وزارة التعليم العالي و البحث العلمي

Université Badji Mokhtar –Annaba-

Badji Mokhtar –Annaba- University

جامعة باجي مختار- عنابة-

Année**: 2014**

Faculté des sciences de l'ingéniorat
Département d'informatique

T H È S E

Pour obtenir le diplôme de
Docteur 3éme cycle

Recherche d'information multicritères

Filière : Informatique
Spécialité : Sciences et Technologies de l'Information et de la Communication

Préparée par

Djalila Boughareb

Jury :

Président: Mr Med Tarek Khadir	Pr.	Université Badji Mokhtar Annaba
Directeur de thèse : Mr Nadir Farah	Pr.	Université Badji Mokhtar Annaba
Examinateur : Mr Hamid Seridi	Pr.	Université de Guelma
Examinatrice : Mme Hassina Seridi	Pr.	Université Badji Mokhtar Annaba
Examinatrice : Mme Labiba Souici	Pr.	Université Badji Mokhtar Annaba

وزارة التعليم العالي و البحث العلمي

Université Badji Mokhtar –Annaba-

Badji Mokhtar –Annaba- University

جامعة باجي مختار- عنابة-

Année**: 2014**

Faculté des sciences de l'ingéniorat
Département d'informatique

T H È S E

Pour obtenir le diplôme de
Docteur 3éme cycle

Recherche d'information multicritères

Filière : Informatique
Spécialité : Sciences et Technologies de l'Information et de la Communication

Préparée par

Djalila Boughareb

Jury :

Président: Mr Med Tarek Khadir	Pr.	Université Badji Mokhtar Annaba
Directeur de thèse : Mr Nadir Farah	Pr.	Université Badji Mokhtar Annaba
Examinateur : Mr Hamid Seridi	Pr.	Université de Guelma
Examinatrice : Mme Hassina Seridi	Pr.	Université Badji Mokhtar Annaba
Examinatrice : Mme Labiba Souici	Pr.	Université Badji Mokhtar Annaba

Remerciement

Tout d'abord, je remercie notre clément Dieu qui m'a donné la puissance, le courage et la détermination nécessaire pour finaliser ce travail de thèse.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse, Professeur Nadir Farah pour ses conseils judicieux et son encadrement qualifié qui m'ont permis d'améliorer grandement la qualité de ce mémoire. Qu'il trouve ici l'expression de mon très grand respect et du plaisir que j'ai à travailler avec lui.

Je souhaite exprimer toute ma reconnaissance aux professeurs Med Tarek Khadir, Hassina Seridi, Hamid Seridi et Labiba Souici pour l'honneur qu'ils me font en acceptant d'être les rapporteurs de ce mémoire.

Mes chaleureux remercîments s'adressent également à professeur Hassina Seridi et au docteur Abdellah Benouareth pour leur soutien moral et leur encouragement.

Je tiens également à remercier très sincèrement ma très chère mère, mon très cher père et mes chers frères et sœurs, qui m'ont beaucoup soutenu durant les moments difficiles.

Un grand remerciement à mon très cher frère Ali pour le temps qui m'a consacré, sa bonté et sa générosité.

Merci aussi à tous mes collègues de la première promotion LMD qui m'ont accompagné par leurs encouragements et leur continuel sourire.

Enfin, je tiens à exprimer ma gratitude aux personnes qui ont contribué de près ou de loin au bon déroulement de mon cursus et à l'élaboration de ce mémoire.

Dédicace

Je dédie ce travail à:

Ma mère et mon père.

Mes frères Ali, Sofiane et Abd erraouf et mes sœurs Amel, Rima et Abir.

Aux petits: iyed, loujain et Mohamed Amine.

Mon beau frère Ahmed et ma belle sœur Samiha.

Mon oncle Mohamed Chebini et sa petite famille.

Mes amies et à tous les étudiants de la première promotion LMD.

Tous ceux qui portent le nom de Boughareb et Chebini.

Résumé

La problématique abordée dans cette thèse se situe dans le cadre de la recherche d'information contextuelle et elle vise à contribuer dans l'amélioration des résultats de recherche. Ceci, à travers la considération de plusieurs dimensions de contexte pouvant affecter le comportement de recherche de l'utilisateur du Web.

Le contexte de recherche peut inclure plusieurs dimensions telles que le temps, le lieu, l'historique de navigation, la tâche en cours, etc. Dans le domaine de la recherche d'information, il a pris une part de recherche importante visant à améliorer la pertinence des résultats de recherche qui est le même objectif de ce travail de thèse. Dans ce cadre, nous proposons de combiner plusieurs dimensions contextuelles qui sont le temps, l'évènement pouvant influé la recherche et l'activité de navigation récente de l'utilisateur dans le but d'identifier automatiquement le domaine cible et récupérer ainsi un contenu pertinent qui correspond à ce que l'utilisateur attend de recevoir sans exiger son implication explicite.

À cette fin, la première partie du travail consiste à proposer un modèle contextuel de l'utilisateur qui doit identifier son besoin informationnel à partir de ce qu'il a recherché récemment tenant compte de l'évènement affectant la recherche.

Après cela, une approche d'expansion de requête est proposée pour suggérer des mots-clés pertinents et aider l'utilisateur à mieux exprimer son besoin de recherche et lui rapprocher ainsi de l'information désirée. Les résultats expérimentaux sont encourageants et suggèrent que l'emploi des dimensions contextuelles considérées peut contribuer à l'amélioration de la recherche sur le Web.

Mots-clés: recherche d'information contextuelle, recherche d'information sur le web, contexte de recherche, contexte temporel, expansion des requêtes, intérêts récents de l'utilisateur, modélisation de l'utilisateur, évènement affectant.

Abstract

The problem addressed in this thesis takes part from the contextual information retrieval field and it aims to contribute in improving the search results through the consideration of several contextual dimensions which can affect user's search behavior on the Web.

The search context may include several dimensions such as time, location, history of interaction, current task, etc. In the information retrieval field, it has taken a very important part of research aiming to improve the relevance of the search results that is the perspective of the current work. In this area, we propose to combine two contextual dimensions which are time, event and the recent navigation activity of the user in order to automatically identify the domain targeted by the user query and retrieve relevant contents that match what he expects to receive without requiring his explicit involvement.

To this end, the first part of the work consists of proposing a contextual user model which should identify the user's search need based on what he looked for a few times ago and taking into account the event affecting the search.

After that, a query expansion algorithm is proposed to suggest relevant keywords and to help the user to better express his information need. The experimental results are encouraging and suggest that the use of the considered contextual dimensions can contribute to the improvement of the search on the Web.

Keywords: contextual information retrieval, web information retrieval, search context, temporal context, query expansion, user's recent interests, user modeling, influential event.

الملخص

الإشكالية التي تناولتها هذه الأطروحة تندرج تحت إطار البحث عن المعلومات عبر الانترنت والتي تأخذ بعين الاعتبار سياق البحث و ذلك لتحسين نتائجه و الحصول على معلومات تتناسب و احتياجات المستعمل. هذا من خلال دراسة أبعاد السياق التي يمكن أن تؤثر على سلوك مستخدم شبكة الانترنت.

ويمكن أن يشمل سياق البحث عدة أبعاد مثل التوقيت، المكان، تاريخ التصفح، المهمة الحالية، وما إلى ذلك. في مجال البحث عن المعلومات، استفاد السياق من حصة كبيرة من الأبحاث وهو ما تقوم عليه هذه الأطروحة.

في هذا السياق، فإننا نقترح الجمع بين عدة أبعاد سياقية هي الوقت، نشاط المستعمل والحدث الذي يمكن أن يؤثر على الإبحار على الانترنيت وذلك من أجل تحديد الهدف المراد من إجراء البحث والحصول على المعلومات المطلوبة و المناسبة لما يتوقع المستخدم تلقيه دون الحاجة إلى تدخل صريح.

تحقيقا لهذه الغاية، الجزء الأول من العمل يهدف إلى تقديم نموذج سياقي للمستخدم يساعد على تحديد الاحتياجات المعلوماتية وهذا من خلال النشاط الذي أجراه المستخدم في الآونة الأخيرة مع الأخذ بعين الاعتبار الأحداث التي من شانها أن تؤثر على البحوث.

بعد ذلك، يقترح العمل مساهمة في مجال إثراء السؤال و اقتراح كلمات ملائمة ذات صلة مع السؤال تساعد المستخدم على التعبير عن حاجته المعلوماتية بشكل أفضل وذلك لتقريب المعلومات المطلوبة منه. النتائج التجريبية مشجعة وتشير إلى أن استخدام هذه الأبعاد السياقية يمكن أن يساهم في تحسين البحث على شبكة الإنترنت.

كلمات البحث: البحث السياقي عن المعلومات ، البحث على شبكة الانترنت، سياق البحث، السياق الزماني، إثراء الأسئلة، الاهتمامات الحالية للمستعمل، تصميم المستعمل، الحدث.

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

Abréviations

ANNIE A Nearly-New Information Extraction System

API Interface de Programmation d'Applications

CLEF Cross-language evaluation forum

DGC Discounted Cumulative Gain

GATE General Architecture for Text Engineering

IDE Integrated Development Environment

MAP Median Average Precision

NTCIR NII Test Collection for IR Systems

RI Recherche d'information

RIC Recherche d'Information Contextuelle

RICW Recherche d'Information Contextuelle sur le Web

RIG Recherche d'information géographique,

RIP Recherche d'Information Personnalisée

RIS Recherche d'Information Social

RIW Recherche d'Information sur le Web

RNA Réseaux de Neurones Artificiels

RTE Real Time Event

SRI Système de Recherche d'Information

TF.IDF Term Frequency-Inverse Document Frequency

TREC Text REtrieval Conference

URL Uniform Resource Locator

WIR Information Retrieval on the Web

WSD Word Sens Disambiguation

PARTIE I : CONTEXTE ET MOTIVATIONS

CHAPITRE 1 : INTRODUCTION GÉNÉRALE

1.1. INTRODUCTION

Dans le Web d'aujourd'hui caractérisé par la nouvelle génération de techniques d'usage de l'information, une grande variété de possibilités de recherche a été ouverte dans différents domaines. Le principe de production, d'étiquetage et de partage d'informations sans aucune restriction a mené la communauté experte en Recherche d'Information (RI) à comprendre davantage le comportement de navigation des utilisateurs du Web à travers leurs sites favoris, leurs commentaires, les étiquettes qu'ils attribuent aux documents, les relations sociales établies,.., etc. En conséquence de quoi, il est devenu beaucoup plus facile de présenter l'information appropriée à son demandeur.

Cependant, et à la lumière de l'évolution éprouvée dans le domaine de la recherche d'information sur le Web, le problème classique de l'identification du besoin en information de l'utilisateur qui est exprimé souvent par le biais de requêtes courtes et ambiguës n'a pas été entièrement résolu.

Dans le domaine de la recherche d'information, l'historique de navigation de l'utilisateur et son feedback de pertinence représentent les principales sources d'information exploitées dans le but de mieux comprendre son comportement et faciliter davantage son accès à l'information qui l'intéresse. L'usage de ces sources [GAR05; BIA09; BAE04; ZHA06; LV10] a permis d'augmenter la pertinence des résultats de recherche. Néanmoins, il faut noter que leur collecte et leur maintien dans un état cohérent constituent un véritable défi, sachant que le comportement de navigation des utilisateurs peut changer d'un jour à l'autre et d'une période de temps à l'autre en fonction de plusieurs paramètres regroupés sous le terme du *contexte* qui peut inclure le temps [GAR05; JAI05; ZHA06; SAI11; BOU13b, BOU12b], la localisation [SAI11;

BOU12b], la tâche en cours [ASF12; SHE05], l'événement pouvant influé la recherche [BOU12a, 13a], etc.

Les travaux abordés dans cette thèse s'inscrivent dans le cadre de la recherche d'information en générale et la recherche d'information contextuelle sur le Web en particulier. Ces travaux visent à contribuer dans l'amélioration de l'accès à l'information pertinente dispersée au sein d'une masse informationnelle gigantesque et variée en s'appuyant sur le contexte de la recherche. L'objectif majeur est tout d'abord, l'identification du besoin en information de l'utilisateur afin de lui assister par la suite dans sa recherche en lui proposant une nouvelle formulation de sa requête pouvant lui assurer de rapprocher de l'information désirée.

1.2. MOTIVATIONS

L'utilisation du contexte de l'utilisateur dans le domaine de la recherche d'information sur le Web est une voie de recherche prometteuse, et elle faisait l'objet de plusieurs travaux [ALO07; DIA09; JIN11; PAS08; DIN11 ; BOU11, 12a, 13a,b] parmi lesquels le présent travail de thèse qui vise à bénéficier du minimum de renseignements recueillis implicitement sur l'activité de navigation de l'utilisateur dans le but de mieux identifier son besoin en information, puis le servir avec du contenu pertinent.

Pour commencer, il est essentiel de définir clairement les points de départ pour une bonne atteinte des objectifs soulignés. En fait, l'analyse du trafic de recherche du moteur de recherche le plus populaire Google nous a permis d'extraire un ensemble d'heuristiques à propos du comportement global des utilisateurs à travers le Web à savoir :

- L'aspect périodique qui caractérise un nombre important de requêtes sur le Web [ALF09; SAN07; VLA04] ;
- La présence d'une relation de dépendance recherche-événements agissants un peu partout à travers le monde ce qui reflète la baisse ou l'augmentation de la fréquence de soumission des requêtes connexes au fil du temps.

1.3. PROBLÉMATIQUE ET CONTRIBUTIONS

La problématique abordée dans cette thèse se situe dans le cadre de la Recherche d'Information Contextuelle sur le Web (RICW) et vise à contribuer dans l'amélioration des résultats de recherche, ceci à travers la considération de plusieurs dimensions contextuelles pouvant affecter le comportement de recherche de l'utilisateur du Web. Selon le principe de l'approche proposée, nous avons traité de manière similaire tous les utilisateurs ayant des comportements de navigation semblables dans un contexte de recherche donnée. Donc, il s'agit de traiter l'utilisateur tenant compte de son contexte de recherche courant qui inclut l'ensemble des centres d'intérêt récemment fréquentés.

La première partie de notre travail consiste à construire un modèle de contexte qui doit aider à identifier le besoin de recherche de l'utilisateur en se basant sur ce qu'il cherchait quelque temps auparavant et prenant en compte la dimension temps et évènement pouvant influé la recherche. Par exemple, un utilisateur qui a effectué une recherche sur les "réseaux de neurones" le mardi, en plus qu'il a mené une recherche sur "l'acide nucléique" quelque temps auparavant. Il est fort probable qu'il visait les réseaux de neurones biologiques et non pas les réseaux de neurones artificiels et qu'il est intéressé à la biologie et non pas à l'informatique.

Après l'identification du domaine visé par l'utilisateur, une approche d'expansion de requête est proposée dans le but de suggérer des mots-clés pertinents et pour aider l'utilisateur à mieux se rapprocher de l'information désirée, ce qui représente la deuxième partie de notre contribution.

Les travaux sur la modélisation de l'utilisateur et l'expansion de requête qui tentent de fournir un contenu satisfaisant et pertinent sont nombreux [SCH04; BAE04; GAR05; SHE05; ZHA06; QIU06; DAO07; DIN11; MOG11; SAI11; ASF12; BOU13b]. Les principales différences entre notre travail et les travaux qui le précèdent, c'est que nous essayons de traiter ces sujets suivant une nouvelle démarche.

- En premier temps, la dimension tâche ou application a été utilisée dans [SHE05; ASF12] différemment à ce travail. Son utilisation en association avec le paramètre d'événement en temps réel a également donné de bons résultats [BOU12a, 13a]. Effectivement, les documents visités par l'utilisateur quelques

temps avant la soumission d'une requête de recherche qui construisent ce que nous appelons dans ce memoire les intérêts récents de l'utilisateur ont été exploités dans ce travail comme étant une dimension contextuelle en vue de déterminer ses centres d'intérêt courants, ce qui pourra aider à identifier son besoin de recherche actuel. Notre principe est proche de celui de [SHE05; ASF12] où ils ont proposé d'employer les requêtes et les documents visités récemment qui sont liés à la tâche ou l'activité courante de l'utilisateur, ce qui permet d'améliorer la pertinence des résultats des requêtes correspondant à l'activité courante de l'utilisateur et non pas celle des autres requêtes. Pour éviter cela, nous avons pensé à construire un modèle qui tient compte de toute l'activité de navigation récente de l'utilisateur qui implique souvent plus d'une tâche et vise plusieurs domaines d'intérêt.

- La dimension temps a été amplement étudiée dans plusieurs ouvrages et a été utilisée de différentes manières [ZHA06; SAI11; BOU12b]. Dans le présent travail et par l'emploi de la dimension temps, nous essayons d'investir le critère périodicité qui peut caractériser un grand nombre de requêtes de recherche [BOU11, 13b].

- Davantage, nous proposons d'utiliser la session de recherche qui inclut une seule requête et qui vise à combler un besoin en information unique comme l'élément de clustering de base, plutôt que de faire un clustering basé utilisateur [GAR05; QIU06; MOG11] ou un clustering basé session de navigation [SCH04; DAO07]. Effectivement, la suivie de l'une de ces deux démarches engendre la production d'un nombre important de clusters en raison du fait que le regroupement d'utilisateurs ayant un comportement de navigation similaire sur le Web pendant une période de temps considérable ne semble pas être évident sachant que si ce cas se présente, il n'est pas possible de déterminer si le flux de requêtes en provenance d'une telle adresse IP appartient à un utilisateur unique.

- Les sessions de recherche similaires ont représenté une source de termes essentielle pour l'expansion des requêtes. En effet, et dans une perspective de résoudre les problèmes de brièveté et d'ambigüité des requêtes, nous avons proposé de les enrichir en se basant sur les feedbacks des utilisateurs sur les

sessions de recherche similaires ayant des contextes de recherche semblables. Outre du feedback de pertinence, les pages visitées quelque temps avant la soumission d'une requête sont également utilisées pour identifier les sessions de recherche similaires. En fait, nous avons considéré similaires, toutes les sessions de recherches conduites dans des contextes similaires et ayant des feedbacks de pertinence similaires.

- L'impact des événements sur les tendances de recherche des utilisateurs a été exploité, dont la présence/l'absence d'un événement (régional ou mondial) a été considérée comme un paramètre contextuel.

1.4. PLAN DE LA THÈSE

Ce mémoire est divisé en trois parties : la première partie s'intitule contexte et motivation, elle comporte le chapitre 1 qui présente le contexte général de la thèse prenant part de la recherche d'information contextuelle sur le Web.

La deuxième partie présente un état de l'art organisé en trois chapitres dont le chapitre 2 présente les concepts clés de la recherche d'information y compris le processus de recherche, les modèles de base ainsi que le protocole d'évaluation adopté dans le domaine. Le chapitre 3 est consacré à une branche de la RI fondée sur la notion du contexte qui est la Recherche d'Information Contextuelle (RIC) où nous décrivons les différentes dimensions du contexte employées dans le domaine de la RI, nous détaillons également les étapes du processus de modélisation contextuelle avant de finir par présenter les différentes méthodes d'évaluation adoptées en RIC. Dans le chapitre 4, nous présentons un aperçu des méthodes d'expansion de requêtes existantes.

La troisième partie ou la partie contribution comporte trois chapitres (les chapitres 5, 6 et 7). Dans le chapitre 5, nous présentons une étude sur le comportement de recherche des utilisateurs du Web. Dans cette étude, nous focalisons sur les données extraites depuis le trafic du moteur de recherche Google. Dans le chapitre 6, nous présentons notre première contribution menée dans une perspective d'améliorer l'identification du besoin d'information et le domaine d'intérêt visé par la requête de l'utilisateur. Le dernier chapitre présente une deuxième contribution dans l'axe d'expansion de requêtes dans laquelle nous sommes basés sur le contexte immédiat de la tâche de navigation de

l'utilisateur comme une source pour l'enrichissement contextuelles des requêtes. Enfin nous dressons les conclusions à retirer de ce travail ainsi que les perspectives envisageables pour nos recherches futures.

Le mémoire contient également trois annexes organisées comme suit :

L'annexe A présente un extrait des données d'historique de navigation utilisées dans les expérimentations ainsi qu'un tableau des adresses des sites à partir desquelles nous avons extrait les terminologies de chacun des domaines d'intérêt traités dans ces travaux. L'annexe B présente la description générale des outils utilisés dans l'implémentation avec des captures écran du système développé. L'annexe C présente quelques notions des réseaux de neurones artificiels.

PARTIE II : ÉTAT DE L’ART

CHAPITRE 2 : CONCEPTS CLÉS DE LA RECHERCHE D'INFORMATION

2.1. INTRODUCTION

Le terme de recherche d'information a été introduit par Calvin Mooers en 1950 [MOO50] et il signifie la discipline informatique qui traite la problématique d'accès à l'information pertinente dans une masse de données souvent importante. Elle peut se définir comme l'ensemble d'opérations, méthodes et procédures qui permettent de retrouver à partir d'une collection de documents, l'information pouvant répondre à une question sur un sujet précis.

Lorsque la recherche s'effectue dans un espace documentaire important, la possibilité de retrouver l'information désirée diminue comme dans le cas du Web qui représente la source d'information numéro un dans le monde. Bien que les recherches évoluées dans le domaine pluridisciplinaire de la Recherche d'Information sur le Web (RIW) aient apporté une grande rénovation aux techniques et modèles sous-jacents, la RIW incite constamment de nombreuses recherches développées conjointement avec la croissance incessante d'informations diverses et hétérogènes.

Le présent chapitre de la partie état de l'art commence par la définition des concepts de base de la RI et la description des étapes majeures du processus de recherche à savoir l'indexation et l'interrogation. Un aperçu des modèles de RI existants est également exposé par la suite ainsi que les avantages et les limites de chacun. Ensuite, nous nous achevons le chapitre par la description du mécanisme d'évaluation des performances d'un Système de Recherche d'Information (SRI) ainsi que la présentation des différentes mesures d'évaluation de la pertinence.

2.2. CONCEPTS DE BASE DE LA RECHERCHE D'INFORMATION

2.2.1. Système de recherche d'information

Un système de recherche d'information (information retrieval system) est tout outil qui permet de retrouver à partir d'une collection de documents, l'information qui répond à un besoin utilisateur exprimé à l'aide d'une requête.

2.2.2. Document

Le document (document) constitue l'élément d'information de base dans un SRI. Il peut être un texte, une image, une vidéo, un son ou encore une combinaison des objets précédents, dans ce cas on parle des pages Web. L'ensemble de documents compris dans un SRI s'appelle la *collection de documents* ou *corpus*.

2.2.3. Requête

La requête (query) représente l'expression d'un besoin en information selon le formalisme d'interrogation d'un SRI d'où, on distingue dans la littérature quatre formalismes d'interrogations [BAZ05c]: par le biais de mots clés, en langage naturel, booléen ou graphique. Elle est considérée comme le médiateur entre l'utilisateur et l'information recherchée.

2.2.4. Pertinence

La pertinence (relevance) mesure le degré de ressemblance entre la requête et le document renvoyé en se référant aux deux concepts : bruit et silence. Tel que, le silence correspond aux documents pertinents qui n'apparaissent pas dans le résultat de la recherche, alors que le bruit correspond aux documents ramenés en réponse, mais qui ne sont pas pertinents par rapport à la question posée [LEL98]. Dans [SAR75, 96, 97, 07], Saracevic a traité le concept de pertinence avec une grande importance où il a défini dans [SAR97] cinq types de pertinence.

- Pertinence système ou algorithmique, qui définit la capacité du système à comparer entre documents et requêtes et à quel point il a réussi à retrouver les documents adéquats à cette requête;

- Pertinence thématique, elle dépend du sujet exprimé dans la requête et celui des documents retrouvés;
- Pertinence cognitive, qui est la correspondance entre le besoin d'information de l'utilisateur, l'état de ses connaissances, et les documents retrouvés; elle est déduite à partir du degré d'accord cognitif entre le niveau de compréhension de l'utilisateur et l'information, de son informativité, de sa fraîcheur, de sa qualité et des préférences de l'utilisateur;
- Pertinence situationnelle ou contextuelle, elle représente la relation entre le problème à résoudre et les documents retrouvés. Elle est déterminée en fonction de l'utilité de ces documents et leur adéquation au but de l'utilisateur;
- Pertinence affective ou motivationnelle, qui est déduite à partir du degré de satisfaction de l'utilisateur vis-à-vis de l'information obtenue. Elle est définie autour de plusieurs critères d'évaluation sont : les intentions, les buts, les motivations et les gouts de l'utilisateur.

2.3. PROCESSUS DE RECHERCHE D'INFORMATION

L'accès à l'information paraît à l'utilisateur une tâche simple pouvant être récapitulée en quelques clics, tandis que derrière cette simplicité se cache un processus sophistiqué. En fait le processus de RI comporte deux grandes étapes sont : l'indexation et l'interrogation. La figure (2.1) illustre le processus en U de recherche d'information proposé par [BEL92].

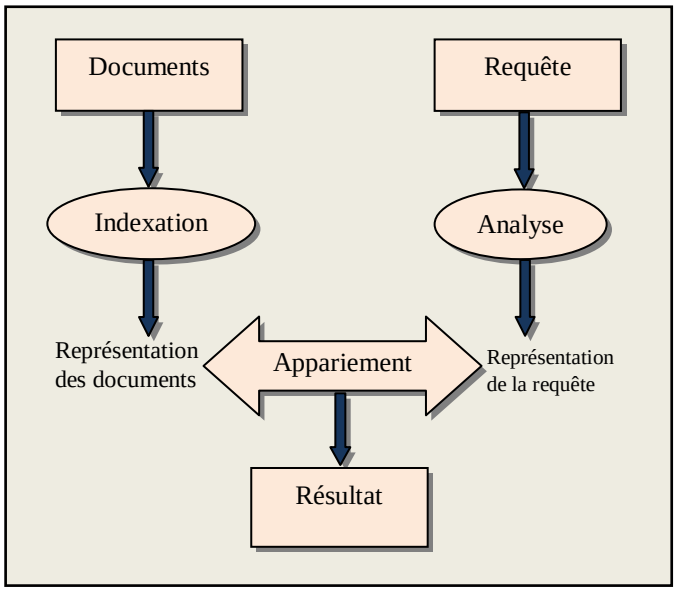

Figure 2.1. Processus en U de la recherche d'information [BEL92]

2.3.1. L'indexation

L'indexation représente une étape cruciale dans le processus global de la RI. Son objectif est la représentation des ressources (requêtes et documents) dans un format exploitable par un SRI dans le but de construire ce qu'on appelle index. La première étape du processus d'indexation est l'analyse de la ressource afin d'en extraire les caractéristiques les plus importantes et les structurer par la suite dans l'un des modèles de recherche d'information discutés dans la section (2.4). L'indexation des documents du corpus a pour objectif d'éviter au système de les analyser à chaque interrogation, en effet l'index créé permet d'établir le lien vers les documents indexés à travers les mots-clés représentatifs de leurs contenus. Salton [SAL83a] distingue trois types d'indexation: l'indexation manuelle, semi-automatique et automatique.

2.3.1.1. L'indexation manuelle

Ce type d'indexation est confié par des personnes qui doivent lire le contenu de la ressource et choisir les termes qui la décrivent en se basant sur un vocabulaire contrôlé (thesaurus[1] ou ontologie[2]). La mise en œuvre de l'indexation manuelle par des

[1] *Un thésaurus est un vocabulaire d'un langage d'indexation contrôlé, organisé formellement de façon à expliciter les relations à priori entre les notions. [ISO86]*
[2] *Une ontologie est une spécification formelle explicite d'une conceptualisation partagée [GRU93]*

spécialistes ayant une bonne connaissance du domaine donne de bons résultats. Cependant, ceci est achevé au détriment du temps de réalisation, de l'effort requis par un nombre important de personnes, et de l'objectivité souvent dominée par le caractère subjectif de l'être humain.

2.3.1.2. L'indexation semi-automatique

Elle consiste en un premier temps à indexer automatiquement les documents en s'appuyant sur un vocabulaire contrôlé comme un thésaurus ou n'importe quelle base terminologique. Ce type d'indexation requiert le contrôle manuel du processus par un spécialiste du domaine afin de valider le résultat obtenu et pour établir des relations sémantiques entre les termes d'indexation.

2.3.1.3. L'indexation automatique

Elle est assurée automatiquement par des programmes spécifiquement conçus pour choisir des termes d'indexation et leur affecter des valeurs reflétant leur pouvoir discriminatif. Le processus d'indexation se résume en quatre étapes, à savoir:

1) L'analyse lexicale, qui permet de découper le texte du document en unités élémentaires séparées par le caractère espace et les signes de ponctuation afin de produire une liste de mots.
2) L'élimination des mots vides de sens à partir de la liste de mots résultants de l'étape d'analyse lexicale, tels que les pronoms personnels, les articles et les prépositions, soit en utilisant une liste de mots vides appelée anti-dictionnaire[1], soit en tenant compte de leurs fréquences d'occurrences en se basant sur le principe de la loi de Zipf [ZIP49], qui stipule que les mots fréquemment apparus dans le texte sont vides de sens.
3) La lemmatisation, qui vise à éliminer les différences de formes non significatives entre les mots de la même famille et de les mettre dans leurs formes de base. En conséquence, la requête « sportif » conduira une recherche en utilisant le mot *sport* et doit permettre de retourner les documents contenant les mots *sport, sports, sportif, sportive, sportivement*,…etc.
4) La pondération des termes consiste à affecter à chacun un poids qui mesure son importance dans le document, de sorte d'attribuer un poids faible aux termes qui

[1] *Un anti-dictionnaire est une liste de termes ne portant aucun sens (mots vides).*

apparaissent dans la plupart des documents, ceci dit qu'un tel terme possède une faible aptitude discriminative. Les mesures de pondération les plus répondues dans la littérature sont :

- La loi de Zipf [ZIP49], elle considère que les termes des documents s'organisent suivant une loi inversement proportionnelle à leur fréquence d'apparition dans le corpus appelé *rang*. Elle est donnée par la formule (2.1).

$$\text{Fréquence} * \text{Rang} \approx \text{Constante} \tag{2.1}$$

- La conjecture de Luhn [LUH57], elle repose sur le principe d'éliminer les termes ayant un rang inférieur à un seuil minimal ou supérieur à un seuil maximal et de ne maintenir que les termes ayant un rang intermédiaire. L'idée sous-jacente est que les rangs faibles sont affectés souvent aux mots vides de sens marquant une fréquence d'apparition élevée dans le corpus alors que les rangs élevés sont attribués aux termes rarement apparus dans le corpus et dans les deux cas de figure, ces termes sont vus comme peu pertinents.
- le TF.IDF [SAL68; SPA79] est le schéma de pondération le plus utilisé, il combine deux facteurs de pondération sont : le facteur TF (Term Frequency), qui mesure la fréquence d'un terme dans le document où il apparaît [SAL68] et le facteur IDF (Inverse Document Frequency), qui mesure sa fréquence dans tout le corpus [SPA79]. Elle est donnée par l'équation (2.2) :

$$TF * IDF = log(1 + TF) * IDF \tag{2.2}$$

Où

$$TF(t_i, d_j) = \frac{n_{(t_i,d_j)}}{\sum_k \ n_{(t_k,d_j)}} \tag{2.3}$$

Tel que $TF(t_i, d_j)$ représente la fréquence du terme t_i dans le document d_j, $n_{(t_i,d_j)}$est le nombre d'occurrences de t_i dans $d_j$et $\sum_k n_{(t_k,d_j)}$ donne la somme des occurrences de chaque terme t_k dans d_j.

$$IDF(t_i, d_j) = log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (2.4)$$

Tel que $IDF(t_i, d_j)$ est l'inverse de la fréquence du terme t_i dans le document d_j, $|D|$ est le nombre total de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ calcule le nombre de documents où le terme t_i apparaît.

2.3.2. L'interrogation

Cette étape inclut un ensemble d'actions qui commence par la formulation du besoin d'information de l'utilisateur en une requête avant qu'elle soit analysée, indexée et représentée suivant le modèle adopté par le système. Il vient après la phase de recherche proprement dite qui consiste à faire l'appariement entre la requête soumise et la collection de documents maintenue sous forme d'index afin de repérer ceux qui peuvent répondre à la requête. Cette correspondance est faite par le biais des mesures de similarité telles que la mesure du cosinus [SAL83a], le produit scalaire [GRA44],...etc.

2.4. LES PRINCIPAUX MODÈLES DE RI

Dans la RI un modèle permet de créer une représentation interne d'un document ou d'une requête, afin de pouvoir créé une correspondance entre les deux. [BAE99] a formellement décrit un modèle de RI par le quadruplet $[D, Q, F, R\ (q_i, d_j)]$ qui consiste en un ensemble de documents D, d'une liste de requêtes Q, d'un schéma de représentation des documents et des requêtes ainsi que les relations qui leur ont associé F, et d'une fonction d'évaluation de la pertinence des résultats (q_i, d_j). Selon la classification présentée dans [BAE99], nous pouvons distinguer trois grandes familles de modèles dans la RI sont : les modèles ensemblistes, algébriques et probabilistes.

Dans cette section, nous essayerons de présenter les principaux modèles dans chaque famille.

2.4.1. Les modèles ensemblistes

Cette classe englobe deux types de modèles sont: le modèle booléen et le modèle flou. Le point commun entre ces deux modèles est qu'ils sont fondés sur la théorie des ensembles.

2.4.1.1. Modèle booléen (boolean model)

Proposé par [SAL71a], il représente le premier modèle utilisé en RI. Dans ce modèle les ressources d'information (documents ou requêtes) sont représentées sous forme d'une équation reliant des termes par le biais de connecteurs logiques. Les SRI basés sur le modèle booléen réalisent des recherches par mots-clés où les documents qui incluent l'ensemble ou un sous-ensemble des termes de la requête sont récupérés. Bien qu'il permet de faire une recherche très restrictive et obtenir ainsi, une information spécifique, le modèle booléen connaît plusieurs limites discutées dans les travaux de Cater et Kraft [CAT87], Cooper [COO88] et Salton et al., [SAL83a] dus principalement à l'absence d'une pondération des termes et en conséquence ils sont considérés d'une même importance. Outre que les résultats de recherche ne sont pas triés, leur nombre est difficile à contrôler. Des améliorations de ce modèle ont été proposées par [SAL83b] dans le modèle booléen étendu qui consistent en l'emploi des poids pour différencier entre les termes selon leur importance, ceci afin de pouvoir classé les résultats de recherche par ordre de pertinence croissant.

2.4.1.2. Modèle flou (fuzzy set model)

Il s'appuie sur la théorie des ensembles flous proposé par Zadeh en (1965) [ZAD65]. Dans ce modèle, le poids affecté à un terme reflète la mesure dans laquelle ce terme décrit le contenu du document où il apparaît. L'intégration de cette théorie dans la RI a permis de traiter l'ambigüité des requêtes, l'imprécision qui caractérise le processus d'indexation ainsi que la divergence de pertinence entre les documents résultats. L'inconvénient repéré par les SRI fondés sur le modèle flou est qu'ils ne permettent pas l'ordonnancement des résultats selon leur pertinence.

2.4.2. Les modèles algébriques

Ils sont fondés sur l'algèbre et ils englobent deux modèles capitaux sont le modèle vectoriel et le modèle d'indexation sémantique latente.

2.4.2.1. Le modèle vectoriel (vector model)

Introduit par Salton en 1971 [SAL71b], ce modèle propose de représenter les documents et les requêtes sous forme de vecteurs dans un espace multidimensionnel, où chaque dimension correspond à un terme d'indexation.

Chaque terme est pondéré selon son degré d'importance par rapport aux autres termes du même document ou de la même requête. Pour déterminer les documents pertinents vis-à-vis d'une requête donnée, le modèle vectoriel repose sur le calcul de la similarité entre le vecteur document et le vecteur requête. Plus les deux vecteurs sont proches, plus probable que le document soit pertinent pour la requête. Les mesures de similarité généralement employée pour cette fin sont :

- Le cosinus de l'angle qui sépare les deux vecteurs qui est obtenu par l'équation (2.5).

$$Cos\left(\vec{Q},\vec{D}\right) = \frac{\vec{Q}.\ \vec{D}}{\left|\vec{Q}\right| \times \left|\vec{D}\right|} = \frac{\sum_{i=1}^{n} w_{i,Q} \times w_{i,D}}{\left(\sum w_{i,Q}^2\right)^{1/2} \times \left(\sum w_{i,D}^2\right)^{1/2}} \quad (2.5)$$

  Ici la similarité entre le vecteur de la requête $\vec{Q}$ et le vecteur de document $\vec{D}$ est calculée à partir de la distance entre les termes de la requête $w_{i,Q}$ et ceux du document $w_{i,D}$.

- La mesure de Jaccard qui donne la similarité entre deux ensembles d'éléments obtenu par le quotient de l'intersection et l'union des éléments comparés. La similarité de Jaccard entre la requête Q et le document D peut être obtenue par l'équation (2.6).

$$J(Q,D)=\frac{|Q \cap D|}{|Q \cup D|} = \frac{\sum_{i=1}^{n} w_{i,Q} \times w_{i,D}}{\sum \mathrm{w_{i,Q}} + \sum \mathrm{w_{i,D}} - \sum \mathrm{w_{i,Q}} \times \mathrm{w_{i,D}}} \quad (2.6)$$

– Le produit scalaire, de la requête Q et du document D représentés par leur vecteur de termes $\vec{Q}$, $\vec{D}$ respectivement est défini par l'équation (2.7).

$$D_E(\vec{Q},\vec{D}) = \left(\sum_{i=1}^{n} \left|w_{i,Q} - w_{i,D}\right|^2\right)^{1/2} \quad (2.7)$$

Les avantages du modèle vectoriel consistent premièrement dans le tri des résultats renvoyés dans une recherche selon leur pertinence ainsi qu'à l'augmentation de la performance des résultats grâce à la pondération des termes. Néanmoins, le traitement de chaque terme indépendamment du reste représente sa majeure limite.

2.4.2.2. L'indexation sémantique latente (Latent Semantic indexing)

Ce modèle apporte une solution au problème d'indépendance de termes repéré dans le modèle vectoriel par la réduction de l'espace dimensionnel de représentation des documents en reliant les termes afin de traiter les "concepts" plutôt que les mots simples. Les recherches menées sur le modèle d'indexation sémantique latente (LSI) peuvent renvoyer des documents ne contenant aucun mot de la requête. Mais ils contiennent la même sémantique émergée à partir de la structure globale des documents et les relations entre les termes qu'ils incluent.

2.4.3. Les modèles probabilistes

Les modèles fondés sur les probabilités datent depuis (1976) et ils se sont basés sur le principe de classement des probabilités (*probability ranking principle)* proposé par Robertson en (1976) [ROB76] qui estime qu'il ya une incertitude dans la représentation de la requête et des documents de la collection. Les SRI fondés sur le modèle probabiliste classifient les documents en fonction de leur probabilité de pertinence pour la requête en deux classes sont : $P(R/d)$ qui mesure la probabilité que le document d inclut l'information pertinente pour la requête q et $P(\neg R/d)$ qui représente la probabilité que le document d n'inclut pas l'information pertinente pour q.

Contrairement au modèle booléen qui permet de faire une recherche restrictive en faisant un appariement exact, le modèle probabiliste emploie des mesures de similarité fondée sur une estimation probabiliste de pertinence.

$$P(R/d) = P(t_1/R) * P(t_1) \ldots . * P(t_n/R) * P(t_n) \quad (2.8)$$

$$P(\neg R/d) = P(t_1/\neg R) * P(t_1) \ldots . * P(t_n/\neg R) * P(t_n) \quad (2.9)$$

Tel que

$$P(t_1/R) = \frac{\mathrm{r_i}}{\mathrm{R}} \quad (2.10)$$

$$P(t_1/\neg R) = \frac{\mathrm{n_i} - \mathrm{r_i}}{\mathrm{N} - \mathrm{R}} \quad (2.11)$$

Où $\mathrm{r_i}$ étant le nombre de documents pertinents dans lesquels le terme t_i apparaît et R étant le nombre de documents pertinents pour la requête q, N représente le nombre total de documents dans la collection et $\mathrm{n_i}$ est le nombre total de documents dans lesquels le terme t_i apparaît.

2.5. ÉVALUATION D'UN SYSTÈME DE RECHERCHE D'INFORMATION

L'évaluation d'un SRI peut être faite suivant plusieurs critères tels que le temps de réponse, la facilité d'utilisation et notamment selon la qualité des résultats et la manière dont ils sont présentés. Dans cette section, nous abordons le protocole d'évaluation des SRI y compris les différentes compagnes et mesures d'évaluation qui existent dans la littérature.

2.5.1. Collection de test

Une collection d'évaluation ou de test contient une composition de documents et de requêtes ainsi qu'un protocole d'évaluation de pertinence tel que, l'ensemble de documents pertinents pour chaque requête est déterminé préalablement. Dans une collection de test les documents et les requêtes sont représentés sous forme d'index.

2.5.2. Compagne d'évaluation

Elle consiste en un système d'évaluation basé sur les collections de test. Il existe plusieurs compagnes d'évaluation qui ont marqué l'historique de la RI [SAN10]. Fondées sur le projet Cranfield [CLE67], TREC[1], CLEF[2] et NTCIR[3] sont considérées comme étant les compagnes d'évaluation les plus expérimentées dans la littérature [MAN08].

2.5.2.1. La compagne TREC

La compagne *Text REtrieval Conference (TREC)* organise des conférences annuelles pour évaluer les différentes méthodes et techniques évoluées dans la RI sur des collections de test volumineuses. Elle date depuis (1992) où ses tout premiers thèmes traités étaient le routage et la recherche ad hoc. Par la suite il y a eu, le filtrage de l'information, la RI non anglais, la question-réponse, la RI multimédia, la recherche d'information sur le Web RIW (Web track) ainsi que la recherche d'information contextuelle à travers les workshops HARD track et Contextual Suggestion Track.

2.5.2.2. La compagne CLEF

La compagne *Cross-language evaluation forum (CLEF*) date depuis l'an (2000), elle est développée pour des tâches spécifiques telles que le multilingues (Cross-Language) et la recherche inter-langues. Ensuite, ses activités ont été élargies pour inclure d'autres tâches comme la recherche d'information sur le Web, la recherche d'information géographique (RIG), la recherche vidéo et image et récemment la tâche d'évaluation de la recherche XML en (2012). CLEF offre des collections de test pour les langues européennes.

2.5.2.3. La compagne NTCIR

La compagne NII Test Collection for IR Systems (NTCIR) est dédiée aux technologies linguistiques spécifiques pour les langues asiatiques et les recherches inter-langues entre ces langues ainsi que l'anglais. Elle fournit des collections de test à grande échelle réutilisables pour les expérimentations en plus d'une infrastructure d'évaluation

[1] *http://trec.nist.gov/*
[2] *http://www.clef-initiative.eu//*
[3] *http://research.nii.ac.jp/ntcir/index-en.html*

commune permettant des comparaisons inter-systèmes. Cette compagne organise une série d'ateliers d'évaluation depuis (1997) dans les thèmes de question/réponse, le résumé de texte, la recherche d'information chinois et japonais, la recherche d'information sur le web, le multilingue…, etc.

2.5.3. Mesures d'évaluation

Traditionnellement, la pertinence des résultats d'une recherche se mesure en comparant les réponses du système avec les réponses idéales que l'utilisateur s'attend à recevoir sur la base de certaines mesures. L'étude présentée par [BAC10] définit plus de 20 mesures d'évaluation. Dans ce qui suit, nous présentons les mesures les plus utilisées.

2.5.3.1. Rappel

Le rappel mesure la capacité du système à retrouver tous les documents pertinents à une requête. Il est donné par le ratio entre le nombre de documents pertinents retrouvés p_t et le nombre de documents pertinents présents dans le corpus P, voir l'équation (2.12). Il mesure donc le silence, et plus le rappel est proche de 100%, meilleure est la réponse du SRI [LEL98]. Le silence peut être calculé par l'équation (2.13).

$$Rappel = \frac{p_t}{P} \tag{2.12}$$

$$Silence = 1 - Rappel \tag{2.13}$$

2.5.3.2. Précision

La précision mesure la capacité du système à ne pas classer les documents non pertinents à une requête. Elle est donnée par le ratio entre le nombre de documents pertinents trouvés p_t et le nombre total de documents trouvés D_t, voir l'équation (2.14). Elle mesure donc le bruit, et plus la précision est proche de 100%, meilleure est la réponse du SRI [LEL98]. Le bruit peut donc être calculé par l'équation (2.15).

$$Précision = \frac{p_t}{D_t} \tag{2.14}$$

$$Bruit = 1 - Précision \tag{2.15}$$

- Exemple applicatif:

Supposons que nous obtenons 40 réponses à une requête sur un corpus qui comporte 100 documents. Après examen des résultats nous apercevons que 10 documents ne sont pas pertinents. Nous savons par ailleurs que dans le corpus, 60 documents sont pertinents. Ainsi, nous pouvons calculer les deux mesures d'évaluation comme suit :

- Le rappel = (40-10)/60, soit 50%.
- La précision = (40-10)/40, soit 75%.

Sauf le quart des documents retournés ne sont pas pertinents c.à.d. le bruit est égal à 25% ce qui explique le taux de précision de 75 % qui peut être considéré élevé par rapport au rappel qui est égal à 50% et qui indique que sauf la moitié de documents pertinents ont été classé, ainsi le silence égal aussi 50%, tel que le nombre de documents pertinents non classés est égale à 30 parmi un total de 60 documents pertinents. Dans le cas d'un système idéal, le taux de précision et le taux de rappel sont égaux, c'est-à-dire que, tous les documents pertinents sont classés et aucun document non pertinent n'est sélectionné.

2.5.3.3. Autres mesures

En plus de la précision et du rappel, d'autres mesures sont utilisées pour évaluer la qualité d'une recherche établie par un SRI comme la précision à X documents, tandis que certaines autres sont formulées à partir de leurs combinaisons telles que la F-mesure, la moyenne harmonique et la précision moyenne.

1) La précision à X documents qui peut être calculée par l'équation (2.16) et qui représente le nombre de documents pertinents p_t sur les X premiers sélectionnés (X peut prendre différentes valeurs telle que 5; 10; 15; 20;….; 1000).

$$P@X = \frac{p_t}{X} \tag{2.16}$$

2) La F-mesure *(F-score)* qui est calculée par l'équation (2.17) :

$$F_\beta = \frac{(\beta^2 + 1) \times Precision \times Rappel}{\beta^2 \times Precision + Rappel} \quad (2.17)$$

Tel que β prend des valeurs réelles positives traduisant l'importance relative du rappel et de la précision.

3) La moyenne harmonique du rappel et de précision qui est un cas particulier de la $Fscore$ et elle est définit quand β prend la valeur 1 par l'équation (2.18).

$$F = \frac{Precision \times Rappel}{Precision + Rappel} \quad (2.18)$$

4) La précision moyenne (*ou MAP pour Median Average Precision*), elle est donnée par l'équation (2.19) qui représente la moyenne des précisions calculées pour chaque document pertinent dans la liste ordonnée du résultat au rang de ce document.

$$P(rp) = \sum_{i=1}^{N_q} \frac{P_i(rp)}{N_q} \quad (2.19)$$

Où $N_q$est le nombre de requêtes et $P_i(rp)$ représente la précision de la i^{ème} requête au niveau de rappel rp .

2.6. CONCLUSION

Dans ce chapitre, nous avons essentiellement présenté d'une manière générale les notions de base de la recherche d'information, mettant l'accent sur le processus de recherche et les différents modèles de la RI existant dans la littérature en essayant de donner brièvement les avantages et les limites de chacun. Nous avons conclu le

chapitre par la présentation des compagnes d'évaluation les plus connues dans le domaine de la RI ainsi que les mesures d'évaluation d'un SRI.

La recherche d'information contextuelle (RIC) est l'un des principaux axes traités dans notre thèse qui consiste à intégrer le contexte dans le processus de la RI. Dans le prochain chapitre, nous présenterons un aperçu sur les différentes approches de la RIC.

CHAPITRE 3: LE CONTEXTE ET LA RECHERCHE D'INFORMATION

3.1. INTRODUCTION

Les recherches menées par [BEL92] ont montré que le besoin en information d'un utilisateur est lié au moment où la recherche est effectuée, en d'autres mots, le besoin en information est influé par le contexte de la recherche. Motivés par cette démonstration, beaucoup de travaux et de conférences (IIiX[1], TREC[2], IRIX[3]) ont émergé depuis 1990 autours de l'exploitation du contexte dans le but d'améliorer l'accès à l'information pertinente [ABO97; RUT03; FON05; VOO06; ZHA06; SON07; KAN08; WAN09; LV10 ; SAI11; BOU11,12a,13a,b; BOU12b]. Ces travaux se focalisaient sur une ou plusieurs dimensions contextuelles comme la tâche ou l'activité de l'utilisateur, le temps, la localisation, la démographie,…etc.

Dans ce chapitre, nous présenterons les différentes dimensions de contexte ainsi que les travaux leaders dans l'axe de recherche sous-jacente appelé la recherche d'information contextuelle qui a apporté beaucoup de rénovations au domaine de la RI. Selon [ALL03], la RI contextuelle combine les techniques de recherche et les connaissances sur la requête et le contexte de l'utilisateur dans un cadre unique afin de fournir la réponse la plus appropriée pour ses besoins d'information.

3.2. NOTION DE CONTEXTE

Le contexte est un concept qui a été amplement étudié et défini en informatique [RYA97; DAV98] et en particulier en RI [CRE07; KOF03 ; ING05]. Il s'avère convenable de présenter quelques définitions de cette notion parmi tant d'autres pouvant

[1] *http://iiix2010.org/*
[2] *http://trec.nist.gov/*
[3] *http://ir.dcs.gla.ac.uk/context/*

être retrouvées dans le travail intitulé « Understanding Context Before Using It » de Bazire et Brézillon [BAZ05a].

Selon [ABO97], le contexte est toute information qui peut être utilisée pour caractériser la situation d'une entité. Telle que l'entité peut être une personne, un lieu ou un objet qui est considéré comme pertinent pour l'interaction entre un utilisateur et une application, y compris l'utilisateur et les applications elles-mêmes. Dans leur article [MYL08], Mylonas et al., ont spécifié que "le contexte peut être assimilé à tous les facteurs qui peuvent décrire les intentions de l'utilisateur et les perceptions de son entourage".

Tandis que Dourish [DOU04], le définit selon deux points de vue: en premier temps, en tant qu'un problème de représentation où il est considéré comme une forme d'information délimitée, stable et indépendante de l'activité qui se compose d'attributs implicites permettant de décrire l'utilisateur et l'environnement dans lequel les activités d'information se produisent. Dans une seconde perspective, le contexte est considéré comme un problème d'interaction où il découle de l'activité à partir de laquelle il ne peut pas être séparé. Une autre définition caractérisée par la brièveté et l'exhaustivité a été proposée par Goker et Myrhaug [GOK02] qui spécifient que "le contexte est la description des aspects d'une situation".

Dans [CRE07], Crestani et Ruthven ont aussi étudié ce concept et ils l'ont défini dans un cadre spécifique au domaine de la RI comme suit : *Le contexte influe sur tous les aspects de la recherche d'information. Le contexte des chercheurs influe sur la façon dont ils interagissent avec un SRI, quel type de réponse ils attendent d'un SRI et comment ils prennent des décisions sur les informations qu'ils reçoivent*.

Outre qu'il possède plusieurs définitions, plusieurs taxinomies ont été proposées pour mieux définir les différentes dimensions contextuelles et c'est ce que nous avons discuté dans la suite de ce chapitre.

3.3. TAXINOMIES DE CONTEXTE

Le grand intérêt envisagé vis-à-vis du contexte, ses applications et son utilisation dans la résolution des problèmes de la recherche d'information a permis l'émergence de maintes idées et propositions d'un modèle de contexte saisissant ses différents facteurs.

Nous détaillons dans cette section les taxinomies les plus adoptées dans la littérature à savoir : la taxinomie de Fuhr [FUH00], la taxinomie de Myrhaug et Göker [MYR03], la taxinomie de Ingerwersen et Jarvelin. [ING05] et la taxinomie de Tamine et al. [TAM09].

3.3.1. Taxinomie de Fuhr [FUH00]

La taxinomie proposée par Fuhr [FUH00] représentée par la figure (3.1), comporte trois dimensions principales sont : la dimension sociale, la dimension de l'application et la dimension temps.

1) Dimension sociale traite l'aspect selon lequel l'utilisateur est lié à son environnement social et elle lui traite selon trois cas de figure possibles soit en tant qu'individu vacant soit en tant qu'individu appartenant à une communauté ou à un groupe.

2) Dimension application détermine trois fins possibles peuvent être visées par une tâche utilisateur : soit la réalisation d'une application workflow, une recherche adhoc ou la résolution des problèmes.

3) Dimension temps définit le besoin de l'utilisateur selon le temps en : temps passé (batch), besoin instantané (interactif) et besoin persistant (personnalisation).

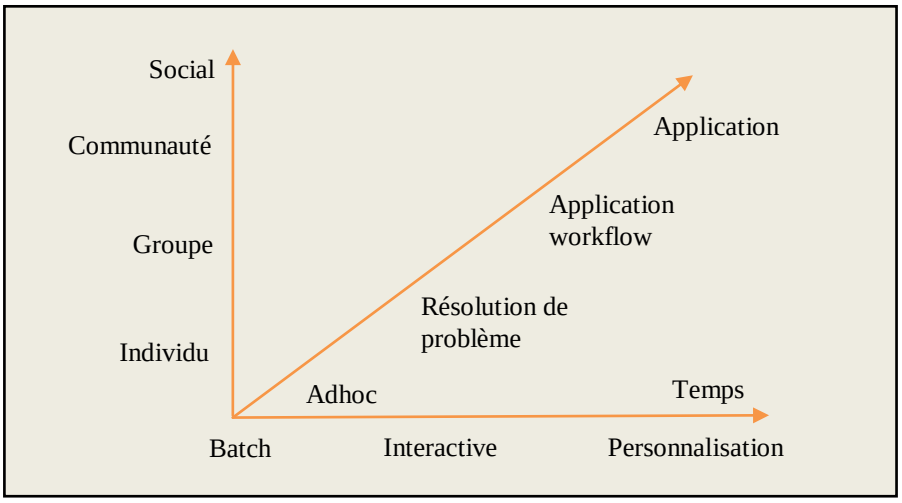

Figure 3.1. Contexte multidimentionnel de Fuhr [FUH00]

3.3.2. La taxinomie de Myrhaug et Göker [MYR03]

Myrhaug et Göker [MYR03], ont détaillé la taxinomie de contexte proposée dans le cadre du projet "Ambisense"[1] qui est composée de cinq éléments sont :

1) Contexte de la tâche (task context) qui peut être décrit par des objectifs explicites, des actions, des activités ou des événements. Comme il peut aussi inclure aussi les tâches d'autres personnes qui se trouvent dans le même contexte spatial.

2) Contexte environnemental (environnmental context) qui définit les différentes entités qui entourent l'utilisateur comme les objets, les services, le degré de la température, la lumière, l'humidité, le bruit et les personnes, ainsi que l'information accessible actuellement par l'utilisateur [SCH95].

3) Contexte personnel (personal context) qui inclut deux sous dimensions à savoir le contexte physiologique (physiological context) qui peut contenir des informations comme les impulsions, la tension artérielle, le poids, le niveau de glucose, le motif de la rétine, la couleur des cheveux…..etc, ainsi que le contexte mental (mental context) qui inclut des informations comme l'humeur, l'expertise, la colère, le stress, etc.….

4) Contexte social (social context) qui traite des informations au sujet des amis, des voisins, des collègues et des parents.., etc. En plus du rôle que l'utilisateur joue dans le contexte qui peut être aussi attribué à une zone géographique et une arène sociale comme par exemple le nageur X fait ses entrainements dans la piscine olympique de la ville Y.

5) Contexte spatio-temporel (spacio-temporal context), cette dimension du contexte décrit les aspects relatifs au temps et à localisation de l'utilisateur. Ses attributs peuvent être: l'heure, le lieu, la direction, la vitesse, l'entourage (des objets / bâtiments / terrain), les vêtements de l'utilisateur, …, etc.

La figure (3.2) montre les différentes dimensions de cette taxinomie.

[1] *Projet à base d'une modélisation multi-agent, qui vise à fournir de l'information contextuelle aux utilisateurs mobiles à propos des voyages d'affaires ou touristiques.*

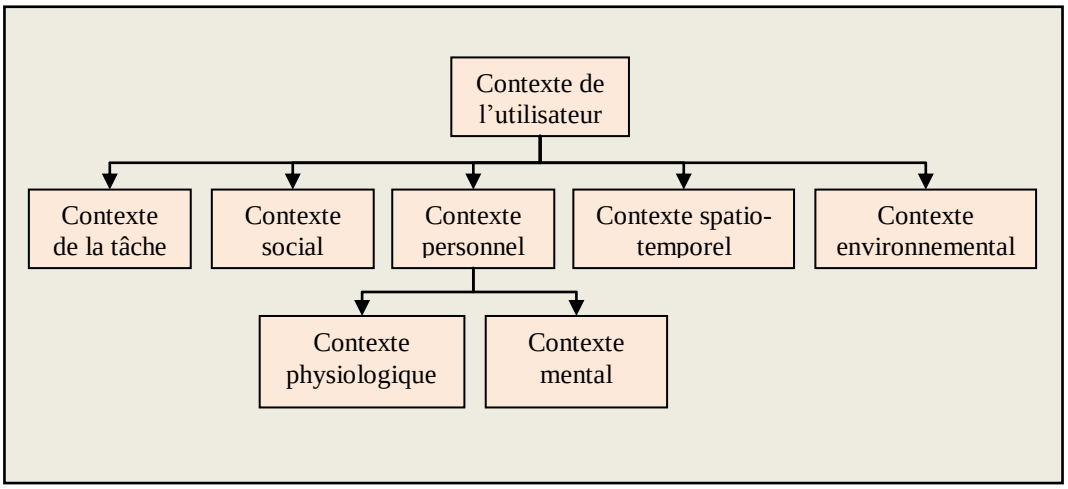

Figure 3.2. Taxinomie de Myrhaug et Göker [MYR03]

3.3.3. Taxinomie de Ingerwersen et Jarvelin [ING05]

La taxinomie proposée par Ingerwersen et Jarvelin [ING05] et représentée dans la figure (3.3) comporte six dimensions :

1) Contexte intra-objet, il fait référence au contenu de chaque document ainsi que sa structure (titre, résumé, paragraphes, conclusion.., etc.).

2) Contexte inter-objet correspond aux propriétés pouvant identifier les différents liens entre les documents appartenant au système comme les références et les citations.

3) Contexte d'interaction, concerne l'ensemble des interactions et des activités qui se produisent à l'intérieur de la session de recherche.

4) Contexte socio-systémique et organisationnel regroupe les aspects : personnel, social, environnemental et thématique, tels que le niveau de connaissances et d'expériences et les tâches perçues.

5) Contexte des infrastructures techno-culturelles et politico-économiques de la société correspond aux aspects globaux des événements réels comme par exemple la crise économique.

6) Contexte historique englobe les actions passées de tous les utilisateurs du système.

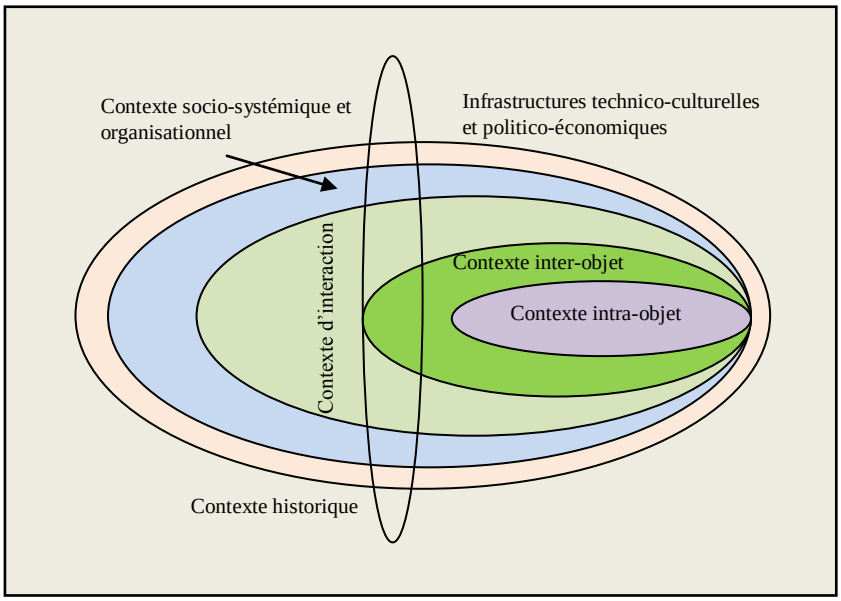

Figure 3.3. Taxinomie de Ingerwersen et Jarvelin [ING05]

3.3.4. La taxinomie de Tamine et al. [TAM09]

C'est une taxinomie qui a été proposée récemment dans [TAM09] qui introduit le contexte du dispositif d'accès à l'information (Device) en plus du contexte de document (document context), le contexte spatio-temporel, le contexte utilisateur (user context) et le contexte de la tâche comme dimensions essentielles. Cette taxinomie qui est représentée dans la figure (3.4) sera abordée avec plus de détails vu qu'elle englobe presque la totalité des dimensions existantes et, qui plus est, elle les présente de manière plus organisée.

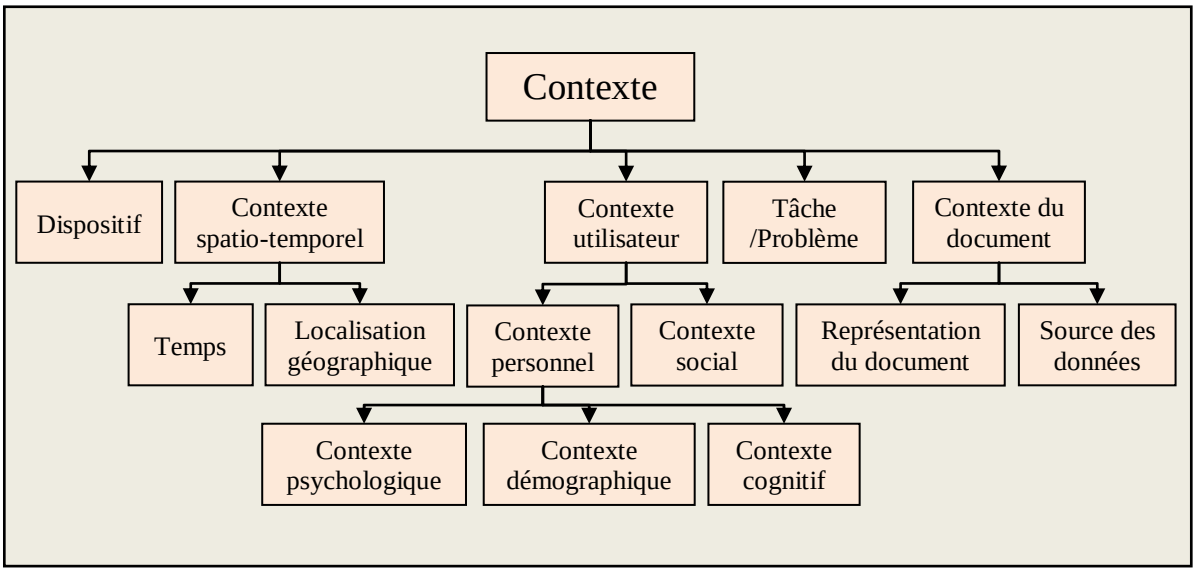

Figure 3.4. Taxinomie de Tamine et al. [TAM09]

1) Dispositif d'accès à l'information, c'est toute infrastructure matérielle (ordinateur, téléphone mobile, PDA, etc.) qui permet à l'utilisateur d'accéder à l'information. En effet, les caractéristiques du dispositif (type, taille, capacité du mémoire,…etc) doivent être prises en compte lors du processus de recherche en particulier dans le cas d'utilisateurs mobiles qui disposent de ressources mémoires limitées et de zones d'affichage réduites [GOK02].

2) Contexte spatio-temporel, selon [TAM09], cette dimension de contexte est relative à la situation géographique et au facteur temporel. En effet, les données et / ou les objets de requête changent leurs emplacements et ils ne sont pas valables dans le temps comme dans le cas de guide touristique.

3) Contexte de l'utilisateur, il comprend deux sous dimensions qui sont : le contexte personnel et le contexte social.

a. *Contexte personnel*, il représente un élément clé dans la Recherche d'Information Personnalisée (RIP), il est exploité afin de mettre en œuvre des techniques d'accès à l'information centrés utilisateurs. Il peut inclure des données démographiques, psychologiques et cognitives.

– Contexte démographique, cette sous-dimension est utilisée dans les approches de modélisation contextuelles explicites [AKT06; GAR05]. La prise en considération des informations telles que l'âge, le genre, la langue, …etc,

peuvent être utiles afin de présenter du contenu pertinent [FRI07a; HUP06]. Par exemple, connaitre l'âge du client d'un système de vente de livres en ligne, permet d'éviter de suggérer des ouvrages réservés aux adultes aux clients de jeunes âges.

- Contexte psychologique, il est similaire au contexte mental dans la taxinomie de Myrhaug et Göker [MYR03] et il concerne le facteur psychologique et les caractéristiques affectives de l'utilisateur (sentiments, humeur,…etc.) qui ont aussi un impact sur son comportement de recherche et son jugement de pertinence [BIL00; KIM08].

- Contexte cognitif, il porte sur le niveau d'expertise de l'utilisateur et ses centres d'intérêt à court terme [SHE05; JOA07; BOU12a, 13a, b] ou à long terme [TAM08a; LIU04].

b. *Contexte social*, cette dimension a été abordée dans toutes les taxinomies précédentes et elle définit l'appartenance possible de l'utilisateur à un groupe ou communauté à travers des relations d'amitié ou de voisinage. En Recherche d'Information Social (RIS) [KAR98; FID00; SMY06], être relié à un groupe de personnes qui partagent les mêmes préférences et les mêmes centres d'intérêt implique l'adaptation de la recherche aux préférences de la communauté et non pas à ceux de l'individu.

4) Contexte du document, appelé aussi contexte informationnel, cette dimension tente d'exploiter la poly-représentation des documents introduite par [ING94] pour améliorer la recherche. Selon le principe de poly-représentation de l'information, l'accès au contenu du document peut être fourni par un ensemble de critères qui constitue les deux sous dimensions distinguées par [TAM09], à savoir la dimension représentation qui concerne la forme, les couleurs, les éléments structurels, les citations, les métadonnées (nom de l'auteur, nom du journal, etc.) [TOM05] et la dimension sources de données qui concerne les caractéristiques de la source d'information et sa perception par les utilisateurs [XIE08].

5) Contexte de la tâche, ou de problème fait référence à l'objectif derrière l'activité accomplie comme dans le cas des requêtes de recherche qui peuvent avoir une intention transactionnelle, informationnelle ou navigationnelle [BRO02]. La tâche pourrait aussi se référer à une application dont la réalisation nécessite le recours à d'autres informations [SCH03].

3.4. SYNTHÈSE

Comme susmentionné, la taxinomie de Tamine et al., [TAM09] englobe presque la totalité des dimensions marquant leurs existences dans l'axe de la modélisation contextuelle. Bien que ces différentes taxinomies essayent de répondre aux besoins de modélisation et aux caractéristiques du web d'aujourd'hui, nous pouvons constater soigneusement qu'il ya une certaine imprécision dans la définition de certaines dimensions de contexte. Prenant la taxinomie de Fuhr [FUH00] qui classifie trois types de contexte distincts, nous remarquons que le facteur temps est inclus dans la dimension application, et ceci dans le but de traiter l'évolution des tâches au fil du temps, ceci dit que le facteur temporel est étudié de manière dépendante de la tâche. Tandis que, dans Myrhaug et Göker [MYR03], nous constatons que le contexte social et le contexte spatio-temporel partagent le facteur localisation. En effet, nous trouvons que le contexte social englobe le facteur localisation géographique qui est également traité au sein de la dimension spatio-temporelle. Davantage, l'attribut entourage (des objets / bâtiments / terrain) issu du contexte spatio-temporel est aussi présent comme attribut du contexte environnemental. Dans la taxinomie de Ingerwersen et Jarvelin [ING05], l'entité document est introduite et traitée au sein de deux dimensions contextuelles à savoir le contexte intra-objet, qui fait référence au contenu du document et sa structure (titre, résumé, paragraphes, conclusion.., etc.) ainsi que le contexte inter-objet traitant les liens entre les documents du système déduits à partir des références communes et des citations. Concernant le contexte socio-systémique et organisationnel, il englobe des sous-dimensions diverses disant hétérogènes tel que la dimension personnelle, la dimension environnementale et la dimension thématique, ce qui signifie que cette dimension traite à la fois les données personnelles de l'utilisateur (âge, poids, taille, etc.) son environnement (objet qui l'entoure, température,…... etc.) en plus de son niveau de connaissances et d'expériences et les tâches qu'il est en train d'accomplir. Tandis que, la taxinomie proposée par Tamine et al., [TAM09] réunit toutes les

dimensions discutées à part la dimension physiologique qui traite des données biologiques telles que les impulsions, la tension artérielle, le niveau de glucose, le motif de la rétine, la couleur des cheveux, etc…

3.5. TAXINOMIE PROPOSÉE

Le développement abondant du hardware et du software, ouvre des voies pour explorer de nouvelles dimensions contextuelles, de ce fait et à partir des taxinomies que nous avons étudiées, nous proposons dans ce qui suit notre taxinomie de contexte dans laquelle nous avons introduit toutes les dimensions discutées auparavant en essayant de définir la portée de chacune, la taxinomie proposée est représentée par la figure (3.5).

En fait, cette taxinomie se fonde principalement sur celle de Tamine et al., [TAM09] avec l'introduction du contexte environnemental et le contexte événementiel ainsi que le contexte physiologique comme sous dimension du contexte utilisateur.

- Contexte environnemental, il définit l'entourage de l'utilisateur y compris les objets, le degré de la température, la lumière, l'humidité, le bruit et les personnes, etc.….

- Contexte événementiel, il définit les événements (économiques, culturels, politiques, sportifs, …etc.) agissants un peu partout dans le monde comme les élections, les manifestations sportives, les désastres naturels, …etc.

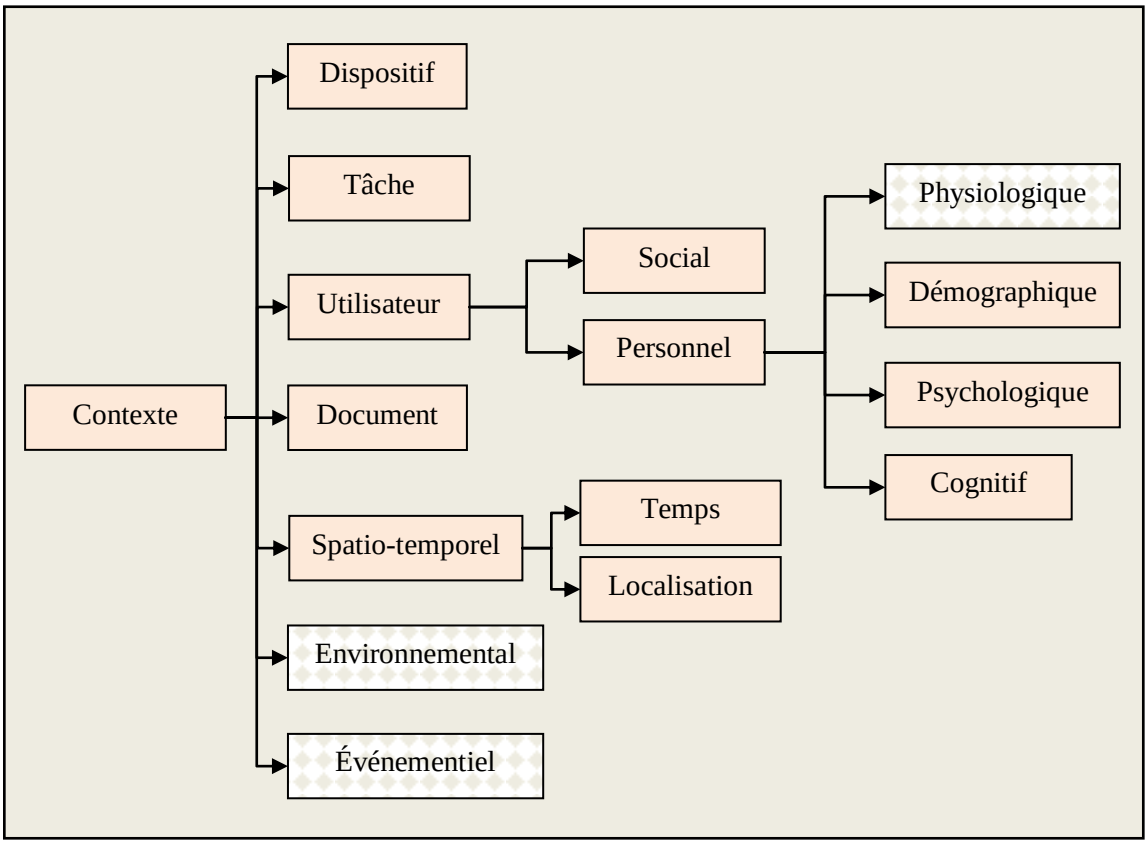

Figure 3.5. Taxinomie proposée

Nous présenterons dans le tableau (3.1) une synthèse des différentes dimensions contextuelles abordées ainsi que quelques travaux de référence les ayant utilisés.

T1 : Taxinomie de Fuhr [FUH00]

T2 : Taxinomie de Myrhaug et Göker [MYR03]

T3 : Taxinomie de Ingerwersen et Javelin [ING05]

T4 : Taxinomie de Tamine et al. [TAM09]

T5 : Taxinomie proposée

CONTEXTE			TAXINOMIE	TRAVAUX DE RÉFÉRENCE
Dispositif			T4 et T5	[CHI03; 08 ; GOK11 ; WAN14]
Spatio-temporel	Temps		T1, T2, T3, T4 et T5	[GAR05; ZHA06; PAS08; JIN11; ALO07; BIE06; BOU11, 13b, BOU12b]
	Localisation		T2, T3, T4 et T5	[SAI11; PAN05; BIE06; BOU12b; BIL08; HAT06]
Utilisateur	Personnel	Psychologique	T2, T4 et T5	[KIM08; BIL00]
		Démographique	T2, T4 et T5	[FRI07a; HUP06]
		Cognitif	T3, T4 et T5	[SHE05; JOA07; BOU12a; TAM08a; LIU04]
		Physiologique	T2 et T5	[COL11 ; ALH13]
	Social		T1, T2, T3, T4 et T5	[KAR98; FID00; SMY06]
Tâche			T1, T2, T3, T4 et T5	[SHE05; BOU12a,13a,b;ASF09, 12 ; FRE05]
Document			T3, T4 et T5	[TOM05, MEL08; XIE08; JEN05a,b; SAR04]
Environnemental			T2, T3 et T5	[SCH95; HAT07; LAN10].
Événementiel			T3 et T5	[BOU13a, 12a ; ALO13]

Tableau 3.1.Synthèse des différentes dimensions de contexte

3.5. MODÉLISATION CONTEXTUELLE

La modélisation ou profilage contextuelle de l'utilisateur est le processus de création et de mise à jour d'un modèle utilisateur (user profile) dont ses caractéristiques sont dérivées à partir des données extraites de son contexte. La modélisation comporte deux étapes intimement liées à savoir : l'acquisition de données, la construction et la représentation des modèles.

3.5.1. Sources de données

Les données de modélisations sont classifiables sous deux grandes catégories sont :

1) Données restreintes à la machine telles que, l'historique des clics, les logiciels utilisés, le mouvement des yeux... etc.

2) Données de navigations, qui comprennent l'historique de navigation et de recherche y compris le contenu des documents consultés et/ou imprimés et/ou sauvegardés, les liens explorés, les clics et le mouvement des yeux, les signets, les pages fréquemment visitées et les dernières pages visitées, le résumé des premières pages web retournés par un moteur de recherche..., etc.

3.5.2. Stratégies d'acquisition

Les informations recueillies pour la modélisation peuvent être collectées de manière explicite ou implicite.

3.5.2.1. Acquisition explicite

L'acquisition explicite s'effectue par la saisie de données nécessaires à la modélisation via le remplissage d'un formulaire, la réponse à des questionnaires ou encore via le jugement explicite des résultats renvoyés par un système d'accès à l'information. Ce retour informationnel est connu par la *réinjection de pertinence* (*relevance feedback*). [AKT06; GAR05] ont tous suivi une approche d'acquisition explicite de données assurant ainsi d'une modélisation contrôlée et précise. Néanmoins, en raison du temps et des efforts supplémentaires exigés de la part de l'être humain, en plus de l'incohérence des données au fil du temps, l'approche d'acquisition implicite s'impose comme alternative prévenant les limites de l'approche explicite.

3.5.2.2. Acquisition implicite

Lorsque le contexte et les préférences de l'utilisateur sont inférés à partir de son interaction avec un système d'accès à l'information en utilisant des modèles statistiques et des techniques d'apprentissage automatique [KEL03; KEL04; SHE05; ASF12; BOU12a, 13a, b; QIU06; SPE05] on parle donc de la modélisation implicite. Cette approche de modélisation est plus flexible et mieux adaptée pour faire face à d'énormes quantités de données. Un exemple des données explorées par cette méthode est le temps

de visite d'une page, le nombre de clics dessus, le mouvement de la barre de défilement, etc.

Afin d'améliorer la qualité des données recueillies et par conséquence la qualité des modèles créés, certains travaux combinent l'approche de modélisation explicite et implicite. Les résultats obtenus par Quiroga et Mostafa [QUI00] montrent que la combinaison des deux approches permet d'améliorer la pertinence des résultats retournés par un SRI. En fait, ils ont atteint une précision de 63% en procédant à une acquisition explicite seul, et 58% de précision via l'approche implicite seule. Néanmoins, par la combinaison de deux méthodes un environ de 68% de précision a été atteint. Cependant, White [WHI01] prouve qu'il n'y a pas de différence significative entre les modèles construits par le biais des deux approches.

3.5.3. Construction et représentation du modèle utilisateur

La construction du modèle ou profil utilisateur consiste à instancier sa représentation à partir de l'ensemble de données recueillies en utilisant généralement les outils d'apprentissage automatique tels que les algorithmes génétiques [YAN06], les réseaux de neurones artificiels [JAI05; TAK09; MOG11; BOU12a, 13a, b], les réseaux bayésiens [GAR07], etc. Les profils créés sont représentés selon l'un des types de représentation suivants [ZEM08] :

1) Représentation ensembliste, qui permet de représenter les centres d'intérêt de l'utilisateur sous forme de vecteur de mots-clés [PAZ96; DUM03] ou classe de mots-clés [MCG03] qui sont pondérés ensuite en utilisant la formule de TF.IDF [SAL68; SPA79].

2) Représentation sémantique, qui vise à lever l'ambigüité des termes qui décrient les centres d'intérêts et les préférences des utilisateurs en se basant sur des réseaux sémantiques [GEN03], des ontologies [BLA08 ; LIU04; SIE07] ou des réseaux sémantiques probabilistes [LIN05; WEN04].

3) Représentation multidimensionnelle, qui consiste à structurer le profil utilisateur selon plusieurs dimensions représentées suivant divers formalismes [BOU05], à titre d'exemple les dimensions domaine d'intérêt, données personnelles, données de sécurité, …etc.

3.6. L'ÉVALUATION EN RECHERCHE D'INFORMATION CONTEXTUELLE

Comme nous l'avons déjà vu dans le deuxième chapitre, les expérimentations et la validation des systèmes en recherche d'information s'effectuent par le biais des outils offerts par les compagnies d'évaluation. Les tests d'évaluation s'appuient sur des collections de documents, de requêtes et de jugements de pertinence fournis par des experts. L'avantage de ce type d'évaluation est qu'il permet de comparer les performances de plusieurs systèmes sur la même base. Toutefois, dans le cadre de la RI traitant l'utilisateur comme étant une composante de la recherche (RIS, RIC et RIP) cette approche a subi plusieurs critiques [HAR96] dues au fait qu'elle ne prend pas en considération ni le profil de l'utilisateur ni la situation de recherche outre qu'elle ne porte que sur l'évaluation de la pertinence thématique. Nous pouvons classer les méthodes d'évaluation contextuelles en méthodes basées sur les collections de test, méthodes basées sur la simulation du contexte et méthodes basées sur des contextes réels.

3.6.1. Méthodes basées sur les collections de test

La création des tâches « interactive track » [HAR95] et « hard track » [ALL05] dans les conférences TREC a ouvert la voie pour la prise en compte du composant utilisateur lors de l'évaluation des SRI. Ceci, à travers l'incorporation des métadonnées dans les documents et les requêtes d'évaluation comme la durée d'interaction, le niveau d'expertise. Les résultats obtenus n'étaient pas significatifs [SPA05 ; BEL08] en comparaison avec le jugement donné par les utilisateurs en temps réel [TUR06].

Les problèmes majeurs liés à l'évaluation contextuelle sont dus à la difficulté de créer des bases de test pouvant comprendre toutes les interactions possibles de l'utilisateur avec un SRI, ainsi que l'incapacité de couvrir les différentes dimensions contextuelles ce qui est quasiment difficile vu la variabilité des utilisateurs et la diversité des centres d'intérêt. En plus de la variation du besoin informationnel en fonction du contexte courant.

3.6.2. Méthodes basées sur la simulation du contexte

Les situations de contextes simulées ont été beaucoup employées lors de l'évaluation des SRI contextuels. Le principe de cette approche repose sur la définition de scénarios d'évaluation qui simulent des situations de recherche réelles à travers l'introduction des centres d'intérêt des utilisateurs ainsi que leurs interactions avec le SRI [BOU13a ; RUT01, 02; SIE07; RYE05]. Les mesures de rappel et de précision sont communément utilisées pour mesurer l'efficacité du modèle contextuel. L'avantage de l'évaluation basée sur la simulation du contexte c'est qu'elle n'implique pas des utilisateurs réels en plus qu'elle n'est pas couteuse en temps, en plus qu'elle permet d'effectuer une évaluation comparative [WHI05].

3.6.3. Méthodes basées sur des contextes réels

Ces méthodes tentent d'incorporer des utilisateurs réels dans le processus d'évaluation ce qui permet d'améliorer les expérimentations en matières d'efficacité et d'exactitude. En effet, cette approche d'évaluation vise la mise de l'utilisateur dans des situations de recherche réelles [LIU04; SPE05; SHE05]. Il existe deux types d'évaluation basée sur des contextes réels en RI. Le premier type d'évaluation consiste à utiliser une interface de recherche qui est branchée à une collection de test où l'utilisateur formule des requêtes liées à l'un des sujets prédéfinis par la collection. Ensuite, les résultats des requêtes sont comparés aux résultats pertinents prédéfinis par la collection [SHE05]. Le second type se fonde sur les interfaces de recherche (comme l'API de Google) qui offre à l'utilisateur la possibilité d'effectuer une recherche habituelle [CHA07; LIU04 ; BOU12a]. Les limites majeures que cette approche connaît sont principalement le cout élevé en temps, en plus qu'elle ne permet pas de faire une évaluation comparative en raison de la difficulté de séparer les éléments contextuels de l'utilisateur (centres d'intérêt, expertises, familiarité avec le sujet de la requête, etc.) de ceux du modèle, ainsi qu'à l'absence des jugements de pertinence préalablement associés aux documents résultants de l'évaluation d'une requête. Davantage, les résultats des expériences ne sont pas reproductibles.

En outre, les mesures classiques adoptées dans l'évaluation de la RIC à savoir le rappel, la précision et la précision pour les X premiers documents retournés [DIN07; SHE05],

la mesure DGC (Discouted Cumulative Gain) est largement utilisée dans ce contexte [CHA07; JAR02; SPE05].

- La mesure DGC (Discouted Cumulative Gain) : le principe du DCG est la sanction des documents ayant une valeur de pertinence élevée et qui ne sont pas bien classés dans le résultat de sorte que le degré de pertinence du document se diminue de manière proportionnelle à sa position dans le résultat. Le DCG d'un document ayant la position p est défini par l'équation (3.1).

$$DCG_p = \sum_{i=1}^{p} \frac{2rel_i - 1}{\log_2(i+1)} \tag{3.1}$$

  Ici, rel_i représente la pertinence graduée du resultat à la position i.

3.7. CONCLUSION

Il existe un nombre important de travaux et de conférences qui ont été consacrés à l'utilisation du contexte dans l'amélioration de l'accès à l'information sur le Web. Une grande partie de ces recherches portent sur la modélisation de l'utilisateur ou profilage en utilisant une grande variété de sources d'informations contextuelles locales telles que l'historique des clics et les logiciels utilisés sur la machine, en plus des sources d'informations résultantes de l'interaction de l'utilisateur avec le web telles que les données de navigations y compris l'historique de navigation, les signets et les pages fréquemment visitées.

L'ensemble de données utilisé pour la modélisation utilisateur dans une application donnée dépend de l'application elle-même et de l'objectif visé, ce qui explique l'importance de l'étape de collecte de données dans le processus global de modélisation et son aspect déterminant pour la qualité des profils créés.

Bien que l'évaluation contextuelle connaisse maints problèmes liés principalement à la variabilité des utilisateurs et la diversité des centres d'intérêt en plus de la variation du besoin informationnel en fonction de plusieurs facteurs contextuels, les méthodes basées sur la simulation du contexte et suite au fait qu'elles n'impliquent pas l'intervention du

facteur utilisateur dans le processus d'évaluation ainsi qu'elles ne possèdent pas un cout élevé, s'avèrent intéressantes pour avoir une évaluation rapide et fiable de la RIC.

CHAPITRE 4 : UN APERÇU DES DIFFÉRENTES TECHNIQUES DE REFORMULATION DE REQUÊTES

4.1 INTRODUCTION

Lorsque l'utilisateur effectue des recherches sur le Web, il emploi des termes qui les jugent expressifs de son besoin en information. Un bon choix de termes en matière d'expressivité et d'adéquatement permet de rapprocher l'utilisateur de l'information désirée. Par ailleurs, la recherche est conduite par le biais d'une requête imprécise ayant un résultat difficilement déterminé, ceci est dû à la manifestation du phénomène de polysémie qui caractérise les termes du langage humain. Une telle requête est appelée *requête ambigüe*.

L'une des principales causes de l'ambigüité des requêtes est leur brièveté. En effet, les requêtes d'un seul mot ou de deux mots sont souvent indécises et elles peuvent se référer à des domaines d'intérêt différents. Il est à noter que les requêtes courtes (de moins de 3 mots-clés) constituent selon la société de mesure d'audience américaine [ONE11], plus de 50% du nombre total de requêtes soumises aux moteurs de recherche.

Afin de résoudre ce problème, la reformulation de requête a été proposée pour aider l'utilisateur à accéder à l'information pertinente plus efficacement, ceci moyennant la modification de la requête initiale par l'ajout de termes plus expressifs et/ou par la réévaluation de leur poids afin de lever l'ambiguïté de la requête et augmenter ainsi la chance de satisfaction de l'utilisateur vis-à-vis de l'information retournée.

Il existe une vaste littérature liée à la reformulation de requêtes [RUT03; FON05; VOO06; SON07; KAN08; WAN09; LV10]. Dans ce chapitre, nous mettons l'accent précisément sur les différentes approches d'expansion que nous avons essayé de les classer suivant plusieurs critères.

4.2. DÉFINITIONS

4.2.1. Ambigüité

C'est un phénomène traditionnel dans le langage naturel dû au fait que la majorité des mots peuvent exprimer plusieurs sens selon leurs contextes d'apparition ce qui est connu sous le nom de polysémie de mots. En RI, une requête ambigüe peut se référer à des domaines d'intérêt différents. Par exemple le mot «sky», signifie pour le dictionnaire Oxford [OXF13] : « the region of the atmosphere and outer space seen from the earth». Cependant, la requête à mot unique «sky» signifie pour Google, le nom de plusieurs chaînes de télévision, une chaîne de radio, un site de nouvelles de sport, un service de Google Earth, un réseau social, une équipe de cyclisme et beaucoup d'autres significations différentes.

4.2.2. Reformulation de la requête

La reformulation de requête est une technique qui permet de générer une nouvelle requête en modifiant la requête initiale par l'ajout de nouveaux termes [BAE99; RUT03; FON05; BIA09] et/ou par la repondération de leur poids [ROB00; TAM03]. Le but de la reformulation est de réaliser une adaptation de la requête aux besoins en information de l'utilisateur. Lorsque la reformulation est limitée seulement à l'ajout de nouveaux termes on parle donc de l'*expansion de requête*.

4.3. CLASSIFICATIONS DES APPROCHES D'EXPANSION DE REQUÊTES

4.3.1. Selon le degré d'implication de l'utilisateur

Selon le degré d'implication de l'utilisateur dans le processus d'expansion de requêtes, nous distinguons deux approches d'expansion, une approche basée sur un processus interactif et une autre basée sur un processus automatique.

4.3.1.1. Approche interactive

Suivant cette approche [EFT96; BAE99; FON05; RUT03], les mots-clés d'expansion sont choisis par l'utilisateur parmi une liste de mots-clés suggérés par le système.

Bien que, l'approche interactive donne de meilleurs résultats en comparaison avec l'approche automatique [KAN08; RUT03], un degré d'expertise est nécessaire pour l'atteinte de ce résultat [RUT03]. L'autre avantage de cette approche est qu'elle donne à l'utilisateur plus de contrôle sur le traitement de sa propre requête, ce qui est un aspect manquant dans l'approche automatique.

4.3.1.2. Approche automatique

Suivant cette approche d'expansion, la requête est étendue automatiquement sans l'intervention directe de l'utilisateur. Ceci, par l'ajout de termes issus des ressources linguistiques existantes [GON06; SON07] ou bien des ressources construites à partir des collections de documents comme Wikipedia [CRO92; PAR07; MIL07]. Le retour de pertinence des utilisateurs sur les résultats renvoyés par un SRI constitue également une source importante pour l'expansion automatique de requêtes [BOU13b; BUC94a; BIA09].

4.3.2. Selon la source des termes d'expansion

4.3.2.1. Méthode basée sur la réinjection de pertinence

Selon Rochio [ROC71], la *réinjection* de pertinence est un processus où le SRI fournit à l'utilisateur un ensemble de documents comme réponse à sa requête. L'utilisateur sélectionne ensuite les documents qui correspondent le mieux à son besoin en information. Ensuite, et sur la base de cet ensemble d'interactions, le SRI peut effectuer une recherche de documents plus raffinée afin de fournir plus de résultats pertinents.

Le feedback permet d'évaluer la pertinence de chaque page consultée par l'utilisateur afin de permettre par la suite l'extraction des termes d'expansion depuis les pages qui sont évaluées comme pertinentes [BUC94b; BIA09; BOU13b].

L'inconvénient majeur de cette méthode est que la qualité de reformulation dépend fortement de l'aptitude des utilisateurs à donner des jugements corrects de la pertinence des documents.

4.3.2.2. Méthode basée sur le pseudo réinjection de pertinence

La méthode de pseudo réinjection de pertinence (pseudo-relevance feedback) considère que les k premiers documents récupérés par une requête comme pertinents. Ensuite ces

documents sont utilisés comme un feedback de pertinence ce qui rend le processus de recherche plus rapide, car il n'y a pas d'interaction humaine au cours du feedback. Cependant, cette méthode est moins précise, car il n'est pas garanti que les documents de feedback sont tous pertinents [BUC94b; VOO06; LV10]. Pour surmonter ce problème, Belkin [BEL00] a proposé l'évaluation des documents initiaux par l'utilisateur avant de procéder à une extraction probabiliste de termes d'expansion des documents pertinents.

Parmi les paramètres utilisés pour déterminer la pertinence d'un document donné nous citons : le clic de données, le mouvement de la barre de défilement, le temps de visite, les commandes (imprimer, copier, coller et sauvegarder), …etc.

4.3.2.3. Méthode basée sur les ressources sémantiques

Elle se fond sur la sémantique des mots dans le but d'identifier leur sens dans leur contexte computationnel [NAV09]. Les ressources terminologiques qui sont généralement employées sont les thesaurus [PAR07; AIR04] et les ontologies [BHO07; NAV03; SON07], dont Wordnet représente la ressource la plus utilisée [VOO94; HAR01; LIU08; SON07]. En effet, cette méthode d'expansion nécessite le recours aux techniques de désambigüisation de sens des mots (Word sens désambigüisation) ou WSD pour identifier les termes adéquats à l'expansion.

Le travail de Frias-Martinez et al., [FRI07b] consiste à étendre les requêtes par des mots simples intervenant dans la construction de certains mots composés de la langue suédoise en se basant sur le thésaurus MeSH. Alors que Plovnick et Zeng [PLO04] ont exploité le thesaurus UMLS pour substituer certains mots d'une requête par leurs synonymes.

Dans [VOO94], l'auteur propose d'enrichir la requête par des synonymes extrait depuis Wordnet. Tandis que dans le travail de Navigli et Velardi [NAV05], d'autres caractéristiques ont été exploitées comme les hyperonymes, et les hyponymes des mots. Le problème majeur avec ces approches est lié à l'aspect trop général et vaste de WordNet.

Ainsi, les approches basées ontologies de domaines se présentent pour surmener le problème de généralité qui survient lors de l'utilisation de Wordnet. En effet

l'utilisation d'une ontologie spécifique à un domaine particulier permet d'avoir le sens exact des mots de la requête à partir des relations sémantiques utilisées pour concevoir l'ontologie. De ce fait, les mots qui vont servir à l'expansion de la requête sont très spécifiques. Dans ce contexte, Fu et al., [FU04] ont exploité les relations spatiales (ex. "au nord de", "proche de") obtenues depuis des ontologies géographiques, ceci pour étendre des requêtes dans le cadre de la RIG. Récemment, Bhogal et Macfarlane [BHO13] ont proposé de développer l'utilisation des techniques de réinjection et de pseudo-réinjection de pertinence en utilisant des informations extraites depuis une ontologie de news, leur contribution a été appliquée sur le modèle de recherche d'information probabiliste. Tandis que Segura et al., [SEG14] ont employé l'ontologie GENE pour reformuler les requêtes dans le domaine biomédical. Toutefois, étant donné la grande spécificité de l'approche basée ontologie de domaine, leur application dans un contexte plus général comme le Web s'avère peu fructueuse.

Les limites majeures des méthodes d'expansion basées sur les ressources sémantiques sont d'abord la non-adéquation de la ressource terminologique à la collection de documents, en plus du besoin qu'elle soit mise à jour pour répondre à l'évolution de la langue, sans oublier l'intervention indispensable de l'expert humain.

4.3.3. Selon le principe de génération des termes d'expansion

Selon cette classification, nous pouvons groupées les approches d'expansion en trois groupes principales qui sont: l'approche linguistique, l'approche statistique et l'approche mixte.

4.3.3.1. L'approche linguistique

Elle s'intéresse à la découverte des relations syntaxiques, lexicales et sémantiques entre les mots en s'appuyant sur des ressources terminologiques telles que les thesaurus et les ontologies [VOO93, 94; BHO07; SEG14].

4.3.3.2. L'approche statistique

Elle se base sur le principe de génération des corrélations entre des paires de termes en exploitant la technique de cooccurrence de termes, et ceci dans un contexte global quand elle est appliquée à tout le corpus de documents du système ou local dans le cas où elle est appliquée aux premiers documents résultants de l'exécution de la requête

[Xu96; JON06; BOU13b]. Dans le travail de Kumar et Carterette [KUM13] et en prenant compte du contexte temporel, les termes fréquents dans les premiers tweets[1] retrouvés ont été employés pour étendre les requêtes de recherche exécutées sur le réseau social Twitter[2].

4.3.3.3. L'approche mixte

Cette approche combine les deux approches précédentes et elle se base sur l'analyse de la distribution des mots en bénéficiant d'une ressource terminologique. Dans les travaux de Liu et al. [LIU04] et de Fang [FAN08], qui ont utilisé Wordnet pour désambigüiser les termes de la requête en prenant compte de la distribution des mots dans le corpus étudié, il a été prouvé que l'approche mixte produise des résultats meilleur que l'approche statistique ou l'approche linguistique seules.

4.4. PROCESSUS D'EXPANSION

Il comprend deux étapes intimement liées sont : le prétraitement de données et la sélection des candidats termes.

4.4.1. Prétraitement de données

Cette étape est généralement indépendante de la requête de l'utilisateur, mais elle dépend du type de source de données. Elle consiste à transformer la source de données brutes utilisées pour l'expansion des requêtes dans un format qui peut être traité plus efficacement par la suite. Dans le reste de cette section nous présenterons pour chaque source de données employée dans l'expansion la méthode de prétraitement adoptée.

4.4.1.1. Collection de documents

Comme il est cité un peu plus haut, les informations contenues dans les documents les mieux classés présentés en réponse à la requête constituent une source d'expansion considérablement adoptée. Le traitement de ce type d'information consiste en l'exécution du processus d'indexation –abordé dans le deuxième chapitre- sur la collection de documents considérés pertinents pour la requête afin de produire une

[1] *Un message de 140 mots publié sur les murs des pages du réseau social twitter*
[2] *www.twitter.com*

représentation terminologique pondérée de chaque document [CAR01; BAI05; VOO04; DIA06; CHI07]. Les termes d'expansion sont sélectionnés depuis la terminologie créée.

4.4.1.2. Les textes d'ancrage

Le traitement de cette source consiste à analyser une collection d'hyperlien pour en extraire le texte des balises d'ancrage [KRA04], ce texte subira les prétraitements nécessaires (indexation) pour en extraire les termes d'expansion. Les expérimentations ont confirmé que la technique d'exploration de textes des balises d'ancrage surpasse les techniques similaires qui sont basées sur la fréquence d'occurrence des mots dans le texte des documents.

4.4.1.3. Les fichiers logs

Ce sont des fichiers texte qui enregistrent l'ensemble de données d'interaction d'un système informatique sous forme d'enregistrements généralement datés et classés par ordre chronologique. Les fichiers logs comprennent l'ensemble des requêtes exécutées ainsi que l'ensemble des documents visités par les utilisateurs pendant leurs sessions de navigation. L'analyse des fichiers logs permet d'en identifier les pages pertinentes à une requête donnée de manière implicite et d'éliminer les pages non pertinentes. Ainsi l'extraction des termes sémantiquement liés à la requête sera possible [BEE00; CUI03; BIL03, BOU13b].

4.4.2. Sélection des candidats termes

La sélection des candidats termes consiste à extraire depuis les données traitées dans l'étape précédente les termes pertinents qui peuvent servir à l'enrichissement de la requête initiale. Pour sélectionner une liste de termes appropriés à l'expansion d'une requête, une phase de désambiguïsation de celle-ci est alors nécessaire pour pouvoir identifier le sens qu'elle vise. L'ajout des termes peut s'effectuer soit par l'expansion de chaque terme de la requête séparément aux autres termes soit en considérant la requête comme étant un lot inséparable de termes.

Pour rapprocher l'information désirée de l'utilisateur, il faudrait appliquer une technique de sélection efficace de terme d'expansion à partir de ressources de termes précédemment collectées. La technique de tri de termes qui a été proposée par Harman [HAR92], se base sur une formule de pondération développée par Robertson et Sparck-

Jones [ROB76] qui consiste à sélectionner les termes ayant une valeur de poids importante dans les documents pertinents et ayant une faible probabilité d'apparition dans les documents non pertinents. La formule qui est présentée par l'équation (4.1) a permis d'obtenir de bons résultats de reformulation.

$$W_{ij} = \log_2 \frac{p_{ij}(1 - q_{ij})}{q_{ij}(1 - p_{ij})} \tag{4.1}$$

Où W_{ij} représente le poids du terme i dans la requête$j$, $p_{ij}(1 - q_{ij})$ est la probabilité que le terme i apparaisse dans les documents pertinents pour la requête j et $q_{ij}(1 - p_{ij})$ mesure la probabilité que le terme i apparaisse dans les documents non pertinents pour la requête j.

D'autres approches de tri exploitent le principe de cooccurrences des mots [BOU99, 00 ; BAI06; BOU13b] où les termes employés pour l'enrichissement sont sélectionnés sur la base d'un seuil de cooccurrence avec les termes de la requête initiale. À titre d'exemple, dans la paire de mots "Java, Voyage", chacun sert d'un contexte à l'autre qui permet de contraindre les termes connexes. Dans ce cas, les mots «hôtel» et «île» auront une probabilité de cooccurrence beaucoup plus importante avec "(Java, Voyage)" que celle du mot "programmation" et "langage".

4.5. CONCLUSION

Dans ce chapitre, Nous avons essayé de présenter les différentes approches et méthodes de reformulation de requête qui existent dans la littérature. Nous avons classé ces techniques selon trois critères principaux :

Selon le degré d'implication de l'utilisateur en, approche interactive où l'utilisateur constitue un acteur principal dans le processus de reformulation et approche automatique qui n'implique pas l'intervention directe de l'utilisateur.

Selon la source des termes d'expansion, nous avons distingué la méthode basée sur la réinjection de pertinence, la méthode basée pseudo-réinjection de pertinence et celle basée sur les ressources sémantiques.

Suivant le principe de génération des termes d'expansion, nous avons distingué l'approche linguistique qui se fonde sur les sources terminologiques et l'approche statistique qui se fonde sur la découverte des relations statistiques telles que la cooccurrence et de la corrélation entre les mots.

Nous avons également discuté les étapes du processus d'expansion à savoir la génération et la sélection des termes d'expansion.

PARTIE III : CONTRIBUTION

CHAPITRE 5 : COMPRENDRE LE COMPORTEMENT DE NAVIGATION DE L'UTILISATEUR DU WEB

5.1. INTRODUCTION

Depuis plusieurs décennies, le Web est considéré comme une source d'information caractérisée par sa richesse et sa diversité informationnelle ce qui a incité une grande audience à y conduire des recherches quotidiennes. Le développement important d'une masse d'informations de qualité sur le Web et le progrès technologique du software et du hardware également a entrainé un trafic de recherche énorme marqué par des outils ayant des performances importantes, appelés moteurs de recherche.

L'objectif principal de l'analyse de l'ensemble d'interaction utilisateur-Web est la compréhension du comportement de navigation de l'utilisateur afin de mettre l'accent sur les problèmes d'accès à l'information inhérents au facteur usager.

Dans ce chapitre nous parlerons du comportement de navigation de l'utilisateur dans les différentes générations du Web en mettant l'accent sur les caractéristiques mises en avant dans ces travaux.

5.2. LA RECHERCHE D'INFORMATION DANS LES DIFFÉRENTES GÉNÉRATIONS DU WEB

5.2.1. Dans le Web 1.0

La toute première génération du Web nommée Web 1.0 comprenait des pages statiques qui constituaient l'équivalent numérique des documents papiers existants interconnectés via des liens hypertextes pouvant être visitées via internet [CER90]. Pour atteindre l'information désirée, l'utilisateur devrait avoir recours au service d'un outil de recherche d'information à savoir un moteur de recherche, un multi-moteur, un

annuaire,…., etc. La recherche se lance par le biais des mots-clés saisis par l'utilisateur et vise à retourner les pages Web concernées.

À cette époque la recherche d'information a connu le développement d'une littérature solide qui a assuré une bonne gestion de l'information par les SRI. Nous pouvons noter les modèles de recherche d'information, les formules de pondération de termes dont la plus connue est la TF.IDF [SAL68; SPA79], les mesures d'évaluation de pertinence,….etc, qui seront discutées en détails dans le chapitre (2).

5.2.2. Dans le Web 2.0

Depuis l'avènement du Web 2.0 en 2004, de nouveaux axes de recherche ont été ouverts devant la communauté de la recherche telle que l'indexation basée tags [MIO13; CHO09], la recherche basée tags [WAN11 ; CHE12], la recherche au sein des réseaux sociaux [HOT06]…etc. En effet, les termes Web 2.0, Web social, Web participatif ou encore Web collaboratif désignent toutes les nouvelles techniques d'usage du Web centrées sur la gestion et le partage de données dans un contexte plus étendu où la production de l'information n'est pas monopolisée par une audience particulière. En revanche, il est devenu opportun à tout utilisateur qui devrait avoir certaines connaissances de manipulation de l'ordinateur d'ajouter différents types de contenu (texte, graphique, vidéo et son), l'annoter par le biais de tags et de participer ainsi à enrichir la masse informationnelle sur la toile.

L'importance des tags est acquise depuis leur capacité à repérer les ressources sur le Web [MAC06] et permettre ainsi de les retrouver par l'interrogation d'un SRI ou bien à travers la navigation dans un nuage de tags. Le Web social a donné lieu à la construction des fonds documentaires riches à travers les blogs, les tweets et les flux RSS, favorisant la recherche sur le Web où l'information est largement disponible, communément décrite et facile à exploiter par tous les usagers du Web.

5.2.3. Dans le Web 3.0

À l'époque du Web 3.0 ou Web sémantique, la recherche basée tags persiste, tandis que les tags qui décrivent une ressource donnée sont attribués automatiquement par les moteurs de recherche qui transforment les pages Web 1.0 et 2.0 en micro contenus. Cela permet de retrouver n'importe quelle information à partir de quelques fragments

significatifs de celle-ci [SHA02]. Le but de la démarche sémantique était l'obtention d'informations plus intelligentes en faisant une représentation structurée du contenu sémantique de celle-ci.

En fait, l'application des technologies du web sémantique a contribué positivement à l'amélioration des divers axes de la RI. Nous pouvons citer l'indexation basée sur les ontologies [BAZ05b ; GUA99; KHA13; ALF14] et la désambigüisation et l'expansion des requêtes par le biais des ontologies [BHO07; NAV03; SON07; SEG14].

5.2.4. Dans le Web 4.0

Le Web4.0 est la vision du Web encore en cours de développement qui consiste à considérer le web comme un système d'exploitation d'où vient le concept du WebOS (Web Operating System) ou système d'exploitation Web qui n'est rien qu'un système d'exploitation omniprésent, intelligent et qui s'adapte aux habitudes des internautes dont les outils sont disponibles en ligne à la demande. En web 4.0, les ressources (matérielles et logicielles) sont offertes à l'utilisateur quand il les demande et pour le temps qu'il veut. L'interaction de l'utilisateur devient plus en plus importante car tous les types de ressources qu'il s'habituait à utiliser localement et qui sont offerts par tous les systèmes d'exploitation traditionnels sont désormais disponibles sur le Web.

L'informatique en nuage (cloud computing) représente le concept clé du Web4.0. Il fait référence à la mise en service de l'être humain les différentes infrastructures, plateformes et applications comme étant des services web électroniques [BAU11]. De ce fait, l'accès à l'information sur le cloud [RUS10] est devenu plus facile et rapide car cette dernière est enregistrée en plusieurs copies dans le nuage ce qui la rend disponible à n'importe quel moment et à partir du n'importe quel endroit.

Ci-dessous, nous pressentons dans le tableau 5.1 une comparaison des caractéristiques de la tâche de recherche dans les différentes générations du Web.

Caractéristiques	Web 1.0	Web 2.0	Web 3.0	Web4.0
Interaction utilisateur/Web	Non significative	Importante	Importante	Très importante
Interrogation	Basée mot-clé	Basée mot-clé Basé tags	Basée mot-clé Basée tags Basée sur le langage naturel	Basée filtres intelligents
Résultats de recherche	Pages web qui incluent l'information recherchée	Pages web qui incluent l'information recherchée	Peut retourner une partie de la page seulement	La bonne information est délivrée au bon moment et dans le bon endroit
Indexation	Nécessite l'intervention des experts ce qui explique son cout élevé en matière de temps et d'effort	L'intervention des experts n'est pas nécessaire parce qu'elle peut être effectuée par les experts et les non experts ce qui réduit	L'intervention humaine n'est pas nécessaire à cause de l'aspect automatique qui caractérise ce type d'indexation ce qui implique un cout réduit	La capacité d'analyser des comportements et de les traduire en données utiles

Tableau 5.1.Caractéristiques de la recherche d'information dans les différentes générations du Web

5.3. INTENTION DE LA REQUÊTE

L'accès à l'information s'effectue généralement par le biais des requêtes, l'utilisateur qui effectue une recherche sur le web a certainement ses propres motivations, ces motivations ont permis selon [BRO02] de classer les requêtes de recherches en trois types sont : les requêtes navigationelles, les requêtes informationnelles et les requêtes transactionnelles.

5.3.1. Requête navigationelle

La conduite d'une recherche en utilisant une requête dite navigationnelle n'a pas forcément un but particulier excepté la navigation sur le net, les requêtes navigationnelles sont souvent consacrées à se divertir dans les sites de la toile. À titre d'exemple les requêtes *voyage, site de vidéos, jouer aux dames, tourisme en Algérie,*…etc., sont du type navigationnel. Brenes et Gayo-Avello [BRE08] ont développé un travail sur la détection des requêtes navigationnelles en combinant la plupart des techniques déjà proposées dans la détection automatique d'intention de requêtes [BRE09]. Jansen et al., [JAN07] confirment qu'environ 10% des requêtes sur le Web sont du type navigationnel, tandis que [BRO02] estiment leur fréquence à 25%, pour Rose et Levinson [ROS04], le nombre de ces requêtes est mesuré à 15%.

5.3.2. Requête informationnelle

Le but d'une requête informationnelle est la recherche d'une information particulière qui induit souvent la consultation de plusieurs résultats dans la page de réponse de l'outil de recherche afin de saisir l'information désirée, à titre d'exemple les requêtes : *le système solaire, débit internet, définition d'internet, comparatif ordinateur portable, biographie victor hugo*. Dans le but d'atteindre un taux important de détermination d'intention, Kang et Kim [KAN03] ont proposé quatre méthodes différentes pour distinguer les requêtes navigationnelles des requêtes informationnelles. En combinant les quatre méthodes, ils ont atteint de bons résultats (91.7% de taux de précision et 61.5% de taux de rappel en se référant aux conducteurs des recherches). Les travaux développés à propos de l'identification des intentions des requêtes sur le web confirment que le type informationnel est majoritairement répondu, où [BRO02] l'ont

mesuré à 40% du total des requêtes sur le Web. Aussi, [JAN07] ont trouvé que ce type de requêtes constitue 80% des requêtes sur le Web, selon Rose et Levinson [ROS04] elles constituent 60%. D'après ces valeurs, nous pouvons repérer que la plupart des requêtes du Web sont de type informationnel.

La différence entre les requêtes navigationnelles et les requêtes informationnelles, Uichin et al. [UIC05], est qu'en utilisant une requête navigationnelle, l'utilisateur attend une seule réponse alors qu'il attend plusieurs réponses en utilisant une requête informationnelle.

5.3.3. Requête transactionnelle

La soumission de ce type de requête engendre la génération d'une transaction qui permet à l'utilisateur de se rendre sur un site donné afin d'acheter un article en ligne, comme l'assurent les requêtes : *laptop pas cher, achat Ipad,* ou se bénéficier d'un service comme la consultation des météos, comme la requête *météo annaba*, ou encore pour effectuer une opération de téléchargement d'un logiciel ou d'un film, comme le garantissent les requêtes *download eclipse, harry potter7 download.* Les termes transactionnels comme (download et buy) ont constitué pour Tamine et al. [TAM08b] en plus de plusieurs autres éléments (comme le taux d'utilisation des termes dans le titre des documents, la longueur de la requête, etc.) les paramètres clés pour déterminer l'intention de la requête. Selon Jansen et al. [JAN07], Rose et Levinson [ROS04] et Broder et al. [BRO02], les requêtes transactionnelles constituent 10%, 25% et 35% respectivement du nombre total des requêtes sur le Web.

Le tableau 5.2 montre les résultats des travaux de Jansen et al. [JAN07], Rose et Levinson [ROS04] et Broder [BRO02] sur la répartition des requêtes sur le Web selon leur intention.

Type de requête	Navigationnel	Informationnel	Transactionnel
Broder [BRO02]	25%	40%	35%
Rose et Levinson [ROS04]	15%	60%	25%
Jansen et al. [JAN07]	10%	80%	10%

Tableau 5.2. Répartition des requêtes sur le Web selon leur intention

5.4. COMPORTEMENT DE RECHERCHE DE L'UTILISATEUR DU WEB

Nous avons commencé le travail avec l'observation exploratoire et l'analyse de comportement de recherche des utilisateurs du Web afin de découvrir les caractéristiques de leurs comportements de navigation en général et leurs tendances de recherche en particulier, et vérifier si leurs recherches dépendent des facteurs externes pouvant avoir un impact sur l'information recherchée.

5.4.1. Déroulement de l'observation

Nous avons choisi d'explorer de manière assez générale les caractéristiques de la tâche de recherche à partir des données de trafic offertes par Google qui est le moteur de recherche le plus populaire sur Internet[1]. En effet, Google présente à travers le service qu'il fournit appelé Google Insights[2] les statistiques sur la fréquence des recherches, leur localisation, leur temps et leur popularité et en plus des tendances de recherches des utilisateurs du Web. En effet, les données sont mises à l'échelle en prenant la valeur 100 pour le mot-clé ayant la plus grande fréquence. Le but de ces observations est la compréhension du comportement de recherche des utilisateurs à travers leurs requêtes.

Tout d'abord, nous nous sommes penchés vers l'inspection des requêtes portant sur des sujets d'intérêt communs, pour ne pas créer de différence sur le plan de l'âge, de la localisation, des connaissances ou de compétences particulières. En effet, les utilisateurs du Web effectuent des recherches sur une infinité de sujets. Nous avons examiné comment les utilisateurs recherchent sur des sujets particuliers tels que la musique, le sport et les jeux comme étant des sujets d'intérêts communs, ainsi que la recherche sur des sujets d'intérêts métiers tels que les mathématiques, la physique et la biologie.

Puis, nous avons consulté la fréquence de soumission des requêtes à propos des sujets cités un peu plus haut sans faire une restriction géographique ceci dit que les résultats présentés représentent tout le trafic de recherche marqué par le moteur de recherche à travers le monde entier.

[1] *http://www.comscore.com/Press_Events/Press_Releases/2012/2/comScore_Releases_January_2012_U.S._Search_Engine_Rankings*
[2] *http://www.google.com/insights/search/*

Ensuite, nous avons essayé de visualiser le trafic de recherche sur les différents sujets d'intérêt dans une zone géographique particulière que c'était l'Algérie.

Les travaux de Jansen [JAN07] pour classifier les requêtes traitées selon leurs intentions, nous ont beaucoup influencés. En effet ils ont présenté dans leurs travaux des spécifications qui permettent de distinguer entre les différents types de requêtes.

5.4.2. Résultats généraux

5.4.2.1. Aspects périodique des requêtes

1) La périodicité journalière des recherches à propos des sujets d'intérêts communs : comme le montre la figure (5.1), la répartition des recherches à propos des sujets d'intérêts communs que nous avons choisi de les inspecter à savoir (les jeux et les multimédias) selon le jour sur Google au cours du mois de janvier 2012 dépend du type du jour. Effectivement, les recherches sur ces sujets d'intérêts communs s'avèrent être périodiques (BOU13b), c.à.d. les mêmes requêtes se produisent souvent dans les mêmes jours. Nous pouvons remarquer aussi que ce type de recherches atteint leurs meilleurs scores dans les week-ends entre le vendredi et le dimanche.

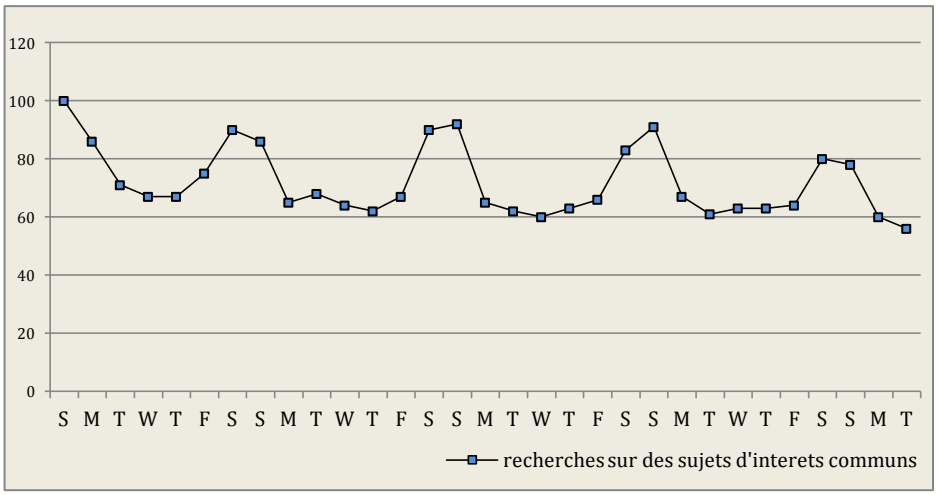

Figure 5.1.Exemple extrait des recherches établies sur Google sur des sujets communs au cours du mois de janvier 2012 et leur répartition selon la journée

2) La périodicité journalière des recherches à propos des sujets d'intérêts métiers : nous appelons un sujet d'intérêt métier, tout centre d'intérêt qui peut permettre de mieux identifier les différentes communautés d'intérêt sur le Web. À cette étape, nous avons inspecté les requêtes spécifiant le type des documents recherchés à travers l'utilisation des extensions (pdf, doc, ppt, ps, etc) ainsi que celles contenant une interrogation ("comment", "c'est quoi", etc.). En effet, en plus du fait que la nature de la tâche courante de l'utilisateur influe son comportement de recherche [LI08], les documents recherchés par ce type de requêtes incluent majoritairement du texte. Donc, une telle spécification dans la requête de l'utilisateur lui attribue l'intention informationnelle [JAN07]. Comme le montre la figure (5.2), les recherches sur des centres d'intérêt métiers atteignent leur maximum de soumission durant les jours de travail.

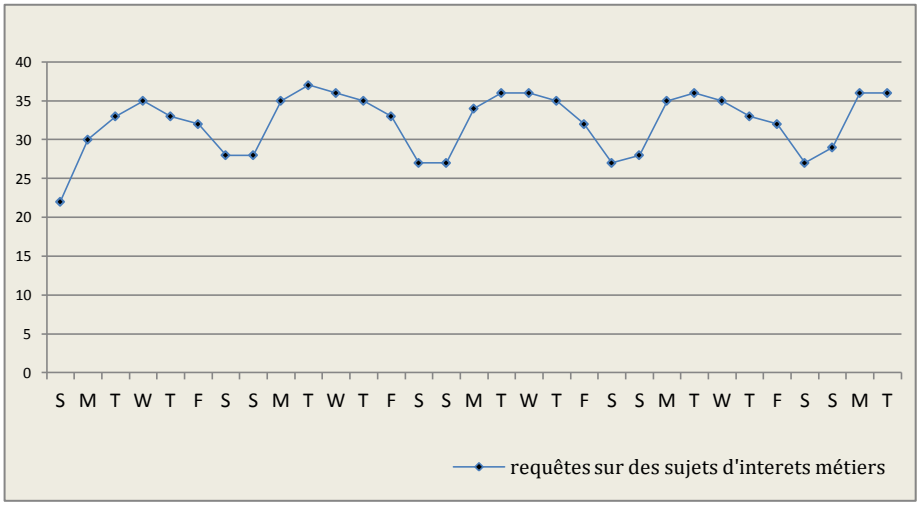

Figure 5.2. Exemple extrait des recherches établies sur Google sur des requêtes métiers au cours du mois de janvier 2012 et leur répartition selon la journée

3) La périodicité saisonnière des recherches: les résultats obtenus sur le trafic de recherche dans une zone géographique spécifique qu'était l'Algérie montrent une baisse repérée pendant la période estivale des requêtes informationnelles. Cette baisse marquée à propos des différents sujets métiers inspectés qui sont: les mathématiques, la physique et la biologie s'explique par la diminution des activités informationnelles des internautes pendant les vacances et les congés. Néanmoins, ils s'intéressent davantage à la recherche de la musique et des jeux, ce qui explique la hausse du nombre

des requêtes transactionnelles durant la même saison. Concernant les requêtes navigationnelles repérées par les recherches sur les noms de plusieurs stars du cinéma américain et celles sur quelques compagnes comme Microsoft et Samsung, nous avons remarqué que leur fréquence connaît plusieurs élévations durant l'année ce qui peut être expliqué par l'influence d'un ou plusieurs événements sur l'activité de navigation des utilisateurs du Web. Il est à noter que l'aspect périodique des requêtes persistait également dans une zone géographique restreinte. Le diagramme présenté dans la figure (5.3) interprète clairement les conclusions obtenues à propos de la périodicité des recherches sur le Web.

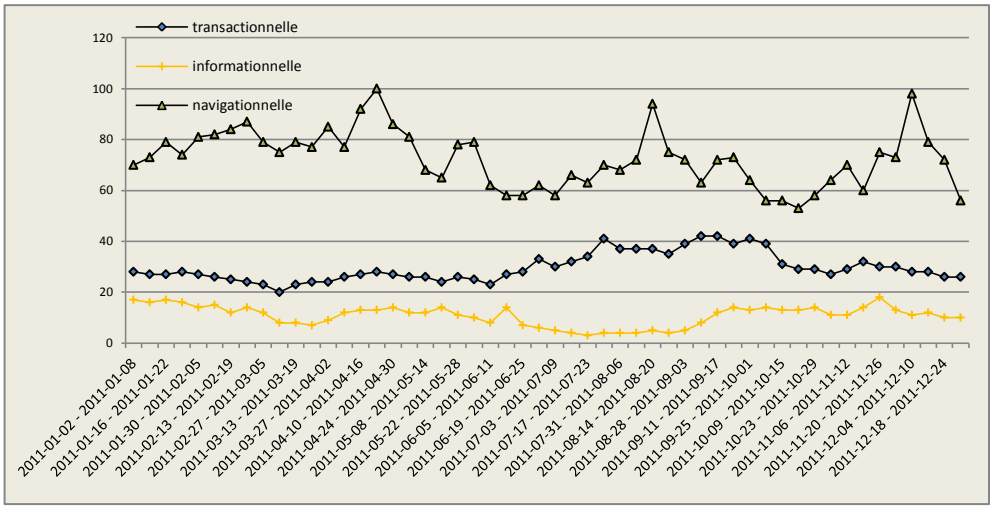

Figure 5.3.Exemple extrait des recherches établies sur Google au cours de l'année 2011 en Algérie classifiées selon l'intention des requêtes

5.4.2.2. Longueur des requêtes

Beaucoup de travaux ont montré que les requêtes soumises aux moteurs de recherche sont souvent courtes [JAN98 ; SIL99 ; ZIE01; BOU13b] et ils estiment leur longueur à environ 3 mots. Nous pouvons confirmer à partir des requêtes les plus fréquentes les résultats obtenus auparavant dont la figure (5.4) présente un exemple des 10 requêtes les plus fréquentes.

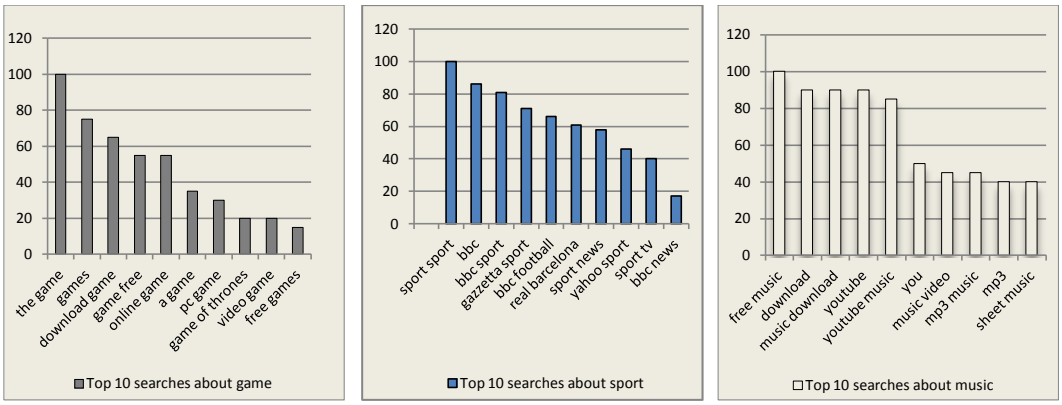

Figure 5.4.Exemple des Top 10 des requêtes soumises au cours du mois de janvier 2012

5.4.2.3. Fréquence des requêtes et événements

Dans l'analyse du trafic Web, nous nous sommes intéressés à discerner si la fréquence des requêtes soit dépendante des événements particuliers agissants un peu partout dans le monde, ceci en inspectant la recherche d'information sur des événements particuliers à savoir une compétition mondiale, la sortie d'un film particulier, un sinistre ou un phénomène naturel. L'idée de base s'agissait de trouver la fréquence des recherches à propos d'un événement particulier dans des périodes particulières (BOU12a, 13a). Si nous prenons l'exemple des recherches effectuées sur Google pendant juin et juillet 2010, on observe comme le montre la figure (5.5) que les requêtes "*word cup*", *"world cup 2010"*, *" FIFA world cup"* et *"soccer world cup"* ont dominés les sujets recherchés. En effet, la période considérée est marquée par le déroulement de l'événement mondial « la coupe du monde de football 2010 ». En faisant une comparaison du trafic de recherche sur cet événement en 2010 avec celui de la même période de l'année 2011 nous pouvons remarquer une baisse considérable de ces requêtes.

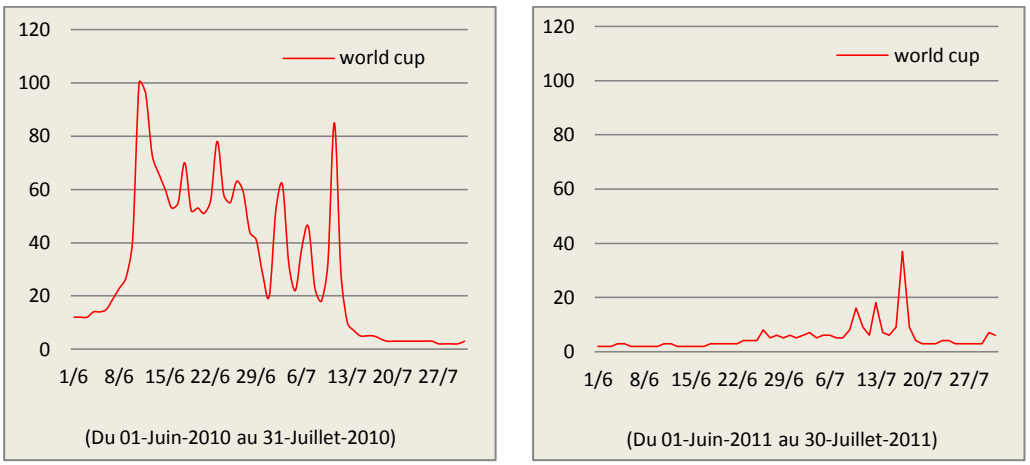

Figure 5.5.La fréquence des requêtes en présence d'un événement particulier

5.5. CONCLUSION

De nombreuses études menées au profit du comportement de l'utilisateur sur le Web ont permis de mieux cerner les problématiques de la RI et ainsi les résoudre. Le besoin d'avoir une information donnée par le biais d'une requête brève dans un temps réduit et l'aspect de consultation des résultats de recherche limitée aux premières pages Web retournées sont les caractéristiques capitales du comportement des usagers.

En effet, la présence d'un événement mondial quelconque a un impact important sur les tendances de recherche à travers le monde. En même temps, la présence d'un événement dans une zone géographique restreinte influe également le comportement des usagers dans cette région. Finalement, les requêtes soumises aux moteurs de recherche se caractérisent par leurs aspects périodiques qui les rendent fréquentes sur des périodes de temps particulières.

CHAPITRE 6 : CONTEXTE DE RECHERCHE ET BESOIN INFORMATIONNEL DE L'UTILISATEUR DU WEB

6.1. INTRODUCTION

En raison de la brièveté des requêtes et l'aspect polysémique manifestant dans le langage humain, un utilisateur qui cherche à télécharger la dernière version de l'environnement de développement intégré (IDE) Eclipse, peut être servi avec du contenu sur l'éclipse lunaire, l'IDE Eclipse ou encore sur le film éclipse qui a eu beaucoup de succès pendant plusieurs saisons.

Effectivement, le comportement de navigation de l'utilisateur dépend de plusieurs paramètres comme le temps, la localisation, la tâche courante, l'historique d'interaction sur le Web, etc., ce qui constituent ce qu'on appelle le contexte de recherche. Dans cette première contribution et dans le but d'identifier l'intérêt cible de l'utilisateur visé à partir de sa requête. Nous avons exploité les intérêts de l'utilisateur à court terme déduits à partir de son historique de navigation récent [BOU12a, 13a, b]. Aussi, nous avons proposé de détecter la présence d'un événement particulier lié à la recherche en cours dans le but de mieux cerner cet intérêt.

6.2. Contribution

Tandis que, les usagers du Web ne soutiennent pas généralement l'idée d'exprimer leurs besoins en information par des requêtes longues et précises [JAN98 ; SIL99 ; ZIE01 ; HOC08], ni de fournir explicitement des informations sur leurs domaines d'intérêt, la compréhension du besoin informationnel de l'utilisateur devient plus sophistiquée ce qui rend ainsi l'accès à l'information désirée moins évident. Dans une tentative d'augmenter la pertinence des résultats renvoyés dans une recherche, nous avons pensé à développer une méthode basée contexte ayant pour objectif d'identifier l'intérêt cible

de l'utilisateur parmi l'ensemble des intérêts susceptibles d'être visés [BOU12a, 13a]. Pour arriver à cet objectif, nous avons proposé une modélisation du comportement de navigation des utilisateurs traités par groupes de comportements, ceci en prenant en considération l'intérêt métier de l'utilisateur pour discriminer entre les différents groupes et en exploitant des informations sur le contexte de la recherche. À travers les modèles créés, nous cherchons à déterminer l'intérêt visé par la requête de l'utilisateur en se basant sur des comportements de navigation passés des utilisateurs ayant effectuait des recherches similaires dans des contextes similaires. Par exemple dans une session de navigation les deux utilisateurs u_1 et u_2 qui s'intéressent à un ensemble de centres d'intérêt, $u_1 = \{d_1, d_2, d_3, d_5, d_6\}$ et $u_2 = \{d_3, d_4, d_5, d_8\}$, effectuent des recherches similaires à propos des deux sujets d'intérêt d_3 et d_5. Selon le principe de base de notre approche, les deux utilisateurs seront traités de la même manière durant cette session indépendamment du reste de leur activités de navigation.

Chaque requête est traitée comme étant un lot inséparable de mots reliés par le connecteur AND. Soit donc une requête $q = \{w_1, \ldots . w_m\}$ tel que $m \in N^*$, qui vise un ensemble d'intérêts $D = \{I_1, .., I_n\}$ tel que $n \in N^*$. Notre objectif est de répondre aux questions suivantes : parmi cet ensemble de domaines d'intérêt quel est le domaine visé par la requête? Comment peut-on le trouver et sur quel critère sera-t-il identifié?

Tout d'abord, nous avons commencé à acquérir les données qui nous intéressent depuis l'historique d'interaction de l'utilisateur avec le Web. Puis, nous avons traité les sessions de navigation similaires au sein d'un même groupe de comportement. L'idée de base est que le traitement de la requête dépend de la session de navigation et du contexte de recherche immédiat et qu'au-delà des intérêts métiers, les comportements de navigation en général et de recherche en particulier des utilisateurs peuvent être semblables. La figure (6.1) ci-dessous présente l'architecture générale de la démarche proposée.

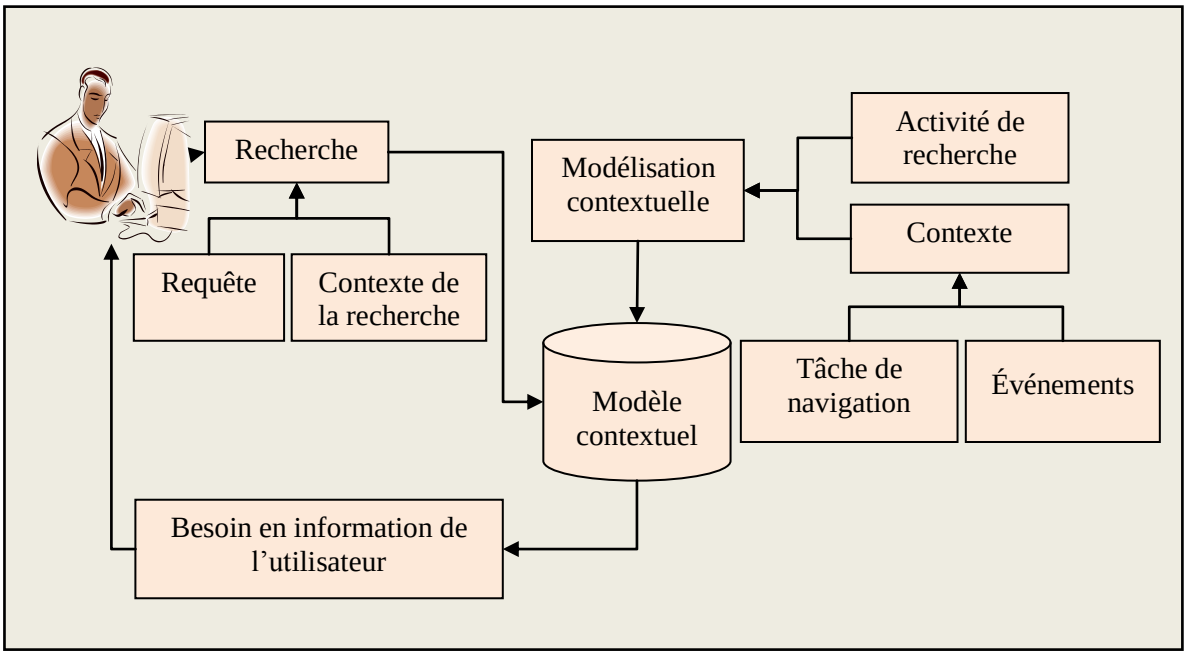

Figure 6.1.Architecture générale de l'approche

6.3. PRINCIPAUX CONCEPTS

6.3.1. Intérêt métier

Nous appelons intérêt métier, le domaine d'expertise ou d'étude de l'utilisateur. Il constitue un paramètre de discrimination entre les groupes de comportements.

6.3.2. Intérêt générique

Il représente un intérêt majoritairement partagé au sein d'une communauté.

6.3.3. Groupe comportemental

Nous appelons groupe comportemental, chaque groupe de comportements de navigation similaires, et qui partagent le même intérêt métier. Soit G_i un groupe comportemental défini comme suit (formule 6.1):

$$G_i = \begin{Bmatrix} \langle\langle b_1, TI_1\rangle, \dots, \langle b_n, TI_n\rangle\rangle \\ / \; b_1 \approx .. \approx b_n, TI_1 = .. = TI_n \end{Bmatrix} \quad (6.1)$$

Tels que $b_i / i = 1..n$ représentent un ensemble de comportements de navigation et $TI_i / i = 1..n$ correspondent au même intérêt métier.

6.4. ETAPES DE RÉALISATION

6.4.1. Première étape : l'acquisition de données

Les fichiers logs représentent une source de données largement étudiée dans la fouille de données d'usage du Web. En effet, on peut distinguer trois types de logs qui sont les logs serveurs, les logs proxy et les logs clients [SRI04].

- Les logs des serveurs Web sont utilisés le plus souvent dans la découverte des motifs séquentiels communs [COO97 ; SRI00 ; MOB07] ou bien pour la classification des utilisateurs [GAR05; MOG11; QIU06],….etc. L'inconvénient de cette approche est la difficulté d'appréhender les configurations d'accès des utilisateurs sur plusieurs serveurs Web. En plus de l'impossibilité de garantir que si le flux qui parvient d'une telle adresse IP appartient à un utilisateur unique. Un type particulier des logs des serveurs est les logs qui sont extraits des moteurs de recherche tel que AOL Log[1]. En fait, ces logs ne couvrent que l'activité de recherche établie sur le moteur en question. Ainsi, ils ne vérifient pas l'aspect d'interrogations multiples de plusieurs moteurs de recherche à la fois.

- Les logs des serveurs proxy [WU02 ; PUR04] vérifient l'acquisition implicite des interactions de l'utilisateur avec le Web. L'avantage du proxy est qu'il permet de contrôler le flux d'interaction des machines qui y sont connectées. Cependant, son traitement pose de vrais problèmes d'identification des différents utilisateurs ayant une même adresse d'accès ou encore l'identification d'un utilisateur unique ayant plusieurs adresses d'accès.

- Les logs clients, permettent d'obtenir l'historique des sessions de navigation qui appartient souvent à un utilisateur unique. Les études réalisées sur la base des données capturées à partir de la machine cliente sont plus authentiques et précises [QUI00] dans la modélisation utilisateur. Cependant, la collecte de ce type de données nécessite l'installation d'un navigateur Web personnalisé, ou un logiciel espion sur la machine cliente comme le Framework de recherche

[1] *http://www.gregsadetsky.com/aol-data/*

Tracker [CHO00] et Wrapper [JAN06b] ou encore à travers les cookies [LIU02; MOB07] que sont des fichiers texte envoyés par un serveur Web au navigateur client afin d'identifier les utilisateurs de manière individuels. Il est à noter que les données enregistrées dans les logs client et ceux des serveurs proxy ne peuvent pas être complètes en raison de la mise en cache des données de navigation, ce qui représente une autre limitation dans l'utilisation de ce type de logs. Dans le but d'exploiter les données des fichiers logs, différentes approches de traitement ont été proposée que Wang et al., ont essayé à les synthétiser dans [WAN03].

6.4.1.1. Source de données

La réalisation de cette première contribution requière initialement la collecte de données de navigation au cours d'une période de temps plus ou moins longue marquée par un ou plusieurs événements que ce soit régional ou mondial. Pratiquement, nous sommes intéressés à l'activité de navigation habituelle des utilisateurs afin de pouvoir traité leurs comportements naturels. En effet l'utilisation des logs des moteurs de recherche sur le Web tel que AOL et MSN [MIC08],...etc., ne satisfait pas nos besoins de couvrir la totalité des activités de navigation y compris l'accès direct aux sites Web. En plus, les données de ces logs ne vérifient pas l'aspect d'interrogations multiples de plusieurs moteurs de recherche à la fois. Suite au problème de l'impossibilité d'identification de l'activité de navigation propre à un utilisateur unique, l'utilisation des logs des serveurs Web et des serveurs proxy également ne permet pas une bonne réalisation de l'approche proposée.

Pour les expérimentations, nous avons construit notre propre ensemble de données en utilisant un plug-in pour le navigateur Firefox[1] qui est chargé d'enregistrer l'historique de navigation à partir de huit ordinateurs mis à la disposition d'un ensemble d'utilisateurs composé des étudiants universitaires de différentes spécialités. Chaque utilisateur peut naviguer pendant une heure par jour; le plug-in installé sur chaque machine enregistre toute l'activité de navigation des utilisateurs capturée à partir du navigateur.

[1] *http://www.mozilla.org/fr/firefox/new/*

L'un des problèmes majeurs envisagés lors du traitement des logs est la difficulté d'identifier des utilisateurs pouvant avoir des flux d'interaction multiples depuis des hôtes différents. Pour le log collecté durant le mois de novembre, 2011, nous avons recueilli 12 Mo de log y compris 396 sessions de navigation d'une heure de temps où chaque session appartient à un utilisateur unique et elle comprend 4 requêtes en moyenne. La figure (6.2) et le tableau (6.1) montrent le trafic capturé représenté par le nombre de sessions et le nombre de requêtes respectivement par semaine.

	Semaine (1)	Semaine (2)	Semaine (3)	Semaine (4)	Total
#sessions	90	75	51	180	396
#requêtes	345	305	196	716	1562

Tableau 6.1.Une description de l'ensemble de données

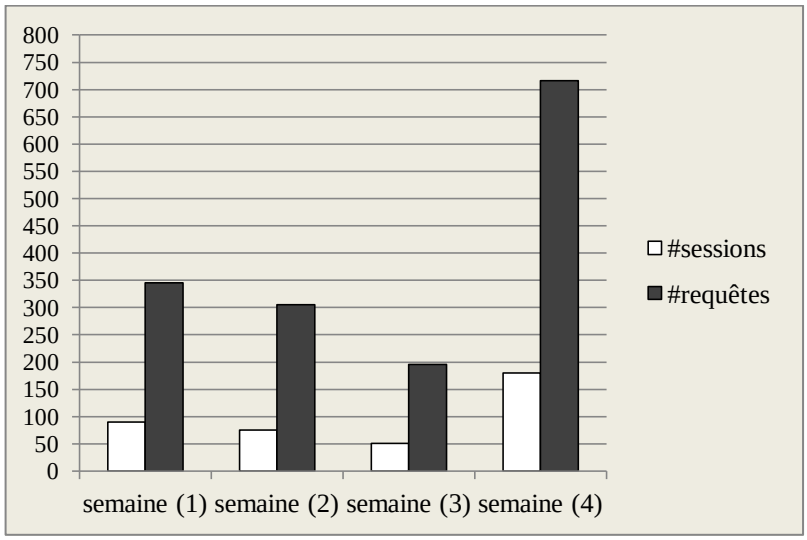

Figure 6.2.Une description de l'ensemble de données

6.4.1.2. Requêtes populaires

Parmi l'ensemble des requêtes marquées dans la période de temps traitée, nous avons noté un nombre important de requêtes sollicitant des sites de messageries et de détente. Nous avons classé ces requêtes selon leurs fréquences de soumissions et nous avons présenté les 20 premières requêtes populaires dans le tableau (6.2) tel que le premier rang est affecté à la requête la plus recherchée.

Rang	Fréquence	Requête
1	31	aljazeera
2	29	exercice
3	27	mail
4	25	mp3
5	13	pdf
6	13	tp link
7	12	video de
8	12	yahoo mail
9	12	youtube
10	11	aljazeera sport
11	11	doc pdf
12	11	music mp3
13	11	exercice de math
14	11	music 2011
15	11	jeux
16	11	math algerie
17	10	telecharger mp3
18	10	tp chimie
19	10	google
20	10	hotmail

Tableau 6.2.les vingt premières requêtes fréquentes

6.4.1.3. Descripteur de domaine

Ce sont des vecteurs qui contiennent les mots fréquents dans chacun des domaines traités dans ce travail et qui se manifestaient majoritairement dans les données de l'historique que nous avons exploitées, à savoir l'informatique, les mathématiques, la biologie, le sport, les news et les multimédias.

6.4.2. Deuxième étape: construction des groupes comportementaux

Nous avons commencé à partir de l'analyse des comportements de navigation des utilisateurs en vue d'arriver à faire un regroupement cohérent et homogène. Ce regroupement se fonde sur le principe que les utilisateurs qui partagent le même intérêt métier et qui possèdent des comportements de navigation similaires sont rassemblés dans le même groupe comportemental.

Plus spécifiquement, le groupe comportemental comprend un ensemble de données de recherche autour d'un ou plusieurs intérêts. Ces données sont subdivisées en données d'intérêt métiers et données d'intérêt génériques, en plus des données événementielles représentées par le paramètre événement en temps réel pouvant affecter la recherche, ainsi que des données décrivant l'activité de navigation de l'utilisateur.

6.4.2.1. Comportement de navigation

Un comportement de navigation b_i est l'ensemble des requêtes soumises et des pages Web visitées par l'utilisateur dans une session de navigation, il est représenté sous forme de vecteur de termes pondérés. Le poids de chaque terme t_i est calculé par l'équation (6.3) qui est une variation de la formule de pondération *tf.idf* [SAL68; SPA79] qui prend en compte la fréquence du terme dans la session de navigation ainsi que son degré de discrimination avec une faveur attribuée aux termes qui expriment un domaine d'intérêt métier. Cette représentation comporte l'ensemble des termes de recherche, les termes significatifs extraits des pages web pertinentes visitées au sein de la session ainsi que leurs poids. Le comportement b_i est représenté par le n-uplet suivant (formule 6.2):

$$b_i = \{(t_1, w_1), \dots, (t_n, w_n)\} \tag{6.2}$$

$$w_i = log(1 + TF_i) * s_i \tag{6.3}$$

TF_i mesure la fréquence d'un terme dans la session de navigation et le facteur s_i représente un seuil décrivant le degré de discrimination du terme, sachant que 50% des sessions traitées incluent au moins une consultation d'un intérêt commun. Tandis que

environ 95% des sessions incluent au moins une activité de navigation à propos d'un intérêt métier. Le seuil s_i prend les valeurs suivantes :

$$s_i = \begin{cases} 1 & si\ le\ terme\ décrit\ un\ domaine\ métier \\ 0.5 & si\ le\ terme\ décrit\ un\ domaine\ commun \end{cases} \quad (6.4)$$

L'idée sous-jacente à l'affectation de telles valeurs pour le seuil consiste à attribuer moins d'importance aux intérêts communs par rapport aux intérêts métiers. Nous présentons dans ce qui suit un exemple d'un comportement de navigation d'un utilisateur qui a effectué pendant une session de navigation une seule recherche et il a consulté deux pages web différentes.

- La recherche est effectuée par le biais de la requête $q = communication$. L'utilisateur a consulté 2 pages Web renvoyées en résultat.

- En utilisant Treetagger[1] les concepts clés de chaque page consultée sont extraits.

- Les poids des termes de la requête et des pages consultées sont calculés par la formule (6.3) en prenant compte de leur fréquence et de leur intérêt par rapport à l'utilisateur. b_i = {(communication, 18), (reseau social, 28.52), (technologies de communication, 14), (telephone mobile, 8), (facebook, 8), (conversation, 7), (twitter, 6.33), (reseau personnel, 5)}

Suivant l'ensemble de données disponibles nous avons construit 10 groupes comportementaux différents. Ces groupes sont décrits dans le tableau (6.3).

Groupe	Centres d'intérêt
G_1	Informatique et technologies
G_2	Mathématiques
G_3	Sciences de la vie
G_4	Informatique, technologies et mathématiques
G_5	Informatique, technologies et sciences de la vie
G_6	Mathématiques, informatique et technologies
G_7	Sciences de la vie, informatique et technologies

[1] Site du logiciel : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

G_8	Sport et loisir
G_9	News
G_{10}	Images, musique et vidéo

Tableau 6.3.Description des groupes comportementaux

6.4.2.2. Contexte de la tâche de navigation

Les travaux de Li et al., [Li08] confirment que la nature des activités de l'utilisateur influe son comportement de navigation. Par exemple la tâche d'achat d'un ordinateur génère le besoin de s'informer sur les différentes marques commerciales de PC et les caractéristiques des produits qu'elles fabriquent afin de pouvoir comparer et choisir l'ordinateur qui convient en matière de qualité et prix. Cela mène l'utilisateur à conduire une recherche par le biais d'une ou plusieurs requêtes à propos de ce sujet, dans ce cas, la recherche est guidée par la tâche d'achat. L'activité ou la tâche se définit comme une série d'actions à accomplir pour atteindre un but particulier. En recherche d'information la tâche peut être considérée comme le stimulateur qui favorise un besoin d'information qui conduit, à son tour, à effectuer l'interrogation d'un SRI [VAK03].

Les travaux de Shen et al., [SHE05] ont mis en évidence l'exploitation de la tâche courante de l'utilisateur pour améliorer les réponses de recherche en proposant une approche de modélisation et d'expansion de requêtes dans le but de personnaliser l'accès à l'information sur le Web. Leur contribution consiste dans un premier temps à déduire les intérêts de l'utilisateur en mettant l'accent sur les activités exercées pendant la recherche. Le contexte exploité pour l'enrichissement des requêtes, se composait des requêtes qui précèdent immédiatement la recherche en cours, ainsi que les résultats correspondants. En conséquence, le système à base d'agent de recherche développé appelé UCAIR [UCA10] a généré des résultats de recherche meilleurs que ceux du moteur de recherche Google. Dans ce même contexte, Asfari et al. [ASF09, 12] se sont focalisés sur l'état de la tâche que l'utilisateur tente d'accomplir quand il conduit une recherche afin de construire un modèle de tâche pour la reformulation de requêtes. Tandis que, [FRE05] s'intéressaient également à la notion de tâche pour améliorer la recherche d'information dans les lieux de travail.

Le présent travail propose de prendre en compte toute l'activité de navigation récente de l'utilisateur et ne pas se limiter à l'ensemble des activités liées seulement à la requête traitée [BOU12a, 13a, b], et ceci dans une perspective de déterminer le besoin de

l'utilisateur à partir d'un comportement similaire étudié auparavant. En effet, l'historique de navigation récent de l'utilisateur peut fournir des informations sur ses centres d'intérêt ainsi que son domaine d'étude ou d'expertise. Dans le présent travail, la tâche de navigation ou comme nous l'appelons les intérêts récents de l'utilisateur sont obtenus à partir de l'analyse de l'historique de navigation récent de celui-ci. Il est à noter que la période de temps analysée peut atteindre une heure de navigation en continu pour chacun, ce qui constitue dans notre cas la durée maximale de navigation en continuité d'un utilisateur dans une journée entière. Le traitement de l'historique de navigation à court terme passe par les étapes suivantes :

a. Identification des pages pertinentes

La pertinence de chaque page visitée pendant la session de navigation est calculée selon le nombre de clics dessus et le temps de visite par l'équation (6.5).

$$R(p,S) = \frac{Tv.C}{T.C_T} \tag{6.5}$$

Ici : Tv mesure le temps de visite de la page p.

C mesure le nombre de clics sur la page p.

C_T correspond au nombre total de clics sur toutes les pages visitées durant la session correspondante.

T est la durée de la session de recherche.

b. Identification de l'intérêt de la page

Pour chaque page pertinente, le domaine est par la suite identifié grâce aux descripteurs de domaines.

c. Représentation linéaire des intérêts de l'utilisateur

Selon l'ensemble de données, nous avons obtenu des requêtes sur 6 domaines d'intérêt différents, les intérêts récents de l'utilisateur sont représentés sous forme d'un vecteur

binaire de 6 valeurs obtenues par l'exécution de l'algorithme *interets_recents*. Par exemple, si l'utilisateur a récemment visité le site Yahoo[1] actualités et le site microbe-edu.org avant de soumettre sa requête, l'algorithme *Interets_recents* retourne le vecteur $I = (0,0,0,1,0,1)$.

Algorithme 1.Interets_recents
Entrées : $S_i = \sum(p_i, q_j)$ D : descripteur de domaine Sortie : I_i DEBUT Pour chaque page visitée avant q_j Faire Calculer sa pertinence $R(p_j, S_i)$ Si la page est pertinente alors Calculer le domaine d'intérêt de la page T=topic(p_i); I_i[T]=1; Retourner I_i FIN

La fonction $topic(x)$ retourne le domaine pouvant être ciblé par la requête ou la page Web, si la requête ou la page Web réfère à plus d'un domaine d'intérêt, elle choisit un domaine aléatoirement.

6.4.2.3. Événement affectant la recherche

L'observation du trafic de recherche des utilisateurs montre qu'en présence d'un événement particulier, le nombre de requêtes connexes augmente considérablement.

L'objectif envisagé dans ce travail n'est pas la détection des évènements incitant la recherche d'information mais plutôt, bénéficier de la relation de dépendance évènement/requêtes connexes afin d'améliorer la recherche sur le Web [BOU12a, 13a]. En effet, il n'est pas évident de détecter tous les évènements qui occurrent et qui peuvent influer les tendances de recherches des utilisateurs du web que ce soit à travers le monde ou bien dans une zone géographique restreinte.

[1] *http://www.yahoo.com/*

La présence d'un évènement E implique l'augmentation du nombre de recherches effectuées sur cet événement, ce qui nous a amené à définir le score RTE (Real Time Event) qui permet de détecter la présence d'un événement lié à une requête donnée depuis l'augmentation des recherches similaires dans une période de temps particulière. Le score RTE est défini par l'équation (6.6).

$$\mathrm{RTE} = \left(\sqrt[3]{\frac{A + B}{B}} - 1 \right) \tag{6.6}$$

Où, A représente le nombre de requêtes connexes dans n unités de temps, B représente le nombre des requêtes durant m unités de temps antérieures, tel que n>=2 et m=n/2. Prenant une semaine comme unité de temps, si n = 2, le score RTE va mesurer l'augmentation du nombre de requêtes soumises à propos d'un sujet donnée dans deux semaines. Dans ce travail, nous avons considéré qu'il y a une augmentation significative dans le nombre de requêtes si le RTE est supérieur à 0,16.

- Exemple applicatif

Selon les statistiques de Google insights illustrées dans la figure (6.3) qui permettent de constater clairement la dépendance événement/requêtes connexes en citant l'exemple de l'événement "élection présidentielle" en France qui s'est déroulé le 22 Avril et le 6 mai 2012.

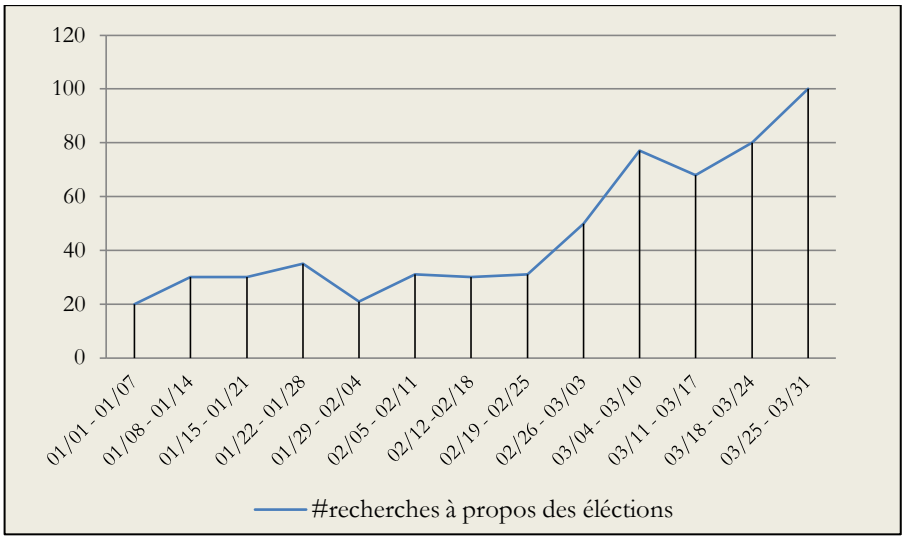

Figure 6.3.L'évolution des recherches utilisateurs au fil du temps en faveur d'un événement particulier

Dans la période considérée, du 01 janvier 2012 jusqu'à le 31 mars 2012, le nombre de requêtes effectuées sur Google depuis la France par le biais du terme "élection présidentielle" a suivi une tendance croissante.

Le nombre de recherche dans la semaine du 16-02-2012/22-02-2012 est $B = 566$ et le nombre de recherches dans la semaine du 23-02-2012/29-02-2012 est $A = 613$, ainsi, le RTE est calculé comme suivant :

$$\mathrm{RTE} = \left(\sqrt[3]{\frac{\mathrm{A}+\mathrm{B}}{\mathrm{B}}} - 1 \right) = \left(\sqrt[3]{\frac{1179}{566}} - 1 \right) = 0.27 \geq 0.16$$

6.4.3. Troisième étape : la modélisation

Dans ce travail, nous avons opté pour une modélisation par les réseaux de neurones artificiels (RNA) avec un apprentissage automatique supervisé.

6.4.3.1. L'apprentissage connexionniste

Le Perceptron Multicouches (PMC) [ROS58] est l'un des réseaux de neurones les plus utilisés pour des problèmes d'approximation [CYB88] et de classification [SHE06 ; UM00]. Il est classé parmi les réseaux à "propagation vers l'avant", c'est-à-dire qu'en mode normal d'utilisation, l'information se propage dans un sens unique, des entrées vers les sorties sans aucune rétroaction. Son apprentissage est de type supervisé, par

correction des erreurs. Il est habituellement constitué de deux ou trois couches de neurones totalement connectées.

Dans les problèmes de classification, le rôle du PMC consiste à arranger les données élémentaires dans une ou plusieurs classes ou catégories prédéfinies [FAY96]. Ses performances dépendent principalement de la représentation des données utilisées pour l'apprentissage et elles peuvent être équivalentes à celles des meilleurs classifieurs en termes du taux de reconnaissance si le traitement des données, leur représentation et la configuration du réseau sont bien choisies.

Nous avons opté pour le PMC pour l'apprentissage des sessions de navigation des utilisateurs qui ont été réunis en groupes de comportements. L'ensemble de données utilisé comprend 396 entrées subdivisées en deux sous-ensembles un pour l'apprentissage et l'autre pour les tests. L'ensemble des données d'apprentissage comporte 250 entrées, tandis que l'ensemble des données de test contient 146 entrées.

a. Phase d'apprentissage

Dans cette étape, les caractéristiques de chaque groupe de comportement ont été apprises afin de générer des modèles. Par manque d'une méthode formelle de configuration des PMC, le choix du nombre de couches ainsi que le nombre de neurones dans chacune a été déterminé en faisant des expériences sur plusieurs configurations, jusqu'à arriver à la plus optimale en matière du temps d'apprentissage, de la précision et de la cohérence des réponses. Après avoir testé de nombreuses configurations, nous avons choisi la configuration (13-10-11-10).

Le PMC édité de 13 entrées, 10 sorties et deux couches cachées de 10, 11 neurones, respectivement, a été entrainé en utilisant l'algorithme de rétro-propagation de gradient [ROJ96]. Les entrées du réseau sont la requête de l'utilisateur, les domaines d'intérêt récemment visités et l'événement connexe s'il est détecté. Tout d'abord, nous avons traité chaque requête comme étant un ensemble de mots et nous avons essayé d'estimer le domaine cible par le biais des descripteurs de domaines. L'estimation initiale du domaine cible permet l'attribution de valeurs réelles entre [0, 1] pour le paramètre requête. Les six entrées suivantes correspondent aux domaines d'intérêt traités. La

dernière entrée correspond à l'événement lié à la requête en cours, sa valeur dépend du score RTE.

b. Phase de test

La phase de test est effectuée pour évaluer le pouvoir de généralisation des modèles neuronaux, 146 entrées sont utilisées à cette fin. Quoiqu'on ait marqué une diminution de l'erreur; cette étape ne fournit qu'une estimation de l'efficacité des modèles. En effet, l'évaluation réelle est obtenue en analysant les jugements des utilisateurs ce qui est présenté dans la section suivante.

6.5. EXPÉRIMENTATIONS ET ÉVALUATION

Après avoir crée les modèles de comportements, leur capacité à identifier le groupe de comportement approprié à l'utilisateur ainsi que ses besoins en information sont évalués. En effet, l'évaluation de ce type d'approche requiert soit l'utilisation des collections de test [HAR95 ; ALL05], soit par l'intervention des utilisateurs à travers des jugements de pertinence donnée en temps réel [LIU04; SPE05; SHE05], ce qui a été fait dans la présente contribution [BOU12a,13a], ou encore par l'utilisation des situations de recherche simulées [RUT01, 02; SIE07; RYE05], ce qui a été fait dans la deuxième contribution [BOU13b]. En effet, les résultats obtenus en procédant à une évaluation basée collection de test standard n'étaient pas significatifs [SPA05] en comparaison avec les jugements donnés par les utilisateurs en temps réel [TUR06]. Cette dernière méthode d'évaluation nécessite de mettre l'utilisateur dans son environnement de navigation naturel afin de pouvoir capturer les données de contexte qui interviennent dans l'approche et permettre ainsi l'exécution du processus d'évaluation. Dans le travail de Speretta et Gauch, [SPE05], les expérimentations ont été effectuées en sollicitant 6 utilisateurs (étudiants de l'université du Kansas[1]). Pendant une période de 6 mois, chaque utilisateur a soumis 47 requêtes au moteur de recherche Google. L'évaluation de l'approche s'effectuait sur 42 requêtes. Tandis que dans le travail de Aktas et al., [AKT06], l'évaluation a été faite par 5 utilisateurs et sur 10 requêtes.

[1] *http://www.ku.edu/*

Le but envisagé par la présente approche de modélisation est l'identification du domaine d'intérêt visé par l'utilisateur à travers la requête de recherche soumise, ceci afin de bien cerner son besoin en information. Les résultats présentés ci-dessous concernent un échantillon de 10 utilisateurs ayant différents centres d'intérêt. Les utilisateurs ont été invités à interagir avec le SRI et à conduire des recherches. Après cela, le système tente de déterminer le domaine cible pour chacune. L'évaluation contient les étapes suivantes:

1) L'utilisateur soumet une requête initiale au système qui ensuite essaye de déduire les centres d'intérêt de l'utilisateur en analysant les mots clés décrivant son besoin en information, on effectue une estimation initiale du domaine cible de l'utilisateur en utilisant les descripteurs de domaines définit dans la section (6.4.1.3).
2) Le degré d'ambigüité de la requête est calculé par l'équation (6.7) à partir du nombre de domaines qu'elle peut viser.

$$Amb(q) = \frac{\#TD}{\#D} \tag{6.7}$$

   Tel que $\#TD$ correspond au nombre de domaines probablement visés par la requêtes et $\#D$ représente le nombre de domaines traités, on a considéré ambigüe toute requête ayant un degré d'ambigüité supérieur à 0.3.
3) Les jugements de l'utilisateur sur la première estimation sont obtenus.
4) Ensuite, le système construit un vecteur de valeurs composé à partir de la requête soumise et les données contextuelles. Le vecteur créé est soumis au réseau de neurones précédemment entrainé afin d'assigner l'utilisateur actuel au modèle approprié. Cette estimation est jugée par l'utilisateur à nouveau et les résultats obtenus sont présentés dans la section suivante.

6.6. RÉSULTATS

L'évaluation a montré que sur 296 requêtes, le système a réussi dans l'identification de 206 des intérêts visés par les utilisateurs, c'est-à-dire près de 70%, dont 80% d'entre elles étaient ambigües et elles visaient plus d'un domaine d'intérêt comme le montre le tableau (6.4) et les figures (6.5 et 6.6).

On remarque que 74% des requêtes évaluées en présence de l'activité de navigation récente de l'utilisateur ont été bien classées et que leurs intérêts cible a été bien identifié. Tandis que 26% des intérêts cibles n'ont pas été identifiés, suite au fait que les requêtes de recherche ont été soumises dans des sessions de navigations caractérisées par la diversité des intérêts sollicités par l'utilisateur. En plus du fait que ces requêtes ont été introduites au milieu ou à la fin des sessions comprenant plusieurs intérêts qui ne sont pas liés à l'intérêt visé par la requête.

Nous citons un exemple de ces requêtes avec les vecteurs d'intérêts correspondants :

$I =$ {(antivirus, 22), (video converter, 18), (gratuit, 17), (avast, 15), (download, 14), (visa schengen, 8), (telecharger mp3, 8), (multimedia, 8), (assurance, 2), (informations, 2)}, $q = \{(\text{fonction differentielle}, 1)\}$.

Dans les cas où les données de navigation récentes n'étaient pas disponibles, 55% des requêtes ont été bien classées alors que, 45% ont été mal classées ce qui reflète la puissance de génération plus ou moins acceptable des modèles neuronaux en cas d'absence des données contextuelles. Il est à noter que parmi les 36 requêtes bien classées, nous n'avons marqué que 47% de requêtes ambigües comme $\{mobile, reseaux\ de\ neurones, the\ killer\}$. Tandis que 53% des requêtes sont non ambigües et elles visent un intérêt unique comme $\{gtk\ games, jeux\ pacman, elwatan\}$. Les requêtes soumises au début des sessions de navigation (démarrage à froid) qui ont été mal classées par le système constituent 45% du total des requêtes, ce qui reflète l'impact du contexte de la tâche de navigation récente dans l'identification des besoins des utilisateurs.

	#requêtes	Estimation initiale		Estimation du modèle de contexte	
		+E	-E	+E	-E
#requêtes ambiguës	252	128	124	165	87
#requêtes non ambigües	44	38	6	41	3

Tableau 6.4.Évaluation de l'efficacité des modèles utilisateur en fonction de l'ambigüité des requêtes

Après le calcul de l'ambigüité de chaque requête, nous avons marqué 252 requêtes ambigües parmi les 296 soumises. L'estimation du domaine cible retournée par le modèle a marqué un taux de succès de 70%, ce qui signifie qu'il y avait une amélioration de 14% dans l'identification du domaine cible en comparaison avec l'estimation initiale.

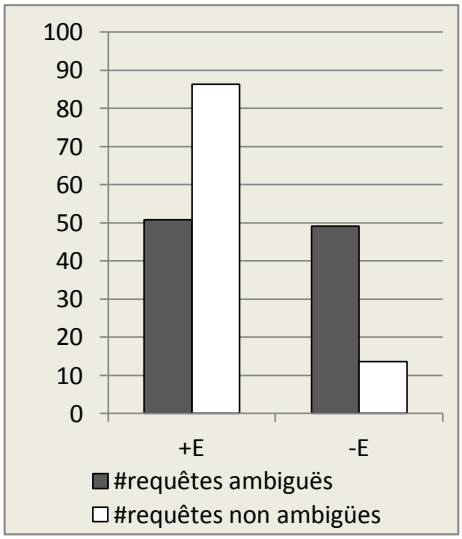

Figure 6.5.Estimation initiale du domaine visé par l'utilisateur

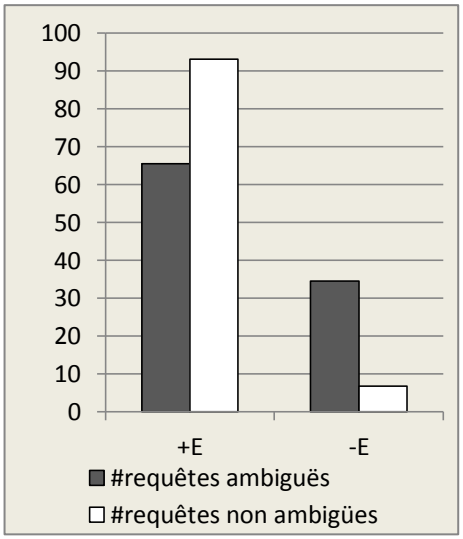

Figure 6.6.Estimation du modèle du domaine visé par l'utilisateur

Les requêtes où le système a échoué dans l'identification de leurs intérêts cibles constituent 30% du nombre total des requêtes soumises, dont 9% parmi elles ont été introduites au début des sessions de navigation. En effet, la prise en compte de l'activité de navigation récente de l'utilisateur permet de limiter l'ensemble des domaines d'intérêt susceptible d'être visés par la requête de recherche, voir le tableau (6.5) et la figure (6.7).

Il est à noter que l'évaluation a été effectuée dans une période marquée par un TP (travaux pratiques) de programmation réalisé en java, qui constitue le seul événement détecté dans la période de test. En effet, l'augmentation du nombre des requêtes connexes durant les deux semaines d'évaluation considérées a permis d'avoir un RTE qui est égal à 0.51, où le nombre des requêtes (TP java, java eclipse, download eclipse, java code) a été augmenté de 4 dans la première semaine à 10 dans la deuxième semaine. Le taux de succès d'identification des intérêts cibles de ces requêtes était de 100%.

intérêts de navigation récents	#requêtes	Les jugements des utilisateurs	
		(+)	(-)
disponible	231	170	61
non disponible	65	36	29

Tableau 6.5.L'apport de l'activité de navigation récente dans l'identification du besoin utilisateur

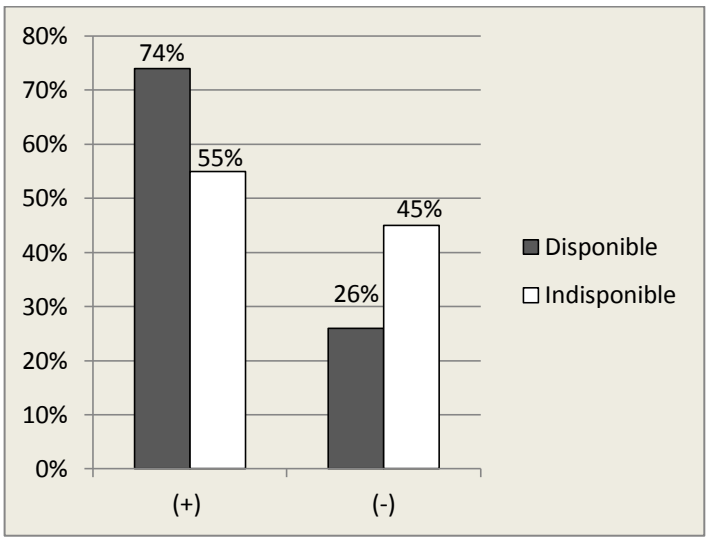

Figure 6.7.L'apport de l'activité de navigation récente dans l'identification du besoin utilisateur

6.7. CONCLUSION

L'identification du besoin en information des utilisateurs à partir des informations contextuelles est l'objectif principal de ce travail. On peut observer que la tâche de recherche dépend des événements en temps réel. Encore, les domaines d'intérêt récemment visités par l'utilisateur ont une forte probabilité de réapparaître quelque temps plus tard, ce qui peut être déduit en examinant les résultats d'identification du besoin d'utilisateur, où le taux de réussite est d'autant plus élevé quand l'historique de navigation récent est disponible.

En plus du fait que les deux dimensions du contexte utilisées dans ce travail n'ont pas été utilisées ensemble auparavant, l'événement en temps réel également n'a pas été utilisé comme une dimension contextuelle investie pour l'amélioration de l'accès à l'information sur le Web. L'intérêt récent a été utilisé dans certains travaux passés, mais dans le présent travail, il a été exploité d'une manière qui permet de prendre en compte

la diversité de sujets et de domaines d'intérêt, ce qui caractérise la session de navigation de l'utilisateur.

CHAPITRE 7 : EXPANSION CONTEXTUELLE DES REQUÊTES

7.1. INTRODUCTION

L'exécution d'une requête ambigüe par un système de recherche d'information engendre la génération d'un résultat bruité qui comporte une liste de documents dont le contenu est souvent loin de ce que l'utilisateur suggère de recevoir.

Les travaux dans le domaine de la recherche d'information contextuelle qui tentent d'améliorer l'accès à l'information pertinente sur le Web sont nombreux [ALO07; DIA09; JIN11; PAS08; BOU12a, 13a, b, BOU12b]. Dans ce domaine, l'axe d'enrichissement de requêtes a pris une partie importante de ces recherches [BOU13b; RUT03; FON05; VOO06; SON07; KAN08; WAN09; LV10]. La principale différence entre ces travaux et le présent travail, c'est que nous essayons de proposer une approche d'expansion de requêtes en considérant les aspects suivants: (1) D'une part, nous soulignons qu'il faut utiliser la session de recherche qui vise à satisfaire un besoin d'information unique comme l'unité de traitement de base. (2) La dimension temporelle a été étudiée dans plusieurs travaux [ZHA06; SAI11;BOU12b]. Ici, et en adoptant cette dimension de contexte, nous essayons d'exploiter le critère de périodicité qui peut survenir dans un grand nombre de recherche afin d'identifier les sessions de recherche similaires pouvant servir à en extraire des termes d'expansion.

7.2. CONTEXTE TEMPOREL EN RECHERCHE D'INFORMATION

Dans le domaine de la recherche d'information contextuelle, la dimension temps a bénéficié d'une part importante d'études [BOU13a, b ; BOU12; SAI11; JIN11; PAS08; ALO07; ZHA06 ; GAR05]. Cette dimension a été souvent combinée avec la dimension de localisation telle qu'il a été étudié dans les travaux de [SAI11] qui ont proposé un modèle utilisateur basé contexte pour la recommandation des films. Les éléments

contextuels sur lesquels se sont focalisés dans leurs travaux étaient le lieu de vision du film "*où un film a été vu? Au cinéma, ou à la maison*" ainsi que le temps de regarde "*quand est-ce que le film a été montré?*" en plus de l'heure de création de la notation sur le film. De la même façon, [PAN05; BIE06 ; BOU12 ; BIL08 ; HAT06] ont employé les dimensions temps et localisation en combinaison avec d'autres dimensions pour la reformulation des requêtes [HAT06], dans l'amélioration de la précision des résultats de recherche [BIL08] et leur personnalisation dans un environnement mobile [KOF03; BIE06; PAN05; BOU12].

Le facteur temps a été utilisé à différentes étapes du processus de recherche d'information [GAR05; ZHA06] dans le but d'identifier les documents pertinents et les classer en fonction de leur fraîcheur. Par exemple, pour récupérer une réponse pertinente à la requête "*coupe du monde*", il est recommandé de rechercher l'information souhaitée dans les pages Web créées en 2010 qui parlent de la dernière coupe du monde. Dans un contexte différent, [PAS08] a proposé un système de questions-réponses chargé de répondre aux questions temporelles du type *« en quelle année l'Algérie a eu son indépendance* ?» grâce à l'extraction de motifs qui ciblent l'information temporelle dans les documents en s'intéressant aux phrases qui intègrent des locutions temporelles comme *année, décennie* et *date*.

Jin et al., [JIN11] ont proposé d'élaborer un index dans lequel le temps d'actualisation et l'information temporelle incluse dans les pages Web ont été prises en compte. Alors que Alonso et al., [ALO07] ont proposé d'exploiter toutes les informations temporelles incorporées dans un document pour mieux identifier son contenu. Ils ont défini trois catégories d'expressions temporelles: les expressions temporelles explicites comme la date exacte et l'année, les expressions implicites qui sont identifiées en utilisant une ontologie du temps et les expressions temporelles relatives, qui sont des expressions qui peuvent être ancrées en se référant à une expression implicite au sein du document comme *aujourd'hui* et *la semaine prochaine*. Le facteur temps a aussi son impact sur les termes de recherche. En fait, dans la dernière décennie, *facebook, Baidu, MySpace, Wikipedia, harry potter et coupe du monde* figuraient dans le top-25 des recherches effectuées sur le Web [WEB01]. Cependant, après dix ou vingt ans, ces mots-clés n'auront certainement pas une telle popularité.

Basés toujours sur la dimension temporel, [BOU11, 12a; BOU12b] ont affiné le facteur temps en plusieurs sous dimensions. Dans [BOU11], le contexte temporel se composait du jour (Dimanche, Lundi, Mardi, Mercredi, Jeudi, Vendredi, Samedi) ainsi que le temps du jour divisé en quatre intervalles horaires : matin (de 6:00 à 11:59), après midi (de 12 :00 à 15:59), soir (de 16:00 à 21:59) et nuit (22:00 à 5:59) et dans [BOU12a], l'aspect périodique qui caractérise beaucoup de recherches Web a été exploité pour mieux identifier le besoin d'information de l'utilisateur. Tandis que dans le travail de Bouidghaghen et Tamine [BOU12b], le facteur temps a été raffiné en : saison (automne, hiver, printemps et été), jour (jour de travail, fin de semaine, jour férié et vacances) et moment de la journée (matin, midi, après-midi, soir et nuit).

De nombreux travaux ont employé les différentes dimensions contextuelles citées dans le chapitre 3, mais pour autant que nous sachions, il n'y a pas de travaux qui utilisent la combinaison des deux dimensions exploitées comme dans la présente approche à savoir: le temps représenté par le paramètre *jour* et les intérêts récents de l'utilisateur afin de cerner un ensemble de termes pertinents pouvant servir à l'expansion contextuelle des requêtes.

7.3. L'ENSEMBLE DE DONNÉES

Dans cette deuxième contribution, le facteur temps représenté par le paramètre "jour" qui prend les valeurs "jour de travail et weekend", constitue un paramètre clé pour éprouver l'aspect périodique. Ce qui implique l'emploi d'un trafic de navigation de plus d'une semaine. Pour ce faire, nous avons focalisé sur le même principe de collecte de données de navigation suivi dans la première contribution, tel qu'une activité de navigation d'une durée d'une heure par jour est offerte à un public constitué des étudiants de l'université. Pour éviter la non-disponibilité du trafic de recherche pendant les weekends, nous avons installé le plugin chargé de collecter l'historique de navigation sur 2 autres machines de bureau où le principe de navigation par heure a été respecté. Nous avons obtenu 10 MO de log durant le mois de février, 2012. Le tableau (7.1) et la figure (7.1) montrent les données obtenues représentées en nombre de sessions et de sessions de recherche respectivement par semaine. Tandis que, le tableau (7.2) et la figure (7.2) présentent le trafic de navigation selon le jour.

	semaine (1)	semaine (2)	semaine (3)	semaine (4)	Total
#sessions	21	18	24	32	95
#sessions de recherche	62	59	73	62	256

Tableau 7.1.L'ensemble de données de navigation obtenues durant le mois de Février 2012

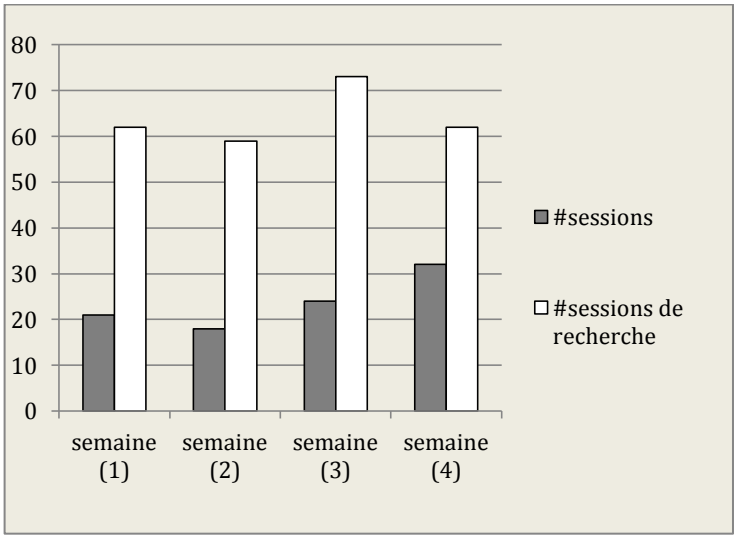

Figure 7.1. L'ensemble de données de navigation obtenues durant le mois de Février 2012

Jours de la semaine	#utilisateurs	#requêtes
Dimanche	9	60
Lundi	10	51
Mardi	8	41
Mercredi	9	34
Jeudi	11	45
Vendredi	5	10
Samedi	4	15

Tableau 7.2.Trafic de navigation en fonction du jour

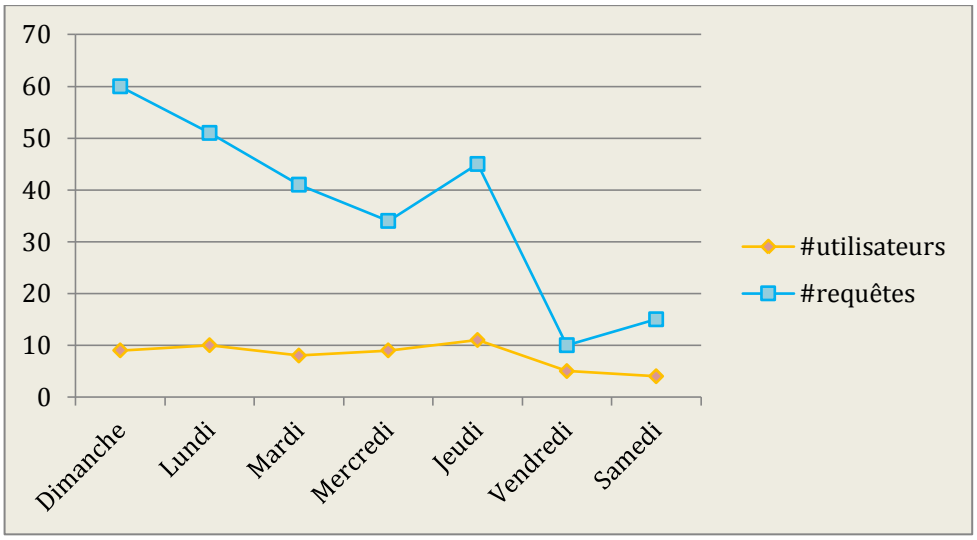

Figure 7.2. Trafic de navigation en fonction du jour

7.4. SESSION ET SESSION DE RECHERCHE

Le traitement de l'historique d'interaction des utilisateurs sur le Web vise généralement à sa subdivision en unités d'activités courtes en matière de temps et réduites en matière du volume traité, ces unités sont appelées sessions. Dans notre travail, nous utilisons les notions de session et de session de recherche différemment. Une session peut être définie comme étant une séquence de requêtes issues par un utilisateur unique dans un intervalle de temps réduit. Bien que, la session de recherche soit une séquence d'instructions logs qui traite une seule requête de recherche et vise à combler un besoin en information unique. Nous définissons formellement la session de recherche QS qui appartient à la session de navigation S par la formule (7.1) comme suit :

$$\forall S \in [t, t'], \exists QS \in \left[t, t^{"}\right] / t^{"} \leq t', QS = \{I, q, t_1, FB\} / t_1 \in \left[t, t^{"}\right] \quad (7.1)$$

Tel que QS inclue la requête unique q issue à l'heure $t_1$et une séquence de pages pertinentes retrouvées par la requête et visitées par l'utilisateur qui a soumis la requête, qui représente son retour de pertinence FB. Toute l'activité de navigation de l'utilisateur conduite avant la soumission de la requête constitue ses intérêts récents qui comprend l'activité de navigation accomplie entre $[t, t_1 - 1]$.

7.5. IDENTIFICATION DES SESSIONS

Afin de subdiviser l'historique de navigation en sessions, la méthode communément utilisée dans la littérature consiste dans la spécification d'un seuil global pour toutes les sessions et en fonction de ce seuil les sessions sont identifiées. Sa durée diffère de 5 à 30 minutes [SIL99 ; HE00]. Tandis que Huynh et Miller [HUY09] ont proposé un modèle mathématique pour identifier la longueur des sessions.

Le choix du seuil a son impact sur l'analyse du comportement utilisateur. En effet, ceci permet de déterminer la longueur des sessions, le nombre de requêtes dans chaque session ainsi que les activités accomplies dans chacune. De ce fait, et pour assurer qu'un flux de navigation parvient d'un utilisateur unique et pour réduire le problème des sessions longues contenant des intérêts multiples et hétérogènes, nous avons choisi un seuil égal à 30 minutes. Comme notre objectif est l'amélioration de la recherche sur le Web, nous étions intéressés par les sessions qui incluent au moins une requête de recherche.

7.6. SUBDIVISION DES SESSIONS EN SESSIONS DE RECHERCHE

Bien que la session de navigation inclue un ensemble d'activités de navigation pouvant concerner plusieurs domaines d'intérêt, par exemple la session S prise de l'ensemble de données explore plusieurs intérêts sont : le sport, les news, les mathématiques et les multimédias. Davantage, nous avons constaté que la focalisation sur de telles sessions pour faire la modélisation engendre la production d'une quantité de classes importante. De ce fait, nous avons été guidés par l'idée de traiter l'activité de recherche uniquement et non pas l'activité de navigation dans son intégralité. Effectivement, ce que nous avons admis c'est que le comportement de navigation y compris les recherches effectuées d'un utilisateur intéressé par un domaine métier A et celui d'un utilisateur intéressé par un domaine métier B, peuvent être similaires à propos des intérêts communs C, D et E. Ces recherches communes peuvent être conduites par le biais de requêtes proches visant les mêmes besoins informationnels [BOU13b]. Partant des points cités, nous avons vu que le traitement d'une seule requête de recherche à la fois s'avère plus fructueux pour notre approche d'expansion qui est basée essentiellement sur le contexte immédiat de la recherche. Ce contexte comprend la dimension temps

représentée par le paramètre type de jour et la tâche de navigation immédiate représentée par les documents récemment visités. La session de recherche a une durée variable, elle commence du début de la session de navigation et elle se termine par le dernier document visité des résultats retrouvés par la requête. Par exemple, soit S la session de navigation composée d'un ensemble d'activités de navigation effectué dans l'intervalle de temps $[t, t']$. Soit la requête q soumise en t_1, la session de recherche QS qui traite la requête unique q et qui est définie dans l'intervalle $[t, t"]$ est composée des intérêts récents de l'utilisateur représentés par la formule (7.2), en plus du retour de pertinence des résultats de la requête FB qui est défini par la formule (7.3).

$$\mathrm{I} = \{(\mathrm{p}, \mathrm{C}, \mathrm{Tv})_\mathrm{i} \ / \mathrm{i} = \ 1..\mathrm{n}, (\mathrm{p}, \mathrm{C}, \mathrm{Tv})_\mathrm{i} \in [\mathrm{t}, \mathrm{t}_1 - 1]\} \tag{7.2}$$

$$\mathrm{FB} = \left\{(\mathrm{p}, \mathrm{C}, \mathrm{Tv})_\mathrm{j} \ / \mathrm{j} = \ 1..\mathrm{m}, (\mathrm{p}, \mathrm{C}, \mathrm{Tv})_\mathrm{j} \in \left[\mathrm{t}_1 + 1, \mathrm{t}"\right]\right\} \tag{7.3}$$

La longueur des sessions de recherche est comprise entre une minute et trente minutes, ce qui représente la durée de la dernière session de recherche effectuée au sein de la session de navigation qui admet toute l'activité de navigation qui précède la requête comme étant des intérêts à court terme. Cependant, la majorité des recherches proprement dites qui commencent par l'instant de soumission de la requête jusqu'à la soumission de la prochaine requête ont des durées comprises entre 4 et 8 minutes comme il le montre la figure (7.3).

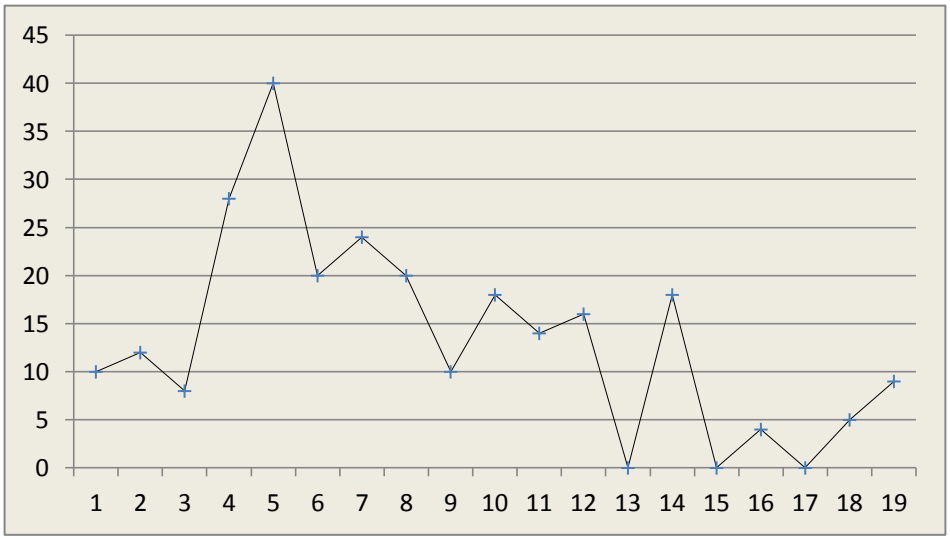

Figure 7.3.Longueurs des sessions de recherches

Comme nous le constatons, 51% des sessions de recherche ont une durée comprise entre 4 et 8 minutes. Alors que, 11% d'entre elles ont une longueur comprise entre 1 et 3 minutes, ce qu'on a considéré comme des sessions de recherche courtes. Tandis que, plus de 36% des recherches effectuées avaient une durée plus longue et qui dépasse les 8 minutes.

7.7. REPRÉSENTATION DES SESSIONS DE RECHERCHE

Chaque session de recherche dispose de deux représentations, une terminologique et l'autre linéaire.

7.7.1. Représentation terminologique

Dans un premier temps, chaque session de recherche est représentée comme une liste de termes extraits des pages pertinentes visités au sein de la session. L'algorithme *terminology-QS* permet de déterminer la pertinence des pages visitées en cours de la session de recherche QS. Ensuite, les termes déterminant du contenu de chaque page sont extraits en utilisant Gate[1]. Sauf les dix premiers termes ayant des scores de fréquence élevés sont conservés.

Le retour de pertinence de la requête q_i inclut les pages visitées durant la période entre la soumission de q_i et la requête q_j soumise immédiatement après. Pour s'assurer que la page fait partie du feedback de la requête traitée, nous avons utilisé une fonction $topic(x)$ ayant pour but de vérifier si la page contient les termes de la requête.

Algorithme 2. terminology-QS

Entrées : QS={I, q,t_1,FB}∈[t,t"]

Sortie : QS_t//vecteur de termes

DEBUT

Pour chaque page p_i visitée dans l'intervalle de temps[t,t"]Faire

Identifier la pertinence de p_i $(is - relevant(p_i))$

Ordonner les pages selon leurs degrés de pertinence

Considérer les deux premiers tiers comme pertinentes

[1] *Site du logiciel: http://gate.ac.uk/*

Pour chaque page pertinente $p_i \in [t, t_1\text{-}1]$ faire

$I_t \leftarrow$ termes décrivant p_i

Pour chaque page pertinente $p_j \in [t_1, t'']$ faire

début

si $topic(q) = topic(p_j)$ alors

$FB_t = \leftarrow$ termes décrivant p_j

fin

$QS_t \leftarrow \{I_t, q_t, FB_t\}$;

Retourner QS_t

FIN

La représentation terminologique d'une QS est obtenue à partir de l'union des deux vecteurs de termes I_t et FB_t ainsi que les termes de la requête. La fonction $topic(x)$ retourne le domaine susceptible d'être visée par une requête ou aussi le thème d'une page web en se référant aux descripteurs de domaine. Après cela, l'ensemble des pages visitées au sein de la session de recherche QS_x est classé selon le degré de pertinence des pages où les deux tiers premières pages pertinentes sont conservées. La pertinence d'une page visitée lors d'une session de recherche est retournée par la fonction is-relevant(x) en se servant de l'équation $R(p, QS)$ donnée par l'équation (6.5).

7.7.2. Représentation Linéaire

La représentation linéaire est dérivée de la représentation terminologique, l'étape de transformation du vecteur de termes en vecteur binaire consiste en la préparation des données à l'étape de classification. Cette représentation permet de construire un vecteur de 20 valeurs qui sont définies dans l'intervalle binaire [0,1]. Les six premières valeurs correspondent aux domaines traités dans ce travail, la valeur 1 signifie que le domaine apparaît dans l'activité de navigation récente de l'utilisateur et 0 autrement. Les six valeurs suivantes correspondent à une estimation initiale du domaine cible de la requête. Ensuite, les deux valeurs suivantes permettent d'introduire le paramètre de temps qui permet de vérifier l'aspect périodique en précisant le type de jour. La dernière séquence de valeurs correspond au feedback de pertinence de l'utilisateur sur les résultats

renvoyés lors de la recherche, telle que la valeur 1 est attribuée aux domaines visités et jugés pertinents par l'utilisateur.

7.8. CLASSIFICATIONS DES SESSIONS DE RECHERCHE

L'idée sur laquelle nous nous sommes appuyés dans la classification basée session de recherche demeure sur le fait que cette dernière vise à satisfaire un besoin en information unique et que l'activité de navigation qu'elle inclut est limitée au sujet de la requête. La classification basée session de recherche permet d'éviter d'avoir un nombre important de classes qui peuvent se produire en faisant une classification basée utilisateur [GAR05; MOG11; QIU06], ou une classification basée session [DAO07; SCH04], car il n'est pas évident de retrouver un nombre important de sessions similaires. Qui plus est, si ce cas se présente il n'est pas possible d'identifier si ce flot d'activités de navigation appartient à un utilisateur unique.

Afin d'avoir un nombre de classes manipulable, nous nous sommes dirigés à faire un regroupement des sessions de recherche similaires en ensembles séparés appelés $SimQS_i$ en utilisant la mesure de similarité cosinus donnée par l'équation (2.5). Ce qui nous a permis d'éviter l'obtention d'un plus grand nombre de groupes produits en utilisant une stratégie de classification non supervisée sans faire le regroupement.

Soit S une session de navigation comprenant au moins une requête q:

- Nous avons divisé S en k sessions de recherche dont chacune traite une seul requête de recherche $S = \langle\langle I_1, q_1, t_1, FB_1\rangle, \ldots \langle I_k, q_k, t_k, FB_k\rangle\rangle$, tel que $(\mathrm{k} \geq 1)$ est le nombre de requêtes dans la session.

- Nous avons préparé à cette étape chaque classe de sessions de recherche $SimQS_j$ pour l'étape de classification supervisée en utilisant le perceptron multicouches.

7.9. APPRENTISSAGE AUTOMATIQUE

Dans nos expériences, nous avons opté pour une stratégie d'apprentissage supervisé en utilisant le perceptron multicouches, qui est l'un des classifieurs permettant l'atteinte de taux de précision importants dans les problèmes de classification [CIA08; KLA10]. De plus, il nous a généré de bons résultats d'identification des besoins utilisateurs dans

notre première contribution. Le réseau créé a été utilisé pour l'apprentissage des sessions de recherche regroupées par leurs degrés de similarité sous forme d'ensembles distincts de sessions de recherche similaires. L'ensemble de données utilisé comprend 256 entrées, correspondant chacune à une session de recherche.

L'ensemble de données est subdivisé en apprentissage et test. L'ensemble d'apprentissage est composé de 170 entrées consacrées à la construction du modèle de contexte. Tandis que l'ensemble de test contient 86 entrées qui ont été utilisés pour estimer l'erreur de test. La configuration du réseau de neurones comprend 32 entrées, telle que la première séquence de 14 valeurs correspond aux entrées du réseau et la dernière séquence représente les sorties désirées, en plus de deux couches cachées de 16, 15 neurones respectivement. L'apprentissage se fait par le biais de l'algorithme de rétro-propagation de gradient.

7.10. L'ENRICHISSEMENT DES REQUÊTES

L'enrichissement des requêtes signifie l'ajout de nouveaux termes à la requête initiale pour lever son ambigüité et augmenter ainsi le nombre de réponses pertinentes.

Comme susmentionné, le retour de pertinence et les requêtes similaires passées ont été traités et proposés largement comme source pour l'enrichissement des requêtes [FON05; RUT03 ; WAN09; SON07; KAN08; VOO06; LV10]. Dans le présent travail, le choix des termes d'expansion dépend de l'ensemble d'activités de navigation récentes de l'utilisateur, l'idée sous-jacente consiste à suggérer des termes qui co-occurrent contextuellement avec les termes des requêtes similaires ayant le même contexte de recherche. C'est-à-dire, l'ensemble des termes qui co-occurrent souvent dans des sessions de recherche similaires sont utilisés pour étendre les requêtes des utilisateurs comme il est expliqué par les étapes suivantes :

- Partant de la représentation terminologique de chaque session de recherche $\mathrm{QS_t}$ nous avons regroupé les QS similaires en groupes séparés appelés $\mathrm{SimQS_i}$ en utilisant la mesure de cosinus.

- Soit $SimQS_j = \{QS_1, \dots, QS_n\}$ un ensemble de sessions de recherche similaires et soit $V_1, \dots, V_n$ est l'ensemble des vecteurs de termes décrivant l'activité de

recherche dans $\{QS_1, \ldots, QS_n\}$ respectivement, où chaque vecteur $V_i$est composé des termes de la requête $q_i \in QS_i$ en plus des termes extraits depuis les retours de pertinence des utilisateurs sur les pages retrouvées FB_i, depuis chaque vecteur $V_i = \langle w_j / w_j \in q_i \vee w_j \in FB_i \rangle$ on construit la matrice des termes co-occurrents correspondante M_n.

- Pour chaque nouvelle requête q_i issue dans une session de recherche QS_j, on identifie la classe ou le modèle approprié.

- Ensuite, on extrait depuis la matrice de cooccurrence correspondante les 5 premiers termes ayant une haute fréquence de cooccurrence avec les termes de la requête, et les utilisés pour l'expansion de celle-ci.

7.10.1. Matrice de cooccurrence

Une matrice de cooccurrence de termes est à l'origine une matrice carrée de N lignes et N colonnes, où les colonnes et les lignes correspondent aux termes fréquents qui sont extraits depuis la source de données traitée. Cette source peut être une phrase, un document, ou une collection de documents.

Dans cette phase du travail et après avoir regroupé les sessions de recherche similaires au sein d'un même groupe $SimQS_i$ en utilisant la mesure de cosinus qui nous a permis d'obtenir 18 $SimQS_i$ différents. Nous avons construit pour chacune d'elles la matrice de cooccurrence correspondante.

Soit $SimQS_{\mathrm{i}}$ un ensemble de sessions de recherche similaires et soit n le nombre de termes dans $SimQS_{\mathrm{i}}$. La matrice de cooccurrence notée par $\mathrm{M_n}$correspond à une matrice carrée de n lignes et n colonnes où n prend une valeur qui correspond au nombre de termes dans chaque classe de sessions de recherche. Nous avons attribué à n une valeur fixe après avoir réduit le nombre de termes de chaque classe à 245, le principe de réduction consiste à éliminer de chaque matrice les termes ayant des relations de cooccurrence avec une minorité de termes. Chaque ligne et chaque colonne de la matrice correspond à un terme de la classe, et l'entrée de la x^{ieme} ligne et la y^{ieme} colonne représente le nombre de fois le $x^{ième}$ terme apparaît avec le y^{ieme} terme. Le tableau (7.3)

montre un exemple de termes qui co-occurrent dans la SimQS_1 triés dans un ordre de fréquence décroissant.

Paire	Fréquence
(tp-pascal)	27
(exercice-pascal)	20
(exercice-algorithme)	19
(pascal-programme)	18
(algorithm-pascal)	18
(turbo-pascal)	15
(programme, java)	10

Tableau 7.3.Les co-occurrents fréquents dans $SimQS_1$

7.11. ÉVALUATION

Traditionnellement, la qualité des résultats d'un SRI se mesure en comparant la réponse du systtème avec celles que l'utilisateur attend de recevoir, la pertinence de ces résultats est déduite depuis les valeurs de mesures de rappel et de précision communément utilisées en recherche d'information. Le rappel mesure la capacité d'un système de recherche d'information à localiser les documents pertinents dans l'index du système. Tandis que la précision mesure sa capacité de ne pas renvoyer les documents non pertinents.

La mesure du rappel requiert le calcul pour chaque requête le nombre total de documents pertinents au sein du système de recherche d'information, ce qui n'est pas permis au sein d'une masse de documents énorme tels que le Web. Dans ce cas, la précision à des points spécifiques s'envisage comme étant une mesure alternative permettant le calcule de documents pertinents au sein d'un ensemble de documents spécifique. Cette mesure est adéquate pour l'évaluation de la présente contribution qui traite la recherche d'information sur le Web. Pour cette fin, nous avons défini un ensemble de scénarios stimulants des situations de recherche à travers l'introduction de sessions de recherche réelles effectuées sur trois systèmes de recherche d'information en ligne à savoir : Google, Aol et Altavista. La pertinence thématique a été mesurée en matière de précision à 5, 10 et 15 premiers documents (cette mesure a été discutée dans

la section 2.5.3.3). Nous nous sommes appuyés dans le choix des points de calcul sur les comportements de navigation étudiés dans ce travail, où nous avons constaté que les utilisateurs n'avancent pas dans la consultation des résultats au-delà des 15 premiers documents. Dans beaucoup d'autres travaux, il a été démontré que les utilisateurs ne consultent que les liens présentés en haut de la liste des résultats [KEA08 ; JAN06a ; LOR08 ; SAI10].

Tout d'abord, nous avons identifié les documents pertinents à partir du feedback de pertinence fournis avec les situations de recherche employées pour l'évaluation [RUT01, 02; SIE07]. Après cela; l'algorithme d'expansion est exécuté et les termes sélectionnés sont utilisés pour enrichir la requête avant qu'elle soit soumise au moteur de recherche à nouveau. La précision à 5, 10 et 15 premiers documents est calculée et les résultats obtenus avant et après expansion sont représentés dans la figure (7.4) et tableaux (7.4 ; 7.5 ; 7.6).

Precision	P@5	P@10	P@15	MAP
Avant expansion	0.58	0.55	0.53	0.29
Apres expansion	0.59	0.57	0.57	0.33

Tableau 7.4.Précision P@n et précision moyenne avant et après l'expansion des requêtes soumises à Google

Précision	P@5	P@10	P@15	MAP
Avant expansion	0.57	0.57	0.49	0.25
Après expansion	0.60	0.59	0.55	0.31

Tableau 7.5. Précision P@n et précision moyenne avant et après l'expansion des requêtes soumises à Aol

Precision	P@5	P@10	P@15	MAP
Avant expansion	0.52	0.46	0.43	0.26
Après expansion	0.55	0.53	0.53	0.35

Tableau 7.6. Précision P@n et précision moyenne avant et après l'expansion des requêtes soumises à Altavista

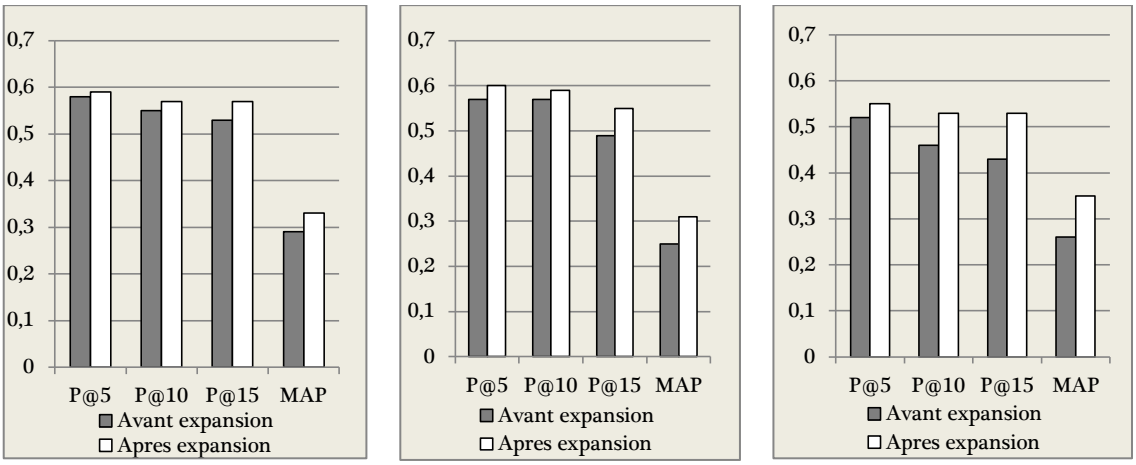

Figure 7.4. Précision P@n et précision moyenne avant et après l'expansion des requêtes soumises à Google, Aol et AltaVista respectivement

Les résultats de l'évaluation de la démarche d'expansion des requêtes montrent qu'il y a une amélioration de la pertinence des résultats de recherche. Cette amélioration augmente de 0.1 dans les 5 premiers documents à 0.4 dans les 15 premiers documents en ce qui concerne les résultats de Google. Tandis que pour Aol, la précision marquée était de 0.4. Concernant AltaVista, la précision augmente de manière considérable de 0.3 dans les 5 premiers documents à 0.6 dans les 15 premiers documents.

Les sessions de recherche évaluées sont en nombre de 58 où la pertinence se diffère pour chacune car les termes utilisés dans l'expansion ont des degrés de pertinence divergents par rapport aux termes de la requête, ce qui est dû au fait que la fréquence de cooccurrence de deux termes dans un contexte donné n'indique pas toujours que chacun peut être utile pour lever l'ambiguïté de l'autre. Par exemple, l'enrichissement de la requête « *pascal* » par les termes {tp, exercice, algorithm, programme et turbo} a été fructueux et a donné des résultats pertinents par les trois moteurs. Cependant, les termes utilisés pour l'expansion de la requête "*sport news"* n'avaient pas tous la même adéquation avec le besoin en information de l'utilisateur. En effet, la requête étendue comprenait les termes : {football, europeen, ligue des champions, actualite}. Ici les termes { actualite, europeen } avaient un degré de pertinence et d'expressivité acceptable pour la requête néanmoins les termes { football, ligue des champions} n'avaient pas le même degré de pertinence. En effet, l'ajout d'un ensemble de termes d'expansion ayant des degrés de pertinence hétérogènes engendre l'augmentation du taux de bruit dans les résultats retournés. Ce qui diminue les taux de précision marqués.

L'augmentation repérée dans le taux de précision indique qu'il y avait une amélioration significative dans la pertinence d'un nombre important des résultats des requêtes enrichies.

7.11.1. Apport de la dimension temps

Le facteur temps représenté dans ce travail par le paramètre type de jour a eu un impact important dans l'amélioration de pertinence contextuelle. Ceci a été prouvé par la séparation des requêtes en deux types différents qui sont: les requêtes métiers (*TQ*) qui sont celles qui ciblent l'un des domaines d'intérêt techniques traités dans ce travail; et les requêtes génériques (*GQ*) qui ciblent l'un des domaines d'intérêt communs traités.

Le système a réussi dans l'enrichissement de 40% des requêtes techniques soumises dans les weekends, et 64% des requêtes génériques. Dans le reste des jours de la semaine, le taux de bonnes expansions était en faveur des requêtes techniques avec 75%. Les résultats obtenus montrent que la prise en compte du type de jour peut être très effective tant que l'aspect périodique est envisagé, le tableau (7.7) et la figure (7.5) présentent respectivement les résultats obtenus.

Type du jour	#TQ	#ITQ(+)	Taux de succès (%)	#GQ	#IGQ(+)	Taux de succès (%)
Weekend	3	2	66,66	4	3	75
Jour de semaine	46	36	78,26	5	3	60

Tableau 7.7.Évaluation de l'impact du facteur temps

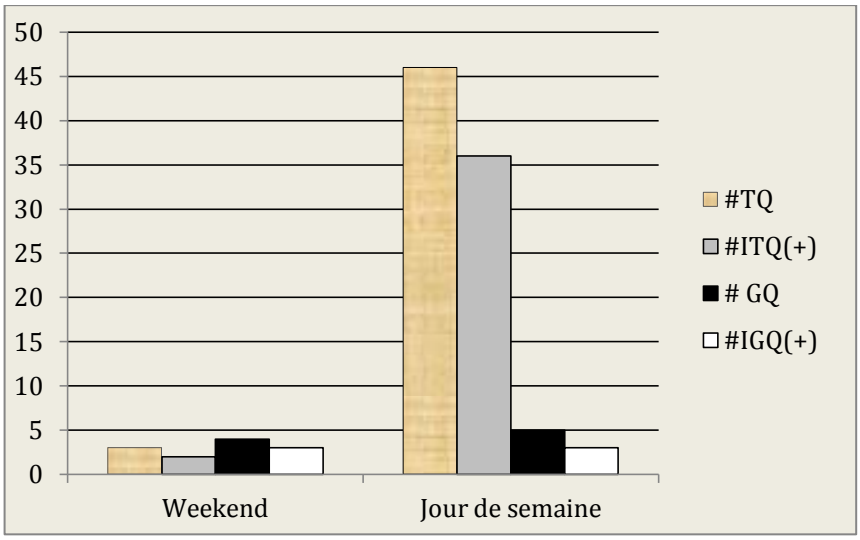

Figure 7.5.Évaluation de l'impact du facteur temporel

7.12. CONCLUSION

L'approche d'expansion présentée dans ce chapitre repose sur le contexte de l'activité de navigation immédiatement capturée ainsi que le contexte temporel dans une perspective de suggérer une expansion des requêtes en fonction du contexte dans lequel elles sont soumises. La méthode proposée a permis d'aider l'utilisateur à mieux exprimer son besoin d'information et d'améliorer en outre les résultats de la recherche sur le Web. Notre étude a porté sur le type de jour et les intérêts récents de l'utilisateur qui représentaient les paramètres exploités dans la modélisation contextuelle.

Les résultats de l'approche ont permis de démontrer que la pertinence des termes d'expansion est relativement dépendante de la tâche de recherche impliquée dans le système de recherche.

Également, plus le contexte terminologique est raffiné, davantage la richesse de vocabulaire et des sujets traités sont évités, ce qui influe positivement la qualité des termes d'expansion.

7.13. SYNTHESES DES DIFFÉRENTS TRAVAUX

Dimension du contexte	Contributions
Temps	– Prise en compte du temps de création du document dans le classement des résultats de la requête [SAI11 ; QAM06] – Prise en compte des données temporelles incluses dans les documents (date explicite, locutions temporelles, etc.) pour déterminer les documents pertinents. [PAS08 ; JIN11 ; ALO07] – Temps de soumission de la requête pour améliorer la recherche dans un environnement mobil. [BOU12b].
Tâche	– Prise en compte de la tâche ou l'activité en cours d'accomplissement (projet, dissertation, …,etc) pour déterminer le besoin stimulant la recherche [SHE05 ;FRE05 ;ASF09 ;12].
Évènement	/

Notre contribution	Travaux récents
– Apport de l'aspect périodique dans l'identification du besoin d'information de l'utilisateur. [BOU13b] – périodicité appliquée selon le type de jour (jour de semaine, weekend). [BOU13b] –	– [LIN14] proposent un algorithme de classement de documents basé sur la pertinence temporelle et textuelle en se référant aux expressions temporelles présentes dans les documents.
– Exploitation des intérêts utilisateur à court terme dans une période de temps qui varie entre 4 et 8 minutes pour construire un ensemble de terme liés à la tâche de navigation en cours et proche du besoin actuel de l'utilisateur. – Identifier l'intérêt cible de l'utilisateur en s'aidant de la tâche de navigation récente de l'utilisateur. [BOU11, 12a, 13a,b]	– Pour améliorer l'accès et l'échange d'informations et la satisfaction d'un besoin d'information commun dans les milieux de travail du monde réel, [BOH14] présentent les premiers résultats empiriques dans un projet de recherche en cours qui étudie l'utilisation des technologies logicielles utilisées au sein du groupe pour réaliser la collaboration, la recherche et le partage d'information.
– Détecter la présence d'un événement depuis les tendances de recherche de l'utilisateur et utiliser ce contexte pour identifier le domaine cible de la requête. [BOU12a,13a]	– Un plan de génération de calendriers qui permet de capturer les événements les plus saillants dans le domaine sportif avec les mots-clés et les commentaires les plus populaires. Cette approche qui permettra selon [ALO13] de présenter des informations très pertinentes aux amateurs de sport qui sont intéressés aux détails des conversations lancées sur l'événement a été appliquée aux données de twitter. [ALO13] – [SAH14] propose d'améliorer la recherche des vidéos liées à des événements fouillées à partir des tendances de recherche sur le Web, chaque événement est représenté par un ensemble de tags liés aux adresses des vidéos pertinentes.

Tableau 7.8. Synthèse des différents travaux

CONCLUSION GÉNÉRALE

L'évolution de la recherche d'information est intimement liée à l'évolution du Web qui a permis l'accès à diverses sources d'information. Bien que, l'apport du contexte a été considérablement étudié, le développement abondant du hardware et du software, ouvre des voies à explorer de nouvelles dimensions contextuelles.

Le travail abordé dans cette thèse vise à contribuer dans l'amélioration de l'accès à l'information pertinente sur le Web. En effet, la démarche mise en œuvre permet l'obtention de bons résultats d'identification des besoins d'information de l'utilisateur et en outre de lui aider à mieux exprimer ce besoin, ceci, en vue d'améliorer les résultats renvoyés. L'approche suivie s'appuie notamment sur le contexte dans lequel s'effectue la recherche, où nous nous sommes intéressés à la tâche de navigation actuelle de l'utilisateur, le facteur temporel ainsi que l'événement pouvant affecter le comportement de navigation de l'utilisateur du Web.

- Efficacement, la dimension tâche de navigation immédiate a été exploitée en vue de déterminer les centres d'intérêt courants de l'utilisateur et pouvoir identifier ainsi le besoin de recherche actuel. Nous avons vu que l'étude des intérêts de l'utilisateur à court terme permet d'obtenir des informations contextuelles en temps réel liées à l'état du besoin informationnel actuel de l'utilisateur, ceci peut aider à rapprocher l'information désirée de son demandeur.

- En d'autres termes, nous avons saisi l'intérêt d'étudier le contexte de la recherche dans une période de temps courte et proche du temps de soumission de la requête afin de borner le nombre d'intérêts susceptibles d'être visés par l'utilisateur dans une période de temps plus longue ainsi que le changement des intérêts au fil du temps.

- Partant de l'heuristique qui dit que les mêmes requêtes se produisent souvent dans les mêmes jours, l'exploitation de la périodicité, qui caractérisent un nombre important de requêtes sur le Web a contribué positivement dans l'obtention de ces résultats.

- La présence d'une relation de dépendance recherche-événements agissants un peu partout à travers le monde explique la baisse ou l'augmentation de la fréquence des requêtes connexes. La prise en compte de cette particularité a aidé à bien cerner le domaine d'intérêt visé par l'utilisateur en période de déroulement d'un événement donnée.

Qui plus est, le comportement de navigation en général et de recherche en particulier des utilisateurs en Algérie est similaire au comportement de tout autre utilisateur de la toile à travers le monde, et il se caractérise par :

- La périodicité des sujets d'intérêts consultés en fonction du type de jour, tel que nous avons prouvé que les requêtes similaires se produisent souvent dans les mêmes jours. Aussi, les requêtes qui concernent des sujets d'intérêts métiers qui sont souvent de type informationnel se produisent fréquemment dans les jours de travail, alors que les requêtes sur des sujets d'intérêt communs qui peuvent être classifié comme étant des requêtes transactionnelles ou navigationelles se produisent le plus souvent dans les weekends.

- l'influence par le contexte événementiel qui a un impact déterminant sur l'activité de navigation de l'utilisateur.

LIMITES ET PERSPECTIVES

Au terme de ce travail, nous pouvons donc dire que l'état, dans lequel l'utilisateur effectue une recherche sur le Web a un effet perceptible sur son comportement de recherche. Pour cette raison, le contexte de recherche doit être étudié plus en vue de découvrir de nouvelles dimensions pouvant aider à comprendre davantage le comportement de l'utilisateur du Web, et améliorer par conséquent la recherche d'information.

De plus, le degré de fréquence seule ne peut pas être un paramètre efficace pour choisir les termes pertinents de l'ensemble des cooccurrences. Pour surmonter ce problème et d'améliorer la précision, nous nous prévoyons d'affiner chaque groupe comportemental et d'étendre leur nombre à plus de 18 et subdiviser chaque domaine en sous-domaines

afin d'obtenir une terminologie solide au sein du groupe et améliorer encore la qualité des termes d'expansion.

Les problèmes majeurs liés à l'évaluation contextuelle sont dus à la difficulté de créer des bases de test pouvant comprendre toutes les interactions possibles de l'utilisateur avec un SRI ainsi que l'incapacité de couvrir les différentes dimensions contextuelles ce qui est quasiment difficile vu la variabilité des utilisateurs et la diversité des centres d'intérêt en plus de la variation du besoin informationnel en fonction du contexte courant.

Le domaine qui reste à explorer serait, d'abord, les dimensions de contexte abordées dans ce travail qui peut être encore étendues et combinées avec d'autres dimensions. Nous essayons d'affiner plus chaque groupe de comportement afin d'améliorer la qualité des relations de cooccurrence entre les termes d'un même groupe.

Concernant les événements (régionaux, mondiaux) affectant le comportement de navigation de l'utilisateur en général et de recherche, leur apport reste à être bien étudié dans l'axe d'expansion, dont nous sommes en train de développer.

Nous envisageons également, d'employer l'ontologie générique WordNet dans l'estimation initiale du domaine cible de la requête. De plus, nous avons pensé à enrichir la base de données de navigation afin d'élargir l'application des approches proposées à plus de six domaines d'intérêt et pour étendre le test sur un ensemble d'intérêts plus large.

PRODUCTIONS SCIENTIFIQUES

Boughareb, D. Farah, N. (2011). Toward a Web Search Personalization Approach Based on Temporal Context. In H., Cherifi, J., Mohamad Zain, E.,El-Qawasmeh (Eds.): Digital Information and Communication Technology and Its Applications– International Conference, DICTAP 2011. Proceedings, Part I. Springer 2011 Communications in Computer and Information Science ISBN 978-3-642-21983-2, pp.33-44.

Boughareb, D., Farah, N. (2011). Une Approche de Modélisation de l'Utilisateur Basée sur le Contexte Temporel pour la Personnalisation de la Recherche d'Information, In the International Workshop on Information Technologies and communications (WOTIC) 2011. Casablanca-Morocco.

Boughareb, D. Farah, N. (2012). Contextual Modelling of the User Browsing Behaviour to Identify the User's Information Need. Proceeding of INTECH'12. Casablanca-Morocco, pp. 268-273. IEEE, DOI: 10.1109/INTECH.2012.6457773.

Boughareb, D., Farah, N. (2012). Modélisation Contextuelle du Comportement de Navigation de l'Utilisateur pour Améliorer la Recherche sur le WEB. In the 2nd. International Symposium ISKO-Maghreb. Tunisia.

Boughareb, D. Farah, N. (2013). Identify the User's Information Need Using the Current Search Context, International Journal of Enterprise Information Systems (IJEIS). DOI: 10.4018/ijeis.2013100103. IGI global publishing. 9(4):30-45.

Boughareb, D., and Farah, N. (2013). A Query Expansion Approach Using the Context of the Search, Advances in Intelligent and Soft Computing (Springer), Book chapter. In the 4th International Symposium on Ambient Intelligence, Salamanca-Spain ISami'13, Volume 219, DOI. 10.1007/978-3-319-00566-9_8. ISBN. 978-3-319-00566-9, pp. 57-63.

BIBLIOGRAPHIE

[ABO97] Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R. and Pinkerton. M. (1997). Cyberguide: A mobile context-aware tour guide. Wireless Networks, 3(5),pp. 421–433.

[AIR04] Airio,E., Järvelin, K., Saatsi, P., Kekäläinen, J., Suomela, S. (2004). CIRI - An Ontology-based Query Interface for Text Retrieval. In Hyvönen, E, Kauppinen, T, Salminen, M, Viljanen, K and Ala-Siuru, P, eds. Web Intelligence. Helsinki: Finnish Artificial Intelligence Society, pp. 73-82.

[AKT06] Aktas M., Nacar M., and Menczer F. (2006). Personalizing PageRank based on domain profiles. Advances in Web Mining and Web Usage Analysis, Proc. 6th SIGKDD Workshop on Web Mining and Web Usage Analysis WebKDD 2004, volume 3932 of LNAI, Springer, pp. 104-115.

[ALF09] Alfonseca, E., Ciaramita, M. and Hall, K. (2009). Gazpacho and summer rash: Lexical relationships from temporal patterns of Web search queries. In Proceedings of EMNLP 2009, pp.1046-1055.

[ALF14] Alfred, R., Chin, K-O., Anthony, P., San, P-W., Im, T-L., Leong, L-C., Soon, G-K. (2014). Ontology-Based Query Expansion for Supporting Information Retrieval in Agriculture. The 8th International Conference on Knowledge Management in Organizations, Springer Proceedings in Complexity, pp. 299-311.

[ALH13] Alhamid, MF., Rawashdeh, M., Al Osman, H., El Saddik, A. (2013). Leveraging biosignal and collaborative filtering for context-aware recommendation. Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, pp.41-48.

[ALL05] Allan. J. (2005). Hard track overview in TREC 2005 high accuracy retrieval from documents. In Proceedings of the Fourteenth Text Retrieval Conference, TREC 2005, pp.1-11.

[ALL03] Allan J. and al (2003). Challenges in information retrieval and language modeling. Report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst,. SIGIR Forum, 37(1), pp. 31–47.

[ALO13] Alonso, O. and Shiells, K. (2013).Timelines as Summaries of Popular Scheduled Events.International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media. WWW 2013 Companion, Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05.pp.1037-1044.

[ALO07] Alonso, O., Gertz, M., Yates, R.B. (2007). On the value of temporal information in information retrieval. In: Proc. of SIGIR'07, pp. 35–41.

[ASF12] Asfari O., Doan B-L., Bourda Y., Sansonne J-P.(2012). Personalized Access to Contextual Information by using an Assistant for Query Reformulation: International Journal on Advances in Intelligent Systems 4, 3-4, pp.128-146.

[ASF09] Asfari O., Doan B.-L., Bourda Y. and Sansonnet J.-P. (2009). Personalized Access to Information by Query Reformulation Based on the State of the Current Task and User Profile. Third International Conference on Advances in Semantic Processing, pp. 113-116.

[BAC10] 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.

[BAE04] Baeza-Yates, R. Carlos, A., Hurtado, A., Mendoza, M. (2004). Query Recommendation Using Query Logs in Search Engines. EDBT Workshops, pp. 588-596.

[BAE99] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.Bai, J., Nie, J.-Y., and Cao, G. 2006. Context- dependent term relations for information retrieval.In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Sydney, Australia, pp.551– 559.

[BAI05] Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In Proceedings of the 14th ACM international conference on Information and knowledge management. ACM Press, Bremen, Germany, pp. 688–695.

[BAI06] Bai. J., Nie, J.-Y., and Cao, G. (2006). Context-dependent term relations for information retrieval. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Sydney, Australia, pp.551–559.

[BAU11] Baun,C., Kunze, M., Nimis, J., Tai,S. (2011). Cloud Computing :Web-basierte dynamische IT-Services. Springer Netherlands ISBN 978-3-642- 18435-2.

[BAZ05a] Bazire, M. and Brézillon, P. (2005a). Understanding Context before Using It. A. Dey et al. (Eds.): CONTEXT 2005, LNAI 3554, pp. 29–40. Springer-Verlag Berlin Heidelberg.

[BAZ05b] Baziz, M. Boughanem, M. Aussenac-Gilles N. et Chrisment, C. (2005b). Semantic Cores for Representing document in IR. In SAC'2005- 20th ACM Symposium on Applied Computing. Santa Fe, New Mexico, USA, pp. 1011-1017.

[BAZ05c] Baziz, M. (2005). Indexation conceptuelle guide par ontologie pour la recherche dinformation. Ph.D. thesis, Universit de Toulouse, Uni-versit Toulouse III-Paul Sabatier.

[BEE00] Beeferman, D. and Berger, A. (2000). Agglomerative clustering of a search engine query log. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, Boston, MA, USA, pp. 407–416.

[BEL08] Belkin, N.J. (2008). Some (what) grand challenge for information retrieval. SIGIR forum 42. pp.47-54.

[BEL00] Belkin, N.J. (2000) Prospects for information "selection". Presentation for UCAO, March 2000.

[BEL92] Belkin, N.J. and Croft, W. (1992). Information Retrieval and Information Filtering: Two Sides of the same Coin. Communications of the ACM. 35(12), pp. 29-38.

[BEN06] Bennani, Y. (2006). Apprentissage connexionniste, http://www.lavoisier. fr/, édition Lavoisier.

[BHO13] Bhogal, J., and Macfarlane, A. (2013). Ontology Based Query Expansion with a Probabilistic Retrieval Model. 6th Information Retrieval Facility Conference, IRFC 2013, Limassol, Cyprus. Multidisciplinary Information Retrieval. Lecture Notes in Computer Science Volume 8201, 2013, ISBN 978-3-642-41057-4. Springer International Publishing, pp 5- 16.

[BHO07] Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. Information Processing and Management 43(4), pp.866–886.

[BIA09] Biancalana, C., Lapolla, A., & Micarelli, A. (2009). Personalized Web Search Using Correlation Matrix for Query Expansion, in J. Cordeiro et al. (Eds.): WEBIST 2008, LNBIP 18, pp. 186–198.

[BIE06] Bierig,R., Göker,A. (2006). Time, Location and Interest: An Empirical and User-Centred Study. Information Interaction in Context, IIiX, Copenhagen Denmark Copyright 2006 ACM 1-59593-482-0/06/10, pp. 79-87.

[BIL08] Bila, N. Cao, J. Dinoff R., Ho T., Hull R., Kumar B., and Santos P. (2008). Mobile user profile acquisition through network observables and explicit user queries. In 9th Int'l conference on Mobile Data Management, pp. 98–107.

[BIL00] Bilal. D. (2000). Children's use of the yahooligans! Web search engine: cognitive, physical, and affective behaviors on fact-based search tasks. Journal of the American Society for Information Science, 51(7):646-665.

[BIL03] Billerbeck, B., Scholer, F., Williams, H. E., and Zobel, J. (2003). Query expansion using associated queries. In Proceedings of the 12th ACM international conference on Information and knowledge management. ACM Press, New Orleans, Louisiana, USA, pp.2–9.

[BLA08] 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. pp. 720-735.

[BOH14] Böhm,T., Klas, C-P., and Hemmje, M.(2014). Collaborative Information Seeking in Professional Work-Settings: A Study of Equipment Utilization. Datenbank-Spektrum Journal. DOI.10.1007/s13222-014- 0145-2, Springer Berlin Heidelberg. Online ISSN.1610-1995.

[BOU99] Boughanem, M. Chrisment, C. and Soule-Dupuy. C. (1999). Query modification based on relevance backpropagation in adhoc environment. Information Processing and Management, 35. pp. 121-139.

[BOU11] Boughareb, D. Farah, N. (2011). Toward a Web Search Personalization Approach Based on Temporal Context. In H.,Cherifi, J., Mohamad Zain, E.,El-Qawasmeh (Eds.): Digital Information and Communication Technology and Its Applications - International Conference, DICTAP 2011. Proceedings, Part I. Springer 2011 Communications in Computer and Information Science ISBN 978-3-642-21983-2, pp.33-44.

[BOU12a] Boughareb, D. Farah, N. (2012a). Contextual Modelling of the User Browsing Behaviour to Identify the User's Information Need. Proceeding of INTECH'12. Casablanca-Morocco, pp. 268-273. IEEE, DOI: 10.1109/INTECH.2012.6457773.

[BOU13a] Boughareb, D. Farah, N. (2013a). Identify the User's Information Need Using the Current Search Context, International Journal of Enterprise Information Systems (IJEIS). DOI: 10.4018/ijeis.2013100103. IGI global publishing. 9(4):30-45.

[BOU13b] Boughareb, D., and Farah, N. (2013b). A Query Expansion Approach Using the Context of the Search, Advances in Intelligent and Soft Computing (Springer), Book chapter. In the 4th International Symposium on Ambient Intelligence, Salamanca-Spain ISami'13, Volume 219, DOI. 10.1007/978-3-319-00566-9_8. ISBN. 978-3-319-00566-9, pp. 57-63.

[BOU12b] Bouidghaghen, O., and Tamine, L. (2012b). Spatio-Temporal Based Personalization for Mobile search, Engineering, and Intelligent Technologies. Next Generation Search Engines: Advanced Models for Information Retrieval, PA: IGI Global publishing, pp.386-409.

[BOU05] Bouzeghoub M. and Kostadinov D. (2005). Personnalisation de l'information : Aperçu de l'état de l'art and définition d'un modèle flexible de définition de profils, Actes de la 2nde Conférence en Recherche d'Information and Applications CORIA, pp. 201-218.

[BRE09] Brenes, D. J., Gayo-Avello, D. and K., Perez-Gonzalez. (2009). Survey and evaluation of query intent detection methods. In Proceedings of the 2009 workshop on Web Search Click Data, WSCD' 09. ACM New York, NY, USA 2009. ISBN: 978-1-60558-434-8 doi.10.1145/1507509.1507510, pp.1-7.

[BRE08] Brenes, D.J., and Gayo-Avello, D. (2008). Automatic detection of navigational queries according to Behavioural Characteristics. LWA'08 Workshop proceedings, pp. 41-48.

[BRO02] Broder, A.Z. (2002). A Taxonomy of web search. SIGIR Forum, 36(2), pp. 3-10.

[BUC94a] Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic Query Expansion Using SMART, Overview of the Third Retrieval Conf. (TREC-3), Nov. 1994, pp. 69-80.

[BUC94b] Buckley, C., Salton, G., and Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. Proceedings of the 17th annual ACM international conference on research and development in information retrieval (SIGIR'94), Dublin, UK, pp. 292– 300.

[CAR01] Carpineto, C., De Mori, R., Romano, G., and Bigi, B. (2001). An Information Theoretic Approach to Automatic Query Expansion. ACM Transactions on Information Systems (TOIS) 19(1):1–27.

[CAT87] Cater, S. C., and D. H. Kraft. (1987). TIRS: A Topological Information Retrieval System Satisfying the Requirements of the Waller-Kraft Wish List. Presented at the 10th Annual Int'l ACM-SIGIR Conference on RetD in Information Retrieval, pp.171-180.

[CER90] Tim Berners-Lee, un informaticien du CERN inventa le World Wide Web en 1990. Le berceau du web. http://home.web.cern.ch/fr/about/birth-web

[CHA07] Challam, V. Gauch S., and Chandramouli, A. (2007). Contextual search using ontology-based user profiles. In Proceedings of RIAO ‘07, Pittsburgh USA, pp. 612-617.

[CHE12] Chen,L., Xu,D.,Tsang, I.W. Luo,J. (2012). Tag-Based Image Retrieval Improved by Augmented Features and Group-Based Refinement. IEEE Transactions on Multimedia, 14(4).pp.1057-1067.

[CHI07] Chirita, P.-A., Firan, C. S., and Nejdl, W. (2007). Personalized query expansion for the web. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Amsterdam, The Netherlands, pp. 7–14.

[CHI08] Chittaro L. (2008). Interactiong with visual interfaces on mobile devices. International federation of information processing IFIP. Vol. 272. Human-Computer Interaction Symposium (HCIS 2008). Paternò, Fabio; Mark Pejtersen, Annelise (Eds.) Milano, Italy, pp. 1-5.

[CHI03] Chittaro L. (ed.) (2003) Human–computer interaction with mobile devices and services. Lecture Notes in Computer Science, vol 2795. Springer, Berlin.

[CHO00] Choo, C., and Turnbull, D. (2000). Information seek-ing on the Web: An integrated model of browsing and searching. First Monday, 5(2),http://firstmon-day.org/issues/issue5_2/choo/index.html.

[CHO09] Chow, K.O., Fan, K., Chan, A., Wong, G. (2009). Content-Based Tag Generation for the Grouping of Tags. International Conference of Mobile, Hybrid, and On-line Learning ELML '09. Mexique. pp.7-12.

[CIA08] Ciaramita, M., Murdock, V., Plachouras, V. (2008). Online Learning from Click Data for Sponsored Search. Proceedings of the International World Wide Web Conference (IW3C2). WWW 2008, Beijing, China, ACM, ISBN: 978-1-60558-085-2 doi.10.1145/1367497.1367529, pp. 227-236.

[CLE67] Cleverdon, C. (1967). The cranfield test on index language devices. In: Aslib, pp. 173–194.

[COL11] Cole, M.J. Gwizdka,J. Belkin, N.J.(2011). physiological Data as Metadata, A Position Paper, SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011), July 28, 2011, Beijing, China.

[COO97] Cooley, R. Mobasher, B. Srivastava, J. (1997). Web mining: information and pattern discovery on the World Wide Web. Proceedings, Ninth IEEE International Conference on Tools with Artificial Intelligence. DOI:10.1109/TAI.1997.632303 ISBN: 0-8186-8203-5, pp. 558-567.

[COO88] Cooper, W. S. (1988). Getting Beyond Boole. Information Processing and Management 24(3), pp.243-48.

[CRE07] Crestani F. and Ruthven, I. (2007). Introduction to special issue on contextual information retrieval systems. Journal of Information Retrieval, 10 (2):111–113.

[CRO92] Crouch, C. and Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Copenhagen, Denmark, pp. 77–88.

[CUI03] Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y. (2003). Query expansion by mining user logs. IEEE Transactions on Knowledge and Data Engineering 15(4):829–839.

[CYB88] Cybenko, G. (1988). Continuous Valued Neural Networks with Two Hidden Layers are Sufficient, Technical Report, Department of Computer Science, Tufts University.

[DAO07] Daoud, M., Tamine, L., Boughanem, M. and Chebaro B. (2007). Learning implicit user interests using ontology and search history for personalization. Proceedings of the2007 international conference on Web information systems engineering, pp. 325-336.

[DAV98] Davies, N. Mitchell K., Cheverest K., and Blair G. (1998). Developing a context sensitive tourist guide. In Proceedings of the First Workshop on Human Computer Interaction for Mobile Devices Glasgow, pp. 64–68.

[DIA09] Diaz, F. (2009). Integration of news content into Web results. Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM), pp. 182–191.

[DIA06] Diaz, F. and Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Seattle, Washington, USA, pp.154– 161.

[DIN11] Dinh, D., Tamine, L. (2011). Towards a context sensitive approach to searching information based on domain specific knowledge sources. Preprint submitted to Special Issue of the Journal of Reasoning with Context in the Semantic Web, 12 (13).

[DIN07] Ding C. and Patra. J. C. (2007). User modeling for personalized web search with self-organizing map. Journal of American Society in Information Science and Technology, ISSN 1532-2882. 58(4). pp. 494– 507.

[DOU04] Dourish, P. (2004). What we talk about when we talk about context, Personal Ubiquitous Comput. 8(1), pp. 19–30.

[DUM03] Dumais S., Cuttrel E.,Cadiz J., Jancke G., Sarin R.,Robbins D.(2003). StuffI've seen: A system for a personal information retrieval and re-use. Proceedings of the 26th ACM SIGIR International Conference on Research and Development, pp. 72-79.

[EFT96] Efthimiadis, E. N. (1996). Query expansion. In Annual Review of Information Systems and Technology (ARIST), M. E. Williams, Ed. ASISetT'96, pp. 121–187.

[FAN08] Fang, F. (2008). A Re-examination of Query Expansion Using Lexical Resources. In Proceedings of the 46th Annual Meetings of the Association for Computational Linguistics (ACL'08).

[FAY96] Fayyad, U.M. Gregory, P.S. Padhraic. S. (1996). From data mining to knowledge discovery in databases, Ai Magazine 17, pp. 37–54.

[FID00] Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S., Grudin, J., and Poltrock, S. (2000). Collaborative Information Retrieval (CIR).The New Review of Information Behaviour Research , pp.235–247.

[FON05] Fonseca, B. M., Golgher, P. B., Possas, B.,Ribeiro-Neto, B. A., and Ziviani, N. (2005). Concept-based interactive query expansion. In CIKM, pp. 696-703.

[FRE05] Freund L., Toms G., Clarke C. (2005). Modeling Task-Genre Relationships for IR in the Workplace. ACM New York, NY, USA ©2005, ISBN: 1-59593-034-5 doi.10.1145/1076034.1076110. SIGIR'05, pp. 441-448.

[FRI07a] Frias-Martinez, E. Chen, S. Y. Macredie R. D., and Liu. X. (2007). The role of human factors in stereotyping behavior and perception of digital library users: a robust clustering approach. User Modeling and User- Adapted Interaction, 17(3), pp. 1573–1391.

[FRI07b] Friberg, K. (2007). Query expansion using domain information in compounds. In Proceedings of the NAACLHLT'07 Doctoral Consortium, Rochester, New York. Association for Computational Linguistics, pp.1–4.

[FU04] Fu, L., Goh, D. H.-L., Foo, S. S.-B. and Supangat, Y. (2004). Collaborative querying for enhanced information retrieval. In Heery, R. and Lyon, L., éditeurs : Research and Advanced Technology for Digital Libraries, 8th European Conference, ECDL 2004, volume 3232 de Lecture Notes in Computer Science, pp.378–388.

[FUH00] Fuhr, N. (2000). Information retrieval: introduction and survey. Post-graduate course on information retrieval, university of Duisburg-Essen, Germany.

[GAR07] Garcia P., Amandi A., Schiaffino S. (2007). Campo M. Evaluating Bayesian Networks' Precision for Detecting Students' Learning Styles, Computers and Education, 49, pp. 794-808.

[GAR05] Garrigos I., Casteleyn S. and Gómez J. (2005). A Structured Approach to Personalize Websites Using the OO-H Personalization Framework, Y. Zhang et al. (Eds.): APWeb, LNCS 3399, pp. 695–706.

[GEN03] Gentili G., Micarelli A., Sciarrone F. (2003). Infoweb: An Adaptive Information Filtering System for the Cultural Heritage Domain. Applied Artificial Intelligence. 17(8):715-744.

[GOK11] Goker, A. (2011). Information in Context: The Mobile Environment. ELPUB2011. Digital Publishing and Mobile Technologies, 15th International Conference on Electronic Publishing 22-24 June 2011, Istanbul, Turkey/ Edited by: Yasar Tonta, Umut Al, Phyllis Lepon Erdogan and Ana Alice Baptista. ISBN 978-975-491-320-0, pp. 1-2.

[GOK02] Goker, A. and Myrhaug,H. I.(2002). User context and personalisation. In Workshop on Case Based Reasoning and Personalization, in conjunction of the 6th European Conference on Case Based Reasoning ECCBR, Mehmet H. G¨oker, Barry Smyth (ed.), pp.1–8.

[GON06] Gong, Z., Cheang, C.-W., and U, L. (2006). Multi-term web query expansion using wordnet. In Proceedings of 17th International Conference on Database and Expert Systems Applications -DEXA '06. Springer, Krakow, Poland, pp.379–388.

[GRA44] Grassmann, H.G. (1844). La Science de la Grandeur Extensive, traduction et préface de D. Flament et B. Bekemeier, Blanchard, Paris (1994).

[GRU93] Gruber, Thomas R. (1993). A translation approach to portable ontology specifications (PDF). Knowledge Acquisition 5 (2), pp.199–220.

[GUA99] Guarino, N. Masolo C. and Vetere G. (1999). OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs. National Research Council, LADSEBCNR, Padavo, Italy.

[HAR01] Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Grju, R., Rus, V., and Morarescu, P. (2001). The role of lexico-semantic feedback in open-domaintextual question-answering. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-01). Association for Computational Linguistics, Toulouse, France, pp.282–289.

[HAR95] Harman D.K. (1995). Overview of the the 1st text retrieval conference (trec-4).Dans Proceedings of the 1st text retrieval conference (TREC-4). National Institute of Standards and Technology, NIST special publication, pp. 1–24.

[HAR92] Harman. D. (1992). Relevance feedback revisited. In 15th Annual International ACM SIGIR Conference on Research and development in Information Retrieval, pp. 1-10.

[HAR96] Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness, Journal of the American Society for Information Science, 47 (1), pp.37-49.

[HAT07] Hattori, S. Tezuka, T. and Tanaka, K. (2007). Context-aware query refinement for mobile web search. In SAINT-W '07: Proceedings of the International Symposium on Applications and the Internet Workshops, Washington, DC, USA. IEEE Computer Society, pp. 15.

[HAT06] Hattori, S. Tezuka, T., and Tanaka, K. (2006). Activity-based query refinement for context-aware information retrieval. The 9th Int'l conference on Asian digital libraries, LNCS, pp. 474–477.

[HAY94] Haykin, S. (1994). Neural Networks-A comprehensive Foundation, Macmillan College Publishing Company, New York.

[HE00] He, D., Goker. A. (2000). Detecting session boundaries from web user logs. Proceedings of the 22nd Annual Colloquium on Information Retrieval Research, pp. 57–66.

[HOC08] Höchstötter N., and Koch M. (2008). Standard parameters for searching behaviour in search engines and their empirical evaluation. Journal of Information Science 39: 346-358.

[HOT06] Hotho, A., Jäschke, R., Schmitz,C., Stumme, G.(2006). Information retrieval in folksonomies: search and ranking Proceeding ESWC'06 Proceedings of the 3rd European conference on The Semantic Web: research and applications. Springer-Verlag Berlin, Heidelberg ISBN: 3- 540-34544-2 978-3-540-34544-2 doi.10.1007/11762256_31. pp. 411- 426.

[HUP06] Hupfer M. E. and Detlor. B. (2006). Gender and web information seeking: A self-concept orientation model. Journal of the American Society for Information Science and Technology, 57(8), pp. 1105–1115.

[HUY09] Huynh, T. and Miller, J. (2009). Empirical observations on the session timeout threshold. Information Processing and Management, 45(5):513– 528.

[ING05] Ingwersen, P. and Jarvelin (2005). The TURN: Integration of Information Seeking and Retrieval in Context. SPRINGER, August 2005 ISBN 978-1-4020-3851-8, Vol. 18.

[ING94] Ingwersen P. (1994). Poly-representation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 1994. Springer-Verlag New York, Inc. pp. 101–110.

[ISO86] ISO 2788:1986: Documentation--Principes directeurs pour l'établissement et le développement de thésaurus monolingues.

[JAI05] Jaimes A., and Liu, J. (2005). Sit straight (and tell me what I did today): A Human Posture Alarm and Activity Summarization System, CARPE'05, Singapore, pp. 23-34.

[JAN07] Jansen, B.J., Booth, D.L., Spink, A. (2007). Determining the informational, navigational, and transactional intent of Web queries. Information Processing and Management 44 (2008) Elsevier Ltd.doi:10.1016/ j.ipm.2007.07.015, pp.1251–1266.

[JAN06a] Jansen, B.J. and Spink A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs, Information Processing and Management 42: 248-263.

[JAN06b] Jansen, B.J., Ramadoss, R., Zhang, M., and Zang, N. (2006). Wrapper: An application for evaluating exploratory searching outside of the lab, SIGIR 2006 Workshop on Evaluating Exploratory Search Systems. The 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR2006). Seattle, Washington, USA.

[JAN98] Jansen, B. J. Spink, A. Bateman, J. and Saracevic, T. (1998). Real life information retrieval: a study of user queries on the web. SIGIR Forum, 32(1): 5–17.

[JAR02] Jarvelin K. and Kekalainen. J. (2002).Cumulative gain-based evaluation of IR techniques. ACM Transactions on Information Systems (ACM TOIS), 20(4) : 422–446.

[JEN05a] Jeng, J. (2005a). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction,and learnability. International Journal of Libraries and Information Services, 55:96–121.

[JEN05b] Jeng, J. (2005b). What is usability in the context of the digital library and how can it be measured? Information Technology and Libraries, 24(2):47–56.

[JIN11] Jin, P., Chen,H., Lin,S., Zhao,X. and Yue, L. (2011). Hybrid Index Structures for Temporal-Textual Web Search, X. Du et al. (Eds.): APWeb 2011, LNCS 6612, © Springer-Verlag Berlin Heidelberg, pp. 271–277.

[JOA07] Joachims, T, Granka, L, Hembrooke H, Radlinski F, Gay G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Trans Inf Syst 25(2):7.

[JON06] Jones, R., Rey, B., Madani, O., and Greiner, W. (2006). Generating query substitutions. In Proceedings of the 15th international conference on World Wide Web. ACM Press, Edinburgh, Scotland, pp.387–396.

[KAN08] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., and Bani-Ismail, B. (2008). Interactive and Automatic Query Expansion: A Comparative study with an Application on Arabic. American Journal of Applied Sciences 5, 11: 1433–1436.

[KAN03] Kang, I.H. and Kim, G.C. (2003). Query type classification for web document retrieval. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, (2003), pp. 64-71.

[KAR98] Karamuftuoglu, M. (1998). Collaborative information retrieval: Towards a social Informatics view of IR interaction. JASIS, 49(12):1070–1080.

[KEA08] Keane M. T., O'Brien M., Smyth B. (2008). Are people biased in their use of search engines?, Magazine -Alternate reality gaming CACM, 51(2):49-52.

[KEL04] Kelly, N. J. (2004). Understanding implicit feedback and document preference: a naturalistic study. In PHD dissertation. Ritgers University, New Jersey.

[KEL03] Kelly D. and Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. SIGIR Forum, 37(2):18–28.

[KHA13] Khan,A., Martin, D., and Tiropanis, T.(2013).Using Semantic Indexing to Improve Searching Performance in Web Archives. WEB2013: The First International Conference on Building and Exploring Web Based Environments. Copyright (c) IARIA. ISBN: 978-1-61208-248-6, pp.101-104.

[KIM08] Kim. K. (2008). Effects of emotion control and task on web searching behavior. Information Processing and Management, 44(1) :373–385,

[KLA10] Klassen, M., and Paturi, N. (2010). Web Document Classification by Keywords Using Random Forests, in F. Zavoral et al. (Eds.): NDT 2010, Part II, CCIS 88, pp. 256–261.

[KOF03] Kofod-Petersen A. and Aamodt, A. (2003). Case-Based Situation Assessment in a Mobile Context-Aware System. In A. KrÄuger and R. Malaka, editors, Artificial Intelligence in Mobile Systems (AIMS), pp. 41-49,

[KRA04] Kraft, R. and Zien, J. (2004). Mining anchor text for query refinement. In Proceedings of the 13th international conference on World Wide Web. ACM Press, New York, NY, USA, pp. 666–674.

[KUM13] Kumar, N., Carterette, B.A. (2013). Time Based Feedback and Query Expansion for Twitter Search. ECIR 2013, pp.734-737.

[LAN10] Lane, N.D. Lymberopoulos, D., Zhao, F. Andrew T. Campbell. (2010). Hapori: context-based local search for mobile phones using community behavioral modeling and similarity, proceeding of the Ubicomp '10 Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 109-118.

[LEL98] Leloup, C. (1998). Moteurs d'indexation et de recherche : Environnement client-serveur, internet et intranet, édition Eirolles 1998, ISBN : 2-212-08976-7, pp. 35-97.

[LER04] Leroux, J.D. (2004). Les réseaux de neurones artificiels, Explique-moi ça! : Concours de propagation scientifique, Rapport de recherche de l'université de Sherbrooke.

[LI08] Li Y. and Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. Information Processing and Management, 44(6) :1822–1837.

[LIN05] Lin C., Xue G., Zeng H., Yu Y. (2005). Using probabilistic latent semantic analysis for personalised Web search. Proceedings of the APWeb Conference, pp. 707-711.

[LIN14] Lin, S., JIN, P., ZHAO, X.and Yue, L. (2014). Exploiting temporal information in Web search. Expert Systems with Applications, 41(2, 1) : 331–341.

[LIU08] Liu, Y., Li, C., Zhang, P., and Xiong, Z. (2008). A query expansion algorithm based on phrases semantic similarity. In Proceedings of the 2008 International Symposiums on Information Processing. IEEE Computer Society, Moscow, Russia, pp. 31–35.

[LIU04] Liu, F. Yu, C. and Meng. W. (2004). Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16(1):28–40.

[LIU02] Liu, F. Yu, C. and Meng. W. (2002). Personalized web search by mapping user queries to categories. In Proceedings of the 11th International Conference on Information and Knowledge Management, Mclean, Virginia. ACM, pp. 558–565.

[LOR08] Lorigo, L. Haridasan, M. Brynjarsdottir, H. Xia L., Joachims, T. Gay, G. Granka, L. Pellacini, F. and Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead, Journal of the American Society for Information Science and Technology 59: 1041-1052.

[LUH57] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. jounal of IBM-JRD, 1(4) :309-317.

[LV10] Lv, Y. and Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pp. 579-586.

[MAC06] Macgregor, G., and McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, Emerald Group Publishing. 55(5):291–300.

[MAN08] Mandl, T. (2008): Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance. Informatica (32): 27–38.

[MCG03] Mc Gowan J., (2003). A multiple model approach to personnalised information access, Master thesis in computer science, Faculty of science, University College Dublin.

[MEL08] Melucci, M. (2008). A Basis for Information Retrieval in Context. ACM Transactions on Information Systems, 26(3):14.

[MIC08] Microsoft Research Microsoft Live Labs: Accelerating Search in Academic Research 2006, Request for Proposals, (2006). Available at: http://research. microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx (accessed 24 November 2008).

[MIL07] Milne, D. N., Witten, I. H., and Nichols, D. M. (2007). A knowledge-based search engine powered by wikipedia. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM Press, Lisbon, Portugal, pp.445–454.

[MIO13] Miotto, R. and Weng,C. (2013). Unsupervised mining of frequent tags for clinical eligibility text indexing. Journal of Biomedical Informatics. Article in Press.

[MOB07] Mobasher, B. (2007). Data Mining for Web Personalization. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) the Adaptive Web: Methods and Strategies of Web Personalization. LNCS Springer, Heidelberg, vol. 4321, pp. 90–135.

[MOG11] Moghaddam S., and Helmy A. (2011). Multidimensional Modelling and Analysis of Wireless Users Online Activity and Mobility: A Neural-networks Map Approach, MSWiM'11, ACM, USA, pp. 401-408.

[MOO50] Mooers, C. (1950). Information retrieval viewed as temporal signaling. Proceedings of the International Congress of Mathematicians.1:572-573.

[MYL08] Mylonas, Ph., Vallet, D., Castells, P., Fernández, M., and Avrithis, Y. (2008). Personalized information retrieval based on context and ontological knowledge. In Knowledge Engineering Review, special issue on Contexts and Ontologies, 23 (1):73-100.

[MYR03] Myrhaug, H. I., and Goker, A. Ambiesense – interactive information channels in the surroundings of the mobile user. In 2nd International Conference on Universal Access in Human-Computer Interaction (Crete, Greece, 2003), vol. 4, Lawrence Erlbaum Associates, pp. 1158–1162.

[NAV09] Navigli, R. (2009). Word Sense Disambiguation: A Survey. ACM 0360-0300/2009/02-ART10. ACM Computing Surveys, DOI. 10.1145/1459352. 1459355. http://doi.acm.org/10.1145/ 1459352.1459355. 41(2), Article 10.

[NAV05] Navigli, R. And Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. Patt. Anal. Mach. Intell. 27 (7):1075–1088.

[NAV03] Navigli, R. and Velardi, P. (2003). An analysis of ontology-based query expansion strategies. In Proceedings of the ECML/PKDD-2003 Workshop on Adaptive Text Extraction and Mining. Cavtat-Dubrovnik, Croatia.

[PAN05] Panayiotou C., Andreou M., Samaras G., and Pitsillides A. (2005). Time based personalization for the moving user. In Proceedings of the 4th Int. Conference on Mobile Business (ICMB'05), pp. 128–136,

[PAR07] Park, L. A. F. and Ramamohanarao, K. 2007. Query expansion using a collection dependent probabilistic latent semantic thesaurus. In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007). Springer, Nanjing, China,pp. 224–235.

[PAR04] Parizeau, M. (2004). Réseaux de neurones, Cours de l'université laval.

[PAS08] Pasca, M. (2008). Towards Temporal Web Search. Proceeding of the 2008 ACM symposium on Applied computing SAC'08, ISBN: 978-1-59593-753-7 doi.10.1145/1363686.1363946, pp.1117-1121.

[PAZ96] Pazanni M., Muramatsu J., Billsus D. (1996). Syskill and Webert: Identifying interesting Web sites. Proceedings of the 13th National Conference on Artificial intelligence, pp. 54-61.

[PLO04] Plovnick, R.M. and Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: a pilot study. Journal of medical Internet research, 6(3).

[PUR04] Purnima Chandrasekaran, A. J. (2004). Mobileiq: A framework for mobile information access. In Proceedings of the 3rd Int'l Conference on Mobile Data Management (MDM'02), pp.43–50.

[QAM06] Qamra, A. Tseng, B. and Chang, E. (2006). Mining Blog Stories Using Community-based and Temporal Clustering. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06), pp 58–67.

[QIU06] Qiu F., and Cho J. (2006). Automatic identification of user interest for personalized search. In Proc. WWW 2006, ACM, Scotland, 2006, pp. 727-736.

[QUI00] Quiroga, L., Mostafa, J. (2000). Empirical evaluation of explicit versus implicit acquisition of user profiles in information filtering systems. In: Proceedings of the 63rd Annual Meeting of the American Society for Information Science and Technology, Medford, vol. 37, pp. 4–13.

[ROB00] Robertson, S. Walker, S. and Beaulieu. M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing and Management, 36(1):.95-108.

[ROB76] Robertson S.E. and Sparck-Jones. J.K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129- 146.

[ROC71] Rocchio, J.J. (1971). The SMART Retrieval System: Experiments in Automatic Document Processing, chapter Relevance Feedback in Information Retrieval, pp. 313–323.

[ROJ96] Rojas, R. (1996). Neural Networks - A Systematic Introduction. Foreword by Jerome Feldman Springer-Verlag, Berlin, New-York.

[ROS04] Rose, D.E., Levinson, D. (2004), Understanding user goals in Web search. In Proceedings of the Thirteenth Int'l. World Wide Web Conf.' 2004.

[ROS58] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Journal of Psychological Review, 65(6): 386-408.

[RUS10] Rusitschka, S., Eger, K. Gerdes, C. (2010). Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain. First IEEE International Conference on Smart Grid Communications (SmartGridComm), Gaithersburg, MD. ISBN: 978-1-4244- 6510-,pp.483–488.

[RUT03] Ruthven, I. (2003). Re-examining the Potential Effectiveness of Interactive Query Expansion. Proceedings of ACM SIGIR'03. Toronto, Canada. ISBN:1-58113-646-3 doi.10.1145/860435.860475, pp. 213-220.

[RUT01] Ruthven, I. (2001). Abduction, explanation and relevance feedback. University of Glasgow. Doctoral dissertation. Technical report: TR-2002-115.

[RYA97] Ryan N, Pascoe J, Morse D. (1997). Enhanced reality fieldwork: the context-aware archaeological assistant. In: Gaffney V, van Leusen M, Exxon S (eds) Computer Applications in Archeology.

[RYE05] Ryen, W.W., Ruthven, I., Jose, J.M. and Van Rijsbergen, C. J. (2005). Evaluating implicit feedback models using searcher simulations. ACM Transactions on Information Systems, 23(3) :325–361.

[SAH14] Sahuguet,M., and Huet, B. (2014). Mining the Web for Multimedia-Based Enriching. MultiMedia Modeling, 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part II, DOI.10.1007/978-3-319-04117-9_24, Lecture Notes in Computer Science, pp. 263-274.

[SAI11] Said, A., De Luca, E.W., and Albayrak, S. (2011). Inferring Contextual User Profiles–Improving Recommender Performance. 3rd RecSys Workshop on Context-Aware Recommender Systems, Chicago, IL, USA.

[SAI10] Saint-Michel, S-H. (2010). Comportement des internautes et moteurs de recherche. Publié le 02.04.2010 : http://www.marketing-professionnel.fr/chiffre/comportement-internautes-moteurs-recherche-seo-marques.html

[SAL83a] Salton, G. and McGill, M (1983a). An Introduction to Modern Information Retrieval. McGraw-Hill, New York.

[SAL83b] Salton, G., E. A. FOX, and H. WU. (1983b). "Extended Boolean Information Retrieval." Communications of the ACM, 26(12): 1022-36.

[SAL71a] Salton, G. (1971a). A comparison between manual and automatic indexing methods Journal of American Documentation, 20(1): 61-71.

[SAL71b] Salton, G., (1971b). The SMART Retrieval System- Experiments in Automatic Document Processing. Prentice Hall, Englewood, Cliffs, New Jersey.

[SAL68] Salton, G. (1968). Search and retrieval experiments in real-time information retrieval. IFIP Congress (2) 1968: 1082-1093.

[SAN10] Sanderson M. (2010). Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4(4):247-375.

[SAN07] Sanderson M. and Dumais S. (2007). Examining repetition in user search behaviour. In Proceedings of the 29th European Conference on IR Research, Advances in Information Retrieval. Amati, G., Carpineto, C., and Romano, G., Eds. Springer. Berlin, Germany, pp.597–604.

[SAR04] Saracevic.T. (2004). Evaluation of digital libraries: An overview. In WP7 Workshop on the Evaluation of Digital Libraries, DELOS Network of Excellence on Digital Libraries, pp. 13–30.

[SAR97] Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and Applications, In Proceedings of the American Society for Information Science meeting, Vol. 34, pp. 313-327.

[SAR96] Saracevic, T. (1996). Relevance reconsidered - Information Science: Integration in perspective, In Proceedings of the second conference on Conception of Library and Information Science, Copenhague, Danemark, pp. 201-218.

[SAR75] 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, 26(6): 321-343.

[SHE06] Shenouda, E.A.M.A. (2006). A Quantitative Comparison of Different MLP Activation Functions in Classification. In J. Wang et al. (Eds.): ISNN 2006, LNCS 3971. Springer-Verlag Berlin Heidelberg, pp. 849–857.

[SCH04] Scherbina, A., and Kuznetsov, S. (2004). Clustering of Web sessions using Levenshtein metric. Proceedings of the 4th Industrial Conference on Data mining (ICDM 2004) .LNCS 3275. San Jose, CA, pp. 127-133.

[SCH03] Schilit, BN., LaMarca, A., Borriello, G., Griswold, WG., McDonald, D., Lazowska, E., Balachandran, A., Hong J., Iverson, V. (2003). Ubiquitous location-aware computing and the place lab initiative challenge. In: WMASHE'03, The first ACM international workshop on wireless mobile applications and serviceson WLAN (WMASH 2003), San Diego, CA, ACM, New York, NY, USA, pp.29-35.

[SCH95] Schilit, B.N. (1995). A System Architecture for Context-Aware Mobile Computing. PhD thesis, Columbia University.

[SEG14] Segura, A., Vidal-Castro, C., Ferreira-Satler, M., Sánchez, S. (2014). Domain Ontology-Based Query Expansion: Relationships Types-Centered Analysis Using Gene Ontology. Proceedings of the 2nd Colombian Congress on Computational Biology and Bioinformatics (CCBCOL). DOI. 10.1007/978-3-319-01568-2_27. Springer International Publishing Switzerland, pp. 183-188

[SHA02] Shah, U., Finin, T., Joshi,A., Scott Cost, R., Matfield. J. (2002). Information retrieval on the semantic web Proceedings of the eleventh international conference on Information and knowledge management CIKM '02, ACM New York, NY, USA ©ISBN: 1-58113-492-4 doi-10.1145/584792.584868, pp. 461-468.

[SHE05] Shen, X., Tan, B, Zhai, C (2005) Context-sensitive information retrieval using implicit feedback. In:Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, pp 43–50.

[SIE07] Sieg A, Mobasher B, Burke R. (2007). Web search personalization with ontological user profiles. In proceedings of the 16th ACM conference on conference on information and knowledge management, ACM, New York, NY, USA, pp 525–534.

[SIL99] Silverstein, C., Henzinger, M., Marais, H., Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR forum 33(1): 6–12

[SMY06] Smyth B., Balfe E. (2006). Anonymous personalization in collaborative web search. Information Retrieval 9(2):165-190.

[SON07] Song, M., Song, I. -Y., Hu, X., and Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. Journal of Data and Knowledge Engineering. 63(1):63–75.

[SPA05] Sparck Jones, K. (2005). Meta-reflections on TREC, TREC: Experiment and Evaluation in Information Retrieval, Cambridge, MA: MIT Press, pp. 421-448.

[SPA79] Sparck-Jones, K. (1979). Experiments in relevance weighting of search terms. Inf. Process. Manage, 15(3):133–144.

[SPE05] Speretta, M. and Gauch, S. (2005). Personalized search based on user search histories. Dans Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp 622–628.

[SRI04] Srivastava,J., Desikan, P., Kumar, V.(2004). Chapter3: Web Mining–Accomplishments and Future Directions, pp.51-70.

[SRI00] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P-N. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. SIGKDD Explorations, 1(2):12–23.

[TAK09] Takács, G., Pilászy, I., Németh, B., and Tikk D. (2009). Scalable Collaborative Filtering Approaches for Large Recommender Systems, in Paolo Frasconi, KristianKersting, HannuToivonen and Koji Tsuda, Journal of Machine Learning Research 10:623-656.

[TAM09] Tamine,L., Boughanem M., and Daoud M. (2009). Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems, 24(1):1–34.

[TAM08a] Tamine, L, Boughanem M, Zemirli WN. (2008a). Personalized document ranking: exploiting evidence from multiple user interests for profiling and retrieval. J Digit Inf Manag (in press).

[TAM08b] Tamine, L., Daoud, M., Dinh, B.D., and Boughanem, M. (2008b). Contextual query classification in web search LWA 2008 Workshop Proceedings, pp.65-68.

[TAM03] Tamine, L. Chrisment, C. and Boughanem. M. (2003). Multiple query evaluation based on an enchanced genetic algorithm. Information Processing and Management, 39(2): 215-231.

[TOM05] Tombros, A., Ruthven I, Jose JM. (2005). How users assess web pages for information seeking. J Am SocInf Sci Technol 56(4):327–344.

[TUR06] Turpin, A. and Scholer, F. (2006). User performance versus precision measures for simple search tasks. SIGIR 2006, Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, pp. 11-18.

[UIC05] Uichin, L. Zhenyu, L., Junghoo, C. (2005). Automatic identification of user goals in Web search. In Proceedings of the 14th International World Wide Web Conference (WWW) '05, Chiba, Japan, disponible en ligne: http://www2005.org/cdrom/docs/p391.pdf.

[UM00] Um, I-T, Ra, J-H., Kim, M-H. (2000). Comparison of clustering methods for MLP-based speaker verification. In proceeding of Pattern Recognition, IEEE Xplore, 02/2000; 2: vol.2. DOI:10.1109/ICPR.2000.906115 ISBN: 0-7695-0750-6, pp. 475 – 478.

[VAC03] Vakkari, P. (2003). Task-based information searching. Annual Review of Information Science and Technology. 37(1):413-464.

[VLA04] Vlachos,M., Meek,C., Vagena, Z., Gunopulos. D. (2004). Identifying Similarities, Periodicities and Bursts for Online Search Queries. SIGMOD 2004 Copyright 2004 ACM, June, Paris, France.

[VOO06] Voorhees, E. (2006). Overview of the trec 2005 robust retrieval track. In E.M. Voorhees and L. P. Buckland, editors, The Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, MD,NIST.

[VOO04] Voorhees, E. (2004). Overview of the trec 2004 robust track. In Proceedings of the 13th Text REtrieval Conference (TREC-7), NIST Special Publication. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, pp. 500-261.

[VOO94] Voorhees, E. (1994). Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Dublin, Ireland, pp. 61–69.

[VOO93] Voorhees, E. (1993). Using wordnet to disambiguate word senses for text retrieval. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Pittsburgh, Pennsylvania, USA, pp.171–180.

[WAN14] Wang, X., Hong , Z., Xu, Y., Zhang, C. and Ling, H. (2014). Relevance judgments of mobile commercial information Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23060

[WAN11] Wang, J-C.; Wu, M-S.; Wang, H-M.; Jeng. S-K. (2011).Query by multi-tags with multi-level preferences for content- based music retrieval. IEEE International Conference on Multimedia and Expo (ICME), pp.1-6. Barcelona, Spain.

[WAN09] Wang, H., Liang, Y., Fu, L., Xue, G.-R., and Yu, Y. (2009). Efficient query expansion for advertisement search. Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press. Boston, MA, USA, pp. 51–58.

[WAN03] Wang, P., Berry, M., and Yang, Y. (2003). Mining longitudinal web queries: trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), pp.743–758.

[WEN04] Wen J., Lao N.,Ma W. Y. (2004). Probabilistic model for contextual retrieval, Proceedings of the 27th annual international ACM SIGIR Conference on Research and development in Information retrieval, pp. 57-63.

[WHI05] White R., Ruthven I., Jose J., and Van Rijsbergen C. (2005). Evaluating implicit feedback models using searcher simulations. ACM Transactions on Information Systems, 23(3):325–361, 2005.

[WHI01] White, R.W., Jose, J.M., Ruthven, I. (2001). Comparing explicit and implicit feedback techniques for web retrieval. In: Proceedings of the Tenth Text Retrieval Conference, Gaithersburg, pp. 534–538.

[WU02] WU, Y-H. and CHEN, A. L. P. (2002). Prediction of Web Page Accesses by Proxy Server Log. World Wide Web: Internet and Web Information Systems. 5:67–88.

[XIE08] Xie, H., (2008) 'Users' evaluation of digital libraries (dls): their uses, their criteria, and their assessment. Inf Process Manag 44(3):1346–1373.

[XU96] Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Zurich, Switzerland, pp.4–11.

[YAN06] Yannibelli, V., Godoy D., Amandi A. (2006). A Genetic Algorithm Approach to Recognize Students' Learning Styles. Interactive Learning Environments.14, pp.55-78.

[ZAD65] Zadeh, L. A. (1965). Fuzzy Sets, Information and Control, 8, 338-353.

[ZEM08] Zemirli, W. N. (2008). Modèle d'accès personnalisé a l'information base sur les diagrammes d'influence intégrant un profil multidimensionnel. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

[ZHA06] Zhao, Q., Hoi, S. C. H., Liu, T.Y., Bhowmick, S. S., Lyu, M. R., Ma, W.Y. (2006) Time-Dependent Semantic Similarity Measure of Queries Using Historical

Click-Through Data. Proceedings of the 15th international conference on World Wide Web. ACM. EDINBURGH, SCOTLAND, pp.543-552.

[ZIE01] Zien, J. Meyer, J. Tomlin, J. and Liu. J. (2001). Web query characteristics and their implications on search engines. In Proceedings of the 10th International WWW Conference'01, Hong Kong.

[ZIP49] Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Eds. Addison WESLEY PUBLISHING.

BIBLIOGRAPHIE WEB

[ONE11] http://www.onestat.com/

[OXF13] http://www.oxforddictionaries.com/

[UCA10] http://code.google.com/p/ucair/

[WEB01] Ten years in search - a look back at the top 25 Web searches http://websearch.about.com/od/enginesanddirectories/tp/top-web-searches-of-the-decade.htm

[WIK13] Wikipedia (2013):

http://fr.wikipedia.org/wiki/Architecture_g%C3%A9n%C3%A9rale_pour_le_traitement_de_texte

ANNEXES

ANNEXE A : SOURCE DE DONNÉES

FICHIER LOG

La figure A.1 représente un extrait capturé depuis le fichier d'historique de navigation sur lequel nous avons travaillé.

05/11/2011 10:52:08 Ecouter et télécharger la musique en mp3 http://www.arazik.net/

QBJREFUeNpizKvsZCASiAKxOhBrAbExjM3y//9/XBoygdgAqgGEhbApwmfANGKcxfL/3z8GSgDLv/8UGkCkC74D8TUgvgzEd0D48+evjLy83MtZ/uEx4Nv3H+lcnBw7
gMxH6HIfPn4y5ObmBAUibgPev/94mpOD7REueZBevC4AAXzyIDnKDSAUiEjy/EDcC8TeII6IEP+u/0guQJEEgq1YXLAIiP2QzI4DpUGYC9Alk6H0HSQXeGBzIRMoIQ
GxB5RGxr8E+Dg/I/F3YFGziQkUFUC8A0ojY3QxoJP/zQXid1AMYscBvQDOTHHQMAiEumw9EBejufYjEKdAMRwABBgALz3NOYpyfjcAAAAASUVORK5CYII=
05/11/2011 10:52:06 http://www.google.dz/url?sa=t&rct=j&q=musique
mp3&source=web&cd=4&cad=rja&ved=0CDkQFjAD&url=http://www.arazik.net/&ei=wK1uUJKQHcrOsgaS_oC4CA&usg=AFQjCNFcfpCOQtfwKLKaRvNbNFtSdtBzM
w
http://www.google.dz/url?sa=t&rct=j&q=musique%20mp3&source=web&cd=4&cad=rja&ved=0CDkQFjAD&url=http%3A%2F%2Fwww.arazik.net%2F&ei=wK1uUJKQH
crOsgaS_oC4CA&usg=AFQjCNFcfpCOQtfwKLKaRvNbNFtSdtBzMw

mD4BuCVgfFKmitCl0s+FKhvoEgVvsyWKuRS9JLcvm7tcplSHW44e6e5/c8x91JAaKFZJXWFELRzZBVWgsQLST9JfknInlt9ExRJLMMqSOG67ID7gLb5xbG100h1hNIFy
zM51gbu61wnN7Znl14Al+GC7LTas9nMi20bPgHPnUXmatOxbE1E89v3D8wd8DAbGBiw0R/XMfupY3RJcM/oBCKkUUDiUMGF/h1HN+AQiiC0xSa4aL04mBgVvcPTK
ZNbBYspHIMy3mGJnXx+s4xmBARAVg4Ybh4ctAb66wNJXSUGxx7RfEqBaDa5EgdMSEwmWXIlnwA+Qcb5QbHcLLTbjBGcfboILLq4yX2xXVsFSzUP1zcVzmOb2zsF
21EVsRkhVD89zPVJTmqhWWV1rsGVFqRo1r4G6iM33AbQTj+AAAAAElFTkSuQmCC
05/11/201110:36:16 Dj Compilation 2011. Dj Compilation 2011 . Dj Compilation 2010 . Dj Compilation 2009 . Dj Compilation Live . Dj Compilation Album . Melody.Ma
Telecharger Et Ecouter Touts Les Types De Musique http://www.melody.ma/Dj_Compilation_MUSIC_0.html

Ny1kF6TJJsbL/gXevBi96M3oycSDMTEx8aAHjx6MAcq8dFqgpYVAEQpE42pCFNGYaKCdztdDTTebePgcnufJIwBqXhXXA7fugedxcXrCScag/NH7ZB7MsqSGsPtlMoqE
pQVw1CCHn3yI53kIt+7haBKGJmOEZUy9+7FwCEu/+gQ7cq1pLXUPsRntwRl6LDt8nezwdZxbvU3ZF280WYONnBvpozhyE5GLa2zEBijcVrkwBbl4P/m40jCmNhXjAxR
HBjn/tpX8mEphXKMU1xHFySG2EkMUp4ZwLcGj9/yUEoNsTwxTmhxmayrK1lSUnQkdzxG4Rgvb01G2p2+xcyeK2E3GKCdH2Z0dwTUFriWovBKnnBylnBxlbzZGeWa
cUvI2ru2jbvooz8Uoz8XYm44j9u/foTKfYH/hLjW7lZopcJfaOHg1RmU+QWU+wUFqjHq2hZrdgmsJ9hYan+O5BOJgcYKj1AzHC9OcL13ir69kqrbgh8WXOUrNcLg4zX7qJ
X55uxfPfArXEhynGlvl3iRiWQuSViSWe2XOPm9jRe2iavqo2W2saJexdImaKTAUP9WlVuq2IH2zs/FRgohVLURG8WNHJB69cxlL6WQ52k7dFhjDnRQmAuQetOIoHZx+
0UbdFpjKFWzFT0Z9DrGmXyOvy+w9vML3rz3Lmi6zHg7g5p7mb0dw4QjWwwE2wn5+frcd1xLkNT/rmsRGuAuxMdjD5lAfZ5+2URqX2NS7KUaC7Gi9nFuCjfvPUIwEK
UaCHL7eQc0UFAcauRAJIczBq6TvXqJu+rgwfOSVHvJ6kLwe5OjNVgqREHk9SE6/wY9v+fBMwe58OyW9j82IhHD6Azjak7JhmWxYJqf9P0eRWQt3s651IKp1D0cPYagS
phLAGJBYVbpJqxKm6sf+rzOVAJYqNxlagMMvP0N4dY/f7RXO7BVOrRV+s9L8YRv8mTE5+eZrKh9/wNYbD1kd1fnu+S6MfgnzhS7++fUnAP4FwFNcc17X+DwAAAA
ASUVORK5CYII=
05/11/2011 10:35:54 Animation for iphone game | Elance Job https://www.elance.com/j/animation-iphone-game-animation-
cartoon/33478404/?utm_medium=partner&utm_source=jobrapido&utm_campaign=jobrapido&rid=1YX1Q

A8eZEzf8ks7f/li+eOc9ucIdD9Bx8/rrlTgN0K3ee0G5ZNftq48/qjx79knk/qsvEuj41atXPDgN0Knce1mrcs/VVatWMZMVBjqV+y6rV+y9jtcWvAZU7LysU7HrMlmaGRgYGM
wrNh1nDpv7Szpl6WO5tKUP0LFr07adeF1nWrD0JE/w5M9+TRs2BrZuWouMfRrWb86fua/v2bNnXLi9kDj5slbCpKsMDAyMZHlBJ6LtsnpoKwWBGFh1Wd23ggIDPHMu
63jmUhALfrnH2VU8v+u4pFzWd087jw33TF9ZiNOA9IoJ0839co/j0qzrlnahd/Zq3JmJHAAAWau8BZBk7+QAAAAASUVORK5CYII=
05/11/2011 10:35:43 oriflame est Annaba annaba http://www.ouedkniss.com/fr/annonces/details/?id=1888565

GgUFBALlUjC1QaQWeqPQlgKtLWyv0ALbbru75/HDLnKZ5M2ZM3PmPTlz5pWE0DEhADAQmN65XcwLDAx0DMMAQAKDMxIwTAoh/hdcJBBCYKCjC82KCySse
xE6oJkQjQvFl8mumgQqcAriELQsqGkQWXNPDRWoWtR1QAFOrJhqEmRBiVHP/6CRm0YUplDlCeqFJeAAdyLNp18xksc6ZWAmUeHlfIjXTi/LyQoSp6uo8hyqPEFDH
kWTv1H9O4SS6qeUX2RwxY/DkyKkwEoRni1E6ZmNMBbcJw9IQh6glv5INTsHWgT0OEbJRW3nOZnEW/rcHrrcGQbjGl3OTbpnwgyHspSBGiDpe93Ukg5E2QMcgChCI
4Cy2U4m1suL+QUeujI8mE3R5tzmlTdH2uya2UQ93kJ1uwuqq0DRalEUJWInFemkZ9rJPWeGu9O72KbitM1sERTmKQDJiN7keMOOyI0AYSABqouKv4ndcDudE5PcntzD
EarSNhXBNh6idzFO6ewJItZCJXCDk9B9OHRAtZ9avBVl7RrbgVY6xie5Nb7FSgXcObB99WMbC/IlKHMASJx+ILf+iNLaHUq+ZoprzRR81yl4mziUB3g6Mop95A/+vEY
RGAweYR/+zZPPS3hyOhLswPEyR/E37Ps6SHsfU4j0Us9+BzK8X/LQtxggWdFQgH1g6OcG71zrjPmTZ5NYAqKAz0IYkDGoUwHK1tRp1noCHFlVkqkBFahYv1A0fVED
DBpW1hAmdKBhjbU1ylhiMSxFnovoql1SqIV/xkqoj3zWJgQAAAAASUVORK5CYII=
05/11/2011 10:35:36 Offre d'emploi Responsable Commercial Algérie - Cveno.com http://www.cveno.com/annonce/detail/504f3041333f00.95308449

AkdRIQ2NDIFIcLUMLU6EYW1O59UJbtYZoCBSjIq1TFpTLEMAoISQsaAhigQokYGew7cww///93+VxMaNxYTjJuzs55108R4DBWo21lowUaEhKiYsPypLyJL5W+Cgk
BgwkrAZAWHzAZA+wcPncFayRoA3WSDQ+vnSxnsYCvu/zj4Q2Hv8NGYwnWTG3hN8v/wE++CqNRXE9nsQFrAQLYDOFIvVoEF+lMVYCBjchebdoJ2uXVvLgZizjtl
DTeYbOa1fhUabIaC8TkB66w92+LSCHSfkeFti24mtCRU0UTHuD+3f+xE0PM76qkdG7f+D0/SFwJUMYjDEIXA/VX8rD3tXEb0fQnmT9/EaCy/aS/9xi5jy/lLd2HsT57AhO
ywVGtkQoPX4KXINSCoFrGb62CntjCd3NLxL77QIbFjXwZmENhQVlTJ2ykunbIzzZepJRLSfJ/eIsuXs6CP/Sn/3ApLh76SW49DI9rYJw6UJKX61m2cwwr495m3GrqxnXfh
yx7yii9TSiKcrTbT8RaOqi/ddeRDLWS+qig+0ay9H6HNYsXMLaWUGWz3iPCQUlTPn0AKLhMM5XJxD7uxnRdp6R+3sQzT1E+wcQ+HGufzuJ5HeCb8ons2Z2BSVzy3kh
fx4Tq+sQtbsINB5CNERw2roItJxgRHMUp66TBCASSQ/+6uVGxzSOvf8MJbNWUVS4jfFlH5Ib2kKgfjdPbP+SvI9/JGfvEUR7J3n7IoSit0CD0FbiA6Tv0ffzJm4eeo2VC8qYs
Hw9YnMFo8I7EBsrCYTrEbUHyKltp/jg9yiZRRnSpPHQAHqIq8cW83lwHYULKphYXMuz5fU4Wz9BVFWRt2sPU8M1eIDOoi/QGdgUGgVIdwBMnNambl6ZX0Z+8Qc8
FQwhguU47xRzy02ClHg6G2CsBA1YBRY8L53ZhWuoC33EgIbZWytxghvpA0gMMpiI/bs9weOkFRqIKUXH2ej/Wv4GJJAhvumWKkMAAAAASUVORK5CYII=
05/11/2011 10:35:33 Emploinet http://www.emploinet.net/details.php?o=8256
euAo91ch9ZbhtZ1nN3D0en09vf+/v7+/v7////8/Pz6+vr+/v78/Pzz9fWfvOY7a80zZdgxZt0zZtkyZeE3ac6Rr+Dx8vP7+/v+/v7////+/v7+/v76+vq0zuc3at4zZtk2bNlijd1jj9w4btIzZ
t00Z92ju+Tz8fD+/v7////////////2+PlhmNcyZuc/ctK7zuXr7u7s7+/K2etKetgzZtdSiNro7O7+/v7////////////t9Pw/i88zZtqCqeDw8PH7/Pz8/f3y8/OfveMzaNY5gdDZ5e7+/v7//////////
/o8/48kswyadGlweD09PX+/v7+/v719/e61eszcNU3jMzO3+f+/v7////////////v9ftGnM8yd993rdzr7e/6+vv7/Pzv8fKPvOAzgN4/mtDg6fD+/v7////////////3+fl2tNstiug0l9mUwuTd6
O7i6e2pzOU8ktszlexkrNfu8fL+/v7////////////6+vrS3+lEodkll+wgl+g7nd47ntcjmM4emOowmOWZx+Py8vL+/v7////////////+/v78/P3H3upZqdormOonme0nme8omeoxmOYhmf
CAut7w8fH+/v7+/v7+/v7////////+/v75+fnp8/S52+yTxeCRwd+s0eexzuRzud1AndS81en09fX+/v7+/v7+/v7////+//7+/v7+/v78/Pz3+Pj29vb5+fn6+/vt8fKMwOE8ntd6vNumzufn
7O/+/v7/////////////////////+/v7+/v7+/v7+/v76+vrc5+4qnOIMmPkomemIvt/9/v7///////////////////////////////////////+/v7g6fAmnOkJmfgLmfpdrt79/v7//////////////////////////////////////+/v7z+Phquu
ENmOYWmuqXzOn+/v7//////////////////////////////////////+/v77/P3p8fW33ezI4O7w9fb+/v4AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
05/11/2011 10:34:42 Jobrapido | Emploi Annaba, recrutement
http://dz.jobrapido.com/?w=&l=annaba&r=auto&utm_source=jobalert&utm_medium=email&mid=20110915003828884803
NjY7c3N3//+qqqs0MzQ9PD1mZWb5+fn//5+fm1tbVISEhMS0zh4eH//h4eFJSUp
JSUnd3d3////46Nzy2cX9+/n///g4OBGRkdHRkfh4eL////stInnnmby3c3//g4OBCQkNFRUXm5ub////rsoXooWvx2MX////////////////////////////////
///////f398/PkBFREXX19f////ts4jmnmjy2cf///f3988PD1EQ0Tc3Nz////ssofim2Xy2sj///e3t85ODpAP0Hc3Nz////psIXfmGLx18T////////////////
////////////////////////e3t41NDY6Ojvc3Nz////lq4Dcl2Hlr4f67OH++/n/////////////////////////////////c3NwtLC8wLzHm5ub////fpHfamGTdmmfalV/hqHz/////////////////////////////t7e1KSUtMTE7w8O//////jsIr
dnm7y1sDcnWzenm3/////////////////////////////////////ExMTExMT//////////89vH68Oj////78+z78en/////////////////////////////j4+MtLC4uLjDi4uL///r6+tEQ0ZGRkjq6ur/
//y8vLz8/P///8AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAA
05/11/2011 10:34:39 http://dz.jobrapido.com/?j=jmvzt3fxverp8j8gm394vvpda25yc4z http://dz.jobrapido.com/?j=jmvzt3fxverp8j8gm394vvpda25yc4z

Figure A.1. Extrait de fichier log utilisé dans les expérimentations

DESCRIPTEURS DE DOMAINE

Domaine	#termes	URL
Informatique	1024	techterms.com
Biologie	785	didier-pol.net
Mathématiques	1424	bibmath.net
Sport	502	linternaute.com
Multimedias	326	fin.ucar.edu/it/mag/printgloss.htm webopedia.com/Multimedia
Infos	572	thenewsmanual.net/Resources/glossary.html

Tableau A.1. Descripteurs de domaine : Adresse des sites sources

ANNEXE B : IMPLÉMENTATION

Nous avons adopté le langage de programmation JAVA. Ce choix se justifie par la simplicité avec laquelle il permet d'analyser et de traiter efficacement du texte. En effet, Java met en œuvre plusieurs astuces de programmation qui facilite l'extraction des données d'historique. S'ajoute à cet argument, la richesse de ses bibliothèques, la portabilité et la fiabilité de ce langage.

Comme librairie de RNA, nous avons choisi *JOONE*[1] (Java Object Oriented Neural Engine) qui est un framework permettant de créer, entraîner et tester des réseaux neuronaux. Du fait qu'il est écrit en java, il peut donc être utilisé sur n'importe quel système disposant d'une machine virtuelle, il peut être aussi intégré à Eclipse. Il dispose d'un éditeur graphique *joonepad* qui permet de créer des PMCs et de les sauvegarder sous deux formats possibles snet/xml.

Pour le traitement du contenu de l'historique de navigation (pages Web et requêtes et URL), nous avons utilisé GATE (General Architecture for Text Engineering) ou l'Architecture générale pour le traitement de texte. GATE est une boîte à outils logicielle écrite en Java consacrée au traitement du langage naturel. Il offre en plus d'un environnement de programmation graphique, une interface de programmation d'applications (API).

GATE comporte un système d'extraction d'information, ANNIE (A Nearly-New Information Extraction System), un analyseur lexical, une base de toponymes (gazetteer), un analyseur syntaxique (segmentation de phrases, avec désambigüisation), un étiqueteur, un module d'extraction d'entités nommées et un module de détection de coréférences. Il est mis en œuvre pour plusieurs langues parmi lesquelles l'anglais le français et l'arabe. Il accepte en entrée divers formats de texte comme le texte brut, .HTML, .XML, .Doc, .PDF, etc. [WIK13]

[1] *http://sourceforge.net/projects/joone/*

Treetagger est un outil d'étiquetage communément employé dans le domaine du traitement du langage naturel, il qui a été utilisé au départ, avant de le remplacer par GATE qui offre plus d'outils y compris l'étiquetage.

Dans nos expérimentations, nous avons utilisé Google comme moteur de recherche et ce, grâce à l'API Google Search (Interface de Programmation d'Applications).

B.1. Identification

L'utilisateur soumet une requête initiale au système qui ensuite essaye de déduire le domaine cible en analysant les mots clés décrivant son besoin informationnel et tenant compte les informations correspondantes au contexte de la recherche.

Ensuite, le système retourne l'estimation du domaine cible de l'utilisateur qui sera jugé par lui-même comme étant soit bonne ou mauvaise estimation.

B.2. Recherche et reformulation

En fonction du domaine d'intérêt estimé, le système sélectionne un ensemble de termes contextuellement pertinents pour reformuler la requête initiale de l'utilisateur.

Dans les deux cas, les résultats sont jugés en donnant une estimation de pertinence (good ou bad) pour chaque page retournée. Les interfaces assurant cette évaluation sont représentées ci-dessous.

B.3. Interfaces

À titre d'expérimentation nous avons pu interroger le moteur de recherche Google grâce à son API (figure B.1), qui permet à l'utilisateur d'introduire sa requête et d'effectuer la recherche sur la base d'index de Google. Le résultat de la recherche est par la suite retourné (figure B.2) et jugé avant et après expansion (figure B.3). La figure (B.4) montre l'interface d'évaluation des résultats.

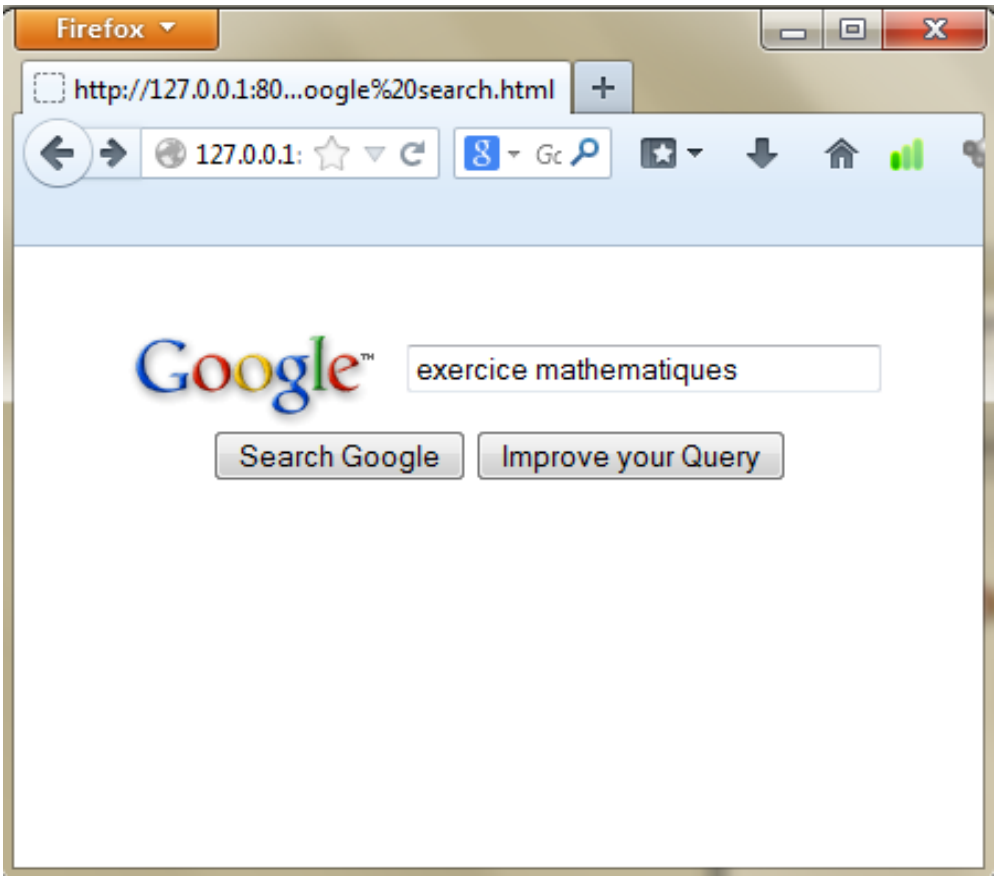

Figure B.1. Interface d'évaluation

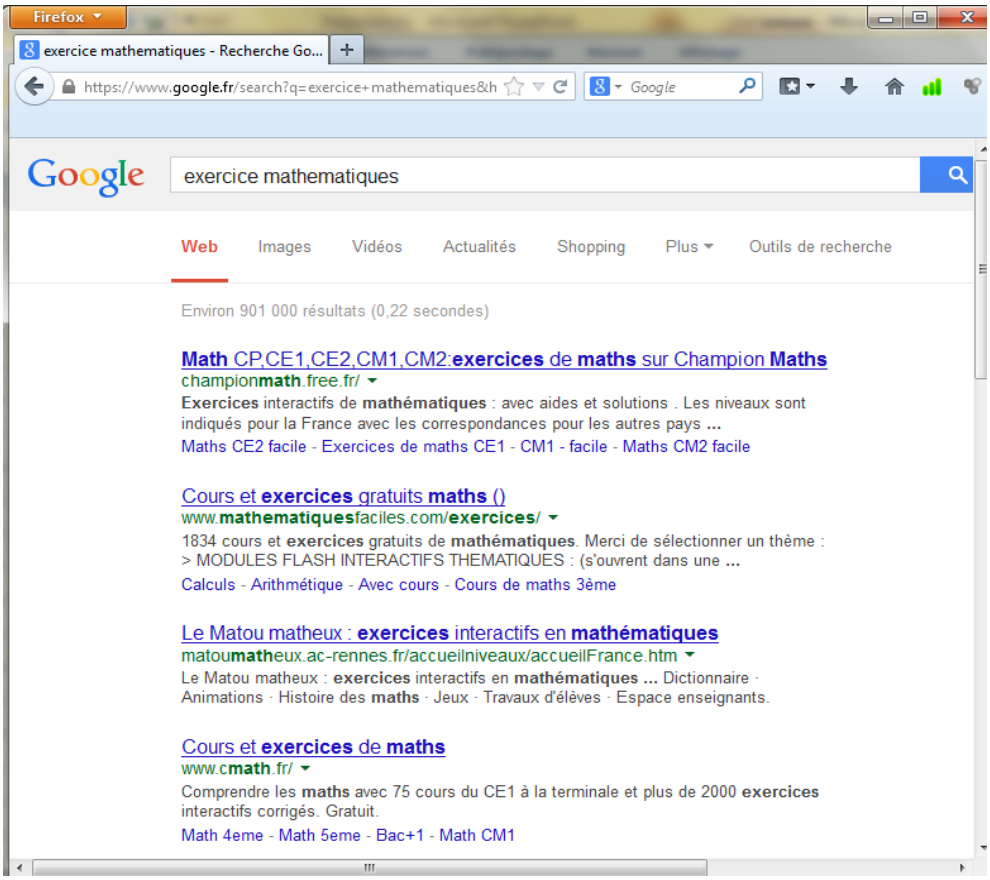

Figure B.2. Résultats de la recherche

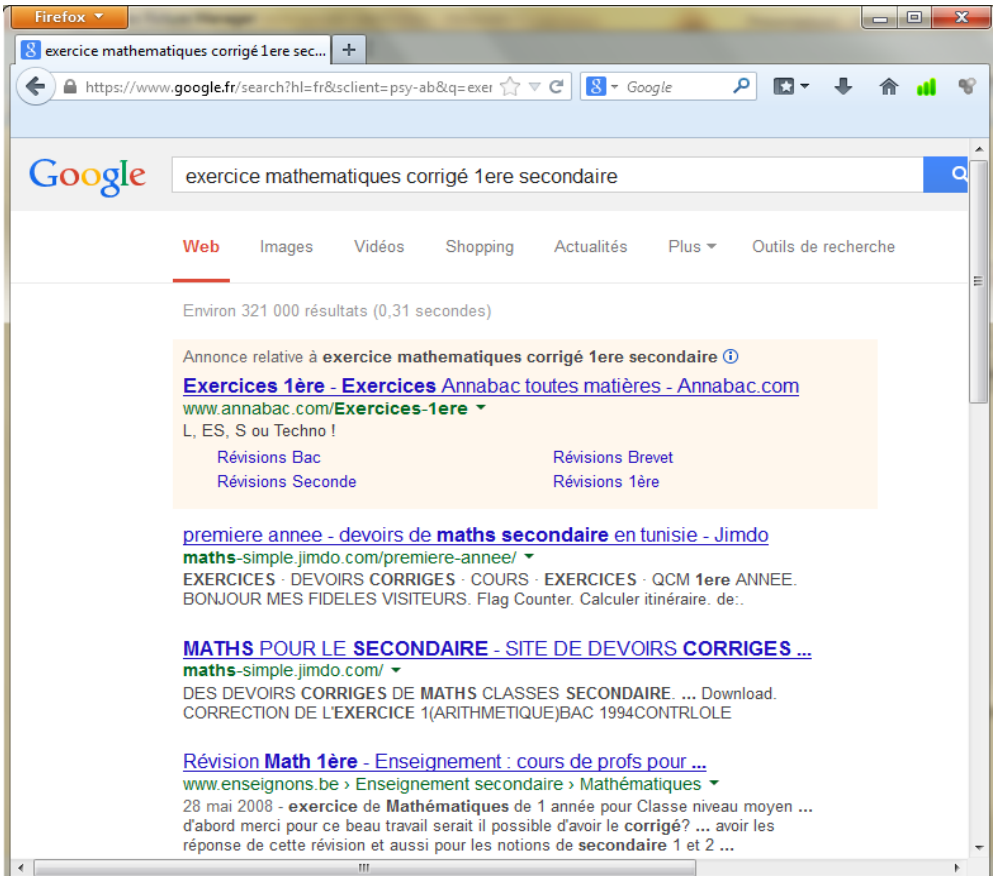

Figure B.3. Résultats de la recherche après expansion de requête

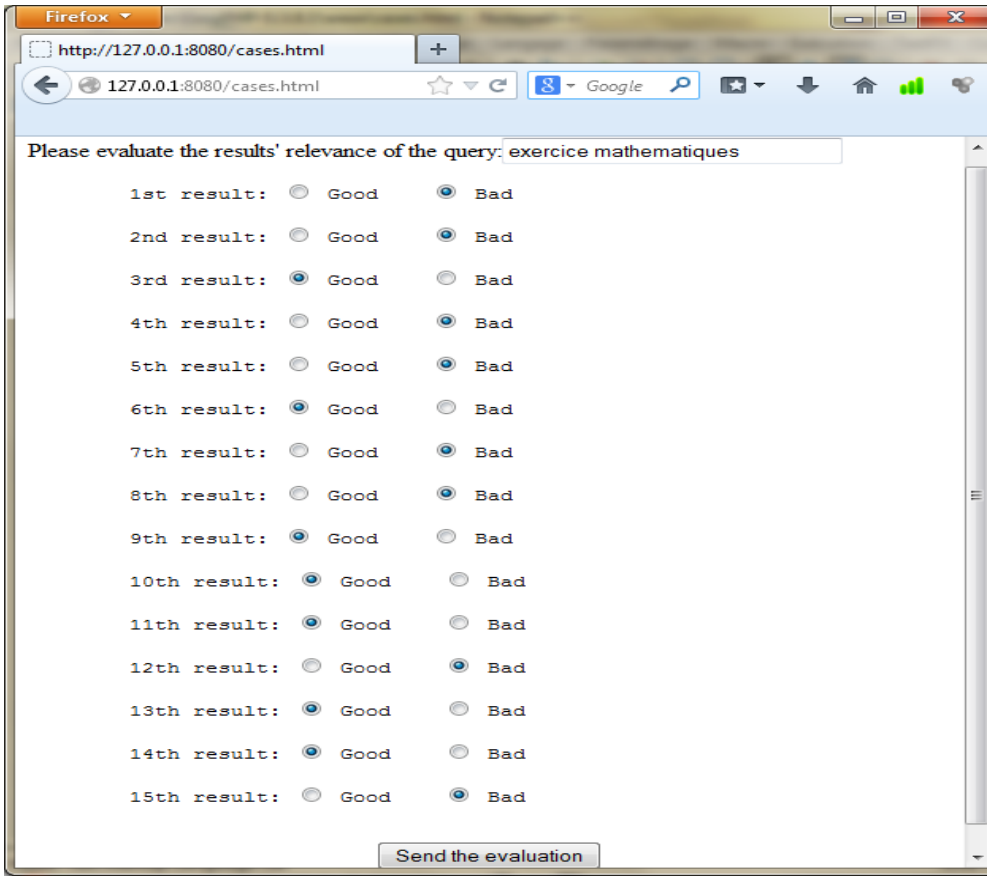

Figure B.4. Interface d'évaluation des résultats

ANNEXE C : INTRODUCTION AUX RÉSEAUX DE NEURONES ARTIFICIELS

C.1. LE NEURONE FORMEL

Un neurone formel ou artificiel est un processeur élémentaire. Comme le montre la figure (C.1), le neurone reçoit un nombre d'entrées en provenance de neurones amont. À chacune de ces entrées est associé un poids *w* (weight) qui représente la puissance de la connexion. Chaque neurone est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones avales.

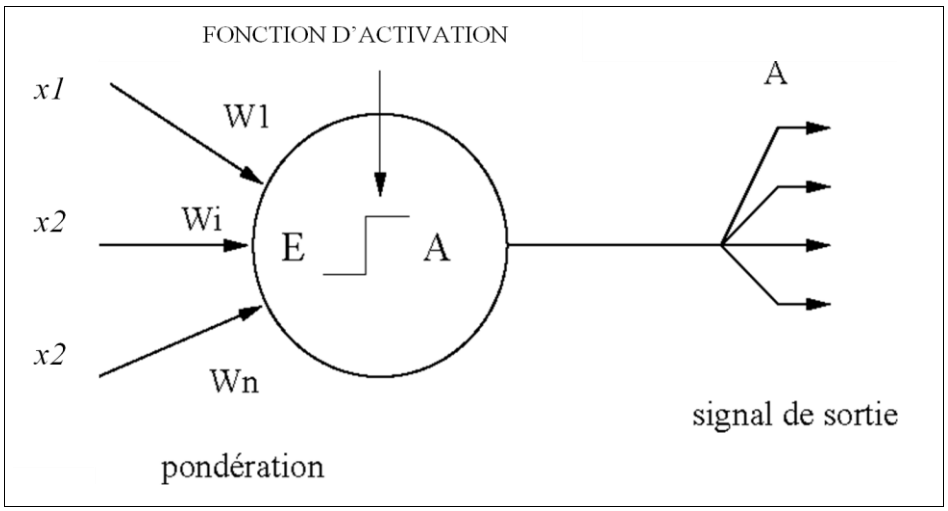

Figure C.1. Neurone formel.

C.2. DÉFINITION D'UN RNA

Selon [HAY94], un réseau de neurones est un processus distribué de manière massivement parallèle, qui a une propension naturelle à mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour utilisation. Il ressemble au cerveau en deux points:

1) La connaissance est acquise au travers d'un processus d'apprentissage;
2) Les poids des connexions entre les neurones sont utilisés pour mémoriser la connaissance.

De manière générale, un réseau de neurones comporte :

- Des neurones d'entrée, auxquels on attribue une excitation en fonction des données que le réseau doit traiter ;
- D'autres neurones au travers desquels l'excitation des neurones d'entrée se propage et est modifiée ;
- Des neurones de sortie dont l'état d'excitation fournit une réponse au problème posé en entrée, la figure C.2 ci-dessous montre l'architecture d'un RNA.

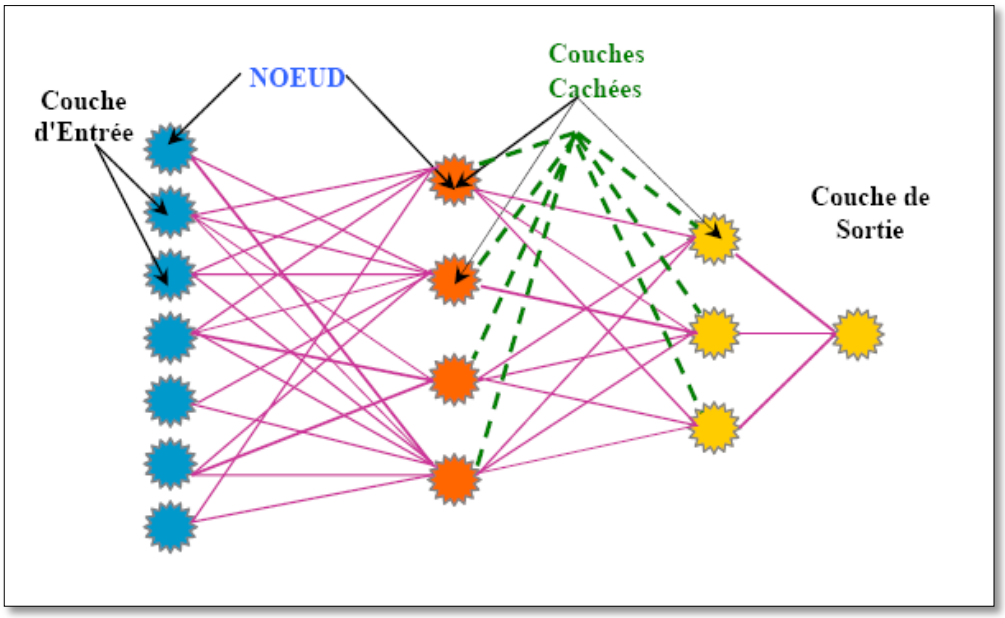

Figure C.2 : Architecture d'un réseau de neurones artificiels.

C.3. DIFFÉRENTES CONFIGURATIONS DES RÉSEAUX

Réseau multicouche où les neurones sont arrangés par couches. Il n'y a pas de connexion entre neurones d'une même couche et les connexions ne se font qu'avec les neurones des couches avales.

1) Réseau à connexions locales, qui s'agit d'une structure multicouche, ou chaque neurone est connecté à quelques neurones localisés dans son entourage. Les

connexions sont donc moins nombreuses que dans le cas d'un réseau multicouche classique.

2) Réseau à connexions récurrentes, les connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouche. Ces connexions sont le plus souvent locales.

3) Réseau à connexion complète, c'est la structure d'interconnexion la plus générale. Chaque neurone est connecté à tous les neurones du réseau (et à lui-même).

C.4. APPRENTISSAGE DES RNAs

Définition

L'apprentissage est un processus dynamique et itératif permettant de modifier l'efficacité synaptique, il se traduit par un changement dans la valeur des poids qui relient les neurones d'une couche à l'autre [PAR04].

Apprentissage supervisé / non supervisé

Les réseaux de neurones se divisent en deux principales classes, les réseaux à apprentissage supervisé (supervised learning) et les réseaux à apprentissage non supervisé (unsupervised learning).

Pour les réseaux à apprentissage supervisés (Perceptron, Adaline, etc.), on présente au réseau des entrées, et au même temps les sorties que l'on désirerait pour cette entrée. Par exemple on lui présente en entrée une lettre " a " manuscrite et en sortie un code correspondant à la lettre " a ".

Le réseau doit alors se reconfigurer, c'est-à-dire calculer ses poids afin que la sortie qu'il donne corresponde bien à la sortie désirée. Ce type de réseaux est généralement destiné à reproduire un processus quelconque (chimique, mécanique, financier...) dont on connaît seulement quelques variables et les résultats correspondants.

Pour les réseaux à apprentissage non supervisé (*Hopfield*, *Kohonen*, etc.), on présente une entrée au réseau et on le laisse évoluer librement jusqu'à ce qu'il se stabilise. Ils sont utilisés par exemple en classification lorsque les classes auxquelles doivent appartenir

les données ne sont pas connues à priori. Il existe aussi des réseaux à apprentissage dit semi-supervisé (reinforcement learning) qui ne tient compte que d'une évaluation partielle ou qualitative des sorties.

C.5. LES RÉSEAUX MULTICOUCHES ET LA RETRO-PROPAGATION DU GRADIENT

L'incapacité démontrée du perceptron à traiter les problèmes non linéaires a été résolue en associant des neurones dotés d'une fonction d'activation non linéaire en un réseau multicouches. Ce type de réseau est dans la famille générale des réseaux à «propagation vers l'avant», c'est-à-dire qu'en mode normal d'utilisation, l'information se propage dans un sens unique, des entrées vers les sorties sans aucune rétroaction. Son apprentissage est de type supervisé, par correction des erreurs.

Le perceptron multicouche est un des réseaux de neurones les plus utilisés pour des problèmes d'approximation, de classification et de prédiction. Il est habituellement constitué de deux ou trois couches de neurones totalement connectées. L'algorithme *de rétro-propagation*, nécessite que les fonctions d'activations des neurones soient continues et dérivables.

Voici les différentes étapes à suivre lors de l'apprentissage d'un réseau de neurones à propagation avant avec l'algorithme « Back propagation » [LER04].

1er étape : Initialiser les poids des liens entre les neurones. Souvent une valeur entre 0 et 1, déterminée aléatoirement, est assignée à chacun des poids.

2e étape : Application d'un vecteur entrées-sorties à apprendre.

3e étape : Calcul des sorties du RNA à partir des entrées qui lui sont appliquées et calcul de l'erreur entre ces sorties et les sorties idéales à apprendre.

4e étape : Correction des poids des liens entre les neurones de la couche de sortie et de la première couche cachée selon l'erreur présente en sortie.

5e étape : Propagation de l'erreur sur la couche précédente et correction des poids des liens entre les neurones de la couche cachée et ceux en entrées.

6e étape : Boucler à la 2e étape avec un nouveau vecteur d'entrées-sorties tant que les performances du RNA (erreur sur les sorties) ne sont pas satisfaisantes.

C.6. DOMAINES D'APPLICATION

Les RNAs offrent une panoplie de techniques pour la fouille de données et la reconnaissance des formes comme la classification, le clustering, la modélisation, la prévision, la prédiction…etc.

Concernant les problèmes de classification, les RNAs permettent de classifier les données élémentaires dans une ou plusieurs classes prédéfinies. Lorsque l'on cherche à résoudre un problème de classification, on a affaire à des formes (par exemple des chiffres manuscrits, des pièces mécaniques, des individus, etc.) susceptibles d'appartenir à des catégories, ou classes, différentes. Un classifieur est capable d'attribuer une classe à une forme inconnue qui lui est présentée, ou de refuser de lui attribuer une classe si la forme inconnue est trop éloignée des formes utilisées pour l'apprentissage.

Le prétraitement des données brutes consiste une étape primordiale permettant de déterminer la qualité et les performances d'un classifieur. Les réseaux de neurones peuvent offrir des performances équivalentes à celles des meilleurs classifieurs en matière de taux de reconnaissance. Ceci dépend du choix de données, leur représentation et leur traitement, sans oublier la conception du réseau lui-même.