité, notamment dans les services géolocalisés, appelés "location based systems" (LBS). Par exemple, (Tahir et al., 2010) construisent des profils de groupe en se basant sur l'interaction de l'utilisateur avec les objets d'une carte ainsi que sur la proximité géographique entre ces objets. (Gupta and Lee, 2010) exploitent la localisation comme point de départ pour la recommandation collaborative.

2.5.2 Le filtrage à base de contenu

Les approches de filtrage à base de contenu se basent sur la similarité des documents visités par l'utilisateur. La similarité est prise en compte en effectuant un rapprochement

des centres d'intérêt des utilisateurs (explicites ou implicites) avec les métadonnées des documents. Le filtrage à base de contenu ne considère pas les avis des autres utilisateurs (Peis et al., 2008). Deux fonctionnalités peuvent découler de cette approche :

- La sélection des documents pertinents par rapport à un profil ;
- La mise à jour du profil en utilisant le retour de pertinence explicite sur les documents reçus par l'utilisateur.

Le filtrage à base de contenu a commencé récement à faire l'objet de recherches dans le domaine de la personnalisation utilisant les données spatiales (Mac Aoidh et al., 2009), (Ballatore et al., 2010).En effet, L'acquisition implicite des données e navigation ayant pour objectif de recommander un contenu spatial aux utilisateurs se base principalement sur l'exploitation des cartes géographiques (Mac Aoidh et al., 2009).

2.5.3 Les approches de reformulation de la requête

Lors de la saisie de sa requête, l'utilisateur formule son intérêt dans son langage naturel. En effet, pour dire qu'il veut avoir des documents ou des liens concernant l'hôtellerie de luxe à un prix raisonnable, l'utilisateur aurait tendance à écrire la requête suivante : "hôtel de luxe et pas trop cher". Cette requête est certes claire mais présente des termes de liaison qui pourraient être enlevés : et, de. Reformuler une requête, dans ce contexte, a donc pour objectif de coordonner le langage de recherche naturel (utilisateur) et le langage d'indexation des systèmes de recherche d'information. Cela consiste à générer une nouvelle requête en utilisant d'autres données. Pour la personnalisation, les données utilisées sont principalement relatives à l'utilisateur. Il s'agit alors d'un enrichissement de la requête. Dans ce cadre, les données intégrées sont les éléments des centres d'intérêt ainsi que des préférences. (Gauch et al., 2003) ajoutent à la requête les préférences implicites de l'utilisateur. (Li and Zaiane, 2004) déduisent les différentes catégories des profils stockés. Dans leurs travaux en bases de données, (Koutrika and Ioannidis, 2004), (Koutrika and Ioannidis, 2005) utilisent le poids de prédicats pondérés qui exprime l'intérêt relatif pour l'utilisateur. (Messai et al., 2005) enrichissent les requêtes en leur ajoutant de nouvelles propriétés à partir des ontologies de domaine disponibles.

2.5.4 La personnalisation dans les systèmes géolocalisés

L'accroissement des données géo-référencées présentes dans le web ainsi que leur liaison avec des localisations géographiques ont conduit à l'intérêt de considérer les données spatiales dans le domaine de la recherche d'information dans le contexte du web (Winter and Tomko, 2006) . En effet, les approches de RI personnalisée classique, telles que le filtrage collaboratif ou celui basé sur le contenu, n'assument pas la complexité de l'information spatiale disponible sur le web. Des informations comme l'adresse IP ou un positionnement fourni par l'utilisateur lui-même sont, pour la plupart, reliées à un emplacement géographique (Buyukkokten et al., 1999). Ces spécifications spatiales peuvent en effet exister dans les requêtes des utilisateurs, et dénoter ainsi de leur intérêt (Silva et al., 2006). Certains définissent l'information géographique comme un ensemble de trois facettes : thème, espace et temps (Gaio, 2001).Les systèmes tels que SPIRIT (Vaid et al., 2005) travaillent en utilisant les adresses postales et gèrent, de ce fait, les relations spatiales exprimées explicitement dans des champs de l'interface utilisateur. Le système STEWARD (Lieberman et al., 2007), utilise deux gazetteers (GNIS - Geographic Names Information System pour les USA et GNS - GEOnet Names Serveur pour le reste du monde) mais n'a pas accès à des opérateurs spatiaux pour traiter des relations spatiales. GéoSem (Bilhaut et al., 2007), malgré l'absence de système d'information géographique, gère certaines relations spatiales (orientation, proximité) définies à base de règles. Ce qui n'est pas le cas de la plateforme PIV (Lesbegueries and Loustau, 2006) dans laquelle un système d'information géographique est utilisé. Le tableau 2.1 présente quelques uns de ces systèmes.

TABLE 2.1 – Présentation de quelques systèmes basés sur la localisation

GIS	Orientation	Presentation
CoMPASS (Weakliam et al., 2005)	Information mobile implicite, context-aware	Construit des modèles des préférences utilisateur. Les recommendations sont basées sur la localisation courante de l'utilisateur. Inclue le temps et l'intérêt utilisateur dans le contexte.
FLAME2008 (Weissenberg et al., 2004)	Information mobile explicite	Extraction explicite des préférences utilisateur. La personnalisation tient compte de la localisation spatiale de l'utilisateur
Genie (O'Grady and O'Hare, 2004)	Système de recommendation mobile	Fourni les informations sur les sites de loisirs pour les touristes.
GeoNotes (Petra, 2002)	Similaire au context Graffiti	Application context-aware. Utilise la navigation sociale.
Graffiti (Burrell and Gay, 2001)	Mobile	Permet à l'utilisateur de définir ce qui l'intéresse pour une localisation donnée en utilisant des notes électroniques.

2.6 Synthèse

Nous présentons dans ce paragraphe une synthèse de l'ensemble des systèmes de personnalisation abordés précédemment. Pour une meilleure lisibilité, nous les avons représentés dans le tableau 2.2. Nous présentons dans ce tableau pour chaque système le type de personnalisation offerte, le mode d'acquisition des données nécessaires au modèle de l'utilisateur, la représentation du modèle et finalement sa construction.

TABLE 2.2: Synthèse de l'état de l'art

Nom	**Type de personnalisation**	**Acquisition/Données du modèle de l'utilisateur**	**Représentation du M_u**	**Construction du M_u**
OntobUMf (Razmerita, 2008)		Explicite : éditeur de profil utilisateur	Sémantique : Ontologie	techniques à base d'heuristiques et de logique floue
Gecko (Bohnert, 2008)	Recommandation	implicite : tuple(item, durée de visite)	Hybride	Filtrage collaboratif et basé sur le contenu
Wifs (Micarelli and Sciarrone, 2004)	Filtrage	feedback explicite	réseau de termes	Gestion directe par l'utilisateur Création des noeuds par des experts mise à jour par raffinement
Infoweb (Gentili et al., 2003)	Filtrage Basé sur le contenu	Feedback explicite	réseau sémantique : concepts/termes	Ajustement utilisateur Création de concepts Ajustement des noeuds et arcs
WebSail (Chen and Shahabi, 2001)	Filtrage Basé sur le contenu	Feedback explicite (aime/n'aime pas)	Vecteur booléen de caractéristiques	
Let's Browse (Lieberman et al., 2001)	Filtrage Basé sur le contenu	Implicite (Liens parcourus, temps de lecture)	Vecteur de mots-clés pondérés	
Websift (Cooley et al., 1999)	Recommandation	Implicite : Historique navigation		Induction de règles, modèles, statistiques
Webmate (Chen and Sycara, 1998)	Recommandation	Implicite : historique de recherche	ensembliste : vecteurs de termes	Extraction d'ensemble de termes des pages web

La suite à la page suivante

TABLE 2.2 – suite de la page précédante

Nom	**Type de personnalisation**	**Acquisition/Données du modèle de l'utilisateur**	**Représentation du M_u**	**Construction du M_u**
SiteIF (Stefani, 1998)	filtrage Basé sur le contenu	Implicite : liens parcourus	réseau sémantique pondéré	Extraction des termes de plus fort poids Associer les termes aux concepts WordNet
Fab (Balabanovic and Shoham, 1997)	Recommandation de pages web	Explicite (évaluations)	ensembliste : VEP	extraction de termes à partir des documents pertinents
Amalthaea [Moukas, 1997]	filtrage Basé sur le contenu	Explicite : votes	Vecteur de mots-clés pondérés	
Syskill & Webert (Pazzani, 1999)	filtrage Basé sur le contenu	Explicite (évaluations)	Vecteurs de mots clés pondérés	
Personal WebWatcher (Mladenic, 1999)	Recommandation	implicite	Vecteur probabiliste de caractéristiques	Analyse es pages web visitées par l'utilisateur lors de sa recherche
ifWeb (Asnicar and Tasso, 1997)	assistant personnel à la navigation	feedback explicite	Connexionniste : réseau sémantique de termes pondérés	Construction d'un réseau de termes reliés par des arcs selon cooccurrence terme/document
WebWatcher (Armstrong et al., 1995), (Joachims et al., 1997)	Recommandation	Explicite (But atteint), Implicite (Liens parcourus)	Vecteur booléen de caractéristiques	
Letizia (Lieberman, 1998)	Aide à la navigation	Implicite : analyse de pages web.	ensembliste : VEP	extraction de termes à partir des de l'analyse des documents pertinents

La suite à la page suivante

TABLE 2.2 – suite de la page précédante

Nom	**Type de personnalisation**	**Acquisition/Données du modèle de l'utilisateur**	**Représentation du M_u**	**Construction du M_u**
GroupLens 1994	Filtrage Collaboratif	Explicite (évaluations), implicite (temps passé)	Couples (objet, note). Matrice d'estimations utilisateur/article.	
OBIWAN		Implicite	Conceptuelle	Analyse es pages web visitées par l'utilisateur lors de sa recherche

2.7 Conclusion

Nous avons présenté au cours de ce chapitre l'état de l'art des principaux axes de recherche de notre thèse. Nous avons en premier lieu abordé la recherche d'information , ses concepts, les modèles représentatifs et son évolution du classique vers le web. Nous avons ensuite présenté la personnalisation de la recherche d'information. Dans cette section nous avons abordé la modélisation de l'utilisateur ainsi que les approches de personnalisation.
Nous avons pu constater qu'une personnalisation efficace dans le contexte du web dépend de la représentation du modèle de l'utilisateur et de la prise en compte des données spatiales de plus en plus présentes dans les documents du web. A ces contraintes, nous ajoutons celle de la collaboration entre les utilisateurs et leurs données.
Ces éléments sont à la base du système proposé dans le cadre de cette thèse. Les chapitres suivants (2 et 3) font l'objet de cette proposition. Nous abordons le système dans sa globalité, la modélisation de l'utilisateur proposée ainsi que l'optimisation des résultats de recherche. Et nous finissons dans le troisième chapitre avec la représentation des modèles des utilisateurs en réseau.

Chapitre 3

Système de Recherche d'Information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur

3.1 Introduction

Les systèmes de recherche d'information (SRI) ont pour but d'offrir des moyens permettant de retourner les informations pertinentes relatives à un besoin en information d'un utilisateur à travers des collections de documents. A cet effet, la RI personnalisée intègre la modélisation de l'utilisateur dans le processus de recherche afin de lui offrir une meilleure réponse à ses besoins en information. En effet, l'utilisation d'un modèle de l'utilisateur permet la connaissance et la prise en compte des intérêts et des intentions de recherche de ce dernier. Le chapitre 2 de l'état de l'art a donné un aperçu des différents travaux de recherche dans le domaine de la recherche d'information personnalisée. Ceci nous a permis de constater que la modélisation de l'utilisateur est au coeur de la RI personnalisée. L'objectif de la RI personnalisée est alors d'intégrer le modèle de l'utilisateur dans le processus de RI à des fins de personnalisation des résultats de la recherche. Cette étude nous a aussi permis de constater que la question principale qui se pose lors de la conception de systèmes de recherche d'information personnalisée est de savoir comment représenter ce qui caractérise l'utilisateur, et comment l'intégrer dans le processus de la recherche d'information. Nous avons aussi constaté le manque de considération de l'information spatiale dans le cadre de

recherche d'information sur le web
Notre contribution dans le domaine de la RI personnalisée porte sur la proposition d'un système basé sur la construction d'un modèle multidimensionnel de l'utilisateur et son exploitation dans la construction d'un réseau de modèles des utilisateurs offrant ainsi une collaboration implicite entre modèles (Hadjouni et al., 2008), (Hadjouni et al., 2009b), (Hadjouni et al., 2010), (Hadjouni et al., 2011). La personnalisation est de ce fait basée sur l'exploitation combinée du modèle de l'utilisateur et du réseau de modèles des utilisateurs.

Nous présentons dans ce chapitre notre système de Recherche d'Information personnalisée ainsi que le modèle de l'utilisateur proposé et son exploitation à travers la construction d'un réseau de modèles multidimensionnels des utilisateurs. Ce chapitre est organisé comme suit. La section 3.2 présente les problématiques et les motivations.La section 3.3 présente le cadre général ainsi que la terminologie utilisée. La section 3.4 décrit le système proposé suivi, dans la section 3.5, par la présentation de la démarche de personnalisation. La dernière section conclut le chapitre.

3.2 Problématique et motivations

Le processus de personnalisation dans les systèmes de recherche d'information est principalement confronté à la question de la définition des informations nécessaires concernant l'utilisateur. En effet, la question de savoir comment représenter ce qui caractérise l'utilisateur, et comment l'utiliser dans le processus de l'extraction de l'information est toujours d'actualité. L'introduction de l'utilisateur dans ce processus nécessite une modélisation de ce dernier et une fiabilité de son profil (Kobsa, 2007). En effet, on constate que l'une des principales raisons du manque de performances des techniques de personnalisation est typiquement l'application d'un profil utilisateur hors contexte (Gauch et al., 2007). Les utilisateurs peuvent avoir des préférences générales, récurrentes et stables. Cependant, l'ensemble des informations contenues dans le profil n'est pas forcément approprié à toutes les situations de recherche. Le plus souvent, les systèmes n'utilisent seulement qu'un sous-ensemble de ces informations, qu'ils supposent pertinents pour la recherche en cours. Et dans ce cas, ces systèmes ont recours à des informations explicitement fournies par les

utilisateurs eux-mêmes. Ce qui n'est pas toujours significatif.

Ainsi, l'objectif de cette thèse est d'apporter des solutions nouvelles et efficaces à la problématique exposée. Pour atteindre cet objectif, les travaux réalisés sont :

- la proposition d'une architecture pour une recherche personnalisée sur les données du web.
- l'intégration de l'utilisateur en tant que composante principale du processus de recherche.

Après le survol des travaux reliés à notre problématique, nous estimons que l'utilisateur peut être représenté par quatre niveaux d'abstraction. Le premier niveau est la dimension textuelle où sont représentées les navigations ainsi que les recherches textuelles effectuées par l'utilisateur. Le second niveau est le niveau de la dimension spatiale qui représente les données spatiales du modèle de l'utilisateur. Ce niveau d'abstraction considère les informations spatiales extraites des navigations ainsi que des intérêts de l'utilisateur. Sa construction se base sur l'hypothèse que si nous extrayons les intérêts d'ordre géo-spatial de l'utilisateur, nous avons une meilleure visibilité quant à l'orientation de ces intérêts. En effet, dans l'état de l'art, nous avons constaté que pour une situation de recherche en contexte non mobile, les données spatiales ne sont pas prises en compte. Ces données sont principalement exploitées dans le domaine de la recherche d'information en contexte mobile (Bila et al., 2008), (Bouidghaghen et al., 2011). Le troisième niveau est le niveau intérêt où une déduction implicite des centres d'intérêt de l'utilisateur est effectuée. Les objets de cette dimension sont représentés par les liens sémantiques pondérés existant entre eux. Ainsi, si nous n'utilisons que les concepts (auxquels appartiennent les objets), les intérêts de l'utilisateur seront représentés par un sous-graphe d'intérêt général. Le quatrième et dernier niveau représente la dimension statique et explicite du modèle de l'utilisateur. En effet, cette dimension profil contient les données d'identification de l'utilisateur. Se contenter d'une seule dimension des trois premières citées ne permet d'avoir qu'une visibilité partielle de l'utilisateur. Nous estimons qu'afin d'aboutir à une personnalisation plus efficace des données fournies à l'utilisateur, l'ensemble des connaissances implicites de l'utilisateur devrait être exploité.

La proposition considère particulièrement les aspects suivants :

- La représentation et la maintenance des informations concernant l'utilisateur et ses manipulations sous forme de modèle. Ce modèle de l'utilisateur est caractérisé par une récolte d'une part implicite et évolutive des données, et d'autre part à partir de ses interactions avec l'architecture proposée.
- La prise en compte, des informations spatiales demandées et/ou fournies par l'utilisateur. Ces données peuvent provenir aussi bien des recherches effectuées que des navigations et sélections des utilisateurs dans les résultats fournis.
- Le troisième aspect concerne l'intégration de ces modèles de l'utilisateur dans un graphe représentatif des utilisateurs du système global dont l'objectif est de construire un graphe d'utilisateurs est de proposer une collaboration implicite entre les modèles.

D'une manière comparative aux travaux présents dans le domaine, l'architecture que nous proposons se distingue par :

- La construction d'un modèle multidimensionnel de l'utilisateur à partir de données implicites. Ce modèle prend en charge les informations spatiales en tant que dimension à part entière. Ce choix repose sur le fait que les informations spatiales existantes ne concernent pas uniquement l'utilisateur, mais aussi les documents et les données manipulés.
- Construction d'un réseau de modèles utilisateurs : l'objectif de l'utilisation de ce réseau est de proposer à l'utilisateur des informations personnalisées en se basant sur un filtrage collaboratif implicite.
- Le calcul implicite de l'intérêt de l'utilisateur pour un document donné. Ce calcul est effectué en temps réel et se base principalement sur les données en cours et à long terme du modèle de l'utilisateur.

3.3 Cadre général

Le cadre général proposé est articulé autour des éléments suivants :

1. L'utilisateur : Le système a pour objectif final de fournir à cet utilisateur des résultats personnalisés. Cet utilisateur interagit avec le système en fournissant des requêtes de manière textuelle ou à travers le positionnement sur une carte géographique. Il fournit au système des informations implicites à travers ses navigations dans le jeu

de résultats fournis. Nous distinguons deux types d'utilisateur :

- Utilisateur connu : Dans notre proposition, un utilisateur connu est enregistré dans le système de personnalisation. Son modèle de l'utilisateur contient ses informations d'accès. Le détail de ces données est fourni dans la section 4.2.
- Utilisateur inconnu : Un utilisateur inconnu est affecté à un stéréotype existant s'il y a une certaine similarité des intérêts (c.f. section 5.2.2). Dans le cas échéant, il y a création d'un stéréotype.

2. L'ensemble des modèles utilisateurs : La construction du système autour d'un réseau de modèles utilisateurs a pour principal objectif d'améliorer la pertinence des résultats personnalisés.
3. Le processus de personnalisation : Ce processus repose sur la construction implicite du modèle de l'utilisateur et sur l'utilisation d'un filtrage collaboratif implicite à travers le réseau des modèles utilisateurs.

La figure 3.1 illustre le cadre général de notre système de recherche d'information personnalisée et spatiale.

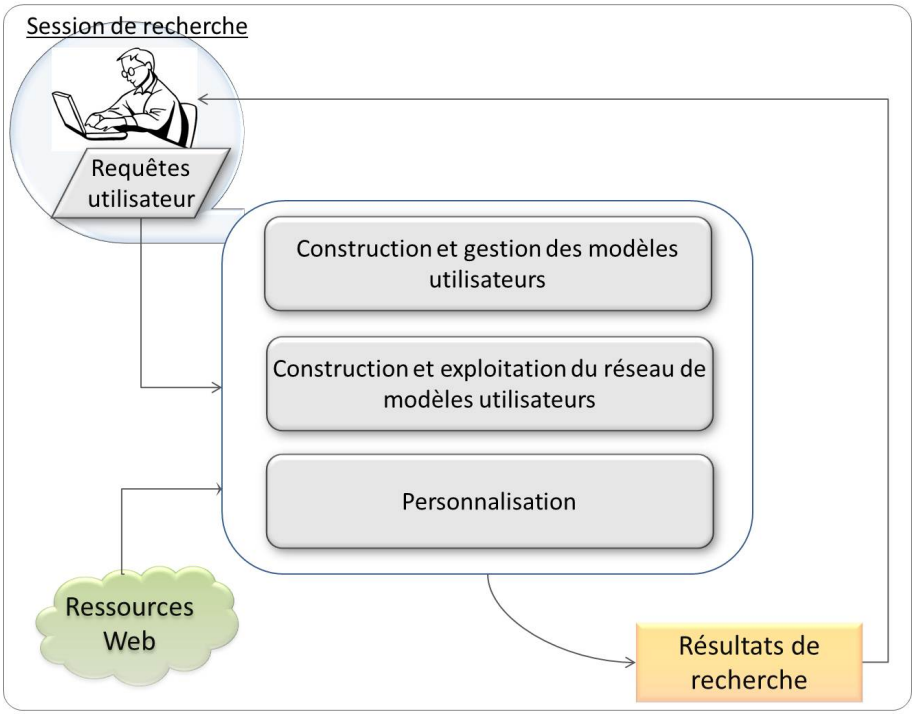


FIGURE 3.1 – Cadre général du système de personnalisation

Les notions de requête de l'utilisateur et de session de recherche sont présentées dans ce qui suit.
Soit U l'ensemble des utilisateurs.

Session de recherche : Une session de recherche est composée d'une séquence de requêtes ayant démarrée au premier accès au système et finissant quand l'utilisateur quitte SyPRISS. La session de recherche est aussi caractérisée par une ou plusieurs itérations de recherche. A l'instant t donné, pour un utilisateur $u \in U$, une session de recherche S contient n requêtes et s'écrit comme suit :

$$S = \{q_u^0, q_u^1, ..., q_u^n\} \tag{3.1}$$

Requête utilisateur : La requête utilisateur est la concaténation d'une, de deux ou de trois types de requêtes : textuelle, spatiale et d'intérêt. A un instant t donné, la requête q de l'utilisateur u s'écrit comme suit :

$$q = q_{tx} + q_{sp} + q_{in} \tag{3.2}$$

La requête textuelle q_{tx} est l'ensemble des mots saisis par l'utilisateur. Un mot est noté par w_i :

$$q_{tx} = \{w_1, w_2, ..., w_n\} \tag{3.3}$$

La requête spatiale q_{ps} représente la/les position(s) géographique(s) p_i sélectionnée(s) par l'utilisateur u sur la carte. En effet, nous mettons à disposition de l'utilisateur une carte géographique pour qu'il puisse sélectionner l'emplacement de recherche voulu (au lieu de l'écrire textuellement). Une position géographique étant caractérisée par une longitude l, une latitude L et un nom N, la requête spatiale est alors la combinaison de ces trois informations :

$$q_{sp} = \{l, L, N\} \tag{3.4}$$

La requête d'intérêt q_{in} est issue de la sélection , à travers l'interface du système, d'un intérêt enregistré dans le modèle de l'utilisateur.

Itération de recherches : Une itération de recherche est l'ensemble des actions correspondantes à une requête soumise. Elle est caractérisée par la requête q, la liste des résultats R_u^q et les actions de l'utilisateur à travers ces résultats.

3.4 Présentation des composants du système de recherche d'Information Personnalisée

Le système proposé est un système de Personnalisation de la Recherche d'Information Spatiale et Sémantique sur le Web qui intègre la modélisation multidimensionnelle de l'utilisateur ainsi que la construction d'un réseau de modèles utilisateurs. Ce système, que nous appellerons SyPRISS, est représenté sous la forme de trois composants :

Composant de collecte de données utilisateur et de reformulation de requêtes : Ce composant a pour rôle principal l'acquisition des données de navigation de l'utilisateur à travers les résultats de recherche. Il prend aussi en charge la reformulation de la requête utilisateur.

Composant de la construction des modèles utilisateurs : Ce composant prend en charge la construction du modèle de l'utilisateur et sa mise à jour en se basant sur les données de navigation de l'utilisateur. Ce composant prend ensuite en charge le calcul implicite de l'intérêt de l'utilisateur pour un lien donné.

Composant de la construction du réseau d'utilisateurs : Les tâches dans ce composant reposent sur la connaissance, ou non, de l'utilisateur. En effet, à l'accès d'un utilisateur au système de personnalisation, le composant de collecte de données utilisateur détecte si ce dernier est identifié ou pas. Pour un utilisateur identifié, le modèle correspondant est chargé et sera utilisé afin d'effectuer les recherches à travers les noeuds du réseau de modèles, détaillé à la section 5.2. Si l'utilisateur effectuant une recherche n'est pas identifié par le système, on exploite sa requête soumise pour l'affecter à un stéréotype du système. Dans le cas de la création d'un nouveau modèle de l'utilisateur, ce composant a la charge de recherche à travers le réseau le noeud dont le modèle est le plus similaire au

nouveau modèle.

La figure 3.2 représente ces trois composants.

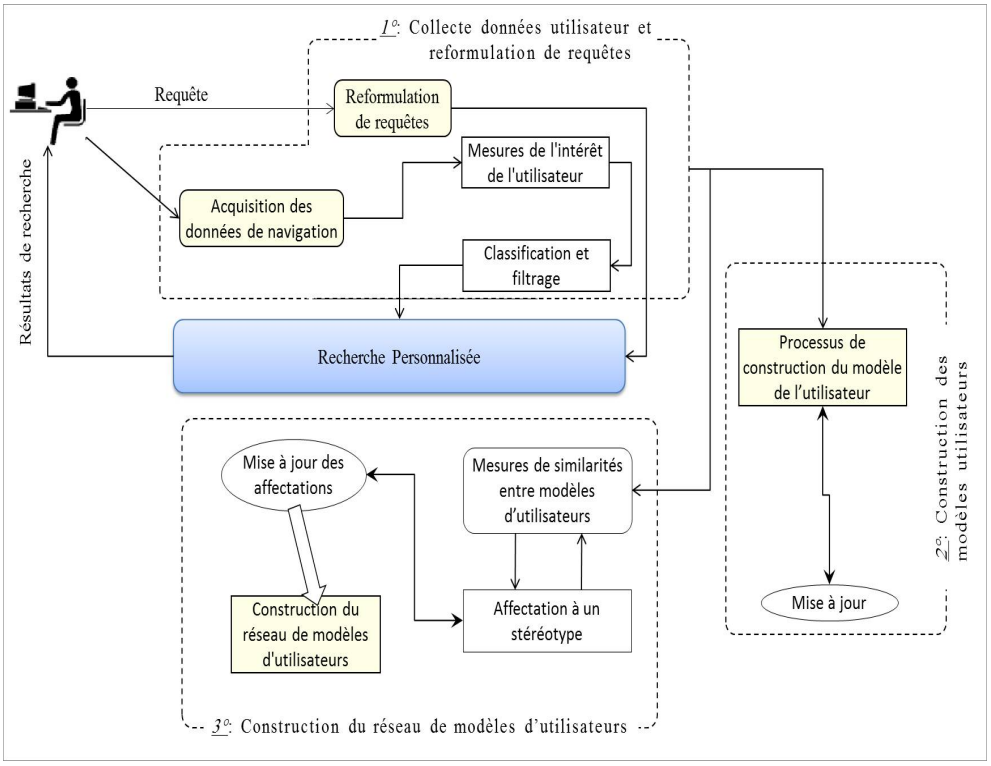


FIGURE 3.2 – Composants de l'architecture du système de personnalisation

Dès que l'utilisateur soumet sa requête, le composant de collecte de données utilisateur et de reformulation de requêtes entame l'étape de reformulation de la requête utilisateur. Une fois la reformulation effectuée (ou non), la recherche personnalisée est lancée. Cette dernière exploite les composants de construction des modèles utilisateurs et de construction du réseau d'utilisateurs afin de récupérer les résultats personnalisés des modèles utilisateurs. Ces deux composants effectuent les mises à jour des modèles au fur et à mesure des recherches et des interactions des utilisateurs avec le système.

L'algorithme de construction des résultats personnalisés est présenté dans la section 5.4.2, p. 106 du chapitre 5

Le modèle de l'utilisateur Le modèle de l'utilisateur, présenté dans le chapitre 4, est multidimensionnel et est composé de quatre dimensions :

- Dimension textuelle
- Dimension centres d'intérêt
- Dimension spatiale
- Dimension des données personnelles de l'utilisateur

La dimension textuelle, présentée dans la section 4.2.1 p.71, contient les données de recherche textuelle, l'historique des navigations ainsi que les requêtes émises de l'utilisateur à travers les différentes sessions de recherche. Elle représente une bibliothèque de données de la recherche textuelle de l'utilisateur. Les informations contenues dans cette dimension vont servir à inférer les données des dimensions d'intérêt et spatiale. Les données de cette dimension sont représentées sous forme vectorielle.

La dimension centres d'intérêt, sect. 4.2.2 p. 74, contient les données d'intérêt de l'utilisateur sous la forme d'une ontologie de concepts pondérés par un degré d'intérêt implicite de l'utilisateur, cf. sect. 4.3.1 p. 85. Cette dimension est exploitée pour l'enrichissement des requêtes de l'utilisateur et pour le réordonnancement des résultats de recherche. La construction de cette dimension est présentée dans la section 4.2.2.2 p. 77.

La dimension spatiale contient les données spatiales du modèle de l'utilisateur et est présentée dans la section 4.2.3 p. 80. Elle est composée d'entités spatiales dont la représentation est sous forme d'arbre de concepts pondérés. Dans le cadre d'une information à caractère spatial, la distribution spatiale des lieux influent les choix de l'utilisateur ainsi que son évaluation de la pertinence des résultats de recherche personnalisés qui lui sont fournis. Dans ce contexte, nous proposons d'exploiter la notion d'accessibilité dans notre proposition cf. sect.4.4 p.89. Nous l'exploitons dans l'objectif d'améliorer la qualité des informations personnalisées fournies à l'utilisateur.

La dimension des données personnelles de l'utilisateur, sect.4.2.4 p.82, contient les données personnelles relatives au profil de l'utilisateur. Dans notre proposition, basée sur la prise en compte implicite des données utilisateur, cette dimension contient seulement les informations d'identification de l'utilisateur, explicitement fournies par l'utilisateur.

La figure 3.3 récapitule les données circulant entre les différentes dimensions du modèle de l'utilisateur :

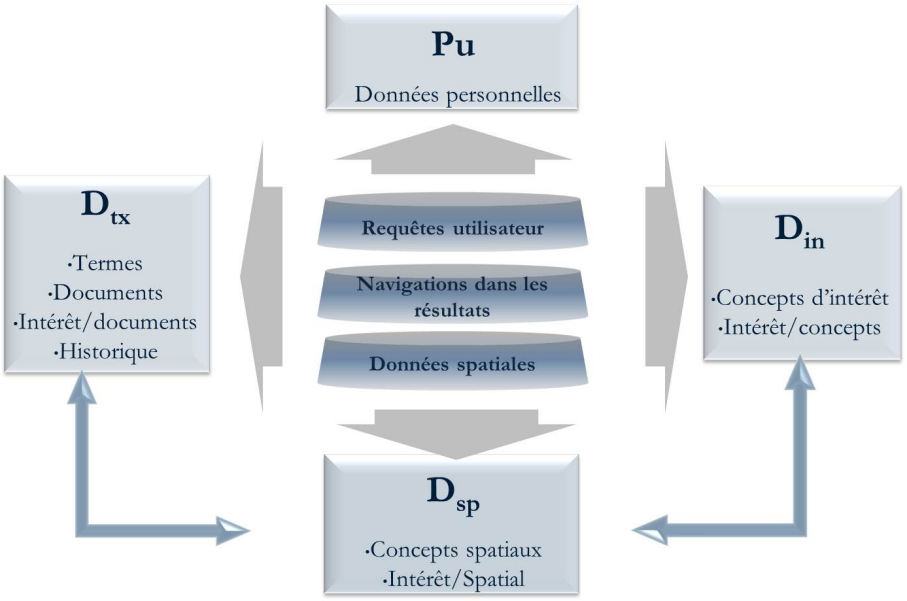


FIGURE 3.3 – Représentation des données circulant entre les dimensions du modèle de l'utilisateur

Le réseau de modèles utilisateurs dans SyPRISS En partant de l'hypothèse que, lors de la recherche d'un document pertinent, le système devrait, en plus des besoins spécifiques des utilisateurs et de leurs recherches antérieures, exploiter les connaissances extraites des autres modèles utilisateurs existants, nous nous proposons de construire un réseau de modèle des utilisateurs, présenté dans le chapitre 5. Ce réseau va permettre une collaboration implicite entre les différents modèles le constituant. Les noeuds du réseau sont représentés par les modèles des utilisateurs. Afin de pouvoir exploiter le voisinage de chaque noeud, nous utilisons des mesures de similarité spatiales et sémantiques entre les différents intérêts des modèles (cf. sect.5.2.2, p. 100). Le choix de ces deux mesures est basé sur le fait qu'un utilisateur effectue généralement une recherche textuelle et peut avoir besoin d'informations d'ordre spatial. Leur objectif est de renseigner sur le degré de correspondance entre deux modèles d'utilisateurs du réseau. En effet, la distance (inverse de la similarité) entre deux noeuds du réseau est la combinaison de la distance sémantique et de la distance spatiale cf. sect. 5.2.2.3, p. 103. La construction du réseau est présentée dans la section 5.3 p. 103.

Le système de personnalisation proposé se base sur la collecte implicite d'informations de l'utilisateur. De ce fait, chacune des définitions de modèle utilisateur est représentée par une instanciation des éléments suivants : les sources d'information, la stratégie de collecte des informations et les ressources de modélisation, que nous décrivons ci-dessous.

Les sources d'information

Afin de modéliser les informations des utilisateurs et de construire le réseau de modèles, section 5.2, nous nous basons sur les sources d'information suivantes :

- les documents cliqués par l'utilisateur comme source représentant son intérêt en réponse à sa requête,
- les données de navigation telles que le nombre de cliques, le nombre de visites et la durée de consultation d'un document cliqué.
- les coordonnées géographiques pouvant être extraites des adresses des documents cliqués.

La stratégie de collecte d'information

Notre stratégie de collecte d'information est totalement implicite. En effet, pour la modélisation de l'utilisateur, cette collecte est basée sur la construction de données historiques des recherches. Ces données sont spécifiques à chaque utilisateur. Il est à noter que, pour des raisons de sécurité des données personnelles, cette historisation des traces de navigation nécessiterait l'autorisation préalable de l'utilisateur.

Les ressources de modélisation

Comme ressource pour la modélisation des intérêts de l'utilisateur, nous exploitons des ressources sémantiques externes : une ontologie thématique générale, l'ODP, et un dictionnaire de données, WordNet.

3.5 La démarche de personnalisation dans SyPRISS

Les étapes de la personnalisation sont les suivantes :

- Reformulation de la requête de l'utilisateur
- Recherche personnalisée
- Affichage des résultats triés à l'utilisateur
- Acquisition des données de navigation de l'utilisateur à travers les résultats de re-

cherche

- Construction et Mise à jour des données du modèle de l'utilisateur
- Construction et Mise à jour du réseau de modèles utilisateurs.

Reformulation de la requête de l'utilisateur Cette étape fait appel au composant de collecte de données utilisateur et de reformulation de requêtes. Lors de cette étape, la requête émise par l'utilisateur est reformulée en utilisant le modèle de ce dernier. Il s'agit d'y ajouter des termes extraits du modèle de l'utilisateur. En effet, nous utilisons pour cela la dimension centres d'intérêt du modèle. Cette dernière étant construite sur la base d'une ontologie, nous en extrayons les termes qui correspondent le plus à ceux utilisés dans la requête de l'utilisateur.

Recherche personnalisée Les données utilisées lors de cette étape sont de trois types : données du modèle de l'utilisateur courant, données des modèles d'utilisateurs jugés les plus proches dans le réseau de modèles utilisateurs du système et finalement les résultats fournis à travers la recherche sur le web. Ce module prend comme données de départ la requête reformulée. L'affichage des données résultats est trié par l'ordre de pertinence suivant : (1) liens précédemment visités par l'utilisateur lui-même et implicitement jugés comme pertinents, (2) liens parcourus par les utilisateurs les plus proches du réseau de modèles utilisateurs du système et finalement (3) les liens du web.

Acquisition implicite de données de navigation Cette étape fait appel au composant de collecte de données utilisateur et de reformulation de requêtes. Les données de navigation utilisées correspondent au temps de lecture des résultats, à la fréquence d'accès ainsi qu'aux différentes sélections que peut effectuer l'utilisateur à travers l'interface du système. Lors de cette étape un calcul implicite de l'intérêt de l'utilisateur pour un lien donné. En effet, nous basons notre approche sur un ensemble de mesures qui nous servira déduire les centres d'intérêt de l'utilisateur et qui sont calculées au fur et à mesure de la navigation.

Modélisation des données de l'utilisateur Cette étape, prise en charge par le composant de construction des modèles utilisateurs, comprend aussi bien la construction initiale

du modèle de l'utilisateur que sa mise à jour lors des différentes recherches et navigations de l'utilisateur. Notre stratégie se déroule en 3 étapes :

- Construction du modèle de l'utilisateur à la première session d'accès
- Mise à jour et évolution de l'historique de recherche à partir des données collectées par le module d'acquisition implicite de données de navigation
- Extraction des données d'intérêt et des données spatiales et prise en charge de leur évolution à travers les différentes sessions de recherche.

La construction du réseau d'utilisateurs Cette étape est prise en charge par le composant de la construction du réseau d'utilisateurs. Partant des données collectées et traitées par chacun des composants du système, le composant de construction du réseau d'utilisateurs se charge de construire et de mettre à jour un réseau de modèles utilisateurs reliés entre eux par une similarité sémantique et spatiale. La construction du système autour d'un réseau de modèles utilisateurs a pour principal objectif d'améliorer la pertinence des résultats à fournir à l'utilisateur. Les noeuds constituant le réseau sont les modèles utilisateurs, et la distance séparant chaque noeud est calculée en fonction des données spatiales et sémantiques extraites des modèles eux-mêmes.

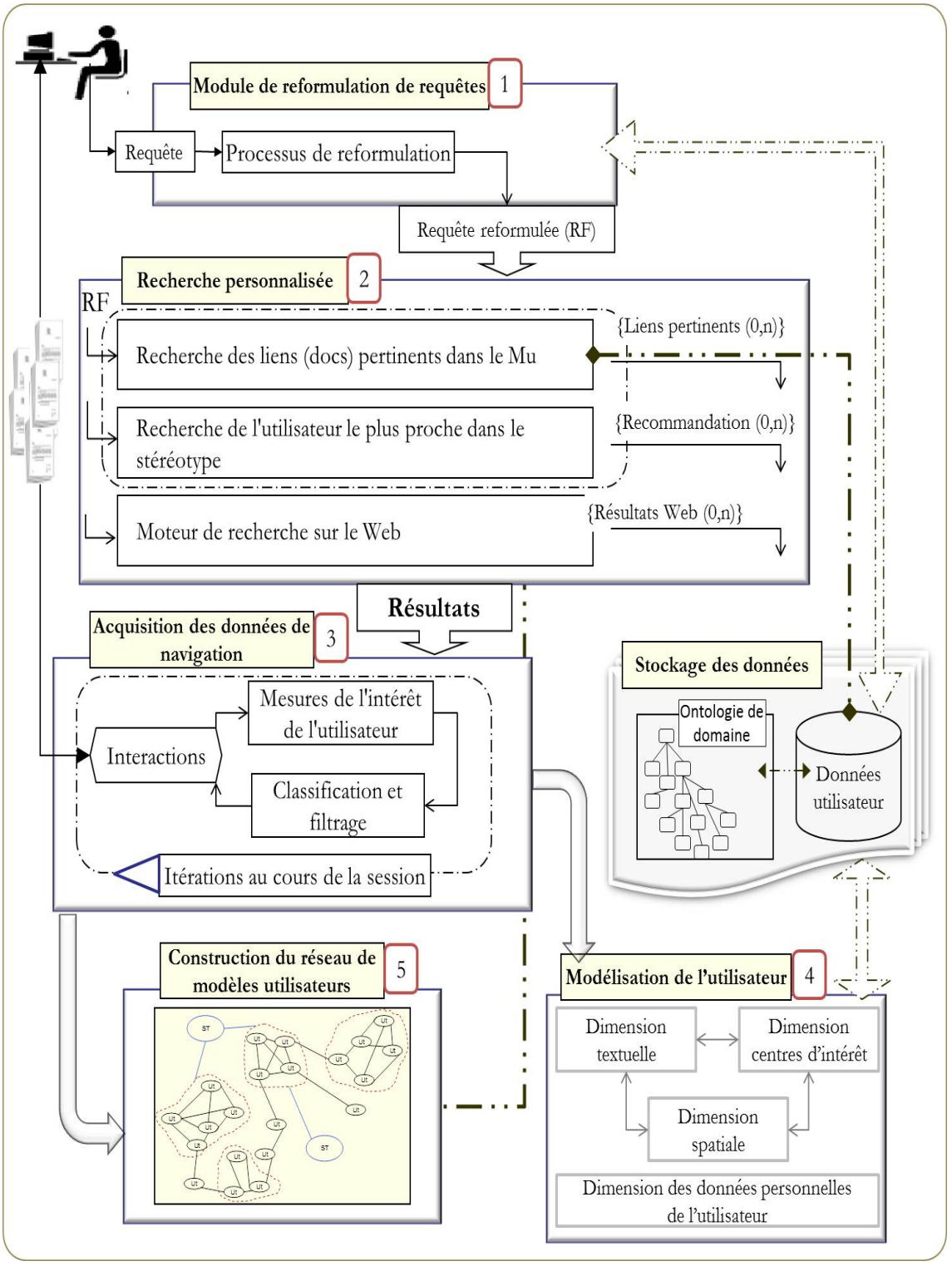


FIGURE 3.4 – Démarche de personnalisation dans SyPRISSr

En partant de la requête de l'utilisateur, la première étape de la personnalisation est de procéder à la reformulation de la requête de l'utilisateur, c'est l'étape 1 de la figure 3.4. Cette reformulation de la requête est basée sur l'algorithme 5 proposé dans la sect. 5.4.1 p. 105 chap. 4.
Une fois la requête reformulée , l'étape de recherche personnalisée est lancée (étape 2 de la figure). Dans cette étape, trois types de recherche sont effectués : une recherche à travers les données du modèle de l'utilisateur courant, une recherche dans le modèle utilisateur le plus similaire du voisinage et une recherche dans le web. Dans la première recherche, l'objectif est d'extraire les documents visités par l'utilisateur courant et qui pourraient répondre à la requête émise. En effet, en exploitant la dimension centres d'intérêt qui contient les concepts pondérés par l'intérêt implicite de l'utilisateur et la dimension textuelle de laquelle vont être extraits, s'ils existent, les documents pertinents, le processus de recherche d'information personnalisée faciliter l'accès à l'information pertinente pour l'utilisateur. Le second type de recherche est effectué dans le modèle de l'utilisateur le plus proche. Le réseau de modèles utilisateurs est ainsi exploité pour fournir les résultats de recherche pertinents du modèle de l'utilisateur dont la dimension d'intérêt contient des concepts d'intérêt similaires à ceux de l'utilisateur courant. Ensuite, une recherche sur le web est effectuée.
Afin de procéder à l'affichage des résultats ordonnancés, le système calcule le score personnalisé des documents en utilisant l'équation 4.9 p. 80.

L'acquisition des données de navigation de l'utilisateur à travers les résultats de recherche est activée dès que l'utilisateur interagit avec le système. Cette étape est active tant que la session de recherche de l'utilisateur n'a pas été terminée. Les composants de modélisation des données de l'utilisateur et de construction du réseau d'utilisateurs interagissent avec le composant de collecte de données utilisateur et de reformulation de requêtes afin de mettre à jour les mesures d'intérêt pour les documents visités ainsi que les mesures de similarité entre les noeuds du réseau.

3.6 Conclusion

Nous avons présenté dans ce chapitre le système de Recherche d'Information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur et sur la construction d'un réseau de modèles utilisateurs. Etant dans le cadre de l'utilisation d'informations à caractère spatial, nous avons proposé de prendre en considération l'accessibilité aux entités spatiales choisies et à choisir par l'utilisateur. A cet effet, nous avons proposé une mesure d'accessibilité basée sur les préférences de l'utilisateur afin de lui fournir un meilleur résultat que nous exploitons dans la construction de dimension spatiale du modèle de l'utilisateur. Nous procédons, dans le chapitre 4, à présenter le modèle multidimensionnel de l'utilisateur en détaillant les dimensions qui le constituent ainsi que leur construction.

Chapitre 4

Modélisation multidimensionnelle de l'utilisateur pour la personnalisation

4.1 Introduction

La modélisation de l'utilisateur est au coeur de la recherche d'information personnalisée dont l'objectif est d'intégrer les données de l'utilisateur dans le processus de RI à des fins de personnalisation des résultats de la recherche. En effet, la question principale qui se pose lors de la conception de systèmes de recherche d'information personnalisée est de savoir comment représenter ce qui caractérise l'utilisateur, et comment l'utiliser dans le processus de la recherche d'information.

Nous présentons dans ce chapitre la modélisation multidimensionnelle de l'utilisateur proposée. Ce modèle intègre une représentation sémantique et spatiale des intérêts utilisateur. En effet, les centres d'intérêts sont représentés par deux dimensions représentant respectivement les intérêts et leur représentation spatiale. Nous commençons ce chapitre par présenter les dimensions du modèle de l'utilisateur, section 4.2. La section 4.3 présente la construction des dimensions du modèle de l'utilisateur. Nous présentons ensuite, dans la section 4.4, la mesure d'accessibilité proposée et exploitée dans la dimension spatiale du modèle de l'utilisateur. Nous finissons le chapitre par une conclusion.

4.2 Présentation du modèle multidimensionnel de l'utilisateur

L'étude des différentes modélisations de l'utilisateur nous a fait aboutir à la conclusion qu'il est nécessaire de capturer un ensemble d'informations assez important et caractérisant l'utilisateur. Leur représentation doit de ce fait être de manière départagée en plusieurs dimensions. En effet, l'utilisateur est déterminé par plusieurs niveaux de connaissances allant des besoins en information au moment de la requête aux préférences et centres d'intérêts. Notre approche générale pour la personnalisation de la recherche repose sur la construction d'un modèle de l'utilisateur dynamique et multidimensionnel qui évolue en fonction de ses recherches et navigations à travers les résultats. La représentation du modèle est sous quatre dimensions : les données utilisateurs, les positions spatiales, les recherches textuelles et les centres d'intérêt. A l'aspect multidimensionnel, nous ajoutons l'utilisation du web sémantique dans l'objectif de contourner les ambigüités sémantiques possibles. Les ambigüités pourraient apparaitre lors de la collecte d'informations de navigation de l'utilisateur : la sélection d'un même terme dans différentes pages à travers la requête utilisateur ne signifie pas la même chose à chaque utilisation. La représentation des intérêts de l'utilisateur est effectuée sous la forme d'une ontologie de concepts. Cette ontologie est construite à partir de l'ontologie de référence ODP et est enrichie en utilisant le dictionnaire de données Wordnet.

La composante minimale du modèle de l'utilisateur est l'information du login d'accès et du mot de passe, répertoriés dans la dimension des données utilisateur. En effet, nous allons vers l'implication minimale de l'utilisateur dans le processus entier de personnalisation. Plus spécifiquement, le modèle de l'utilisateur, noté M_u, comprend un ensemble de données de recherche, caractéristiques de la dimension textuelle, auxquelles sont associés des centres d'intérêt, dimension centres d'intérêt, et des données spatiales, dimension spatiale. Nous avons donc, à un instant t donné :

$$M_u = (D_{tx}, D_{sp}, D_{in}, P_u) \tag{4.1}$$

Avec :

D_{tx} est la dimension contenant les données de recherche textuelle (cf. section 4.2.1 p. 71). Cette dimension contient l'ensemble des termes de recherche auxquels sont associées les durées de consultation des résultats ainsi que l'historique des navigations. En notant W, D, H respectivement l'ensemble des termes, les durées de consultations minimale et maximale associées et l'historique des navigations, nous aurons :

$$D_{tx} = \{W, D, H\} \tag{4.2}$$

D_{sp} est la dimension contenant les données spatiales du modèle de l'utilisateur (cf. section 4.2.3 p. 80). Cette dimension est composée d'entités spatiales, notées E_{sp}, dont la représentation est sous forme d'arbre de concepts pondérés. La pondération est issue du calcul d'intérêt, I porté par l'utilisateur.

$$D_{sp} = \{E_{sp}, I_i\} \tag{4.3}$$

D_{in} est la dimension représentative des données d'intérêt de l'utilisateur. La représentation de ses intérêts est sémantique selon un ensemble de concepts sémantiquement liés à travers l'utilisation de l'ontologie de référence ODP. A chaque concept est associée une pondération de l'intérêt porté par l'utilisateur. Pour un concept C_i, nous aurons donc l'information des termes associés ainsi que l'intérêt porté par l'utilisateur :

$$D_{in} = \{C_i, w_j, I_i\}, j \in [1, n] \tag{4.4}$$

P_u représente les données de l'utilisateur. Cette dimension contient les données personnelles relatives au profil de l'utilisateur. Dans notre proposition, basée sur la prise en compte implicite des données utilisateur, cette dimension contient seulement les informations d'identification de l'utilisateur.

4.2.1 La dimension textuelle

La dimension textuelle du modèle de l'utilisateur est représentative des différentes sessions de recherche effectuées, de ses requêtes ainsi que de l'historique des navigations. Elle représente une bibliothèque de données de la recherche textuelle de l'utilisateur. Les informations contenues dans cette dimension vont servir à inférer les données des dimensions

d'intérêt et spatiale.

Les données de cette dimension sont représentées sous forme vectorielle. L'équation générale de cette dimension, éq.4.2 p. 71, est détaillée dans ce qui suit. W, qui représente l'ensemble des termes de recherche est un n-uplet sous la forme :

$$W = \{(S_1, w_{s_1}^{q_1}), ..., (S_i, w_{s_i}^{q_i}), ..., (S_n, w_{s_n}^{q_n})\} \tag{4.5}$$

Avec S_i la session de recherche présentée par l'équation 3.1, p. 58,

$w_{s_i}^{q_i}$ l'ensemble des termes de la requête q_i de la session S_i

D présente les différentes durées de consultation ainsi que les fréquences d'accès. D est représenté par le n-uplet suivant :

$$D = (F_s, dc_{max}, dc_{min}, T_t) \tag{4.6}$$

Où :

F_s représente la fréquence de saisie du terme w

$$F_s = \{(w_1, f_1), ..., (w_i, f_i), ..., (w_n, f_n)\}$$

$dc_{max/min}$ est la durée maximum/minimum de consultation des résultats de recherche pour le terme w :

$$dc_{min/max} = \{S_i, \{d_{w_j}^{Q_i}\}\}$$

$d_{w_j}^{q_i}$: durée de consultation pour chaque terme w_j de chaque requête q_i de la session de recherche S_i.

T_t est le temps total de consultation des documents par requête.

H est une représentation de l'historique des documents visités. Chaque terme w_u^i est relié à une requête q_u émise par l'utilisateur u et est associé à l'ensemble des documents consultés D_u^i. Cet historique de recherche, pour une requête donnée q_u, s'écrit :

$$H_q = \{w_u^i, D_u^i, q_u, S_u\} \tag{4.7}$$

Où

$D_u^i = \{(d_j, t_j), 0 < j < n\}$ est l'ensembles des documents consultés par l'utilisateur u et de leurs temps de consultation respectifs et

n le nombre de documents consultés par requête.

Nous présentons dans ce qui suit un exemple de données de la dimension textuelle. Considérons l'utilisateur u qui a effectué à travers le système trois sessions de recherche, nous avons à un instant t :

$$W = \{(S_1, q_1, moto, 3roues), (S_1, q_2, montagne, hotel, fac), (S_2, q_3, moto, hotel), (S_3, q_4, fac, Paris)\}$$

L'ensemble des fréquences de saisie associées à chaque terme est alors :

$$F_s = \{(moto, 2), (3roues, 1), (montagne, 1), (hotel, 1), (fac, 2), (Paris, 1)\}$$

Pour les durées minimales et maximales, elles sont présentées par l'ensemble des t-uplets suivants :

$$dureeMax = \{(S_1, q_1, moto, 20), (S_1, q_1, 3roues, 20), (S_1, q_2, montagne, 125), (S_1, q_2, hotel, 125), (S_1, q_2, fac, 125), (S_2, q_3, moto, 35), (S_2, q_3, hotel, 35), (S_3, q_4, fac, 300), (S_3, q_4, Paris, 300)\}$$

Sachant que le temps de lecture d'un document est relatif à la requête, donc, toute session confondue, nous aurons :

$dc_{max} = \{(moto, 35), (3roues, 20), (montagne, 125), (hotel, 125), (fac, 300), (Paris, 300)\}$

Respectivement pour les durées minimales nous aurons donc :

$dc_{min} = \{(moto, 10), (3roues, 10), (montagne, 100), (hotel, 10), (fac, 100), (Paris, 250)\}$

Le temps total de lecture par requête découle des données sur le temps de lecture estimé pour chaque document consulté. En utilisant l'historique des recherches H_q, un exemple est présenté dans ce qui suit, nous aurons l'information sur le temps total de lecture par

requête T_t.

$H_q = \{(moto, ((url1, 15), (url2, 20), (url3, 10)), q_1, S_1),$

$(3roues, ((url1, 15), (url2, 20), (url3, 10)), q_1, S_1),$

$(montagne, ((url3, 105), (url4, 100), (url5, 122), (url6, 125), (url7, 120)), q_2, S_1),$

$(hotel, ((url3, 105), (url4, 100), (url5, 122), (url6, 125), (url7, 120)), q_2, S_1),$

$(fac, ((url3, 105), (url4, 100), (url5, 122), (url6, 125), (url7, 120)), q_2, S_1),$

$(moto, ((url8, 35), (url9, 25), (url10, 10), (url11, 35), (url12, 29), (url13, 31),$

$(url14, 24), (url15, 22), (url16, 19)), q_3, S_2),$

$(hotel, ((url8, 35), (url9, 25), (url10, 10), (url11, 35), (url12, 29), (url13, 31),$

$(url14, 24), (url15, 22), (url16, 19)), q_3, S_2),$

$(fac, ((url17, 300), (url18, 255), (url19, 250), (url20, 285), (url21, 290), (url22, 291),$

$(url23, 257), (url24, 282), (url25, 279)), q_4, S_3),$

$(Paris, ((url17, 300), (url18, 255), (url19, 250), (url20, 285), (url21, 290), (url22, 291),$

$(url23, 257), (url24, 282), (url25, 279)), q_4, S_3),$

4.2.2 La dimension centres d'intérêt

La dimension centres d'intérêt, seconde dimension du modèle de l'utilisateur, a pour objectif la représentation des centres d'intérêts de l'utilisateur. Notre approche intègre une représentation sémantique basée sur l'exploitation d'une ontologie de domaine prédéfinie, l'ODP, comme source de base des domaines sémantiques. Les données de cette dimension sont construites sous forme d'une ontologie de concepts, très étroitement associés aux termes des utilisateurs.

Les sources de données sont les requêtes et les navigations de l'utilisateur à travers les résultats de ses recherches. Ces informations sont extraites de la dimension textuelle qui a pour objectif d'alimenter les différentes dimensions du modèle de l'utilisateur. De ces données, nous dégageons implicitement le degré d'intérêt porté par l'utilisateur (cf. section 4.3.1 p. 85) et que nous exploitons en tant que poids des différents concepts représentatifs

de cette dimension.

Les intérêts des utilisateurs représentent les concepts de l'ontologie d'intérêt construite. Ces concepts contiennent un nombre de propriétés comme les synonymes extraits du dictionnaire de données WordNet. Notre but est de proposer une représentation efficace des intérêts afin de permettre une meilleure précision lors de l'extraction des résultats de recherche.

4.2.2.1 Présentation des ressources utilisées

Les deux ressources sémantiques exploitées sont : l'ontologie de domaine ODP libre d'accès du projet Open Directory Project [7] ainsi que le dictionnaire de données Wordnet [8]. Ces deux sources représentent les ressources de base du domaine de la recherche scientifique et particulièrement de l'information sémantique.

Présentation de la ressource sémantique

L'ODP, pour Open Directory Project, représenté comme le plus complet des répertoires du Web édités par des êtres humains, est souvent utilisé comme source de connaissance sémantique. C'est à ce même effet que nous avons choisi de l'utiliser afin de construire la dimension centres d'intérêt du modèle de l'utilisateur de notre système. Nous nous proposons dans cette section de présenter les concepts de l'ODP ainsi que leur utilisation pour la représentation de la dimension centres d'intérêt du modèle de l'utilisateur.
Les données de l'ODP sont disposées dans deux fichiers RDF. Le premier fichier sert de représentation de la structure arborescente de l'ontologie de l'ODP. Quant au second, il contient la liste des ressources web associées à chacun des concepts. Chaque concept de l'ODP représente un domaine d'intérêt des utilisateurs du web. Les concepts de plus haut niveau sont des concepts généraux. Les concepts de plus bas niveau sont plus spécifiques. Chacun des concepts est représenté par un titre et une description générale du contenu

7. http ://www.dmoz.org/
8. http ://wordnet.princeton.edu/. Ce dictionnaire est téléchargeable à l'adresse : http ://wordnet.princeton.edu/wordnet/download/

des pages web associées. Les concepts de l'ontologie sont reliés entre eux par des relations de type "is-a", "symbolic" et "related". Le premier type de liens permet un passage hiérarchique des concepts génériques à ceux plus spécifiques. Le type "symbolic" permet la multi classification des pages dans plusieurs concepts et de ce fait, offre à l'utilisateur la possibilité de naviguer entre des concepts sémantiquement liés sans revenir à chaque fois vers les concepts généraux. Quant au type "related", il permet de pointer vers des concepts traitant de la même thématique sans pour autant avoir de pages web communes.
Les concepts dans ODP peuvent être représentés de deux manières : la première considère chaque concept comme un vecteur de termes pondérés et issus des pages web associées ; la seconde représente chaque concept par le titre associé dans l'ontologie. Chacune de ces deux méthodes présente un inconvénient. Dans la représentation par vecteurs de termes, certains concepts de haut niveau de l'ontologie n'ont pas d'association avec des pages Web. Dans la seconde représentation il y a un manque de termes représentatifs de chaque concept.
Afin d'exploiter l'ODP pour la construction de la dimension centres d'intérêt, nous adoptons la représentation de concepts par vecteurs de termes. Nous suivons les étapes suivantes :

- Pour chaque concept C_i de l'ODP, nous sauvegardons les titres et les descriptions des 40 premières adresses associées dans un fichier dc_i.
- Pour chaque document dc_i, nous effectuons une lemmatisation à travers l'utilisation de l'algorithme de Porter.
- La fréquence, f_i, des termes, t_i, présents dans le document dc_i est ensuite calculée.
- Chaque document dc_i est ensuite représenté par un vecteur $\overrightarrow{vc_i} = \{w_{1,i}, w_{2,i}, ..., w_{n,i}\}$ où $w_{n,i}$ est le poids du terme t_i dans le document dc_i.

Présentation du dictionnaire de données

WordNet, base de données lexicales, a été développé depuis 1985 à l'université de Princeton (Fellbaum, 1998). Ce dictionnaire de données offre l'avantage d'être librement et gratuitement utilisable. Il couvre la majorité des noms, verbes, adjectifs et adverbes de la langue anglaise. Ces termes sont structurés sous forme d'un réseau de noeuds et de liens entre eux. Chaque noeud est constitué par un ensemble de synonymes, appelés syn-

sets. La relation de base entre les termes d'un même synset est la synonymie. Un synset contient trois parties : le terme pour lequel le synset est identifié, les termes synonymes et un glossaire qui contient une définition sysnet et éventuellement un ou plusieurs exemples du monde réel.
L'utilisation de WordNet nous sert à la désambiguïsation des termes ainsi qu'à leur extension.

4.2.2.2 Construction de la dimension centres d'intérêt

Afin de construire l'ontologie d'intérêt, nous avons exploité la technique d'apprentissage d'ontologies développée dans OntoCoSemWeb (Mustapha et al., 2007), (Baazaoui et al., 2007). OntoCoSemWeb permet de représenter les connaissances d'un domaine à l'aide d'ontologies. Il se base sur le fait qu'une seule ontologie n'est pas adaptée à la modélisation de toutes les connaissances relatives à un domaine donné sur le Web. L'architecture ontologique sur laquelle il repose est composée d'une méta-ontologie générant trois ontologies interdépendantes :

- ontologie de domaine qui spécifie les connaissances statiques du domaine,
- ontologie des services de domaine qui spécifie les connaissances opérationnelles du domaine et
- ontologie de structure des sites Web qui conceptualise la sémantique de la structuration des connaissances sur le Web. Cette méta-ontologie comprend les éléments abstraits de chacune de ces trois ontologies.

Construction de l'ontologie La construction de l'ontologie est effectuée en trois phases : initialisation, apprentissage incrémental d'ontologie de domaine et analyse des résultats :

- La phase d'initialisation a pour rôle de préparer et de prétraiter des sources de données qui sont constituées d'une ontologie minimale, d'une métaontologie et de l'ontologie générale Wordnet.
- La phase d'apprentissage est caractérisée par son aspect incrémental et itératif. Chaque itération se déroule en deux étapes successives : alimentation de la métaontologie et application des axiomes relatifs à l'apprentissage des éléments de l'ontologie

de domaine.

- La phase d'analyse des résultats a pour rôle de vérifier la cohérence de la métaontologie en effectuant l'analyse des résultats d'apprentissage. Lors de cette phase, l'enrichissement des concepts est effectué en exploitant WordNet. A cet effet, le dictionnaire est consulté pour le concept d'intérêt courant et l'extraction des termes synonymes est alors effectuée. Une fois les termes extraits, il y a vérification de leur existence ou pas dans l'ontologie. Si un terme synonyme existe, il n'est pas alors transcrit dans l'ontologie d'intérêt.

Dans le cadre de notre proposition, nous avons procédé à quelques adaptations du processus décrit ci-dessus. En effet, la première phase d'initialisation n'est sollicitée qu'au premier accès de l'utilisateur connu ou à la première construction du modèle d'un stéréotype d'utilisateurs. Les incrémentations de la seconde phase sont relatives aux soumissions des requêtes de l'utilisateur. Cette phase s'active de ce fait une fois que l'utilisateur soumet une $(i+1)me$ requête. Quant à la troisième phase, elle est activée à la fin de chaque session de recherche.

De ce fait, les étapes de construction de la dimension centres d'intérêt sont :

- Initialisation de l'ontologie : cette étape est effectuée lors de l'initialisation du modèle de l'utilisateur.
- L'alimentation de l'ontologie est effectuées à chaque requête soumise et une fois que l'utilisateur débute sa navigation à travers les résultats de la recherche. Les documents de concepts issus de l'exploitation de l'ODP sont utilisés lors de cette étape qui exploite aussi les mesures d'intérêt de l'utilisateur pour les documents consultés afin d'assigner les pondérations associées aux concepts. Ces pondérations étant elles-mêmes les valeurs de l'intérêt utilisateur.
- L'apprentissage des données de l'ontologie est activé suite à chaque soumission de requête.
- Analyse de cohérence de l'ontologie construite : cette étape est réalisée en mode hors ligne, dès que l'utilisateur se déconnecte du système.

Un extrait d'une ontologie d'intérêt et des ressources utilisées est présenté dans la figure 4.1.

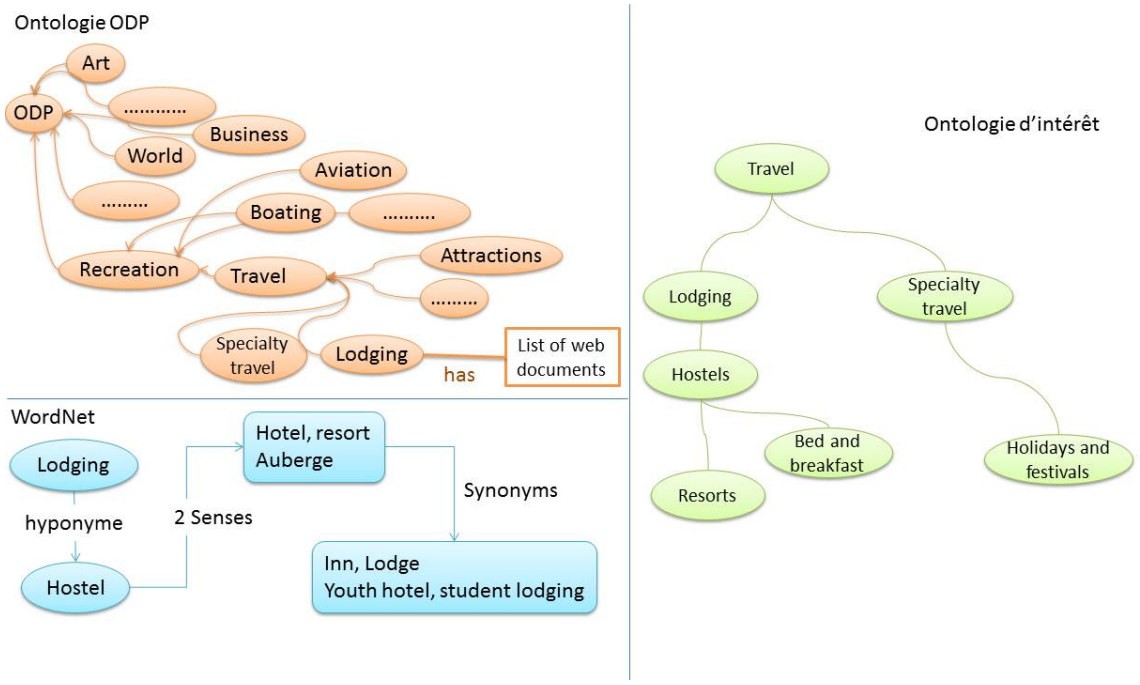


FIGURE 4.1 – Exemple d'utilisation de l'ODP et de WordNet pour la construction de la dimension centres d'intérêt

4.2.2.3 Exploitation de la dimension centres d'intérêt

La dimension centres d'intérêt est exploitée pour l'enrichissement des requêtes de l'utilisateur et pour le réordonnancement des résultats de recherche.

Enrichissement des requêtes de l'utilisateur : Une fois que l'utilisateur soumet sa requête, et comme introduit dans la section 3.5 p. 63, une étape de reformulation de la requête est déclenchée. Lors de cette étape, sont exploités les concepts de la dimension centres d'intérêt ainsi que leur pondération relative à l'intérêt utilisateur. L'enrichissement de la requête proposé est présenté dans la section 5.4.1 p.105.

Réordonnancement des résultats de recherche : A chaque document (lien) est associé un score original et un score personnalisé. Pour une requête q_{t+1}, le ré-ordonnancement des résultats de recherche utilise la combinaison du score initial (S_i) et du score personnalisé (S_p) de chaque document. Le score final du document d_j pour la requête q_{t+1} et le

modèle de l'utilisateur à exploiter M_u^s est calculé suivant la formule suivante (Daoud et al., 2009) :

$$S_{f(d_j)} = \gamma * S_i(q, d_j) + (1 - \gamma) * S_p(d_j, M_u^s) \tag{4.8}$$

Avec $\gamma \in]0, 1[$
Si $\gamma = 0$, l'ordonnancement des résultats sera effectué en tenant compte du score personnalisé en se basant sur leur similitude avec les concepts d'intérêt du modèle de l'utilisateur.
Si $\gamma = 1$, alors les résultats seront ordonnancés suivant leur score original (score calculé sans intégration de la personnalisation).
Le score personnalisé S_p est calculé selon une mesure de similarité cosine entre le document et les concepts de la dimension centres d'intérêt. Cette formule est la suivante :

$$S_p(d_j, M_u^s) = \frac{1}{h} \sum_{j=1..h} score(c_j) * \cos(\overrightarrow{d_k}, \overrightarrow{c_j}) \tag{4.9}$$

Où c_j est un concept de la dimension centres d'intérêt de l'utilisateur, $score(c_j)$ la pondération de c_j et h est le nombre le nombre de concepts considérés dans le processus de personnalisation. h sera paramétré dans la phase expérimentale, section 6.3.

4.2.3 La dimension spatiale

Cette dimension est composée d'entités informationnelles spatiales dont la représentation est sous forme d'arbre de concepts pondérés. Cette structuration est basée sur l'utilisation de concepts spatiaux et d'attributs caractéristiques de ces concepts.

Les concepts spatiaux : Les entités informationnelles sur les intérêts spatiaux sont classées en un ensemble de concepts spatiaux auxquels sont associés des attributs. Ces derniers ont pour objectif de décrire les entités leur appartenant.

Les attributs caractéristiques des concepts spatiaux : Dans l'objectif de ne pas limiter la dimension spatiale à un ensemble de relations avec les concepts spatiaux, nous introduisons une description plus concrète de chaque concept par l'ajout d'un ensemble

d'attributs décrivant les entités spatiales. Ceci nous semble nécessaire car les entités spatiales d'un même concept diffèrent par les valeurs de leurs attributs et de ce fait par l'intérêt que leur porte l'utilisateur. En effet, les intérêts des utilisateurs pour un même concept spatial changent en fonction donc des préférences pour les attributs.

Afin de clarifier ces différents composants de la dimension spatiale, nous proposons de continuer d'exploiter l'exemple utilisé dans la section 4.2.1 à la page 73. Dans cet exemple, l'utilisateur a effectué 3 sessions de recherche avec un total de 4 requêtes émises. Les concepts spatiaux extraits de ces requêtes sont *Paris* et *montagne*. En plus de l'information d'emplacement fournie dans la requête, à savoir la ville de *Paris*, les informations spatiales extraites à travers les différentes navigations de cet utilisateur concernent tous, toute session confondue, des emplacements autour de *Paris*.

Une quatrième session de recherche est effectuée et qui permet d'avoir d'autres attributs pour le concept *Paris* dont une instance est :

$$Paris = \{hotel, ChampsElysees\}$$

Où les informations concernant le concept *Paris* est représenté comme suit :

TABLE 4.1 – Instance d'un concept de la dimension spatiale de l'utilisateur u

Concept	Attribut	Caractéristiques	Entités-Liens
Paris	Hotel	URL26	
		Standing	4étoiles
		Piscine	Oui
		Distance Plage	500m
		URL27	
		Standing	3étoiles
		Piscine	Couverte
		Distance Plage	1500m

Chaque entité spatiale (l'équivalent d'une instance d'objet) de ce concept devrait donc avoir, en plus de sa localisation, une valeur caractéristique pour chaque attribut de ce concept pour pouvoir ainsi distinguer entre les entités par comparaison de leurs valeurs.

La pondération associée à chaque concept spatial est correspondante à l'intérêt implicite de l'utilisateur envers cette entité (Hadjouni et al., 2009b). L'enrichissement de cette dimension est aussi effectué en utilisant le dictionnaire de données Wordnet. Le schéma de

la figure 4.2 montre un exemple de représentation d'une partie de la dimension spatiale.

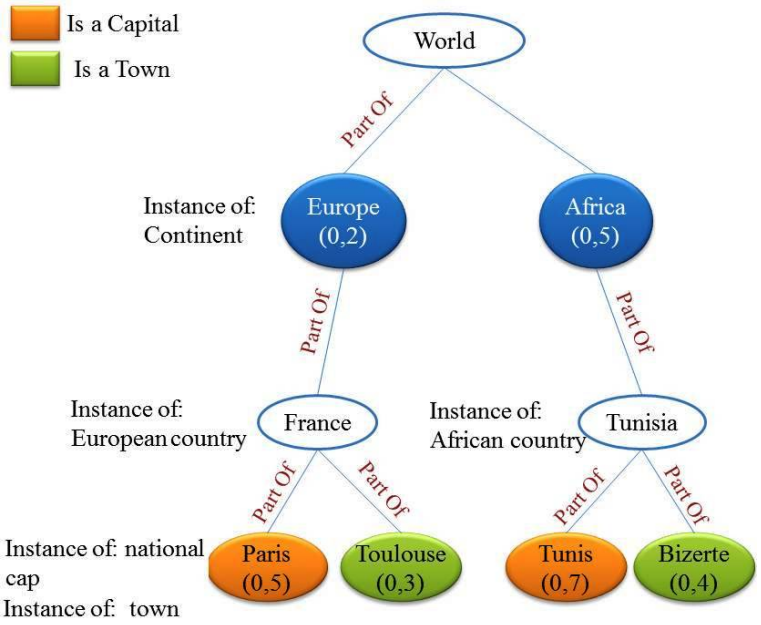


FIGURE 4.2 – Exemple de dimension spatiale du modèle de l'utilisateur

4.2.4 La dimension des données personnelles de l'utilisateur

La dimension des données personnelles de l'utilisateur est une dimension qui contient les données explicites fournies par l'utilisateur. Ces données sont son code d'accès au système ainsi que son mot de passe. Afin d'avoir plus de détails de l'utilisateur, nous sauvegardons aussi dans cette dimension les différentes adresses IP utilisées ainsi que les données des cookies (dans le cas où le navigateur de l'utilisateur le permet). Cette dimension ne concerne que les utilisateurs enregistrés au système. En effet, n'étant pas inscrit, un utilisateur va être associé à un stéréotype et de ce fait bénéficier du modèle de stéréotype.

4.3 Construction du modèle de l'utilisateur

La construction du modèle de l'utilisateur est basée sur la prise en compte implicite et interactive des données enregistrées à partir des navigations de l'utilisateur à travers le système. Ces données sont dites implicites car l'utilisateur n'est pas amené à fournir des informations sur ses préférences, et interactives car nous utilisons l'historique de la

navigation pour mesurer son intérêt pour une entité donnée. Ces mesures sont basées sur :

- Les similarités qui pourraient exister entre les attributs et les entités d'intérêt : Pour les similarités entre les valeurs des attributs, nous distinguons les différents types qui peuvent exister (les valeurs numériques, des intervalles numériques, les ensembles ...). Pour le calcul de la similarité entre entités, nous considérons l'agrégation des degrés de similarités existant entre les attributs de ces entités.
- La déduction des intérêts des utilisateurs à partir de l'ensemble de leurs navigations : nous considérons pour cela la fréquence et les durées des visites effectuées. Nous avons appelé ce données indicateurs d'intérêt (Hadjouni et al., 2009a).
- Le calcul de la pertinence d'un éventuel déplacement spatial. En effet, dans notre approche, nous considérons également que si un utilisateur demande ou sélectionne une zone spatiale, c'est qu'il a l'envie de s'y déplacer. Cette déduction est effectuée en corrélation avec la déduction des intérêts des utilisateurs à partir de l'ensemble de leurs navigations (Hadjouni et al., 2009b).

Les deux principales étapes de la construction du modèle de l'utilisateur sont (a) l'initialisation du modèle de l'utilisateur et (b) la construction des différentes dimensions constituant le modèle. Le processus de construction global est décrit par l'algorithme 1.

Algorithme 1 Algorithme de construction du modèle de l'utilisateur

Entrées : $Q_t = \{w_1, w_2, ..., w_n\}$
$u \in U$
U l'ensemble des utilisateurs
Sortie : $M_u = \{D_u^{tx}, D_u^{sp}, D_u^{in}\}$: le modèle de l'utilisateur
DEBUT
**Cas d'un utilisateur enregistré dans le système*
Initialisation modèle de l'utilisateur Mu
pour chaque lien accédé **faire**
  Enregistrer les navigations
  Calculer l'intérêt
  Remplir la dimension textuelle
  ** Inférer la dimension centres d'intérêt*
  Calculer I_e
  $InfDInteret(lien, I_e, d_x, Q_t, Mu)$
  **Inférer la dimension spatiale*
  $InfDSaptiale(lien, I_e)$
fin pour
**Cas d'un utilisateur non enregistré dans le système*
Associer l'utilisateur à un stéréotype
pour chaque lien accédé **faire**
  Enregistrer les navigations
  Mettre à jour les données du modèle de stéréotype
fin pour
Retourner M_u
FIN

Les fonctions d'inférence des dimensions centres d'intérêt (InfDInteret) et spatiale (InfDSaptiale) sont respectivement présentées par les algorithmes 2 p. 88 et 3 p.89.

La construction du modèle de l'utilisateur est effectuée de manière itérative. Les itérations débutent avec la première interaction de l'utilisateur et la construction du modèle évolue avec les recherches et les navigations au cours de la session de recherche (figure 4.3). Deux cas de figures, pour un utilisateur, se présentent : soit qu'il est connecté au système à travers ses nom d'utilisateur est mot de passe, et de ce fait il est reconnu, soit qu'il commence sa session de travail de manière anonyme, nous disons alors qu'il est inconnu du système. Pour le premier cas, le niveau modèle de l'utilisateur du système active le modèle de l'utilisateur correspondant et en effectue la mise à jour au cours de la navigation.
Si l'on considère $M_u(t_0 = 0)$ ce modèle de l'utilisateur au moment de l'activation ($t_0 = 0$),

nous aurons à $(t_1 = t_0 + \lambda)$:

$$M_u(t_1 = t_0 + \lambda) = M_u(t_0) + Z_i + Sim(x) + I_e(u, x) \tag{4.10}$$

Où l'opérateur + signifie la prise en compte de : la zone de recherche, la similarité et l'intérêt.

Avec :

Z_i la zone de recherche de l'utilisateur

$Sim(x)$ la similarité pouvant exister entre les attributs des résultats de la recherche et les entités d'intérêt x (Hadjouni et al., 2009b).

$I_e(u, x)$ la valeur d'intérêt de l'utilisateur

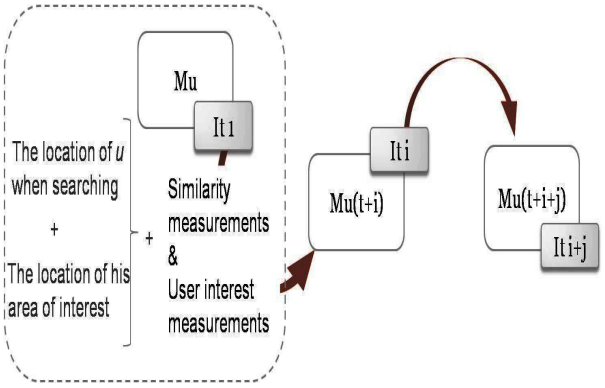


FIGURE 4.3 – Construction du modèle de l'utilisateur

4.3.1 Inférence de la dimension centres d'intérêt

La dimension centres d'intérêt, comme présentée dans la section 4.2.2, contient les concepts d'intérêt relatif à l'utilisateur. Ces concepts sont construits en se basant sur les différentes requêtes émises par l'utilisateur ainsi que sur l'utilisation de l'ontologie ODP et du dictionnaire de données WordNet. L'inférence de la dimension centres d'intérêt est de ce fait effectuée en utilisant le calcul d'intérêt de l'utilisateur pour les documents consultés. Nous distinguons un ensemble de mesures ayant pour objectif de déduire les centres d'intérêt de l'utilisateur et qui sont mesurées tout au long de sa navigation. En effet, lors des différents parcours de l'utilisateur à travers les résultats de recherche, nous mémorisons les actions déclenchées par celui-ci. Les principales mesures employées sont la fréquence et

la durée des visites pour un résultat (Hadjouni et al., 2009a). L'intérêt de l'utilisateur est estimé à deux niveaux :

- par rapport aux concepts pour déterminer ceux qui l'intéressent le plus et dans un niveau de granularité plus fin
- par rapport aux valeurs caractéristiques des entités appartenant à chaque concept. Pour cela nous, proposons la formule suivante et qui estime le degré d'intérêt d'un utilisateur pour une entité visitée

4.3.1.1 Intérêt de l'utilisateur pour les concepts

Pour pouvoir classer les concepts selon le degré d'intérêt de l'utilisateur, nous utilisons les mesures d'intérêt pour les entités (spatiales ou non) relatives aux concepts. En effet, la pertinence d'un concept par rapport aux besoins de l'utilisateur se calcule en prenant en considération les degrés d'intérêt qu'il a montré à l'égard des entités de ce concept.
Soient C un concept et x_i une entité visitée par l'utilisateur u et appartenant à C. Le degré d'intérêt que montre un utilisateur u pour ce concept, noté $I_c(U, C)$, est :

$$I_c(U, C) = \sum_{i=1}^{n} I_e(U, x_i) \tag{4.11}$$

n étant le nombre des entités visitées du concept C.
$I_e(U, x_i)$: degré d'intérêt de l'utilisateur U à x_i.

4.3.1.2 Intérêt de l'utilisateur pour les entités visitées

Nous adoptons les notations suivantes : Soit x une entité (spatiale ou non) visitée par l'utilisateur.
$d_v(U, x)$: Durée de visites de l'utilisateur u pour l'entité x.
$D_v(U)$: Durée totale de toutes les visites de l'utilisateur u.
$n_v(U, x)$: Nombre de visites de l'utilisateur u pour l'entité x.
$N_v(U)$: Nombre total de visites de l'utilisateur u.
Le degré d'intérêt de l'utilisateur u pour l'entité x, noté $I_e(U, x)$ est :

$$I_e(U,x) = \left(\frac{d_v(U,x)}{D_v(U)} + \frac{n_v(U,x)}{N_v(U)} \right) \tag{4.12}$$

4.3.1.3 Intérêt de l'utilisateur pour les valeurs caractéristiques des entités

Après avoir introduit le critère de classement des concepts par degrés d'intérêt, il nous reste à ordonner par pertinence les entités appartenant à un même concept en exploitant une pondération basée sur l'intérêt porté à ces entités. Ceci nécessite la déduction des valeurs caractéristiques des entités qui intéressent l'utilisateur. Cette étape se fait par analogie aux concepts, c'est-à-dire que la note d'intérêt pour une valeur donnée est calculée en fonction de la durée et de la fréquence des visites pour les entités ayant cette valeur. Cette note pourra être raffinée par la prise en compte des votes de l'utilisateur pour ces entités. Soient F_i une caractéristique du concept C et v_k une valeur possible de F_i. Soit $X(F_i, v_k)$ l'ensemble des entités visitées par l'utilisateur u et qui ont la valeur v_k pour l'attribut F_i. Le calcul de l'intérêt de u pour la valeur v_k est :

$$I_f(U,F_i,v_k) = \sum_{x_j \in X(F_i,v_k)} I_e(U,x_j) \tag{4.13}$$

Après avoir calculé l'intérêt de l'utilisateur pour les différentes valeurs possibles des attributs, nous procédons à leur normalisation.

L'algorithme de la fonction d'inférence de la dimension centres d'intérêt est :

Algorithme 2 Algorithme de la fonction InfDInteret

Variables : lien consulté
d_x : durée de consultation du lien
Q_t : la requête de l'utilisateur
M_u : le modèle de l'utilisateur courant
I_e

Sortie : M_u mis à jour
DEBUT
Rechercher le concept associé dans D_{in}
{Mettre à jour la pondération du concept}
si le concept existe dans D_{in} **alors**
  Mettre à jour I_c
sinon
  Ajouter le concept
  Mettre à jour I_c
finsi
FIN

4.3.2 Inférence de la dimension spatiale

Afin d'inférer les données de la dimension spatiale, nous introduisons la notion de satisfaction des contraintes spatiales. Les contraintes spatiales doivent prendre en considération les règles suivantes :

1. Les entités spatiales les plus proches de la position actuelle supposée de l'utilisateur ou bien de son choix sur la carte spatiale sont les plus probables à choisir
2. Les relations spatiales et de similarité sémantique entre les entités sont asymétriques
3. La relation spatiale entre deux entités est affaiblie quand le nombre d'entités proches de l'entité référence augmente.
4. L'intérêt pour une entité spatiale augmente lorsque celle-ci est entourée par des entités similaires.
5. L'intérêt pour une entité spatiale augmente lorsque le nombre d'entités préférées qui l'entourent augmente.

Lors de la personnalisation nous supposons que l'utilisateur est positionné sur une entité x (la dernière visitée) pour calculer les entités les plus probables à visiter après celle-ci en

la considérant comme référence.

L'algorithme de la fonction d'inférence de la dimension spatiale :

Algorithme 3 Algorithme de la fonction InfDSpatiale

Variables : lien consulté
M_u : le modèle de l'utilisateur courant
I_e

Sortie : M_u mis à jour
DEBUT
Rechercher le concept spatial associé dans D_{sp}
{Mettre à jour la pondération du concept}
si le concept n'existe pas dans D_{sp} **alors**
  Ajouter le concept
finsi
Calculer la mesure d'utilité
Calculer l'accessibilité du concept
Mettre à jour la pondération du concept
FIN

4.4 Mesure d'accessibilité spatiale proposée

Comme présenté dans l'état de l'art, un processus de personnalisation doit prendre en considération les préférences, les contraintes et les besoins de l'utilisateur. Néanmoins, dans le cadre d'une information à caractère spatial, la topologie et la distribution spatiale des lieux ont une influence non négligeable sur les choix de l'utilisateur et sur son évaluation de la pertinence des résultats de recherche personnalisés qui lui sont fournis. En effet, les notions de coût, de distance, de contraintes de déplacement et de localisation peuvent jouer un rôle important lors d'une recherche locale d'une information spatiale. Dans ce contexte, la notion d'accessibilité est, à notre avis, un facteur à prendre en considération lors de la proposition de services personnalisés à caractère spatial.

L'accessibilité est définie comme la facilité avec laquelle un lieu peut être atteint, elle est aussi utilisée pour étudier la topologie de l'espace et son utilisation (Batty, 2004). Les mesures d'accessibilité permettent d'étudier et de comprendre la manière avec laquelle l'espace est utilisé. C'est dans ce contexte que nous présentons la mesure d'accessibilité proposée afin d'améliorer la qualité des informations personnalisées fournies à l'utilisateur.

4.4.1 Prérequis à la mesure d'accessibilité

La mesure d'accessibilité ayant été présentée comme la facilité avec laquelle un lieu peut être atteint, nous nous proposons d'associer à cette notion la notion d'utilité. En effet, une première étape dans la proposition d'une mesure d'accessibilité est la définition des facteurs spatiaux qui engendrent la manière avec laquelle l'espace est utilisé. Dans cette optique, les règles de départ sont :

R(1) Les entités les plus proches sont les plus probables à être visitées.

R(2) La relation spatiale entre deux entités est affaiblie lorsque le nombre d'entités proches de l'entité référence augmente.

R(3) L'intérêt pour une entité spatiale augmente lorsque celle-ci est entourée par des entités similaires.

R(4) L'intérêt pour une entité spatiale augmente lorsque le nombre d'entités préférées qui l'entourent augmente.

R(5) Afin de mettre en évidence les quatre règles ci-dessus, une mesure d'utilité est proposée.

Lors de la personnalisation, nous supposons que l'utilisateur est positionné sur une entité x (ou le cas échéant, nous considérons la dernière entité visitée) afin de retrouver les entités les plus probables à visiter.
Afin de répondre à ces règles définies, nous utiliserons les mesures suivantes :

- Proximité contextuelle : correspondant et répondant aux règles $R(1)$ et $R(2)$
- Proximité des entités similaires : utilisée pour répondre à la règle $R(3)$
- Proximité des entités préférées de l'utilisateur : utilisée pour répondre à la règle $R(4)$

Proximité contextuelle : En partant de la règle $R(2)$, nous nous nous proposons d'utiliser une fonction de calcul de la proximité contextuelle entre deux entités (Yang and Chan, 2005). Cette fonction, sous forme de distance inversée, est dans l'intervalle $[0, 1]$. Soient les entités spatiales x et y, y étant l'entité sur laquelle est positionné l'utilisateur et x l'entité candidate, la formule de proximité contextuelle est :

$$P_c = \frac{1}{1 + D_c(x, X)} \tag{4.14}$$

D_c est la distance contextuelle D proposée par Worboys (Worboys, 1996) comme une formulation de la distance relativisée :

$$D_c(x, y) = \frac{d(x, y)}{d(x, X)} \tag{4.15}$$

Avec, $d(x, y)$ la distance euclidienne entre x et y
$d(x, X)$ la distance moyenne entre x et les entités de l'ensemble des entités spatiales X.
X est l'ensemble de toutes les entités dans le voisinage de x réalisable par rapport aux contraintes de déplacement de l'utilisateur ($R(1)$ et $R(2)$.

Proximité des entités similaires : De manière analogue à la mesure précédente, nous évaluons cette mesure en utilisant la proximité moyenne entre l'entité candidate x ainsi que l'ensemble des entités qui lui sont similaires dans son entourage noté $Sim(x)$. Ainsi, la proximité moyenne, notée $P_m(x, Sim(x))$ augmente lorsque des entités similaires à x sont présentes dans son entourage, respectant ainsi la règle R(3).

Proximité des entités préférées : Le but de cette mesure est d'évaluer la proximité des entités préférées par l'utilisateur u autour d'une entité candidate. Pour cela, nous avons recours au calcul de la proximité moyenne, notée $P_m(x, Pref(u))$ entre cette entité x et l'ensemble des entités susceptibles d'être préférées par l'utilisateur $Pref(U)$. $Pref(U)$ représente l'ensemble des entités préférées par l'utilisateur, dont l'intérêt calculé par la fonction I_e (cf. section 4.3.1.2) est supérieur à la moyenne de ses notes d'intérêt aux entités déjà visitées. La proximité moyenne augmente avec l'augmentation du nombre d'entités préférées par l'utilisateur et proches de x.

4.4.2 Proposition d'une mesure d'utilité orientée utilisateur

N'étant pas dans le cadre d'une recommandation collaborative, nous nous sommes penchés sur la possibilité de fournir un résultat personnalisé en fonction du calcul des

mesures d'intérêt pour les entités spatiales visitées et des prédictions sur les entités utiles à fournir. A cet effet, et en partant des règles ($R(1)$–$R(4)$) définies dans la section précédente, nous introduisons l'utilisation de la notion d'utilité d'une entité. Le calcul de l'utilité d'une entité est basé sur les données suivantes :

- La correspondance de l'entité aux préférences de l'utilisateur (cf. section 4.3.1.3 formule 4.13, p. 87).
- La proximité contextuelle de l'entité (cf. section 4.4.1 formule 4.14, p. 91).
- La proximité moyenne de l'ensemble des entités préférées de l'utilisateur (cf. section 4.4.1)
- La proximité moyenne de l'ensemble des entités similaires à une entité (cf. section 4.4.1)

La formule d'utilité a donc pour objectif de permettre de prédire les entités passibles d'intéresser l'utilisateur. La formule est la suivante :

$$P_{di}(u,x,y) = \alpha \times I_e(u,x) + \beta \times P_c(x,y) + \lambda \times P_m(x, Pref(u)) + \gamma \times P_m(x, Sim(x)) \quad (4.16)$$

Avec

$\alpha + \beta + \lambda + \gamma = 1$

x étant l'entité candidate à la personnalisation pour laquelle la formule prédit le degré d'intérêt de l'utilisateur u.

y est l'entité référence : entité où se positionne l'utilisateur.

$I_e(u,x)$ est le degré de satisfaction des préférences de l'utilisateur par l'entité x.

$P_c(x,y)$ est la mesure de proximité contextuelle (cf. section 4.4.1 formule 4.14).

$P_m(x, Pref(u))$ est la proximité moyenne de l'ensemble des entités préférées de l'utilisateur

$P_m(x, Sim(x))$ est la proximité moyenne de l'ensemble des entités similaires à une entité

α, β, λ et γ sont les coefficients qui normalisent le résultat et qui déterminent l'importance et le degré de contribution de chaque facteur.

Le cas pour lequel l'utilisateur ne s'est pas encore localisé est traité par la formule (4.17). Cette formule est caractérisée par l'absence du calcul de la proximité contextuelle.

$$P_{di}(u,x,y) = \alpha \times I_e(u,x) + \lambda \times P_m(x, Pref(u)) + \gamma \times P_m(x, Sim(x)) \quad (4.17)$$

Dans la définition de la mesure d'utilité, les pondérations α, β, λ et γ sont les coefficients qui déterminent l'importance de chaque terme de la mesure. Ces coefficients dépendent du contexte d'utilisation et peuvent varier pour un même utilisateur. Par exemple, dans les voyages thématiques, l'utilisateur a tendance à visiter les entités entourées par d'autres entités similaires et cela dans le cadre de la thématique recherchée. Par contre, dans un autre contexte, l'utilisateur peut donner plus d'importance à des facteurs comme la pertinence ou la proximité d'entités préférées mais pas dans la même thématique. Afin de tenir compte des variations propres à chaque utilisateur, ces coefficients sont déterminés de manière dynamique. Nous appliquons un algorithme d'optimisation qui, selon la qualité de la personnalisation, agit sur ces valeurs afin de les améliorer (Hocquet and Bazin, 2008). La figure 4.4 présente le processus de paramétrage de la mesure d'utilité optimisant la qualité de la personnalisation.

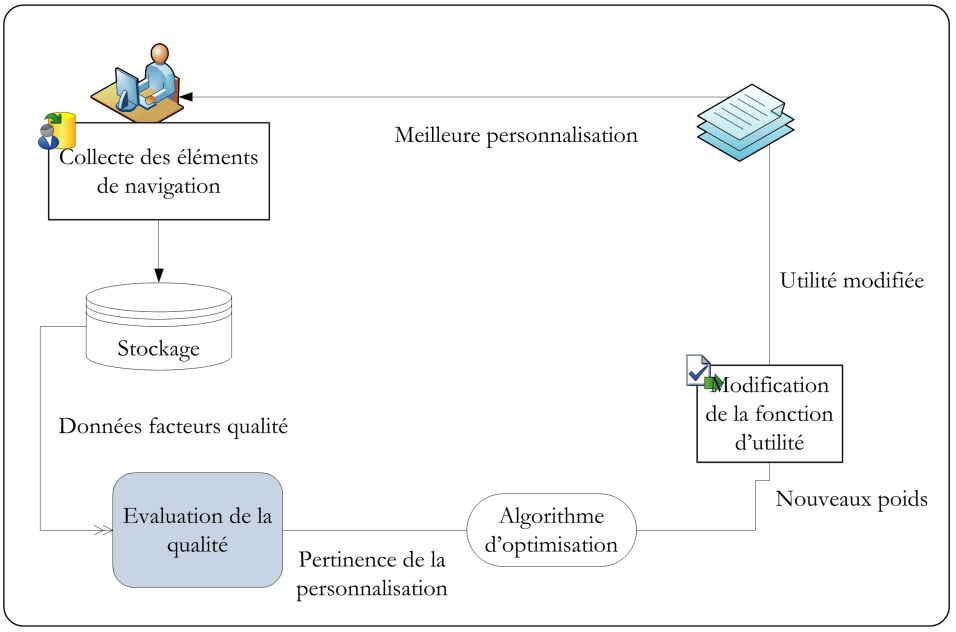


FIGURE 4.4 – Processus de paramétrage de la mesure d'utilité

Le processus d'optimisation et de personnalisation des pondérations est réalisé par l'algorithme de recuit simulé. Les tâches effectuées sont les suivants :

- Récupération de la configuration courante(α, β, λ et γ).
- Sélection des entités les plus accessibles par rapport à la configuration courante de

la mesure d'utilité.

- Récupération des éléments relatifs à l'interaction de l'utilisateur avec les résultats affichés (temps de visite, vote, etc. . .)
- Evaluation de la qualité des résultats personnalisés par agrégation des mesures de qualité.
- Mesurer l'écart de qualité entre la configuration courante et la configuration optimale.
- Déduction de la nouvelle configuration ($\alpha\prime$, $\beta\prime$, $\lambda\prime$ et $\gamma\prime$) à adopter comme configuration courante en utilisant le principe du recuit simulé.

La fonction de transformation FT utilisée pour calculer la nouvelle configuration ($\alpha\prime$, $\beta\prime$, $\lambda\prime$ et $\gamma\prime$)se base sur les caractéristiques de la recherche personnalisée R acceptée par l'utilisateur afin mettre à jour la configuration courante. Soient V_{ci} la valeur prise par l'entité R par rapport à une des composantes de la mesure d'utilité ($I_e(u,x)$,$P_c(x,y)$,$P_m(x,Pref(u))$ ou $P_m(x,Sim(x))$). Le calcul du nouveau poids x_0 lié à V_c se fait à partir de la valeur de l'ancien poids x, comme le montre la formule suivante :

$$FT(pi) = \frac{1}{2} * (p_i + \frac{V_{ci}}{\sum_i V_{ci}}) \tag{4.18}$$

4.4.3 Présentation de la mesure d'accessibilité proposée

Nous adoptons dans cette proposition les notions d'attraction et d'impédance car elles reflètent la logique de déplacement d'un utilisateur et font référence respectivement à la personnalisation et à l'aspect spatial. Cette agrégation permet de prendre en considération les aspects sémantique et spatial de l'information mise à disposition sur le Web. En effet, elle permet de mieux modéliser le contexte d'utilisation et le processus de prise de décision de l'utilisateur et ainsi de l'aider à filtrer la masse d'information disponible en déterminant le contenu pertinent recherché.

L'approche de personnalisation proposée repose sur l'utilisation des navigations et sélections des utilisateurs afin de construire leurs modèles. Nous mesurons à cet effet les intérêts de l'utilisateur sur les données visitées (cf. section 4.2.2), et plus particulièrement sur les entités visitées, et procédons à la prédiction des entités les plus probables à visiter (cf. section 4.4.2).

La mesure d'utilité proposée dans la section 4.4.2 sera considérée comme une valeur d'attractivité dans la mesure d'accessibilité. Ceci va servir pour l'évaluation de l'attractivité (pertinence) d'une entité par rapport au modèle de l'utilisateur. L'attractivité devient ainsi une notion plus personnelle à l'utilisateur.
Quant à la notion d'impédance, nous considérons qu'elle peut être une combinaison de plusieurs facteurs liés au déplacement et non pas seulement la distance du parcours (Handy and Niemeier, 1997). Cette mesure peut aussi être personnalisée selon les préférences de l'utilisateur pour mieux approcher sa logique d'évaluation de l'impédance et donc améliorer la qualité de la mesure d'accessibilité.
Analogiquement aux mesures étudiées, nous considérons que la valeur d'attraction d'une entité en un point donné est affaiblie en fonction de l'impédance du déplacement à ce point. Soient $V(x_i)$ un ensemble d'entités appartenant au voisinage de x_i et d_{ij} l'impédance du déplacement de x_i à x_j. La mesure d'accessibilité proposée est la suivante :

$$A(x_i) = \sum_{x_j} P_u(x_j) * e^{-\frac{d_{ij}^2}{\beta_u^2}} \tag{4.19}$$

Avec β_k la permissivité (tolérance) par rapport coût du déplacement de l'utilisateur u.

Plus β_k augmente, plus les déplacements coûteux sont permis. Nous avons ajouté à l'utilisateur la possibilité de fournir au système cette information afin de lui permettre d'avoir des résultats plus adaptés spatialement.

4.5 Conclusion

Le contenu de ce chapitre a été découpé en trois grandes parties qui ont débuté par la présentation du système de recherche d'information personnalisée sémantique et spatiale (SyPRISS). Ce système se base sur la construction d'un modèle multidimensionnel de l'utilisateur et sur l'utilisation de ce dernier afin de construire un réseau de modèles utilisateurs. Ceci a fait l'objet de la seconde partie du chapitre. Etant dans le cadre de l'utilisation d'informations à caractère spatial, nous avons proposé de prendre en considération l'accessibilité aux entités spatiales choisies et à choisir par l'utilisateur. A cet effet,

nous avons utilisé une mesure d'accessibilité basée sur les préférences de l'utilisateur afin de lui fournir un meilleur résultat que nous exploitons dans la construction de dimension spatiale du modèle de l'utilisateur.

Nous procédons, dans le chapitre suivant, à présenter le réseau de modèles utilisateurs et sa construction. Ce réseau a pour objectif de proposer à l'utilisateur des informations personnalisées en se basant sur un filtrage collaboratif implicite à travers les modèles des utilisateurs qui le constituent. Nous détaillerons aussi la construction globale des résultats de recherche personnalisés proposés à l'utilisateur.

Chapitre 5

Construction d'un réseau de modèles utilisateurs pour une recherche d'information personnalisée sémantique et spatiale

5.1 Introduction

Comme présenté dans le chapitre 2, le système de personnalisation proposé repose sur la construction d'un modèle multidimensionnel de l'utilisateur ainsi que sur la construction d'un réseau de modèles des utilisateurs. La présentation du modèle de l'utilisateur a fait l'objet du chapitre 3. Ce chapitre a pour objectif de présenter le réseau de modèles ainsi que de sa construction. L'objectif de l'exploitation d'un tel réseau est de permettre une collaboration implicite entre les utilisateurs et cela à travers l'exploitation de leurs modèles. Nous présentons aussi à la suite du réseau la construction globale des résultats fournis à l'utilisateur. Cette construction est de ce fait basée sur la modélisation de l'utilisateur ainsi que sur la construction et l'exploitation du réseau de modèles d'utilisateurs.
Ce chapitre est organisé comme suit. La section 5.2 présente le réseau de modèles utilisateurs, sa structure globale ainsi que les mesures qui relient les noeuds du réseau entre eux. La section 5.3 présente le processus de construction du réseau. Nous abordons la construction des résultats fournis à l'utilisateur dans la section 5.4. Cette construction se base sur l'intégration de la totalité des éléments présentés dans les deux chapitres 4 et 4. La section 5.5 présente une proposition de retour d'expérience de l'utilisateur. La dernière section

conclut le chapitre.

5.2 Le réseau de modèles utilisateurs dans SyPRISS

5.2.1 Structure globale du réseau

Le système repose sur la construction d'un réseau basé sur les modèles des utilisateurs. Notre hypothèse est que, lors de la recherche d'un document pertinent, le système devrait, en plus des besoins spécifiques des utilisateurs et de leurs recherches antérieures, exploiter les connaissances extraites des autres modèles utilisateurs existants. Les noeuds du réseau représentent des modèles d'utilisateurs qui sont interconnectés par des distances spatiales et sémantiques (cf. section 5.2.2). Un tel réseau permet aux utilisateurs :

- d'avoir des résultats correspondant à leurs propres préférences (implicites)
- de bénéficier des résultats du voisin le mieux correspondant sémantiquement (via les critères de recherche) et spatialement (à travers les différentes positions spatiales de recherche de ce noeud).

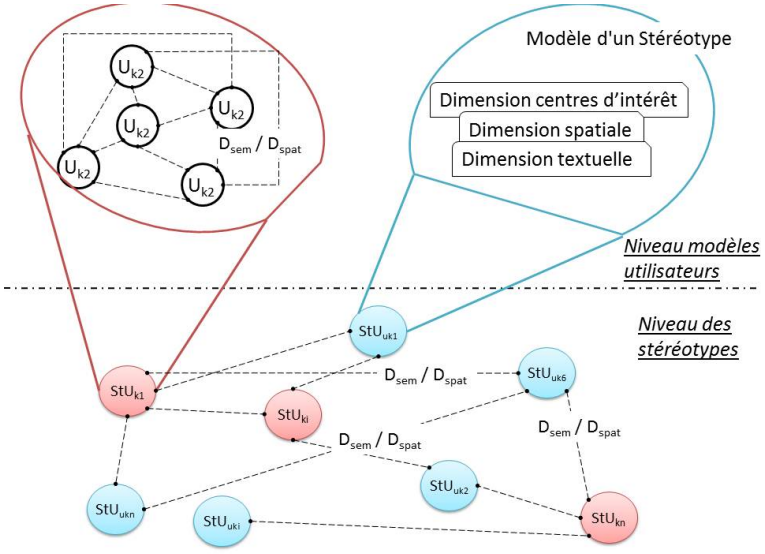


FIGURE 5.1 – Le réseau d'utilisateurs
Sur cette figure, les deux niveaux sont mis de manière superposée. Su représentent les noeuds des modèles des utilisateurs non enregistrés et u représente les noeuds des modèles des utilisateurs enregistrés dans le système.

Notre contribution réside dans la construction d'un réseau représenté sur deux couches : une première couche donnant accès aux modèles des utilisateurs connus, et une seconde couche pour les modèles des stéréotypes des utilisateurs inconnus, c.f. figure 5.1. Les arcs séparant les noeuds de la représentation visuelle en graphe sont calculés en fonction de distances sémantique et spatiale (Hadjouni et al., 2009b). Le choix de ces deux distances est basé sur le fait qu'un utilisateur effectue généralement une recherche textuelle et peut avoir besoin d'informations d'ordre spatial. Nous avons, de ce fait, une représentation sous forme de graphe valué dans lequel les distances sont associées aux arcs.
La proximité entre deux entités spatiales au sein du graphe sera donc exprimée à travers les relations et les autres entités de ce graphe. Nous définissons donc la représentation du réseau en graphe par :

Définition 1 (Le graphe représentatif du réseau d'utilisateurs) *Soient U l'ensemble des utilisateurs, M l'ensemble des modèles utilisateurs associés, nous définissons $G_M(V, E, \vartheta, \sigma, \varphi)$ le graphe associé au réseau de modèles utilisateur, avec :*

1. $V = \{M_u^i, tq.M_u^i \in M\}$ *est l'ensemble des noeuds*
2. $E = \{s_i, tq.s_i \in V\}$ *est l'ensemble des arcs*
3. $\vartheta :\longrightarrow (E, M)$ *est la fonction de recherche du noeud le plus proche*
4. $\sigma \longrightarrow [0, 1]$ *est la fonction de pondération des noeuds*
5. $\varphi \longrightarrow [0, 1]$ *est la fonction de pondération des arcs*

La fonction de recherche du noeud le plus proche ϑ permet, à partir du noeud courant, i.e. le modèle de l'utilisateur effectuant une recherche dans le système, de retrouver le noeud le plus proche en fonction de la pondération de similarité existante.

La fonction de pondération des noeuds σ a pour objectif de mettre à jour la pondération du noeud courant en fonction des navigations courantes de l'utilisateur à travers les résultats de recherche.

La fonction de pondération des arcs φ calcule et met à jour les pondérations en similarité des arcs. L'ensemble des noeuds similaires au noeud courant est mis à jour en fonction des

changements effectués par φ.

Nous proposons de représenter le graphe par une matrice d'adjacence. Les lignes et les colonnes de cette matrice représentent les noeuds. Les valeurs contenues dans la matrice sont les distances entre les noeuds. Nous définissons cette matrice comme suit :

Définition 2 (Matrice représentative du graphe) *Soient U l'ensemble des utilisateurs, M l'ensemble des modèles utilisateurs associés, n le nombre de modèles utilisateurs. Nous définissons la matrice carrée à n lignes $M_x(n*n)$, d'ordre $|V(Mx)|$ dont les valuations sont les pondérations des arcs reliant les noeuds telle que : M_x=$(p_{i,j})$ tq. $1 \leq i \leq n$ et $1 \leq j \leq n$ et avec p la pondération des noeuds i et j*

- *p est calculé avec la fonction φ de la définition 1*
- *Si $M_x[i][j] = 0$, les noeuds i et j sont dits indépendants.*
- *Le graphe étant non orienté, M_x est symétrique par rapport à sa diagonale ascendante*[9].

Dans ce qui suit, nous présentons les mesures de similarité utilisées dans le réseau.

5.2.2 Les mesures de similarité entre les noeuds

L'objectif des mesures de similarité à introduire est de renseigner sur le degré de correspondance entre deux noeuds du réseau qui représentent respectivement deux modèles d'utilisateurs. Etant donné que nous sommes dans le cadre d'une recherche combinant les informations spatiales et sémantiques des modèles des utilisateurs, nous nous proposons de fournir une mesure de similarité prenant en considération ces deux aspects.

5.2.2.1 La similarité sémantique

La mesure de similarité sémantique proposée a été inspirée de la mesure de Wu et Palmer (Wu and Palmer, 1994). En effet, étant dans le cadre d'une comparaison de deux concepts au sein d'un graphe, il nous parait intéressant d'exploiter aussi bien la profondeur des concepts considérés que la distance sémantique entre deux concepts indépendamment

9. Dans ce cas, nous pouvons ne mémoriser que la composante triangulaire supérieure de la matrice d'adjacence

du graphe.

La similarité sémantique est calculée en considérant pour chaque noeud de modèle de l'utilisateur les concepts d'intérêt pondérés. Ces concepts sont valués à l'aide de l'intérêt implicite de l'utilisateur.

Deux noeuds de modèles utilisateurs sont dits sémantiquement similaires si la mesure de sim_{sem} est > 0. La similarité entre le noeud n_1 et le noeud n_2 repose sur la formule suivante :

$$Sim_{sem}(n_1, n_2) = \frac{1}{Nb_c} * \frac{Sim(C_1, C_2)}{(Prof(C_1) + Prof(C_2))} \tag{5.1}$$

Où Nb_c est le nombre total de concepts des deux noeuds à comparer,
$Sim(C_1, C_2)$ la similarité les entre deux concepts d'intérêt des noeuds considérés
et $Prof(C_i)$ la profondeur du concept dans la dimension centres d'intérêt du modèle utilisateur concerné.

Pour calculer la valeur de la similarité entre deux concepts d'intérêt considérés, nous nous sommes basés sur la mesure de similarité proposée par (Jiang and Conrath, 1997) combinée à la mesure de calcul du contenu informatif -entropie- proposée par (Seco et al., 2004). Ces derniers considèrent que plus un concept a de descendants, moins il est informatif et de ce fait utilisent les hyponymes [10] des concepts pour calculer l'entropie de ceux-ci. Cette technique combine les techniques de calcul de similarité basées sur les arcs et celles basées sur les noeuds. Cette formule est définie par l'inverse de la distance sémantique :

$$D_{sem}(C_1, C_2) = IC(C_1) + IC(C_2) - 2 * IC(L_{super}(C_1, C_2)) \tag{5.2}$$

Où $L_{super}(C_1, C_2)$ est le concept commun le plus spécifique qui subsume (situé à un niveau hiérarchique plus élevé) les deux concepts C_1 et C_2. Le contenu informationnel d'un concept C, qui utilise WordNet, est calculé par la formule suivante :

$$IC_{wn}(C) = 1 - \frac{\log(hypo(C) + 1)}{\log(\max_{wn})} \tag{5.3}$$

10. Rapport d'inclusion entre des unités lexicales, considéré comme orienté du plus spécifique au plus général. Inverse de l'hyperonymie. www.larousse.fr, consulté en décembre 2010

Où $hypo(C)$ est le nombre d'hyponymes dont dispose le concept C et $\max_{wn}$ le nombre de concepts de la taxonomie.

Dans le cadre de notre travail, la taxonomie représente la dimension centres d'intérêt du modèle de l'utilisateur. De ce fait, $\max_{wn}$ représente le nombre de concepts contenus dans la dimension centres d'intérêt.
La formule de calcul de la distance sémantique entre deux noeuds proposée est donc :

$$Sim_{sem}(n_1, n_2) = \frac{1}{Nb_c * D_{sem}(C_1, C_2) * (Prof(C_1) + Prof(C_2))} \tag{5.4}$$

C_1 et C_2 étant les concepts considérés au moment de la recherche d'information.

5.2.2.2 La similarité spatiale

Les données spatiales utilisées sont de deux types : les données spatiales extraites des documents et celles fournies implicitement ou explicitement par l'utilisateur. Les informations spatiales explicitement fournies sont celles se trouvant dans les requêtes de l'utilisateur. Quant à celle implicites, elles sont déduites des navigations de l'utilisateur à travers les résultats de recherche.
Nous partons de l'hypothèse d'utilisation d'une distance réelle entre les différentes entités spatiales demandées à travers les requêtes utilisateurs. Cette hypothèse implique un calcul réel des distances : nous allons utiliser la distance du grand cercle, plus connue sous le nom distance orthodromique. Cette distance représente la plus petite distance entre deux points sur une sphère. La sphère représente la terre, dont la forme est plus proche d'une sphère. En l'appliquant sur les points spatiaux, il est possible d'utiliser les coordonnées représentées par la longitude et la latitude.
La formule, pour deux points $A\{lat_1, lon_1\}$ et $B\{lat_2, lon_2\}$, est donc :

$$d(A, B) = R \times \arccos(\cos(lat_1)\cos(lat_2)\cos(lon_1 - lon_2) + \sin(lat_1)\sin(lat_2)) \tag{5.5}$$

Afin de calculer la similarité spatiale entre deux noeuds du réseau, nous exploitons donc les données suivantes :

– La distance du grand cercle entre les deux concepts considérés lors du calcul de la similarité spatiale.
– Le nombre de concepts communs, Nb_{Comm}, entre les noeuds

La formule est la suivante :

$$Sim_{Spa}(n_1, n_2) = \frac{\min(d(n1, n_2))}{Nb_{Comm}} \tag{5.6}$$

Avec :
Nb_{Comm} le nombre de concepts communs.

5.2.2.3 Calcul de la similarité entre deux noeuds du réseau

La distance entre deux noeuds du réseau est une combinaison de la distance sémantique et de la distance spatiale. Nous proposons d'utiliser la moyenne de ces deux distances pondérées au nombre de concepts communs afin de définir la distance entre deux noeuds :

$$Sim(n_1, n_2) = \frac{1}{2} * (\theta * Sim_{spa}(n_1, n_2) + \sigma * Sim_{sem}(n_1, n_2)) \tag{5.7}$$

Avec θ le nombre de concepts spatiaux en commun
σ le nombre de concepts sémantiques en commun

5.3 Construction du réseau

La construction du réseau est effectuée en prenant en compte les modèles utilisateurs associés aux utilisateurs connus et aux stéréotypes regroupant les données des utilisateurs inconnus. Un utilisateur inconnu étant un utilisateur dont la dimension des données personnelles n'a pas été renseignée.

L'algorithme suivant illustre cette construction :

Algorithme 4 Algorithme de construction du réseau de modèles utilisateurs

Entrées : Q_{tx}
$u \in U$
M_u
Sortie : Réseau G construit ou mis à jour
DEBUT
**Etat initial : Le réseau ne contient aucun noeud*
si $(G_M = \emptyset)$ **alors**
$AjouterNoeud(M_u, G_M)$
CalculerPonderationNoeud$(M_u^i, \sigma(M_u^i))$
sinon
si $(u \notin G)$ **alors**
pour (chaque noeud V_j de G_M) **faire**
Calculer la similarité entre V_j et M_u^i
Construire l'arc pondéré entre V_j et M_u^i
fin pour
sinon
si (u est connu) **alors**
{Suivant les navigations courantes}
$MettreAJour(M_u^i, \sigma(M_u^i))$
$MettreAJour(M_u^i, \varphi(M_u^i))$
sinon
Rechercher le modèle utilisateur correspondant à un stereotype le plus correspondant
Affecter u
finsi
finsi
finsi
Retourner G
FIN

5.4 Construction de l'ensemble des résultats fournis à l'utilisateur

Nous avons présenté jusqu'à présent les deux principaux composants du système proposé, section 3.4, à savoir la construction du modèle de l'utilisateur et la construction du réseau de modèles utilisateurs. Afin d'être construits et de fournir un résultat personnalisé à l'utilisateur, ces deux composants interagissent avec un troisième qui est celui de la collecte des données utilisateurs et de reformulation de requêtes. En effet, et comme schématisé dans la figure 3.2, une fois que l'utilisateur soumet sa requête, le composant de

collecte des données utilisateurs et de reformulation de requêtes vérifie l'existence du modèle de l'utilisateur courant et procède à la reformulation de sa requête. Une fois cette étape effectuée, le processus de recherche personnalisée est alors activé. C'est dans ce contexte que nous présentons dans cette section les algorithmes de reformulation de la requête de l'utilisateur et de la recherche personnalisée du système.

5.4.1 Reformulation de la requête utilisateur

De manière générale, la requête de l'utilisateur peut être exprimée de diverses manières : à l'aide de mots clefs, en utilisant le langage naturel ou en sélectionnant ou en navigant dans un ensemble de catégories prédéfinies. Il a été prouvé que la reformulation de requêtes a des effets positifs en RI (Harman, 1992).
Dans notre système, la requête est exprimée en combinant les mots clefs à la navigation spatiale sur une carte. En partant de cette formulation par l'utilisateur et dans l'objectif d'avoir une requête répondant au mieux à la demande de ce dernier, nous procédons à son enrichissement en utilisant les dimensions d'intérêt et spatiale du modèle de l'utilisateur. En effet, la première dimension nous apporte les concepts construits au fur et à mesure des différentes recherches de l'utilisateur ainsi que l'enrichissement à travers l'utilisation des relations sémantiques de WordNet. Quant à la seconde dimension, elle nous permet de répondre, si besoin est, aux différentes demandes d'ordre spatial de l'utilisateur.
Les étapes du traitement de la requête sont, dans leur ordre d'écriture :

- Analyse sémantique simple lors de cette analyse, nous procédons à l'élimination des mots vides.
- Analyse sémantique utilisant WordNet : il s'agit ici d'exploiter le dictionnaire afin d'enrichir la requête. Il est important de noter ici l'existence de termes ambigus.
- L'étape suivante est donc une désambigüisation afin de ne garder que les termes les plus correspondants.

Nous avons choisi de ne pas passer par l'étape de lemmatisation[11] de la requête et d'utiliser le dictionnaire wordnet qui nous permet d'enrichir les termes soumis par l'utili-

11. La lemmatisation a pour objectif de transformer un terme de la forme fléchie ou conjuguée vers sa forme canonique.

sateur.

La reformulation de la requête considère la sémantique simple ainsi que la sémantique spatiale. A la seconde étape précédemment décrite, les termes considérés peuvent aussi bien être spatiaux, tels les noms de villes. Nous proposons l'algorithme de traitement de la requête suivant :

Algorithme 5 Algorithme de traitement de la requête utilisateur

Entrée : $Q_t = \{w_1, w_2, ..., w_n\}$, M_u
Sortie : Requête enrichie

Données exploitées : q_{tx} la requête textuelle de l'utilisateur
La dimension d'intérêt D_{in} du modèle de l'utilisateur M_u

DEBUT
pour chaque terme w_i de la requête **faire**
**Enrichissement en consultant M_u et la dimension D_{in}*
si (Enrichissement q_{tx} effectué) **alors**
**Ajout des concepts dont l'intérêt répond au besoin à la requête*
$q_{tx} = q_{tx} \cup c_i^u$
finsi
**Consultation de WordNet*
si ($hypernymes(w_i) \in \{country, state, land\}$) **alors**
**Consultation D_{sp}*
si CalculAccessibilité(synonyme(w_i),c_s^u) $\in [0, 1]$ **alors**
**Consultation des méronymes de w_i*
$Q_t = Q_t \cup synonyme(w_i) \cup meronyme(w_i)$
finsi
finsi
fin pour
FIN

5.4.2 Approche de recherche personnalisée

Nous détaillons l'approche de recherche personnalisée dans l'algorithme 6. Il s'agit de :

1. Traiter et reformuler la requête de l'utilisateur au début d'une session de recherche ainsi qu'à chaque soumission de requête.
2. Effectuer la recherche d'information en utilisant le contenu du modèle de l'utilisateur ainsi que celui du réseau de modèles des utilisateurs.
3. Au cours des navigations de l'utilisateur à travers les résultats de recherche et pour

une même session, il y a mise à jour et construction du modèle de l'utilisateur ainsi que du réseau de modèles utilisateurs. Ceci implique :

Dans cet algorithme, le système implique le scénario suivant pour chaque requête soumise par l'utilisateur : un utilisateur u soumet sa requête q_t à un instant donné t au système de recherche SyPRISS. Ce dernier retourne un ensemble résultats. Partant des documents cliqués par l'utilisateur, le système construit le modèle de l'utilisateur et le réseau de modèle d'utilisateurs selon les algorithmes 1 et 4.

Algorithme 6 Approche de recherche personnalisée

Entrées : $Q_t = \{w_1, w_2, ..., w_n\}$
$u \in U$
U l'ensemble des utilisateurs
DEBUT
pour chaque nouvelle session de recherche **faire**
  pour chaque nouvelle requête soumise **faire**
    Effectuer le traitement de Q_t en utilisant la dimension centres d'intérêt de M_u
    si ResultatsPertinentModele==true **alors**
      Extraire les résultats pertinents de M_u
    finsi
    si ExisteUtilisateurVoisin==true **alors**
      Extraire les résultats pertinents de $V(M_u)$
    finsi
    Effectuer une recherche sur les données du web
    Ordonner tous les résultats de recherche selon M_u
    pour chaque résultat considéré **faire**
      Enregistrer les données de navigation
      Calculer les intérêts implicites
    fin pour
  fin pour
fin pour
FIN

L'extraction des résultats pertinents est effectuée en fonction des pondérations d'intérêt. En effet, et comme présenté à travers le modèle de l'utilisateur, à chaque concept d'intérêt de la dimension centres d'intérêt, respectivement de la dimension spatiale, est associée une pondération de l'intérêt. Pour chaque concept d'intérêt, les adresses des documents consultés sont sauvegardées dans la dimension textuelle. De ce fait extraire les résultats pertinent du modèle de l'utilisateur ou du modèle avoisinant revient à sélectionner de la

dimension textuelle les adresses des documents ayant un intérêt pour l'utilisateur.

Dans le but d'illustrer la construction des résultats personnalisés fournis, nous présentons le scénario de recherche suivant effectué par l'utilisateur u à l'instant t.

A l'instant t, nous avons la collection de documents suivante :

$D = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\}$

A chaque document est associé un ensemble de termes tels que

$W_d = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8\}$.

La fréquence à laquelle apparait chaque terme dans chaque document est comme suit :

$f_{d1} = \{4w_1, 3w_5, 6w_6, 8w_8\}$,

$f_{d2} = \{15w_1, 35w_3, 20w_5, 7w_8\}$,

$f_{d3} = \{5w_2, 1w_5, 24w_6\}$,

$f_{d4} = \{2w_1, 4w_2, 3w_5, 26w_8\}$,

$f_{d5} = \{7w_2, 53w_7, 81w_8\}$,

$f_{d6} = \{17w_3, 3w_5, 1w_6, 19w_8\}$,

$f_{d7} = \{27w_4, 5w_5, 8w_7, 1w_8\}$,

$f_{d8} = \{9w_1, 26w_2, 11w_4, 31w_6, 16w_7\}$,

$f_{d9} = \{5w_2, 3w_3, 13w_4\}$,

$f_{d10} = \{3w_3, 21w_5, 8w_7, 18w_{10}\}$.

Dans cet exemple, nous supposons le contenu des différentes dimensions du modèle de l'utilisateur tel que :

TABLE 5.1 – Contenu du modèle de l'utilisateur u

D_{tx}	$W_i = \{(w_1, 2), (w_3, 4), (w_5, 4), (w_8, 2))\}$ $D_j = \{(w_1, 8, 1), (w_3, 4, 2), (w_5, 140, 90), (w_8, 200, 15))\}$ $T = 460ms$ $H = \{(w_1, d_1, d_2, d_8, q_1, 1), (w_3, d_2, d_9, q_1, 1), (w_5, d_1, d_6, q_2, 1),$ $(w_4, d_2, d_6, d_7, q_2, 1), (w_7, d_5, q_1, 2), (w_4, d_8, q_2, 2)\}$
D_{sp}	$\{w_1, 0.8\}$
D_{in}	$\{c_1, (w_1, w_5), 0.63\}$ $\{c_2, (w_8), 0.83\}$
P_u	login : user password : user

Dans le but d'illustrer l'algorithme, nous présentons le scénario de recherche suivant effectué par l'utilisateur u à l'instant t_1 :

- la requête soumise est $q_{t1} = \{w_3, w_4, w_1\}$
- les résultats affichés dans l'ordre sont : $d_1, d_2, d_7, d_8, d_{10}, d_9, d_4$
- les documents consultés ainsi que leur durée de consultation sont :

  $\{(d_2, 156), (d_7, 80), (d_8, 50), (d_9, 35), (d_4, 3)\}$

Nous présentons dans la figure 5.2 un exemple d'exécution de l'algorithme. La requête de départ est "hôtel+piscine".

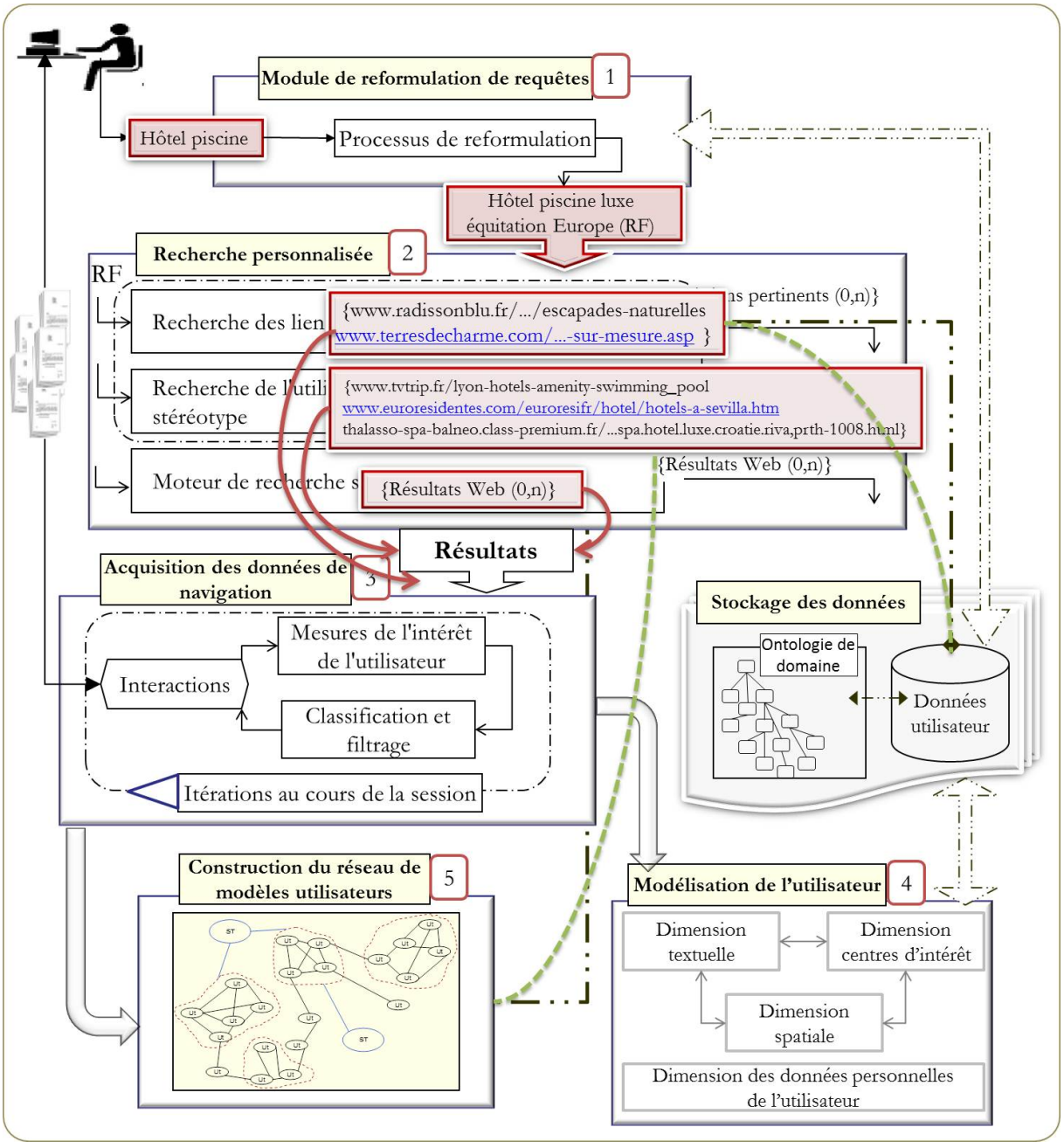


FIGURE 5.2 – Exemple d'exécution de l'algorithme de recherche personnalisée

En partant de la requête "hôtel+piscine", le processus de reformulation, après consultation des concepts du modèle de l'utilisateur courant, soumet la requête reformulée "hôtel+piscine+luxe+équitation+Europe". Le processus de recherche d'information person-

nalisée est alors activé. Dans l'exemple considéré par cette figure, la recherche dans les résultats pertinents du modèle de l'utilisateur courant a donné un lien. Quant à la recherche dans le voisinage du modèle de l'utilisateur courant, elle a permis d'avoir trois liens de plus. Suite à ses deux recherches à travers les modèles du système, la recherche sur le Web est alors effectuée. Les résultats sont ensuite soumis à l'utilisateur. L'exemple de la figure 5.2 a pour objectif de visualiser les résultats pouvant être extraits des modèles des utilisateurs.

5.5 Retour d'expérience de l'utilisateur sur les résultats de la personnalisation

La précision et l'exactitude du modèle l'utilisateur sont déterminantes pour obtenir une bonne personnalisation. En effet, le modèle créé et alimenté tout au long de la navigation de l'utilisateur permet au système de proposer à ce dernier des résultats de recherche correspondants à ses attentes. En s'appuyant sur les mesures d'intérêt pour les concepts, le système a la possibilité de proposer les meilleurs candidats à l'utilisateur. A l'intérieur d'un même concept, les mesures d'intérêt des attributs permettent aussi de déterminer les caractéristiques préférées et susceptibles d'être les plus visitées.

Nous nous proposons, dans ce contexte, d'avoir un retour d'expérience de l'utilisateur en introduisant un ensemble de critères d'évaluation explicites qui concernent principalement les données spatiales. En effet, notre hypothèse est que les informations spatiales telles que la localisation de l'utilisateur ainsi que la localisation spatiale de ses centres d'intérêt exprimés ont un fort impact sur la qualité des résultats retournés.

Dans ce qui suit, nous utilisons la terminologie suivante :

- localisation de l'utilisateur : n'étant pas dans le cadre d'une application mobile, la localisation de l'utilisateur au moment de l'émission de sa requête sera déduite de son adresse IP ou bien de sa requête spatiale q_{sp}. Nous notons cette localisation par L_u.
- localisation de la zone d'intérêt de l'utilisateur : cette localisation correspondant à la zone-cible de recherche de l'utilisateur, elle peut être contenue dans la requête textuelle de l'utilisateur q_{tx} ou dans sa requête spatialeq_{sp}. Cette information sera

notée Z_i.

- objet(s) d'intérêt : type de données recherché (par exemple, des hôtels, des restaurants et/ou des informations touristiques). Cette information est extraite des requêtes de l'utilisateur.
- point et chemin de rayonnement : le déplacement de l'utilisateur de sa localisation de recherche vers sa zone d'intérêt implique d'avoir un point de chute de ce dernier, tel un aéroport ou une gare routière. Nous avons appelé ce point le point de rayonnement. La distance séparant le point d'arrivée à l'objet d'intérêt du point de rayonnement est appelée chemin de rayonnement. L'objectif est de tendre à afficher des résultats de recherche optimisant le chemin de rayonnement en le diminuant.

Partant de l'hypothèse forte que l'utilisateur qui se trouve dans une localisation L_u et effectue une recherche ciblant une zone Z_i pourrait avoir l'intention de s'y déplacer, l'objectif de ce retour d'expérience est d'estimer le degré de correspondance des résultats affichés avec le besoin réel de l'utilisateur. Les mesures sont : (1) évaluation du chemin d'arrivée à la zone d'intérêt de l'utilisateur et (2) évaluation de l'arrivée par rapport à un chemin de rayonnement.

5.5.1 Chemin d'arrivée à la zone d'intérêt

L'évaluation du chemin d'arrivée à la zone d'intérêt sert de critère de qualité. En effet, le chemin pris, tel que nous le considérons, a un impact sur les points d'arrivée dans la zone d'intérêt et par conséquent influe sur l'emplacement des objets d'intérêt à afficher à l'utilisateur. Dans cette optique, nous effectuons une évaluation qualitative explicite basée sur une pondération de critères tels que le coût ou le type du chemin. L'utilisateur pouvant dans un deuxième temps ajouter ses propres choix d'évaluation. Les critères minimums que nous avons pris sont le type de chemin (voie ferrée, avion, à pied,..), son coût et la distance parcourue. Les pondérations prises sont des valeurs comprises entre 0 et 1, valeurs qui seront données sous forme de note par l'utilisateur. L'équation du calcul du chemin emprunté est la suivante :

$$EvalPath = \frac{\alpha_{type} * \beta_{count} * \lambda_{distance}}{n} \tag{5.8}$$

Avec α, β et λ dans l'intervalle $]0, 1]$ et n le nombre de critères choisis.

5.5.2 Evaluation de l'arrivée par rapport à un chemin de rayonnement

Le déplacement de l'utilisateur de sa localisation de recherche vers sa zone d'intérêt peut nous fournir deux types d'information : (1) le point d'arrivée, donnée pouvant être déduite du chemin pris et (2) un emplacement repère qui constitue le point de chute de l'utilisateur, nous l'avons appelé point de rayonnement. Le point de rayonnement peut être la station métro la plus proche de l'objet d'intérêt ou l'endroit de résidence de l'utilisateur. La distance séparant le point d'arrivée du point de rayonnement est appelée chemin de rayonnement. Cette distance est égale à 1 dans le cas où le point d'arrivée et le point de rayonnement sont confondus. Les hypothèses de départ sont donc les suivantes :

- Existence d'un point d'arrivée dans la zone Z_i
- Si l'utilisateur s'est déplacé vers la zone d'intérêt Z_i, nous avons alors les paramètres suivants : un point de départ et un chemin de rayonnement qui est la distance entre le point d'arrivée et le point de rayonnement : $d(pointd'arrive, pointderayonnement)$.

L'évaluation s'effectue en fonction du chemin de rayonnement et de la sémantique de l'objet de recherche :

$$Eval(P_Arr, C_r) = \alpha_{(d_{P_{Arr}, P_r})} * \frac{1}{n} \sum \beta_{i(d_{P_r, O_{rech}})} \tag{5.9}$$

Avec :

α et β dans l'intervalle $]0, 1]$

P_Arr le point d'arrivée,

C_r le chemin de rayonnement,

$\alpha_{d_{P_{Arr}, P_r}}$: évaluation donnée par l'utilisateur sur la distance séparant le point d'arrivée du point de rayonnement,

n : nombre d'objets de recherche évalués par l'utilisateur,

$\beta_{i(d_{P_r, O_{rech}})}$: évaluation donnée par l'utilisateur sur la distance séparant le point de rayonnement et l'objet de recherche considéré.

5.6 Conclusion

Nous avons présenté dans ce chapitre le réseau de modèles utilisateurs ainsi que le processus global de la personnalisation de la recherche proposé. La construction d'un réseau de modèles utilisateurs a pour objectif de proposer à l'utilisateur courant des résultats de recherche personnalisés en exploitant les intérêts contenus dans les modèles du voisinage. Nous avons aussi proposé un retour d'expérience sur les résultats fournis à l'utilisateur. Nous procédons maintenant à la présentation de la phase expérimentale de notre thèse. Cette phase fait l'objet du chapitre suivant dans lequel nous détaillons les principes de l'évaluation et les résultats expérimentaux.

Chapitre 6

Expérimentation et évaluation de la proposition

Introduction

Afin d'expérimenter l'ensemble de la proposition détaillée dans les chapitres 4, et 5 nous avons développé un prototype que nous décrivons dans le présent chapitre. Les apports de notre proposition de personnalisation de la recherche sur le web par rapport à une recherche classique (à travers un moteur de recherche) seront démontrés expérimentalement. Et étant dans un contexte de recherche d'information, les mesures d'évaluation adaptées sont principalement basées sur la précision et le rappel. Le principe étant de comparer les résultats obtenus en utilisant notre prototype dans son environnement expérimental aux réponses idéales attendues par l'utilisateur et celles fournies par un moteur de recherche pris comme référence. Afin d'élargir le spectre de l'expérimentation et de présenter des résultats réels, l'expérimentation a été effectuée en laboratoire et avec la collaboration d'une entreprise touristique. Le prototype développé s'insère dans le cadre d'un projet de coopération Tuniso-Française DGRST-INRIA STIC et dont l'objectif principal est la modélisation d'un environnement logiciel pour la personnalisation de la recherche d'information géographique sur le Web tout en prenant compte des préférences de l'utilisateur et de son contexte spatial.

Nous décrivons, dans la section 6.1, le prototype expérimental développé. La section 6.2 présente les objectifs de l'évaluation ainsi que le déroulement et les données de l'expérimentation. Les sections suivantes (6.4, 6.5 et 6.6) présentent les différentes évaluations

effectuées. Dans la section 6.7, nous présentons la synthèse des résultats expérimentaux obtenus.

6.1 Architecture de l'environnement logiciel développé

L'environnement développé et supportant le système proposé, décrit par la figure 6.1, repose sur l'utilisation de l'api gratuit Google Web Search [12] . Cet api utilise comme base de documents les liens indexés par le moteur de recherche Google [13].

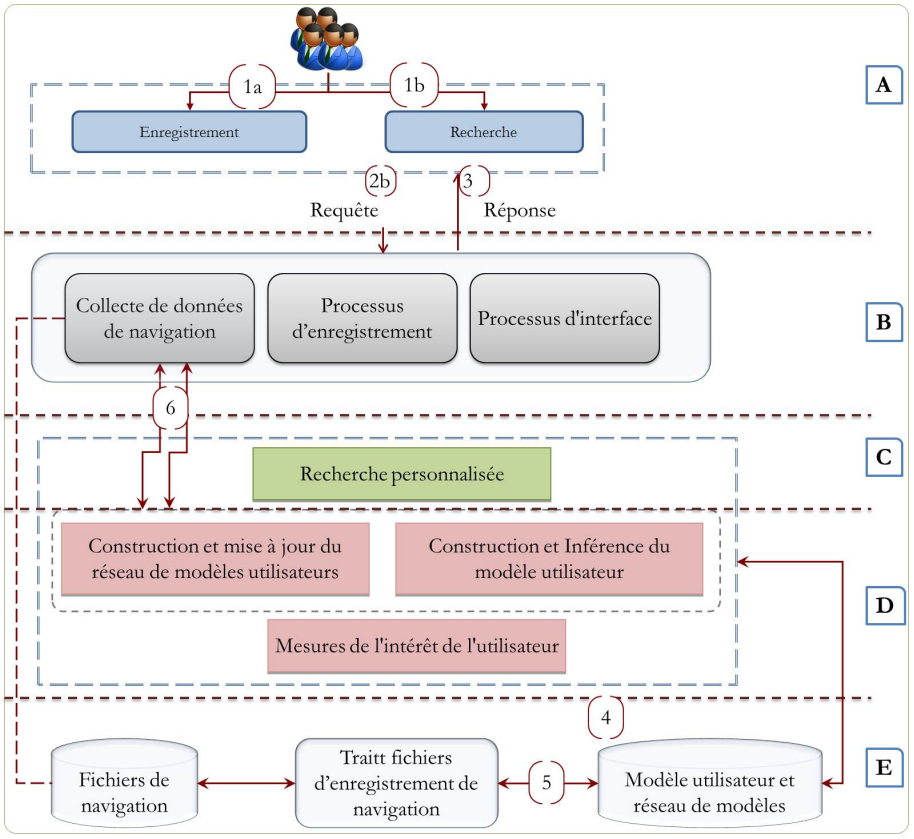


FIGURE 6.1 – Environnement développé

Les utilisateurs accèdent au système à travers une interface de recherche dans laquelle nous leur proposons de s'enregistrer (1.a) ou bien d'effectuer directement leur recherche (1.b).

12. Officiellement remplacé par le JSON/Atom Custom Search API, le 01 novembre 2010. Url : http ://code.google.com/intl/fr/apis/customsearch/v1/overview.html
13. http ://www.google.com/

L'environnement que nous proposons est constitué de cinq niveaux : utilisateur, applicatif, modélisation de l'utilisateur (NMU), recherche d'information (NRI) et stockage (Hadjouni et al., 2008). Nous présentons dans ce qui suit chacun de ces niveaux :

- Le niveau utilisateur, niveau A sur la figure, comporte trois possibilités pour l'utilisateur :
  - Enregistrement dans le système,
  - Accès à la recherche personnalisée,
  - Accès à ses informations du modèle de l'utilisateur.
- Le niveau applicatif permet de déterminer l'utilisateur accédant à l'environnement. Ce dernier considère un utilisateur comme identifié (noté UCn) s'il possède un modèle propre, i.e. le modèle de l'utilisateur relatif est construit.
- Le niveau de la modélisation utilisateur a pour charge :
  - De créer un profil utilisateur.
  - De construire et d'alimenter le modèle de l'utilisateur à travers les critères de recherche émis. Sont aussi pris en considération : la navigation lors de la session courante, les navigations antécédentes, les centres d'intérêt explicitement fournis par le biais de la sélection d'objets représentatifs des intérêts et les choix cartographiques effectués.
  - D'activer le modèle un utilisateur et de charger le contenu nécessaire à la recherche au niveau applicatif. Ce niveau présente la particularité d'être construit autour d'un réseau de modèles utilisateurs qui interagissent entre eux afin d'avoir une meilleure pertinence du résultat de la recherche.
- Le niveau recherche d'information, une fois qu'il a la requête enrichie de l'utilisateur, effectue l'extraction des liens correspondants. Il a aussi la charge de transmettre ses résultats triés suivant leur adéquation au modèle de l'utilisateur.
- Le niveau stockage de données est constitué de sources d'informations hétérogènes. Il contient les données du modèle de l'utilisateur et du réseau d'utilisateurs ainsi que les fichiers logs des navigations.

L'implémentation de l'environnement logiciel est générique, elle est indépendante à tout domaine d'application. En effet, la recherche personnalisée proposée n'est dépendante

d'aucun domaine particulier. L'utilisation d'une ontologie de domaine pour alimenter la dimension centres d'intérêt du modèle de l'utilisateur est la seule liaison avec un quelconque domaine.

6.2 Présentation de l'évaluation et des objectifs expérimentaux

Afin de procéder à l'évaluation, nous avons exploité les éléments suivants :

- Une collection donnée de recherche : dans notre cas, cette collection est le Web accessible à travers l'API de recherche de Google.
- Un ensemble de thèmes de recherche autour desquels vont se construire les requêtes : nous avons choisi quatre thèmes présentés dans la section 6.2.2.
- Les jugements de pertinence des documents : ces jugements ont été demandés aux utilisateurs ayant expérimenté le prototype développé. Ils ont pour objectif d'avoir un nombre de documents pertinents par thème.
- Les utilisateurs : afin de mener à bien notre expérimentation, nous avons demandé à des utilisateurs réels de participer. Ces derniers sont présentés dans le paragraphe 6.2.2.1 et dans le Tableau 6.2.
- Le protocole d'évaluation : ce protocole est présenté et utilisé tout au long des phases d'expérimentation.
- Les mesures d'expérimentation : ces mesures sont principalement le rappels et la précision.

Le processus d'expérimentation, comme le montre la figure 6.2, commence par une phase de manipulation du prototype par les utilisateurs. Ces derniers disposent d'un scénario pour lequel ils sont libres de soumettre les termes de requêtes désirés. Les résultats affichés sont ensuite stockés et indexés. Nous avons demandé aux utilisateurs d'indiquer les réponses pertinentes lesquelles vont être utilisées lors de la troisième phase. Cette phase est celle de l'évaluation du système proposé.

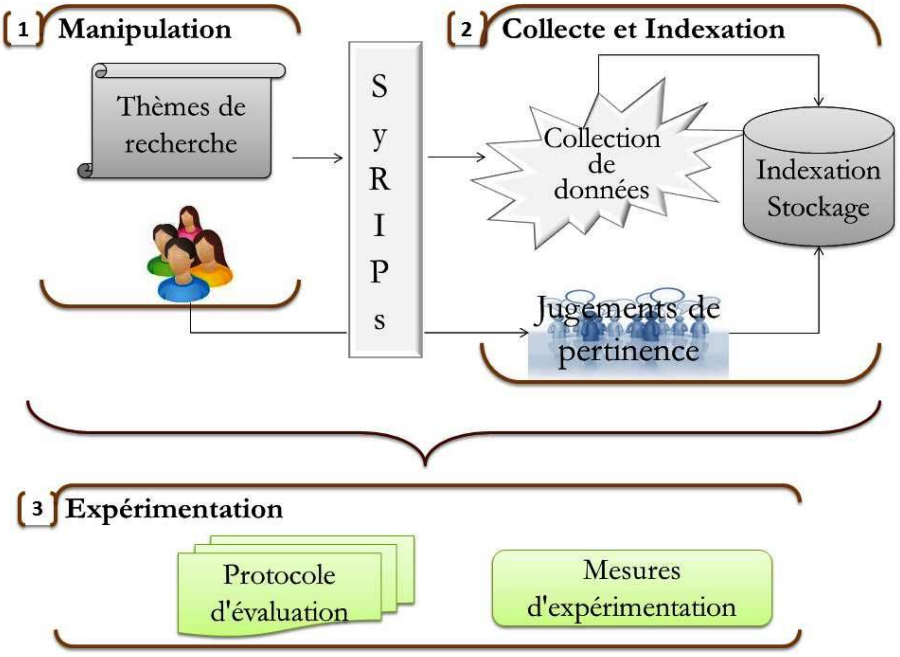


FIGURE 6.2 – Processus expérimental

6.2.1 Objectifs expérimentaux

L'évaluation expérimentale de notre proposition a pour objectifs :

D'évaluer l'efficacité de la recherche : pour effectuer cette étape de l'évaluation, nous prenons comme référence les résultats fournis par le système utilisant uniquement un api nous permettant d'avoir les résultats sans personnalisation. Nous avons utilisé pour cela l'api Google Search. Nous comparons ensuite l'amélioration qu'apporte notre proposition. Cette évaluation fait objet de la section 6.4.

D'évaluer l'efficacité de la recherche avec l'intégration de la dimension spatiale du modèle de l'utilisateur : en partant de l'évaluation précédente, nous ajoutons à cette étape le calcul de la précision des résultats fournis avec la prise en compte des informations spatiales contenues dans le modèle de l'utilisateur. Cette évaluation fait objet de la section 6.5.

D'évaluer la recherche en exploitant le réseau de modèles utilisateurs. Dans cette évaluation, nous avons pour objectif de vérifier deux points : le premier est la construction du réseau d'utilisateurs et le second est l'efficacité des résultats fournis aux utilisateurs. Cette évaluation fait objet de la section 6.6.
Les composantes de notre évaluation, qui seront présentées au fur et à mesure de leur utilisation, sont résumées dans le tableau suivant :

TABLE 6.1 – Composantes de l'évaluation

Déroulement de l'expérimentation	
Documents	Collection de pages Web accessibles à travers l'api du moteur de recherche Google.
Utilisateurs	Utilisateurs réels dans l'environnement académique Utilisateurs du site touristique d'une entreprise.
Requêtes	Requêtes récoltées à travers la manipulation de notre système Requêtes obtenues en collaboration avec une entreprise touristique.
Jugements de pertinence	Jugements obtenus par les utilisateurs afin de pouvoir expérimenter le système.
Evaluation	
Protocole	Evaluation décomposée en quatre étapes ayant chacune son protocole.
Mesures	Précision à n, précision moyenne, taux d'amélioration.

6.2.2 Présentation des données de l'expérimentation

6.2.2.1 Données d'expérimentation en laboratoire

Nous utilisons pour effectuer notre évaluation un corpus de pages Web collecté à partir de réponses de notre prototype à différentes requêtes. Notre expérimentation est effectuée autour du domaine du tourisme. Nous avons défini trois thèmes de recherche qui simulent chacun une situation dans le même thème. A ces sujets, nous en avons ajouté un quatrième considéré comme "chaotique" car n'appartenant à aucun domaine en particulier. Pour ce cas, nous avons demandé aux utilisateurs d'effectuer ce que bon leur semble quant au choix des termes. Ces quatre thèmes sont donc :

Thème(1) Une personne, de type [14] quelconque, recherche un hôtel à proximité de la plage et en Afrique du Nord (Tunisie et Maroc de préférence). Cette personne aime le luxe et les visites de sites historiques où les musées de tout genre.

Thème(2) Un étudiant en stage en Europe, dans une des capitales, recherche un hôtel pas trop cher car il aura la visite d'un parent. L'hôtel devra être proche de la cité universitaire dans laquelle il vit [15].

Thème(3) Une personne recherchant un hôtel sur la méditerranée. Pour ce thème, nous considérerons deux catégories de personnes : une venant de France et la seconde de Tunisie et chacune devra quitter son pays. Cette personne voudrait que l'hôtel soit de luxe et surtout pas très loin de l'aéroport.

Thème(4) Scénario chaotique où les testeurs mettront ce qu'ils veulent comme requêtes.

Les volontaires auxquels nous assignons chaque thème simulé sont libres du choix des termes de leurs requêtes textuelles ainsi que spatiales. Un utilisateur assigné à un thème ne peut pas appartenir à une autre catégorie. Nous effectuons nos tests sur une durée de 20 jours dans l'objectif d'avoir plusieurs sessions de recherche assez éloignées dans le temps. Les utilisateurs impliqués dans cette étape d'évaluation sont des utilisateurs volontaires issus du milieu universitaire. Ces personnes sont indifféremment des étudiants, des doctorants ou appartenant au corps enseignant. Nous distinguons aussi les utilisateurs enregistrés et ceux qui effectuent leurs recherches anonymement. Les cas de tests sont présentés dans le tableau 6.2.

TABLE 6.2 – Organisation des groupes d'utilisateurs testeurs du prototype

Nombre de testeurs	**Connus**	**Non connus**	**Thème assigné**
55	50	5	Thème 1
58	45	13	Thème 2
53	50	3	Thème 3
54	43	11	Thème 4

14. Par type, nous entendons étudiant, retraité ou autre.
15. Nous fournissons aux utilisateurs des noms de cités universitaires des villes choisies dans leurs requêtes.

Les hypothèses de génération des requêtes de notre base de tests sont donc les suivantes :

- Un utilisateur connu assigné à un thème, ne peut pas en changer.
- De ce fait, les vecteurs de termes d'intérêt créés sont respectivement associés aux scénarios assignés.

Les caractéristiques des données expérimentales sont présentées dans le tableau 6.3.

TABLE 6.3 – Caractéristiques des données d'expérimentation

Nombre de domaines (thèmes)	4
Nombre de requêtes	589
Nombre de documents	623514
Nombre de termes distincts des requêtes	156
Nombre de termes distincts des intérêts	224

Lors de l'expérimentation, les utilisateurs ont jugé 15 documents pertinents sur l'ensemble des documents retournés. Les requêtes utilisées pour l'expérimentation ont été étendues suivant l'algorithme proposé dans la section 5.4.1.

6.2.2.2 Données d'expérimentation avec des données du site touristique Addictrip

En parallèle aux thèmes précédemment choisis et à travers la collaboration avec une entreprise de tourisme, nous avons entamé une série de tests avec leurs données. En effet, cette entreprise propose un site de planification de voyages [16] et utilise ses propres données touristiques. Ce site se base sur les votes des utilisateurs pour personnaliser les résultats proposés aux internautes enregistrés. Les données utilisées sont celles des utilisateurs de ce site. Nous proposons d'effectuer une comparaison entre leurs résultats (basés sur les votes) et ceux proposés par notre système. Les thèmes sont les suivants :

Thème(1) Demande de réservation d'hôtel, suite à laquelle l'utilisateur saisi la ville choisie. Une fois la liste des résultats affichés, l'internaute peut spécifier, en plus des dates de réservation :

- Son style : indifférent, couleur locale, classique ou design.

16. http ://www.addictrip.com/fr/home

- Sa situation : indifférent, solo, couple dynamique, couple calme, avec des amis, avec des enfants ou business.
- Son budget : indifférent, économique, confort ou luxe.

Thème(2) Demande de découverte d'une ville. L'utilisateur choisi ensuite la ville à visiter. Les mêmes critères de choix présentés dans le thème 1 sont présentés.

Thème(3) Recherche d'inspiration. Dans ce thème, les villes proposées sont placées suivant les votes des internautes.

Pour les trois thèmes, les internautes peuvent aussi choisir de rechercher des restaurants, des sites touristiques, des bars et clubs ou bien rien que des endroits pour le shopping. Et dans ces cas, le même principe de personnalisation est offert : système de votes des utilisateurs.

Les utilisateurs du site de l'entreprise sont au nombre de 90. L'expérimentation est datée du mois de novembre 2009. Les caractéristiques des données expérimentales sont présentées dans le tableau 6.3.

TABLE 6.4 – Caractéristiques des données d'expérimentation du site touristique

Nombre d'utilisateurs	90
Nombre de domaines (thèmes)	3
Nombre de requêtes	228
Nombre de documents	125639
Nombre de termes distincts des requêtes	40
Nombre de termes distincts des intérêts	53

6.2.3 Indexation de la base de recherche

L'indexation est effectuée de la manière suivante :

Pour chaque recherche effectuée par les utilisateurs expérimentaux, les données sauvegardées sont :

- La requête qi courante. Cette requête est reliée à la session de recherche courante.
- La liste de résultats fournis Ri=r1, r2, ..., rj.
- Les documents sélectionnés par l'utilisateur. A chaque document est assignée la mesure d'intérêt implicite calculée en cours de navigation de la session courante.

En parallèle à cela, nous stockons pour chaque requête utilisateur les résultats fournis par le moteur de recherche Google. Ces données nous servirons comme base d'index.
Nous avons utilisé Lemur Toolkit [17] afin de procéder à cette indexation. LEMUR est paramétré de façon à se comporter comme le système OKAPI (Robertson et al., 1998) : la partie textuelle de chaque document est représentée par plusieurs mots-clefs extraits automatiquement du corpus. Pour chaque document, l'importance accordée à chaque mot-clef est calculée à partir de la formule de pondération BM25 avec les paramètres par défaut.

6.3 Paramétrages préliminaires à l'évaluation

Comme présenté dans la section 4.2.2, la dimension centres d'intérêt du modèle de l'utilisateur représente, sous forme d'ontologie, les centres d'intérêt de l'utilisateur. Ces derniers ont un rôle principal dans la qualité des résultats offerts. De ce fait et afin de procéder à l'évaluation de la personnalisation à travers l'implication des données de cette dimension, nous allons évaluer le nombre de concepts optimal à considérer. L'objectif étant de quantifier le degré de correspondance entre les concepts d'intérêt construits implicitement et ceux annotés expérimentalement par les utilisateurs.
Nous procédons ensuite à la détermination du meilleur paramètre de réordonnancement à utiliser pour le tri des résultats affichés, formule 4.8 p. 80.

6.3.1 Etapes expérimentales

Afin de mener à bien cette expérimentation, nous avons besoin des requêtes des utilisateurs, de leurs intérêts implicites et de leurs jugements expérimentaux explicites. Les points d'entrée de cette expérimentation sont donc :

- Les requêtes des utilisateurs : Comme présentées dans la section 6.2.2, ces requêtes sont associées à des thèmes expérimentaux où chaque utilisateur ne peut appartenir qu'à un seul thème.
- Le calcul d'intérêt implicite issu des navigations des utilisateurs à travers les résultats de recherche. Ce calcul est ensuite associé aux concepts d'intérêt de la dimension centres d'intérêt.

17. http ://www.lemurproject.org/

– Les jugements explicites fournis par les utilisateurs : Afin de pouvoir mesurer la qualité des résultats et des modèles utilisateurs, nous avons demandé aux utilisateurs d'annoter les documents qu'ils considèrent pertinents et correspondant à leurs besoins.

Nous suivons la démarche qui suit :

– Pour chaque thème d'expérimentation et pour chaque utilisateur, le modèle associé est créé et la dimension de termes d'intérêt est construite.
– En considérant les concepts d'intérêt créés, nous procédons au calcul du nombre de concepts optimal utilisés dans le calcul du score personnalisé du document.
– Pour un modèle donné, nous mesurons la précision. Ensuite, nous mesurons la précision moyenne de tous les modèles utilisateurs de notre base de tests.

Nous présentons dans le tableau 6.5 un exemple de contenu de la dimension centres d'intérêt pour le second thème de nos jeux de données.

TABLE 6.5 – Exemple de dimension centres d'intérêt d'un modèle de l'utilisateur appartenant au second thème des jeux de données

Requêtes	Hotel/ cheap/ Roma Hotel/ near university / Roma Hotel/ low cost / student
Concepts d'intérêt du modèle de l'utilisateur	Student/Hotel/ Student/Hotel/Travel/Cheap/ Student/University/stage/accommodation/cheap/breakfast

Les métriques utilisées pour cette première étape sont la précision aux n premiers concepts d'intérêt pour un modèle de l'utilisateur et la précision moyenne pour tous les modèles utilisateurs.

6.3.2 Résultats expérimentaux du paramétrage

L'objectif de cette expérimentation est de calculer le nombre le plus adéquat de concepts d'intérêt allant être utilisé lors d'une recherche courante de l'utilisateur. Pour cela, nous procédons au calcul du score personnalisé des documents retournés à l'utilisateur. Comme

présenté dans la section 4.2.2, p.74 du chapitre 4, ce score est aussi exploité pour le réordonnancement des résultats retournés à l'utilisateur. Nous commençons donc par faire varier le paramètre de réordonnancement γ avec pour nombre de documents pertinents 15. Pour chaque valeur de $\gamma \in [0,1]$, nous avons calculé la précision P@5, P@10, P@20 et la moyenne MAP. Pour $\gamma = 0$, nous avons une recherche uniquement basée sur le score personnalisé du document. Quand $\gamma = 1$, il s'agit de la recherche classique.
Des résultats de la variation du paramètre de réordonnancement, nous constatons que, la meilleure précision est donnée pour $\gamma = 0,3$, et cela aussi bien pour P@5, P@10, P@20 que pour la moyenne des précisions. Ces résultats sont montrés par la figure 6.3 qui nous permet aussi de dire que pour $\gamma = 0$, nous avons une assez faible précision.

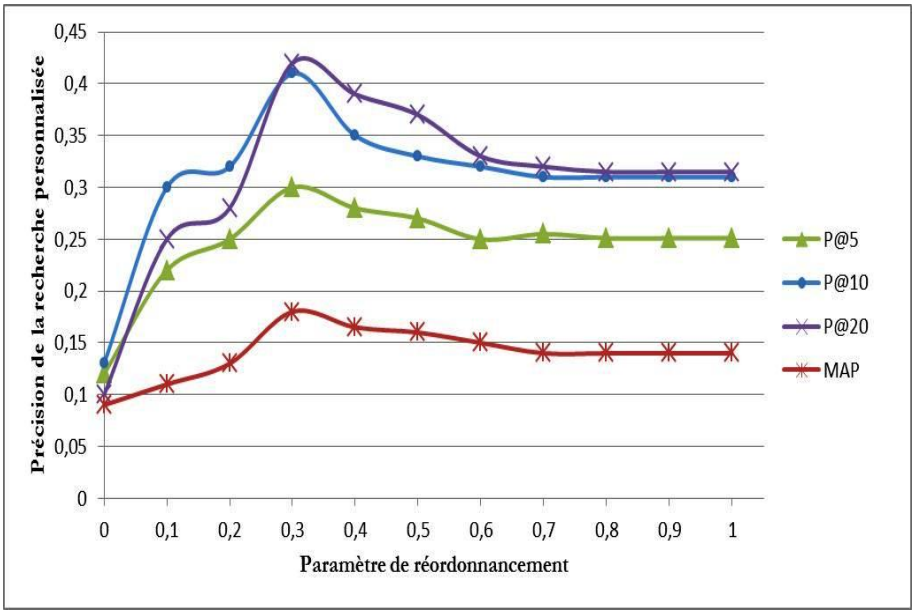


FIGURE 6.3 – Variation de la performance de la recherche en fonction du paramètre de réordonnancement des résultats de recherche pour un modèle de l'utilisateur donné

Nous procédons maintenant au calcul du nombre de concepts d'intérêt optimal à considérer pour une meilleure personnalisation, figure 6.4. Nous avons calculé, pour chaque thème expérimental, la précision de la recherche personnalisée en faisant varier le nombre de concepts utilisé. Nous avons pris comme intervalle $3, 5, 7, 10, 20$, avec toujours le même nombre de documents pertinents 15.

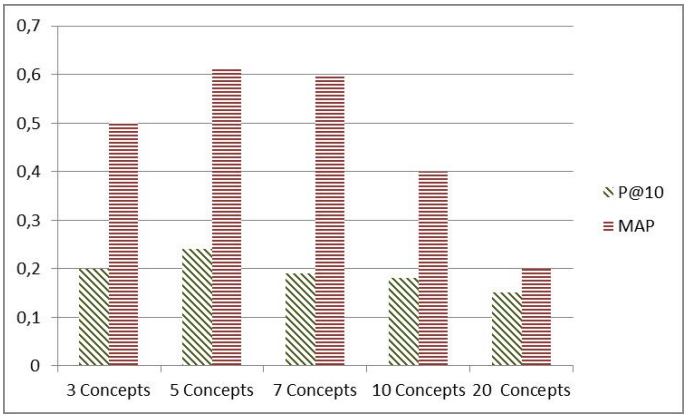


FIGURE 6.4 – Variation de la recherche personnalisée à P10 en fonction des paramètres de construction de la dimension centres d'intérêt

Nous constatons que :

- La meilleure précision à P@10 est obtenue en utilisant 5 concepts d'intérêt dans le calcul du score personnalisé.
- La meilleure valeur de la MAP est obtenue en utilisant 5 ou 7 concepts dans le calcul du score personnalisé.

Dans la suite de l'expérimentation, nous exploitons les meilleures valeurs identifiées dans l'expérimentation précédente :

1. le nombre de concepts utilisé dans le calcul du score personnalisé $h = 5$
2. le paramètre de réordonnancement utilisé pour calculer le score d'un document est $\gamma = 0.3$

6.4 Evaluation de l'efficacité de la recherche

L'objectif de cette section est de proposer une évaluation de l'efficacité des résultats de recherche personnalisés de notre système. Le protocole d'évaluation est basé sur la validation croisée permettant de créer le modèle de l'utilisateur sur la base d'un ensemble de requêtes d'apprentissage et est inspiré des travaux de (Daoud et al., 2008). Dans cette expérimentation, nous exploitons les meilleures valeurs identifiées dans la section précédente.

6.4.1 Protocole d'évaluation de l'efficacité de la recherche

Pour effectuer cette évaluation nous avons fait varier l'ensemble des requêtes des utilisateurs ayant servi de base à la construction des différents modèles des utilisateurs du système. La stratégie utilisée consiste en les étapes suivantes :

- Partir des requêtes des utilisateurs construites lors de la première phase d'évaluation et sélectionner un modèle d'utilisateur à expérimenter. Pour ce modèle, nous subdivisons l'ensemble des requêtes récoltées en un sous-ensemble d'apprentissage de n-1 requêtes et en un sous-ensemble de test contenant la nième requête à tester,
- A partir des requêtes ainsi récoltées, le modèle de l'utilisateur est construit en exploitant un sous-ensemble de documents pertinents par requête. Ce sous ensemble est basé sur le calcul implicite de l'intérêt de l'utilisateur porté aux documents. Comme effectué précédemment avec les concepts d'intérêt, nous avons aussi demandé aux utilisateurs de cocher les réponses jugées pertinentes.
- Et finalement, passer à la phase d'évaluation qui consiste à l'exécution des requêtes et, en utilisant le modèle de l'utilisateur, recalculer l'ordre des résultats. Cette étape comprend aussi une phase d'enrichissement de requêtes en utilisant le modèle de l'utilisateur.

Nous illustrons dans la figure 6.5 les étapes du protocole d'évaluation.

Les métriques utilisées dans cette seconde partie de l'évaluation sont la précision et la précision moyenne.
Pour effectuer la phase expérimentale, nous procédons aussi à l'exécution de recherches basées sur les mêmes requêtes à travers le moteur de recherche Google. Nous construisons ainsi notre base de référence (baseline) et nous en mesurons les précisions à 5 (P@5), 10 (P@10), 15 (P@15) documents. Cette baseline nous servira à comme comparatif afin de visualiser l'apport de notre système à travers le calcul du taux d'amélioration.

6.4.2 Résultats expérimentaux relatifs à l'évaluation de l'efficacité de la recherche

La première partie expérimentale consiste à présenter un exemple de reformulation de la requête. Cette reformulation est effectuée selon le modèle présenté dans la section 5.4.1.

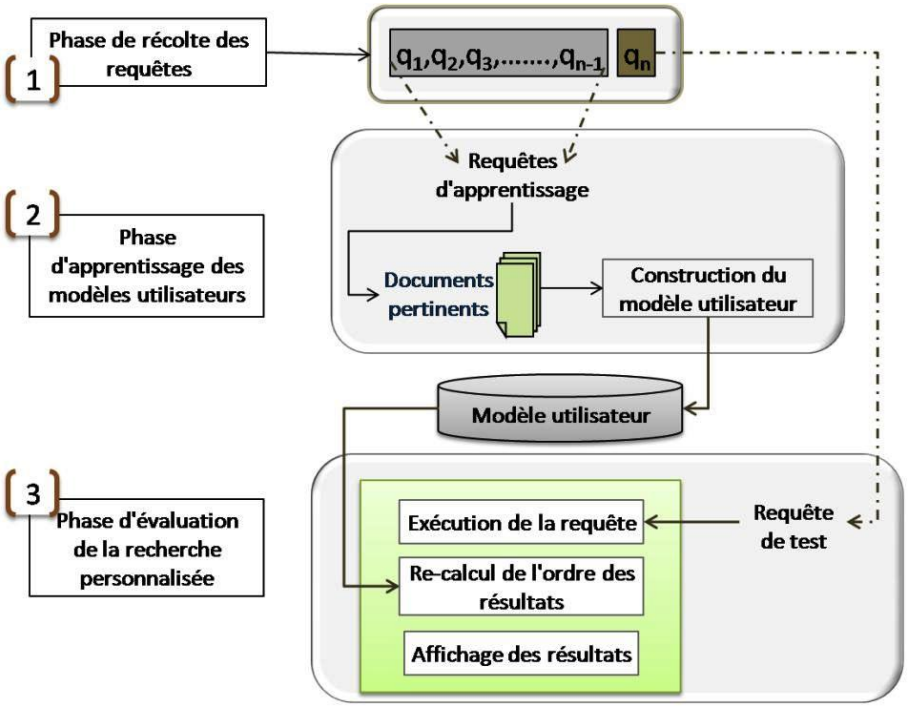


FIGURE 6.5 – Protocole d'évaluation

Nous présentons, pour une requête donnée, les résultats fournis par le moteur de recherche Google (sans reformulation de la requête de départ) et ceux fournis par SyPRISS après reformulation de la requête. Notons que la reformulation considère en plus des vecteurs d'intérêt de l'utilisateur, les termes des requêtes précédentes pour la session de recherche courante. Les résultats de recherche ont été extraits du moteur de recherche Google en juin 2010.

Les abréviations P_q et EU_q signifient respectivement requête précédente et requête utilisateur enrichie.

TABLE 6.6: Exemple de reformulation de requête

	Résultats fournis par Google (requête utilisateur="campus Paris")	**SyPRISS (requête utilisateur ="campus Paris")**	
		Pq = "hotel cheap" EUq = "hotel cheap Paris near campus"	Pq= "mathematics thesis" EUq = "campus sciences mathematics paris thesis"
1	Radio Campus Paris www.radiocampusparis.org/	Hotel Croix de Malte Paris www.croixdemalte-paris-hotel.com	Campus de Jussieu www.upmc.fr/
2	Ecoutez Radio Campus Paris en direct $24h/24$ *www.radiocampusparis.org/.../player_rcp.php*	Accommodation Paris France www.accommodationparisfrance.net	Université Paris-Est Marne-la-Vallée www.univ-mlv.fr/
3	Campus Université Paris 1 eep.univ-paris1.fr/	Hotel Agenor Paris www.agenor-paris-hotel.com	CNRS - Campus de Meudon www.cnrs-bellevue.fr/
4	Campus Paris-Jourdan www.paris-jourdan.ens.fr/	Hotel Aquarelle www.hotel-aquarelle.com	Université Paris II ermes.u-paris2.fr
5	Bienvenue sur le site du Campus du plateau de Saclay www.campus-paris-saclay.fr/	Hotel Lyon Mulhouse Booking Hotel Paris	Universite Paris 6 Pierre Et Marie Cury www.parisetudiant.com
6	IASTAR, le réseau des Radio Campus www.radio-campus.org/	Hôtel Le Quartier Republique, Le Marais www.lequartierhotelrm.com	Escola Normal Superior de Paris www.ens.fr/international/
7	European Management Center - Management hotelier - www.emc-campus.com/	hotel aux 3 porte www.hotel-aux3portes.com	ESPCI ParisTech www.espci.fr/
8	Dép. de Sciences Sociales - Ecole normale supérieure www.sciences-sociales.ens.fr/	Hotel Nord et Champagne Paris www.paris-hotel-nordetchampagne.com	Institut de Physique du Globe de Paris www.ipgp.fr/

La suite à la page suivante

TABLE 6.6 – suite de la page précédante

	Résultats fournis par Google (requête utilisateur="campus Paris")	**SyPRISS (requête utilisateur ="campus Paris")**		
		Pq = "hotel cheap" EUq = "hotel cheap Paris near campus"	Pq= "mathematics thesis" EUq = "campus sciences mathematics paris thesis"	
9	L'Université acteur du Campus Paris-Saclay *www.u – psud.fr/fr/l_universite/plan_campus.html*	Hôtel Beauséjour www.hotel-beausejour-paris.com	Institut Télécom www.institut-telecom.fr/	
10	Rejoignez Radio Campus Paris cet été ... *www.etudiantdeparis.fr/.../*	Hôtel Saint-Sébastien www.hotel-st-sebastien.com	Ecole centrale Paris www.ecp.fr/	

Nous présentons dans la figure 6.6, les mesures de P@10 de la baseline et du système intégrant la dimension centres d'intérêt du modèle de l'utilisateur. Dans cet histogramme, nous avons calculé la P@10 pour chaque scénario expérimental utilisé. Nous rappelons que $Sc1$, $Sc2$, $Sc3$ et $Sc4$ sont respectivement relatifs aux thèmes d'expérimentation 1, 2, 3 et 4 présentés dans la section 6.2.2.1.

Cette représentation affirme l'amélioration de la précision apportée par la dimension centres d'intérêt.

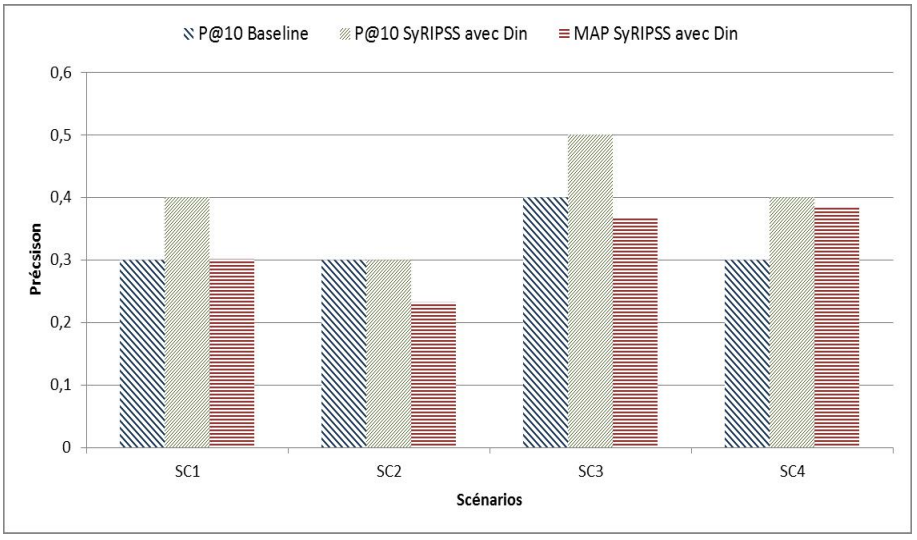


FIGURE 6.6 – Comparaison des mesures de P@10 et MAP entre la Baseline et le système intégrant la dimension centres d'intérêt du modèle de l'utilisateur

Le tableau 6.7 présente le taux d'amélioration de SyPRISS. La formule utilisée est la suivante :

$$Taux_amelioration = \frac{resultatsdusysteme - resultatsdelabaseline}{resultatsdelabaseline} \tag{6.1}$$

TABLE 6.7 – Le taux d'amélioration de SyPRISS

	P@5	**P@10**	**P@15**	**P@20**	**P@25**	**P@30**
SyPRISS(D_{in})	0.6	0.4	0.46	0.5	0.52	0.5
Baseline	0.6	0.3	0.4	0.45	0.44	0.4
Amélioration	0.0%	33.33%	15%	11.11%	18.18%	25%

Ce tableau ainsi que la figure 6.6 confirment l'amélioration des résultats fournis par SyPRISS en ayant intégré la dimension centres d'intérêt du modèle de l'utilisateur.

6.5 Evaluation de l'efficacité de la recherche intégrant les informations spatiales

Nous entamons cette phase de l'expérimentation avec le modèle de l'utilisateur dans son intégralité. Nous présentons le protocole de l'évaluation, ensuite les résultats expérimentaux.

6.5.1 Protocole d'évaluation de l'efficacité de la recherche intégrant les informations spatiales

Pour montrer l'apport de l'utilisation des données spatiales ainsi que les différentes positions prises lors des recherches, nous avons réalisé en plus des expérimentations utilisant uniquement la dimension textuelle et d'intérêt du modèle de l'utilisateur, des expérimentations qui incluent la prise en compte de la dimension spatiale. Cette dernière nous permet d'avoir, en plus des positions spatiales prises lors des différentes recherches, les données spatiales extraites des liens parcourus.
Cette expérimentation est effectuée en deux étapes : la première est l'évaluation de la mesure d'accessibilité proposée dans le cadre de la dimension spatiale et la seconde est l'évaluation de l'efficacité de la recherche en intégrant toute la dimension spatiale. Cette seconde étape est basée sur la même stratégie que pour l'évaluation de l'efficacité de la recherche (section 6.4.1). Nous effectuons, une simulation de soumission des requêtes des utilisateurs en gardant leurs mêmes choix quant aux résultats pertinents.

6.5.2 Evaluation de la mesure d'accessibilité

Comme présenté dans la section 4.4.2, nous avons proposé une mesure d'utilité sur laquelle nous avons basé notre mesure d'accessibilité. Cette mesure a la particularité de reposer sur la pondération dynamique des termes qui la composent. En effet, ces pondérations sont issues d'un processus d'apprentissage qui détermine, selon les réactions de l'utilisateur par rapport aux résultats personnalisés proposés, les degrés d'importance qu'il

accorde à chacun des facteurs adoptés.

Dans le cadre de cette expérimentation, nous avons donnée des noms à ces pondérations :

- Relevancy, représenté par α : La pertinence de l'entité par rapport aux préférences de l'utilisateur.
- Proximity, représenté par β : La proximité de l'entité par rapport à la position de l'utilisateur.
- PrefProx, représenté par λ : Le terme reflétant la proximité des entités préférées ou pertinentes par rapport à une entité référence.
- SimProx, représenté par γ : Le terme mesurant, pour une entité donnée, la densité d'entités similaires dans son entourage.

A chaque interaction entre l'utilisateur et un résultat de recherche, une nouvelle configuration est calculée pour ce tuple afin de déterminer leurs affinités et leurs importances comme perçues par l'utilisateur.

Le diagramme de la figure6.7 présente l'évolution de ces pondérations pour l'utilisateur considéré à cinq stades d'une session de navigation.

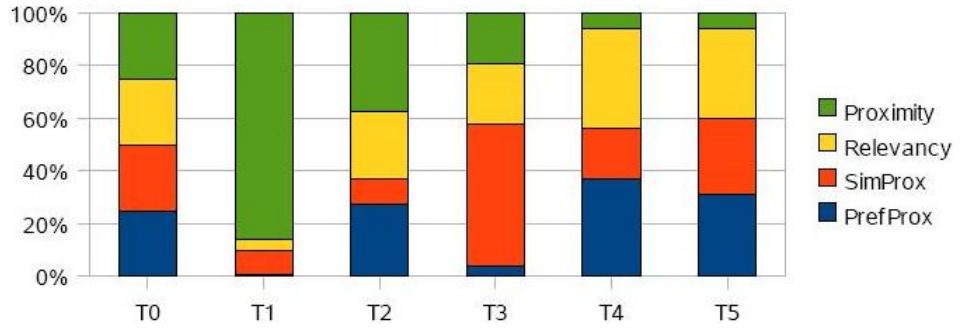
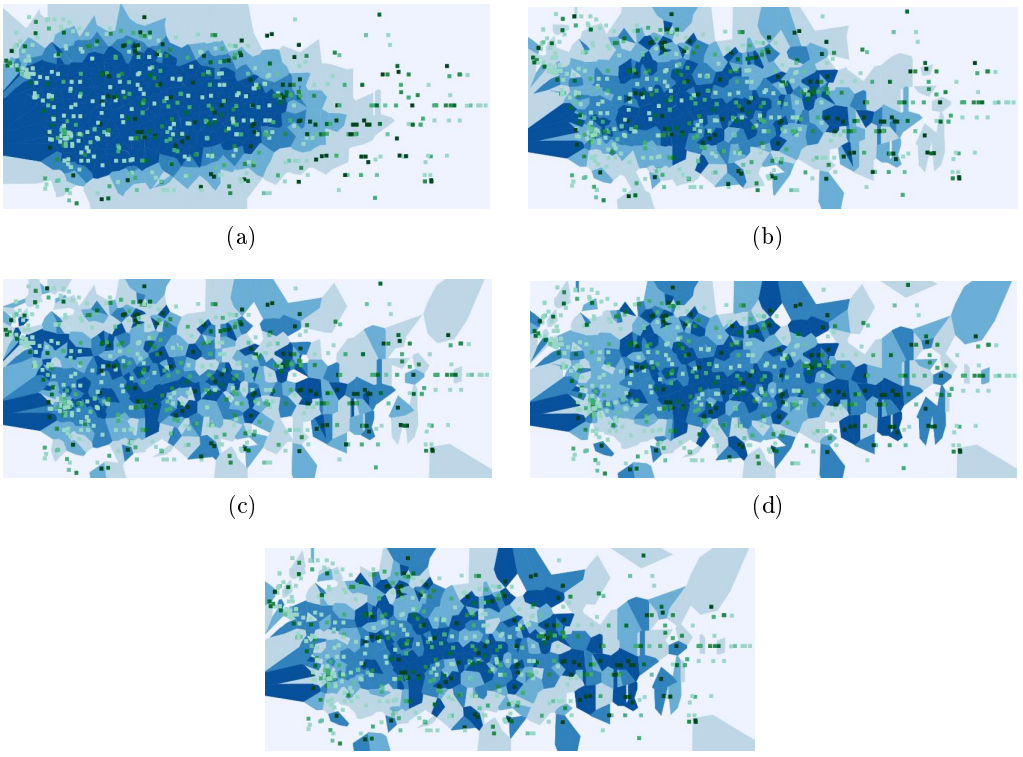


FIGURE 6.7 – Variation des pondérations de la mesure d'utilité

La particularité de la mesure d'utilité est que, pour le même état du modèle de l'utilisateur, peuvent exister plusieurs configurations possibles donnant des évaluations différentes de l'accessibilité. Les figures 6.8(a), 6.8(b), 6.8(c), 6.8(d) et 6.8(e) sont les résultats de l'application de la mesure d'utilité pour le même utilisateur considéré dans les expérimentation pendant les cinq stades représentés dans la figure 6.8.



(a) (b) (c) (d) (e)

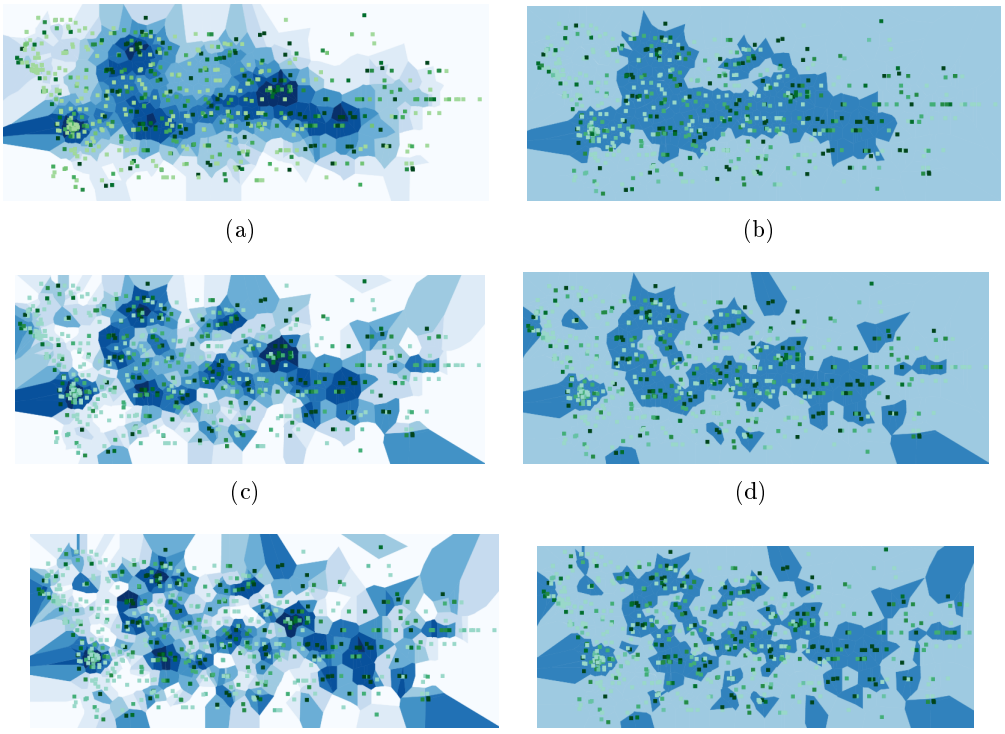
FIGURE 6.8 – Evaluation de la mesure d'utilité

Nous remarquons que cette mesure permet de déterminer et de délimiter les zones qui sont susceptibles d'intéresser l'utilisateur et qui respectent ses préférences et ses contraintes spatiales. L'analyse des propriétés des entités se trouvant dans les zones désignées comme accessibles montre leur adéquation avec les intérêts modélisés dans le modèle de l'utilisateur.

Les figures suivantes présentent les résultats de l'analyse de l'accessibilité pour l'utilisateur considéré en utilisant la mesure d'utilité expérimentée précédemment. Ces visualisations ont été obtenues en faisant varier la valeur de la permissivité du déplacement pour des valeurs correspondantes à des limites de déplacement égales à $30Km$, $15Km$ et $10Km$. Les figures 6.9(a), 6.9(c) et 6.9(e) présentent les niveaux d'accessibilité perçues par l'utilisateur. La couleur des points reflète la pertinence des entités par rapport à leurs caractéristiques. La couleur des zones reflète le degré d'accessibilité des leurs localisations correspondantes.
Les figures 6.9(b), 6.9(d) et 6.9(f) représentent les localisations les plus accessibles par rapport à cette mesure.

Les zones les plus pertinentes et denses sont retenues par la mesure d'accessibilité. Les localisations les plus accessibles par rapport à cette mesure respectent les intérêts de l'utilisateur du point de vue des caractéristiques préférées des entités spatiales puisqu'elles présentent une grande attractivité dans leurs entourages. Ces zones respectent au mêmes temps la contrainte de permissivité aux déplacements de l'utilisateur et s'alignent donc à ses préférences spatiales. L'effet du facteur de permissivité se présente dans la topologie des zones filtrées dont les tailles changent en fonction des voisinages et de leurs pertinences.

Il n'y a pas d'effets de bord dans les résultats obtenus puisqu'il existe des zones proposées à l'utilisateur qui se trouvent au bord de la zone étudiée. La décroissance non rapide de cette formulation donne les meilleurs résultats lors de la détermination des zones pertinentes et denses. Au fur et à mesure de la navigation et de l'exploration du contenu, le modèle de l'utilisateur est mis à jour et enrichi par des nouvelles informations concernant ses préférences. Ces changements d'intérêts spatiaux de l'utilisateur affectent directement la



(a) (b) (c) (d) (e) (f)

FIGURE 6.9 – Evaluation de la mesure d'accessibilité gravitaire

topologie et la distribution de l'espace accessible ainsi que les résultats proposés.

6.5.3 Résultats expérimentaux relatifs à l'évaluation de l'efficacité de la recherche intégrant les informations spatiales

Nous comparons la précision obtenue par SyPRISS intégrant le modèle de l'utilisateur avec toutes ses dimensions avec la recherche classique. Notons que nous exploitons les mêmes valeurs identifiées dans les expérimentations précédentes :

- Le nombre de concepts d'intérêt considérés : 5
- Le paramètre de réordonnancement $\gamma = 0.3$

Les résultats obtenus sont présentés dans la figure 6.10. Nous avons calculé la précision sur plusieurs points de documents restitués : (5,10,..., 100 premiers documents).
Nous présentons ensuite, dans la figure 6.11, une comparaison des mesures de précision à 10 et de MAP entre SyPRISS intégrant la dimension centres d'intérêt du modèle de l'utilisateur et SyPRISS intégrant la totalité du modèle à travers les différents scénarios expérimentaux. La courbe comparative confirme l'amélioration offerte par le modèle de l'utilisateur.

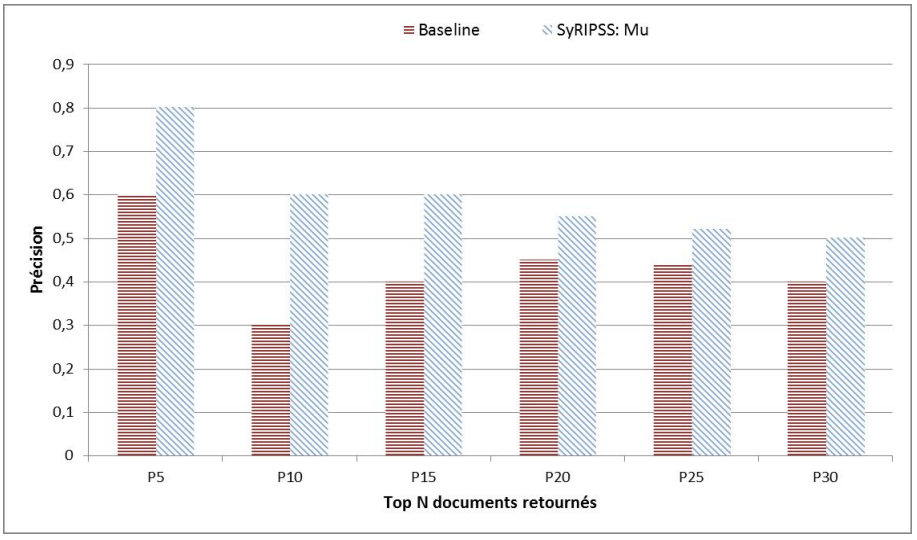


FIGURE 6.10 – Précision de SyPRISS intégrant le modèle de l'utilisateur

Dans la figure 6.11 les scénarios $Sc1$, $Sc2$, $Sc3$ et $Sc4$ sont respectivement relatifs aux thèmes d'expérimentation 1, 2, 3 et 4 présentés dans la section 6.2.2.1

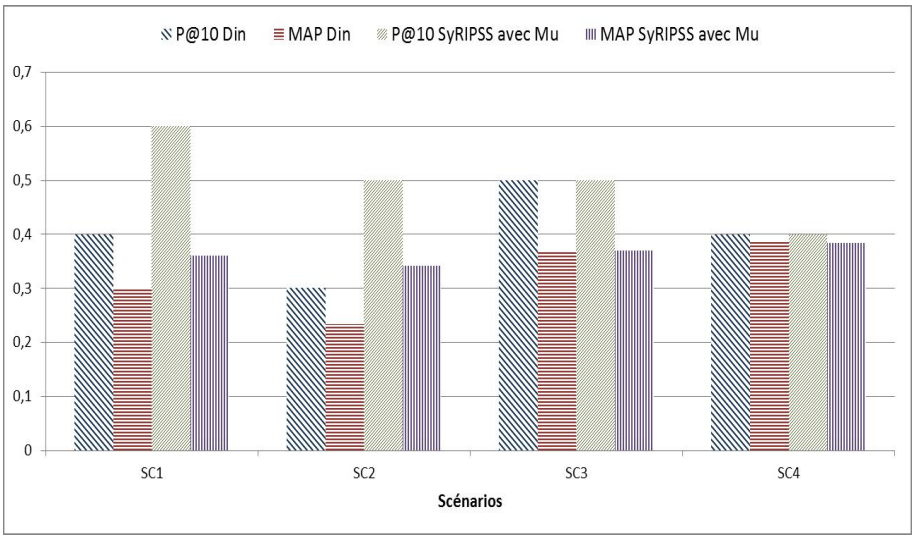


FIGURE 6.11 – Comparaison des mesures de P@10 et MAP entre SyPRISS intégrant la dimension centres d'intérêt du modèle de l'utilisateur et SyPRISS avec la totalité du modèle de l'utilisateur

6.6 Evaluation de la recherche exploitant le réseau d'utilisateurs

6.6.1 Protocole d'évaluation de la recherche exploitant le réseau d'utilisateurs

Afin de montrer l'apport de l'utilisation du réseau de modèles d'utilisateurs dans le système de recherche personnalisé, nous avons réalisé, en plus des expérimentations utilisant uniquement le modèle de l'utilisateur pour la personnalisation, des expérimentations qui incluent la prise en compte de ce réseau.

Cette expérimentation est effectuée en deux étapes : la première est l'évaluation de la classification des modèles des utilisateurs dans le réseau et la seconde est l'évaluation de l'efficacité de la recherche du système dans son intégralité. Cette seconde étape est basée sur la même stratégie que pour l'évaluation de l'efficacité de la recherche (section 6.4.1). Une simulation de soumission des requêtes des utilisateurs gardant leurs mêmes choix quant aux résultats pertinents est aussi effectuée pour cette expérimentation.

Pour la première étape de cette phase expérimentale, l'analyse effectuée pour l'étude des relations entre ces utilisateurs est une analyse conceptuelle à travers les treillis de gallois.

Notre objectif est de visualiser et de comprendre la composition des stéréotypes construits par le biais du calcul d'empreintes conceptuelles à partir de treillis de Galois. Ces empreintes vont nous aider à comprendre la structure et les propriétés des données extraites des requêtes des utilisateurs étudiés.

Les treillis de Gallois ont été introduits par (Birkhoff, 1940) et (Barbut and Monjardet, 1970) et consistent principalement à regrouper les objets en classes qui matérialisent des concepts du domaine d'application. Soit (O, A, I) le contexte correspondant à un treillis de gallois. Selon la terminologie de (Wille, 1992), O est l'ensemble des objets, A l'ensemble des propriétés de O et I est la relation binaire ente O et A : $(I \subseteq OxA)$. Les notions d'objet et de propriété dans notre contexte représentent respectivement les utilisateurs et leurs préférences. Le treillis de Galois est donc constitué de concepts comprenant des ensembles de personnes (objets) décrits par leurs préférences (propriétés). Nous procédons ensuite à l'évaluation des résultats de recherche personnalisée fournis à l'utilisateur en exploitant le réseau de modèles utilisateurs.

6.6.2 Résultats expérimentaux relatifs à de la recherche exploitant le réseau d'utilisateurs

6.6.2.1 Classification des utilisateurs à travers l'utilisation de leurs modèles

L'expérimentation été effectuée sur la base de la navigation des utilisateurs à travers la liste des résultats affichés et sur les différentes localisations spatiales choisies. Nous avons utilisé le même échantillon d'utilisateurs. Leurs préférences ont été traduites en relations binaires, comme le montre le tableau dans la figure 6.12.

	objet de recherche						déplacements				Préférences implicites								
	Hôtel	Restaurant	Auberge	Sport Nautique	Magasin	Banque	Tunis	Madrid	Sousse	Paris	Avion	5*	musées	internet	Proche aéroport	5*	Sauna	Zone touristique	internet
utilisateur 1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
utilisateur 2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
utilisateur 3	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
utilisateur 4	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
utilisateur 5	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
utilisateur 6	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
utilisateur 7	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
utilisateur 8	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1
utilisateur 9	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1
utilisateur 10	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
utilisateur 11	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
utilisateur 12	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
utilisateur 13	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
utilisateur 14	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
utilisateur 15	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
utilisateur 16	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
utilisateur 17	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
utilisateur 18	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

FIGURE 6.12 – Représentation binaire de l'échantillon de requêtes utilisateurs

Nous avons donc :

- O = {utilisateur1,..., utilisateur83}
- A = {Hôtel, Restaurant, Auberge, Sport Nautique, Magasin, Banque, Tunis, Madrid, Sousse, Paris, Avion, musées, internet, Proche aéroport, 5*, Sauna, Zone touristique, internet}

Afin de mieux comprendre l'origine des préférences des utilisateurs, nous avons notifié sur la figure chaque ensemble de propriétés :

- Objets de recherche = {Hôtel, Restaurant, Auberge, Sport Nautique, Magasin, Banque}
- Déplacements = {Tunis, Madrid, Sousse, Paris}
- Préférences implicites = {Avion, 5*, musées, internet, Proche aéroport, 5*, Sauna, Zone touristique, internet}

La figure 6.13 montre le treillis construit.

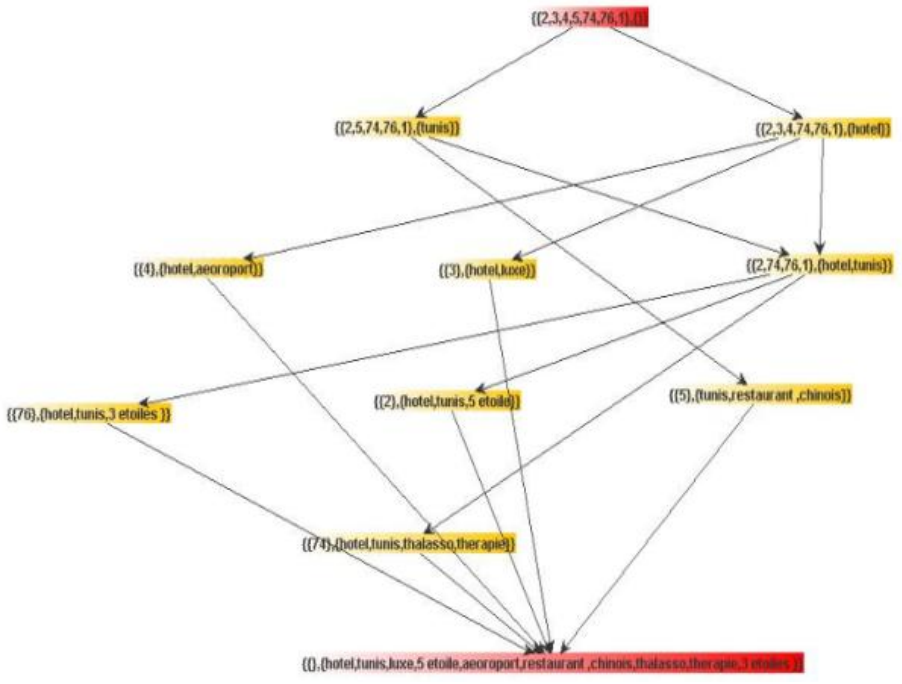


FIGURE 6.13 – Visualisation du treillis

L'objet $o \in O$ est caractérisé par deux paramètres appelés Relatedness (lien) et Closeness (proximité) (Le Grand et al., 2009). Dans notre contexte, la Relatedness indique

si l'objet (l'utilisateur considéré) possède des préférences communes à beaucoup d'autres objets (les utilisateurs du système). Ce paramètre sert donc à exprimer le degré de relation d'un utilisateur avec les autres utilisateurs, et ceci d'un point de vue quantitatif. Quant à la Closeness, elle sert à nous informer sur la quantité de propriétés communes ou non, c'est un paramètre qui exprime d'un point de vue qualitatif la connectivité de l'objet o avec les autres objets.

De ce fait, un utilisateur ayant une Relatedness élevée a plusieurs préférences en commun avec d'autres utilisateurs du système. S'il a aussi la valeur de la Closeness élevée, c'est qu'il possède de nombreuses propriétés en commun avec les utilisateurs du concept auquel il a été attribué. La distribution des utilisateurs est représentée dans la figure 6.14.

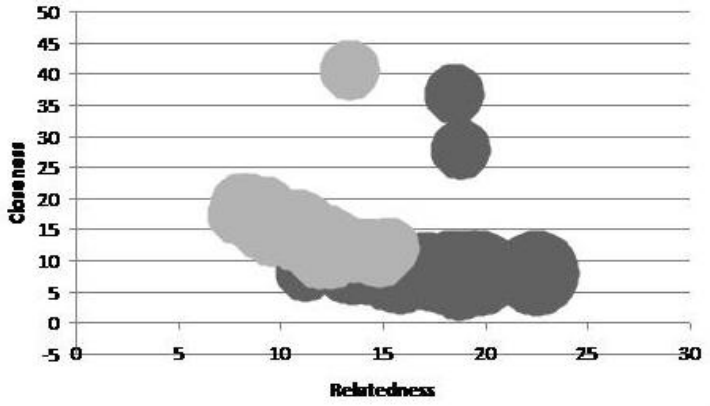


FIGURE 6.14 – Evaluation des Closeness et Relatedness

Chaque utilisateur est représenté par une bulle. La Relatedness est représentée dans l'axe des abscisses et la Closeness dans l'axe des ordonnées. Ces bulles sont homogènes et regroupées ; cela dénote d'un rapprochement entre les utilisateurs du système.

6.6.2.2 Personnalisation utilisant le réseau de modèles utilisateurs

Cette étape de l'expérimentation ayant pour objectif d'évaluer l'utilisation du réseau de modèles utilisateurs dans le système personnalisé, nous gardons le même nombre de documents jugés pertinents par les utilisateurs, à savoir 15. Dans la figure 6.15 nous présentons les mesures de p@10 et MAP en les comparant à celles de SyPRISS sans le réseau

de modèles utilisateurs. Cette figure est analogue de celle présentée dans la section 6.5.3, figure 6.11 dans laquelle nous avons présenté les mêmes courbes pour SyPRISS intégrant le modèle de l'utilisateur. Nous constatons une amélioration de la précision moyenne de SyPRISS ainsi que de la précision à 10.

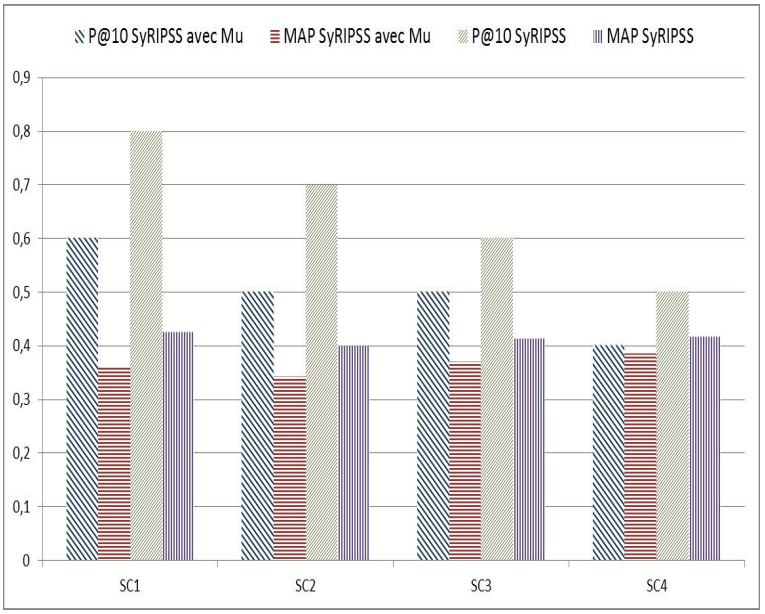


FIGURE 6.15 – Mesures de P10 et MAP de SyPRISS intégré dans sa globalité

6.7 Synthèse de l'évaluation

Nous présentons dans la figure 6.16 la mesure de la précision de la recherche pour les Top N documents retournés. Nous avons superposé les courbes par étape d'évaluation. Nous constatons une nette amélioration de la personnalisation des résultats présentés par SyPRISS comparativement à la recherche initiale. Cette amélioration se fait surtout ressentir entre les 5 et les 25 premiers documents retournés.

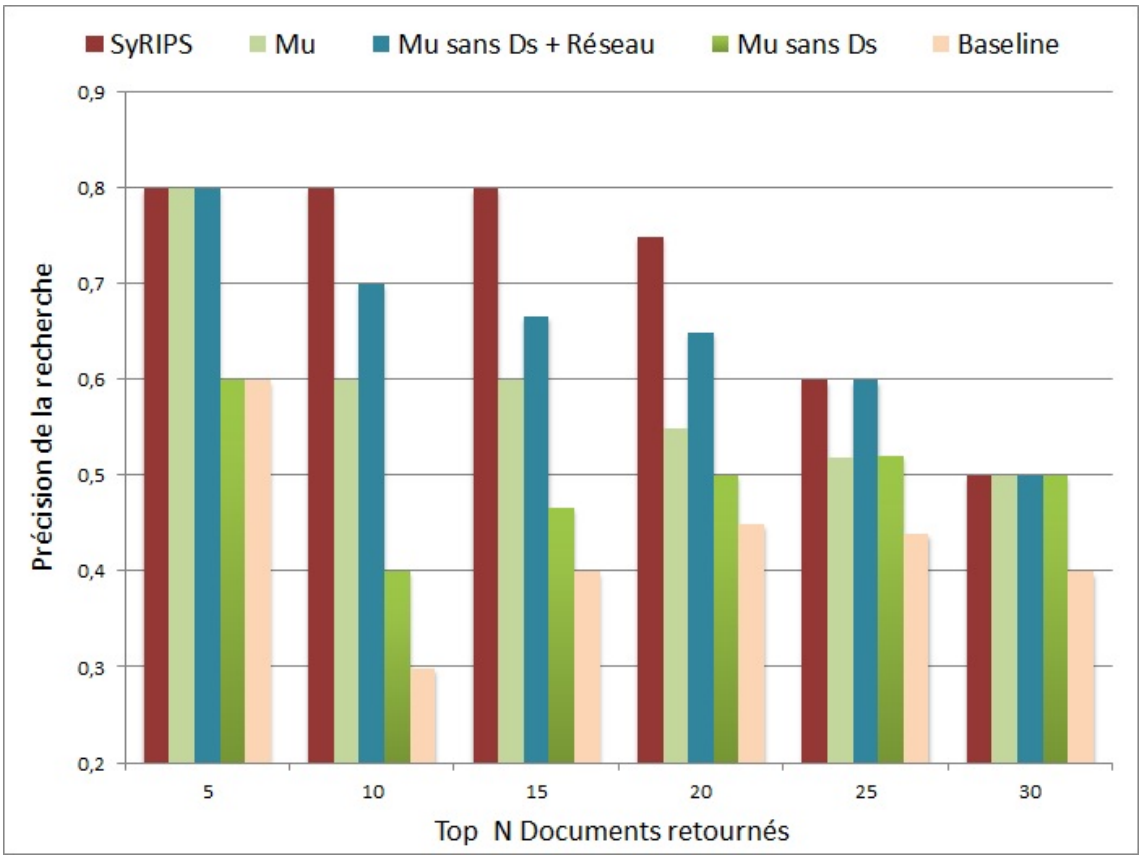


FIGURE 6.16 – Comparaison de la précision des top N documents retournés pour les différentes étapes de l'expérimentation

Nous récapitulons dans le tableau 6.8 les mesures de précision calculées lors des différentes étapes de l'évaluation en précisant, à chaque passage d'une phase expérimentale à une autre, le taux d'amélioration.

TABLE 6.8 – Récapitulatif des mesures de précision de l'expérimentation

	P@5	**P@10**	**P@15**	**P@20**	**P@25**	**P@30**
Recherche Simple (Baseline)	0.6	0.3	0.4	0.45	0.44	0.4
Recherche intégrant le M_u sémantique	0.6	0.4	0.46	0.5	0.52	0.5
Taux d'amélioration	0%	33.33%	15%	11.11%	18.18%	25%
Recherche intégrant les informations spatiales du M_u	0.8	0.7	0.66	0.65	0.6	0.5
Taux d'amélioration	33.33%	75%	43.48%	30%	15.38%	0%
Evaluation de la recherche exploitant le réseau d'utilisateurs	0.8	0.55	0.7	0.75	0.6	0.5
Taux d'amélioration	33.33%	83.33%	75%	66.67%	36.36%	25%

6.8 Conclusion

Dans ce chapitre, nous avons décrit les étapes d'évaluation et d'expérimentation du prototype développé. L'évaluation proposée a été effectuée en plusieurs étapes en commençant par SyPRISS sans la personnalisation et en y ajoutant en premier lieu le modèle de l'utilisateur sans la dimension spatiale ensuite avec la dimension spatiale et en second lieu la construction et l'exploitation du réseau de modèles utilisateur.

Cette évaluation en étapes a pour objectif de démontrer à chaque fois l'utilité de l'intégration et de l'ajout des données utilisateurs ainsi que l'exploitation de la construction du réseau de modèles utilisateur.

Nous pouvons dire que le prototype développé nous a permis de valider notre proposition d'un système de Personnalisation de la Recherche d'Information Spatiale et Sémantique sur le Web intégrant la modélisation de l'utilisateur.

Conclusion Générale

Les travaux développés et présentés dans ce mémoire s'inscrivent dans le cadre de la proposition d'un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur.
L'idée initiale de ce travail de thèse part du constat porté sur la croissance considérable du volume de données présentes dans le Web ainsi que leur diversité. A cela s'ajoute la volonté des utilisateurs d'avoir l'information adéquate et répondant le plus à leurs besoins. Ceci a conduit à la nécessité de prendre en compte l'utilisateur dans le processus de recherche d'information afin de lui fournir la personnalisation souhaitée.
Par ailleurs, une autre constatation porte sur le manque d'utilisation des données géographiques lors des recherches sur le web et cela malgré leurs présences croissantes.
Ainsi, un système de personnalisation de la recherche d'information a été proposé. Il se base sur les éléments suivants :

- La représentation et la maintenance des informations concernant l'utilisateur et ses manipulations sous forme de modèle. Le modèle multidimensionnel de l'utilisateur est caractérisé par la collecte implicite et évolutive des données qui vont le constituer.
- La prise en compte, des informations spatiales demandées et/ou fournies par l'utilisateur. Ces données peuvent provenir aussi bien des recherches effectuées que des navigations et sélections des utilisateurs dans les résultats fournis.
- Le troisième aspect concerne l'intégration de ces modèles de l'utilisateur dans un graphe représentatif des utilisateurs du système global dont l'objectif est de construire un réseau d'utilisateurs et de proposer une collaboration implicite entre les modèles.

Le système repose sur l'utilisation de données implicites venant de l'utilisateur et de ses navigations. En effet, la recherche personnalisée proposée débute par l'utilisation des données formulées dans la requête de l'utilisateur, se poursuit par la recherche des documents répondants aux besoins de l'utilisateur et l'interprétation des navigations à travers les résultats de recherche pour s'achever éventuellement par la construction et la mémorisation du modèle de l'utilisateur. A ces étapes est ajoutée la construction d'un réseau constitué de modèles utilisateurs. Ce réseau a pour objectif d'exploiter les principes de la collaboration implicite des utilisateurs, à travers leurs modèles, dans un réseau construit itérativement lors des différentes recherches.
Le système SyRIPSS, pour Système de Recherche d'Information Personnalisée Sémantique et Spatiale, contribue par :

- la construction d'un réseau de modèles utilisateurs et son exploitation pour la personnalisation de la recherche d'information,
- l'utilisation de distances sémantiques et spatiales entre les noeuds du réseau,
- la combinaison de la recherche spatiale et sémantique dans le web,
- le calcul implicite et en ligne de l'intérêt de l'utilisateur pour les documents visités. Ce calcul d'intérêt sert à la session courante et aux sessions futures de l'utilisateur.

Un prototype supportant le système proposé a été développé. Les différentes approches qui y sont utilisées ont été expérimentées ce qui a permis d'évaluer la proposition. L'évaluation a été effectuée en quatre étapes :

- Evaluation de la qualité du modèle de l'utilisateur contenant uniquement la dimension d'intérêt
- Evaluation de l'efficacité de la recherche en comparant l'amélioration apportée à cette étape par rapport au système sans le modèle de l'utilisateur.
- Evaluation de l'efficacité de la recherche avec l'intégration de la dimension spatiale du modèle de l'utilisateur.
- Evaluation de la recherche en exploitant le réseau de modèles utilisateurs : nous avons commencé par vérifier la construction du réseau d'utilisateurs et sa correspondance aux utilisateurs expérimentaux et aux scénarios d'expérimentation de départ. Nous

avons ensuite évalué l'efficacité des résultats fournis aux utilisateurs.

Ces expérimentations montrent une amélioration des résultats de la recherche personnalisée proposés à l'utilisateur. Cette amélioration a été constatée dans un premier temps par l'utilisation du modèle multidimensionnel de l'utilisateur et dans un second temps par l'ajout du réseau de modèles des utilisateurs.
Parmi les perspectives que nous pouvons envisager, nous citons :

- La construction automatique, ou semi-automatique de l'ontologie de la dimension d'intérêt du modèle de l'utilisateur proposé.
- La prédiction des besoins de l'utilisateur pour pouvoir estimer son type de besoin selon ses requêtes et ses navigations. Cette prédiction pourrait être basée aussi bien sur l'utilisation du modèle de l'utilisateur que sur une collaboration implicite entre les différents modèles proposés.
- Un autre point à proposer concerne la dimension spatiale du modèle de l'utilisateur. Cette dimension, comme présenté dans la proposition ne concerne que les informations spatiales de la donnée géographique. Une extension possible serait d'intégrer les aspects temporels.

Bibliographie

Amato, G. and Straccia, U. (1999). User profile modeling and applications to digital libraries. In *European Conference on Digital Libraries (ECDL)*, Paris. Springer. 37, 38

Asnicar, F. and Tasso, C. (1997). ifweb : A prototype of user model-based intelligent agent for documentation filtering and navigation in the world wide web. In *Proceedings of the 6th International Conference on User Modeling*, pages 3–11, Chia Laguna, Sardinia, Italy. 40, 49

Baazaoui, H., Aufaure, M., and Ben Mustapha, N. (2007). Extraction of ontologies from web pages : conceptual modeling and tourism. *Journal of internet Technologies*. 77

Baccini, A., Déjean, S., Kompaore, N. D., and Mothe, J. (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. *Technique et Science Informatiques*. 21

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. xvii, 14

Balabanovic, M. and Shoham, Y. (1997). Fab : Content-based, collaborative recommendation. *Communications of the ACM*, 40 :pp 66–72. 34, 49

Baldwin, J. F., Martin, T. P., and Tzanavari, A. (2000). User modeling using conceptual graphs for intelligent agents. In *Proceedings of the Linguistic on Conceptual Structures : Logical Linguistic, and Computational Issues*, pages 193–206, London, UK. Springer-Verlag. 36

Ballatore, A., McArdle, G., Kelly, C., and Bertolotto, M. (2010). Recomap : an interactive and adaptive map-based recommender. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 887–891, New York, NY, USA. ACM. 45

Barbut, M. and Monjardet, B. (1970). *Ordre et classification, algèbre et combinatoire, Tome 2.* 140

Batty, M. (2004). A new theory of space syntax. Technical report, CASA working paper no. 75. 89

Baziz, M., Boughanem, M., and Prade, H. (2007). Une approche de représentation de l'information en RI basée sur les sous- arbres. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Saint-Etienne, 28/03/07-30/03/07*, pages 335–350. Université de Saint-Etienne. 11

Bazsalicza, M. and Naim, P. (2001). *Data Mining pour le Web : Profiling, filtrage collaboratif, personnalisation client.* 32

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval : Two sides of the same coin ? *Communications of the ACM*, 35(12) :29–38. 15, 30

Bila, N., Cao, J., Dinoff, R., Ho, T. K., Hull, R., Kumar, B., and Santos, P. (2008). Mobile user profile acquisition through network observables and explicit user queries. In *Proceedings of the The Ninth International Conference on Mobile Data Management*, MDM '08, pages 98–107, Washington, DC, USA. IEEE Computer Society. 55

Bilhaut, F., Dumoncel, F., Enjalbert, P., and Hernandez, N. (2007). Indexation sémantique et recherche d'information interactive. In *COnférence en Recherche d'Infomations et Applications - CORIA 2007, 4th French Information Retrieval Conference, Saint-Etienne, France, March 28-30, 2007. Proceedings*, pages 65–76. Université de Saint-Etienne. 46

Billsus, D. and Pazzani, M. J. (1999). A hybrid user model for news story classification. In *Proceedings of the 7th International Conference on User Modeling*, pages 99–108. 36, 42

Birkhoff, G. (1940). *Lattice theory*, volume 25 of *Colloquium publications.* American Mathematical Society. 140

Blanco-Fernandez, Y., Pazos-Arias, J., Gil-Solla, A., Ramos-Cabrer, M., and M., L.-N.

(2008). Semantic reasoning : A path to new possibilities of personalization. *Proceedings of the 5th European Semantic Web Conference*. 36

Bohnert, F. (2008). Constraint-aware user modelling and personalisation in physical environments. *Adjunct Proceedings of the Sixth International Conference on Pervasive Computing*, pages pp 167–172. 39, 48

Borlund, P. (2003). The iir evaluation model : a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3). 20

Bouidghaghen, O., Tamine, L., and Boughanem, M. (2011). Personalizing mobile web search for location sensitive queries. In *Proceedings of the 2011 IEEE 12th International Conference on Mobile Data Management - Volume 01*, MDM '11, pages 110–118, Washington, DC, USA. IEEE Computer Society. 55

Bouzeghoub, M. and Kostadinov, D. (2005). Personnalisation de l'information : aperéu de l'état de l'art et définition d'un modéle flexible de profils. In *CORIA*, pages pp 201–218. 30, 37

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, 36(2) :3–10. 18

Broder, A. Z. (2007). The next generation web search and the demise of the classic ir model. In Amati, G., Carpineto, C., and Romano, G., editors, *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, volume 4425 of *Lecture Notes in Computer Science*, page 1. Springer. 18

Burrell, J. and Gay, G. (2001). Collectively defining context in a mobile, networked computing environment. In *CHI '01 : CHI '01 extended abstracts on Human factors in computing systems*, pages pp 231–232, New York, NY, USA. ACM. 47

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, pages pp 91–96. 46

Chen, C. C., Chen, M. C., and Sun, Y. (2001). Pva : a self-adaptive personal view agent system. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 257–262, New York, NY, USA. ACM. 43

Chen, L. and Sycara, K. (1998). Webmate : a personal agent for browsing and searching. In *AGENTS '98 : Proceedings of the second international conference on Autonomous agents*, pages pp 132–139. ACM Press. 40, 41, 48

Chen, Y.-S. and Shahabi, C. (2001). Automatically improving the accuracy of user profiles with genetic algorithm. In *Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing*, pages pp 283–288, Cancun, Mexico. 48

Chirita, P. A., Firan, C. S., and Nejdl, W. (2007). Personalized query expansion for the web. In *SIGIR '07 : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages pp 7–14, New York, NY, USA. ACM. 20

Cleverdon, C. W., Mills, J., and Keen, M. (1966). *Factors determining the performance of indexing systems*, volume 1 : Design ; volume 2 : Results. Cranfield, UK : Aslib Cranfield Research Project, College of Aeronautics. 20

Cool, C. and Spink, A. (2002). Issues of context in information retrieval (ir) : an introduction to the special issue. *Inf. Process. Manage.*, 38(5) :605–611. 29

Cotter, P. and Smyth, B. (2000). Ptv : Intelligent personalised tv guides. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA*, pages 957–964. AAAI Press / The MIT Press. 36

Daoud, M. (2009). *Accés personnalisé ? l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines ? travers l'historique des sessions de recherche.* Thése de doctorat, Université Paul Sabatier, Toulouse, France. 35, 41

Daoud, M., Boughanem, M., and Tamine-Lechani, L. (2008). Detecting session boundaries to personalize search using a conceptual user context. In Ao, S. I. and Gelman, L., editors, *World Congress on Engineering (Selected Papers)*, volume 39 of *Lecture Notes in Electrical Engineering*, pages 471–482. Springer. 127

Daoud, M., Tamine-Lechani, L., Boughanem, M., and Chebaro, B. (2009). A session based personalized search using an ontological user profile. In Shin, S. Y. and Ossowski, S., editors, *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*, pages 1732–1736, Honolulu, Hawaii, USA. ACM. 80

Dey, A. K. and Abowd, G. D. (2000). Towards a better understanding of context and context-awareness. In *Workshop on The What, Who, Where, When, and How of Context-Awareness (CHI 2000)*, The Hague, The Netherlands. 28

Di Lascio, L., Fischetti, E., and Gisolfi, A. (1999). A fuzzy-based approach to stereotype selection in hypermedia. *User Modeling and User-Adapted Interaction*, 9 :285–320. 36

Dieterich, H., Malinowski, U., Kuhme, T., and Schneider-Hufschmidt, M. (1993). *State of the Art in Adaptive User Interfaces*, pages 13–48. Elsevier Science Publishers B.V, Amsterdam. 31

Dolog, P. and Nejdl, W. (2007). Semantic web technologies for the adaptive web. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web : Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, chapter 23, pages 697–719. Springer, Berlin, Heidelberg. 36

Dourish, P. (2004). What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1) :19–30. 28

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C. (2003). Stuff i've seen : a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 72–79, New York, NY, USA. ACM. 41

Eirinaki, M. and Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, 3, N°1 :pp 1–27. 40

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. MIT Press. 76

Fink, J. and Kobsa, A. (2002). User modeling for personalized city tours. *Artificial Intelligence Review*, 18(1) 4 :pp 33–7. 38

Gaio, M. (2001). *Traitements de l'Information Géeographique : Représentations et Structures*. PhD thesis, Mémoire d'HDR, Université de Caen. 46

Gauch, S., Chaffee, J., and Pretschner, A. (2003). Ontology based personalized search and browsing. *Web Intelligence and Agent Systems*, 1 :pp 219–234. 35, 45

Gauch, S., Speretta, M., Chandramouli, A., and Micarelli, A. (2007). User profiles for personalized information access. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web : Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, chapter 2, pages pp 54–89. Springer, Berlin, Heidelberg. 3, 41, 54

Gentili, G., Micarelli, A., and Sciarrone, F. (2003). Infoweb : An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9) :715–744. 35, 48

Göker, A. and Myrhaug, H. I. (2002). User context and personalisation. In *ECCBR Workshop on Case Based Reasoning and Personalisation, Aberdeen*. 28

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste, a constant time collaborative filtering algorithm. *Information Retrieval Journal*, 4(2) :pp 133–151. 43

Gowan, J. (2003). A multiple model approach to personalised information access. Master's thesis, Master thesis in computer science, Faculty of science, Universitt de College Dublin. 34

Gupta, G. and Lee, W.-C. (2010). Collaborative spatial object recommendation in location based services. In *39th International Conference on Parallel Processing, ICPP Workshops 2010, San Diego, California, USA, 13-16 September 2010*, pages 24–33. IEEE Computer Society. 44

Hadjouni, M., Baazaoui, H., Aufaure, M., and Ben Ghezala, H. (2009a). Vers un système d'information pour la personnalisation sur le web basé sur la modélisation de l'utilisateur. *IC 2009 : 20èmes Journées Francophones d'Ingénierie des Connaissances, Hammamet, Tunisie, Mai 25-29, 2009.* 83, 86

Hadjouni, M., Baazaoui, H., Aufaure, M., and Ben Ghezala, H. (2010). Système de personnalisation web basé sur la construction d'un réseau d'utilisateurs. *Workshop La recherche d'information personnalisée sur le web in Extraction et Gestion des Connaissances, EGC, Janvier 2010.* 54

Hadjouni, M., Baazaoui, H., Aufaure, M., Claramunt, C., and Ben Ghezala, H. (2008). Towards personalized spatial web architecture. *In Workshop : Semantic Web meets Geospatial Applications, held in conjunction with AGILE 2008, 11th International Conference on Geographic Information Science.* 54, 117

Hadjouni, M., Haddad, M., Baazaoui, H., Aufaure, M., and Ben Ghezala, H. (2009b). Personalized information retrieval approach. *Web Information Systems Modeling (WISM 2009). In conjunction with the 21st International Conference on Advanced Information Systems : CAiSE 2009.* 54, 81, 83, 85, 99

Hadjouni, M., Zghal, H. B., Aufaure, M.-A., and Ben Ghézala, H. (2011). User modeling-based spatial web personalization. In König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R. J., and Jain, L. C., editors, *15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2011)*, volume 6882 of *Lecture Notes in Computer Science*, pages 41–50. Springer. 54

Handy, S. L. and Niemeier, D. A. (1997). Measuring accessibility : an exploration of issues and alternatives. *Environment and Planning - Part A*, 29(7) :1175–1194. 95

Harman, D. (1992). The darpa tipster project. *SIGIR Forum*, 26(2) :26–28. 105

Hersh, W. R., Elliot, D. L., Hickam, D. H., Wolf, S. L., Molnar, A., and Leichtenstien, C. (1995). Towards new measures of information retrieval evaluation. In Fox, E. A., Ingwersen, P., and Fidel, R., editors, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum), pages 164–170. ACM Press. 20

Hinze, A., Voisard, A., and Buchanan, G. (2009). Tip : Personalizing information delivery in a tourist information system. *J. of IT & Tourism*, 11(3) :247–264. 34

Hocquet, F. and Bazin, M. (2008). Optimisation d'un processus de personnalisation dans un systéme de recherche d'information touristique. Technical report, Laboratoire RIADI-Gdl, Tunis and Ecole Navale, Brest. 93

Hölscher, C. and Strube, G. (2000). Web search behavior of internet experts and newbies. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 337–346, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co. 19

Höök, K., Karlgren, J., Wærn, A., Dahlbäck, N., Jansson, C., Karlgren, K., and Lemaire, B. (1996). A glass box approach to adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6 :157–184. 10.1007/BF00143966. 32

Ingwersen, P. (1992). *Information Retrieval Interaction*. Taylor Graham. 18

Ingwersen, P. and Belkin, N. J. (2004). Information retrieval in context - irix : workshop at sigir 2004 - sheffield. *SIGIR Forum*, 38(2) :50–52. 10

Ingwersen, P. and Ja"rvelin, K. (2005). *The Turn : Integration of Information Seeking and Retrieval in Context*. Springer, first edition. 28, 29

Jansen, B., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs : a study and analysis of user queries on the web. *Information Processing & Management*, 36, n ?° 2 :207–227. 18

jason zien, jorg meyer, j. t., editor (2001). *Web query characteristics and their implications on search engines*. 18

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the Intnational Conference on Research in Computational Linguistics*, pages 19–33. 101

Johnson, J. D. (2003). On contexts of information seeking. *Information Processing and Management*, 39 :735–760. 28

Jrad, Z., Aufaure, M., and Hadjouni, M. (2007). A contextual user model for web personalization. In Weske, M., Hacid, M.-S., and Godart, C., editors, *WISE Workshops*, volume 4832 of *Lecture Notes in Computer Science*, pages pp 350–361. Springer. 36

Kekäläinen, J. and Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In *in Proceedings of the CoLIS 4 Conference*, pages 253–270. 20

Kim, K. and Allen, B. (2002). Cognitive and task influences on web searching behaviour. *Journal of the American Society for Information Science*, 53(2) :pp 109–119. 32

Kobsa, A. (2005). User modeling and user-adapted interaction. *User Modeling and User-Adapted Interaction*, 15(1-2) :185–190. 3

Kobsa, A. (2007). Generic user modeling systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web : Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, chapter 4, pages pp 136–154. Springer, Berlin, Heidelberg. 54

Kostadinov, D., Bouzeghoub, M., and Lopes, S. (2007). Accés personnalisé é des sources de données multiples : évaluation de deux approches de reformulation de requétes. In *INFORSID, Congrés Informatique des organisations et systémes d'Information et de Décision*, pages pp 73–88. 38

Koutrika, G. and Ioannidis, Y. (2004). Personalization of queries in database systems. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 597–, Washington, DC, USA. IEEE Computer Society. 45

Koutrika, G. and Ioannidis, Y. (2005). A unified user-profile framework for query disambiguation and personalization. *Workshop on New Technologies for Personalized Information Access in conjonction with the 10th International User Modeling*, pages pp 44–53. 35, 40, 45

Kraft, R., Maghoul, F., and Chang, C.-C. (2005). Y !q : contextual search at the point of inspiration. In Herzog, O., Schek, H.-J., Fuhr, N., Chowdhury, A., and Teiken, W., editors, *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 816–823. ACM. 40

Le Grand, B., Aufaure, M.-A., and Soto, M. (2009). Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux. In Ganascia, J.-G. and Gançarski, P., editors, *Extraction et gestion des connaissances (EGC'2009)*, volume RNTI-E-15 of *Revue des Nouvelles Technologies de l'Information*, pages 349–354. Cépadués-Editions. 141

Lesbegueries, J. and Loustau, P. (2006). Extraction et interprétation d'information géographique dans des données non-structurées. *CORIA (COnférence en Recherche d'Informations et Applications)*. 46

Li, J. and Zaiane, O. R. (2004). Combining usage, content, and structure data to improve web site recommendation. In *Proceedings of the 5th International Conference on Electronic Commerce and Web Technologies (EC-Web)*. 45

Lieberman, H. (1995). Letizia : An agent that assists web browsing. In *IJCAI (1)*, pages 924–929. Morgan Kaufmann. 40

Lieberman, H. (1998). Integrating user interface agents with conventional applications. In *Proceedings of the 3rd international conference on Intelligent user interfaces*, IUI '98, pages 39–46, New York, NY, USA. ACM. 34, 49

Lieberman, H., Fry, C., and Weitzman, L. (2001). Exploring the web with reconnaissance agents. *Commun. ACM*, 44(8) :69–75. 48

Lieberman, M. D., Samet, H., Sankaranarayanan, J., and Sperling, J. (2007). Steward : architecture of a spatio-textual search engine. In Samet, H., Shahabi, C., and Schneider, M., editors, *GIS*, page 25. ACM. 46

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations : item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1) :76–80. 44

Liu, F., Yu, C., and Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16, no. 1 :pp 28–40. 35

Lopes, C. T. (2009). Context features and their use in information retrieval. In *Third BCS-IRSG Symposium on Future Directions in Information Access*. 28

Mac Aoidh, E., McArdle, G., Petit, M., Ray, C., Bertolotto, M., Claramunt, C., and Wilson, D. (2009). Personalization in adaptive and interactive gis. *Annals of GIS*, 15(1) :23–33. 45

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge University Press, New York, USA. 28

Marcus, R. S. (1991). Computer and human understanding in intelligent retrieval assistance. In *Proceedings of the 54st ASIS Annual Meeting*, volume 28, pages 49–59. 15

Marinilli, M., Micarelli, A., and Sciarrone, F. (1999). A hybrid case-based architecture for information filtering on the web. In Schmitt, S. and Vollrath, I., editors, *Challenges for Case-Based Reasoning - Proceedings of the ICCBR'99 Workshops, Seeon Monastery, Germany, July 27-30, 1999*, pages 23–32. University of Kaiserslautern, Computer Science. 36

Messai, N., Devignes, M., Napoli, A., and Smaïl-Tabbone, M. (2005). Méthode sémantique pour la classification et l'interrogation des sources de données génomiques. *Revue des Nouvelles Technologies de l'Information (RNTI), Extraction des connaissances : Etat et perspectives (Ateliers EGC 2005). Ch1 : Modélisation des connaissances*, pages 43–47. 45

Micarelli, A. and Sciarrone, F. (2004). Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Modeling and User-Adapted Interaction*, 14(2-3) :159–200. 35, 48

Mladenic, D. (1999). Text-learning and related intelligent agents : A survey. *IEEE Intelligent Systems*, 14(4) :pp 44–54. 40, 49

Mobasher, B., Dai, H., Luo, T., Sun, Y., and Zhu, J. (2000). Integrating web usage and content mining for more effective personalization. In *EC-Web*, pages pp 165–176. 27

Mustapha, N. B., Aufaure, M.-A., and Zghal, H. B. (2007). Towards an architecture of ontological components for the semantic web. In Frasincar, F., Houben, G.-J., and Thiran, P., editors, *Proceedings of the CAISE 06 Third International Workshop on Web Information Systems Modeling WISM '06, Luxemburg, June 5-9, 2006*, volume 239 of *CEUR Workshop Proceedings*. CEUR-WS.org. 77

O'Grady, M. J. and O'Hare, G. M. P. (2004). Gulliver's genie : agency, mobility, adaptivity. *Computers & Graphics*, 28(5) :677–689. 47

Palacio, D., Cabanac, G., Sallaberry, C., and Hubert, G. (2010). Measuring effectiveness of geographic ir systems in digital libraries. In Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., and Frommholz, I., editors, *Research and Advanced Technology for Digital Libraries*, volume 6273 of *Lecture Notes in Computer Science*, pages 340–351. Springer Berlin / Heidelberg. 10.1007/978-3-642-15464-5_34. 2

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5 - 6) :pp 393–408. 49

Pazzani, M. J., Muramatsu, J., and Billsus, D. (1996). Syskill & webert : Identifying interesting web sites. In *AAAI/IAAI, Vol. 1*, pages 54–61. 41

Peis, E., del Castillo, J. M. M., and Delgado-Lopez, J. A. (2008). Semantic recommender systems. analysis of the state of the topic. *Hipertext.net*, 6 :(online). 45

Petra, F. (2002). Social awareness in a location-based information system. Master's thesis, Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden. 47

Razmerita, L. (2008). Modeling behavior of users in adaptive and semantic enhanced information systems : The role of a user ontology. In *Adaptive Hypermedia and Adaptive Web-Based Systems 2008*, Hannover, Germany. Springer Berlin / Heidelberg. xvii, 36, 37, 48

Razmerita, L. (2009). User modeling and personalization of advanced information systems. pages 3928–3933. 30

Rich, E. (1998). User modeling via stereotypes. In *In M.T. Maybury & W. Wahlster (Eds.), Readings in intelligent user interfaces (pp. 329-341)*, pages 329–342, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 41

Robertson, S. E. and Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146. 16, 17

Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. 17

Robertson, S. E., Walker, S., and Hancock-beaulieu, M. (1998). Okapi at trec-7 : Automatic ad hoc, filtering, vlc and interactive. In *Text REtrieval Conference*, pages 199–210. 124

Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *WWW '04 : Proceedings of the 13th international conference on World Wide Web*, pages 13–19, New York, NY, USA. ACM. 18

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 10, 14

Salton, G., Fox, E. A., and Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, 26(12) :1022–1036. 15, 16

Salton, G. and Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation.*, 29(4) :351–372. 34

Saracevic, T. (1975). Relevance : a review of and a framework for the thinking on the notion in information science. *Journal of The American Society for Information Science and Technology*. 11, 13

Saracevic, T. (1996). Relevance reconsidered. In *Information science : Integration in perspectives.Proceedings of the Second Conference on Conceptions of Library and Information Science, Copenhagen (Denmark)*. 11

Sato, K. (2004). Context-sensitive approach for interactive systems design : modular scenario-based methods for context representation. *Journal of Physiological Anthropology and Applied Human Science*, 23(6) :277–281. 28

Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In de Ma'ntaras, R. L. and Saitta, L., editors, *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004*, pages 1089–1090. IOS Press. 101

Shen, X., Tan, B., and Zhai, C. (2005). Implicit user modeling for personalized search. In *CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, pages pp 824–831, New York, NY, USA. ACM Press. 41

Sieg, A., Mobasher, B., and Burke, R. (2004a). Inferring user's information context : Integrating user profiles and concept hierarchies. In *the 2004 Meeting of the International Federation of Classification Societies*. 42

Sieg, A., Mobasher, B., and Burke, R. (2007a). Representing context in web search with ontological user profiles. In *in Proceedings of the Sixth International and Interdisciplinary Conference on Modeling and Using Context*. 20

Sieg, A., Mobasher, B., and Burke, R. D. (2007b). Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 525–534. ACM. 35

Sieg, A., Mobasher, B., Lytinen, S., and Burke, R. (2004b). Using concept hierarchies to enhance user queries in web-based information retrieval. In *in Proceedings of the International Conference on Artificial Intelligence and Applications, IASTED 2004*. 40

Silva, M. J., Martins, B., Chaves, M. S., Cardoso, N., and Afonso, A. P. (2006). Adding geographic scopes to web resources. *CEUS - Computers, Environment and Urban Systems - Elsevier Science*, 30(4) :pp 378–399. 46

Simonnot, B. (2008). La pertinence en sciences de l'information : des modéles, une théorie ? In *Problématiques émergentes dans les Sciences de l'Information*, pages 161–182. Hermés Lavoisier. 11

Stefani, A., S. C. (1998). Personalizing access to web sites : The siteif project. In *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia HYPERTEXT' 98 Pittsburgh*. 42, 49

Tahir, A., McArdle, G., Ballatore, A., and Bertolotto, M. (2010). Collaborative filtering - a group profiling algorithm for personalisation in a spatial recommender system. In *Proceedings of Geoinformatik 2010*, pages 44–50. 44

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 44

Tanudjaja, F. and Mui, L. (2002). Persona : A contextualized and personalized web search. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 3 - Volume 3*, HICSS '02, pages 67–, Washington, DC, USA. IEEE Computer Society. 42

Teevan, J., Dumais, S. T., and Liebling, D. J. (2008). To personalize or not to personalize : modeling queries with variation in user intent. In *SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 163–170, New York, NY, USA. ACM. 20

Tryfona, N. and Pfoser, D. (2005). Data semantics in location-based services. 3534 :168–195. 37

Usery, E. L. (1996). A feature-based geographic information system model. *Photogrammetric Engineering and Remote Sensing*, 62(7) :833–838. 2

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In Medeiros, C. B., Egenhofer, M. J., and Bertino, E., editors, *SSTD*, volume 3633 of *Lecture Notes in Computer Science*, pages 218–235. Springer. 46

van Rijsbergen, C. J. (1979). *Information Retrieval*. 10, 23

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. pages 315–323. 20

Wahlster, W. and Kobsa, A. (1986). Dialog-based user models. In *Proceedings of the IEEE, Special Issue on Natural Language Processing*, pages 948–960. 30

Weakliam, J., Lynch, D., Doyle, J., Bertolotto, M., and Wilson, D. (2005). Delivering personalized context-aware spatial information to mobile devices. In Li, K.-J. and Vangenot, C., editors, *W2GIS*, volume 3833 of *Lecture Notes in Computer Science*, pages pp 194–205. Springer. 47

Webb, G. I., Pazzani, M. J., and Billsus, D. (2001). Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2) :19–29. 40

Weissenberg, N., Voisard, A., and Gartmann, R. (2004). Using ontologies in personalized mobile applications. In *GIS '04 : Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages pp 2–11, New York, NY, USA. ACM Press. 47

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*, pages 62–67, Palo Alto. 41

Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers and Mathematics with Applications*, 23 :493–515. 140

Winter, S. and Tomko, M. (2006). *Translating the Web Semantics of Georeferences*, volume pp. 297-333, chapter Chapter in Web Semantics and Ontology, pages pp pp. 297–333. Idea Publishing, Hershey, PA. Taniar, D. ; Rahayu, W. (Eds.). 46

Worboys, M. F. (1996). Metrics and topologies for geographic space. *Computer*, 96 :1–11. 91

Wu, D., Zhao, D., and Zhang, X. (2008). An adaptive user profile based on memory model. In *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, WAIM '08, pages 461–468, Washington, DC, USA. IEEE Computer Society. 40

Wu, Z. and Palmer, M. S. (1994). Verb semantics and lexical selection. In Pustejovsky, J., editor, *32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings*, pages 133–138. Morgan Kaufmann Publishers / ACL. 100

Yang, C. and Chan, K. (2005). Retrieving multimedia web objects based on pagerank algorithm. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 906–907, New York, NY, USA. ACM. 90

Zemirli, W. N. (2008). *Modèle d'accés personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif.* PhD thesis, Université Paul Sabatier de Toulouse III. 34, 35, 42